TECHNIQUES D'ENQUÊTE

Nº 12-001-XPB au catalogue

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

JUIN 1999

•

VOLUME 25

•

NUMÉRO 1

Statistique Canada Statistics Canada

Canada

Ca 005

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010290891

TECHNIQUES D'ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

JUIN 1999 • VOLUME 25 • NUMÉRO 1

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, 1999

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 des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Septembre 1999

Nº 12-001-XPB au catalogue

Périodicité: semestrielle

ISSN 0714-0045

Ottawa

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. 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
D. Binder
G.J.C. Hole
F. Mayda (Directeur de la Production)
C. Patrick
R. Platek (Ancien président)
D. Roy
M.P. Singh

COMITÉ DE RÉDACTION

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

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*
D. Binder, *Statistique Canada*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistique Canada*
J. Eltinge, *Texas A&M University*
W.A. Fuller, *Iowa State University*
R.M. Groves, *University of Maryland*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *Central Statistical Office, U.K.*
G. Kalton, *Westat, Inc.*
R. Lachapelle, *Statistique Canada*
P. Lahiri, *University of Nebraska-Lincoln*
S. Linacre, *Australian Bureau of Statistics*
G. Nathan, *Central Bureau of Statistics, Israel*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Bell Communications Research, U.S.A.*
F.J. Scheuren, *Ernst and Young, LLP*
J. Sedransk, *Case Western Reserve University*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
R. Valliant, *Westat, Inc.*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints P. Dick, H. Mantel, B. Quenneville et D. Stukel, *Statistique Canada*

POLITIQUE DE RÉDACTION

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

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes des ménages, 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 Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 $ par année au Canada et de 47 $ US par année à l'extérieur du Canada. Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au (613) 951-7277 ou au 1 800 700-1033, par télécopieur au (613) 951-1584 ou au 1 800 889-9734 ou par Internet : order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research et la Société Statistique du Canada.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 25, numéro 1, juin 1999

TABLE DES MATIÈRES

Dans ce numéro 1

H. KRÖGER, C.-E. SÄRNDAL et I. TEIKARI
L'échantillonnage mixte de Poisson: une famille de plans permettant une sélection coordonnée à l'aide de nombres aléatoires permanents 3

W. R. BELL et M. KRAMER
Vers des variances pour la désaisonnalisation X-11 13

J. DE HAAN, E. OPPERDOES et C. M. SCHUT
Le choix des produits pour l'indice des prix à la consommation: le seuil d'inclusion par opposition au sondage probabiliste 33

P. DUCHESNE
Estimateurs de calage robustes 47

Y. TILLÉ
Estimation dans des enquêtes par sondage avec des probabilités d'inclusion conditionnelles : enquêtes à plan d'échantillonnage complexe 61

N.G.N. PRASAD et J.N.K. RAO
Estimation régionale robuste au moyen d'un modèle simple à effets aléatoires 73

F.A.S. MOURA et D. HOLT
Production d'estimations régionales à partir de modèles multiniveau 81

M. CHATTOPADHYAY, P. LAHIRI, M. LARSEN et J. REIMNITZ
Estimation composite de la prévalence des drogues pour des zones infraétats 91

S. RUBIN BLEUER et M. KOVAČEVIĆ
Étude de certains problèmes posés par l'estimation de la dynamique du revenu 99

P. F. TATE
Utilisation de données reliées longitudinalement de l'Enquête britannique sur la population active 113

S. GABLER, S. HAEDER et P. LAHIRI
Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe 119

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 1-2
Statistique Canada

Dans ce numéro

Chers lecteurs,
Chères lectrices,

Je m'empresse de partager avec vous deux bonnes nouvelles. Tout d'abord, le numéro de décembre marquera le 25e anniversaire de *Techniques d'enquête*. Ce sera un numéro un peu plus volumineux de la revue, comportant des articles d'éminents statisticiens contemporains. Deuxièmement, nous comptons préparer une version électronique de la revue. Nous espérons monter le numéro de décembre 1999 sur un site Web spécial. Tous les abonnés actuels pourront télécharger la revue gratuitement. Les résultats de cet essai nous permettront de déterminer s'il serait préférable d'offrir la version électronique au lieu de la version imprimée, ou en plus de celle-ci. Le prochain numéro vous fournira des renseignements supplémentaires. Vos remarques et vos suggestions sont toujours bien accueillies.

Le présent numéro aborde un choix de sujets : trois articles sur les estimations infrarégionales, quatre sur les estimations en général, deux sur les nouveaux plans d'échantillonnage et deux autres sur l'analyse des données.

Kröger, Särndal et Teikari proposent une nouvelle famille de plans d'échantillonnage, sous la manchette d'échantillonnage mixte de Poisson, qui correspond à un mélange pondéré de plans d'échantillonnage de Poisson et de Bernoulli. Grâce à une étude de Monte Carlo portant sur des données finlandaises, ils montrent empiriquement que, pour divers estimateurs ponctuels, l'échantillonnage mixte de Poisson est plus efficace que l'échantillonnage de Poisson ordinaire.

Bell et Kramer cherchent à résoudre le problème que pose depuis longtemps l'estimation de la variance des estimateurs X-11. Chaque mois, les bureaux de la statistiques du monde entier publient les estimations brutes de diverses variables, ainsi qu'une mesure correspondante de l'erreur, ordinairement l'erreur-type ou le coefficient de variation. Par contre, ils ne sont en mesure de produire aucune mesure de ce genre pour les estimations désaisonnalisées ou estimations de tendance correspondantes obtenues par la méthode X-11. Bell et Kramer proposent une méthode intéressante qui permet de résoudre pratiquement le problème. Ils calculent deux formes d'erreur, l'une résultant de l'erreur d'échantillonnage et l'autre, de l'utilisation d'extrapolations ARMMI pour prolonger la série aux deux extrémités.

De Haan, Opperdoes et Schut se penchent sur l'échantillonnage d'articles dans un groupe de produits, aux fins de leur inclusion dans le calcul de l'indice des prix à la consommation, d'après les données de lecteurs optiques. À l'heure actuelle, la plupart des bureaux de la statistique appliquent une méthode d'échantillonnage raisonné qui biaise naturellement les estimations. Les auteurs s'efforcent de déterminer si l'échantillonnage probabiliste serait une meilleure méthode en ce qui concerne l'erreur quadratique moyenne et obtiennent des résultats intéressants.

Pierre Duchesne considère une nouvelle classe d'estimateurs de calage robustes qui permet d'obtenir des poids restreints à un intervalle donné. La demarche consiste a transformer en poids cales des poids implicites robustes judicieusement choisis. Dans une brève étude empirique, les nouveaux estimateurs sont à la fois illustrés et comparés avec des estimateurs déjà proposés.

Tillé étudie une méthode d'échantillonnage répété qui tient compte de renseignements auxiliaires. Il commence par généraliser l'utilisation des probabilités conditionnelles d'inclusion de façon à pouvoir les appliquer à tout plan d'échantillonnage. Puis, il construit des estimateurs que l'on peut considérer comme étant des estimateurs linéaires optimaux et les compare à l'estimateur de régression généralisé (GREG). Il différencie les divers estimateurs grâce à un ensemble de simulations. Enfin, il discute du problème de l'interaction entre le plan de sondage et les variables auxiliaires.

Prasad et Rao étudient le problème de l'estimation régionale au moyen d'un modèle à effets aléatoires. Alors que les méthodes classiques d'estimation des moyennes régionales sont fondées sur un modèle, Prasad et Rao obtiennent des estimations fondées sur le plan de sondage (assistées par modèle) grâce à l'intégration des poids d'enquête. Ils calculent aussi les estimateurs à base de modèle correspondants de l'erreur quadratique moyenne (EQM) des estimations régionales. Ils montrent, grâce à une simulation, que leur estimateur de l'EQM est assez stable et n'est que légèrement biaisé.

Dans leur article sur l'estimation régionale, Moura et Holt se concentrent sur des modèles à plusieurs niveaux, qui tiennent compte de données auxiliaires au niveau tant de l'unité que de la petite région, ainsi que d'effets aléatoires régionaux en ce qui concerne aussi bien les coordonnées à l'origine que la pente des droites de régression. Ils estiment les paramètres des modèles à effets fixes et aléatoires par une méthode itérative généralisée des moindres carrés restreinte. Ils calculent une valeur approchée de l'erreur quadratique moyenne. Ils montrent, par des simulations, que le modèle permet de produire de meilleurs estimateurs régionaux que ceux fondés sur des modèles plus simples, qu'une spécification trop poussée du modèle n'entraîne aucune perte importante d'efficacité et que l'approximation de l'EQM et l'estimateur connexe de l'EQM donnent de bons résultats.

Chattopadhyay, Lahiri, Larsen et Reimnitz étudient l'estimation des proportions régionales d'événements rares. Ils se servent des données d'une enquête téléphonique sur la consommation d'alcool et de drogues pour illustrer leur méthode et la comparer à d'autres. L'estimateur qu'ils proposent combine des estimations démographiques par groupe âge/sexe/comté fondées sur des données de recensement et des estimations bayesiennes empiriques des proportions dans ces groupes fondées sur des données d'enquête. Ils proposent un estimateur jackknife de l'erreur quadratique moyenne qui tient compte de la variabilité due à l'estimation des paramètres du modèle.

Rubin Bleuer et Kovacevic étudient les problèmes que pose l'estimation des proportions longitudinales de faibles revenus d'après les données d'une enquête longitudinale dont le plan de sondage est complexe. Ils envisagent deux estimateurs fondés sur le plan de sondage : l'un s'appuyant à la fois sur les échantillons longitudinal et transversal, appelé « estimateur mixte » et l'autre, uniquement sur l'échantillon longitudinal. Grâce à une simulation, ils comparent les deux estimateurs en cas d'érosion de l'échantillon en appliquant des modèles de correction pour l'érosion fondés sur l'hypothèse, d'une part, que les unités d'échantillonnage « manquent au hasard » et, d'autre part, qu'elles « manquent complètement au hasard ». Ils illustrent les résultats au moyen de données de deux enquêtes longitudinales.

Tate étudie le couplage des données recueillies sur une même personne au cours de cycles trimestriels subséquents de la British Labour Force Survey, une enquête par panel avec renouvellement d'un cinquième de l'échantillon à chaque cycle. Elle examine les divers facteurs susceptibles de biaiser les analyses réalisées d'après des données couplées de ce genre. Plus précisément, elle étudie les effets éventuels de l'érosion de l'échantillon, des erreurs de réponse et de la réponse par procuration. Elle envisage aussi diverses méthodes de correction pour tenir compte de ces biais.

Enfin, dans une note brève, Gabler, Haeder et Lahiri présentent une justification fondée sur un modèle de la formule bien connue de l'effet de plan de sondage proposée par Kish. Ils montrent que la formule produit en réalité une valeur prudente de l'effet réel de plan de sondage.

Le rédacteur en chef

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 3-12
Statistique Canada

L'échantillonnage mixte de Poisson: une famille de plans permettant une sélection coordonnée à l'aide de nombres aléatoires permanents

HANNU KRÖGER, CARL-ERIK SÄRNDAL et ISMO TEIKARI[1]

RÉSUMÉ

Les auteurs présentent l'échantillonnage mixte de Poisson, qui correspond à une famille de plans d'échantillonnage tirant son nom du fait que chaque membre de la famille est un mélange de deux plans d'échantillonnage de Poisson, c'est-à-dire l'échantillonnage ptπ de Poisson et l'échantillonnage de Bernoulli. Ces deux plans se situent aux extrémités opposées d'un spectre continu indexé en fonction d'un paramètre continu. L'échantillonnage mixte de Poisson est destiné aux populations très asymétriques que l'on rencontre souvent dans les enquêtes-entreprises. Les statisticiens y trouvent un éventail d'options pour l'ampleur de la coordination des échantillons et le contrôle du fardeau de réponse. Certains plans d'échantillonnage mixte de Poisson donnent des estimations nettement plus précises que l'échantillonnage ptπ de Poisson habituel. Ce résultat a de l'importance car l'échantillonnage ptπ de Poisson est très efficace en soi, à supposer qu'il se fonde sur une bonne mesure de la taille.

MOTS CLÉS: Enquêtes-entreprises; population irrégulière; fardeau de réponse; estimateur de régression.

1. OBJECTIFS DE L'ÉCHANTILLONNAGE MIXTE DE POISSON

L'échantillonnage mixte de Poisson est une famille de plans d'échantillonnage qui se prête aux enquêtes-entreprises, avec leur population souvent très asymétrique. L'échantillonnage mixte de Poisson englobe les plans traditionnels d'échantillonnage de Bernoulli et d'échantillonnage ptπ de Poisson comme deux cas spéciaux qui se trouvent aux extrémités d'une gamme de possibilités indexée en fonction d'un paramètre continu. Ce paramètre, appelé largeur de Bernoulli et désigné B, satisfait $0 \le B \le f_R$, où f_R est la fraction d'échantillonnage prévue et préétablie de la portion «tirage partiel» de la population, c'est-à-dire la portion où la sélection aléatoire s'applique.

Les nombres aléatoires, sous forme de réalisations indépendantes de la variable aléatoire Unif (0,1), sont utilisés couramment dans les procédures automatisées modernes de sélection d'un échantillon. Fan, Muller et Rezucha (1962) ont introduit plusieurs mécanismes de tirage séquentiel (unité par unité) fondés sur des nombres aléatoires. À présent, l'échantillonnage mixte de Poisson se fonde sur la technique du nombre aléatoire permanent (NAP), qui suppose dès la création d'une entreprise, l'attribution d'un nombre aléatoire à chaque unité de la base de sondage (le registre des entreprises, dans le cas d'une enquête-entreprise). Le nombre aléatoire est permanent en ce sens qu'il reste associé à l'unité pour toute la durée de vie de celle-ci. La technique du NAP facilite la coordination des échantillons et le contrôle du fardeau de réponse. Parmi les premières mentions de l'échantillonnage axé sur le NAP, mentionnons Brewer, Early et Joyce (1972) et Atmer, Thulin et Bäcklund (1975). On trouvera dans Ohlsson (1995) un récent aperçu des diverses techniques NAP, avec d'importantes applications.

L'échantillonnage ptπ de Poisson offre l'avantage d'une sélection de grandes unités avec une probabilité relativement supérieure à celle des petites unités, dont la part du total estimatif de la population est de toute façon relativement mineure. La coordination des échantillons ptπ de Poisson à l'aide du NAP a été introduite par Brewer et coll. (1972) et elle a été discutée ensuite par plusieurs auteurs, y compris Sunter (1977) et Ohlsson (1995).

Comme pour l'échantillonnage ptπ de Poisson, l'échantillonnage mixte de Poisson permet de contrôler le fardeau de réponse (voir la section qui suit). Les grandes entreprises ont relativement plus de chances d'être choisies que les plus petites. Le contrôle de la sélection se fait par renouvellement de façon à répartir le fardeau de réponse. Un autre objectif de l'échantillonnage mixte de Poisson est d'inclure dans l'échantillon l'ensemble (ou une partie importante) des unités de la population (elles sont donc observées, de sorte que leurs données de base peuvent être mises à jour) de façon régulière au cours d'une certaine période de temps. Il peut s'agir, par exemple, d'inclure chaque entreprise dans l'échantillon au moins une fois durant une période de dix ou douze ans.

2. LA PROCÉDURE DE SÉLECTION LIÉE À L'ÉCHANTILLONNAGE MIXTE DE POISSON

Soit $U = \{1, ..., k, ..., N\}$, une population finie, où le nombre entier k représente la k-ième unité de la population. Soit y la variable d'intérêt et y_k sa valeur pour l'unité k; y_k est inconnue avant la sélection de l'échantillon et

[1] Hannu Kröger et Ismo Teikari, P.O. Box 3A, Fin-00022, Statistics Finland, Finland; Carl-Erik Särndal, 11ième étage, Immeuble R.H. Coats, Statistique Canada, Ottawa, (Ontario) K1A 0T6, Canada.

l'observation. À l'unité $k \in U$ est également associée une mesure positive connue de la taille x_k. Celle-ci a comme rôle dans l'échantillonnage mixte de Poisson d'entraîner une sélection plus fréquente des grandes unités; de plus, la variable de taille devrait servir de variable auxiliaire à l'étape de l'estimation.

Un échantillon s est tiré de la population U. La taille de s peut être aléatoire; sa taille prévue, désignée n, est un nombre établi d'avance. L'échantillon s est constitué de deux parties non chevauchantes, $s = s_C \cup s_R$, où s_C s'appelle la partie certitude de s et s_R la partie randomisation de s. La partie s_C, constituée de très grandes unités avec probabilité de sélection égale à 1 est désignée au cours d'une étape préliminaire, à l'aide des mesures connues de la taille x_k. On trouvera une procédure de ce type à la section 3. Selon les caractéristiques de la population, il peut arriver qu'aucune partie certitude ne soit désignée, de sorte que s_C est la série ensemble, mais cette possibilité est assez exceptionnelle pour la population très irrégulière que l'on trouve normalement dans les enquêtes-entreprises.

Un terme qui est très souvent utilisé comme synonyme de la partie certitude est la strate à tirage complet. Si la strate à tirage complet est désignée U_C, une description probabiliste consiste à affirmer que s_C est tiré de la strate à tirage complet U_C de sorte que $s_C = U_C$ suivant une probabilité de 1. La taille de $s_C = U_C$ est n_C.

Ensuite, l'échantillon de randomisation s_R est tiré du reste de la population $U_R = U - U_C$, de taille $N_R = N - n_C$. Il est constitué d'unités ayant une probabilité d'inclusion de π_k strictement inférieure à 1. Dans le présent exposé, s_R est tiré par échantillonnage mixte de Poisson (donc suivant la technique du NAP). La taille de s_R est aléatoire; sa taille prévue, désignée n_R, est établie en fonction de l'équation $n_R = n - n_C$.

Dans le présent exposé, l'échantillonnage de Poisson désigne une sélection correspondant à des essais de Bernoulli indépendants unité par unité avec des probabilités d'inclusion quelconques de π_k. Plus précisément, l'échantillonnage ptπ de Poisson signifie un échantillonnage de Poisson pour lequel π_k est directement proportionnel à une mesure de la taille. L'échantillonnage de Bernoulli est un cas spécial d'échantillonnage de Poisson dans lequel tous les π_k sont égaux.

Pour l'échantillonnage mixte de Poisson, il nous faut d'autres éléments de notation. Pour l'unité $k \in U_R$, définissons la mesure de la taille relative

$$A_k = n_R x_k \Big/ \sum_{U_R} x_k. \tag{2.1}$$

Nous pouvons désormais supposer que $A_k < 1$ pour tous les $k \in U_R$, parce que si $A_k < 1$ n'était pas vrai pour certaines unités $k \in U_R$, la procédure de la section 3 servant à l'établissement de la partie certitude de l'échantillon aurait en réalité attribué ces unités à la partie certitude s_C.

Nous définissons maintenant l'échantillonnage mixte de Poisson à l'aide d'un diagramme bidimensionnel. En abscisse, le NAP d'une unité est tracé. En ordonnée, une mesure Q_k, liée à la taille est tracée. À chaque cycle d'enquête, une nouvelle région de tirage d'un échantillon est désignée par renouvellement dans ce diagramme, et la coordination de l'échantillon est assurée comme suit.

L'échantillonnage mixte de Poisson est caractérisé par deux paramètres, B et f_R, où f_R tel que $0 < f_R = n_R/N_R < 1$ désigne le taux d'échantillonnage fixe prévu dans $U_R = U - U_C$. Le paramètre B, appelé la largeur de Bernoulli, est tel que $0 \le B \le f_R$. Pour chaque unité $k \in U_R$, définissons

$$Q_k = \frac{f_R - B}{1 - B} \frac{x_k}{\bar{x}_{U_R}} = \frac{1 - B/f_R}{1 - B} A_k \tag{2.2}$$

où $\bar{x}_{U_R} = \sum_{U_R} x_k / N_R$. Pour $B = 0$, nous avons $Q_k = A_k$, qui est la mesure de la taille pour l'échantillonnage ptπ de Poisson habituel. À l'autre extrême $B = f_R$, nous avons $Q_k = 0$ pour tous les $k \in U_R$; dans ce cas, la taille ne joue aucun rôle dans la sélection de U_R, qui se laisse réduire à l'échantillonnage de Bernoulli. Les mesures Q_k sont utilisées dans la sélection mixte de Poisson d'échantillons coordonnés, comme nous le décrivons ci-dessous.

Commençons par un tracé des points (r_k, Q_k) pour $k \in U_R$, où r_k désigne le NAP associé à l'unité k au moment de la création, et Q_k est donné d'après (2.2). Pour ce qui est de la figure 1, l'échantillonnage de Poisson se définit comme suit: inclure dans l'échantillonnage de randomisation, s_R, toutes les unités ayant des NAP r_k qui se trouvent dans l'intervalle (0, B], et inclure également certaines unités ayant des NAP r_k dans l'intervalle (B, 1], c'est-à-dire ceux pour lesquels Q_k est au moins égal à une valeur de seuil qui se trouve sur la ligne reliant les points (B, 0) et (1, 1). La zone de sélection est ainsi la partie ombragée de la figure 1. À noter que puisque $A_k < 1$ pour tous les $k \in U_R$, il s'ensuit d'après (2.2) que $Q_k < (1 - B/f_R)/(1 - B) \le 1$ pour tous les $k \in U_R$, lorsque $0 \le B < f_R$.

Pour l'échantillonnage ptπ de Poisson, la probabilité d'inclusion de l'unité k est $\pi_k = A_k$ d'après (2.1). La coordination des échantillons de Poisson fondés sur le NAP a été introduite par Brewer et coll. (1972) à l'aide d'une représentation graphique correspondant à $B = 0$ dans les figures 1 et 2. À chaque cycle, la zone de sélection est alors un triangle; le NAP de l'unité en abscisse est tracé en fonction de la mesure de la taille de l'unité, A_k, en ordonnée. La coordination est assurée par «le déplacement de la zone de sélection vers» la droite. La coordination des échantillons mixtes de Poisson est assurée de façon semblable.

Les figures 1 et 2 illustrent comment l'échantillonnage mixte de Poisson coordonné à partir de $U_R = U - U_C$ peut avoir lieu pour deux cycles d'enquête consécutifs. Dans chacune des deux figures, l'échantillon se définit comme la série d'unités pour laquelle le point (r_k, Q_k) se trouve à l'intérieur de la zone ombragée. Le «point de départ» sur

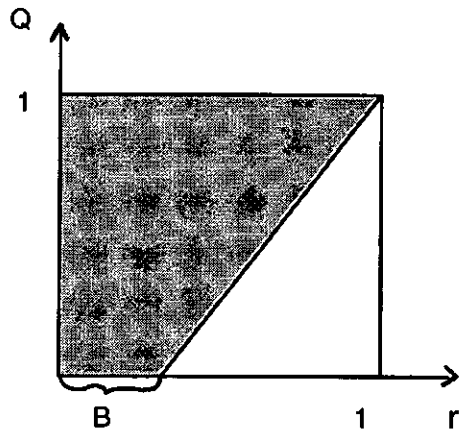


Figure 1. Échantillonnage au moment 1

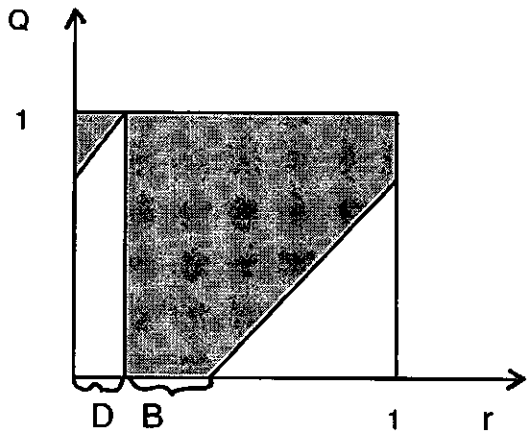


Figure 2. Échantillonnage au moment 2

l'axe du NAP est le point à la droite duquel nous commençons à compter les unités à inclure dans l'échantillon. Au moment 1 (figure 1), le point de départ est 0; au moment 2 (figure 2), le point de départ est D. (En général, le point de départ peut être un point choisi au hasard dans l'intervalle unitaire; autrement dit, l'échantillon identifié à la figure 2 est également celui qui serait choisi au moment $t = 1$ si à ce moment-là le point de départ choisi au hasard sur l'axe du NAP avait été égal à D.)

Une façon pratique d'assurer le renouvellement de l'échantillon est d'avoir recours à la méthode du déplacement constant, qui suppose que le point de départ est déplacé vers la droite suivant une valeur fixe à chaque nouveau cycle d'enquête (voir Ohlsson 1995). La constante D s'appelle le déplacement constant. Le point de départ au moment 3 serait donc $2D$, et ainsi de suite.

Nous examinons ci-dessous l'échantillonnage mixte de Poisson et l'estimation pour un même cycle, et nous pouvons nous concentrer sur la figure 1 (moment 1), avec le point de départ 0 sur l'axe du NAP. L'algorithme de l'échantillonnage mixte de Poisson pour les paramètres B et f_R, et le point de départ 0, est donc le suivant: à la figure 1, l'unité k est comprise dans la partie randomisation, s_R, (i) si $0 < r_k \le B$, ou (ii) si $B < r_k \le 1$ et $Q_k \ge (r_k - B)/(1 - B)$.

Par conséquent, k est compris dans s_R si

$$0 < r_k \le B + Q_k(1 - B).$$

Puisque $r_k \sim \text{Unif}(0,1)$, les probabilités d'inclusion de premier ordre dans le cadre de l'échantillonnage mixte de Poisson sont

$$\pi_k = \begin{cases} B + Q_k(1 - B) & \text{pour } k \in U_R \\ 1 & \text{pour } k \in U_C. \end{cases} \qquad (2.3)$$

Il est facile de voir que les probabilités d'inclusion répondent à l'exigence obligatoire que leur somme soit égale à la taille prévue de l'échantillon qui est fixée d'avance:

$$\sum_U \pi_k = \sum_{U_C} 1 + \sum_{U_R} \{B + Q_k(1 - B)\} = n_C + n_R = n.$$

Nous observons maintenant deux cas extrêmes de la famille de schémas d'échantillonnage mixte de Poisson: l'échantillonnage de Bernoulli est obtenu si $B = f_R$ dans l'algorithme mixte de Poisson, car alors $Q_k = 0$ pour tous les $k \in U_R$, et l'algorithme devient: inclure l'unité $k \in U_R$ dans s_R si $0 < r_k \le f_R$, ce qui représente l'échantillonnage de Bernoulli. L'échantillonnage ptπ de Poisson est obtenu si $B = 0$ dans l'algorithme mixte de Poisson, car alors l'algorithme devient: inclure l'unité $k \in U_R$ dans s_R si $0 < r_k \le A_k$, où A_k est donné d'après (2.1). Or cela représente l'échantillonnage ptπ de Poisson à partir de U_R, la probabilité d'inclusion étant de $\pi_k = A_k$, c'est-à-dire directement proportionnelle à la mesure de la taille x_k.

L'échantillonnage mixte de Poisson est un mélange d'échantillonnage ptπ de Poisson et d'échantillonnage de Bernoulli du fait que la probabilité d'inclusion mixte de Poisson, $\pi_k = B + Q_k(1 - B)$, est égale à une combinaison linéaire des probabilités d'inclusion qui s'appliquent pour les deux plans extrêmes, pondérée en fonction de la largeur de Bernoulli relative, $\lambda = B/f_R$, de sorte que $0 \le \lambda \le 1$. Nous avons $\pi_k = \lambda \pi_k^{\text{BE}} + (1 - \lambda)\pi_k^{\text{pt}\pi}$, où $\pi_k^{\text{BE}} = f_R$ pour tous les k (Bernoulli) et les $\pi_k^{\text{pt}\pi} = A_k$ (ptπ de Poisson).

Le caractère de l'échantillonnage mixte de Poisson est déterminé par ses deux paramètres B et f_R. À titre d'illustration, nous constatons à partir de (2.1) et (2.3) que la probabilité d'inclusion de l'unité $k \in U_R$ est $\pi_k = B + (1 - B/f_R)A_k$. Ainsi, pour une unité k qui est grande (mais pas assez grande pour répondre aux exigences pour $s_C = U_C$), de sorte que A_k se rapproche de 1, nous avons, en approximation étroite, $\pi_k = B + 1 - B/f_R$. Par contre, pour une unité qui est petite, de sorte que sa valeur A_k se trouve très près de zéro, nous avons, en approximation étroite, $\pi_k = B$, indépendamment de la taille. Ainsi, pour

f_R = 10 %, le tableau 1 ci-dessous indique comment la probabilité d'inclusion π_k varie selon B, où $B = 0$ est l'échantillonnage ptπ de Poisson et $B = f_R = 0.10$ est l'échantillonnage de Bernoulli.

Tableau 1
Valeurs de la probabilité d'inclusion π_k en fonction du paramètre B et de la taille relative A_k lorsque le taux d'échantillonnage fixe prévu f_R est de 0,1

	Valeurs de π_k				
Valeur de B	0	0,03	0,05	0,07	0,10
(petite unité)	0	0,03	0,05	0,07	0,10
(grande unité)	1	0,73	0,55	0,37	0,10

On voit donc que pour un plan d'échantillonnage mixte de Poisson se rapprochant de l'échantillonnage de Bernoulli (B près de 0,10), les probabilités d'inclusion des unités à la fois grandes et petites se situent près du taux d'échantillonnage fixe prévu de 0,10. Par contre, pour un plan mixte de Poisson se rapprochant de l'échantillonnage ptπ de Poisson (B près de 0), il est pratiquement certain qu'une petite unité ne fera pas partie de l'échantillon et qu'une grande unité en fera partie. Le tableau montre également comment l'échantillonnage mixte de Poisson comportant une valeur intermédiaire de B modifie les probabilités d'inclusion: les chances qu'une petite unité fera partie de l'échantillon diminuent quelque peu comparativement à l'échantillonnage de Bernoulli, et la sélection d'une grande unité devient moins probable que pour l'échantillonnage ptπ de Poisson.

Les répercussions pour ce qui est du fardeau de réponse sont les suivantes: le fardeau de réponse total à la population demeure le même pour toutes les valeurs de B; on demande toujours une déclaration à un total de $n = n_R + n_C$ unités. Comparativement à l'échantillonnage ptπ ($B = 0$) de Poisson, le fait d'établir une valeur B dans les limites de l'intervalle $[0, f_R]$ aura comme effet de déplacer une partie du fardeau de réponse des unités plus grandes aux unités plus petites; par le fait même, la précision des estimations est accrue dans de nombreux cas (voir les sections 5 et 6).

Enfin, nous devons mentionner les probabilités d'inclusion de second ordre dans le cadre de l'échantillonnage mixte de Poisson, car elles sont nécessaires pour le calcul de la variance en fonction du plan. Elles sont simples. Si π_{kl} désigne la probabilité que les unités k et l se trouvent toutes deux dans l'échantillon, nous avons

$$\pi_{kl} = \pi_k \pi_l \qquad (2.4)$$

pour $k \neq l \in U = U_R \cup U_C$, puisque les NAP r_k sont des réalisations indépendantes de la variable aléatoire Unif(0,1). Pour $k = l$, nous avons $\pi_{kl} = \pi_{kk} = \pi_k$. La fonction multiplicative (2.4) de π_{kl} simplifie énormément le calcul de la variance en fonction du plan. Nous obtenons un estimateur simple à somme unique de la variance, comme en (4.2) ci-dessous.

3. DÉTERMINATION DE LA PARTIE CERTITUDE DE L'ÉCHANTILLON

Si la population est extrêmement irrégulière, un ensemble d'unités (la partie certitude de l'échantillon, s_C) est tirée moyennant une probabilité de 1, et l'échantillonnage mixte de Poisson peut servir à la sélection randomisée pour le reste de la population, U_R. Plusieurs procédures peuvent être envisagées pour l'établissement de la série certitude; nous présentons ici une procédure qui est raisonnable (sans être nécessairement optimale) et qui sert à la simulation de Monte Carlo décrite à la section 5. La série certitude est désignée à l'aide des mesures positives connues de la taille x_k, par l'entremise de la procédure décrite ci-dessous, moyennant une ou plusieurs étapes. La taille d'échantillon prévue n est fixe pour l'ensemble de l'échantillon, $s = s_R \cup s_C$. Au cours de la première étape, les mesures relatives de la taille $A_{k(1)} = nx_k / \sum_U x_k$ sont calculées pour $k \in U$. Les unités k, s'il y en a, pour lesquelles $A_{k(1)} \geq 1$ sont attribuées à la partie certitude. Elles forment une série désignée $U_{C(1)}$; notons sa taille $n_{C(1)}$. La procédure est alors répétée afin que l'on puisse vérifier si d'autres unités devraient être attribuées à la partie certitude. La deuxième étape consiste à calculer les mesures relatives de la taille $A_{k(2)} = (n - n_{C(1)}) x_k / \sum x_k$ où la sommation s'étend à la série $U - U_{C(1)}$. Si $A_{k(2)} < 1$ pour tous les $k \in U - U_{C(1)}$ la procédure cesse et la partie certitude définitive est $s_C = U_{C(1)}$. Si toutefois $A_{k(2)} \geq 1$ pour certaines unités, celles-ci sont également attribuées à la partie certitude, et ainsi de suite jusqu'à ce qu'une étape i soit atteinte où toutes les mesures relatives intermédiaires de la taille $A_{k(i)}$ sont inférieures à 1. La partie certitude ultime s_C contiendra par exemple n_C unités et nous aurons alors $A_k < 1$ pour tous les $k \in U - s_C$, où $A_k = n_R x_k / \sum x_k$ avec $n_R = n - n_C$, et la somme s'étendra à $U - s_C$.

4. ESTIMATION SUBSÉQUENTE À L'ÉCHANTILLONNAGE MIXTE DE POISSON

Même si la variable auxiliaire joue un rôle utile à l'étape de l'échantillonnage, nous recommandons de l'utiliser également à l'étape de l'estimation. Pour estimer le total de la population, $Y = \sum_U y_k$, considérons l'estimateur de régression généralisée (REGG).

$$\hat{Y}_{\text{REGG}} = \sum_s a_k g_k y_k \qquad (4.1)$$

où $a_k = 1/\pi_k$ est le poids d'échantillonnage et le second poids, le poids g, est donné par

$$g_k = 1 + (X - \hat{X})' T_s^{-1} x_k / c_k$$

où x_k désigne la valeur vectorielle auxiliaire pour l'unité k, $\hat{X} = \sum_s a_k x_k$, et $T_s = \sum_s a_k x_k x_k' / c_k$. L'exigence imposée à l'information auxiliaire veut que le total vectoriel $X = \sum_U x_k$ soit connu d'une source fiable. La variable de

taille unidimensionnelle x_k servant à calculer le Q_k dans le schéma de sondage mixte de Poisson peut être une des composantes de $\boldsymbol{x}_k$ ou peut être une combinaison linéaire des composantes de $\boldsymbol{x}_k$. Dans l'étude empirique décrite à la section 5, $\boldsymbol{x}_k$ est unidimensionnel, et $\boldsymbol{x}_k = x_k$. Les constantes c_k, précisées par l'utilisateur, permettent la pondération des données, en plus des poids d'enquête a_k. Un choix répandu est $c_k = 1$ pour tous les k.

Un estimateur répandu de la variance de l'estimateur REGG (voir Särndal, Swensson et Wretman 1992, ch. 6) est donné sous forme quadratique dans $g_k e_k$, où e_k est le résidu de régression $e_k = y_k - \boldsymbol{x}_k'\hat{\boldsymbol{b}}$, où $\hat{\boldsymbol{b}} = \hat{\boldsymbol{T}}_s^{-1} \sum_s a_k \boldsymbol{x}_k y_k / c_k$. Un des avantages de l'échantillonnage mixte de Poisson, c'est que l'estimation de la variance correspondante est simple. Il en est ainsi parce que les NAP r_k sont des réalisations indépendantes de la distribution Unif(0,1). L'équation (2.4) s'applique donc et tous les termes produits sous forme quadratique égalent zéro. Puisqu'il ne reste que les termes mis au carré, l'estimateur de la variance devient simplement $\hat{V} = \sum_s a_k(a_k - 1) g_k^2 e_k^2$. Enfin, puisque $a_k - 1 = 0$ pour tous les $k \in s_C$, nous obtenons l'estimateur de la variance servant à l'étude de Monte Carlo décrite à la section 5, c'est-à-dire,

$$\hat{V} = \sum_{s_R} a_k(a_k - 1) g_k^2 e_k^2. \qquad (4.2)$$

5. UNE ÉTUDE DE MONTE CARLO DE L'ÉCHANTILLONNAGE MIXTE DE POISSON MENÉE À L'AIDE DE DONNÉES FINLANDAISES

Afin d'illustrer différents aspects de l'échantillonnage mixte de Poisson, nous avons mené une étude de Monte Carlo comportant quatre estimateurs du total Y de la population. L'expérience a supposé des tirages répétés d'échantillons aussi bien que des attributions répétées de la série de N NAP aux unités de la population. À noter que, puisque chaque attribution des N NAP aux unités de la population donne un résultat aléatoire, une étude Monte Carlo correcte exige également une répétition des attributions de NAP. Par conséquent, à la suite de la première attribution de N NAP, nous avons tiré 100 échantillons mixtes de Poisson à l'aide d'une valeur fixe de la largeur de Bernoulli B. (Chaque échantillon a été sélectionné en utilisant un nouveau point de départ sélectionné au hasard sur l'axe des NAP.) Ensuite, une nouvelle série de N NAP a été attribuée, 100 échantillons supplémentaires ont été tirés, et ainsi de suite, jusqu'à l'obtention de 100 × 100 = 10 000 paires NAP/échantillon, pour la valeur donnée de B. Pour chacune des 10 000 paires, nous avons calculé les quatre estimateurs ponctuels, les quatre estimateurs correspondants de la variance et les quatre intervalles de confiance correspondants. Vu les 10 000 répétitions, nous nous attendons à ce que l'erreur de Monte Carlo soit plutôt faible.

Les quatre estimateurs utilisés dans l'étude de Monte Carlo comportent les expressions ci-dessous, où $a_k = 1/\pi_k$ est le poids d'échantillonnage de l'unité k, et π_k est donné d'après (2.3):

1. L'estimateur de Horvitz-Thompson,

$$\hat{Y}_1 = \sum_s a_k y_k = \sum_{U_C} y_k + \sum_{s_R} a_k y_k.$$

2. L'estimateur de ratio (combiné),

$$\hat{Y}_2 = X \hat{b}_2$$

où $X = \sum_U x_k$ et $\hat{b}_2 = \sum_s a_k y_k / \sum_s a_k x_k$. C'est un cas spécial de (4.1) tel que $\boldsymbol{x}_k = x_k = c_k$.

3. L'estimateur REGG

$$\hat{Y}_3 = \sum_{U_C} y_k + \sum_{s_R} a_k y_k + \hat{b}_3 \left(X_R - \sum_{s_R} a_k x_k \right)$$

où $X_R = \sum_{U_R} x_k$ et $\hat{b}_3 = \sum_{s_R} a_k(a_k - 1) y_k x_k / \sum_{s_R} a_k(a_k - 1) x_k^2$. C'est un ca spécial de (4.1) tel que $\boldsymbol{x}_k = x_k$ et $c_k = (a_k - 1)^{-1}$.

4. L'estimateur de ratio (distinct),

$$\hat{Y}_4 = \sum_{U_C} y_k + X_R \hat{b}_4$$

où $X_R = \sum_{U_R} x_k$ et $\hat{b}_4 = \sum_{s_R} a_k y_k / \sum_{s_R} a_k x_k$.

Pour l'échantillonnage ptπ de Poisson (B = 0), nous avons $\hat{b}_4 = (\sum_{s_R} y_k / x_k) / n_{s_R}$, où n_{s_R} est la taille aléatoire de s_R; l'estimateur correspondant $\hat{Y}_4$ a été considéré par Brewer et coll. (1972). Maintenant $\hat{Y}_2$ et $\hat{Y}_4$ diffèrent en autant que la courbe de régression est calculée en fonction de l'échantillon groupé s, mais dans $\hat{Y}_4$ la courbe est calculée séparément pour l'échantillon de randomisation s_R. Enfin, $\hat{Y}_3$ diffère de $\hat{Y}_2$ et de $\hat{Y}_4$ en autant qu'il fait appel à la pondération $a_k(a_k - 1)$, et non simplement à a_k. À noter que tous les $\hat{Y}_2$, $\hat{Y}_3$ et les $\hat{Y}_4$ sont membres de la famille REGG d'estimateurs fournis par (4.1). En assimilant $\hat{Y}_2$, $\hat{Y}_3$ et $\hat{Y}_4$ à (4.1), nous trouvons les poids g supposés par chacun des trois estimateurs. Ces poids sont nécessaires pour l'estimation de la variance. Nous pouvons nous attendre à ce que la simulation montre que $\hat{Y}_2$, $\hat{Y}_3$ et $\hat{Y}_4$, qui utilisent la variable auxiliaire tant à l'étape du plan de sondage qu'à l'étape de l'estimation, permettront d'améliorer l'estimateur HT $\hat{Y}_1$, (comporteront une variance inférieure à la sienne), cet estimateur utilisant l'information auxiliaire uniquement à l'étape de l'échantillonnage; l'ampleur de l'amélioration est imprévisible, mais il est intéressant de l'observer.

Nous avons utilisé une véritable population de données pour la simulation de Monte Carlo. Cette population est constituée de $N = 1\,000$ entreprises finlandaises. Pour une entreprise k, $k = 1, ..., 1{,}000$, y_k est le nombre d'employés (équivalents à temps plein) × 10, et x_k est le salaire payé par l'entreprise à ses employés en milliers de MFI (marks finlandais). L'information auxiliaire (salaire payé) provient

du registre TVA des autorités fiscales de la Finlande. La variable de l'emploi est celle qui exige une estimation. Les 1 000 unités ont été tirées (d'une façon essentiellement aléatoire) d'une population initiale plus grande d'entreprises finlandaises. Les unités comportant une valeur de zéro pour y_k ou pour x_k, ont été éliminées afin que les résultats de la simulation ne soient pas perturbés par des facteurs externes. Par conséquent, comme pour les valeurs y_k et x_k, la population utilisée dans la simulation est naturelle, mais à cause de l'élimination d'unités, ses propriétés (moyenne, écart-type, asymétrie, etc.) diffèrent de celles de la population initiale plus grande.

Le total y de la population à estimer est $Y = \sum_U y_k = 169, 168$. La taille prévue de l'échantillon total $s = s_C \cup s_R$, a été fixée à $n = 100$. La procédure décrite à la section 3 a permis de déterminer la partie certitude s_C de l'échantillon. Il en est résulté une partie certitude s_C constituée des $29 = s_C$ unités les plus grandes. Le reste de la population, $U_R = U - s_C$ possède les caractéristiques descriptives suivantes: la taille est de $N_R = 1\,000 - 29 = 971$; le total de y est $Y_R = 46\,138$ (c'est-à-dire 27% du total de la population $Y = 169\,168$); le coefficient de variation (écart-type divisé par la moyenne) est de 1,78 pour la variable y et de 1,94 pour la variable x; le coefficient de corrélation entre x et y est de 0,965. La partie randomisation s_R, de taille prévue $n_R = 100 - 29 = 71$ unités, est mise en oeuvre dans la simulation par sélection répétée de l'échantillon mixte de Poisson à partir de U_R. On trouvera à l'annexe un tracé de (x_k, y_k) pour les unités k dans U_R.

Afin de constater l'effet de la largeur de Bernoulli, nous avons effectué la simulation pour un éventail de valeurs B: $B = 0, 0{,}01, ..., 0{,}07$, et, de plus, $B = f_R = n_R/N_R = 71/971 = 0{,}073$ (d'où l'échantillonnage de Bernoulli). Pour chaque valeur de B, $100 \times 100 = 10\,000$ paires NAP/échantillon ont été réalisées, et les résultats ont servi à calculer, pour chacun des quatre estimateurs ponctuels, cinq statistiques sommaires de Monte Carlo. Ce sont les suivantes, si $\hat{Y}$ désigne un des quatre estimateurs ponctuels, $\hat{V}$ l'estimateur de variance correspondant obtenu de (4.2) et $(\hat{Y} - z_{1-\alpha/2}\sqrt{\hat{V}}, \hat{Y} + z_{1-\alpha/2}\sqrt{\hat{V}})$ l'intervalle de confiance correspondant pour Y au niveau de confiance nominal $1 - \alpha$, où $z_{1-\alpha/2}$ est la valeur pour la distribution, $z_{1-\alpha/2} = 1\,960$ pour $\alpha = 5\%$, et $z_{1-\alpha/2} = 1{,}645$ pour $\alpha = 10\%$:

(1) MCE $\hat{Y}$ = l'espérance de Monte Carlo de l'estimateur ponctuel, c'est-à-dire la moyenne arithmétique des 10 000 estimations ponctuelles;

(2) MCV $\hat{Y}$ = la variance de Monte Carlo de l'estimateur ponctuel, c'est-à-dire la variance des 10 000 estimateurs ponctuels;

(3) MCE $\hat{V}$ = l'espérance de Monte Carlo de l'estimateur de la variance, c'est-à-dire la moyenne arithmétique des 10 000 estimations de la variance;

(4) MCRTE95 = le taux de couverture de Monte Carlo des intervalles de confiance nominaux de 95%, c'est-à-dire le nombre de fois que le paramètre cible Y se trouve dans l'intervalle de confiance, divisé par 10 000 et exprimé en pourcentage;

(5) MCRTE90 = le taux de couverture de Monte Carlo des intervalles de confiance nominaux de 90%; sa définition est analogue à celle de MCRTE95.

On trouve les résultats de la simulation dans le tableau 2 (taille moyenne de l'échantillon, variance de Monte Carlo; espérance de Monte Carlo de l'estimateur de la variance) et le tableau 3 (taux de couverture de Monte Carlo). Les tableaux n'indiquent pas l'espérance de Monte Carlo $\hat{Y}$, puisque dans tous les cas cette quantité se rapproche beaucoup de la valeur du paramètre cible $Y = 169, 168$, confirmant que tous les estimateurs sont essentiellement non biaisés. L'écart de l'espérance de Monte Carlo $\hat{Y}$ par rapport à Y est inférieur dans tous les cas à 0,14%, et dans la plupart des cas nettement inférieur. On constate que la taille moyenne de l'échantillon pour les 10 000 répétitions se rapproche beaucoup de $n = 100$, comme il se doit.

Tableau 2

Résultats de l'étude de simulation pour différentes largeurs de Bernoulli B: taille moyenne de l'échantillon, MCV $\hat{Y}$ et MCE $\hat{V}$; est. j renvoie à l'estimateur $\hat{Y}_j$; $j = 1, ..., 4$. Les valeurs des huit dernières colonnes sont à multiplier par 10^6.)

Bernoulli largeur de B	Taille moyenne de l'échantillon	MCV $\hat{Y}$				MCE $\hat{V}$			
		est. 1	est. 2	est. 3	est. 4	est. 1	est. 2	est. 3	est. 4
0,000	99,95	24,56	3,63	3,43	3,46	24,92	3,67	3,43	3,46
0,010	100,05	22,74	1,96	1,84	1,85	23,53	1,98	1,86	1,87
0,020	100,04	24,75	1,82	1,77	1,78	25,37	1,85	1,77	1,78
0,025	100,09	25,51	1,82	1,78	1,79	26,86	1,87	1,81	1,82
0,030	100,06	28,03	1,83	1,80	1,81	28,58	1,91	1,87	1,88
0,040	99,86	35,17	2,01	2,03	2,06	33,54	2,11	2,11	2,15
0,050	100,02	42,25	2,56	2,64	2,67	41,42	2,48	2,51	2,59
0,060	99,99	56,08	3,42	3,65	3,67	55,70	3,20	3,24	3,44
0,070	100,05	90,73	4,80	5,47	5,59	91,28	4,89	4,72	5,37
0,073	100,02	119,13	6,06	7,09	7,43	116,27	6,00	5,34	6,49

Tableau 3

Résultats de l'étude de simulation pour différentes largeurs de Bernoulli B: taux de couverture en pourcentage des intervalles de confiance nominaux de 95% et de 90%, MCRTE95 et MCRTE90; est. j renvoie à l'estimateur; $\hat{Y}_j$; $j = 1, ..., 4$

Bernoulli	niveau de confiance nominal de 95%				niveau de confiance nominal de 90%			
largeur B	est. 1	est. 2	est. 3	est. 4	est. 1	est. 2	est. 3	est. 4
0,000	94,50	92,03	92,75	92,48	89,70	86,36	87,35	86,74
0,010	95,20	93,13	93,47	93,52	90,43	88,06	87,93	88,02
0,020	95,06	94,23	93,88	93,88	90,36	89,36	88,49	88,55
0,025	95,06	93,73	94,56	94,70	90,64	89,15	89,73	89,72
0,030	94,63	94,12	94,09	94,19	89,85	89,06	88,70	88,86
0,040	93,84	94,44	94,47	94,64	88,77	89,60	89,41	89,60
0,050	93,97	94,38	93,76	93,82	88,67	88,67	88,08	88,53
0,060	93,54	93,57	92,12	92,69	89,10	87,57	85,99	87,27
0,070	92,93	94,99	90,67	92,03	88,40	88,62	84,27	86,11
0,073	91,03	95,02	88,03	90,46	86,53	88,62	81,26	83,86

Les tableaux 2 et 3 donnent lieu aux remarques suivantes:

1. Entamons l'examen du tableau 2 en comparant les variances de Monte Carlo d'un estimateur à l'autre, pour une largeur de Bernoulli B fixe. On constate que, pour chaque valeur de B, il n'y a guère de choix entre $\hat{Y}_2$, $\hat{Y}_3$ et $\hat{Y}_4$, relativement à la variance. Par contre, l'estimateur HT comporte une variance nettement supérieure. À titre d'exemple, le ratio MCV $\hat{Y}_3$/MCV $\hat{Y}_1$ est égal à 3,43/24,56 = 0,140 pour $B = 0$ (ptπ de Poisson), 1,80/28,03 = 0,064 pour $B = 0,03$, et 7,09/119,13 = 0,060 pour $B = 0,073$ (Bernoulli). Cela confirme que l'estimateur HT est un choix médiocre comparativement à l'utilisation de la variable auxiliaire fortement corrélée. Cela s'applique à l'échantillonnage de Bernoulli, ce qui n'a rien de surprenant, mais aussi aux valeurs B se rapprochant de l'extrémité inférieure de l'intervalle $[0, f_R]$, ce qui montre que le plan de sondage n'assure pas à lui seul l'extraction de toute la puissance de la variable auxiliaire, même si avec B pratiquement nul, nous nous rapprochons d'une sélection ptπ rigoureuse (et donc supposément très efficace). La variance comparativement grande de l'estimateur HT s'explique en partie par le fait que le caractère aléatoire de la taille de l'échantillon dans le cadre de l'échantillonnage mixte de Poisson pénalise l'estimateur HT (mais non pas les estimateurs REGG). Puisque l'estimateur HT est inefficace, nous n'en parlons pas davantage.

2. Lorsque nous examinons les petites différentes entre $\hat{Y}_2$, $\hat{Y}_3$ et $\hat{Y}_4$, nous remarquons au tableau 2: Mesuré en fonction de la variance de Monte Carlo, $\hat{Y}_3$ est meilleur que $\hat{Y}_4$ pour toutes les largeurs de Bernoulli B, mais de façon marginale seulement. De plus, $\hat{Y}_3$ et $\hat{Y}_4$ sont meilleurs que $\hat{Y}_2$ à l'extrémité inférieure de la fourchette de valeurs B, possiblement parce que dans $\hat{Y}_2$ nous permettons à la partie certitude de l'échantillon de contribuer à l'estimation de la courbe, de façon quelque peu inappropriée, puisqu'il n'existe un problème d'estimation que pour la partie randomisation. Or à la limite supérieure, la relation est inversée et, pour l'extrémité supérieure $B = 0,073$ (Bernoulli), $\hat{Y}_2$ clairement meilleure que $\hat{Y}_3$. Que les différences entre $\hat{Y}_2$, $\hat{Y}_3$ et $\hat{Y}_4$ soient si faibles n'a rien de surprenant car elles sont toutes des variétés de l'estimateur REGG faisant appel essentiellement à la même information auxiliaire.

3. Le tableau 2 confirme que l'estimateur de la variance $\hat{V}$ proposé fonctionne bien, comme nous nous y attendions; l'espérance de Monte Carlo $\hat{V}$ se rapproche énormément, à peu d'exceptions près, de la valeur cible que $\hat{V}$ tente d'estimer, c'est-à-dire la variance de $\hat{Y}$, mesurée ici en fonction de la variance de Monte Carlo $\hat{Y}$. Cela est vrai pour tous les estimateurs et toutes les valeurs de B, à quelques exceptions remarquables près, c'est-à-dire dans le cas de $\hat{Y}_3$ et de $\hat{Y}_4$ lorsque B se rapproche de l'extrémité supérieure (Bernoulli). L'estimateur de la variance sous-estime alors la variance.

4. Le résultat le plus intéressant du tableau 2 est, à notre avis, le fait que la variance de $\hat{Y}_2$ ou de $\hat{Y}_3$ ou encore de $\hat{Y}_4$, considérée en fonction de la largeur de Bernoulli B, n'atteint pas son minimum pour $B = 0$ (ptπ de Poisson), comme on aurait pu s'y attendre au début, mais plutôt pour une valeur de B qui se situe entre 0,02 et 0,03. De plus, l'amélioration du cas $B = 0,02$ par rapport au cas $B = 0$ est appréciable pour tous les $\hat{Y}_2$, les $\hat{Y}_3$ et les $\hat{Y}_4$. Si nous mesurons cette amélioration à l'aide de la variance de Monte Carlo($\hat{Y}|B = 0,02$) divisée par la variance de Monte Carlo ($\hat{Y}|B = 0$), nous trouvons que ce ratio est de 50% environ seulement pour tous les $\hat{Y}_2$, les $\hat{Y}_3$ et les $\hat{Y}_4$. Plus précisément, pour $\hat{Y}_2$ le ratio est de 1,82/3,63 = 0,501, pour $\hat{Y}_3$ il est de 1,77/3,43 = 0,516, et pour $\hat{Y}_4$ il est de 1,78/3,46 = 0,514. Compte tenu de ces résultats, nous avons ajouté une simulation pour $B = 0,025$, valeur qui n'a pas été examinée lors de

la série originale de simulations. Les résultats, que l'on trouve également au tableau 1, confirment qu'une variance minimale est obtenue, pour les trois estimateurs, $\hat{Y}_2$, $\hat{Y}_3$ et $\hat{Y}_4$, à un point qui se trouve près de $B = 0{,}025$. Une raison qui pourrait expliquer pourquoi il est nettement préférable de considérer B comme une valeur franchement supérieure à $B = 0$, ce qui donne l'échantillonnage ptπ de Poisson, se formule comme suit: lorsque B est égal à 0 (ou très près de 0), les unités comportant les valeurs x les plus petites, lorsqu'elles sont choisies, comportent des poids démesurément grands, ce qui entraîne une forte variabilité. On peut l'éviter en choisissant B nettement éloigné de zéro.

5. Les résultats de Monte Carlo du tableau 3 concernant les taux de couverture montrent que l'estimation de la variance et la procédure d'intervalle de confiance fonctionnent de façon satisfaisante. Comme la théorie nous le laisse entendre, MCRTE95 et MCRTE90 se rapprochent, pour les quatre estimateurs, de leurs valeurs théoriques, 95% et 90% respectivement. C'est uniquement pour $\hat{Y}_3$ et pour $\hat{Y}_4$, lorsque B se rapproche de l'extrémité supérieure (Bernoulli), que nous remarquons nettement une baisse de MCRTE sous la valeur nominale, en partie à cause de la sous-estimation de la variance mentionnée antérieurement.

6. AUTRE INDICATION QUE L'ÉCHANTILLONNAGE MIXTE DE POISSON EST PLUS EFFICACE QUE L'ÉCHANTILLONNAGE ptπ DE POISSON

Initialement, nous n'avions aucune raison particulière de croire que l'échantillonnage mixte de Poisson combiné à un estimateur REGG serait plus efficace pour certaines largeurs de Bernoulli B à l'intérieur de $[0, f_R]$ que pour l'échantillonnage ptπ de Poisson (B = 0). La forte amélioration – une réduction de la variance de 50% environ pour notre population particulière – a été assez surprenante. Pour d'autres populations, la réduction de la variance peut être égale ou inférieure aux 50% que nous avons observés. Puisque notre résultat dépend des données, il est souhaitable de fournir un argument plus général à l'appui de l'affirmation que l'échantillonnage mixte de Poisson comportant une valeur B nettement à l'intérieur de $[0, f_R]$ est meilleur que l'échantillonnage ptπ de Poisson ($B = 0$). Nous allons maintenant présenter un tel argument.

Nous avons examiné la variance de Taylor de $\hat{Y}$, c'est-à-dire la variance de la statistique linéarisée de Taylor. Elle est donnée par (voir Särndal et coll. 1992, ch. 6) $V_{TAY} = \sum_{U_R}(a_k - 1)E_k^2$, où E_k est l'analogue pour la population du résidu e_k fondé sur l'échantillon qui est utilisé dans l'estimateur de la variance (4.2). Pour l'estimateur $\hat{Y}_3$, par exemple, le résidu en question est $E_k = y_k - b_3 x_k$ avec $b_3 = \sum_{U_R}(a_k - 1)y_k x_k / \sum_{U_R}(a_k - 1)x_k^2$; pour $\hat{Y}_4$, $b_4 = \sum_{U_R} y_k / \sum_{U_R} x_k$ remplace b_3.

Il est raisonnable de modéliser le résidu mis au carré comme $E_k^2 = \sigma^2 x_k^p (1 + \delta_k)$, où p satisfait $0 \le p \le 2$, et δ_k se rapproche de zéro. Cela correspond à l'hypothèse d'un modèle de superpopulation $y_k = \boldsymbol{x}_k' \boldsymbol{\beta} + \varepsilon_k$, où les ε_k sont des erreurs indépendantes avec une valeur modèle prévue de zéro pour chaque k, et ε_k comporte la variance modèle $\sigma^2 x_k^p$. Pour l'approximation $E_k^2 \approx \sigma^2 x_k^p$, et pour $a_k = 1/\pi_k$ avec π_k donné par (2.3), nous obtenons

$$V_{TAY} \approx \sigma^2 \sum_{U_R}(a_k - 1)x_k^p = \sigma^2\{H(B,p) - T(p)\}$$

où $H(B,p) = \bar{x}_{U_R} \sum_{U_R} x_k^p [B\bar{x}_{U_R} + (f_R - B)x_k]^{-1}$ et $T(p) = \sum_{U_R} x_k^p$. Considérons maintenant une valeur fixe de p telle que $0 \le p \le 2$. Nous voulons savoir si $H(B, p)$ comporte une valeur plus petite pour certains B à l'intérieur de l'intervalle $[0, f_R]$, comparativement à sa valeur pour $B = 0$, qui représente $H(0, p)$. À cette fin, voyons si la dérivée $H'(B,p) = \partial H(B,p)/\partial B$ est négative pour $B = 0$. Nous trouvons

$$H'(B,p) = \bar{x}_{U_R} \sum_{U_R} x_k^p (x_k - \bar{x}_{U_R}) \left[B\bar{x}_{U_R} + (f_R - B)x_k\right]^{-2}.$$

Sa valeur pour $B = 0$ est de $H'(0,p) = (\bar{x}_{U_R}/f_R^2) \sum_{U_R} x_k^{p-2} (x_k - \bar{x}_{U_R})$. Le signe de $H'(0,p)$ est le même que celui de $\sum_{U_R} x_k^{p-2}(x_k - \bar{x}_{U_R})$. Toutefois, cette quantité correspond, à part le facteur $1/(N_R - 1)$, à la covariance dans U_R entre x_k^{p-2} et $x_k - \bar{x}_{U_R}$ (à noter que $x_k - \bar{x}_{U_R}$ comporte une moyenne nulle). Lorsque p satisfait $0 \le p \le 2$, cette covariance est négative: lorsque x_k augmente, $x_k - \bar{x}_{U_R}$ augmente progressivement, et x_k^{p-2} diminue progressivement (et demeure toujours positif). Le signe de $H'(0,p)$ est donc négatif; par conséquent, ce n'est pas pour $B = 0$ que $H(B,p)$ atteint sa valeur minimale, mais pour un B quelconque à l'intérieur de $[0, f_R]$. Pour $p = 2$, $H'(0,p) = 0$, et $H(B,p)$ atteint son minimum pour $B = 0$.

Compte tenu de ces considérations, on peut se demander si la population utilisée pour la simulation de la section 5 correspond à une valeur de $p \in [0,2]$, mais distinctement inférieure à 2, de façon que l'on puisse s'attendre à des avantages appréciables de l'échantillonnage mixte de Poisson. Pour trouver une réponse, nous avons estimé p en ajustant la version logarithmique du modèle $E_k^2 = \sigma^2 x_k^p (1 + \delta_k)$ aux données disponibles pour $U_R = U - U_C$. Autrement dit, nous avons ajusté $w_k = a + pz_k$, où $w_k = \log E_k^2$; $E_k = y_k - b_4 x_k$ avec $b_4 = \sum_{U_R} y_k / \sum_{U_R} x_k$; $z_k = \log x_k$, et a est un terme de la coordonnée à l'origine. Nous avons obtenu la valeur $p = 1{,}45$, en traitant p comme une courbe de régression linéaire estimée comme $p = \sum_{U_R}(w_k - \bar{w}_{U_R})(z_k - \bar{z}_{U_R}) / \sum_{U_R}(z_k - \bar{z}_{U_R})^2$. Puisque cette valeur p est nettement inférieure à 2, notre population de Monte Carlo est effectivement une population dont nous pouvons attendre des avantages appréciables pour l'échantillonnage mixte de Poisson comportant une valeur de B qui se trouve à l'intérieur de $[0, f_R]$.

7. DISCUSSION ET RECOMMANDATIONS PROVISOIRES

L'échantillonneur peut se demander: si je compte utiliser l'échantillonnage mixte de Poisson pour mon enquête, jumelé à un estimateur REGG, quel serait un choix approprié pour B? Rappelons que, dans le présent exposé, nous avons observé pour une population particulière une efficacité nettement accrue (une réduction de la variance de 50% environ comparativement à l'échantillonnage ptπ de Poisson) lorsque le paramètre mixte de Poisson B est ajusté à 30% de f_R environ. Nous avons cru savoir que le gain de variance se rapporte aux caractéristiques de l'erreur résiduelle, ce qui a été confirmé à la section 6 par la constatation que, lorsque la configuration résiduelle mise au carré se conforme à $E_k^2 = \sigma^2 x_k^p$, où p satisfait $0 \leq p < 2$, comme c'est le cas dans de nombreuses populations d'enquête-entreprise, l'échantillonnage mixte de Poisson avec B dans les limites de l'intervalle $[0, f_R]$ peut être fort avantageux. Toutefois, le présent exposé n'aborde pas la question du choix optimal de B. Dans la pratique, la valeur de B doit être ajustée à l'étape du plan de sondage, et le choix optimal de B dépend de caractéristiques inconnues de la population. Une connaissance préalable de la population, en particulier pour ce qui est de la structure de sa variance résiduelle, peut guider le choix de B.

Les recommandations provisoires du présent exposé sont les suivantes: Si l'information préalable suggère une configuration résiduelle mise au carré se conformant à $\sigma^2 x_k^p$ avec $p < 2$, on utilise l'échantillonnage mixte de Poisson avec $B = 0{,}3 f_R$. Par contre, si en réalité le p inconnu est tel que $p > 2$, même si le meilleur choix dans ce cas pourrait être $B = 0$ (échantillonnage ptπ de Poisson), il serait probablement peu risqué d'utiliser $B = 0{,}3 f_R$, car il est probable que la variance considérée comme une fonction de B augmente tout doucement. Par conséquent, $B = 0{,}3 f_R$ semble représenter une suggestion générale raisonnable. Ces recommandations sont provisoires; la question mérite d'être étudiée plus à fond que ne le permet le cadre du présent exposé.

ANNEXE

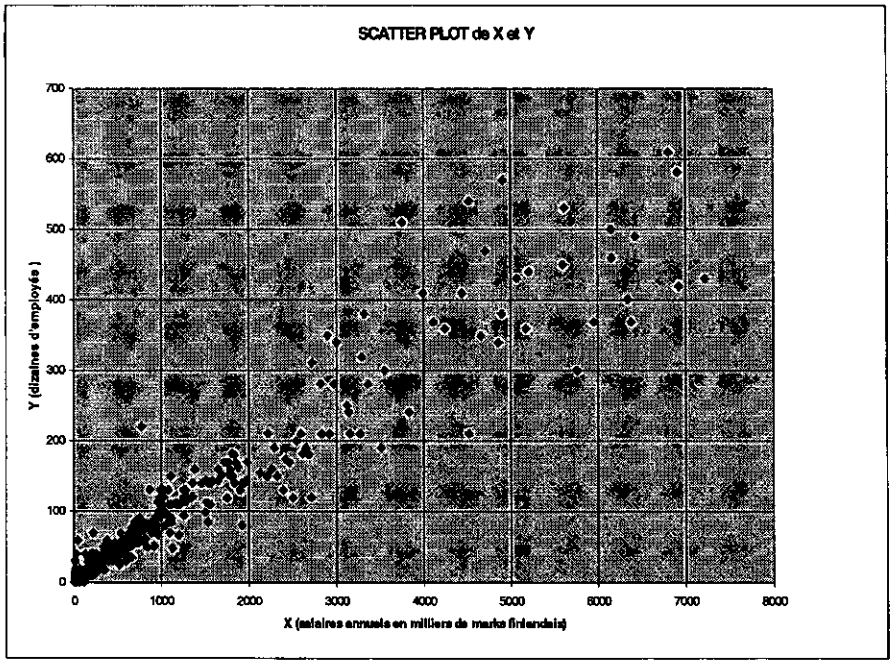


Figure 3. Diagramme de dispersion de x (salaires annuels) par rapport à y (emploi) pour la partie U_R de la population de Monte Carlo de 1 000 entreprises finlandaises

REMERCIEMENTS

Les auteurs tiennent à remercier l'examinateur de son travail, ses suggestions ayant donné lieu à de précieuses améliorations du manuscrit original.

BIBLIOGRAPHIE

ATMER, J., THULIN, G., et BÄCKLUND, S. (1975). Coordination of samples with the JALES technique. *Statistisk Tidskrift*, 13, 443-450.

BREWER, K.R.W., EARLY, L.J., et JOYCE, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 231-239.

FAN, C.T., MULLER, M.E., et REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) techniques and digital computers. *Journal of the American Statistical Association*, 57, 387-402.

OHLSSON, E. (1995). Coordination of samples using permanent random numbers. Dans *Business Survey Methods* (Éds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, et P.S. Kott) New York: Wiley, 153-169.

SÄRNDAL, C.E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SUNTER, A.B. (1977). Response burden, sample rotation, and classification renewal in economic surveys. *Revue Internationale de Statistiques*, 45, 209-222.

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 13-32
Statistique Canada

Vers des variances pour la désaisonnalisation X-11

WILLIAM R. BELL et MATTHEW KRAMER[1]

RÉSUMÉ

Nous élaborons une méthode qui permet d'estimer la variance dans le cas de la désaisonnalisation effectuée à l'aide de la méthode X-11 et qui prend en compte les effets de l'erreur d'échantillonnage et des erreurs liées à l'extension de prévision. Dans notre méthode, l'erreur de désaisonnalisation présente dans les valeurs centrales d'une série chronologique suffisamment longue est causée uniquement par les effets de l'application de filtres X-11 aux erreurs d'échantillonnage. Vers les deux extrémités d'une série, nous prenons également en considération la «contribution» des erreurs d'extrapolation rétrospective et de prévision à l'erreur de désaisonnalisation. Nous élargissons notre méthode afin d'obtenir la variance d'erreurs contenues dans des estimations de tendance X-11 et de prendre en compte l'erreur d'estimation de coefficients de régression utilisés pour modéliser, par exemple, des effets de calendrier. Dans le cas de nos résultats empiriques, la contribution la plus importante à la variance de la désaisonnalisation provenait souvent de l'erreur d'échantillonnage. Cependant, la variance des estimations de tendance augmentait de façon importante aux extrémités des séries, en raison des effets d'erreurs d'extrapolation rétrospective et de prévision. En outre, des composantes non stationnaires des erreurs d'échantillonnage produisaient des patrons singuliers dans la variance de la désaisonnalisation et de l'estimation de tendance.

MOTS CLÉS: Erreur d'échantillonnage; erreur de prévision; jour commercial; modèle ARIMA.

1. INTRODUCTION

La difficulté liée à la façon d'obtenir des variances dans le cas de données désaisonnalisées n'est pas un sujet nouveau (President's Committee to Appraise Employment and Unemployment Statistics 1962). Des méthodes de désaisonnalisation fondées sur des modèles (voir Bell et Hillmer 1984) font appel aux résultats de la théorie de l'extraction du signal pour produire des estimations et les variances d'erreur correspondantes des composantes saisonnières et non saisonnières. Cependant, la plupart des corrections officielles visant à tenir compte des variations saisonnières sont fondées sur des méthodes empiriques, notamment les méthodes X-11 (Shishkin, Young et Musgrave 1967) ou X-11-ARIMA (Dagum 1975). Ces méthodes s'appuient sur des filtres fixes plutôt que sur des modèles; la façon de calculer la variance des erreurs de désaisonnalisation n'est donc pas évidente. Diverses méthodes d'obtention de variances dans le cas de la désaisonnalisation X-11 ont été proposées; elles sont résumées ci-après.

Wolter et Monsour (1981) ont proposé deux méthodes. Ces auteurs ont reconnu qu'un grand nombre de séries chronologiques désaisonnalisées sont des estimations relatives à des enquêtes sur échantillon successives et donc susceptibles d'être entachées d'erreurs d'échantillonnage. La première méthode ne rend compte que de l'effet de l'erreur d'échantillonnage sur la variance associée à la désaisonnalisation. Dans la deuxième méthode, on essaie également de refléter, dans la variance de désaisonnalisation, l'incertitude due à la variation stochastique des séries chronologiques. Toutefois, dans cette deuxième méthode, on présuppose qu'abstraction faite des termes de régression, la série chronologique est stationnaire. Ce type de modèle est maintenant rarement utilisé dans le cas de séries chronologiques saisonnières. Cette méthode comporte également une erreur conceptuelle: elle permet d'obtenir la variance des estimations désaisonnalisées plutôt que la variance recherchée de l'erreur que contiennent ces mêmes estimations désaisonnalisées.

Burridge et Wallis (1985) ont étudié l'utilisation du filtre statique de Kalman pour le calcul des variances de désaisonnalisation fondé sur un modèle et ont appliqué cette méthode à un modèle qu'ils avaient obtenu antérieurement (Burridge et Wallis 1984) pour l'approximation des filtres X-11. Selon ces auteurs, cette méthode pourrait être utilisée pour «obtenir des mesures de la variabilité de la méthode X-11 lorsque celle-ci est appliquée à des données pour lesquelles elle est optimale» (p. 551), mais ils émettent des réserves quant à l'utilisation de cette méthode dans les cas où le filtre X-11 est sous-optimal (c'est-à-dire très différent du filtre fondé sur un modèle optimal). Hausman et Watson (1985) ont proposé une méthode pour estimer l'erreur quadratique moyenne dans le cas de X-11 lorsque cette méthode est utilisée dans des situations sous-optimales. Bell et Hillmer (1984, section 4.3.4) ont attiré l'attention sur une difficulté, liée à l'utilisation d'approximations fondées sur un modèle dans le cas de la méthode X-11, en ce qui a trait au calcul de la variance de la désaisonnalisation. Cette difficulté est que les filtres X-11 (ou, d'ailleurs, tout autre filtre de correction) ne suffisent pas à déterminer de manière unique des modèles pour la série observée et ses composantes.

[1] William R. Bell et Matthew Kramer, Statistical Research Division, Room 3000-4, U.S. Bureau of the Census, Washington, D.C., 20233-9100, U.S.A.

Pfeffermann (1994) a mis au point une méthode qui tient compte de la contribution de l'erreur d'échantillonnage et de la variation irrégulière (variation des séries chronologiques dans la composante irrégulière) à la variance de la désaisonnalisation X-11. Les caractéristiques de l'erreur combinée (erreur d'échantillonnage et variation irrégulière) sont estimées à l'aide de la variation irrégulière estimée de X-11. Ces caractéristiques sont ensuite utilisées pour estimer deux types de variance de désaisonnalisation. Un inconvénient de cette méthode est que celle-ci est fondée sur l'hypothèse que le filtre X-11 supprime la composante saisonnière et reproduit la composante de tendance (voir Pfeffermann 1994, (p. 90), analyse portant sur l'équation (2.7).) Le non-respect de cette présupposition dans la pratique compromet cette méthode à un degré qui paraît difficile à évaluer. Cette hypothèse nous semble donc très douteuse et, dans tout cas particulier, également non vérifiable. Un deuxième inconvénient est que l'un des types de variance qui est proposé par Pfeffermann présuppose que l'on prenne comme estimation de la tendance la série désaisonnalisée X-11 plutôt que l'estimation de la tendance. Breidt (1992) et Pfeffermann, ainsi que Morry et Wong (1993) ont développé la méthode générale proposée par Pfeffermann.

Le but de la présente communication est l'élaboration et l'application d'une méthode permettant d'obtenir la variance de la désaisonnalisation X-11 dans le cas de deux types d'erreur. Le premier type est l'erreur d'échantillonnage; le second, l'erreur qui résulte de la nécessité d'élargir la série chronologique avec des prévisions et des extrapolations rétrospectives, avant l'application des filtres X-11 symétriques. Ce deuxième type d'erreur mène à des révisions des désaisonnalisations (Pierce 1980). Il est à noter que les révisions finissent par disparaître à mesure qu'un nombre suffisant de données peuvent être obtenues au-delà du point temporel qui fait l'objet de la correction. Il faut noter également qu'une série désaisonnalisée ne contiendra pas d'erreur d'échantillonnage si la série non désaisonnalisée correspondante n'en contient pas non plus. C'est le cas notamment de certaines séries chronologiques économiques (par exemple, les statistiques sur les exportations et les importations de la plupart des pays).

Dans notre méthode, on présuppose que l'objectif de la désaisonnalisation effectuée à l'aide de la méthode X-11 (ce que nous supposons être l'estimation visée par l'application de X-11) est ce qui résulterait de l'application du filtre X-11 linéaire symétrique (sans qu'il soit nécessaire d'ajouter des prévisions et des extrapolations rétrospectives) si la série chronologique ne contenait pas d'erreur d'échantillonnage. Bien que cette définition de l'objectif puisse être critiquée parce qu'elle ne tient pas compte de la variation de la série chronologique dans les composantes saisonnière et non saisonnière sous-jacentes, nous sommes d'avis qu'elle pourrait convenir à des utilisateurs types de données désaisonnalisée à l'aide de la méthode X-11. La préoccupation la plus probable de ce genre d'utilisateurs a pour objet l'incertitude reflétée dans les écarts existant entre les corrections initiales et les corrections finales, c'est-à-dire les révisions. Certains de ces utilisateurs sont également conscients du fait que les séries non corrigées consistent en des estimations fondées sur des échantillons des populations réelles sous-jacentes, et se rendront compte que les effets d'erreurs d'échantillonnage sur les corrections devraient également être reflétés dans la variance de la désaisonnalisation.

Notre méthode est fondée sur l'utilisation des filtres X-11 linéaires symétriques. Nous présumons que les filtres symétriques sont appliqués aux séries élargies à des prévisions et des extrapolations rétrospectives comportant un minimum d'erreurs quadratiques moyennes. Dans la pratique, les prévisions et les extrapolations rétrospectives sont obtenues à partir d'un modèle de série chronologique ajusté. Cette façon de procéder est dans l'esprit de la méthode X-11-ARIMA de Dagum (1975), mais avec un ajout complet de prévisions et d'extrapolations rétrospectives, comme le recommandent Geweke (1978), Pierce (1980), ainsi que Bobbitt et Otto (1990). Nos résultats s'appliquent directement à l'utilisation de X-11 additive ou log-additive (avec ajout de prévisions et extrapolations rétrospectives), et les résultats log-additifs sont présumés s'appliquer approximativement (Young 1968) à X-11 multiplicative.

Dans la section 2 de la présente communication, nous présentons notre méthode, qui s'appuie sur la première méthode de Wolter et Monsour (1981). Les différences existant entre ces deux méthodes sont décrites dans la section 2.4. Dans la section 3, nous décrivons trois extensions des résultats présentés dans la section 2. La première extension a pour but de montrer que notre méthode donne de bons résultats tant dans le cas d'estimations saisonnières que dans le cas d'estimations de tendance et d'estimations irrégulières, et qu'une plus grande généralité peut facilement être obtenue en permettant différents choix de filtre pour des mois différents. La deuxième extension donne la variance d'estimations relatives à des variations observées d'un mois à l'autre ou d'une année à l'autre. Enfin, lorsque la désaisonnalisation comporte l'estimation d'effets de régression (par exemple, dans le cas de variations de jours commerciaux ou de jours fériés), les résultats sont élargis afin de prendre en compte une variance supplémentaire attribuable à une erreur dans l'estimation des paramètres de régression.

Dans la section 4, nous présentons plusieurs exemples qui illustrent l'approche fondamentale et les élargissements décrits dans la section 3. Un aspect qui ressort de ces exemples est que dans le cas de séries chronologiques comportant une erreur d'échantillonnage, la variance de la désaisonnalisation que nous obtenons est souvent dominée par la contribution de l'erreur d'échantillonnage. Au centre des séries, nos résultats correspondent en réalité aux résultats obtenus avec la première méthode de Wolter et Monsour. Nos propres résultats tendent à différer de ceux

de Wolter et Monsour près des extrémités des séries. Ce fait est important, étant donné que les valeurs désaisonnalisées les plus récentes sont celles que l'on examine le plus attentivement. En outre l'effet des erreurs de prévision et d'extrapolation rétrospective sur les estimations de tendance peut être très important à la fin d'une série chronologique. D'autres résultats qui présentent un intérêt particulier sont les effets de certaines composantes non station naires des erreurs d'échantillonnage. Les exemples présentés dans la section 4 montrent que des éléments non stationnaires comme la variance d'erreurs d'échantillonnage qui varie avec le temps, ou la sélection périodique indépendante d'un nouvel échantillon, peuvent donner lieu à des modifications frappantes, au fil du temps, du type de variance de données désaisonnalisées ou d'estimations de tendance.

Enfin, la section 5 est consacrée à des remarques conclusives.

2. MÉTHODE

Définissons la série chronologique non désaisonnalisée observée par la notation y_t pour $t = 1, ..., n$. Les séries chronologiques désaisonnalisées sont souvent des estimations obtenues au moyen d'enquêtes sur échantillon successives (effectuées à intervalles mensuels ou trimestriels) et peuvent donc être considérées comme étant composées d'une série chronologique sous-jacente réelle Y_t et d'une série d'erreurs d'échantillonnage e_t présumée non corrélée avec Y_t. (Voir Bell et Hillmer 1990.) En notation vectorielle, $\mathbf{y}_o = \mathbf{Y}_o + \mathbf{e}_o$, où l'indice inférieur o indique que le laps de temps de ces vecteurs est l'ensemble des points temporels observés $1, ..., n$. Dans certains cas, y_t peut provenir de dénombrements successifs (comme dans le cas courant des statistiques nationales sur les exportations et les importations, par exemple); dans de tels cas, il n'y a pas d'erreurs d'échantillonnage, c'est-à-dire que $e_t = 0$.

Dans le développement qui suit, on présuppose que Y_t et e_t suivent des modèles de série chronologique connus. Le modèle pour Y_t comporte généralement le calcul de différences, comme dans le cas de la méthode ARIMA (autoregressive-integrated-moving average) et de modèles (structuraux) de composantes ARIMA. Le modèle pour Y_t peut être élargi afin d'inclure des termes de régression. (Nous abordons cet aspect dans la section 3.3.) Nous supposons que la série e_t ne nécessitera pas de calcul de différences, mais elle pourrait néanmoins présenter certains aspects non stationnaires, comme une variance qui change dans le temps. On présuppose que tout élément non stationnaire de ce genre a été pris en considération dans le modèle de e_t. Dans la pratique, les modèles sont élaborés à partir de données observées, comme l'ont décrit, par exemple, Bell et Hillmer (1990, présenté), Binder et Dick (1989, 1990), et Tiller (1992).

Dans l'application d'un filtre X-11 symétrique d'une longueur $2m + 1$ pour une désaisonnalisation comportant une extension complète avec prévisions et extrapolations rétrospectives, il faut ajouter au vecteur $\mathbf{y}_o$ m extrapolations et des m prévisions. Le vecteur portant les m valeurs de y_t avant les données observées, et les vecteurs $m \times 1$ correspondants pour Y_t, et e_t sont notés respectivement $\mathbf{y}_{er}$, $\mathbf{Y}_{er}$ et $\mathbf{e}_{er}$. Les vecteurs analogues des m valeurs futures de y_t, Y_t et e_t sont notés respectivement $\mathbf{y}_p$, $\mathbf{Y}_p$ et $\mathbf{e}_p$. Par conséquent,

$$\begin{pmatrix} \mathbf{y}_{er} \\ \mathbf{y}_o \\ \mathbf{y}_p \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{er} \\ \mathbf{Y}_o \\ \mathbf{Y}_p \end{pmatrix} + \begin{pmatrix} \mathbf{e}_{er} \\ \mathbf{e}_o \\ \mathbf{e}_p \end{pmatrix}. \tag{2.1}$$

Les vecteurs complets de (2.1), ci-après notés respectivement $\mathbf{y}$, $\mathbf{Y}$ et $\mathbf{e}$, ont une longueur $n + 2m$.

Nous supposons que les extrapolations rétrospectives et les prévisions ajoutées à $\mathbf{y}_o$ sont des prédictions linéaires avec erreur quadratique moyenne minimum (EQMM) de $\mathbf{y}_{er}$ et $\mathbf{y}_p$ (en utilisant $\mathbf{y}_o$) obtenues à partir du modèle de série chronologique connu. (Dans la pratique, le modèle est ajusté aux données $\mathbf{y}_o$.) Dans une situation de normalité, les extrapolations rétrospectives et les prévisions sont respectivement $E(\mathbf{y}_{er} | \mathbf{y}_o)$ et $E(\mathbf{y}_p | \mathbf{y}_o)$. Le vecteur des données observées auquel on a ajouté des extrapolations rétrospectives et des prévisions est noté $\hat{\mathbf{y}} = (\hat{\mathbf{y}}_b', \mathbf{y}_o', \hat{\mathbf{y}}_f')'$, où $\hat{\mathbf{y}}_{er} = E(\mathbf{y}_{er} | \mathbf{y}_o)$ et $\hat{\mathbf{y}}_p = E(\mathbf{y}_p | \mathbf{y}_o)$. Afin de simplifier la notation, nous considérons dorénavant des expressions telles que $(\hat{\mathbf{y}}_{er}, \mathbf{y}_o, \hat{\mathbf{y}}_p)$ comme étant le vecteur-colonne $(\hat{\mathbf{y}}_{er}', \mathbf{y}_o', \hat{\mathbf{y}}_p')'$.

Soit le filtre linéaire symétrique de désaisonnalisation X-11, noté $\omega(B) = \sum_{-m}^{m} \omega_j B^j$, où B est l'opérateur de décalage et où les notations ω_j représentent les poids de filtre ($\omega_j = \omega_{-j}$). Le calcul de ω_j est traité par Young (1968) et Wallis (1982). Ici, on a utilisé les résultats de Bell et Monsell (1992). L'application d'un $\omega(B)$ à la série élargie par des prévisions et des extrapolations rétrospectives peut être écrite sous la forme $\mathbf{\Omega}\hat{\mathbf{y}}$, où $\mathbf{\Omega}$ est une matrice de dimension $n \times (n + 2m)$. Chaque ligne de $\mathbf{\Omega}$ contient les poids de filtre ($\omega_{-m}, ..., \omega_0, ..., \omega_m$), précédés et suivis du nombre de zéros approprié, pour que le poids central du filtre X-11 (ω_0) multiplie l'observation qui est corrigée. Par conséquent, dans la première ligne de $\mathbf{\Omega}$, il n'y a pas de zéros qui précèdent et un nombre $n - 1$ de zéros qui suivent; dans la deuxième ligne, il y a un zéro qui précède et un nombre $n - 2$ de zéros qui suivent, etc. Dans le cas du filtre X-11 par défaut, $m = 84$. Le choix d'autres moyennes mobiles saisonnières ou de tendance, dans X-11, modifie la valeur de m, qui passe alors d'une valeur inférieure de 70 à une valeur supérieure de 149.

La question qui se pose est celle de savoir ce qu'estime $\mathbf{\Omega}\hat{\mathbf{y}}$. Comme nous l'avons mentionné dans l'introduction, nous définissons l'«objectif» de la désaisonnalisation comme étant la série corrigée qui résulterait s'il n'y avait pas d'erreur d'échantillonnage et s'il y avait suffisamment de données avant et après tous les points temporels présentant un intérêt pour l'application du filtre symétrique.

L'objectif est donc $\omega(B)Y_t$, ou, en notation vectorielle, $\mathbf{\Omega Y}$, et le vecteur d'erreur de désaisonnalisation est $\mathbf{v} = \mathbf{\Omega}(\mathbf{Y} - \hat{\mathbf{y}})$. Nous voulons connaître la matrice de variance-covariance $\text{var}(\mathbf{v}) = \mathbf{\Omega}\,\text{var}(\mathbf{Y} - \hat{\mathbf{y}})\mathbf{\Omega}'$. Celle-ci peut être calculée facilement une fois que l'on a obtenu $\text{var}(\mathbf{Y} - \hat{\mathbf{y}})$. Ci-après, et jusqu'à la section 2.3, nous présentons le calcul de $\text{var}(\mathbf{Y} - \hat{\mathbf{y}})$.

Nous écrivons d'abord $\mathbf{Y} - \hat{\mathbf{y}} = (\mathbf{y} - \mathbf{e}) - \hat{\mathbf{y}} = (\mathbf{er}, \mathbf{0}, \mathbf{p}) - \mathbf{e}$, où $\mathbf{er} = \mathbf{y}_{er} - \hat{\mathbf{y}}_{er}$ est le vecteur $m \times 1$ des erreurs d'extrapolation rétrospective, et $\mathbf{p} = \mathbf{y}_p - \hat{\mathbf{y}}_p$ est le vecteur $m \times 1$ des erreurs de prévision. Les modèles pour Y_t et e_t étant donnés, nous calculons $\text{var}(\mathbf{Y} - \hat{\mathbf{y}})$ en calculant de manière distincte $\text{var}(\mathbf{e})$, $\text{var}(\mathbf{er}, \mathbf{0}, \mathbf{p})$ et $\text{cov}[(\mathbf{er}, \mathbf{0}, \mathbf{p}), \mathbf{e}]$, comme nous le décrivons dans les sections 2.1 à 2.3. On obtient ensuite facilement $\text{var}(\mathbf{Y} - \hat{\mathbf{y}})$, sous la forme $\text{var}(\mathbf{er}, \mathbf{0}, \mathbf{p}) + \text{var}(\mathbf{e}) - \text{cov}[(\mathbf{er}, \mathbf{0}, \mathbf{p}), \mathbf{e}] - \text{cov}[(\mathbf{er}, \mathbf{0}, \mathbf{p}), \mathbf{e}]'$. Par conséquent,

$$\text{var}(\mathbf{v}) = \mathbf{\Omega}\{\text{var}(\mathbf{er}, \mathbf{0}, \mathbf{p}) + \text{var}(\mathbf{e}) - \text{cov}[(\mathbf{er}, \mathbf{0}, \mathbf{p}), \mathbf{e}] - \text{cov}[(\mathbf{er}, \mathbf{0}, \mathbf{p}), \mathbf{e}]'\}\mathbf{\Omega}'.$$

Exemple – Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis. Les calculs pour chaque composante de $\text{var}(\mathbf{Y} - \hat{\mathbf{y}})$ étant expliqués, nous présentons sous forme de graphiques les résultats obtenus dans le cas d'une série utilisée comme exemple: les mises en chantier, aux États-Unis, de janvier 1975 à novembre 1988, d'immeubles comprenant au moins cinq unités de logement (167 observations). La série d'origine, la série désaisonnalisée et la tendance estimée sont illustrées dans la figure 1. Dans la pratique, pour effectuer la désaisonnalisation de cette série, le Census Bureau utilise une décomposition multiplicative, avec une moyenne mobile saisonnière 3×9 et un filtre de tendance de Henderson à treize termes. Le modèle qui suit, utilisé pour cette série, a été élaboré dans l'article de Bell et Hillmer (présenté):

$$y_t = Y_t + e_t$$

$$(1-B)(1-B^{12})Y_t = (1-0.67B+0.36B^2)(1-.8753B^{12})a_t,$$

$$\sigma_a^2 = 0.0191$$

$$e_t = (1-0.11B-0.10B^2)b_t, \sigma_b^2 = 0.00714 \quad (2.2)$$

Dans ce cas, y_t représente le logarithme de la série chronologique d'origine (e^{y_t}), de sorte que (2.2) implique une décomposition multiplicative de cette série ($e^{y_t} = e^{Y_t} e^{e_t}$).

2.1 Calcul de Var(e)

Si e_t suit un modèle ARIMA stationnaire, on peut calculer var(e) à partir de résultats standard (par exemple, McLeod 1975, 1977 et Wilson 1979). Si $\text{var}(e_t)$ change au fil du temps, nous écrivons $e_t = h_t \tilde{e}_t$, où $h_t^2 = \text{var}(e_t)$, et où $\tilde{e}_t$ a une variance de un et la même fonction d'autocorrélation que e_t (voir Bell et Hillmer, présenté). Puis, nous écrivons $\mathbf{e} = \mathbf{H}\tilde{\mathbf{e}}$, où $\mathbf{H} = \text{diag}(h_{1-m}, \ldots, h_{n+m})$, et nous obtenons $\text{var}(\mathbf{e}) = \mathbf{H}\,\text{var}(\tilde{\mathbf{e}})\mathbf{H}'$. $\text{Var}(\tilde{\mathbf{e}})$ est la matrice d'autocorrélation de $\mathbf{e}$, et elle peut être calculée comme nous venons de le faire, à l'aide du modèle pour $\tilde{e}_t$.

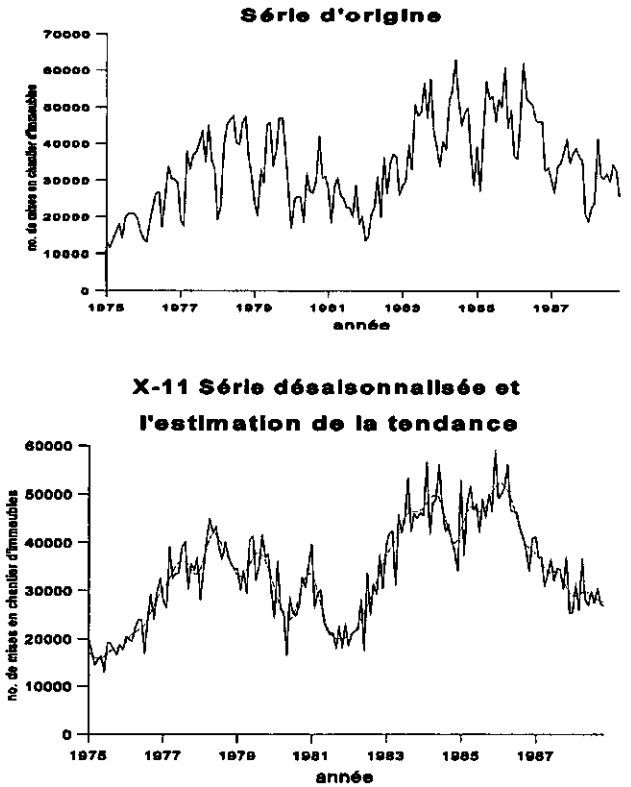


Figure 1. Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis. Le graphique supérieur montre la série d'origine allant de janvier 1975 à novembre 1988. Le fort caractère saisonnier de cette série est mis en évidence par les déflexions annuelles typiques des mois d'hiver. Le graphique inférieur montre la série désaisonnalisée par la méthode X-11 (ligne continue) et l'estimation de tendance (ligne pointillée) pour la même période. La désaisonnalisation est multiplicative, par l'utilisation d'une moyenne mobile saisonnière 3×9 et d'un filtre d'estimations de tendance de Henderson à treize termes.

Si l'échantillon fait l'objet d'un renouvellement indépendant de temps à autre, var (e) sera une matrice diagonale par blocs, et ces blocs correspondent aux points temporels auxquels les divers échantillons distincts étaient en vigueur. Chaque bloc diagonal de var (e) peut être calculé de la façon que nous venons de décrire. Ces deux aspects non stationnaires de e_t (la variance qui change au fil du temps et les «ruptures de covariance» dues à des renouvellements indépendants de l'échantillon) sont ceux que l'on observe dans les exemples de la section 4.

Exemple – Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis (suite). Dans le cas du modèle MA(2) pour e_t donné dans (2.2), l'autocovariance peut être calculée facilement. La var (e) qui est obtenue est une matrice bande, où $\text{var}(e_t) = 0{,}007298$ sur la diagonale, $\text{cov}(e_t, e_{t-1}) = -0{,}000707$ sur la première sub-diagonale et la première supra-diagonale, et $\text{cov}(e_t, e_{t-2}) = -0{,}000714$ sur la deuxième sub-diagonale et la deuxième

supra-diagonale. Le reste de var (**e**) est égal à zéro. Après la pré-multiplication et la postmultiplication par les matrices à filtre de désaisonnalisation $\mathbf{\Omega}$ et $\mathbf{\Omega}'$, l'incidence de l'erreur d'échantillonnage sur la variance des séries désaisonnalisées est constante dans le cas de chacune des observations (figure 2). Il en est ainsi parce que le résultat de l'application d'un filtre linéaire qui ne varie pas avec le temps à une série stationnaire $(\omega(B)e_t)$ est une série stationnaire qui présente une variance constante.

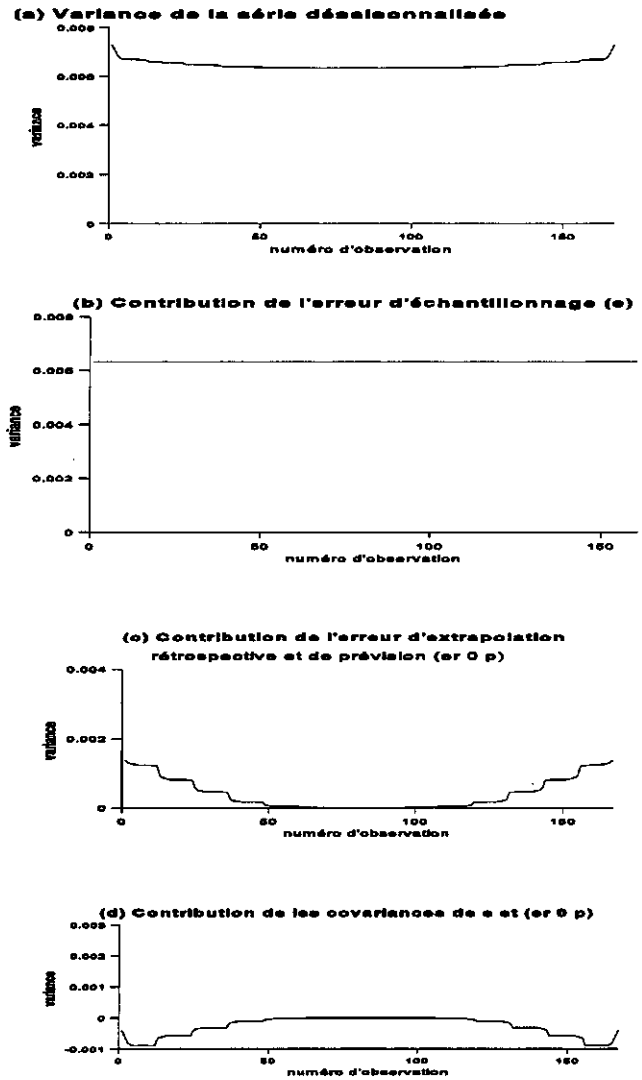


Figure 2. Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis: décomposition de la variance après désaisonnalisation X-11. Le graphique supérieur montre la variance de la série désaisonnalisée, sous forme de total des trois composantes. Le deuxième graphique indique la contribution de l'erreur d'échantillonnage (e), qui est la composante la plus importante; celle-ci est constante tout au long de la série. Le troisième graphique montre la contribution de l'erreur d'extrapolation rétrospective et de prévision (**er 0 p**), dont la valeur est zéro au milieu de la série, où aucune extrapolation rétrospective ou prévision n'est nécessaire, et va en augmentant à chacune des extrémités de la série, à mesure qu'on utilise davantage d'extrapolations rétrospectives et de prévisions. Le graphique inférieur indique la somme des deux termes de covariance (cov(e, (**er 0 p**)) + cov((**er 0 p**), e)), qui tend à neutraliser la contribution de l'erreur d'extrapolation rétrospective et de prévision.

2.2 Calcul de Var (er, 0, p)

Les valeurs des n lignes et des n colonnes centrales de var(**er, 0, p**) sont toutes zéro. Il nous faut calculer var(**er**), var(**p**) et cov(**er, p**) dans le cas des blocs d'angle var(**er, 0, p**). Bien que le calcul de la variance d'erreurs d'extrapolation rétrospective (er) ou de prévision (p) dans le cas de modèles donnés est une opération courante dans l'analyse de séries chronologiques, dans le cas qui nous intéresse elle est compliquée par la représentation de la composante y_t comme $Y_t + e_t$, ainsi que par le calcul des différences dans le modèle utilisé pour Y_t. Bien que les calculs à effectuer dans le cas de tels modèles soient souvent traités à l'aide du filtre de Kalman (Bell et Hillmer, présenté; Binder et Dick 1989, 1990; Tiller 1992), cette façon de procéder est peu commode ici, étant donné que nous devons connaître la covariance de toutes les paires de variables aléatoires parmi m les erreurs de prévision et les m erreurs d'extrapolation rétrospective. C'est pourquoi nous utilisons plutôt une méthode directe à matrice, proposée par Bell et Hillmer (1988).

Supposons que l'opérateur de calcul de différences requis pour rendre Y_t stationnaire est $\delta(B)$, ce qui est de degré d. Étant donné que l'on suppose que e_t ne nécessite pas de calcul de différences, $\delta(B)$ est également l'opérateur de calcul de différences requis par y_t. Si $\delta(B)y_t = w_t$, alors $w_t = \delta(B)Y_t + \delta(B)e_t$. Nous introduisons la matrice $\mathbf{\Delta}$, correspondant à $\delta(B)$, définie de manière à ce que $\mathbf{\Delta y = w}$ soit le vecteur de **y** différencié. Le vecteur $\mathbf{w} = (\mathbf{w}_{er}, \mathbf{w}_o, \mathbf{w}_p)$, qui a une longueur de $n + 2m - d$, est divisé de telle manière que $\mathbf{w}_{er}$ et $\mathbf{w}_p$ sont des vecteurs $m \times 1$, et $\mathbf{w}_o$ est le vecteur des $n - d$ données observées ayant fait l'objet d'un calcul de différences. Ainsi, $\mathbf{\Delta}$ a les dimensions $(n+2m-d)\times(n+2m)$. Il est a noter que parce que d observations sont perdues dans le calcul de différences, $\mathbf{w}_{er}$ et $\mathbf{w}_o$ commencent à d points temporels plus tard que $\mathbf{y}_{er}$ et $\mathbf{y}_o$, respectivement. C'est-à-dire que $\mathbf{y}_{er}$ et $\mathbf{y}_o$ commencent aux points temporels 1-m et 1, mais $\mathbf{w}_{er}$ et $\mathbf{w}_o$ commencent aux points temporels $1 - m + d$ et $d + 1$.

Définissons $\mathbf{u} = (u_{1-m+d}, \ldots, u_{n+m})' = \mathbf{\Delta Y}$. La série chronologique u_t est stationnaire. Étant donné que $\mathbf{w} = \mathbf{u} + \mathbf{\Delta e}$, (**u** et **e** n'étant pas corrélés l'un avec l'autre) var(**w**) = var(**u**) + $\mathbf{\Delta}$var(**e**)$\mathbf{\Delta}'$. Nous découpons var(**w**) comme suit:

$$\text{var}(\mathbf{w}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix},$$

où $\mathbf{\Sigma}_{11}$ est var($\mathbf{w}_{er}$), $\mathbf{\Sigma}_{12}$ est cov($\mathbf{w}_{er}, \mathbf{w}_o$), etc.

Étant donné que **y**, lorsqu'il est différencié par rapport à **w** à l'aide de $\delta(B)$, a perdu d valeurs de données, il ne peut être obtenu à partir de **w** sans connaître également une séquence de d «valeurs de départ». Considérons l'obtention de $\mathbf{y}_p$ à partir de $\mathbf{w}_p$ et des valeurs de départ $\mathbf{y}_* = (y_{n+1-d}, \ldots, y_n)'$. Le théorème 1 contenu dans Bell (1984a) peut être utilisé pour montrer que

$$\mathbf{y}_p = \mathbf{A}\mathbf{y}_* + \mathbf{C}\mathbf{w}_p \qquad (2.3)$$

dans le cas des matrices **A** et **C** déterminées par $\delta(B)$. Les lignes de la matrice $m \times m$ **C** sont composées des coefficients de $\xi(B) = 1 + \xi_1 B + \xi_2 B^2 + \ldots = \delta(B)^{-1}$, sous la forme

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ \xi_1 & 1 & & 0 & 0 \\ \xi_2 & \xi_1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & 1 & 0 \\ \xi_{m-1} & \xi_{m-2} & \ldots & \xi_1 & 1 \end{pmatrix}$$

A est une matrice $m \times d$ qui rend compte de l'incidence des d valeurs de départ présentes dans $\mathbf{y}_*$ sur $\mathbf{y}_p$. La forme exacte de **A** est donnée dans Bell (1984a) et comme elle s'annule exactement dans notre application, nous ne la donnerons pas ici. Dans (2.3) $\mathbf{y}_*$ est connu, étant donné que ce terme fait partie de $\mathbf{y}_o$, c'est-à-dire des données observées. Par conséquent, à partir de (2.3), la prévision d'EQMM de $\mathbf{y}_p$ est $\hat{\mathbf{y}}_p = \mathbf{A}\mathbf{y}_* + \mathbf{C}\hat{\mathbf{w}}_p$, où $\hat{\mathbf{w}}_p$ est la prévision d'EQMM de $\mathbf{w}_p$. Ainsi, $\mathbf{p} = \mathbf{y}_p - \hat{\mathbf{y}}_p = \mathbf{A}\mathbf{y}_* + \mathbf{C}\mathbf{w}_p - (\mathbf{A}\mathbf{y}_* + \mathbf{C}\hat{\mathbf{w}}_p) = \mathbf{C}(\mathbf{w}_p - \hat{\mathbf{w}}_p)$, et $\text{var}(\mathbf{p}) = \mathbf{C}\text{var}(\mathbf{w}_p - \hat{\mathbf{w}}_p)\mathbf{C}'$.

Dans l'hypothèse A de Bell (1984a), qui mène aux résultats standard pour les prévisions relatives aux séries non stationnaires (comme dans Box et Jenkins 1976, chapitre 5), $\hat{\mathbf{w}}_p = \Sigma_{32}\Sigma_{22}^{-1}\mathbf{w}_o$. Il est à noter que dans ce cas, on utilise uniquement les données $\mathbf{w}_o$ ayant fait l'objet d'un calcul des différences dans la prévision de $\mathbf{w}_p$. Ensuite, à partir de résultats standard sur la prédiction linéaire, $\text{var}(\mathbf{w}_p - \hat{\mathbf{w}}_p) = \Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{32}'$. Ainsi, $\text{var}(\mathbf{p}) = \mathbf{C}(\Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{32}')\mathbf{C}'$.

Pour obtenir var (**er**) et cov (**er**, **p**), nous signalons que les résultats obtenus par Bell (1984a, p. 651) impliquent que des calculs similaires sont valides dans le cas d'erreurs d'extrapolation rétrospective **er**. En effet, on peut montrer que $\mathbf{er} = (-1)^r\mathbf{C}'(\mathbf{w}_{er} - \hat{\mathbf{w}}_{er})$, où $\hat{\mathbf{w}}_{er}$ est l'extrapolation rétrospective d'EQMM de $\mathbf{w}_{er}$, et r est le nombre de fois que $(1 - B)$ figure dans le polynome $\delta(B)$. (La présence de $\mathbf{C}'$ dans cette expression, plutôt que de **C** est due à l'indexation de $\mathbf{w}_{er}$ et $\hat{\mathbf{w}}_{er}$ vers l'avant, au fil du temps, même si le processus d'extrapolation rétrospective se déroule en remontant dans le temps.) Ainsi, $\text{var}(\mathbf{er}) = \mathbf{C}'\text{var}(\mathbf{w}_{er} - \hat{\mathbf{w}}_{er})\,\mathbf{C} = \mathbf{C}'(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}')\mathbf{C}$. De manière analogue, $\text{cov}(\mathbf{p}, \mathbf{er}) = (-1)^r\,\mathbf{C}(\Sigma_{31} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{12})\mathbf{C}$. Dans la pratique, afin d'éviter l'inversion de Σ_{22}, var(**p**), var(**er**) et cov(**p**, **er**) peuvent être calculés en appliquant la décomposition de Cholesky à Σ_{22}. (Voir annexe A.)

Exemple – Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis (suite). L'incidence de var(**er**, **0**, **p**) sur la variance de la désaisonnalisation est indiquée dans la figure 2. Celle-ci est égale à zéro, ou essentiellement égale à zéro, dans le cas d'observations effectuées dans le milieu des séries, où peu d'extrapolations rétrospectives ou prévisions, voire aucune, n'ont à être effectuées pour appliquer le filtre symétrique de correction. Vers les extrémités des séries, l'incidence des erreurs d'extrapolation rétrospective ou de prévision devient plus importante, étant donné qu'un nombre croissant d'observations doivent faire l'objet d'une extrapolation rétrospective ou d'une prévision, afin de pouvoir appliquer le filtre. Les sauts dans le graphique surviennent lorsqu'une observation prévue ou extrapolée rétrospectivement est multipliée par un poids du filtre de correction qui est un multiple de la période saisonnière, étant donné que ces poids ont la valeur la plus importante (Bell et Monsell, 1992). Il est à noter que l'incidence de var(**er**, **0**, **p**) aux extrémités proprement dites des séries est moins importante que l'incidence de var(**e**), mais elle n'est pas négligeable.

2.3 Calcul de Cov[(er 0 p), e]

Pour calculer cov(p,e), nous remarquons d'abord, d'après les résultats de la section précédente, que $\mathbf{p} = \mathbf{y}_p - \hat{\mathbf{y}}_p = \mathbf{p} = \mathbf{y}_p - \hat{\mathbf{y}}_p = \mathbf{C}(\mathbf{w}_p - \hat{\mathbf{w}}_p) = \mathbf{C}(\mathbf{w}_p - \Sigma_{32}\Sigma_{22}^{-1}\mathbf{w}_o) = \mathbf{C}[0 \mid -\Sigma_{32}\Sigma_{22}^{-1} \mid \mathbf{I}_m]\mathbf{w} = \mathbf{C}[0 \mid -\Sigma_{32}\Sigma_{22}^{-1} \mid \mathbf{I}_m]\Delta\mathbf{y}$. Étant donné que $\text{cov}(\mathbf{y}, \mathbf{e}) = \text{cov}(\mathbf{Y} + \mathbf{e}, \mathbf{e}) = 0 + \text{var}(\mathbf{e})$, nous constatons que $\text{cov}(\mathbf{p}, \mathbf{e}) = \mathbf{C}[0 \mid -\Sigma_{32}\Sigma_{22}^{-1} \mid \mathbf{I}_m]\,\Delta\,\text{var}(\mathbf{e})$. Cov(**er**, **e**) est calculée d'une façon similaire en observant que: $\mathbf{er} = (-1)^r\mathbf{C}'(\mathbf{w}_{er} - \hat{\mathbf{w}}_{er}) = (-1)^r\mathbf{C}'(\mathbf{w}_{er} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{w}_o) = (-1)^r\,\mathbf{C}'[\mathbf{I}_m \mid -\Sigma_{12}\Sigma_{22}^{-1} \mid 0]\;\Delta\mathbf{y}$, de sorte que $\text{cov}(\mathbf{er}, \mathbf{e}) = (-1)^r\,\mathbf{C}'[\mathbf{I}_m \mid -\Sigma_{12}\Sigma_{22}^{-1} \mid 0]\,\Delta\text{var}(\mathbf{e})$.

Exemple – Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis (suite). La figure 2 montre que l'incidence de cov[(**er**, **0**, **p**), **e**] est zéro ou près de zéro dans le milieu des séries, mais qu'elle devient de plus en plus négative vers les extrémités de celles-ci, d'une façon similaire, mais de signe opposé et d'une moindre importance, à celle qui caractérise var(**er**, **0**, **p**). Aux extrémités proprement dites des séries, cependant, la tendance s'inverse et la covariance augmente. Les éléments de cov[(**er**, **0**, **p**) **e**], sont principalement positifs, de sorte que la contribution de ce terme à la variance de la désaisonnalisation est négative parce que cov[(**er**, **0**, **p**), **e**] et sa transposée sont soustraites de var(**e**) + var(**er**, **0**, **p**). L'effet net est que la soustraction de $\mathbf{\Omega}\{\text{cov}[(\mathbf{er}, \mathbf{0}, \mathbf{p}), \mathbf{e}] + \text{cov}[(\mathbf{er}, \mathbf{0}, \mathbf{p}), \mathbf{e}]'\}\mathbf{\Omega}'$ tend à neutraliser l'effet de l'addition de $\mathbf{\Omega}\,\text{var}(\mathbf{b}, \mathbf{0}, \mathbf{f})\mathbf{\Omega}'$, sauf à proximité des extrémités des séries. Ainsi, le graphique de la variance des séries désaisonnalisées (figure 2) est très semblable au graphique de la contribution de var(**e**), sauf à proximité des extrémités de la série. Nous avons observé ce type d'«effet de neutralisation» dans plusieurs autres exemples, dont ceux présentés dans la section 4.

2.4 Comparaison avec la première méthode de Wolter et Monsour

La première méthode de Wolter and Monsour (1981) proposait l'utilisation de $\mathbf{\Omega}_{wm}\,\text{var}(\mathbf{e}_o)\,\mathbf{\Omega}_{wm}'$ comme matrice de variance-covariance des erreurs de désaisonnalisation

X-11, où $\mathbf{\Omega}_{wm}$ est une matrice $n+n$ dont les lignes contiennent les poids du filtre linéaire X-11, tant symétrique qu'asymétrique. Autrement dit, les lignes centrales (lignes t telles que $m<t<n-m+1$, en présumant que $n>2m$) de $\mathbf{\Omega}_{wm}$ contiennent les poids de filtre X-11 symétrique, mais les m premières et dernières lignes de $\mathbf{\Omega}_{wm}$ contiennent des poids de filtre X-11 asymétrique. Ainsi, les lignes centrales de Ω_{wm} et de Ω contiennent les mêmes poids de filtre, mais non les m premières et dernières lignes. Cela signifie que notre méthode va donner les mêmes résultats que la méthode de Wolter et Monsour dans le cas de $m<t<n-m+1$, c'est-à-dire, dans le cas de points temporels auxquels on applique le filtre symétrique. Les résultats obtenus avec les deux méthodes vont être différents dans le cas des m premiers et derniers points temporels. Étant donné que ce sont les données désaisonnalisées les plus récentes qui reçoivent le plus d'attention, cette différence peut être importante.

Wolter et Monsour ont également considéré l'utilisation d'une matrice $\mathbf{\Omega}^*$ au lieu d'une matrice $\mathbf{\Omega}_{wm}$, où $\mathbf{\Omega}^*$ étant $(n+12)\times(n+12)$, afin d'inclure 12 lignes supplémentaires de poids correspondant à des filtres de désaisonnalisation d'année subséquente. Bien que la correction d'année subséquente fût une pratique courante au début des années 80, cette méthode a maintenant été remplacée en grande partie, aux États-Unis, par la correction concurrente (McKenzie 1984).

La différence entre notre approche et celle de Wolter et Monsour peut être considérée de deux façons. Notre point de vue est le suivant: étant donné que Wolter et Monsour n'ont pas pris en considération l'extrapolation rétrospective et la prévision, leur méthode ne tient pas compte de la contribution des erreurs d'extrapolation rétrospective et de prévision à l'erreur de désaisonnalisation. Cette contribution a une incidence sur les résultats dans le cas des m premiers et derniers points temporels, même si les exemples présentés dans la section 4 montrent que cette contribution est souvent faible. Dans certains cas, elle n'est pas négligeable, notamment dans celui des séries chronologiques qui ne sont pas sujettes à des erreurs d'échantillonnage. Dans le cas de telles séries, la méthode de Wolter et Monsour attribuerait une variance zéro aux corrections, même si les corrections initiales sont révisées à mesure que l'on obtient de nouvelles données.

L'autre façon de considérer la différence entre les deux méthodes est centrée sur la différence existant entre les «objectifs». Dans le cas de la méthode de Wolter et Monsour, l'erreur de désaisonnalisation peut être vue comme $\mathbf{\Omega}_{wm}(\mathbf{Y}_o - \mathbf{y}_o) = -\mathbf{\Omega}_{wm}\mathbf{e}_o$. Étant donné que cette approche donne une erreur zéro dans le cas des séries qui ne comportent pas d'erreurs d'échantillonnage ($\mathbf{Y}_o = \mathbf{y}_o$), Wolter et Monsour définissent implicitement l'objectif de la désaisonnalisation comme étant $\mathbf{\Omega}_{wm}\mathbf{Y}_o$. Cette définition d'objectif présente une caractéristique non souhaitable: la valeur cible pour un point temporel donné varie à mesure que l'on obtient des données supplémentaires, étant donné que les lignes de $\mathbf{\Omega}_{wm}$ contiennent des poids de filtre différents. Notre valeur cible pour tout point temporel t est toujours $\omega(B)Y_t$.

Exemple – Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis (suite). Nous avons comparé les résultats obtenus à l'aide de notre méthode avec les résultats obtenus par Wolter et Monsour à l'aide du filtre de désaisonnalisation X-11 par défaut, même si, comme nous l'avons mentionné précédemment, cette série servant d'exemple est corrigée au moyen du filtre optionnel à moyenne mobile saisonnière 3×9. Dans cette comparaison, nous avons utilisé, pour des raisons de commodité, le filtre par défaut; l'obtention de résultats pour l'application de la méthode de Wolter et Monsour requiert des poids de filtre X-11 asymétrique, et Nash Monsour a mis à notre disposition un programme informatique qui produisait ces résultats seulement pour le filtre par défaut. La figure 3 montre les résultats obtenus avec les deux méthodes. Les variances non constantes au fil du temps que l'on obtient avec la méthode Wolter-Monsour sont dues à l'application de filtres différents à des points temporels différents. Une conséquence intéressante de cet aspect est que malgré le caractère stationnaire de l'erreur d'échantillonnage, la variance de désaisonnalisation obtenue à l'aide de la méthode de Wolter et Monsour est nettement plus élevée au milieu des séries que dans le cas de nombreux points temporels situés aux extrémités (mais non près de celles-ci). Ce fait comporte la conséquence peu plausible que l'utilisation d'une moindre quantité de données produit des estimateurs qui présentent une variance plus faible. Un phénomène similaire peut être observé dans le cas de plusieurs exemples présentés par Pfeffermann (1994).

Les résultats obtenus à l'aide du filtre de désaisonnalisation X-11 par défaut sont également utiles aux fins de la comparaison avec le filtre à moyenne mobile saisonnière 3×9, dont les résultats sont indiqués dans la figure 2. Les différences entre les résultats obtenus à l'aide des deux filtres ne sont pas grandes. L'incidence de l'erreur d'échantillonnage est quelque peu plus faible, et celle de l'erreur d'extrapolation rétrospective ou de prévision est un peu plus élevée lorsqu'on utilise le filtre de désaisonnalisation par défaut.

3. ÉLARGISSEMENT DE LA MÉTHODE

Dans la présente section, nous présentons trois élargissements de la méthode générale décrite dans la section 2. Les deux premiers élargissements sont simples, tandis que le troisième est plus complexe.

3.1 Variances relatives à des estimations saisonnières, de tendance et irrégulières – Variances dans le cas de l'utilisation de filtres qui varient avec le temps

Seuls les poids de filtre placés dans la matrice $\mathbf{\Omega}$ permettent de distinguer la composante non saisonnière (désaisonnalisée) dans la dérivation présentée dans la

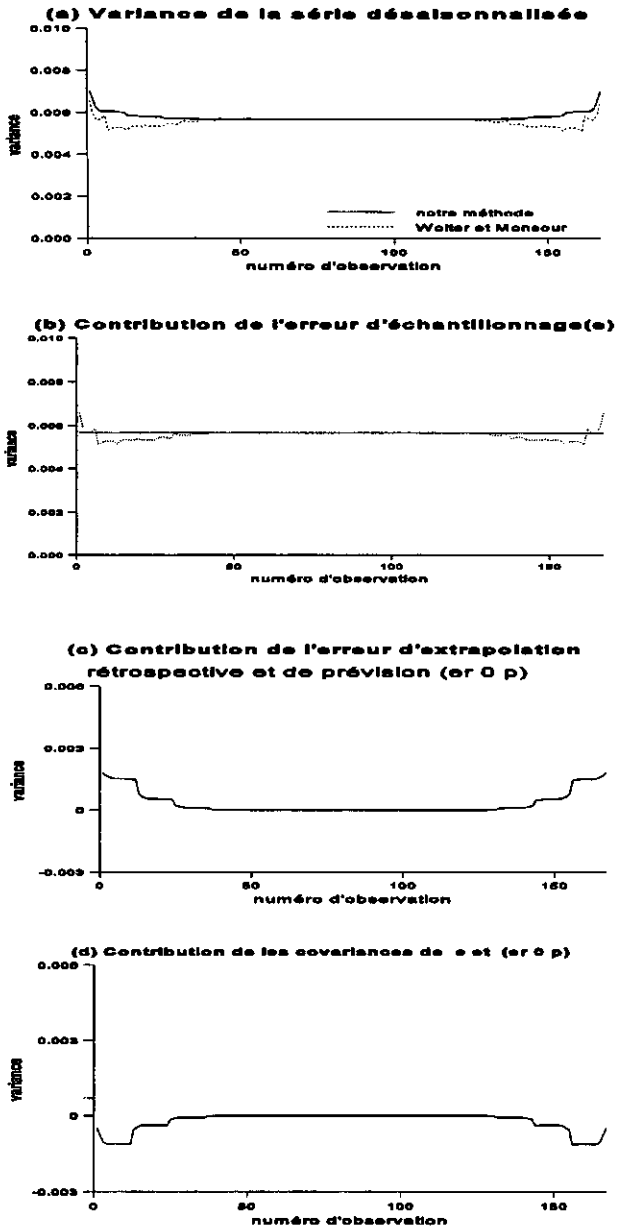


Figure 3. Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis: Comparaison avec la méthode de Wolter et Monsour (1981). La description du graphique est la même que celle de la figure 2. La méthode de Wolter et Monsour (lignes pointillées) fait appel aux filtres X-11 asymétriques pour les extrémités de la série et rend compte uniquement de l'erreur d'échantillonnage. Nos méthodes concordent au milieu de la série, où il n'y a pas de contribution de l'erreur d'extrapolation rétrospective et de prévision. Dans le cas de la méthode de Wolter et Monsour, la variance diminue de manière inappropriée près des extrémités de la série, ce qui laisse penser que l'utilisation d'une moindre quantité de données produit des estimations présentant une variance plus faible. Par contraste avec les figures 1, 2, 4, 5 et 6, les résultats indiqués dans cette figure-ci sont obtenus à l'aide de filtres X-11 par défaut (voir texte).

section 2. Par conséquent, les variances correspondantes relatives aux estimations X-11 des composantes saisonnières, de tendance et irrégulières découlent de la même expression par une simple modification de la matrice Ω afin d'y ajouter les poids de filtre voulus. Cela modifie également la dimension de Ω, étant donné que les longueurs respectives des filtres de désaisonnalisation, de tendance et d'estimations irrégulières diffèrent (dans le cas d'options données), et que la longueur de filtre détermine la taille de Ω.

Un élargissement similaire permet de traiter le cas de moyennes mobiles saisonnières différentes choisies pour des mois (ou des trimestres) différents, qui est une option offerte par la méthode X-11. Cela modifie les filtres de désaisonnalisation (ainsi que les filtres saisonniers, de tendance et d'estimations irrégulières) appliqués au cours de mois différents. Cet élargissement peut être appliqué également aux résultats de la section 2 par une simple modification de Ω. Étant donné que les lignes de Ω correspondent aux points temporels qui font l'objet de la correction, nous établissons simplement que la ligne t de Ω contient les poids (avec un nombre suffisant de zéros) du filtre qui est appliqué au mois t. Il faut veiller à attribuer la bonne taille à Ω si la moyenne mobile la plus longue qui a été choisie n'est pas utilisée pour les premier et dernier mois d'une série.

Exemple – Mises en chantier d'immeubles à cinq unités de logement ou plus aux États.-Unis (suite). La figure 4 montre la variance de l'estimation de tendance X-11 dans le cas de l'utilisation de la moyenne mobile saisonnière 3×9 et du filtre de Henderson à treize termes. La différence la plus évidente par rapport aux résultats de la désaisonnalisation est l'importante incidence de l'erreur de prévision ou d'extrapolation rétrospective aux extrémités de chaque série. Cela se produit parce que les poids les plus grands du filtre de tendance ($\omega^{(T)}(B)$) sont le poids central ($\omega_o^{(T)}$) et les poids adjacents ($\omega_1^{(T)}, \omega_2^{(T)}, \omega_3^{(T)}$), qui sont appliqués aux données qui précèdent et qui suivent immédiatement l'observation faisant l'objet de la correction (Bell et Monsell 1992). Aux extrémités des séries, les poids ($\omega_1^{(T)}, \omega_2^{(T)}, \omega_3^{(T)}$) s'appliquent aux observation qui ont fait l'objet d'une prévision ou d'une extrapolation rétrospective, ce qui entraîne dans ce cas une augmentation importante de l'incidence des erreurs de prévision ou d'extrapolation rétrospective. Il s'ensuit que l'incertitude au sujet de la tendance augmente brusquement aux extrémités des séries. Cependant, au milieu de celles-ci, la variance des estimations de tendance présentées dans la figure 4 est nettement moins importante que la variance de la désaisonnalisation que montre la figure 3, en raison du lissage de l'erreur d'échantillonnage par le filtre d'estimations de tendance.

3.2 Variances relatives à des variations désaisonnalisées observées d'un mois à l'autre et d'une année à l'autre

La variance d'erreurs d'estimations désaisonnalisées de variations observées d'un mois à l'autre est représentée par les quantités $\text{var}(v_t - v_{t-1})$, $t = 2, \ldots, n.$; var(v) étant donné, la matrice complète de covariance des erreurs pour les

variations désaisonnalisées observées d'un mois à l'autre peuvent être calculées de la façon suivante: $\Delta_1 \text{var}(\mathbf{v})\Delta_1'$, où

$$\Delta_1 = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & \ddots & 0 & 0 \\ 0 & 0 & & -1 & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

est de dimension $(n - 1) \times n$. La matrice de la covariance des erreurs relatives aux variations désaisonnalisées observées d'une année à l'autre, dans des séries trimestrielles, est calculée de façon similaire, soit $\Delta_4 \text{var}(\mathbf{v})\Delta_4'$, où

$$\Delta_4 = \begin{pmatrix} -1 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & & & \ddots & 0 & 0 \\ 0 & 0 & \cdots & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

est de dimension $(n-4)\times n$. La matrice $(n-12)\times n$ correspondante, Δ_{12}, dans les séries mensuelles, suit un modèle similaire, avec des zéros supplémentaires.

La variance de variations de tendance observées d'un mois à l'autre ou d'une année à l'autre peut également être obtenue facilement, comme le montrent la présente description et celle contenue dans la section 3.1.

Exemple – Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis (suite). Nous avons produit les erreurs-types relatives à des variations désaisonnalisées observées d'un mois à l'autre et d'une année à l'autre, dans le cas de cette série (figure 5). Étant donné que cette série chronologique a été traitée par transformation logarithmique, les erreurs-types peuvent être interprétées approximativement comme des pourcentages sur l'échelle d'origine (non transformée). Par rapport aux erreurs-types relatives aux séries désaisonnalisées, il y a de légères augmentations dans les erreurs-types des variations observées d'un mois à l'autre près des extrémités des séries, mais les erreurs-types des variations observées d'une année à l'autre ne présentent presque pas d'augmentation de ce genre. Ainsi, dans le cas de cette série et de ce filtre. l'incertitude relative à la variation de pourcentage observée d'un mois à l'autre et d'une année à l'autre dans le cas des données désaisonnalisées est presque constante tout au long de la série. Les erreurs-types des variations observées d'un mois à l'autre et d'une année à l'autre sont, dans les deux cas, environ 50 pour cent plus élevées que les erreurs-types relatives aux séries désaisonnalisées.

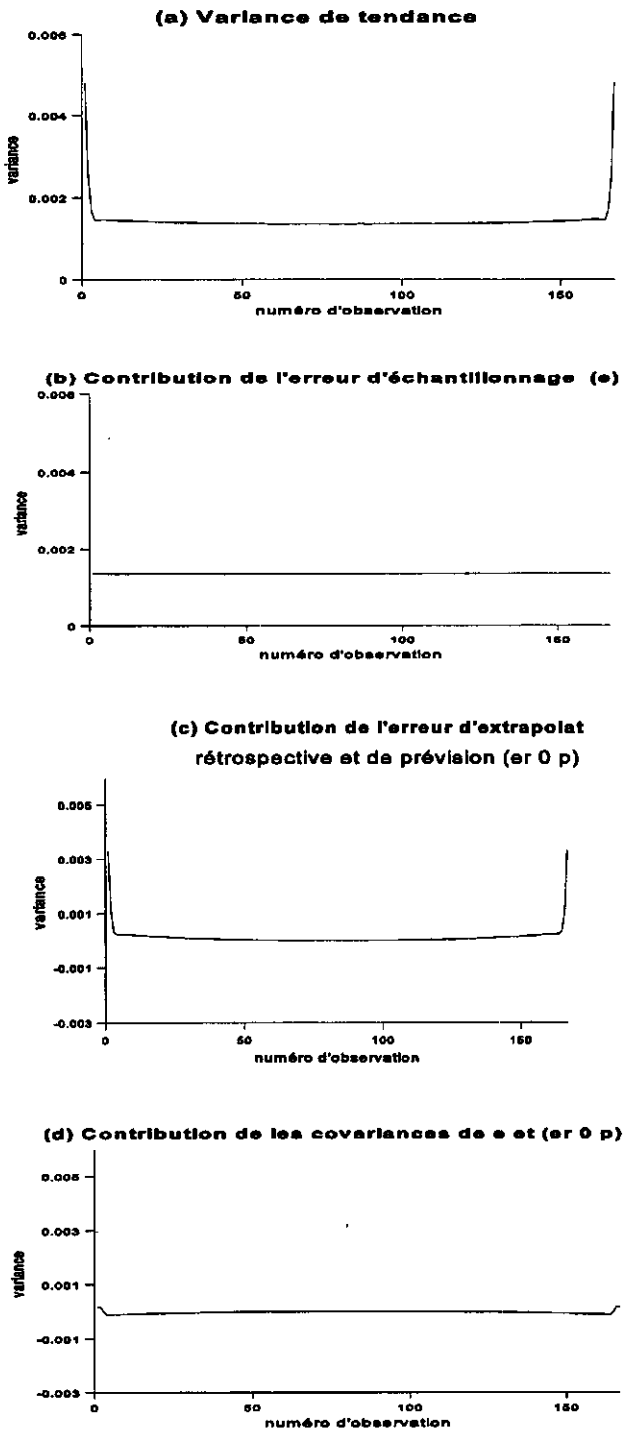


Figure 4. Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis: décomposition de la variance de l'estimation de tendance. La description du graphique est la même que celle de la figure 2. On remarquera le saut important dans la variance de l'estimation de tendance aux extrémités de la série, qui est dû à la contribution de l'erreur d'extrapolation rétrospective et de prévision (troisième graphique).

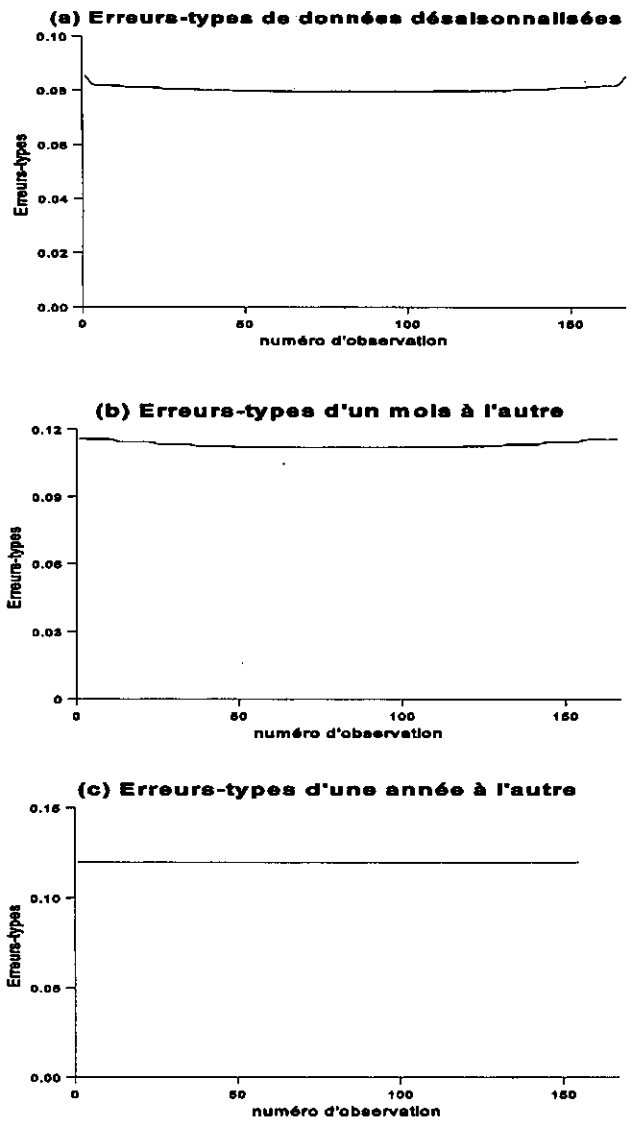


Figure 5. Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis: erreurs-types. Ces graphiques comparent les erreurs-types (et non la variance, comme c'est le cas dans les figures précédentes) relatives aux données désaisonnalisées (graphique supérieur) avec les erreurs-types plus grandes relatives aux estimations désaisonnalisées des variations observées d'un mois à l'autre (graphique médian) et d'une année à l'autre (graphique inférieur).

3.3 Variances des désaisonnalisations X-11 et effets de régressions estimés

La désaisonnalisation comporte souvent l'estimation de certains effets de régression, afin de tenir compte de facteurs comme les changements de calendriers, les interventions connues et les cas aberrants (Young 1965; Cleveland et Devlin 1982; Hillmer, Bell et Tiao, 1983; Findley, Monsell, Bell, Otto et Chen, présenté). (Les effets de cas aberrants sont souvent estimés de la même façon que les interventions connues, même si, idéalement, l'inférence concernant les cas aberrants devrait prendre en compte le fait que la série a été examinée afin de trouver les cas aberrants les plus «importants».) Dans la présente section, nous montrons comment les résultats déjà obtenus peuvent être élargis pour inclure la contribution à l'erreur de désaisonnalisation par l'erreur d'estimation touchant des paramètres de régression. Nous supposons toujours que les autres paramètres de modèle, qui déterminent les structures de covariance de $\mathbf{Y}$ et de $\mathbf{e}$, sont connus. Dans la pratique, ces autres paramètres seront également estimés, mais la prise en compte d'erreurs dans leur estimation est beaucoup plus difficile à faire. Une approche bayesienne qui permettrait d'effectuer cette prise en compte dans le contexte de la désaisonnalisation fondée sur des modèles est étudiée actuellement par Bell et Otto (présenté).

Nous élargissons le modèle utilisé pour Y_t, afin d'inclure les termes de régression, en écrivant $Y_t = \mathbf{x}_t'\beta + Z_t$, où x_t est le vecteur des variables de régression à l'époque t, β est le vecteur des paramètres de régression, et Z_t est la série de quantités de population réelle de laquelle on a enlevé les effets de régression. En élargissant notre notation de matrice-vecteur, nous écrivons $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}$, $\mathbf{Y}_o = \mathbf{X}_o\beta + \mathbf{Z}_o$, etc. La matrice de régression $\mathbf{X}$ peut être fractionnée d'après ses lignes correspondant aux périodes d'extrapolation rétrospective (er), d'observation (o) et de prévision (p): $\mathbf{X} = (\mathbf{X}_{er}' | \mathbf{X}_o' | \mathbf{X}_p')'$. Nous supposons que e_t a une valeur moyenne de zéro, de sorte que son modèle ne comporte aucun effet de régression. Nous avons alors $\mathbf{y} = \mathbf{Y} + \mathbf{e} = (\mathbf{X}\beta + \mathbf{Z}) + \mathbf{e}$, et c'est la division habituelle qui s'applique. En représentant la série y_t, après avoir enlevé les effets de régression, par la notation z_t, nous obtenons $\mathbf{z} = \mathbf{y} - \mathbf{X}\beta = \mathbf{Z} + \mathbf{e}$.

La matrice $\mathbf{X}$ et le vecteur β requièrent une division supplémentaire, parce que certains des effets de régression relatifs à $\mathbf{x}_t'\beta$ pourraient être attribués à la composante non saisonnière, tandis que d'autres effets, comme les effets des jours commerciaux ou des jours fériés peuvent être enlevés dans le cadre de la désaisonnalisation. Voir à ce sujet Bell, (1984b). Divisons $\mathbf{x}_t'$ de la façon suivante: $(\mathbf{x}_{St}' | \mathbf{x}_{Nt}')$ où $\mathbf{x}_{Nt}$ représente les variables de régression attribuées à la composante non saisonnière et $\mathbf{x}_{St}$ les variables dont les effets doivent être enlevés par la désaisonnalisation. De manière correspondante, divisons β, tel que $\mathbf{x}_t'\beta = \mathbf{x}_{St}'\beta_S + \mathbf{x}_{Nt}'\beta_N$ et $\mathbf{X}\beta = \mathbf{X}_S\beta_S + \mathbf{X}_N\beta_N = (\mathbf{X}_S | \mathbf{X}_N)\ (\beta_S' | \beta_N')'$. ($\mathbf{x}_{St}\beta_S$ est attribué à la composante saisonnière «combinée».) La matrice $\mathbf{X}$ peut donc être fractionnée de deux façons: d'après les effets saisonniers, par opposition aux effets non saisonniers, et d'après les périodes d'extrapolation rétrospective, d'observation et de prévision. Ainsi, nous écrivons

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{Ser} & \mathbf{X}_{Ner} \\ \mathbf{X}_{So} & \mathbf{X}_{No} \\ \mathbf{X}_{Sp} & \mathbf{X}_{Np} \end{bmatrix}.$$

Si β était connu, nous pourrions calculer $\mathbf{z}_o = \mathbf{y}_o - \mathbf{X}_o\beta = \mathbf{Z}_o + \mathbf{e}_o$, élargir cette série par extrapolation rétrospective et par prévision (série $\hat{\mathbf{z}}$ dite élargie), corriger $\hat{\mathbf{z}}$ par X-11 ($\boldsymbol{\Omega}\hat{\mathbf{z}}$), et ajouter à nouveau les effets de régression $\mathbf{X}_{No}\beta_{No}$ requis. L'objectif de la désaisonnalisation serait $\mathbf{X}_{No}\beta_{No}$ +

$\boldsymbol{\Omega Z} = \mathbf{X}_{No}\beta_{No} + \boldsymbol{\Omega}(\mathbf{Y} - \mathbf{X}\beta)$, et l'erreur de désaisonnalisation serait alors $(\mathbf{X}_{No}\beta_{No} + \boldsymbol{\Omega Z}) - (\mathbf{X}_{No}\beta_{No} + \boldsymbol{\Omega}\hat{\mathbf{z}}) = \boldsymbol{\Omega}(\mathbf{Z} - \hat{\mathbf{z}})$. Donc, si les paramètres de régression étaient connus, ceux-ci ne contribueraient pas à l'erreur de désaisonnalisation, et les résultats déjà donnés pourraient être utilisés pour calculer $\text{var}(\boldsymbol{\Omega}(\mathbf{Z} - \hat{\mathbf{z}}))$.

Dans la pratique, β est estimé en tant que partie du modèle ajusté, disons, par maximum de vraisemblance, en présumant une situation de normalité. Compte tenu des estimations des autres paramètres de modèle, et en considérant ces paramètres comme s'ils étaient connus, l'estimation de maximum de vraisemblance de β, ainsi que sa variance sont donnés par

$$\hat{\beta} = [\mathbf{X}_o'\boldsymbol{\Delta}_o'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Delta}_o\mathbf{X}_o]^{-1}\mathbf{X}_o'\boldsymbol{\Delta}_o'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Delta}_o\mathbf{y}_o \qquad (3.1)$$

$$\text{var}(\hat{\beta}) = [\mathbf{X}_o'\boldsymbol{\Delta}_o'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Delta}_o\mathbf{X}_o]^{-1}, \qquad (3.2)$$

où $\boldsymbol{\Delta}_o$ a la dimension $(n-d)\times n$, contenant la partie de la matrice $\boldsymbol{\Delta}$ plus grande qui sert au calcul des différences dans le cas de la série observée y_o. Les expressions (3.1) et (3.2) sont des résultats généralisés des moindres carrés, obtenus à l'aide de l'équation de régression dans le cas des données qui ont fait l'objet d'un calcul des différences, $\mathbf{w}_o = \boldsymbol{\Delta}\mathbf{y}_o = (\boldsymbol{\Delta}\mathbf{X}_o)\beta + (\mathbf{u}_o + \boldsymbol{\Delta}\mathbf{e}_o)$, où le terme d'erreur, $\mathbf{u}_o + \boldsymbol{\Delta}\mathbf{e}_o$ a comme matrice de covariance $\text{var}(\mathbf{w}_o) = \boldsymbol{\Sigma}_{22}$, qui est déterminée par les autres paramètres de modèle.

Les paramètres de régression estimés $\hat{\beta}$ étant donnés, la série désaisonnalisée serait obtenue en soustrayant des données les effets de régression estimés (exprimons la série qui en résulte comme suit: $\hat{\mathbf{z}}_o = \mathbf{y}_o - \mathbf{X}_o\hat{\beta}$), en élargissant cette série avec des prévisions et des extrapolations rétrospectives, à l'aide du modèle (exprimons la série qui en résulte comme suit: $\hat{\mathbf{z}} = [\hat{\mathbf{z}}_{er}, \hat{\mathbf{z}}_o, \hat{\mathbf{z}}_p]$), en appliquant la méthode X-11 à la série élargie ($\boldsymbol{\Omega}\hat{\mathbf{z}}$), et en réintroduisant les effets de régression estimés attribués à la composante non saisonnière ($\boldsymbol{\Omega}\hat{\mathbf{z}} + \mathbf{X}_{No}\hat{\beta}_{No}$). L'objectif de la désaisonnalisation est toujours $\mathbf{X}_{No}\beta_{No} + \boldsymbol{\Omega Z}$, tel que décrit précédemment. L'erreur de désaisonnalisation est alors $\mathbf{v} = (\mathbf{X}_{No}\beta_{No} + \boldsymbol{\Omega Z}) - (\boldsymbol{\Omega}\hat{\mathbf{z}} + \mathbf{X}_{No}\hat{\beta}_{No}) = \mathbf{X}_{No}(\beta_{No} - \hat{\beta}_{No}) + \boldsymbol{\Omega}(\mathbf{Z} - \hat{\mathbf{z}})$.

L'expression de $\mathbf{v}$ peut être simplifiée en récrivant $\hat{\mathbf{z}}$. Tout d'abord, posons $\mathbf{G} = [\mathbf{B}'|\mathbf{I}|\mathbf{F}']'$, où $\mathbf{F}$ est la matrice qui produit les prévisions $\hat{\mathbf{y}}_p$ à partir de $\mathbf{y}_o$, et $\mathbf{B}$ la matrice correspondante qui produit les extrapolations rétrospectives $\hat{\mathbf{y}}_{er}$ à partir de $\mathbf{y}_o$. Nous n'avons pas besoin d'expressions explicites pour $\mathbf{F}$ et $\mathbf{B}$. L'application de $\mathbf{G}$ à $\mathbf{z}_o$ produit $\hat{\mathbf{z}}$ tandis que $\mathbf{G}$ appliqué à $\hat{\mathbf{z}}_o$ produit $\hat{\mathbf{z}}$. Par conséquent, $\hat{\mathbf{z}} = \hat{\mathbf{z}} + (\hat{\mathbf{z}} - \hat{\mathbf{z}}) = \hat{\mathbf{z}} - [\mathbf{G}(\mathbf{y}_o - \mathbf{X}_o\beta) - \mathbf{G}(\mathbf{y}_o - \mathbf{X}_o\hat{\beta})] = \hat{\mathbf{z}} + \mathbf{G}\mathbf{X}_o(\beta - \hat{\beta})$. Il est à noter que l'on obtient $\mathbf{G}\mathbf{X}_o$ en appliquant la procédure permettant d'élargir une série par l'ajout de prévisions et d'extrapolations rétrospectives (à partir du modèle choisi pour z_t) à chaque colonne de $\mathbf{X}_o$. La méthode que nous avons utilisée pour effectuer cette opération est décrite dans l'annexe B. Ensuite, nous avons

$$\begin{aligned}\mathbf{v} &= \mathbf{X}_{No}(\beta_{No} - \hat{\beta}_{No}) + \boldsymbol{\Omega}[(\mathbf{Z} - \hat{\mathbf{z}}) - \mathbf{G}\mathbf{X}_o(\beta - \hat{\beta})] \\ &= \boldsymbol{\Omega}(\mathbf{Z} - \hat{\mathbf{z}}) + \{[\mathbf{0}|\mathbf{X}_{No}] - \boldsymbol{\Omega}\mathbf{G}\mathbf{X}_o\}(\beta - \hat{\beta}).\end{aligned}$$

Maintenant, $\mathbf{Z} - \hat{\mathbf{z}} = \mathbf{z} - \mathbf{e} - \hat{\mathbf{z}} = [\mathbf{er}|\mathbf{0}|\mathbf{p}] - \mathbf{e}$. Il est à noter que $[\mathbf{er}|\mathbf{0}|\mathbf{p}]$, le vecteur d'erreur obtenus de la projection de $\mathbf{z}$ sur $\mathbf{z}_o$ ou $\mathbf{y}_o$, est orthogonal (non corrélé) par rapport à $\beta - \hat{\beta}$, étant donné que $\hat{\beta}$ est une fonction linéaire des données $\mathbf{y}_o$. Par conséquent, en écrivant $\mathbf{K} = [\mathbf{0}|\mathbf{X}_{No}] - \boldsymbol{\Omega}\mathbf{G}\mathbf{X}_o$, nous obtenons la matrice de variance-covariance de l'erreur de désaisonnalisation, qui permet de prendre en compte l'erreur d'estimation de β:

$$\begin{aligned}\text{var}(\mathbf{v}) = {} & \boldsymbol{\Omega}\text{var}(\mathbf{Z} - \hat{\mathbf{z}})\boldsymbol{\Omega}' + \mathbf{K}\text{var}(\hat{\beta})\mathbf{K}' \\ & + \boldsymbol{\Omega}\text{cov}(\mathbf{e}, \hat{\beta})\mathbf{K}' + \mathbf{K}\text{cov}(\hat{\beta}, \mathbf{e})\boldsymbol{\Omega}'\end{aligned} \qquad (3.3)$$

où $\text{var}(\hat{\beta})$ est donné par (3.2). Dans (3.3), $\boldsymbol{\Omega}\text{var}(\mathbf{Z} - \hat{\mathbf{z}})\boldsymbol{\Omega}'$ est calculé à l'aide des résultats présentés dans la section 2, et le calcul de $\mathbf{K}\text{var}(\hat{\beta})\mathbf{K}'$ est simple, un fois que l'on a calculé $\mathbf{G}\mathbf{X}_o$. Le calcul des deux autres termes requiert

$$\begin{aligned}&\text{cov}(\hat{\beta}, \mathbf{e}) \\ &= \text{cov}([\mathbf{X}_o'\boldsymbol{\Delta}_o'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Delta}_o\mathbf{X}_o]^{-1}\mathbf{X}_o'\boldsymbol{\Delta}_o'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Delta}_o\mathbf{y}_o, \mathbf{e}) \\ &= \text{cov}([\mathbf{X}_o'\boldsymbol{\Delta}_o'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Delta}_o\mathbf{X}_o]^{-1}\mathbf{X}_o'\boldsymbol{\Delta}_o'\boldsymbol{\Sigma}_{22}^{-1}[\mathbf{u}_o + \boldsymbol{\Delta}_o\mathbf{e}_o], \mathbf{e}) \\ &= [\mathbf{X}_o'\boldsymbol{\Delta}_o'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Delta}_o\mathbf{X}_o]^{-1}\mathbf{X}_o'\boldsymbol{\Delta}_o'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Delta}_o[\mathbf{0}_{n\times m}|\mathbf{I}_{n\times n} \\ &\quad |\mathbf{0}_{n\times m}|\text{var}(\mathbf{e}).\end{aligned} \qquad (3.4)$$

Il est à noter que $[\mathbf{0}_{n\times m}|\mathbf{I}_{n\times n}|\mathbf{0}_{n\times m}]\text{var}(\mathbf{e}) = [\text{cov}(\mathbf{e}_o, \mathbf{e}_b) |\text{var}(\mathbf{e}_o)| \text{cov}(\mathbf{e}_o, \mathbf{e}_f)]$ représente les n lignes centrales de $\text{var}(\mathbf{e})$. À l'aide de (3.4) et des résultats précités, on peut calculer (3.3). Nous pouvons comparer les éléments diagonaux de $\text{var}(\mathbf{v})$ qui sont obtenus avec ceux de la somme des trois derniers termes de (3.3), afin de voir si la prise en compte de l'erreur due à l'estimation des paramètres de régression est importante.

Une précision importante doit être apportée en ce qui a trait aux résultats décrits dans la présente section. Étant donné que le premier terme du côté droit de (3.3), $\boldsymbol{\Omega}\text{var}(\mathbf{Z} - \hat{\mathbf{z}})\boldsymbol{\Omega}'$, représente la variance de la désaisonnalisation que nous obtiendrions en ne tenant pas compte de l'erreur d'estimation des paramètres de régression, il est tentant d'interpréter la somme des trois derniers termes de (3.3) comme étant la contribution à la variance de la désaisonnalisation de l'erreur relative à l'estimation des paramètres de régression. Malheureusement, cette somme n'est pas une variance en elle-même (en fait, elle peut être représentée sous la forme $\text{var}(\mathbf{K}\hat{\beta} + \boldsymbol{\Omega}\mathbf{e}) - \text{var}(\boldsymbol{\Omega}\mathbf{e})$), et elle peut donc être négative. Lorsque c'est le cas, les variances de désaisonnalisation qui tiennent compte de l'erreur relative à l'estimation des paramètres de régression sont en fait inférieures à celles qui ne tiennent pas compte de ce type d'erreur. Nous avons été en mesure d'obtenir un

tel résultat en modifiant artificiellement des paramètres de modèle dans l'exemple qui suit, comprenant des variables de jours commerciaux (bien que, comme le montrent les résultats, les effets soient assez faibles). Cette situation contraste avec des résultats comparables obtenus avec des méthodes fondées sur des modèles qui exprime l'erreur de désaisonnalisation comme la somme de deux termes orthogonaux: l'erreur présente lorsque tous les paramètres sont connus, ainsi que la contribution à cette erreur de l'estimation des paramètres de régression. Dans ce cas, la variance de la désaisonnalisation est donc la somme des variances de ces deux termes; par conséquent, la «contribution» de la régression est toujours non négative. Ce résultat est semblable à $\boldsymbol{\Omega}\text{var}(\mathbf{Z} - \hat{\mathbf{z}})\boldsymbol{\Omega}' + \mathbf{K}\ \text{var}(\hat{\beta})\mathbf{K}'$, dans (3.3). La difficulté dans (3.3) réside dans le fait que l'estimation $\boldsymbol{\Omega}\mathbf{z}$ obtenue à l'aide de X-11 n'est pas un estimateur optimal (EQMM) de l'objectif $\boldsymbol{\Omega}\mathbf{Z}$; c'est pourquoi l'erreur $\boldsymbol{\Omega}(\mathbf{Z} - \hat{\mathbf{z}})$ est corrélée avec $\hat{\beta}$ par l'erreur d'échantillonnage $\mathbf{e}$, ce qui mène aux deux termes de covariance dans (3.3). Cette situation résulte en partie de notre choix d'objectif ($\mathbf{X}\beta + \boldsymbol{\Omega}\mathbf{Z}$) et en partie du fait qu'on ne peut supposer que X-11 produit un estimateur optimal de quoi que ce soit (voir à cet égard les commentaires contenus dans l'introduction).

Exemple – Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis (suite). Nous utilisons le même exemple pour illustrer la contribution à l'erreur de désaisonnalisation de l'ajout de variables de jours commerciaux (Bell et Hillmer 1983), même si les coefficients de régression correspondants n'étaient pas statistiquement significatifs lorsqu'ils ont été estimés dans le cas de cette série. La figure 6a montre les résultats. Sur cette figure, la ligne inférieure représente la «contribution» à l'erreur de désaisonnalisation de l'estimation des effets de jour commerciaux (voir toutefois la remarque ci-dessus). Lorsqu'on ajoute ces effets à l'estimation initiale de la variance (ligne pointillée), nous obtenons la variance de la série désaisonnalisée et corrigée pour tenir compte des effets de jours commerciaux (ligne supérieure continue). On peut voir que, dans le cas de cet exemple, l'augmentation de la variance due à l'inclusion, dans le modèle, des effets estimés de jours commerciaux est faible. La figure 6b montre les résultats dans le cas du filtre d'estimations de tendance. Dans ce cas, la contribution à l'incertitude relative à la tendance attribuable à l'estimation des coefficients de jours commerciaux est certainement négligeable.

La contribution à la variance de la désaisonnalisation de l'ajout de trois variables de cas aberrants et d'une variable de changement de niveau est illustrée dans la figure 6c. Ces variables de régression ont été identifiées comme de possibles effets de cas aberrants à l'aide du programme Regarima (élaboré par le personnel des séries chronologiques (Time Series Staff) du U.S. Census Bureau) avec une statistique t critique de 2,5. Le programme Regarima comprend une méthode de détection de cas aberrants semblable à celles décrites par Bell (1983) ainsi que par Chang, Tiao et Chen (1988). Les contributions des cas aberrants additifs apparaissent sous forme de trois pics, tandis que la contribution du changement de niveau se présente sous la forme d'une seule petite bosse, au milieu de la série. Comparativement aux variables de régression relatives aux jours commerciaux, l'incidence de ces variables de cas aberrants est principalement locale, mais elle est aussi beaucoup plus forte. On constate notamment une incertitude supplémentaire au sujet de la désaisonnalisation, dans le cas d'observations considérées comme des cas aberrants additifs.

4. EXEMPLES

Nous illustrons notre méthode en utilisant plusieurs séries chronologiques économiques supplémentaires dont les erreurs d'échantillonnage respectives suivent des modèles différents. Les modèles utilisés pour ces exemples sont tirés de travaux antérieurs, tel qu'indiqué.

4.1 Ventes au détail des magasins à rayons

Les ventes au détail des magasins à rayons sont estimées dans le cadre de l'enquête mensuelle du Census Bureau qui porte sur le commerce de détail. Essentiellement, toutes les données sur les ventes proviennent de chaînes de magasins à rayons, qui sont toutes incluses dans l'enquête; par conséquent, les estimations ne comportent pratiquement aucune erreur d'échantillonnage. Ainsi, la variance de la désaisonnalisation X-11 provient uniquement d'erreurs de prévision ou d'extrapolation rétrospective, et d'erreurs liées à l'estimation d'effets de régression. (Il est à noter que la variance de la désaisonnalisation Wolter-Monsour serait zéro dans le cas de cette série.) Le modèle utilisé pour cette série (Bell et Wilcox 1993), pour la période allant d'août 1972 à mars 1989 et pour les logarithmes des observations, est le suivant: $(1-B)(1-B^{12})[Y_t - x_t'\beta] = (1-0{,}53B)(1-0{,}52B^{12})a_t$ avec $\sigma_a^2 = 4{,}32\times10^{-4}$, où x_t comprend des variables qui rendent compte des effets de jours commerciaux et de la fête de Pâques, et où $Y_t = y_t$ est le logarithme de la série initiale divisé par les facteurs de durée de mois. Pour effectuer la correction de la série, le Census Bureau utilise le filtre de correction X-11 par défaut et le filtre d'estimations de tendance à treize termes de Henderson.

La figure 7a montre les erreurs-types au fil du temps dans le cas des données désaisonnalisées, avec et sans la contribution des effets de régression. À la différence de la série des mises en chantier d'immeubles de cinq unités de logement ou plus, on constate des augmentations marquées des erreurs-types dans les données désaisonnalisées, aux extrémités de la série, ce qui est dû entièrement aux erreurs de prévision ou d'extrapolation rétrospective. La contribution à l'erreur-type due à l'estimation des effets de régression est également plus prononcée dans le cas de cette série. Une caractéristique intéressante que montre la figure 7 est l'ensemble de petits pics orientés vers le bas qui

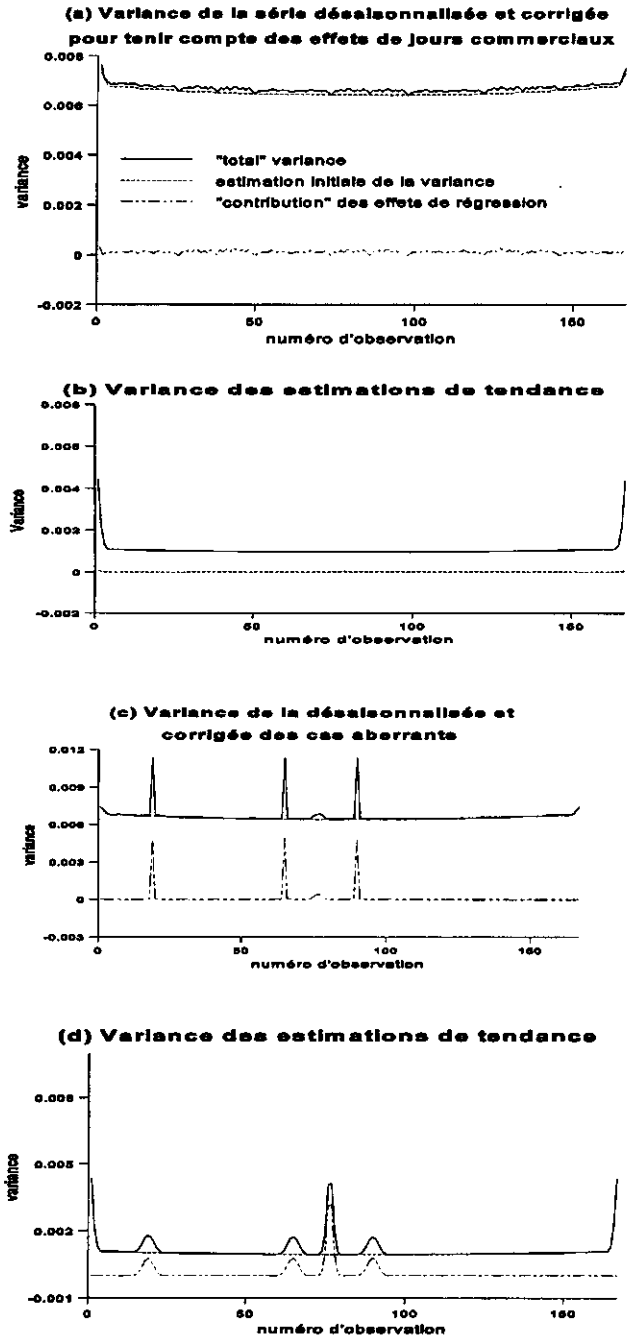


Figure 6. Mises en chantier d'immeubles à cinq unités de logement ou plus aux États-Unis: inclusion de la «contribution» des effets de régression dans les estimations de variance. Le graphique supérieur montre aussi bien la variance initiale tirée du premier graphique de la figure 2 (ligne pointillée) que la variance prenant en compte l'incertitude supplémentaire due à l'estimation des effets de régression de jours commerciaux (ligne continue supérieure). La contribution de variables de régression est également indiquée (ligne continue inférieure). Le deuxième graphique montre les résultats correspondants relatifs à la variance des estimations de tendance. On remarquera que la contribution de variables de régression à la variance de la désaisonnalisation est faible, et dans le cas de la variance des estimations de tendance, elle est pratiquement nulle. Les troisième et quatrième graphiques montrent des résultats analogues lorsque les effets de régression de jours commerciaux sont remplacés par trois cas aberrants additifs et un changement de niveau. Il est à noter que ces cas aberrants et ce changement de niveau ont une incidence locale importante sur la variance de la désaisonnalisation et des estimations de tendance.

surviennent par triplets à un an d'intervalle. Ces pics surviennent au mois de février des années non bissextiles, pour lequel il n'y a pas d'effet de jours commerciaux (les variables de régression de jours commerciaux ont toutes une valeur zéro). Il y a quand même une faible contribution de variables de régression à l'erreur de désaisonnalisation en ces points temporels, étant donné que la correction fait la moyenne dans le cas de ces contributions, à partir de points temporels adjacents. (Des inflexions survenant au mois de février d'années non bissextiles sont également visibles si on examine de près la figure 6a.) En outre, dans le cas de certaines années, l'erreur d'estimation de l'effet de la fête de Pâques produit un pic évident qui touche les mois de mars et avril.

La contribution relative de la régression aux erreurs-types de désaisonnalisation diminue vers les extrémités de la série. Cela est dû à deux facteurs: 1) l'importance de la contribution de variables de régression à var(v_t) diminue quelque peu vers les extrémités des séries; 2) ce qui est encore plus important, var($Z_t - \hat{Z}_t$) augmente de façon marquée vers ces extrémités, ce qui réduit la contribution relative à var(v_t) due à la régression (et cette diminution est accentuée lorsqu'on prend les racines carrés).

Le patron des erreurs-types des variations désaisonnalisées entre deux mois consécutifs (figure 7b) est semblable à celui de l'erreur-type des données désaisonnalisées (figure 7a). La contribution de variables de régression est légèrement plus importante que celle de l'erreur-type des données désaisonnalisées. Les erreurs-types des variations observées d'une année à l'autre (figure 7c) suivent des patrons similaires, mais la contribution de variables de régression est considérablement plus importante que dans le cas des variations observées d'un mois à l'autre, et cette contribution demeure importante aux extrémités des séries.

Nous avons effectué un ensemble de calculs semblables à l'aide du filtre d'estimations de tendance X-11 par défaut; les résultats relatifs aux erreurs-types des estimations de tendance, avec et sans la contribution de variables de régression, sont présentés dans la figure 7d. Les patrons suivis au fil du temps par ces erreurs-types sont semblables à ceux des chiffres correspondants relatifs à la séries des mises en chantier d'immeubles à cinq unités de logement et plus, mais les erreurs-types sont beaucoup plus petites, en raison de l'absence d'erreur d'échantillonnage. La contribution des effets de régression est faible.

Les erreurs-types dans le cas de tous les tracés de la figure 7 sont faibles (aucune n'excède 0,8 pour cent). Dans le cas de cette série, la contribution des effets de régression est peu importante et probablement négligeable près des extrémités de la série, sauf dans le cas des variations observées d'une année à l'autre. Cependant, au milieu de la série, la seule contribution aux erreurs-types est celle due aux effets de régression.

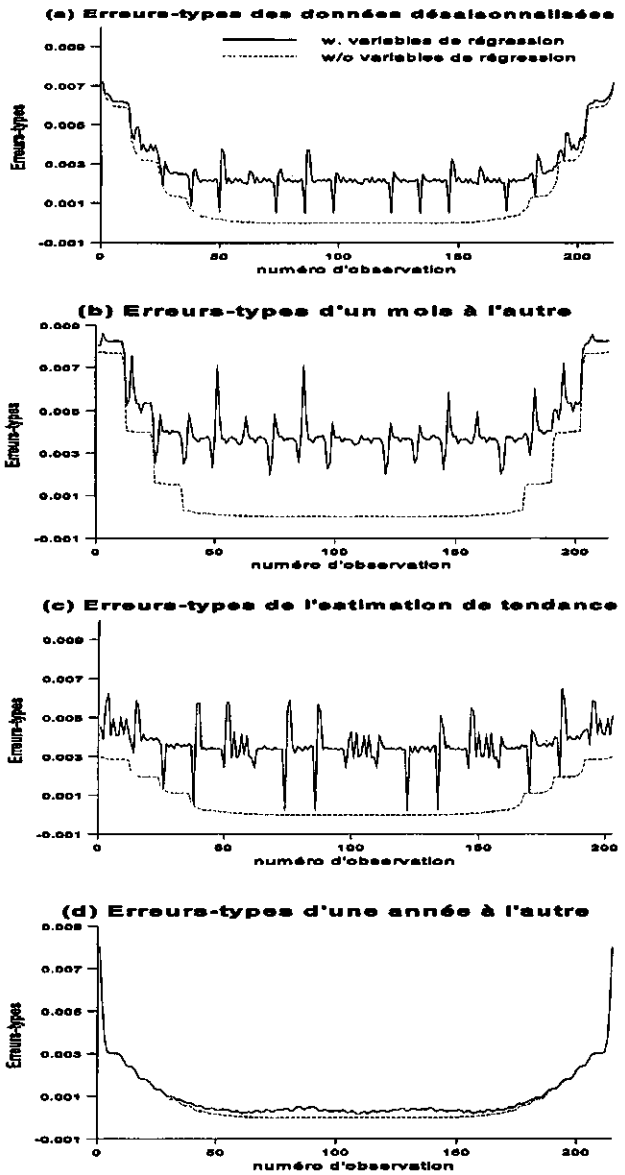


Figure 7. Magasins à rayons américains, avec effets de jours commerciaux et de Pâques. Cette série ne comporte pas d'erreur d'échantillonnage. Les quatre graphiques indiquent les erreurs-types, avec et sans la contribution de l'estimation des effets de régression. Dans le cas des données désaisonnalisées et des variations correspondantes observées d'un mois à l'autre et d'une année à l'autre (trois premiers graphiques) la «contribution» de l'estimation des effets de régression est substantielle et erratique dans le milieu de la série (où cette estimation est le seul élément contributif), mais à chacune des extrémités de la série, elle diminue, pour les raisons expliquées précédemment. La contribution des effets de régression aux erreurs-types de l'estimation de tendance est faible.

4.2 Chômage chez les adolescents

Le Bureau of Labor Statistics (BLS) publie la série chronologique mensuelle du nombre d'adolescents et d'adolescentes américains au chômage, qui est estimé à partir des données de la Current Population Survey (CPS). Les données relatives à la période allant de janvier 1972 à décembre 1983 ($n = 144$) ont été utilisées par Bell et Hillmer (présenté) afin d'élaborer un modèle pour la série en question. La variance de l'erreur d'échantillonnage h_t^2 varie au fil du temps; elle est donc non stationnaire. Le modèle d'erreur d'échantillonnage qu'ont élaboré ces auteurs est le suivant:

$$e_t = h_t \tilde{e}_t \text{ où } (1 - 0{,}6B)\tilde{e}_t = (1 - 0{,}3B)b_t, \qquad (4.1)$$

avec $\sigma_b^2 = 0{,}87671$, de sorte que $\text{var}(\tilde{e}_t) = 1$. La variance de l'erreur d'échantillonnage de la CPS peut être approximée par des fonctions de variance généralisées (Wolter 1985, chap. 5; Hanson 1968). La fonction de variance généralisée qu'ont utilisée Bell et Hillmer dans le cas de la série du chômage chez les adolescents est la suivante:

$$h_t^2 = 1.971\, y_t - (1.53 \times 10^{-5}) y_t^2, \qquad (4.2)$$

où y_t est l'estimation du nombre en milliers d'adolescents en chômage à l'époque t. Le modèle estimatif pour la composante de signal Y_t est

$$(1 - B)(1 - B^{12})Y_t = (1 - 0.27B)(1 - 0.68B^{12})a_t, \qquad (4.3)$$

avec $\sigma_a^2 = 4294$. Ce modèle ne comporte pas d'effets de régression et la série n'est pas transformée. Le BLS utilise le filtre de désaisonnalisation X-11 par défaut (donc $m = 84$). Si on applique la méthode décrite dans le présent article à cet exemple, on rencontre des difficultés parce que la variance (estimée) de l'erreur d'échantillonnage h_t^2 dépend de l'estimation y_t par l'intermédiaire de la fonction de variance généralisée (4.2). Dans les périodes d'extrapolation rétrospective et de prévision, y_t est inconnu. Afin d'obtenir h_t^2 pour ces périodes, nous extrapolons rétrospectivement et prévoyons y_t en utilisant un simple modèle ARIMA$(0\ 1\ 1)(0\ 1\ 1)_{12}$ pour y_t (et non pour Y_t, comme dans (4.3)). Nous avons utilisé par la suite les 84 prévisions et extrapolations rétrospectives ainsi obtenues dans (4.2), afin de produire h_t^2 dans les périodes de prévision et d'extrapolation rétrospective. Des traitements plus affinés sont possibles, comme l'utilisation du modèle de composante donné par (4.1) et (4.3) pour prévoir y_t.

La variance de la désaisonnalisation relative à cette série (figure 8a) est dominée par la contribution de l'erreur d'échantillonnage (figure 8b) à la plupart des points temporels t. Il en est ainsi parce que même si la contribution de var(**er**, **0**, **p**) est substantielle dans le cas de la série en question (figure 8c), elle tend à être neutralisée par la contribution de cov[(**er**,**0**,**p**), **e**] + cov[(**er**,**0**,**p**), **e**]′ (figure 8d), sauf à quelques points temporels initiaux et finals. Les patrons de la variance relative aux variations désaisonnalisées observées d'un mois à l'autre et d'un an à l'autre (non indiquée) sont semblables au patron que montre la figure 8a. La variance relative aux variations observées d'un mois à l'autre est légèrement plus grande que dans le cas de la série désaisonnalisée, et celle relative aux variations observées d'un an à l'autre l'est encore davantage.

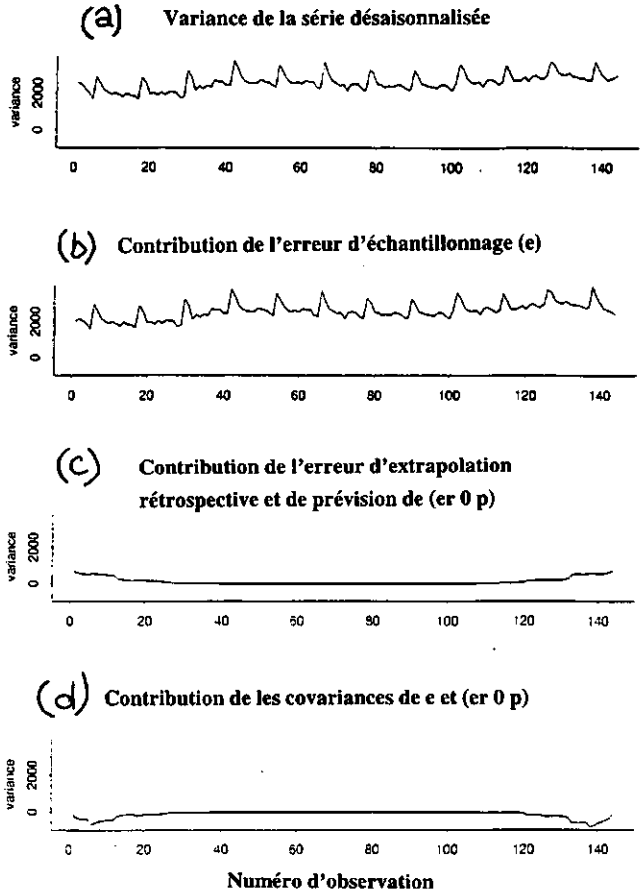


Figure 8. Chômage chez les adolescents. avec options X-11 par défaut. La description des graphiques est la même que pour la figure 2. Le patron saisonnier de la contribution de la variance de l'erreur d'échantillonnage (deuxième graphique) résulte de sa dépendance vis-à-vis du niveau de la série par l'intermédiaire d'une fonction de variance généralisée (voir texte).

4.3 Ventes au détail des débits de boissons

Les ventes au détail des débits de boissons sont estimées dans le cadre de l'enquête mensuelle du Census Bureau qui porte sur le commerce de détail. Dans cette enquête, les échantillons (de non-certitude) sont remaniés de manière indépendante, environ tous les cinq ans, de sorte que la matrice de covariance des erreurs d'échantillonnage est du type diagonale par blocs. Bell et Hillmer (1990) ont élaboré le modèle suivant pour l'erreur d'échantillonnage des séries soumises à une transformation logarithmique à l'intérieur d'un échantillon donné:

$$(1-0{,}75B-0{,}66B^3+0{,}50B^4)(1-0{,}71B^{12})e_t = (1+0{,}13B)b_t, \quad (4.4)$$

avec $\sigma_b^2 = 9{,}301 \times 10^{-5}$. Dans le cas des points temporels t et j, dans des échantillons différents, $\text{cov}(e_t, e_j) = 0$. Bell et Hillmer ont créé un modèle pour la composante de signal des séries précitées à partir d'estimations non étalonnées relatives à la période allant de septembre 1977 à décembre 1986. Nous allons plutôt utiliser le modèle suivant, ajusté par Bell et Wilcox (1993), avec des données supplémentaires recueillies jusqu'au mois d'octobre 1989.

$$(1-B)(1-B^{12})[Y_t - X_t'\beta] = (1-0{,}23B)(1-0{,}88B^{12})a_t,$$

où X_t contient des variables de régression de jours commerciaux, avec $\sigma_a^2 = 4{,}16 \times 10^{-4}$.

Pour la désaisonnalisation de cette série, nous utilisons les filtres X-11 par défaut. La contribution de l'erreur due à l'estimation de paramètres de régression est faible dans le cas de cette série; elle n'est donc pas incluse dans les résultats qui suivront. Étant donné que la contribution de l'erreur d'échantillonnage est beaucoup plus forte que la contribution des erreurs de prévision ou d'extrapolation rétrospective, et plus importante également que l'incidence de $\text{cov}[(\mathbf{er}, \mathbf{0}, \mathbf{p}), \mathbf{e}] + \text{cov}[(\mathbf{er}, \mathbf{0}, \mathbf{p}), \mathbf{e}]'$, nous n'illustrons pas non plus ces autres contributions à la variance. La figure 9a indique l'erreur-type des données désaisonnalisées (indiquée sur 232 observations, afin de mieux illustrer le patron, qui comporte des lignes verticales indiquant la sélection d'un nouvel échantillon), et la figure 9b montre l'erreur-type relative aux variations désaisonnalisées observées d'un mois à l'autre.

On remarquera le patron marqué dans les figures 9a et 9b, qui est dû à la sélection d'un nouvel échantillon tous les cinq ans. Ce remaniement produit notamment un grand pic dans l'erreur-type relative aux variations désaisonnalisées observées d'un mois à l'autre (figure 9b). Des sauts similaires dans les écarts-types relatifs aux variations observées d'une année à l'autre surviennent dans le cas de la première année d'un nouvel échantillon. Nous avons constaté également des patrons similaires dans le cas d'autres séries tirées de l'enquête sur le commerce de détail, en utilisant des modèles proposés par Bell et Wilcox (1993).

L'analyse et les résultats qui précèdent ne prennent pas en considération certains aspects de la façon dont est effectuée l'estimation pour l'enquête sur le commerce de détail. En fait, afin d'éviter des augmentations importantes de la variance relative à des variations observées à l'époque de la sélection d'un nouvel échantillon, comme celles que montre la figure 9, nous apportons une simple modification aux estimations d'un échantillon nouvellement introduit, afin de rendre leur niveau cohérent par rapport à celui de l'ancien échantillon. La version la plus simple de cette modification est décrite ci-après. Soient $z_{(\text{ancien})t}$ ($= \exp(y_{(\text{ancien})t})$) les estimations tirées de l'ancien échantillon, et $z_{(\text{nouveau})t}$ les estimations non modifiées provenant du nouvel échantillon. Présumons que l'ancien échantillon fournit des estimations de $t \le \tau$, et que le nouvel échantillon doit fournir des estimations de $t > \tau$. Afin de fournir des données de chevauchement pour la modification, le nouvel échantillon est démarré un mois plus tôt, de sorte que l'on dispose tant de $z_{(\text{ancien})\tau}$ que de $z_{(\text{nouveau})\tau}$. Les estimations modifiées du nouvel échantillon sont définies comme suit: $z'_{(\text{nouveau})t} = z_{(\text{nouveau})t}(z_{(\text{ancien})\tau} / z_{(\text{nouveau})\tau})$, pour $t \ge \tau$. Cette modification est

effectuée chaque fois qu'un nouvel échantillon est introduit. En ce qui a trait aux estimations enregistrées correspondantes y_t, la modification s'écrit $y'_{(\text{nouveau})t} = y_{(\text{nouveau})t} + (y_{(\text{ancien})\tau} - y_{(\text{nouveau})\tau})$. Étant donné que la modification de y_t est linéaire, il est facile de prendre en compte ses effets dans le calcul de la variance de la désaisonnalisation. La variation d'un mois à l'autre à l'époque $\tau + 1$, avant la modification (et sans désaisonnalisation) est $y_{(\text{nouveau})\tau+1} - y_{(\text{ancien})\tau}$. Il est à noter que cette variation présente une variance importante, étant donné que $y_{(\text{nouveau})\tau+1}$ et $y_{(\text{ancien})\tau}$ proviennent d'échantillons indépendants différents. Après la modification, cette variation s'écrit $y_{(\text{nouveau})\tau+1} - y_{(\text{nouveau})\tau}$, et elle présente une variance beaucoup plus faible, en raison d'une forte corrélation positive entre $y_{(\text{nouveau})\tau+1}$ et $y_{(\text{nouveau})\tau}$ (résultant du modèle d'erreur d'échantillonnage (4.4)). Les estimations non désaisonnalisées de variations observées d'un mois à l'autre dans le cas de points temporels autres que $\tau + 1$ ne sont pas touchées par cette modification.

La figure 9d montre que la modification d'estimations provenant du nouvel échantillon supprime les augmentations importantes de l'écart-type de variations désaisonnalisées observées d'un mois à l'autre aux points de transition vers de nouveaux échantillons. Des effets similaires ont été constatés dans le cas de variations observées d'une année à l'autre sur une période d'un an. Le prix à payer pour cette amélioration est une erreur qui augmente de façon constante dans les estimations de niveau (figure 9c) après l'introduction de nouveaux échantillons. Cela se produit parce que cette modification introduit dans les estimations de niveau une erreur temporaire qui perdure dans l'ensemble du nouvel échantillon. Ainsi, la modification échange de moins bonnes estimations de niveau contre des améliorations dans les estimations de variation. (La figure 9c ne montre aucune augmentation pour les cinq premières années, parce que nous présupposons que ces estimations ne sont pas modifiées pour concorder avec celles d'un échantillon antérieur.) De plus, les patrons marqués que montre la figure 9a se produisent parce que les erreurs d'échantillonnage relatives à des estimations non modifiées d'échantillons adjacents ne sont pas corrélées. D'autre part, les erreurs d'échantillonnage présentes dans les estimations modifiées sont corrélées de manière assez forte entre échantillons adjacents. Cette situation donne lieu, après l'application du filtre de désaisonnalisation, à un patron très différent (une patron presque inexistant) pour les cinq premières années indiquées dans la figure 9c, et de légères oscillations autour de l'augmentation linéaire ultérieure.

Les erreurs-types relatives aux estimations de tendance et aux variations X-11 (non indiquées) ressemblent à des versions lissées de celles indiquées dans la figure 9.

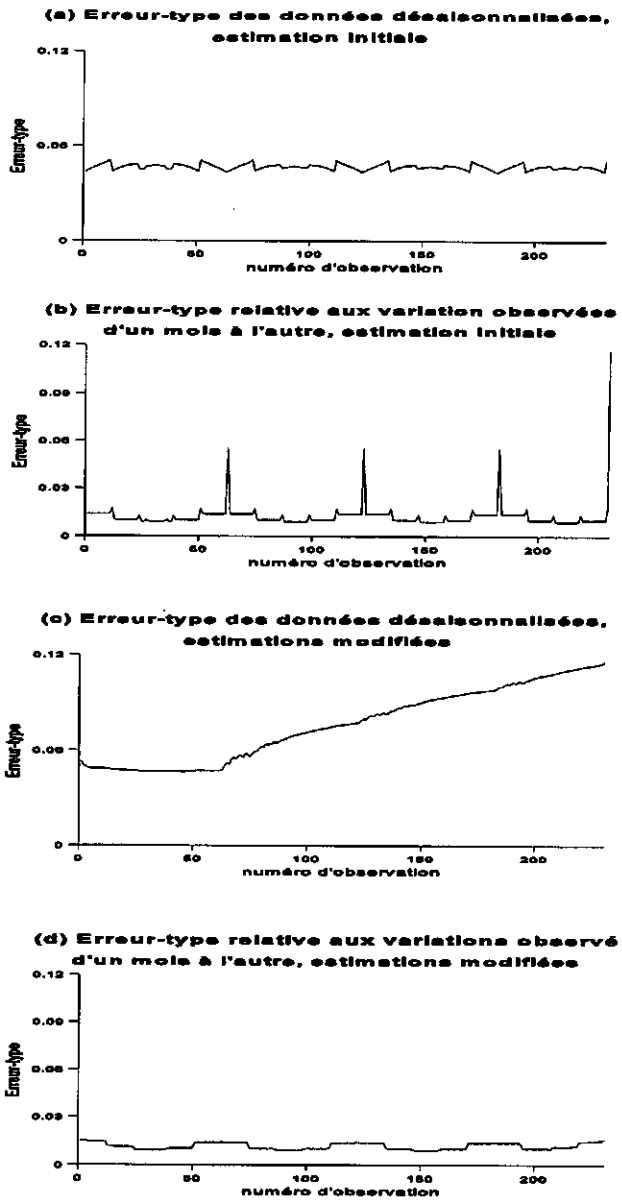


Figure 9. Ventes au détail des débits de boissons: sélection d'un nouvel échantillon tous les cinq ans. Le graphique supérieur montre l'erreur-type des données désaisonnalisées et le deuxième graphique indique l'erreur-type des variations correspondantes observées d'un mois à l'autre. Le patron marqué résulte de la sélection indépendante d'un nouvel échantillon tous les cinq ans (aux points correspondant aux lignes pointillées verticales). Dans le cas des variations observées d'un mois à l'autre, ce fait entraîne d'importantes augmentations de l'erreur-type au moment de la sélection d'un nouvel échantillon. Afin d'éviter cet inconvénient, le nouvel échantillon est sélectionné de manière à chevaucher l'ancien durant un mois ou davantage, et les estimations relatives à ce nouvel échantillon sont modifiées à l'aide de données tirées du chevauchement, afin de rendre leur niveau cohérent par rapport à celui des estimations provenant de l'ancien échantillon (voir texte). Cette façon de procéder permet d'éviter l'augmentation de l'erreur-type d'estimations de variation au moment de la sélection du nouvel échantillon (quatrième graphique), mais elle introduit une erreur temporaire dans les estimations de niveau modifié, dont les effets s'accumulent au fil du temps (troisième graphique)

Dans la pratique, les estimations finales provenant de l'enquête sur le commerce de détail présentent une complexité encore plus grande que celle que nous venons de décrire et d'illustrer. Tout d'abord, on recueille plus d'un mois de données de chevauchement, qui peuvent être utilisées pour modifier des estimations de niveau lors de l'introduction d'un nouvel échantillon. Ensuite, ce qui est encore plus important, les estimations mensuelles sont étalonnées, pour qu'elles concordent avec les totaux annuels obtenus dans le cadre de l'enquête annuelle (plus précise) sur le commerce de détail ou lors du dénombrement économique quinquennal. L'étalonnage devrait donc limiter l'augmentation, que montre la figure 9c, de la variance des estimations de niveau au fil du temps. Cependant, étant donné que l'étalonnage impose des contraintes de somme linéaires sur les estimations d'origine (non-transformées), ses effets sur la désaisonnalisation sont difficiles à étudier dans le cadre de la méthode présentée ici, et nous ne les avons donc pas étudiés. (Nous avons utilisé un modèle pour données non étalonnées, afin d'éviter ce problème.) Durbin et Quenneville (1995) ont élaboré une méthode fondée sur un modèle qui tient compte des non-linéarités que de telles contraintes d'étalonnage imposent aux données soumises à une transformation logarithmique.

5. CONCLUSIONS

Dans la présente communication, nous avons décrit une méthode qui permet de s'attaquer au vieux problème lié à l'obtention de la variance dans le cas de la désaisonnalisation effectuée à l'aide de la méthode X-11. Notre but était la création et l'application d'une méthode permettant d'obtenir une variance qui rend compte de deux sources d'erreurs. La première de ces sources, est l'erreur d'échantillonnage (e_t), qui survient parce que nous n'observons pas la série réelle, Y_t, mais plutôt les estimations $y_t = Y_t + e_t$ tirées d'une enquête répétée. La deuxième source d'erreurs résulte de la nécessité d'élargir la série observée avec des prévisions et des extrapolations rétrospectives pour pouvoir appliquer les filtres X-11 symétriques. Cette deuxième source d'erreurs mène à des révisions des corrections. Afin de tenir compte de ces deux sources d'erreurs, nous avons défini la variance de la désaisonnalisation comme étant la variance de l'erreur qui entache l'utilisation de la correction X-11 visant à estimer une cible précise. Cette cible, $\omega(B)Y_t$, est ce que l'on obtiendrait en appliquant le filtre X-11 symétrique linéaire, $\omega(B)$, à la série réelle, si les valeurs de celles-ci étaient connues suffisamment en avant dans le futur et en arrière dans le passé pour pouvoir utiliser le filtre symétrique. (L'application à X-11 additif avec prévisions et extrapolations rétrospective est immédiate, et X-11 log-additive est considérée comme une approximation de X-11 multiplicative.)

Nous avons également appliqué notre méthode pour obtenir la variance d'estimations X-11 de tendance et la variance de variations observées d'un mois à l'autre et d'une année à l'autre, tant dans le cas des données désaisonnalisées que dans celui des estimations de tendance. Nous avons effectué un autre élargissement pour tenir compte de l'erreur liée à l'estimation de paramètres de régression (par exemple, pour modéliser des effets de calendrier), bien que cette opération soit plus complexe et comporte certaines limites.

Les variances que nous obtenons ne tiennent pas compte de l'incertitude due à la variation des séries chronologiques dans les composantes saisonnières et non saisonnières. Dans la section 2, nous avons fait valoir que cette façon de procéder pourrait être appropriée pour des utilisateurs types de données désaisonnalisées à l'aide de la méthode X-11. Cependant, si l'on souhaite tenir compte de cette variation des séries chronologiques, nous suggérons la prise en considération de méthodes de désaisonnalisation fondées sur des modèles, étant donné que les séries chronologiques offrent un moyen de rendre compte de manière explicite des variations relatives à toutes les composantes. Comme autre approche, Pfeffermann (1994) a élaboré une méthode de traitement de la désaisonnalisation X-11 dans laquelle on essaie de prendre en compte la variation irrégulière et l'erreur d'échantillonnage.

Notre méthode s'appuie sur la première méthode proposée par Wolter et Monsour (1981), en prenant en considération les erreurs d'extrapolation rétrospective et de prévision, qui n'a pas été considérée par ces auteurs. Une autre façon de voir la différence existant entre notre méthode et celle de Wolter et Monsour est que nous définissons un objectif de désaisonnalisation cohérent, tandis qu'en utilisant les filtres X-11 asymétriques, Wolter et Monsour ont utilisé implicitement des objectifs qui varient au fil du temps. C'est pourquoi notre méthode évite la caractéristique irréaliste d'une variance qui diminue vers la fin des séries, qui peut être observée dans les résultats de Wolter et Monsour, et également dans ceux de Pfeffermann.

Dans les résultats empiriques que nous avons présentés, la contribution de l'erreur d'échantillonnage domine souvent la variance relative à la désaisonnalisation. Cette prédominance est due en partie au fait que l'erreur d'échantillonnage était souvent importante par rapport à l'erreur de prévision et d'extrapolation rétrospective, et en partie au fait que la contribution de l'erreur de prévision et d'extrapolation rétrospective avait tendance à être neutralisée par la contribution de la covariance de cette même erreur avec l'erreur d'échantillonnage. D'autre part, les résultats empiriques relatifs à la variance d'estimations de tendance ont montré des augmentations importantes aux extrémités des séries, en raison des effets de l'erreur de prévision et d'extrapolation rétrospective. Étant donné que la «contribution » la plus forte de l'erreur de prévision et d'extrapolation rétrospective a lieu aux extrémités des séries, et que c'est la variance relative aux désaisonnalisations et aux estimations de tendance les plus récentes qui

présente le plus d'intérêt, il faut tenir compte de la contribution de ce type d'erreur.

La contribution relative à la variance de l'erreur d'estimation de coefficients de régression de jours commerciaux ou de jours fériés avait tendance à être faible, sauf si la série ne comportait pas d'erreur d'échantillonnage. L'erreur due à l'estimation des effets des cas aberrants additifs et des changements de niveau était importante autour des points temporels des cas aberrants. L'incidence des cas aberrants additifs était importante sur la variance relative à la désaisonnalisation, tandis que les effets des changements de niveau étaient importants dans le cas de la variance des estimations de tendance.

Les composantes non stationnaires des erreurs d'échantillonnage produisent des patrons intéressants dans le cas de variances relatives à la désaisonnalisation et aux estimations de tendance. Nous avons examiné deux types de composantes non stationnaires d'erreurs d'échantillonnage. Les patrons saisonniers des variances d'erreur d'échantillonnage produisaient des patrons saisonniers correspondants dans les variances relatives à la désaisonnalisation. Les sélections indépendantes de nouveaux échantillons, qui produisaient des erreurs d'échantillonnage corrélés à l'intérieur des échantillons mais non entre ceux-ci, étaient à l'origine, au fil du temps et à l'intérieur d'un échantillon, de patrons erratiques dans les variances relatives à la désaisonnalisation et aux estimations de tendance. Ces patrons se répétaient de façon approximative dans des échantillons différents si ceux-ci restaient en vigueur durant des laps de temps à peu près égaux.

Les calculs pour les exemples que nous avons choisis (compte tenu des modèles ajustés, qui proviennent des sources citées) ont été effectués en programmant les expressions contenues dans les sections 2 et 3 en langage de programmation statistique S+. Le code informatique correspondant est disponible sur demande.

REMERCIEMENTS

Nous remercions Nash Monsour d'avoir mis à notre disposition le programme FORTRAN pour calculer les poids de filtres X-11 asymétriques par défaut, et de nous avoir patiemment expliqué comment sont modifiées les estimations tirées des enquêtes mensuelles sur le commerce de détail lorsqu'on sélectionne de nouveaux échantillons. La présente communication contient les résultats généraux des recherches effectuées par du personnel du Census Bureau. Les opinions exprimées sont celles des auteurs et ne reflètent pas nécessairement celles du Census Bureau.

ANNEXE A

Plusieurs expressions calculées dans la présente communication prennent la forme générale

$$\mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{B} \qquad \text{(A.1)}$$

où $\boldsymbol{\Sigma}$ est une matrice positive définie, et où $\mathbf{A}$ et $\mathbf{B}$ sont conformes à $\boldsymbol{\Sigma}$. Soit $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$ la décomposition de Cholesky de $\boldsymbol{\Sigma}$. Alors $\mathbf{A}\sum^{-1}\mathbf{B} = \mathbf{A}(\mathbf{L}^{-1})'\mathbf{L}^{-1}\mathbf{B}$ et (A.1) peut être calculé comme suit:

(1) Résoudre $\mathbf{L}\mathbf{Q}_1 = \mathbf{B}$ pour $\mathbf{Q}_1$

(2) Résoudre $\mathbf{L}\mathbf{Q}_2 = \mathbf{A}'$ pour $\mathbf{Q}_2$

(3) Calculer $\mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{B} = \mathbf{Q}_2\mathbf{Q}_1$.

Les expressions (1) et (2) peuvent être résolues efficacement puisque $\mathbf{L}$ est triangulaire inférieur.

ANNEXE B

Deux opérations sont nécessaires pour obtenir $\mathbf{G}\mathbf{X}_o$, utilisé dans la section 3.3. La première opération produit des extensions de «prévision» et d'«extrapolation rétrospective» des variables de régression ayant fait l'objet de calculs des différences. Dans la deuxième opération, on utilise ces résultats et l'équation aux différences finies pour produire les extensions de prévision et d'extrapolation rétrospective des variables de régression initiales (n'ayant pas fait l'objet d'un calcul des différences).

Soit $\mathbf{R}_o = \boldsymbol{\Delta}_o\mathbf{X}_o$, où $\boldsymbol{\Delta}_o$ est la partie de la matrice $\boldsymbol{\Delta}$ qui différencie les séries observées $\mathbf{y}_o$. De façon analogue au calcul de $\hat{\mathbf{w}}_p$ et $\hat{\mathbf{w}}_b$ décrit dans la section 2.2, les extensions de prévision des variables de régression différenciés sont calculées comme suit: $\mathbf{R}_p = \boldsymbol{\Sigma}_{32}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{R}_o$; quant aux extensions d'extrapolation rétrospective, elles sont calculées de la façon suivante: $\mathbf{R}_{er} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{R}_o$. $\mathbf{R}_p$ et $\mathbf{R}_{er}$ ont la forme (A.1) et peuvent être calculés avec la méthode précitée.

Pour la deuxième opération, soit x_t n'importe laquelle des variables de régression contenues dans $\mathbf{X}$. Notons les extensions de prévision requises $\hat{x}_{n+l}$ dans le cas de $l = 1, 2, ..., m$. Posons l'opérateur $\delta(B) = 1 - \delta_1 B - ... - \delta_d B^d$, et notons $\hat{r}_{n+l}$ l'extension de prévision de $\delta(B)x_t = r_t$ à l'époque $n+l$ ($\hat{r}_{n+l}$ est un élément de $\mathbf{R}_p$). Les extensions $\hat{x}_{n+l}$ sont déterminées par calcul itératif comme suit:

$$\hat{x}_{n+l} = \delta_1\hat{x}_{n+l-1} + ... + \delta_d\hat{x}_{n+l-d} + \hat{r}_{n+l}, \text{ pour } l = 1, ..., m,$$

où $\hat{x}_{n+j} = x_{n+j}$ si $j \le 0$.

Les extensions requises d'extrapolation rétrospective x_t sont notées $\hat{x}_{1-l}$ dans le cas de $l = 1, ..., m$. Ces extensions sont également obtenues par des calculs récurrents, à partir de l'équation aux différences finies $\delta(B)\hat{x}_t = \hat{r}_t$, en résolvant $\hat{x}_{1-l}$ dans l'expression

$$\hat{x}_{d+1-l} = \delta_1\hat{x}_{d-l} + \cdots + \delta_d\hat{x}_{1-l} + \hat{r}_{d+1+l}$$

et en reportant les extrapolations rétrospectives calculées précédemment. Ainsi

$$\hat{x}_{1-l} = \delta_d^{-1}(\hat{x}_{d+1-l} - \delta_1\hat{x}_{d-l} - \ldots - \delta_{d-1}\hat{x}_{2-l} - \hat{r}_{d+1-l}),$$

pour $l = 1, \ldots, m$,

où $\hat{x}_{1-j} = x_{1-j}$ lorsque $j \leq 0$.

BIBLIOGRAPHIE

BELL, W. R. (1983). A computer program for detecting outliers in time series. Dans *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 634-639.

BELL, W. R. (1984a). Signal extraction for nonstationary time series. *Annals of Statistics*, 12, 646-664.

BELL, W. R. (1984b). Seasonal Decomposition of Deterministic Effects. Rapport de recherche No 84/01. Statistical Research Division, Bureau of the Census.

BELL, W. R., et HILLMER, S. C. (1983). Modeling time series with calendar variation. *Journal of the American Statistical Association*, 78, 526-534.

BELL, W. R., et HILLMER, S. C. (1984). Issues involved with the seasonal adjustment of economic time series (avec discussion). *Journal of Business and Economic Statistics*, 2, 291-320.

BELL, W. R., et HILLMER, S. C. (1988). A matrix approach to signal extraction and likelihood evaluation for ARIMA component time series models. Rapport de recherche No 88/22, Statistical Research Division, Bureau of the Census.

BELL, W. R., et HILLMER, S. C. (1990). Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques. *Techniques d'enquête*, 16, 205-227.

BELL, W. R., et HILLMER, S. C. (1990, soumis). Applying time series models in survey estimation.

BELL, W. R., et MONSELL, B. C. (1992). X-11 Symmetric Linear Filters and Their Transfer Functions. Rapport de recherche 92/15, Statistical Research Division, Bureau of the Census.

BELL, W. R., et OTTO, M.C. (soumis). Bayesian Assessment of Uncertainty in Seasonal Adjustment with Sampling Error Present.

BELL, W. R., et WILCOX, D. W. (1993). The effect of sampling error on the time series behavior of consumption data. *Journal of Econometrics*, 55, 235-265.

BINDER, D. A., et DICK, J. P. (1989). Enquêtes répétées - modélisation 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.

BOBBITT, L., et OTTO, M. C. (1990). Effects of forecasts on the revisions of seasonally adjusted values using the X-11 seasonal adjustment procedure. Dans *Proceedings of the Business and Economics Statistics Section, American Statistical Association*, 449-454.

BOX, G.E.P., et JENKINS, G.M. (1976). Time Series Analysis: Forecasting and Control. San Francisco: Holden Day.

BREIDT, F. J. (1992). Variance estimation in the frequency domain for seasonally adjusted time series. Dans *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 337-342.

BURRIDGE, P., et WALLIS, K.F. (1984). Unobserved components models for seasonal adjustment filters. *Journal of Business and Economic Statistics*, 2, 350-359.

BURRIDGE, P., et WALLIS, K. F. (1985). Calculating the variance of seasonally adjusted series. *Journal of the American Statistical Association*, 80, 541-552.

CHANG, I., TIAO, G. C., et CHEN, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, 30, 193-204.

CLEVELAND, W. S., et DEVLIN, S. J. (1982). Calendar effects in monthly time series: modeling and adjustment. *Journal of the American Statistical Association*, 77, 520-528.

DAGUM, E. B. (1975). Seasonal factor forecasts from ARIMA models. In *Proceedings of the 40th Session of the International Statistical Institute*, Warsaw, Poland, 206-219.

DURBIN, J., et QUENNEVILLE, B. (1995). Benchmarking monthly time series with structural time series models. Dans *Proceedings of the Survey Methods Section, 23rd Annual Meeting of the Statistical Society of Canada*, Montréal, Canada, 13-18.

FINDLEY, D. F., MONSELL, B. C., BELL, W. R., OTTO, M. C., et CHEN, B. (soumis). New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program.

GEWEKE, J. (1978). Revision of Seasonally Adjusted Time Series. SSRI Report No. 7822, Department of Economics, University of Wisconsin.

HANSON, R. H. (1968). The Current Population Survey: Design and Methodology. Rapport technique 40, U.S. Census Bureau, Washington, D.C., Government Printing Office.

HAUSMAN, J. A., et WATSON, M. W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.

HILLMER, S.C., BELL, W.R., et TIAO, G. C. (1983). Modeling considerations in the seasonal adjustment of economic time Series. Dans *Applied Time Series Analysis of Economic Data*, (Éd. A. Zellner, U.S. Department of Commerce, Bureau of the Census, 74-100.

MCKENZIE, S.K. (1984). Concurrent seasonal adjustment with CensusX-11. *Journal of Business and Economic Statistics*, 2, 235-249.

MCLEOD, I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 24, 255-256.

MCLEOD, I. (1977). Correction to derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 26, 194.

PFEFFERMANN, D. (1994). A general method for estimating the variances of X-11 seasonally adjusted estimators. *Journal of Time Series Analysis*, 15, 85-116.

PFEFFERMANN, D., MORRY, M., et WONG, P. (1993). Estimation of the Variances of X-11-ARIMA Seasonally Adjusted Estimators for a Multiplicative Decomposition and Heteroscedastic Variances. Document de travail METH-93-005, Time Series Research and Analysis Division, Statistique Canada, Ottawa, Canada.

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

PRESIDENT'S COMMITTEE TO APPRAISE EMPLOYMENT AND UNEMPLOYMENT STATISTICS (1962). Measuring Employment and Unemployment, U.S. Government Printing Office, Washington, DC.

SHISKIN, J., YOUNG, A. H., et MUSGRAVE, J. C. (1967). The X-11 variant of the Census method II seasonal adjustment program. Rapport technique No. 15, U.S. Department of Commerce, Bureau of Economic Analysis.

TILLER, R. B. (1992). Time series modeling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

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

WILSON, G. T. (1979). Some efficient computational procedures for high order ARIMA models. *Journal of Statistical Computation and Simulation*, 8, 301-309.

WOLTER, K. M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

WOLTER, K. M., et MONSOUR, N. J. (1981). On the problem of variance estimation for a deseasonalized Series. Dans *Current Topics in Survey Sampling*, (Éds. D. Krewski, R. Platek, et J. N. K. Rao, New York: Academic Press, 367-403.

YOUNG, A. (1965). Estimating Trading-day Variation in Monthly Economic Time Series. Rapport technique 12, Bureau of the Census.

YOUNG, A. H. (1968). Linear approximations to the Census and BLS seasonal adjustment methods. *Journal of the American Statistical Association*, 63, 445-471.

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 33-45
Statistique Canada

Le choix des produits pour l'indice des prix à la consommation: le seuil d'inclusion par opposition au sondage probabiliste

JAN DE HAAN, EDDY OPPERDOES et CECILE M. SCHUT[1]

RÉSUMÉ

La plupart des bureaux de la statistique utilisent des techniques non probabilistes pour choisir l'échantillon de produits dont les prix permettent de calculer les indices des prix à la consommation. Aux Pays-Bas, comme dans beaucoup d'autres pays, ce genre de sondage raisonné se rapproche en quelque sorte de la sélection par seuil d'inclusion, une bonne partie de la population (normalement les produits suscitant le moins de dépenses) étant délibérément exclue des observations. Bien sûr, cette méthode donne lieu à des chiffres biaisés pour l'indice des prix. On peut se demander si un échantillonnage probabiliste donnerait de meilleurs résultats quant à l'erreur quadratique moyenne. Les auteurs ont considéré l'échantillonnage aléatoire simple, l'échantillonnage stratifié et l'échantillonnage systématique proportionnel aux dépenses. Ils ont mené des simulations de Monte Carlo à l'aide de données de lecture optique pour le café, les couches de bébés et le papier hygiénique afin d'évaluer le rendement des quatre plans d'échantillonnage. Il est assez surprenant de constater que la sélection par seuil d'inclusion est une bonne stratégie d'échantillonnage des produits pour l'indice des prix à la consommation.

MOTS CLÉS: Indice de prix de Laspeyres; simulation de Monte Carlo; échantillonnage; données de lecture optique; biais de substitution.

1. INTRODUCTION

Les non-initiés pourraient croire que la mesure de l'inflation est une tâche facile: il suffit de visiter des magasins, de recueillir une foule de prix et d'en établir la moyenne. Toutefois, les statisticiens chargés de calculer l'indice des prix à la consommation (IPC), la mesure la plus répandue de l'inflation, doivent surmonter de nombreux problèmes théoriques et pratiques. Dans la plupart des pays, l'IPC est essentiellement un indice de prix de Laspeyres. Cet indice établit le poids des indices de prix partiels des divers produits selon des proportions de dépenses fixes pour une période de référence. Il faut des procédures d'échantillonnage pour estimer la valeur de la population. Idéalement, l'erreur quadratique moyenne de l'estimateur est minimisée. La formule de l'indice de Laspeyres est extrêmement simple, mais les procédures d'estimation appliquées à l'IPC en font une statistique relativement complexe. Décrite de façon stylisée, l'estimation comporte trois types d'échantillons. Un échantillon des ménages participant à l'enquête sur les dépenses permet d'estimer le poids des dépenses par groupe de produits. Pour chaque groupe de produits, on choisit un échantillon de marchandises (produits). On recueille le prix de ces articles pour un échantillon de points de vente.

Les auteurs du présent article se penchent sur l'échantillonnage des produits. Il existe seulement quelques bureaux de la statistique, par exemple le Bureau of Labor Statistics des États-Unis, qui utilisent un échantillonnage probabiliste pour choisir les produits dont le prix sera calculé. La plupart des autres bureaux, y compris Statistique Pays-Bas, font appel au raisonnement de spécialistes travaillant au bureau central afin de déterminer quels produits devraient représenter le groupe de marchandises. Dans le passé, on a pu justifier cette méthode en soulignant l'absence de bases de sondage appropriées. Compte tenu cependant de l'automatisation de plus en plus poussée du commerce de détail, des registres de biens de consommation sont de plus en plus accessibles, de sorte qu'un échantillonnage probabiliste est possible. Toutefois, avant de passer à une nouvelle stratégie d'échantillonnage, il est bon d'examiner d'autres possibilités afin d'évaluer leur effet sur l'exactitude des indices de prix estimatifs. Il s'agit de déterminer si les méthodes actuelles de sélection non probabiliste donnent des résultats qui, relativement à l'erreur quadratique moyenne, sont inférieurs à ceux des techniques probabilistes. C'est là le thème principal du présent article. Des études de simulation ont été menées pour trois groupes de produits, c'est-à-dire le café, les couches jetables de bébés et le papier hygiénique.

Il n'y a pas si longtemps, le calcul empirique des indices de prix était d'autant plus difficile que les données hautement désagrégées sur les dépenses et les quantités au niveau des points de vente étaient inexistantes ou tout au plus accessibles pour de petits échantillons. De nos jours, des compagnies d'études de marché ont réussi à établir de vastes bases de microdonnées sur la vente de biens de consommation, en particulier dans le secteur des produits de grande consommation. Il s'agit de données obtenues par balayage électronique (lecture optique d'un code à barres ou inscription d'un code à barres à la caisse). Bradley, Cook, Leaver et Moulton (1997) présentent un aperçu des applications possibles des données de lecture optique pour le calcul de l'IPC. Le traitement de grandes bases de

[1] Jan de Haan, Eddy Opperdoes et Cecile M. Schut, Division Socio-economic, Statistics, Statistics Netherlands; P.O. Box 4000, 2270 JM, Voorburg, The Netherlands, courrier electronique: jhhn@cbs.nl.

données de lecture optique est une tâche assez fastidieuse. Pour ce qui est du calcul de l'IPC, c'est peut-être un obstacle à son application pour le moment. Toutefois, les données de lecture optique représentent certainement une riche source de renseignements pour les analyses empiriques. Les données de lecture optique permettent non seulement d'examiner l'échantillonnage, mais aussi de calculer les indices de prix en fonction de différentes formules. L'indice de prix de Laspeyres (poids fixe) ne tient pas compte des réactions des ménages aux changements de prix. Les auteurs ont donc examiné le degré de biais des indices de prix de Laspeyres (population) pour les trois groupes de produits relativement à l'indice de prix de Fisher, qui lui tient compte de la substitution des produits.

On trouvera à la section 2 un aperçu des données de lecture optique utilisées par les auteurs. À la section 3, il est question de quatre plans d'échantillonnage de produits. Trois d'entre eux (l'échantillonnage aléatoire simple, l'échantillonnage stratifié et l'échantillonnage proportionnel à la taille) relèvent de techniques probabilistes, tandis que le quatrième (échantillonnage par seuil d'inclusion) relève d'une méthode raisonnée qui reflète la stratégie officielle des Pays-Bas. La section 4 décrit les expériences de Monte Carlo utilisées par les auteurs pour déterminer l'exactitude des indices de prix des groupes de produits estimés à l'aide des plans d'échantillonnage mentionnés. La section 5 traite de l'utilisation des indices de Fisher pour des produits et des groupes de produits. Il y est question du biais de substitution intra-groupe des indices de prix de Laspeyres (groupes de produits). À la section 6, enfin, les résultats sont résumés et discutés.

2. DONNÉES DE LECTURE OPTIQUE (CODE À BARRES)

2.1 Aperçu

En Europe, les produits lisibles optiquement sont définis en fonction du numéro européen des articles (NEA). Les fabricants devraient attribuer un seul NEA à chaque variété, taille et type d'emballage d'un produit. Il en résulte deux phénomènes. Premièrement, les NEA changent parfois très rapidement, par exemple à cause d'un nouvel emballage. Bien sûr, il est alors difficile de suivre l'évolution d'un article. Deuxièmement, certains NEA comportent des données négligeables. Il semble que le système de classification soit trop détaillé; ce qui représente en réalité un article donne lieu à une multitude d'articles classés. Dans une étude fondée sur des données de lecture optique pour le café, Reinsdorf (1995) a trouvé que des articles pratiquement identiques portaient parfois un CUP (code universel des produits des États-Unis) différent. Une certaine agrégation relativement aux NEA est nécessaire. Heureusement, les séries de données de lecture optique englobent plusieurs caractéristiques de produits comme le nom commercial et le sous-nom. Les NEA comportant des caractéristiques de produits identiques sont traités ici comme un seul article. Si le nombre de caractéristiques est insuffisant, il existe bien sûr un risque de suragrégation, c'est-à-dire de regroupement d'articles hétérogènes.

De la société A.C. Nielsen (Nederland), nous avons reçu des séries de données de lecture optique sur les chiffres de vente hebdomadaires du café, des couches jetables de bébés et du papier hygiénique dans des supermarchés. Les séries de données initiales comportaient 320, 569 et 294 NEA, respectivement. Pour chaque NEA, nous avions le nombre d'emballages vendus et la valeur correspondante, de même que plusieurs caractéristiques des produits et des points de vente. Les prix n'étaient pas indiqués; les prix moyens (valeurs unitaires) ont dû être calculés à partir des valeurs et des quantités. Les données sur la vente de café couvraient une période de deux ans et demi, depuis la première semaine de 1994 jusqu'à la vingt-quatrième semaine de 1996, pour un échantillon de 20 supermarchés situés dans une région urbaine néerlandaise non connue des auteurs. Les données sur la vente des deux autres produits se rapportaient à un échantillon de 149 magasins répartis dans l'ensemble du pays et couvraient une période de deux ans, depuis la première semaine de 1995 jusqu'à la cinquante-deuxième semaine de 1996 inclusivement.

Par souci de simplification, nous avons supprimé les marques moins importantes. Dans le cas du café, seules les 15 marques se vendant le mieux durant toute la période d'observation ont été retenues parmi les 55 marques vendues. Après l'étape d'agrégation des NEA comportant des caractéristiques de produits identiques, nous avons limité la population aux articles vendus au cours de l'année de référence 1994 et durant chaque mois par après de façon à constituer une série de données complètes pour chaque mois. Nous avons ainsi obtenu 68 produits (à l'exclusion des graines de café), dont 40 produits de café moulu (y compris le café décaféiné) et 28 articles de café instantané, représentant 94,5% des dépenses totales en café pour l'année de référence dans le fichier initial. Pour ce qui est des couches et du papier hygiénique (à l'exclusion du papier hygiénique humide), nous avons éliminé les marques dont le chiffre des ventes représentait moins de 1% du total. De plus, nous avons retenu uniquement les articles vendus pendant l'année de référence 1995 et au moins huit mois par après. Nous avons ainsi obtenu 58 produits de couches et 70 produits de papier hygiénique, représentant 90% et 86% respectivement du total des dépenses de 1995 dans les fichiers initiaux.

2.2 Statistiques descriptives

L'aspect le plus surprenant des articles de dépense est l'asymétrie de la distribution. La figure 1 indique l'inégalité des dépenses de la période de référence dans nos fichiers de données rajustés à l'aide de courbes dites de Lorenz. Le total cumulatif des dépenses figure en ordonnée, tandis que le nombre cumulatif de produits figure en abscisse, en pourcentage dans les deux cas. Les produits

sont triés dans l'ordre croissant des dépenses. En cas de dépenses égales, les courbes de Lorenz décrivent une diagonale. Plus la répartition est inégale et plus sa position est inférieure. La distribution des articles de dépense pour le café est très inégale, les trois articles les plus importants représentant plus de la moitié du total des dépenses en café pour l'année de référence (1994). Pour ce qui est des couches de bébés et du papier hygiénique, les six et huit articles les plus importants, respectivement, représentent près de la moitié du total des dépenses pour l'année de référence (1995).

La figure 2 fournit des chiffres d'indice de valeurs unitaires, c'est-à-dire le changement de valeur par emballage, peu importe la quantité, le nom commercial, le type, etc. pour tous les points de vente. On y trouve une première indication du changement de «prix» au cours de la période à l'étude. Pour le café, une baisse remarquable a été observée au cours du deuxième semestre de 1995 à la suite d'une forte hausse des prix survenue en 1994 à cause de mauvaises récoltes au Brésil. Le prix du café est largement déterminé par le prix des graines de café sur le marché mondial. Nous n'avons trouvé aucune différence appréciable de changement de prix d'un point de vente à l'autre. Les couches de bébés font exception à cet égard. Une forte concurrence sévissait entre les différents producteurs (ce qui explique peut-être le déclin des valeurs unitaires en 1996), tandis que les rabais et autres mesures spéciales étaient fréquents. Par conséquent, la valeur unitaire pour tous les produits et points de vente donne un aperçu inexact du changement de prix agrégé des couches de bébés.

3. ESTIMATION DES INDICES DE PRIX DE TYPE LASPEYRES

Nous présentons d'abord des éléments de notation. Soit A, le groupe de marchandises constitué d'un nombre fini N de marchandises (produits); $g \in A$ signifie que le produit g appartient au groupe A. A est supposé fixe dans le temps. En réalité, ce n'est pas vrai: certains produits disparaissent du marché, tandis que d'autres y apparaissent. À court terme, toutefois, l'hypothèse de groupes de produits constants semble raisonnable. À noter que notre fichier de données initial a été rajusté en conséquence, car nous voulions considérer uniquement l'aspect de l'échantillonnage. L'indice de prix de Laspeyres (poids fixe) pour le groupe de marchandises A au cours de la période t est

$$P^t = \frac{\sum_{g \in A} e_g^0 P_g^t}{\sum_{g \in A} e_g^0} = \sum_{g \in A} w_g^0 P_g^t, \qquad (1)$$

où P_g^t désigne l'indice de prix du produit g, e_g^0 les dépenses pour g au cours de la période de référence 0 et w_g^0 la part de dépenses correspondante de g au sein du groupe de produits A. Au cours de la période de référence, un échantillon $\hat{A}$ de taille fixe n est tiré de A. Puisque A est supposé fixe dans le temps, il semble naturel que $\hat{A}$ soit fixe également.

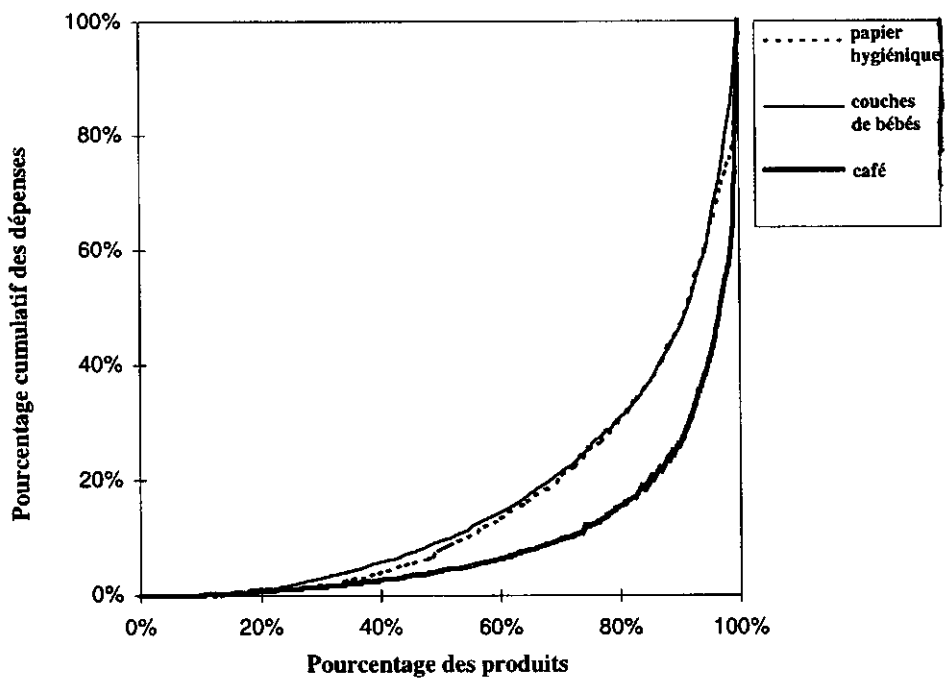


Figure 1. Distribution des articles de dépense pour l'année de référence

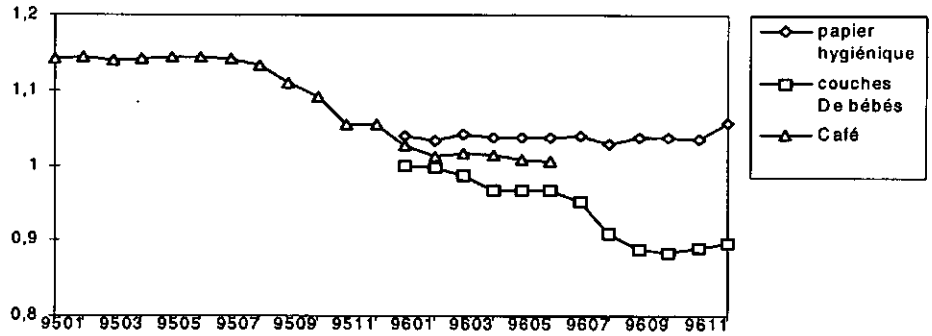


Figure 2. Chiffres d'indice de valeurs unitaires (1994=100 pour le café, 1995=100 pour les couches de bébés et le papier hygiénique)

3.1 Échantillonnage aléatoire simple

L'échantillonnage probabiliste renvoie à des situations comportant une probabilité connue de sélection de tous les échantillons éventuels. Pour l'échantillonnage aléatoire simple (sans remise), il existe une probabilité égale de sélection de tous les échantillons éventuels. L'estimateur de Horvitz-Thompson $\hat{P}^t_A = (N/n)\sum_{g\in\hat{A}} w^0_g P^t_g$ n'est pas biaisé pour P^t, c'est-à-dire que $E(\hat{P}^t_A) = P^t$ où l'espérance $E(.)$ désigne la moyenne de tous les échantillons éventuels pour un plan d'échantillon donné, dans le cas présent un échantillonnage aléatoire simple. Malgré son caractère non biaisé, $\hat{P}^t_A$ ne sera pas utilisé dans la pratique à cause de deux propriétés indésirables. Premièrement, si les indices de prix de tous les produits échantillonnés sont égaux, l'indice du groupe de produits estimatif diffère de cette valeur-là, à moins que la population et les moyennes empiriques des dépenses ne coïncident. Les personnes qui calculent les indices de prix n'apprécient probablement pas cet aspect. Deuxièmement, et surtout, $\hat{P}^t_A$ manifeste inévitablement une variance d'échantillonnage extraordinairement grande. Afin de surmonter ces deux difficultés, on prépare une estimation de P^t en établissant des estimateurs non biaisés du numérateur et du dénominateur:

$$\hat{P}^t_B = \frac{(N/n)\sum_{g\in\hat{A}} e^0_g P^t_g}{(N/n)\sum_{g\in\hat{A}} e^0_g} = \sum_{g\in\hat{A}} \hat{w}^0_g P^t_g, \qquad (2)$$

où $\hat{w}^0_g$ est la part de dépenses du produit g de l'échantillon. À l'aide d'une linéarisation de Taylor de premier ordre (Särndal, Swensson et Wretman 1992, pp. 172-176), on peut écrire la variance de $\hat{P}^t_B$ sous la forme

$$V(\hat{P}^t_B) \approx V(\hat{P}^t_A) - (P^t)^2(2\mu^t\rho^t - 1)(\sigma^0)^2,$$

où σ^0 désigne le coefficient de variation ou l'erreur type relative de la moyenne empirique des dépenses pour l'année de référence, μ^t est le rapport des erreurs types relatives des dépenses moyennes pour la période de référence exprimé sous forme de prix pour la période t et 0, et ρ^t est le coefficient de corrélation entre les dépenses moyennes pour la période de référence en prix de t et 0 (présumé de signe positif). Le choix de $\hat{P}^t_B$ au lieu de $\hat{P}^t_A$ s'explique donc par le fait que le premier bénéficie du caractère panel de l'échantillon; pour $\rho^t > 1/(2\mu^t)$, une réduction appréciable de la variance est attendue. La variance de $\hat{P}^t_B$ peut également s'exprimer comme suit:

$$V(\hat{P}^t_B) \approx \frac{1-f}{n}\frac{N^2}{N-1}\sum_{g\in A} (w^0_g)^2(P^t_g - P^t)^2, \qquad (3)$$

que l'on peut estimer à l'aide de données d'échantillon à la condition que la fraction de sondage $f = n/N$ soit connue. Cette formule, déjà mentionnée par Balk (1989), montre que la variance dépend de la dispersion intra-groupe des indices de prix des produits. Par conséquent, on pourrait réduire la variance soit en établissant des groupes de produits comportant des changements de prix semblables, soit en élargissant l'échantillon. Särndal et coll. (1992, p. 176) font remarquer toutefois que la méthode de linéarisation de Taylor risque d'entraîner une sous-estimation de la variance pour des échantillons qui ne sont pas très grands. Les échantillons de produits de l'IPC sont généralement assez petits. Pour certains groupes de produits, il n'y a parfois qu'un ou deux produits représentatifs. Ainsi, en plus d'être instable, comportant elle-même une grande variance, la variance sera probablement sous-estimée si elle se fonde sur (3).

À noter que l'estimateur $\hat{P}^t_B$, du fait qu'il est un rapport, subit pour de petits échantillons un biais à peu près égal à $o(1/n)$. On peut facilement vérifier que sa valeur absolue $|B(\hat{P}^t_B)| \le \sigma^0\sqrt{V(\hat{P}^t_B)}$. Si σ^0 est petit, par exemple inférieur à 0,1, le biais de $\hat{P}^t_b$ peut sans danger être considéré comme négligeable relativement à son erreur type. Par contre, pour un petit échantillon de produits et une grande variabilité des dépenses pour la période de référence, σ^0 peut facilement dépasser 0,1 de beaucoup. À noter également qu'il est peu probable que l'IPC pour tous les produits soit biaisé dans

une large mesure à cause de cela, puisque le biais est une moyenne (pondérée) de biais positifs et négatifs de différents indices de groupes de produits.

3.2 Échantillonnage proportionnel à la taille

L'échantillonnage proportionnel à la taille a ceci d'avantageux que les produits les plus importants ont une bonne chance d'être sélectionnés. Nous nous limitons ici à l'échantillonage de taille fixe sans remise, car ce semble être la méthode la plus probable en cas d'échantillonnage proportionnel à la taille du produit (voir par exemple le cas suédois décrit par Dalén et Ohlsson 1995). Les dépenses pour la période de référence nous servent de mesure de la taille; la probabilité d'inclusion de premier ordre requise pour le produit g est $\pi_g = ne_g^0/e^0 = nw_g^0$, où $e^0 = \sum_{g\in A} e_g^0$. Il s'ensuit que $\sum_{g\in \hat{A}} p_g^t/r$ est un estimateur non biaisé de P^t.

L'échantillonnage proportionnel à la taille sans remise, combiné à l'estimateur de Horvitz-Thompson ou π, s'appelle parfois l'échantillonnage ptπ. La plupart des schémas existants pour l'échantillonnage ptπ de taille fixe sont de type tirage-progressif et passablement compliqués. Nous allons donc nous en tenir plutôt à une sélection ptπ systématique. On peut décrire ce schéma en imaginant que les dépenses $e_g^0 (g \in A)$ sont étalées de façon cumulative et horizontalement, depuis l'origine jusqu'à e^0. Un nombre réel est choisi de façon aléatoire dans l'intervalle $(0, e^0/n]$, et nous continuons systématiquement en prenant les produits g identifiés par des points séparés par la distance constante e^0/n. Cette méthode fournit exactement la taille d'échantillon souhaitée. Pour des groupes de marchandises comportant une grande variation des dépenses pour la période de référence, il n'est pas toujours possible de tirer un échantillon de produits strictement proportionnel aux dépenses. Bien sûr, $\pi_g \le 1$ doit être respecté pour tous les g. Si $n > 1$ et si certaines valeurs e_g^0 sont extrêmement grandes, il est possible que pour certains produits $ne_g^0/e^0 > 1$, ce qui va à l'encontre de l'exigence $\pi_g \le 1$. La difficulté est surmontée comme suit. Les N produits sont rangés dans l'ordre descendant des dépenses. D'abord, si $e_1^0 > e^0/n$, nous prenons $\pi_1 = 1$. Ensuite, si $e_2^0 > (e^0 - e_1^0)/(n-1)$, nous prenons également $\pi_2 = 1$. La procédure est répétée jusqu'à ce que l'exigence d'un échantillonnage proportionnel aux dépenses pour la période de référence soit respectée pour tous les produits qui restent. Notre stratégie récursive diffère quelque peu de la méthode proposé par Särndal et coll. (1992, p. 90), qui suggèrent de prendre $\pi_g = 1$ pour tous les g comportant $e_g^0 > e^0/n$. Dans nos fichiers de données, cela entraînerait des nombres inutilement grands des produits comportant $\pi_g = 1$. Le sous- groupe A_H de produits comportant les dépenses les plus élevées pour la période de référence, qui est choisi avec certitude, s'appelle la partie auto-sélection de l'échantillon. À partir du sous-groupe A_L à faibles dépenses qui reste, on tire un échantillon $\hat{A}_L$ de taille n_L strictement proportionnel aux dépenses. L'estimateur non biaisé qui en résulte est une moyenne pondérée en fonction des dépenses de $P^t(H)$, l'indice de prix de Laspeyres (population) de A_H, et de $\sum_{g\in \hat{A}_L} P_g^t/n_L$, l'indice de prix estimatif de A_L.

3.3 Échantillonnage stratifié

L'échantillonnage aléatoire simple offre un avantage évident relativement à l'échantillonnage proportionnel aux dépenses: en effet, à part l'utilisation d'un registre de produits comme base de sondage, aucune autre donnée n'est nécessaire (voir aussi Balk 1994). En présence d'articles de dépense à distribution très inégale, il y a de fortes chances que les chefs de file du marché se trouvent hors de l'échantillon, situation qui est intuitivement jugée indésirable. Nous estimons qu'il serait effectivement préférable qu'ils soient sélectionnés. Rappelons que la variance de l'indice de prix des groupes de produits dans le cadre de l'échantillonnage aléatoire simple dépend de la dispersion intra-groupe des indices de prix des produits. La variance pourrait être réduite s'il était possible de stratifier le groupe de produits en sous-groupes homogènes en fonction des changements de prix. Toutefois, les changements de prix des produits ne sont pas connus d'avance. Une autre façon de réduire la variance serait de stratifier le groupe de produits en deux sous-groupes, le premier (A_H) comportant de fortes dépenses pour la période de référence, que l'on observe intégralement, l'autre (A_L) comportant de faibles dépenses, dont on tire un échantillon aléatoire $\hat{A}_L$. Le nouvel estimateur de l'indice de prix des groupes de produits est une moyenne pondérée en fonction des dépenses de $\hat{P}_B^t(L)$, l'indice de Laspeyres du sous-groupe à faibles dépenses, estimé en conformité avec (2), et de $P^t(H)$. Sa variance d'échantillonnage est de $(1 - \tau_H)^2 V[\hat{P}_B^t(L)]$, où τ_H est la partie dépenses de A_H dans A. Cette méthode ne réduit pas nécessairement la variance de l'indice de prix estimatif, mais elle la réduit probablement dans certaines circonstances. La variance du nouvel estimateur sera plus faible que la variance de $\hat{P}_B^t$ lorsque

$$1 - \tau_H < \frac{\text{e.t.}(\hat{P}_B^t)}{\text{e.t.}[\hat{P}_B^t(L)]}, \qquad (4)$$

où e.t.(.) désigne l'erreur-type. L'inégalité (4) est présumée intacte si les articles de dépense sont distribués de façon extrêmement inégale, puisque $1 - \tau_H$ devient alors nettement inférieur à 1. Il est possible que la stratification soit particulièrement productive à mesure que la taille globale de l'échantillon n augmente.

Le choix de τ_H et donc de la taille N_H de la strate «à tirage complet» A_H est un peu difficile. Il serait préférable d'avoir un certain critère d'optimalité afin de minimiser la variance. Toutefois, puisque les changements de prix des produits ne sont pas connus d'avance et que les tendances antérieures ne permettent pas une prévision très exacte des futurs changements de prix, il n'est guère possible de

calculer la taille optimale de A_H. Pour l'analyse empirique, nous ferons l'essai de deux tailles d'échantillon relatives $\lambda_H = N_H/n$ de A_H, c'est-à-dire $\lambda_H = 1/3$ et $\lambda_H = 2/3$. Ces valeurs permettent d'indiquer clairement le rendement.

3.4 Échantillonnage par seuil d'inclusion

Lorsque la taille de l'échantillon est très petite, il semble assez probable qu'une stratification fondée sur $\lambda_H = 2/3$ donnera une erreur type plus grande de l'indice de prix estimatif qu'une stratification fondée sur $\lambda_H = 1/3$. Qu'arrive-t-il si A_L n'est pas observé du tout, de sorte que $\lambda_H = 1$ et donc $n = N_H$? Nous aurions alors recours à un échantillonnage par seuil d'inclusion (de type spécial). L'estimation de l'indice de prix des groupes de produits se fonde simplement sur $\hat{P}^t_C = P^t(H)$. Dès lors tous les $g \in A_H$ ont une probabilité d'inclusion de 1, tandis que tous les $g \in A_L$ ont une probabilité d'inclusion nulle (Särndal et coll. 1992, pp. 531-533). Puisque nous savons exactement quels produits seront choisis, il n'y a rien d'aléatoire et la variance d'échantillonnage du $\hat{P}^t_C$ est nulle par définition. Le biais correspond à l'erreur réelle, c'est-à-dire à la différence entre la valeur estimée et l'indice de population véritable

$$\hat{P}^t_C - P^t = (1 - \tau_H)[P^t(H) - P^t(L)]. \qquad (5)$$

Lorsque la distribution des articles de dépense est extrêmement inégale, même une petite taille d'échantillon peut entraîner une grande valeur de τ_H. Dans un tel cas, une estimation par seuil d'inclusion peut donner de meilleurs résultats que la stratification, relativement à l'erreur quadratique moyenne. Nous avons le choix de fixer le taux d'inclusion τ_H, de façon que la taille de l'échantillon n soit déterminée par τ_H, ou de fixer la taille de l'échantillon, et alors τ_H dépend du choix de n. Nous avons retenu cette dernière possibilité puisque les plans d'échantillonnage à taille fixe sont un outil répandu pour la sélection des produits de l'IPC, et parce que cela permet une comparaison appropriée à d'autres plans à taille fixe.

Le recours à des procédures de seuil d'inclusion est justifié du fait que i) les frais d'établissement d'une base de sondage fiable pour l'ensemble de la population sont trop élevés et que ii) le biais est jugé négligeable. Bien sûr, l'hypothèse ii) ne peut pas être vérifiée en général. Néanmoins, l'exclusion délibérée d'une partie de la population cible du tirage de l'échantillon peut donner des résultats satisfaisants sous réserve de corrections appropriées. Statistique Pays-Bas fait appel à un échantillonnage par seuil d'inclusion pour diverses autres enquêtes-entreprises, par exemple pour la statistique de la production et du commerce étranger, de très petites entreprises restant inobservées. Quant aux comptes nationaux néerlandais, pour lesquels les données sur la production et le commerce étranger sont un intrant important, on établit des estimations explicites pour les petites entreprises. La méthode du seuil d'inclusion pour la sélection des produits de l'IPC, par contre, n'apporte pas de correction pour les produits exclus. La question du coût mise à part, on défend parfois cette méthode du fait que, au moins à long terme, les changements de prix des produits moins importants ne diffèrent pas tellement de ceux des chefs de file du marché pour la même catégorie de produits à cause de la structure analogue des coûts de production.

4. ESTIMATION EMPIRIQUE

4.1 Simulation de Monte Carlo

Exception faite pour la sélection par seuil d'inclusion, il est difficile de trouver des mesures fiables des distributions d'échantillonnage fondées sur un seul échantillon. Pour ce qui est de l'échantillonnage aléatoire simple, l'estimateur $\hat{P}^t_C$ comporte un biais inconnu, tandis que l'estimation de la variance fondée sur les techniques de linéarisation de Taylor donne des résultats inexacts parce que les échantillons de produits de l'IPC sont généralement très petits. L'échantillonnage ptπ systématique pose la difficulté du choix d'une méthode d'estimation de la variance de l'estimateur puisque les probabilités d'inclusion de second ordre sont inconnues. Afin d'obtenir la distribution d'échantillonnage exacte, il faudrait considérer tous les échantillons $\hat{A}$ qui sont possibles pour un certain plan d'échantillonnage. Pour chaque $\hat{A}$ la probabilité de sélection de $\hat{A}$ et la valeur estimative de l'indice de prix des groupes de marchandises doivent être connues si l'on veut calculer la valeur attendue, le biais et la variance de l'estimateur de façon exacte. Or cela est pratiquement impossible à cause du nombre extrêmement grand d'échantillons possibles. Afin de décrire la distribution d'échantillonnage, nous menons donc des simulations de Monte Carlo. Un grand nombre d'échantillons K est tiré de la (même) population A en fonction du plan choisi et, pour chaque échantillon, on calcule l'estimation. Si K est assez grand, la distribution des K estimations se rapproche étroitement de la distribution d'échantillonnage exacte. Soit $\hat{P}^t_k$, le résultat du k-ième échantillon pour un certain plan d'échantillonnage. Dès lors

$$\bar{\hat{P}}^t = \frac{1}{K}\sum_{k=1}^{K} \hat{P}^t_k$$

est une estimation non biaisée de la valeur attendue $E(\hat{P}^t)$. Nous calculons alors

$$\bar{\hat{P}}^t - P^t,$$

qui est une estimation non biaisée du biais $B(\hat{P}^t)$;

$$S^2_P = \frac{1}{K-1}\sum_{k=1}^{K} (\hat{P}^t_k - \bar{\hat{P}}^t)^2,$$

qui est une estimation non biaisée de la variance $V(\hat{P}^t)$; et

$$\sqrt{(\bar{\hat{P}}^t - P^t)^2 + S^2_P},$$

qui représente une estimation à peu près non biaisée de la racine de l'erreur quadratique moyenne (reqm) de $\hat{P}'$. Särndal et coll. (1992, p. 280) font remarquer que l'imperfection résultant du nombre fini de répétitions est plus sérieuse dans le cas d'une mesure de la variance... que dans le cas de mesures calculées comme moyennes.

4.2 Résultats

Des simulations de Monte Carlo ont été effectuées avec trois tailles d'échantillon: n=3, n=6 et n=12. Le nombre de répétitions (K) par expérience a été fixé à 500 000. Le tableau 1 fournit les résultats pour le café en janvier 1995 (1994=100), tandis que les tableaux 2 et 3 fournissent les résultats pour les couches de bébés et le papier hygiénique, respectivement, en janvier 1996 (1995=100). Le choix de la formule qui permet d'agréger les observations de prix individuelles en un même indice de prix des produits est décrit à l'annexe. Pour ce qui est de la présente section, tous les indices de prix des produits sont calculés sous forme d'indices de valeurs unitaires pour tous les points de vente. L'échantillonnage aléatoire simple donne des résultats particulièrement médiocres. Ainsi, pour n=3, la véritable augmentation du prix du café (Laspeyres) de 17,2 % est sous-estimée de 1,4 point. Jumelée à une erreur type de 5,1 points, la reqm représente 5,3 points, c'est-à-dire près du tiers de l'augmentation véritable des prix. Même pour n=12, qui donne une fraction d'échantillonnage de 0,18 (valeur inhabituellement élevée dans la pratique), la reqm demeure néanmoins considérablement élevée. À noter que, comme prévu, le biais pour les petits échantillons est réduit de moitié lorsque la taille de l'échantillon est doublée. La stratification fonctionne assez bien pour de grandes tailles d'échantillon, mais donne de piètres résultats pour n=3. Dans ce dernier cas, la stratification fait augmenter la reqm comparativement à l'échantillonnage aléatoire simple pour les couches de bébés et le papier hygiénique lorsque $N_H = 2$ (c'est-à-dire lorsque $\lambda_H = 2/3$). Notre plan probabiliste préféré serait un échantillonnage systématique proportionnel aux dépenses parce que les estimations ne sont pas biaisées et que les erreurs types sont relativement faibles. Le résultat le plus surprenant est sans doute le bon rendement de la sélection par seuil d'inclusion. Sauf pour n=3 et n=6 dans le cas des couches de bébés, cette méthode donne les meilleurs résultats.

Tableau 1

Chiffres estimatifs de l'indice de prix de Laspeyres pour le café (1994=100), Janvier 1995 (N=68)

Plan d'échantillonnage	n = 3				n = 6				n = 12			
	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm
A.S. *)	115,7	5,1	-1,4	5,3	116,4	3,4	-0,7	3,5	116,7	2,3	-0,4	2,3
ptπ	117,2	2,2	0	2,2	117,2	1,3	0	1,3	117,2	0,7	0	0,7
Stratifié												
λ_H = 1/3	116,4	3,9	-0,7	4,0	116,6	2,3	-0,5	2,3	117,0	1,2	-0,1	1,2
λ_H = 2/3	115,6	4,5	-1,5	4,7	116,4	2,5	-0,7	2,6	117,0	1,1	-0,2	1,1
Seuil d'inclusion	117,0	0	-0,2	0,2	117,2	0	0,0	0,0	117,5	0	0,3	0,3

*) Aléatoire simple

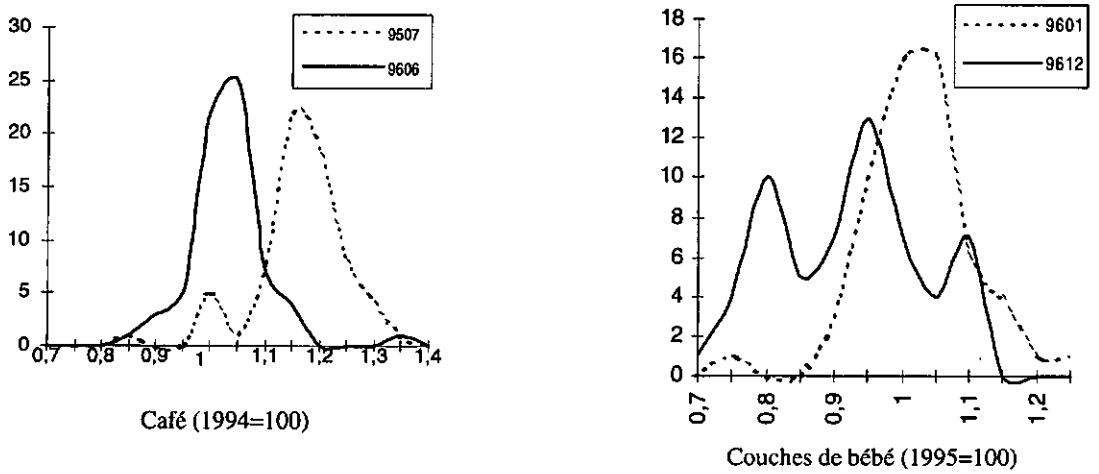
Tableau 2

Chiffres estimatifs de l'indice de prix de Laspeyres pour les couches de bébés (1995=100), Janvier 1996 (N=58)

Plan d'échantillonnage	n = 3				n = 6				n = 12			
	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm
A.S.	99,4	5,0	2,3	5,5	98,7	3,9	1,5	4,2	97,9	2,9	0,8	3,0
ptπ	97,2	2,8	0	2,8	97,2	1,6	0	1,6	97,2	1,5	0	1,5
Stratifié												
λ_H = 1/3	98,9	5,0	1,8	5,3	98,1	3,3	1,0	3,4	97,4	1,7	0,2	1,7
λ_H = 2/3	98,3	5,8	1,1	5,9	97,4	3,3	3,3	3,3	97,0	1,6	-0,2	1,6
Seuil d'inclusion	92,0	0	-5,1	5,1	93,4	0	-3,8	3,8	95,5	0	-1,6	1,6

Tableau 3
Chiffres estimatifs de l'indice de prix de Laspeyres pour le papier hygiénique (1995=100), Janvier 1996 (N=70)

Plan d'échantillonnage	n = 3				n = 6				n = 12			
	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm
A.S.	103,9	4,5	0,1	4,5	103,9	3,5	-0,1	3,5	103,9	2,6	0,1	2,6
$pt\pi$	103,9	3,4	0	3,4	103,9	1,8	0	1,8	103,9	1,2	0,0	1,2
Stratifié												
λ_H = 1/3	103,5	4,3	-0,3	4,3	103,7	3,2	-0,1	3,2	104,0	2,1	0,1	2,1
λ_H = 2/3	103,7	4,6	-0,2	4,6	104,2	3,4	0,4	3,4	103,9	1,6	0,0	1,6
Seuil d'inclusion	105,0	0	1,1	1,1	104,0	0	0,1	0,1	104,0	0	0,1	0,1



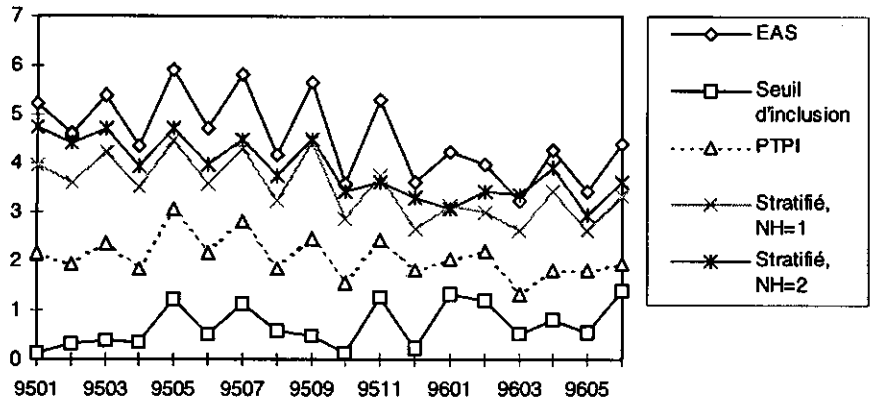
Café (1994=100) Couches de bébé (1995=100)

Figure 3. Distribution de fréquences des chiffres des indices de prix des produits

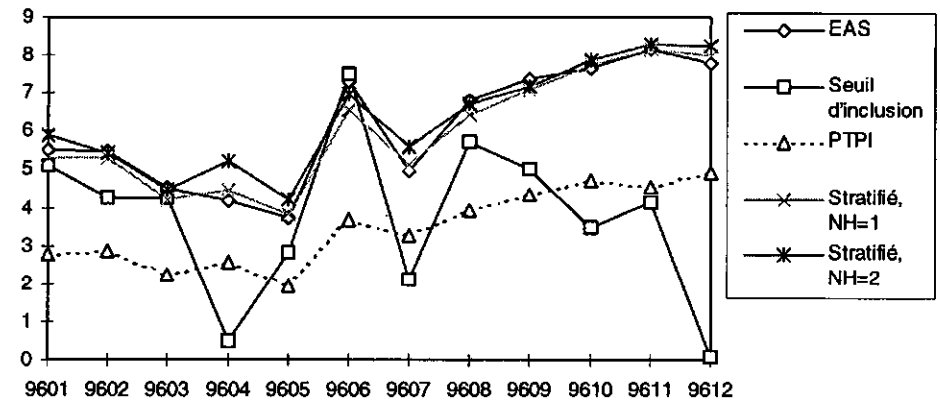
Pour le café, nous avons également fait l'essai d'un autre type d'échantillonnage stratifié. La population globale des produits a été divisée en café moulu et en café instantané et nous avons tiré de chaque strate des échantillons aléatoires. Même si les changements de prix du café instantané sont lissés et prennent du retard comparativement au café moulu, les résultats de Monte Carlo obtenus à l'aide de l'échantillonnage stratifié sont semblables à ceux d'un échantillonnage non stratifié pour l'ensemble des quatre méthodes d'échantillonnage. Cela va à l'encontre de résultats antérieurs (voir de Haan et Opperdoes 1997a). Il en est ainsi parce que nous avons supprimé certains produits de café instantané pour cette étude afin d'obtenir un fichier complet de données pour chaque mois; en fin de compte nous avons obtenu une fraction mineure (8%) de café instantané dans les dépenses totales de café pour l'année de référence.

Il serait dangereux de tirer des conclusions au sujet du rendement des différents plans d'échantillonnage à partir de simulations pour un mois particulier car il est probable que les résultats dépendent de la distribution de fréquences des indices de prix des produits. La figure 3 montre ces distributions pour le café et les couches de bébés pendant deux mois. Les deux distributions se déplacent vers la gauche, indiquant une baisse de la moyenne non pondérée. Autrement, la distribution de fréquences pour le café demeure assez stable. Le profil de la courbe pour les couches, par contre, change considérablement; il y a eu accroissement de la variance des indices de produits. Des expériences de Monte Carlo ont été menées pour chaque mois de la période à l'étude. La figure 4 montre la reqm pour n=3. Le profil qui se dégage pour le café et le papier hygiénique est remarquablement robuste: la sélection par seuil d'inclusion est toujours la meilleure. Il semble que, pour de faibles tailles d'échantillon, l'exclusion des produits moins importants n'a guère d'importance. C'est ce que de nombreux bureaux de la statistique ont compris depuis longtemps, sans avoir pu le vérifier de façon empirique jusqu'à présent. Le meilleur rendement de la sélection par seuil d'inclusion relativement à l'échantillonnage proportionnel aux dépenses s'explique, dans le cas du papier hygiénique, en partie par le fait qu'il n'existe pas d'autosélection pour ce dernier schéma d'échantillonnage. Pour de grands échantillons, les résultats de la sélection par seuil d'inclusion et de l'échantillonnage proportionnel à la taille sont très semblables. Pour les couches de bébés, les résultats sont quelque peu différents. À cause de la grande volatilité des indices du produit, l'reqm varie considérablement pour ce qui est de la sélection par seuil d'inclusion; elle semble se promener autour de la reqm provenant d'un échantillonnage systématique proportionnel aux dépenses. La forte variabilité de l'erreur peut être considérée comme un désavantage de la sélection par seuil d'inclusion.

Café



Couches de bébés



Papier hygiénique

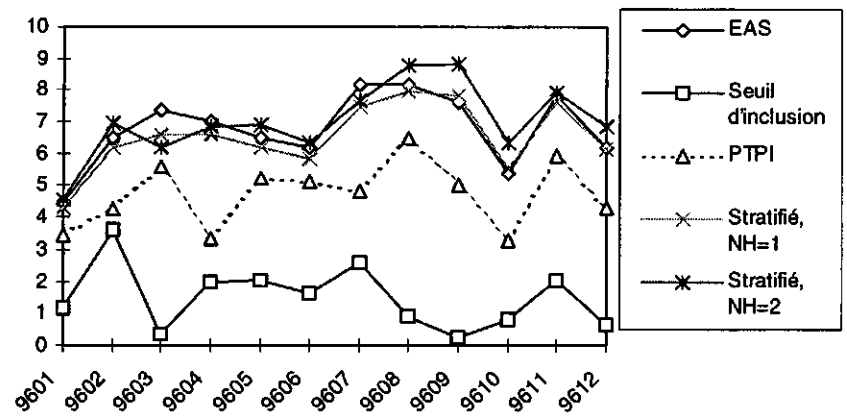


Figure 4. Racine de l'erreur quadratique moyenne (reqm) des indices de prix de Laspeyres estimatifs

5. UTILISATION DES INDICES DE FISHER

5.1 Indices de produits de valeurs unitaires/ de Fisher

À la section 4, les indices de prix des produits ont été calculés comme indices de valeurs unitaires pour tous les points de vente. Comme moyen d'évaluer l'effet du choix de la formule de l'indice de produits sur le résultat de l'étude de simulation, le tableau 4 compare les résultats de Monte Carlo pour n=3 d'après les chiffres de l'indice de produits de valeurs unitaires (comme dans les tableaux 1-3) et les chiffres de l'indice de prix de Fisher (on trouvera des détails à l'annexe). Pour le café, il n'y a guère de différence. Pour les couches et le papier hygiénique, par contre, la reqm diminue lorsqu'on fait appel aux chiffres de l'indice de Fisher, surtout pour l'échantillonnage aléatoire simple. Il en est ainsi parce que les indices de valeurs unitaires manifestent un comportement plus irrégulier. Si des types de couche ou de papier hygiénique physiquement

identiques sont jugés hétérogènes d'un point de vente à l'autre, de sorte que la formule de Fisher serait plus appropriée, le recours à des indices de valeurs unitaires entraîne une surestimation de la variabilité des prix des produits exceptionnellement petits et exagère le mauvais rendement de l'échantillonnage aléatoire simple. Il reste cependant que l'échantillonnage aléatoire simple ne fonctionne pas très bien.

5.2 Biais de substitution intra-groupe

De nombreux bureaux de la statistique et utilisateurs estiment que l'IPC devrait se rapprocher du véritable indice du coût de la vie. Cette notion théorique tirée de la micro-économie sert à mesurer le changement, pour un consommateur ou un ménage représentatif, des coûts minimaux nécessaires au maintien du même niveau de vie ou d'utilité. Puisque l'utilité ne se laisse pas mesurer, il convient de choisir une formule d'indice se rapprochant étroitement de la notion. Diewert (1976) a montré que (ce qu'il appelle) des indices superlatifs fournissent des approximations de second ordre de l'indice du coût de la vie. L'aspect le plus important des indices de prix superlatifs, c'est qu'ils tiennent compte de l'évolution des consommateurs vers des biens et services manifestant des augmentations de prix relativement faibles. Ces formules d'indice font appel à des données sur les dépenses tant pour la période de référence 0 que pour la période courante t. Concrètement, il faut un certain temps avant que les données sur les dépenses soient connues, de sorte que les indices superlatifs ne peuvent pas être compilés en temps réel. Par souci de rapidité de diffusion, la plupart des bureaux de la statistique nationaux adoptent la formule de Laspeyres (poids fixe) pour l'établissement de leur IPC.

L'indice de Fisher est un des indices superlatifs les mieux connus. Lorsqu'on l'applique au niveau des groupes de produits, la différence entre les indices de prix (population) calculés en fonction des formules de Laspyres et de Fisher peut être interprétée comme un biais de substitution de produits intra-groupe (figure 5). Pour le café, le biais est inférieur à 1 point par année. Pour le papier hygiénique et pour les couches en particulier, le biais est très considérable, c'est-à-dire environ 1,5 à 3 points par année. Le biais de substitution intra-groupe est généralement positif et il augmente avec le temps. À noter, toutefois, que pour les couches de bébés durant quelques mois au cours du premier semestre de 1996, les chiffres de l'indice de Laspeyres sont inférieurs à ceux de l'indice de Fisher. Ce phénomène inattendu et peut-être aussi l'ampleur du biais positif au cours d'autres mois sont possiblement dus à un défaut du fichier de données qui ne comporte que des supermarchés. Il est bien connu que les couches de bébés aux Pays-Bas s'achètent aussi dans d'autres magasins comme les pharmacies, qui n'ont pas recours à la lecture optique de codes à barres. La substitution des points de vente inclus et exclus de la base de données pourrait porter atteinte à nos chiffres d'indice de la population comme approximation exacte des valeurs véritables. Nous sommes toutefois persuadés qu'elle ne porte pas sérieusement atteinte à l'évaluation des méthodes d'échantillonnage présentées à la section 4.

6. DISCUSSION

Même si les données tirées de la lecture optique de codes à barres comportent certaines lacunes, elles sont un excellent outil de recherche empirique pour diverses questions d'échantillonnage se rapportant aux IPC. Nos simulations montrent que, du moins pour le café, les couches jetables de bébés et le papier hygiénique, l'échantillonnage aléatoire simple des produits n'est pas à conseiller. Nous estimons qu'une telle affirmation s'applique à tous les groupes de produits lorsque la distribution des dépenses est très asymétrique. Si les bureaux de la statistique veulent avoir recours à un échantillonnage probabiliste, ils auraient avantage à utiliser l'échantillonnage proportionnel aux dépenses. Toutefois, la sélection par seuil d'inclusion serait un choix convenable ou même meilleur lorsque les groupes de produits ne comportent pas de changements de prix trop volatils. En effet, autant que nous sachions, la présente étude est la première à fournir des arguments empiriques à l'appui des méthodes de sélection des produits de l'IPC par seuil d'inclusion. Les données de lecture optique agrégées relativement aux points

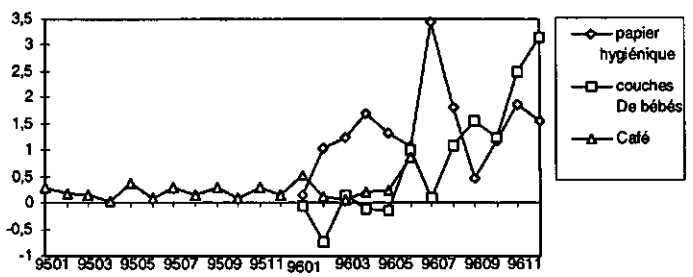


Figure 5. Différence entre les indices de prix (population) de Laspeyres et de Fisher

Tableau 4
Chiffres estimatifs de l'indice de prix de Laspeyres résultant d'autres indices de produits (n=3)

Plan d'échantillonnage	Café, janvier 1995 (1994=100)				couches, janvier 1996 (1995=100)				papier hygiénique, janvier 1996 (1995=100)			
	(1)		(2)		(1)		(2)		(1)		(2)	
	valeur attendue	reqm	valeur attendue	reqm	valeur attendue	reqm	valeur attendue	reqm	valeur attendue	reqm	valeur attendue	reqm
A.S.	115,7	5,3	115,8	5,3	99,4	5,5	100,4	4,2	103,9	4,5	104,0	3,5
$pt\pi$	117,2	2,2	117,2	2,2	97,2	2,8	98,6	2,1	103,9	3,4	104,3	3,6
Stratifié												
$\lambda_H = 1/3$	116,4	4,0	116,5	4,0	98,9	5,3	100,1	4,1	103,5	4,3	103,3	3,6
$\lambda_H = 2/3$	115,6	4,7	115,6	4,8	98,3	5,9	99,5	4,6	103,7	4,6	103,2	3,9
Seuil d'inclusion	117,0	0,2	117,0	0,2	92,0	51,	94,8	3,8	105,5	1,1	104,7	1,0

(1) D'après les chiffres d'indice de produits de valeurs unitaires
(2) D'après les chiffres d'indice de prix de Fisher

de vente devraient fournir une indication claire du seuil d'inclusion requis. Statistique Pays-Bas a déjà fait appel à des données de Nielsen agrégées pour différents groupes de marchandises afin de choisir des produits pour l'échantillon de l'IPC. Les méthodes de seuil d'inclusion sont répandues aux Pays-Bas et dans d'autres pays européens (Boon 1997). Aux Pays-Bas, la sélection des produits comme telle est un peu plus complexe que la situation décrite ci-dessus. Tout d'abord, la méthode du seuil d'inclusion sert à choisir des sous-groupes de produits et non pas des produits particuliers. Ensuite, des produits particuliers sont choisis dans chaque sous-groupe à l'aide d'un échantillonnage dit raisonné. La sélection de ces produits représentatifs se fonde sur le raisonnement de spécialistes travaillant au bureau central et connaissant à fond le marché de consommation en question. Normalement on choisit les produits qui sont achetés le plus souvent ou qui comportent le taux de rotation le plus élevé, de sorte que le schéma d'échantillonnage est une procédure d'inclusion à deux degrés. Il est peu probable qu'une telle méthode à deux degrés donne des résultats bien différents de la procédure à un degré utilisée par les auteurs du présent exposé.

Dans d'autres pays européens, par exemple, au Royaume-Uni, la sélection par seuil d'inclusion ne se fait pas au bureau central, mais bien par le personnel régional là où les prix sont mesurés. Nous choisissons, par exemple, un produit par point de vente, c'est-à-dire le produit le plus vendu durant la période de référence pour le point de vente en question. Dans le cas du café, des couches de bébés et du papier hygiénique, cela donne 2, 12 et 24 produits différents, respectivement. L'estimation de l'indice de Laspeyres du groupe de produits se fait en conformité avec l'expression (2), où les indices de prix des produits sont calculés sous forme d'indices de valeurs unitaires particuliers aux points de vente et pondérés en fonction du point de vente. La figure 6 montre la reqm résultant de cette méthode. Si nous comparons cette figure à la figure 4 (sélection par seuil d'inclusion effectuée au bureau central pour n=3), l'exactitude des deux méthodes de sélection par seuil d'inclusion semble être «en moyenne» du même ordre de grandeur, même si le schéma est un peu plus irrégulier pour la sélection qui se fait aux points de vente. Une telle comparaison est cependant assez arbitraire. Pourquoi ne pas comparer la sélection par seuil d'inclusion qui se fait aux points de vente à la sélection par seuil d'inclusion qui se fait au bureau central pour n=6, ou encore n=12, ou même pour une autre taille d'échantillon quelconque? Une autre difficulté, c'est que nous avons traité les indices de prix des produits comme s'ils étaient connus de façon certaine. En réalité, ils se fondent sur un échantillon de points de vente, de sorte que nos résultats dépendent de l'échantillon en question. Pour bien évaluer les deux procédures de sélection par seuil d'inclusion, il faudrait tenir compte à la fois de l'échantillonnage des produits et de l'échantillonnage des points de vente, ce qui toutefois dépasse le cadre du présent exposé.

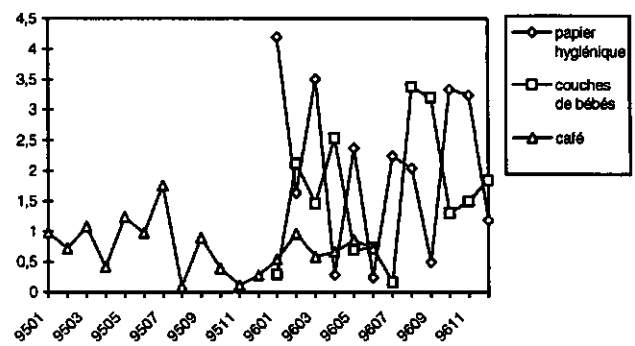


Figure 6. Racine de l'erreur quadratique moyenne (reqm) résultant de la sélection par seuil d'inclusion aux points de vente

Les données de lecture optique ouvrent des perspectives pour la recherche statistique dans le domaine de l'échantillonnage en vue de l'IPC, et servent aussi à compiler toutes sortes de chiffres d'indice, y compris les indices superlatifs, à l'aide de données réelles aussi bien que fortement désagrégées au niveau des points de vente particuliers. Nous avons montré que les indices de prix de

Laspeyres pour des groupes de produits utilisés par les bureaux de la statistique peuvent être biaisés de plus d'un point sur une base annuelle relativement à l'indice de prix de Fisher (superlatif) qui tient compte de la substitution des produits. Un type connexe de biais, résultant d'une considération insuffisante des produits introduits après la période de référence (voir par exemple Boskin, Dulberger, Gordon, Griliches et Jorgenson 1996), n'a pas été étudié dans le présent exposé. Or les données de lecture optique sont un bon moyen d'examiner ce type de biais lié aux nouveaux produits.

REMERCIEMENTS

La présente recherche a été appuyée en partie par Eurostat (le bureau de la statistique des Communautés européennes) dans le cadre de SUP-COM 1996, Lot 1: élaboration de méthodologies pour les indices de prix à la consommation et les parités de pouvoir d'achat. Les auteurs tiennent à remercier la société A.C. Nielsen (Nederland) d'avoir bien voulu fournir des données de lecture optique à un coût marginal. Ils tiennent également à remercier Bert M. Balk, Leendert Hoven et deux lecteurs anonymes de leurs précieuses remarques au sujet d'une version provisoire antérieure.

ANNEXE: LE CHOIX D'UNE FORMULE D'INDICE DE PRODUITS

Les simulations d'échantillonnage supposent des chiffres d'indice de produits. Quelle formule d'indice choisir? Les bureaux de la statistique doivent généralement calculer les indices au niveau d'agrégation le plus faible basé sur les seules données sur les prix à cause de l'absence de données sur les quantités ou les dépenses. On pourra consulter à ce sujet les travaux détaillés de Szulc (1987), de Dalén (1991), de Balk (1994) et de Diewert (1995). Grâce à la présence de microdonnées de lecture optique, nous nous trouvons dans une position unique de pouvoir établir des indices de prix authentiques (Silver 1995, Hawkes 1997). Soit B_g, une série de points de vente supposée fixe dans le temps, où g est un produit achetable, $b \in B_g$ signifie que g peut être acheté au point de vente b. Le prix de g au point de vente b pour la période s (s=0,t) et le nombre correspondant d'articles vendus sont P_{gb}^s et x_{gb}^s, respectivement. Le produit est considéré comme le niveau d'agrégation le plus faible pour lequel des indices de prix sont établis. Pour commencer, nous nous limitons aux indices de produits que nous pouvons formuler comme des coefficients de prix moyens arithmétiques pondérés pour la période t et la période 0:

$$P_g^t = \frac{\sum_{b \in B_g} w_{gb}^s P_{gb}^t}{\sum_{b \in B_g} w_{gb}^u P_{gb}^0}, \qquad (6)$$

où $w_{gb}^z = x_{gb}^z / \sum_{b \in B_g} x_{gb}^z$ désigne la part du point de vente b dans la vente globale du produit g pour la période z (z=s,u). Si u=0 et s=t, les prix au cours de la période 0 et de la période t sont pondérés à l'aide des quantités relatives correspondantes. Les prix moyens sont appelés des valeurs unitaires, et P_g^t est un indice de valeurs unitaires. Les mérites de cette méthode sont évalués par De Haan et Opperdoes (1997b) et par Balk (1998). L'addition des quantités n'a de sens que si le produit g peut être considéré comme homogène, c'est-à-dire identique pour l'ensemble des $b \in B_g$. Les valeurs unitaires donnent alors les prix de transaction moyens appropriés et l'indice de valeurs unitaires est l'indice de prix approprié des produits.

Bien sûr, le problème consiste à définir l'homogénéité. Il est possible de soutenir que des produits identiques physiquement mais vendus à différents endroits ne sont pas des produits identiques puisque différents services accompagnent les transactions, de sorte que l'homogénéité d'un point de vente à l'autre n'est jamais atteinte. Une autre formule d'indice devrait donc être choisie. Si u=s dans l'expression (6), P_g^t peut constituer un indice de prix de quantité fixe avec u comme période de référence de la quantité. Pour u=s=0, P_g^t devient l'indice de prix de Laspeyres et, pour u=s=t, P_g^t est l'indice de prix de Paasche. Théoriquement, nous ne pouvons préférer l'un ou l'autre. Pour des raisons de symétrie, il semble naturel de retenir la moyenne géométrique (non pondérée) des indices de Paasche et de Laspeyres, qui est l'indice de prix(idéal) de Fisher.

BIBLIOGRAPHIE

BALK, B.M. (1989). On Calculating the Precision of Consumer Price Indices. Rapport, Department of Price Statistics, Statistics Netherlands, Voorburg.

BALK, B.M. (1994). On the first step in the calculation of a consumer price index. *Proceedings of the First International Conference on Price Indices*, Statistique Canada, Ottawa.

BALK, B.M. (1998). On the use of unit value indices as consumer price subindices. *Proceedings of the Fourth International Conference on Price Indices*, U.S. Bureau of Labor Statistics, Washington, D.C.

BOON, M. (1997). Sampling Designs in Constructing Consumer Price Indices: Current Practices at Statistical Offices. Research Paper no. 9717, Research and Development Division, Statistics Netherlands, Voorburg.

BOSKIN, M.J., DULBERGER, E.R., GORDON, R.J., GRILICHES, Z., et JORGENSON, D. (1996). Toward a More Accurate Measure of the Cost of Living. Rapport final à la U.S. Senate Finance Committee, Washington, D.C.

BRADLEY, R., COOK, B., LEAVER, S.G., et MOULTON, B.R. (1997). An overview of research on potential uses of scanner data in the U.S. CPI. *Proceedings of the Third International Conference on Price Indices*, Statistics Netherlands, Voorburg.

DALÉN, J. (1991). Computing elementary aggregates in the Swedish consumer price index. *Journal of Official Statistics*, 8, 129-147.

DALÉN, J., et OHLSSON, E. (1995). Variance estimation in the Swedish consumer price index. *Journal of Business & Economic Statistics*, 13, 347-356.

DIEWERT, W.E. (1976). Exact and superlative index numbers. *Journal of Econometrics*, 4, 115-145.

DIEWERT, W.E. (1995). Axiomatic and Economic Approaches to Elementary Price Indexes. Working Paper no. 5104, National Bureau of Economic Research, Cambridge.

HAAN, J. De, et OPPERDOES, E. (1997a). Estimation of the Coffee Price Index Using Scanner Data: The Sampling of Commodities. Rapport de recherche, Socio-economic Statistics Division, Statistics Netherlands, Voorburg.

HAAN, J. De, et OPPERDOES, E. (1997b). Estimation of the coffee price index using scanner data: The choice of the micro index. *Proceedings of the Third International Conference on Price Indices*, Statistics Netherlands, Voorburg.

HAWKES, W.J. (1997). Reconciliation of consumer price index trends with corresponding trends in average prices for quasi-homogeneous goods using scanning data. *Proceedings of the Third International Conference on Price Indices*, Statistics Netherlands, Voorburg.

REINSDORF, M. (1995). Constructing Basic Component Indexes For The U.s. Cpi From Scanner Data: A Test Using Data On Coffee. Présenté à NBER Conference on Productivity, Cambridge, Mass., 17 juillet, 1995.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*, New York: Springer Verlag.

SILVER, M. (1995). Elementary aggregates, micro-indices and scanner data: some issues in the compilation of consumer price indices. *Review of Income and Wealth*, 41, 427-438.

SZULC, B.J. (1987). Price indices below the basic aggregation level. *ILO Bulletin of Labour Statistics*, Ré-imprimé dans Turvey (1989).

TURVEY, R. (1989). Consumer Price Indices: An ILO Manual. Geneva: International Labour Office.

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 47-60
Statistique Canada

Estimateurs de calage robustes

PIERRE DUCHESNE[1]

RÉSUMÉ

Nous considérons l'utilisation d'estimateurs de calage en présence de valeurs aberrantes. Une extension de la classe des estimateurs de calage de Deville et Särndal (1992) reposant sur les estimateurs QR de Wright (1983) est obtenue. Elle s'obtient aussi en minimisant une métrique générale sous des contraintes sur les variables de calage et sur les poids. Comme application, cette classe d'estimateurs nous permet de considérer des estimateurs de calage robustes en choisissant judicieusement ses paramètres. Il est ainsi possible, pour des besoins éventuellement cosmétiques, de restreindre les poids robustes à un intervalle spécifié à l'avance. L'utilisation d'estimateurs robustes avec haut point de rupture est considéré. Dans le cas particulier de la métrique quadratique, l'estimateur que nous suggérons est une généralisation d'une proposition de Lee (1991). Une brève étude de simulation illustre la nouvelle méthodologie.

MOTS CLÉS: Estimateur de calage; estimateur par la régression; pondération restreinte; robustesse.

1. INTRODUCTION

Le problème des valeurs aberrantes est important dans toutes les branches de la statistique. En théorie des sondages, le contexte est différent de celui de la statistique paramétrique puisque l'on s'intéresse souvent à l'estimation du total d'une variable d'intérêt y. Or une valeur aberrante peut avoir son plein poids dans le total de la population. De plus, le méthodologiste qui procède à la phase d'estimation peut supposer que les valeurs des unités sont enregistrées sans erreur, puisque souvent les unités une fois ramassées sont traitées dans un système d'édition (Särndal, Swensson et Wretman 1992, section 1.7). Cette étape fait partie de la réalisation d'un sondage dans les grandes agences statistiques, comme par exemple à Statistique Canada. Lee (1995) fait un survol des développements de la robustesse en théorie des sondages.

Cependant, puisqu'il arrive fréquemment que les populations dans les enquêtes économiques sont asymétriques, quelques unités risquent d'être extrêmes par rapport aux autres, sujet abordé dans Kish (1965). Retirer complètement ces unités entraîne des estimations biaisées, alors que si on les conserve avec leur plein poids un estimateur comme celui par la régression généralisée (GREG) risque d'être très variable. Ceci suggère un compromis entre le biais et la variance. En présence de valeurs aberrantes, le défi est de proposer des estimateurs robustes du total peu affectés par certaines unités très à l'écart des autres. Ces estimateurs devront avoir un petit biais et une petite erreur quadratique moyenne. Traditionnellement, la théorie de l'échantillonnage a consacré beaucoup d'efforts à développer des estimateurs sans biais ou asymptotiquement sans biais (ADU). Voir par exemple Särndal et coll. (1992, section 7.12). Cependant, la propriété ADU n'est peut-être pas souhaitable dans un contexte de valeurs aberrantes. Voir la discussion de Chambers et Kokic (1993) qui montrent qu'un conflit est présent entre la propriété ADU et la robustesse d'un estimateur.

Considérons l'estimateur Horvitz-Thompson (HT), défini par $\hat{T}_{y\mathrm{HT}} = \sum_s d_k y_k$, où $d_k = \pi_k^{-1}$, π_k étant la probabilité d'inclusion (Si A est un ensemble d'unités, $A \subset U$, alors $\sum_A$ est une notation signifiant $\sum_{k \in A}$). Supposons la variable d'intérêt y positive et la population asymétrique. L'estimateur HT étant une moyenne pondérée par les d_k il est vulnérable aux grandes valeurs de y. Une unité avec un poids d_k élevé peut aussi avoir un grand impact sur l'estimation en entraînant des estimations variables. Lee (1995) définit ces unités comme influentes. Une unité extrême n'est pas forcément influente si son poids d_k est suffisamment petit. Traditionnellement, les méthodologistes cherchent à limiter l'impact des unités influentes lorsqu'elles sont connues avant le sondage, en donnant par exemple des poids d'échantillonnage proche de un aux unités extrêmes. Gambino (1987) et Lee (1995) discutent cependant de situations où nous ne pouvons pas agir ainsi. Dans un article important, Hidiroglou et Srinath (1981) considèrent l'altération des poids d'échantillonnage en présence de valeurs aberrantes. Leur approche a grandement légitimé la modification des poids dans le monde des sondages.

Parmi les premières alternatives robustes du total, plusieurs reposent sur les M-estimateurs et les GM-estimateurs. Cependant, beaucoup d'attention a été consacrée récemment à des estimateurs possédant en plus de bonnes propriétés de robustesse globale, telle que mesurée par le point de rupture d'un estimateur. Ces notions sont discutées par exemple dans Donoho et Huber (1983), Hampel, Ronchetti, Rousseeuw et Stahel (1986) et Rousseeuw et Leroy (1987). Le point de rupture mesure le pourcentage de valeurs aberrantes dans l'échantillon que l'estimateur peut tolérer tout en donnant tout de même une bonne estimation d'une certaine caractéristique de la

[1] Pierre Duchesne, Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal, (Québec), H3C 3J7.

population. Lee, Ghangurde, Mach et Yung (1992) réclament des estimateurs du total basés sur des estimateurs robustes avec un haut point de rupture.

Nous allons considérer des estimateurs de calage du total T_y s'écrivant comme $\sum_s w_k y_k$. Ces estimateurs sont développés par exemple dans Deville et Särndal (1992). On cherche les poids w_k aussi près que possible des poids d'échantillonnage $d_k = \pi_k^{-1}$ mais en respectant les contraintes d'étalonnage, notée CE (dites aussi contraintes de calage),

$$\sum_s w_k x_k = T_x, \tag{1.1}$$

où x_k est un vecteur de dimension m correspondant à l'information auxiliaire disponible de total $T_x = \sum_U x_k$ connu. Ces estimateurs sont populaires car ils s'interprètent aisément, puisque les méthodologistes sont habitués d'apposer des poids w_k aux unités y_k. Plusieurs métriques sont étudiées afin de mesurer la proximité entre d_k et w_k. L'estimateur GREG est un exemple important avec $w_k = d_k(1 + (T_x - \hat{T}_{x\mathrm{HT}})' M_s^{-1} x_k / c_k)$, où $M_s = \sum_s d_k x_k x_k' / c_k$. Il est obtenu en minimisant la métrique quadratique $\sum_s c_k (w_k - d_k)^2 / d_k$. Les constantes c_k sont des facteurs de pondération qui peuvent prendre en compte (par exemple) des problèmes d'hétéroscédasticité. Särndal (1996) discute du choix de ces constantes. Cependant, puisque les poids-g $g_k = w_k / d_k$ du GREG ne sont pas bornés en général, d'autres métriques sont proposées afin de les borner pour qu'ils satisfassent certaines restrictions applicables à la fourchette des valeurs (RAFV). Ceci permet en particulier d'éviter la présence indésirable de poids w_k négatifs. Voir Deville et Särndal (1992), Singh et Mohl (1996) et Stukel, Hidiroglou et Särndal (1996).

Les estimateurs de calage ont un lien avec les méthodes robustes comme noté dans Fuller, Loughin et Baker (1994, p. 81). Cependant, il est inexact de prétendre que les estimateurs de calage possèdent nécessairement de bonnes propriétés de robustesse, puisqu'il suffit de se rappeler que tous les estimateurs de calage considérés par Deville et Särndal (1992) sont asymptotiquement équivalents au GREG qui étant ADU n'est pas robuste. De plus, on voit qu'un estimateur traditionnel de calage n'est pas robuste puisqu'il dépend linéairement de w_k, et w_k ne prend pas en compte y_k.

Le but de cet article est de construire des estimateurs de la forme $\sum_s w_k y_k$ où les poids w_k assurent en même temps la robustesse et le respect des contraintes sur les variables de calage et sur les poids w_k. Le point de départ de notre approche est la classe des estimateurs QR de Wright (1983). Supposons que nous disposions des constantes $\{(q_k, r_k), q_k > 0, r_k \geq 0, \forall\, k \in U\}$, telles que $\sum_U \pi_k q_k x_k x_k' > 0$ et $\sum_s q_k x_k x_k' > 0, \forall s$. (Si A est une matrice symétrique, $A>0$ signifie que A est définie positive). Les estimateurs QR sont définis à partir de q_k et r_k par la relation

$$\hat{T}_{\mathrm{yQR}} = T_x' \hat{B}_q + \sum_s r_k e_k, \tag{1.2}$$

où $\hat{B}_q$ admet une forme pondérée par les q_k

$$\hat{B}_q = \left(\sum_s q_k x_k x_k'\right)^{-1} \sum_s q_k x_k y_k, \tag{1.3}$$

et

$$e_k = y_k - x_k' \hat{B}_q. \tag{1.4}$$

Nous montrons à la section 2 que les estimateurs QR sont des estimateurs de calage et introduisons une nouvelle classe d'estimateurs, notée RQR, reposant également sur le choix des constantes q_k et r_k. Elle généralise d'une certaine façon les estimateurs QR ainsi que la classe des estimateurs de Deville et Särndal (1992). L'intérêt de la classe RQR est qu'elle permet d'obtenir des poids w_k restreints à un intervalle donné, disons $[L, U]$. Quelques propriétés des classes QR et RQR sont données à la section 2.

La section 3 contient des applications de la classe RQR pour la construction d'estimateurs de calage robustes. L'idée principale consiste à modifier des poids implicites robustes pour qu'ils respectent les contraintes de calage. Nous discutons à la section 3.1 du choix des constantes q_k et r_k en utilisant un raisonnement propre aux estimateurs de calage. Cette approche s'avère nouvelle et unificatrice, et nous guide à la section 3.2 dans le choix des q_k et r_k en présence d'information auxiliaire. Un élément important est l'utilisation d'un estimateur robuste admettant la forme pondérée (1.3), fournissant les q_k. Nous remarquons que c'est le cas pour les GM-estimateurs. Les estimateurs avec un haut point de rupture n'ont habituellement pas une forme pondérée. Une repondération de ces estimateurs est considérée, permettant de conserver le point de rupture sous contrôle et d'avoir des estimateurs s'écrivant sous la forme (1.3). Voir Rousseeuw et Leroy (1987) et Simpson et Chang (1997). Nous discutons alors des choix de q_k et r_k permettant de calculer un estimateur RQR afin d'obtenir un estimateur de calage robuste avec des poids bornés. Nous comparons dans la section 4 divers estimateurs robustes, dont l'estimateur de Lee (1991) et celui de Chambers (1986), aux estimateurs RQR ainsi qu'au GREG et à un estimateur de calage considéré dans Deville et Särndal (1992) dont les poids sont restreints. L'estimateur de Lee (1991) peut être considéré comme un cas particulier de notre approche. Il nous permet aussi de considérer un nouvel estimateur avec des poids bornés. Quatre populations déjà étudiées dans la littérature sont considérées. On notera que les estimateurs sans contrainte sur les poids sont sujets à des problèmes de pondération négative. Les estimateurs de la classe RQR permettent d'obtenir des estimateurs robustes possédant des poids positifs, se comparant bien aux estimateurs sans contrainte sur les poids. Finalement, nous concluons dans la section 5. Notons que l'appendice B contient une liste des abréviations et que l'appendice C contient une liste des diverses constantes contenues dans l'article avec leurs définitions.

2. ESTIMATEURS DE LA CLASSE RQR

Soit une population finie $U = \{1, 2, ..., N\}$ de taille N dont nous désirons estimer le total $T_y = \sum_U y_k$ pour une variable d'intérêt y positive. On sélectionne un échantillon s de taille n_s selon un plan d'échantillonnage $p(s)$. La probabilité d'inclusion d'une unité k est notée π_k et les probabilités d'inclusion d'ordre deux sont notées π_{kl}. Nous supposons que l'information auxiliaire x_k est unitaire, c'est-à-dire que x_k est connu de source sûre, $\forall k \in U$.

Wright (1983) introduit la classe des estimateurs QR s'écrivant sous la forme (1.2) avec pour objectif principal d'unifier un grand nombre d'estimateurs familiers. On retrouve le meilleur estimateur linéaire sans biais (BLUP) de Royall (1970) issu de la théorie basée sur les modèles, obtenu en posant $(q_k, r_k) = (1/c_k, 1)$ et le GREG de Cassel, Särndal et Wretman (1976) en considérant le choix $(q_k, r_k) = (d_k/c_k, d_k)$. Alternativement, (1.2) peut s'écrire comme

$$\hat{T}_{y\mathrm{QR}} = \sum_s d_k g_k y_k,$$

où $d_k g_k$ satisfait

$$d_k g_k = r_k + (T_x - \hat{T}_{xr})' \left(\sum_s q_k x_k x_k' \right)^{-1} q_k x_k, \qquad (2.1)$$

avec $\hat{T}_{xr} = \sum_s r_k x_k$. Si on pose $r_k = d_k$, g_k correspond au poids-g du GREG.

Les estimateurs QR sont des estimateurs de calage, obtenus de la minimisation de la métrique quadratique sujet aux CE

$$\min \frac{1}{2} \sum_s (w_k - r_k)^2/q_k, \text{ tel que } \sum_s w_k x_k = T_x. \qquad (2.2)$$

Les poids w_k sont choisis aussi près que possible des r_k et les q_k sont des facteurs de pondération. Autrement formulé, les poids de départ r_k sont transformés en poids calés w_k. La solution au problème (2.2) est $w_k = d_k g_k$, où $d_k g_k$ est donné dans la formule (2.1).

Cependant, rien ne garantit que les poids w_k de l'estimateur QR soient positifs, ce qui peut être indésirable pour le praticien. Voir Brewer (1994), qui formalise l'interprétation des poids. Afin de restreindre les poids w_k dans $[L, U]$, nous considérons de résoudre

$$\min \sum_s G(w_k; q_k, r_k),$$
$$\text{tel que } \sum_s w_k x_k = T_x \text{ et } w_k \in [L, U]. \qquad (2.3)$$

L'estimateur de calage du total est

$$\hat{T}_{y\mathrm{RQR}} = \sum_s w_k y_k, \qquad (2.4)$$

où les w_k sont obtenus de la résolution du problème (2.3). La fonction $G(w; q, r)$ est supposée strictement convexe et dérivable en w pour r et q fixés. Nous notons $g(u; q, r) = G'(u; q, r)$ et $h(u; q, r) = g^{-1}(u; q, r)$. On suppose de plus que $h(0; q, r) = r$ et $h'(0; q, r) = q$. Nous appelons les estimateurs obtenus les estimateurs de calage restreints QR (RQR).

Fuller et coll. (1994) privilégient des estimateurs de régression possédant des propriétés raisonnables d'invariance. On peut montrer que les estimateurs RQR sont équivariants de régression et d'échelle lorsque les constantes q_k et r_k sont invariantes aux transformations. Voir pour les définitions utiles Bolfarine et Zacks (1992).

L'existence d'une solution au problème (2.3) n'est pas garantie. Voir l'étude de simulation de Stukel et coll. (1996). Il peut arriver par exemple des réalisations de l'échantillon pour lequel on ne peut même pas satisfaire les CE (1.1). Ainsi, l'échantillon est tellement débalancé qu'il est impossible que la somme pondérée des composantes de chaque dimension puisse donner le total de la population correspondant. Le seul recours du praticien est alors de relaxer les contraintes en réduisant la dimension du nombre de variables auxiliaires. Voir aussi la discussion de Fuller et coll. (1994). Dans le cas des estimateurs de calage considérés dans Deville et Särndal (1992), les auteurs montrent dans leur résultat 1 qu'une solution existe avec probabilité tendant vers un. Sous certaines conditions, ce résultat peut être adapté aux estimateurs de la classe RQR.

La métrique qui retiendra notre attention afin que les poids satisfassent les RAFV est une légère modification du cas 7 de Deville et Särndal (1992). Nous l'appelons la métrique quadratique restreinte. La fonction G correspondant au choix de cette métrique est

$$G(w_k; q_k, r_k) = \begin{cases} \frac{1}{2}(w_k - r_k)^2/q_k & \text{si } w_k \in [L, U], \\ \infty & \text{sinon,} \end{cases}$$

alors que la fonction h est

$$h(x_k'\lambda; q_k, r_k) = \begin{cases} L & r_k + q_k x_k'\lambda < L, \\ r_k + q_k x_k'\lambda & r_k + q_k x_k'\lambda \in [L, U], \\ U & r_k + q_k x_k'\lambda > U. \end{cases}$$

Avec cette modification, c'est le poids w_k qui est contraint et non pas seulement w_k/d_k comme dans le cas 7 de Deville et Särndal (1992). Dans notre situation, w_k peut «corriger» un poids initial qui est aberrant. Notons que telle que formulée la métrique de Deville et Särndal (1992) insère subtilement les contraintes sur les w_k dans la fonction G. Pour calculer l'estimateur (2.4) selon cette métrique, il suffit de respecter la même démarche que Deville et Särndal (1992), ce qui nous amène à résoudre avec la méthode de Newton l'équation suivante en λ

$$\sum_s h(x_k'\lambda; q_k, r_k) x_k = T_x. \qquad (2.5)$$

L'estimateur final est $\hat{T}_{y\mathrm{RQR}} = \sum_s h(x_k'\lambda_\infty; q_k, r_k) y_k$ où λ_∞ est la solution de l'équation (2.5).

Il est d'intérêt de savoir si la contrainte des poids change les propriétés de l'estimateur comparativement à l'estimateur QR sans contrainte de poids. Le résultat suivant (dont la preuve est en annexe) montre sous certaines conditions que les deux estimateurs sont asymptotiquement équivalents. En pratique, avec la métrique quadratique restreinte, nous n'avons pas observé d'écarts importants.

Proposition 1. Sous les hypothèses C_1 et C_2 données en appendice,

$$N^{-1}|\hat{T}_{yQR} - \hat{T}_{yRQR}| = o_p(n^{-1/2}). \tag{2.6}$$

Ce résultat s'obtient possiblement de la démarche pour obtenir le résultat 5 dans Deville et Särndal (1992) portant sur l'équivalence asymptotique entre le GREG et les estimateurs de calage considérés par ces auteurs. Cependant, la proposition 1 peut être utile pour comprendre le type de conditions nécessaires pour avoir le résultat annoncé dans notre contexte.

Puisque (1.2) a le même comportement asymptotique que la quantité $T_x' B_q + \sum_s w_k E_k$, où $E_k = y_k - x_k' B_q$ et $B_q = (\sum_U \pi_k q_k x_k x_k')^{-1} \sum_U \pi_k q_k x_k y_k$, ceci suggère pour estimateur de variance

$$v_L = \sum \sum_s \check{\Delta}_{kl}(w_k e_k)(w_l e_l), \tag{2.7}$$

où $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, $\check{\Delta}_{kl} = \Delta_{kl}/\pi_{kl}$ et e_k est donné dans (1.4). Voir Särndal et coll. (1989) et Särndal et coll. (1992, p. 234). On peut montrer que le biais asymptotique d'un estimateur QR est donné sous des conditions générales par

$$E_p(\hat{T}_{yQR} - T_y) = \sum_U (\pi_k r_k - 1) E_k.$$

Un estimateur possible du biais est alors $b = \sum_s d_k(\pi_k r_k - 1) e_k$, que l'on peut utiliser conjointement avec la formule (2.7) pour construire un estimateur de l'erreur quadratique moyenne pour un estimateur QR et RQR, en utilisant la proposition 1.

Les estimateurs RQR permettent d'obtenir des estimateurs de calage avec des poids contraints. Avec q_k et r_k déterminées, il suffit de procéder à la résolution du problème (2.3). Les sections qui suivent sont des applications de la classe RQR dans un contexte de robustesse. Nous verrons comment devrait s'orienter le choix des constantes q_k et r_k, choisies en pratique avec l'aide de l'échantillon s.

3. CONSTRUCTION D'ESTIMATEURS ROBUSTES ET CALÉS

3. Méthodes reposant sur la réduction de poids et la modification des valeurs

Lee (1995) discute de diverses propositions reposant sur la méthode de réduction de poids dans un tirage aléatoire simple. Une fois les observations aberrantes détectées, ces méthodes consistent à réduire les poids des observations extrêmes. Ces méthodes sont préférables à celles qui retirent complètement les observations douteuses puisque toutes les observations dans l'échantillon sont légitimes, comme discuté dans Lee et coll. (1992).

Traduit dans le langage des estimateurs de calage, commençons par considérer la situation où nous ne disposons pas d'information auxiliaire et où la seule contrainte est $\sum_s w_k = N$. Ce cas sert à motiver notre démarche. Considérons l'estimateur QR avec $q_k = r_k$. Pour les fins de notre discussion, on considère les constantes r_k connues et fixées. Les poids minimisant (2.2) sujet à $\sum_s w_k = N$ sont $w_k = C_s(r) r_k$, où $C_s(r) = N/\sum_s r_k$, de sorte que $\hat{T}_{yQR}$ devient

$$\hat{T}_{yQR} = C_s(r) \sum_s r_k y_k. \tag{3.1}$$

Lorsqu'une observation est extrême, elle représente potentiellement peu d'unités comme elle dans la population et on devrait alors peut-être réduire son poids. Afin de satisfaire les CE, ceci suggère de trouver des poids w_k les plus près possible des poids d'échantillonnage d_k pour les unités qui ne sont pas aberrantes mais aussi près que possible d'un facteur de réduction r pour les unités aberrantes, où r est choisi par le statisticien. Plus précisément, notons $s = s_1 \cup s_2$, où s_1 de cardinalité n_1 représente les unités qui ne sont pas déclarées aberrantes, alors que $s_2 = s - s_1$ de cardinalité $n_2 = n - n_1$ représente les unités aberrantes de s. Le facteur de réduction r satisfera typiquement $r \le d_k, \forall k \in s_2$. Par exemple, considérons l'estimateur (3.1) avec $q_k = r_k = B_k = d_k I_{k1} + r(1 - I_{k1})$, où I_{k1} est la variable indicatrice de l'appartenance à s_1. Ainsi les constantes q_k et r_k sont réduites pour les unités de s_2 afin de refléter que les unités de s_2 sont extrêmes. L'estimateur (3.1) devient

$$\hat{T}_{yQR} = C_s(B)\left(\sum_{s1} d_k y_k + r \sum_{s2} y_k\right).$$

Dans le cas du tirage aléatoire simple, $d_k = N/n$ et on obtient

$$\hat{T}_{yQR} = C_s(B) \hat{T}_{yB},$$

où $\hat{T}_{yB} = N/n \sum_{s1} y_k + r \sum_{s2} y_k$ est l'estimateur de Bershad (1960) discuté dans Lee (1995). D'autres méthodes reposant sur la méthode de réduction de poids sont discutées dans Lee (1995), qui discute également du choix de r.

Un désavantage des méthodes reposant sur la réduction de poids est que l'analyste doit lui-même identifier les unités aberrantes. Les méthodes reposant sur la modification des valeurs permettent d'éviter cette difficulté en réduisant graduellement le poids des unités plus extrêmes. Considérons un tirage aléatoire simple. Posons

$$m(y_k; t, a, b) = b + (a - b) \min(1, t/y_k). \tag{3.2}$$

Ainsi, cette fonction attribue un poids de départ valant a pour les $y_k < t$, et graduellement réduit jusqu'à un poids final b, à mesure que y_k devient extrême. La valeur t est appelée le seuil. Les constantes a, b et t sont choisies par le statisticien. Plusieurs valeurs pour a et b ont été considérées dans la littérature. Ainsi, plutôt que d'attribuer un facteur de réduction fixe aux unités de s_2, choisissons $q_k = r_k = W_k = m(y_k; t, N/n, fN/n)$, où f est une constante comprise entre 0 et 1. L'estimateur (3.1) devient

$$\hat{T}_{y\mathrm{QR}} = C_s(W)\sum_s W_k y_k$$
$$= C_s(W)\hat{T}_{yW}.$$

L'estimateur $\hat{T}_{yW}$ est discuté dans Gross, Bode, Taylor et Lloyd-Smith (1986) ainsi que dans Chambers et Kokic (1993), qui l'appellent l'estimateur winsorisé. Il est un cas particulier de l'approche de Chambers (1982, 1986). Lorsque $f = 0$, l'estimateur (3.1) devient $\hat{T}_{y\mathrm{QR}} = C_s(W_I)\hat{T}_{yW_1}$ avec $q_k = r_k = W_{Ik} = m(y_k; t, N/n, 0)$, où $\hat{T}_{yW_1} = N/n(\sum_{s1} y_k + n_2 t)$, s_1 dénotant la partie de s contenant les unités satisfaisant $y_k < t$. L'estimateur $\hat{T}_{yW_1}$ est discuté dans Lee (1995), ainsi que dans Gross et coll. (1986) qui l'appellent l'estimateur winsorisé de type I. Lorsque $f = n/N$, Gross et coll. (1986) l'appellent l'estimateur winsorisé de type II. Il est également discuté dans Bruce (1991).

Dans un plan ptπ, Dalén (1987) a inséré le plan en posant $D_k = m(y_k; \pi_k t, d_k, 1)$. Ainsi, si k et l sont deux observations extrêmes telles que $y_k = y_l$, alors celle dont le poids d'échantillonnage est le plus grand aura un poids D_k plus élevé. Le choix $r_k = q_k = D_k$ permet d'obtenir essentiellement l'estimateur de Dalén, $\hat{T}_{y\,\mathrm{QR}} = C_s(D)\sum_s D_k y_k$. L'estimateur $\hat{T}_{yD} = \sum_s D_k y_k$ été étudié notamment dans Tambay (1988).

Tableau 3.1
Estimateur (3.1) reposant sur la réduction de poids et la modification des valeurs

Estimateur	Valeurs de $q_k = r_k$
Bershad	$B_k = d_k I_{kl} + r(1 - I_{kl})$
Winsorisé	$W_k = m(y_k; t, N/n, fN/n)$
Winsorisé, type I	$W_{Ik} = m(y_k; t, N/n, 0)$
Winsorisé, type II	$W_{IIk} = m(y_k; t, N/n, 1)$
Dalén	$D_k = m(y_k; \pi_k t, d_k, 1)$

Note: $m(y_k; t, a, b) = b + (a - b)\ \min(1, t/y_k)$.

La démarche de cette section suggère que nous pouvons à l'occasion vouloir des estimateurs dont les poids sont près des r_k plutôt que des poids d'échantillonnage d_k. Les constantes r_k seront choisies elle-mêmes près des d_k pour les bonnes unités, mais réduites lorsqu'une unité est déclarée extrême. Les estimateurs QR unifient les méthodes de réduction de poids et de modification des valeurs. Les méthodes reposant sur la modification des valeurs nous amènent à choisir les poids de manière adaptative à l'échantillon particulier s choisi. Comme noté dans Chambers et Kokic (1993) cela n'est pas surprenant puisque le problème des valeurs aberrantes survient après la sélection de l'échantillon s. On doit utiliser l'échantillon à notre disposition pour tenter de le régler. Dans la prochaine section nous généralisons ces méthodes en présence d'information auxiliaire.

3.2 Estimateurs du total basés sur des statistiques robustes

Parmi les premières tentatives vers l'obtention d'alternatives robustes pour le total de la population en présence d'information auxiliaire on retrouve Chambers (1982, 1986) qui propose un estimateur par le ratio robuste basé sur une décomposition de l'estimateur BLUP. Une extension récente des travaux de Chambers se retrouve dans Welsh et Ronchetti (1998). Gwet et Rivest (1992) proposent également une version robuste de l'estimateur par le ratio en basant leur approche sur le plan d'échantillonnage dans un tirage aléatoire simple. Rivest et Rouillard (1991) font une étude comparative de plusieurs estimateurs robustes et étudient plusieurs estimateurs de l'erreur quadratique moyenne. Dans les plans avec probabilités inégales, Hulliger (1995) considère une robustification de l'estimateur HT lorsque les probabilités d'inclusion sont obtenues avec l'information auxiliaire. Gwet et Rivest (1992) et Hulliger (1995) considèrent une version de la fonction d'influence pour populations finies, insistant sur le besoin de procédures possédant de bonnes propriétés de robustesse locales et l'utilisation d'estimateurs ayant des fonctions d'influence bornées. Les fonctions d'influence sont discutées dans un contexte général dans Hampel et coll. (1986).

Nous allons voir dans les sections suivantes comment construire des estimateurs robustes avec des poids contraints. La construction de ces estimateurs repose sur les étapes suivantes:

- Identification des constantes q_k et r_k. Ceci nous donne un estimateur QR.
- Résolution du problème (2.3) afin d'obtenir un estimateur RQR.

D'un point de vue de la robustesse, les coefficients q_k sont choisis de telle sorte que $\hat{B}_q$ est un estimateur robuste. De cette manière, la première partie de l'estimateur QR, $T_x'\hat{B}_q$, donne une bonne valeur prédite pour toute la population. La deuxième partie de l'estimateur QR, $\sum_s r_k e_k$, corrige la première partie pour les y_k observés dans l'échantillon. Les constantes r_k évitent qu'avec cette correction les valeurs aberrantes dans l'échantillon reviennent avec plein poids.

3.2.1. Choix de q_k basé sur un GM-estimateur

Considérons l'estimateur (1.2) dans lequel on remplace $\hat{B}_q$ par un estimateur robuste d'un coefficient de régression.

Ces estimateurs sont discutés par exemple dans Huber (1981) ou Hampel et coll. (1986). On obtient ainsi

$$\hat{T}_g = T_x'\hat{B}_g + \sum_s r_k\left(y_k - x_k'\hat{B}_g\right). \tag{3.3}$$

L'estimateur (3.3) n'est pas sous la forme des estimateurs QR à moins que $\hat{B}_g$ admette une forme pondérée. C'est le cas si $\hat{B}_g$ est un GM-estimateur défini par l'équation

$$\sum_s d_k h_k x_k \psi\left(\left(y_k - x_k'B\right)\Big/\left(\sigma h_k^{\alpha}\sqrt{c_k}\right)\right)\Big/\sqrt{c_k} = 0, \tag{3.4}$$

puisque la solution de (3.4) peut s'exprimer comme

$$\hat{B}_g = \left(\sum_s d_k h_k^{1-\alpha} u_k x_k x_k'/c_k\right)^{-1} \sum_s d_k h_k^{1-\alpha} u_k x_k y_k/c_k,$$

où

$$u_k = \frac{\psi\left(\left(y_k - x_k'\hat{B}_g\right)\Big/\left(\sigma h_k^{\alpha}\sqrt{c_k}\right)\right)}{\left(y_k - x_k'\hat{B}_g\right)\Big/\left(\sigma h_k^{\alpha}\sqrt{c_k}\right)}.$$

Les propriétés des GM-estimateurs sont discutées dans Simpson et Chang (1997). On suppose que σ est connu afin de simplifier notre discussion, et le rôle de c_k est le même que dans le cas du GREG. La fonction ψ est déterminée par l'analyste. Un exemple d'un usage courant est la fonction de Huber

$$\psi_{Hub}(x;c) = \begin{cases} c & \text{si } x > c, \\ x & \text{si } |x| \le c, \\ -c & \text{si } x < -c. \end{cases} \tag{3.5}$$

Une valeur de c autour de 2 est souvent utilisée dans le calcul des GM-estimateurs. Voir par exemple Hampel *et coll.* (1986), Gwet et Rivest (1992) et Hulliger (1995).

Le choix de h_k permet de borner l'influence d'une information auxiliaire trop extrême. La constante $\alpha = 0$ donne lieu au choix de Mallows alors que $\alpha = 1$ permet d'obtenir la version de Schweppe. La version de Schweppe est parfois privilégiée. Voir Coakley et Hettmansperger (1993) et Hampel et coll. (1986, p. 322). En présence d'information auxiliaire minimale, c'est-à-dire lorsque nous disposons que d'une variable réelle $x_k, \forall k \in U$, un choix possible de fonction h_k est

$$h_k = \min\left(1, \frac{t}{x_k/\mathrm{med}(x_k)}\right). \tag{3.6}$$

Dans un plan ptπ, une modification de h_k inspirée de Dalén (1987) pour tenir compte des différents poids d'échantillonnage serait peut-être souhaitable. La constante t doit être spécifiée par le statisticien. Une valeur de t autour de 1,5 est rencontrée dans les applications. Voir par exemple Rivest et Rouillard (1991) qui donnent aussi d'autres choix de fonctions h_k.

L'écriture de $\hat{B}_g$ comme un estimateur pondéré permet d'écrire l'estimateur (3.3) comme un estimateur QR avec

$$(q_k, r_k) = (d_k h_k^{1-\alpha} u_k/c_k, r_k).$$

Nous discuterons du choix des constantes r_k dans la section 3.2.3.

3.2.2 Choix de q_k basé sur un estimateur avec haut point de rupture

Le choix d'un GM-estimateur n'est qu'un premier pas afin d'obtenir un estimateur très robuste du total. En effet, bien que la fonction d'influence des GM-estimateurs soit bornée, il n'en demeure pas moins que ces estimateurs ne possèdent pas un haut point de rupture, qui décroit habituellement avec la dimension de l'information auxiliaire (Rousseeuw et Leroy 1987, p. 13). Nous verrons dans cette section comment construire des estimateurs de calage robustes basés sur des estimateurs avec un haut point de rupture. Comme ces estimateurs ne possèdent habituellement pas la forme pondérée, nous considérons une repondération de ces estimateurs. Ceci nous permettra de dégager, comme dans la section précédente, les constantes q_k nécessaires au calcul de la métrique de l'estimateur RQR. Plus précisément, on considère les poids $\hat{u}_k$ suivants

$$\hat{u}_k = \frac{\psi\left(\left(y_k - x_k'\hat{B}_0\right)\Big/\left(\sigma h_k^{\alpha}\sqrt{c_k}\right)\right)}{\left(y_k - x_k'\hat{B}_0\right)\Big/\left(\sigma h_k^{\alpha}\sqrt{c_k}\right)}, \tag{3.7}$$

où $\hat{B}_0$ est un estimateur équivariant avec un haut point de rupture satisfaisant certaines conditions de régularité. L'estimateur repondéré est

$$\hat{B}_r = \left(\sum_s d_k h_k^{1-\alpha}\hat{u}_k x_k x_k'/c_k\right)^{-1} \times \sum_s d_k h_k^{1-\alpha}\hat{u}_k x_k y_k/c_k. \tag{3.8}$$

Les propriétés asymptotiques de ce type d'estimateurs sont étudiées dans Simpson et Chang (1997).

L'estimateur $\hat{B}_0$ que l'on considère est le GM-estimateur en une étape de Coakley et Hettmansperger (1993). Cet estimateur possède un haut point de rupture. Il s'obtient comme la première itération de la formule de Newton dans l'équation (3.4), où l'on adopte la version de Schweppe en posant $\alpha = 1$. D'autres choix d'estimateurs robustes auraient pu être utilisés. Cependant, l'efficacité jumelée avec les propriétés de robustesse de l'estimateur de Coakley et Hettmansperger (1993) en font un bon choix. Ainsi, la constante q_k que l'on propose est

$$q_k = d_k h_k^{1-\alpha}\hat{u}_k/c_k,$$

avec $\hat{B}_0 = \hat{B}_{CH}$, $\hat{B}_{CH}$ désignant l'estimateur de Coakley et Hettmansperger (1993).

3.2.3 Choix de r_k.

Ayant déterminé les constantes q_k, nous devons maintenant choisir les constantes r_k. Si $d_k = r_k$ alors l'estimateur QR est sous des conditions générales un

estimateur ADU. Cependant ce choix de r_k donne un estimateur sensible aux valeurs aberrantes. Alternativement le choix $r_k = 0$ procure un estimateur robuste qui risque d'être très biaisé comme souligné dans Gwet et Rivest (1992, p. 1180). Lee (1991) suggère le choix $r_k = \theta d_k$, où $\theta \in [0,1]$. Le biais asymptotique devient sous des conditions générales $(\theta - 1)\sum_U E_k$, où E_k représente les résidus obtenus en ajustant un estimateur robuste sur toute la population. Le choix de θ permet de contrôler le biais de l'estimateur. La discussion de la section 3 nous amène à suggérer des constantes r_k proches des d_k pour les bonnes unités et graduellement réduites pour les observations suspectes. Nous suggérons le choix

$$r_k = d_k u_k^*, \tag{3.9}$$

où

$$u_k^* = \frac{\psi^*\left[(y_k - x_k'\hat{B}_r)/(\sigma h_k^a \sqrt{c_k})\right]}{(y_k - x_k'\hat{B}_r)/(\sigma h_k^a \sqrt{c_k})}.$$

La fonction ψ^* que nous allons considérer est une modification de la fonction de Huber

$$\psi^*(x) = \begin{cases} x & \text{si } |x| \le a, \\ a\ \text{sign}(x) & \text{si } |x| > a \text{ et } |x| < a/b, \\ bx & \text{si } |x| > a/b. \end{cases} \tag{3.10}$$

Nous choisissons $a = 9$, $b = 1/4$. La raison de cette modification est que nous ne voulons pas que les valeurs aberrantes possédant de gros résidus ou une information auxiliaire extrême aient des poids trop réduits. Ainsi le poids d'échantillonnage est pleinement conservé lorsque l'argument de u_k^* est entre -9 et 9, et graduellement réduit jusqu'au quart. Si on réduit trop le poids des grands résidus alors le biais devient trop grand, ce qui motive le choix de ψ^*. Le choix des constantes r_k s'est effectué de manière empirique et semble bien fonctionner en pratique.

Ainsi, on se propose de considérer le choix des constantes q_k et r_k suivant

$$(q_k, r_k) = (d_k h_k^{1-\alpha} \hat{u}_k / c_k, d_k u_k^*). \tag{3.11}$$

Notre suggestion est une généralisation de la proposition de Lee (1991) puisque plutôt que de considérer $r_k = d_k\theta$ où θ est fixe, $r_k = d_k u_k^*$ s'adaptera de manière automatique (ou adaptative) avec l'échantillon. Ayant ce choix des constantes q_k et r_k à notre disposition et avec la métrique quadratique usuelle, nous obtenons un estimateur QR. Il est cependant sujet à des problèmes de pondération négative. Cependant avec les constantes (3.11) nous pouvons considérer la métrique quadratique restreinte, résoudre le problème (2.3) et obtenir un estimateur robuste satisfaisant les CE et les RAFV.

Le comportement asymptotique de l'estimateur RQR résultant comparativement à l'estimateur QR sans contrainte de poids trouve des éléments de réponse dans la proposition 1. Cependant, puisque les constantes (3.11) dépendent de s de façon compliquée, ceci ne nous permet pas de conclure automatiquement à l'équivalence asymptotique. Cependant, l'étude de simulation de la section 4 semble suggérer un comportement très comparable entre l'estimateur avec et sans contrainte sur les poids, en ce qui a trait à l'erreur quadratique moyenne de Monte-Carlo. Ainsi, l'évidence empirique montre que si les q_k et r_k sont choisies de telle sorte que l'estimateur sans contrainte sur les poids est robuste, alors la version avec des contraintes de poids sera aussi robuste.

Finalement, résumons les étapes de la méthode proposée afin d'obtenir un estimateur RQR robuste.

1. Choix des constantes q_k et r_k. Nous suggérons les constantes que l'on retrouve à l'équation (3.11). Cette étape nécessite le calcul de $\hat{B}_{CH}$.
2. Choix de la métrique. Si besoin est, choix des constantes L et U. On choisit ces constantes telles que $L \le r_k \le U$, $\forall k \in s$.
3. Résolution avec la méthode de Newton de l'équation (2.5).
4. Poser $w_k = h(x_k'\lambda_\infty; q_k, r_k)$ pour λ_∞ solution de l'étape 3.
5. Poser $\hat{T}_{yr} = \sum_s w_k y_k$, qui est l'estimateur RQR que l'on propose.

La procédure nécessite un certain nombre de constantes. On retrouve les constantes α, t et c dans le calcul de q_k et r_k. Le choix de ces valeurs est cependant justifié avec l'aide de la théorie de la robustesse, ce qui permet de guider le praticien. Par exemple, la valeur du c de la fonction de Huber peut être obtenue en tenant compte de préoccupations d'efficacité sous erreurs normales. Voir Hampel et coll. (1986, p. 333) et Gwet et Rivest (1992). On retrouve aussi les constantes a et b, qui sont plus directement reliées aux estimateurs proposés. La constante b représente la réduction maximale de poids que l'on est prêt à accorder dans la spécification des poids implicites r_k, et a dans ce sens un lien avec la suggestion de Lee (1991). La constante la plus importante à spécifier est possiblement la valeur de a. Nous suggérons ici $a = 9$. Cependant, dans nos simulations, une valeur de a entre 6 et 12 donne des résultats relativement comparables. Le choix des bornes L et U repose sur des considérations cosmétiques, afin d'avoir des poids restreints à un intervalle. Cette dernière considération est peut-être secondaire pour le praticien. Il semble donc que l'aspect le plus important est de choisir une valeur de r_k proche de d_k pour les bonnes valeurs, et réduite à mesure qu'une observation est extrême, et c'est ce but qui a motivé le choix de r_k de cette section. Cependant il serait utile d'avoir un choix de r_k satisfaisant un certain critère d'optimalité.

3.3 Estimateur modéliste de Chambers

Une autre approche est basée sur une décomposition proposée par Chambers (1982, 1986) que nous appliquons maintenant aux estimateurs QR. Remarquons que nous pouvons toujours écrire un estimateur QR sous la forme

$$\hat{T}_{y\mathrm{QR}} = \sum_s r_k y_k + (T_x - \hat{T}_{xr})' B + \sum_s z_k \sqrt{q_k}\,(y_k - x_k' B),$$

où $z_k = (T_x - \hat{T}_{xr})'(\sum_s q_k x_k x_k')^{-1} x_k \sqrt{q_k}$, $\hat{T}_{xr} = \sum_s r_k x_k$ et B est arbitraire. Chambers (1986) avait considéré le cas particulier $(q_k, r_k) = (1/\sigma_k^2, 1)$ pour l'estimateur par le ratio. Afin de limiter l'influence des unités aberrantes, Chambers propose

$$\hat{T}_{y\mathrm{CHAM}} = \sum_s r_k y_k + (T_x - \hat{T}_{xr})' B + \sum_s z_k \psi_p\left(\sqrt{q_k}\,(y_k - x_k' B)\right). \tag{3.12}$$

La fonction ψ_p permet de limiter l'influence des gros résidus. Le choix de B est un estimateur robuste, comme par exemple $\hat{B}_g$. Une fonction ψ_p considérée dans Chambers (1986) est

$$\psi_p(t) = t \exp(-0.25(|t| - 6)^2). \tag{3.13}$$

Il est intéressant de noter que (3.12) peut s'écrire comme

$$\hat{T}_{y\mathrm{CHAM}} = T_x' B + \sum_s (r_k + (d_k g_k - r_k)\lambda_k) e_k(B),$$

où $e_k(B) = y_k - x_k' B$, g_k est défini dans la formule (2.1) calculé à l'aide de q_k et r_k, et

$$\lambda_k = \frac{\psi_p\left(\sqrt{q_k}(y_k - x_k' B)\right)}{\sqrt{q_k}(y_k - x_k' B)}. \tag{3.14}$$

Ainsi, les résidus $e_k(B)$ sont pondérés par une relation rappelant la formule (3.2). Si $\lambda_k \equiv 1$, alors $d_k g_k$ est appliqué aux résidus $e_k(B)$ et on vérifie facilement que nous retrouvons l'estimateur $\hat{T}_{y\mathrm{QR}}$. Alternativement, si $\lambda_k \equiv 0$, nous obtenons (3.3) si nous posons $B = \hat{B}_g$. Si dans (3.12) on pose $(q_k, r_k) = (1/c_k, 1)$ et $B = \hat{B}_g$, alors l'estimateur de Chambers représente un compromis entre le BLUP et un estimateur robuste basé sur un GM-estimateur. Remarquons que (3.12) est formellement un estimateur QR avec

$$(q_k^*, r_k^*) = \left(d_k h_k^{1-\alpha} u_k / c_k,\ r_k + (d_k g_k - r_k)\lambda_k\right).$$

Cependant, puisque r_k^* n'est pas forcément positif, il n'est pas toujours possible de procéder à un changement de métrique dans ce cas.

4. ÉTUDE EMPIRIQUE

Afin d'étudier la performance des estimateurs de calage robustes, nous avons entrepris une étude de simulation de Monte Carlo. Nous avons considéré quatre populations composées de données provenant d'ouvrages facilement accessibles en théorie de l'échantillonnage. Pour chacune des populations, K = 2 000 échantillons ont été tirés par tirage aléatoire simple pour différentes tailles échantillonnales. Notre objectif principal est de savoir s'il est possible d'obtenir des estimateurs possédant de bonnes propriétés empiriques (biais, erreur quadratique moyenne) mais satisfaisant aussi les CE et les RAFV. Notons que tous les programmes ont été écrit en S-PLUS (Statistical Sciences 1991) et sont disponibles en communiquant avec l'auteur.

4.1 Populations à l'étude

Les graphiques des populations se retrouvent dans la figure (4.1). La première population possède 51 unités et se retrouve dans Mosteller et Tukey (1977, p. 560). Elle consiste en la population américaine en 1960 et en 1970 pour chacun des 50 états et du district fédéral de Columbia. Nous la notons POPUSA. Du nuage de points de la population de 1970 en fonction de celle de 1960, nous remarquons que toutes les unités semblent être sur la même droite, avec la présence de quelques bons points de levier. Un exemple d'un bon point de levier est le point entouré dans cette population. La seconde population possède 34 unités et se retrouve dans Singh et Chaudhary (1986, p. 177). Il est question de superficie de champs ensemencés en 1971 et en 1974. Nous appellons cette population AREA. Un mauvais point de levier (voir le point entouré) est présent dans cette population puisque le point (4170,99) ne respecte pas la tendance linéaire de la majorité des unités. Des échantillons de taille 10 et 15 sont tirés dans POPUSA et AREA. La troisième population est la population MU284 de Särndal et coll. (1992) qui porte sur les 284 municipalités de Suède. Nous considérons les variables x = *S82* concernant le nombre total de sièges dans le conseil municipal et y = *P85* la population suédoise en 1985. On note la présence d'aberrants verticaux (voir par exemple le point entouré) et d'un mauvais point de levier. Finalement, nous considérons la population MU281 consistant de MU284 de laquelle les trois plus grosses municipalités ont été exclues. Les variables considérées sont x = REV84 qui représente les valeurs des biens fonciers suivant l'évaluation de 1984 et y = RMT85 qui représente les revenus des taxes municipales en 1985. L'unité de mesure est le million de kronors pour ces deux variables. Plusieurs mauvais points de levier semblent présents dans cette population. Des échantillons de taille n = 30 et n = 60 sont tirés dans MU284 et MU281. Le tableau (4.1) contient les totaux des diverses populations.

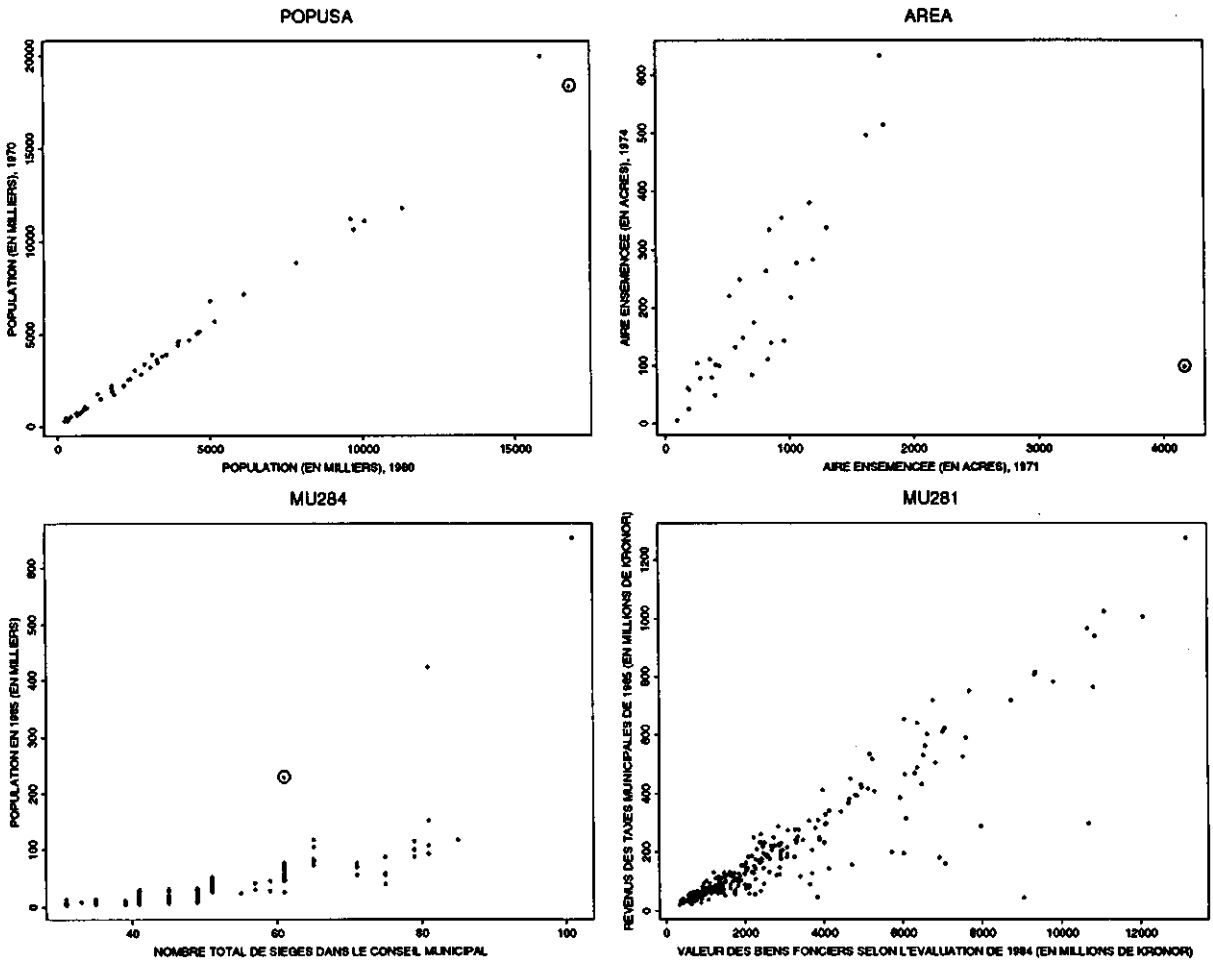


Figure 4.1. Les quatre populations à l'étude

Tableau 4.1
Totaux des diverses populations et totaux connus de l'information auxiliaire

Population	T_x	T_y	N
POPUSA	179 972	203 923	51
AREA	29 118	6 781	34
MU284	13 500	8 339	284
MU281	757 246	53 124	281

4.2 Description des estimateurs

Les deux estimateurs de base sont le GREG ainsi que l'estimateur obtenu en considérant le cas 7 de Deville et Särndal (1992) qui est un GREG avec des poids bornés. Nous notons ces estimateurs GREG/U et GREG/R respectivement. Nous avons choisi $c_k \equiv 1$ pour les populations POPUSA et AREA, alors que nous avons optés pour $c_k = x_k$ pour les populations MU284 et MU281. Notre choix des c_k est motivé par la relation existant entre ces constantes et l'hétéroscédasticité du modèle de superpopulation. Parmi les estimateurs robustes, nous avons étudié l'estimateur de Chambers (1986) en considérant

$$(q_k^*, r_k^*) = (d_k \hat{u}_k(\hat{B}_{\mathrm{CH}})/c_k,\ 1 + (d_k g_k - 1)\lambda_k(\hat{B}_r)),$$

où dans la formule (2.1) $(q_k, r_k) = (1/c_k, 1)$, noté CHAM, basé sur $\hat{B}_r$. Les constantes $\hat{u}_k(\hat{B}_{CH})$ sont obtenus de la formule (3.7). Le choix $\alpha = 1$ a été adopté tout au long de la simulation. La fonction ψ de Huber a été utilisé avec la constante $c = 1{,}345$ pour $\hat{B}_r$. Les fonctions h_k sont celles données par la formule (3.6), où nous avons choisis $t = 1{,}46$. La fonction λ_k est définie par l'équation (3.14). La fonction ψ_p considérée est celle donnée dans l'équation (3.13). L'échelle a été estimée comme dans Coakley et Hettmansperger (1993). Nous avons également considéré l'estimateur modéliste BLUP dans lequel nous avons remplacé l'estimateur par les moindres carrés généralisés par l'estimateur $\hat{B}_r$, que nous appelons MODEL. De plus, l'estimateur de Lee (1991) a été considéré basé sur $\hat{B}_r$ où $r_k = 0{,}25\, d_k$, en utilisant la métrique quadratique. Nous étudions également une extension de l'estimateur de Lee (1991) en considérant la métrique quadratique restreinte. Nous notons ces estimateurs LEE25/U et LEE25/R respectivement. Finalement, nous avons considéré la nouvelle méthode de la section 3.2.3 avec le choix (q_k, r_k) donné dans l'équation (3.11) selon la métrique quadratique et la métrique quadratique restreinte. Nous les notons QRROB/U et QRROB/R respectivement. Le choix de la fonction ψ^* est donné dans la formule (3.10).

Tableau 4.2
Résultats de Monte-Carlo lors de l'échantillonnage dans la population POPUSA

Estimateurs	VAR_M	EQM_M	CV_M	BR_M	MIN	MAX	RAFV [1]	CONV
$n = 10$								
GREG/U	34,90	34,92	2,90	-0,07	-6,24	26,75	86,7	
GREG/R	35,29	35,30	2,91	-0,04	0,20	32,00	100,0	98,4
CHAM	32,43	33,75	2,85	-0,56	-19,61	40,96	84,0	
MODEL	27,66	30,69	2,72	-0,85	-19,71	40,86	82,8	
LEE25/U	27,48	30,07	2,69	-0,79	-19,38	39,64	83,2	
LEE25/R	28,67	30,90	2,73	-0,73	0,20	32,00	100,0	98,4
QRROB/U	27,40	28,40	2,61	-0,49	-15,68	40,10	83,2	
QRROB/R	28,33	29,18	2,65	-0,45	0,20	32,00	100,0	98,4
$n = 15$								
GREG/U	21,90	21,95	2,30	-0,10	-3,13	15,32	94,7	
GREG/R	22,12	22,15	2,31	-0,09	0,20	16,00	100,0	99,5
CHAM	18,11	20,14	2,20	-0,70	-5,79	16,44	92,4	
MODEL	15,43	19,03	2,14	-0,93	-6,09	16,92	91,0	
LEE25/U	15,44	19,54	2,17	-0,99	-6,19	17,06	90,8	
LEE25/R	15,72	19,68	2,18	-0,98	0,20	16,00	100,0	99,5
QRROB/U	14,68	16,44	1,99	-0,65	-4,48	16,41	90,9	
QRROB/R	14,85	16,56	2,00	-0,64	0,20	16,00	100,0	99,5

[1] Les bornes pour les RAFV sont [0,20, 32] pour $n = 10$ et [0,20, 16] pour $n = 15$.

4.3 Mesures fréquentistes

Pour chaque échantillon, les huit estimateurs de la section (4.2) sont calculés. Les résultats sont dans les tableaux 4.2, 4.3, 4.4 et 4.5. Puisqu'un attrait des nouvelles méthodes concerne les RAFV, des statistiques sont calculées sur ces poids. Les colonnes MIN et MAX des tableaux de résultats contiennent les valeurs minimales et maximales des poids calculées lors de la simulation et ce pour chaque estimateur. Nous rapportons aussi le pourcentage d'échantillons pour lesquels les poids sont à l'intérieur des RAFV dans la colonne RAFV des tableaux de résultats. Nous avons également considéré le pourcentage d'échantillons pour lesquels les estimateurs restreints ont convergé dans la colonne CONV. Les intervalles adoptés $[L, U]$ pour les intervalles restreints sont précisés dans les divers tableaux. Dans tous les cas, les diverses statistiques ont été calculées sur les échantillons pour lesquels tous les estimateurs ont convergé.

Un autre aspect important est relié au biais et à l'efficacité des méthodes proposées. Soit $\hat{T}$ un estimateur du total T_y. Posons $\hat{T}_i$ l'estimateur du total calculé à l'aide de l'échantillon $i, i = 1, ..., K$. Le biais relatif de Monte Carlo BR_M, la valeur moyenne E_M et la variance V_M sont donnés par les formules usuelles, à savoir

$$BR_M = \left(E_M(\hat{T}) - T_y\right)/T_y \times 100, E_M(\hat{T}) = \frac{1}{K}\sum_{i=1}^{K}\hat{T}_i \text{ et } V_M = \frac{1}{K}\sum_{i=1}^{K}\left(\hat{T}_i - E_M(\hat{T})\right)^2.$$

Notre critère principal d'efficacité sera l'erreur quadratique moyenne de Monte Carlo, définie par $EQM_M = 1/K\sum_{i=1}^{K}(\hat{T}_i - T_y)^2$. Les coefficients de variation CV_M sont calculés selon $\sqrt{EQM_M}/T_y$. Les variances et erreurs quadratiques moyennes sont exprimées en millions. Les coefficients de variation, les biais relatifs, les RAFV et la convergence des versions restreintes sont exprimés en pourcentage.

4.4 Discussion

La population POPUSA ne comportait pas de valeurs aberrantes ne respectant pas le modèle linéaire. Lors de l'échantillonnage, les coefficients de variation des estimateurs étaient petits, ce qui était prévisible compte tenu de l'allure de la population. Les colonnes EQM et VAR sont très semblables, confirmant que le biais n'est pas un problème dans cette population. Tous les biais relatifs sont inférieurs à 1%. L'estimateur QRROB/U offre une réduction de variance par rapport à GREG/U de plus de 21% lorsque $n = 10$, et de plus de 30% lorsque $n = 15$.

La population AREA est de petite taille. Elle comportait un mauvais point de levier qui a occasionné des biais relatifs empiriques très élevés pour tous les estimateurs. L'estimateur GREG/U affiche un biais relatif de plus de 7% malgré un échantillonnage de 44% dans cette population. Les estimateurs robustes affichent les biais les plus importants, mais relativement comparables au biais de l'estimateur GREG/U. La réduction de variance la plus importante a été enregistrée par l'estimateur QRROB/U, mais au prix d'un biais relatif autour de 10%.

La population MU284 contient un aberrant vertical et des mauvais points de levier. Les estimateurs robustes réduisent de manière radicale la variance, puisqu'ils ne sont pas affectés par les trois unités extrêmes en y qui s'éloignent de façon marquée de la tendance linéaire. Les

estimateurs CHAM, QRROB/R et QRROB/U sont plus que quatre fois moins variables que l'estimateur GREG/U. Cependant cela occasionne un biais négatif qui est beaucoup plus élevé. Tous les estimateurs robustes sont sévérement biaisés. L'estimateur MODEL affiche un biais négatif de plus de 13% alors que QRROB/U possède un biais négatif de l'ordre de 11%. Dans le cas de QRROB, un meilleur choix des constantes dans la fonction ψ^* pourrait réduire une plus grande part du biais au prix d'une moins grande réduction de variance. L'augmentation de la taille échantillonnale à $n = 60$ a permis de réduire le biais sous les 10% pour les estimateurs CHAM et QRROB, mais les autres estimateurs robustes restent plus biaisés.

La population MU281 contient un nombre assez grand de mauvais points de levier. La variance domine la part du EQM dans cette population. L'estimateur LEE25 est le moins variable, avec une réduction de plus de 35% comparativement à GREG/U. Cependant, bien que $\theta = 0{,}25$ fonctionne bien dans cette population, notre étude illustre bien que ce n'est pas toujours le meilleur choix.

On remarque que tous les estimateurs robustes sont plus efficaces que le GREG ou sa version restreinte. Comme confirmé par les résultats de Deville et Särndal (1992), la version restreinte du GREG se comporte essentiellement de la même manière que le GREG, tant au niveau du biais que de la variance Monte Carlo dans chacune des populations. Parmi tous les estimateurs considérés, GREG/U et GREG/R sont les estimateurs les moins biaisés. Les versions robustes présentent toutes un biais plus important. Cependant la réduction de variance compense plus que largement de telle sorte que l'efficacité des estimateurs robustes est toujours plus grande que GREG/U ou GREG/R.

Concernant les restrictions sur les poids, on remarque que les estimateurs GREG/U, CHAM, MODEL, LEE25 et QRROB/U sont tous sujets aux problèmes de pondération négative comme en fait foi la colonne MIN. Les estimateurs restreints évitent ce problème. La colonne RAFV montre que les restrictions ne furent pas satisfaites relativement fréquemment, dépendant de la population et de la taille de l'échantillon, allant de 5% à 60%. Les deux estimateurs robustes restreints se comportent en général de façon comparable à leur version non-restreinte. De plus, QRROB/R, en plus de satisfaire les RAFV offre des propriétés d'efficacité intéressantes, lorsque comparé aux autres estimateurs robustes. Les versions restreintes ont été moins sujettes aux problèmes de convergence lorsque les tailles échantillonnales étaient plus grandes. Notons que nous avons dû prendre des bandes plus larges dans le cas de POPUSA afin d'obtenir des taux de convergence satisfaisants.

5. CONCLUSION

Le but de cet article est d'introduire des estimateurs de calage possédant de bonnes propriétés de robustesse. Les estimateurs traditionnels de calage sont faciles d'emploi, puisqu'il suffit d'un ensemble de poids de départ, habituellement les poids d'échantillonnage d_k, qui sont transformés en poids calés. La démarche de cet article suit ces mêmes étapes, où les poids implicites robustes r_k sont transformés en poids calés, et où les constantes q_k sont choisies de sorte que $\hat{B}_q$ soit un estimateur robuste. Le choix de r_k proposé est donné par la formule (3.9), avec $a = 9$, $b = 1/4$. Une théorie pour un choix de r_k optimal reste à faire. Il est suggéré dans les applications de varier la constante a, entre

Tableau 4.3
Résultats de Monte-Carlo lors de l'échantillonnage dans la population AREA

Estimateurs	VAR_M	EQM_M	CV_M	BR_M	MIN	MAX	RAFV [1]	CONV
$n = 10$								
GREG/U	1,334	1,700	19,23	8,92	-3,35	14,94	86,6	
GREG/R	1,295	1,629	18,82	8,53	0,20	14,00	100,0	99,0
CHAM	1,187	1,541	18,30	8,77	-4,09	14,90	87,2	
MODEL	1,291	1,580	18,54	7,93	-5,23	16,75	86,8	
LEE25/U	1,279	1,593	18,61	8,26	-5,28	16,89	86,6	
LEE25/R	1,284	1,596	18,63	8,24	0,20	14,00	100,0	99,0
QRROB/U	1,026	1,440	17,70	9,50	-4,74	15,38	87,6	
QRROB/R	1,028	1,437	17,68	9,43	0,20	14,00	100,0	99,0
$n = 15$								
GREG/U	0,940	1,178	16,00	7,18	-1,40	7,03	93,0	
GREG/R	0,928	1,154	15,85	7,01	0,20	6,00	100,0	99,8
CHAM	0,708	0,989	14,67	7,82	-1,52	7,92	93,7	
MODEL	0,757	0,997	14,73	7,22	-1,66	8,39	93,1	
LEE25/U	0,672	1,059	15,18	9,18	-1,68	9,40	92,0	
LEE25/R	0,671	1,056	15,15	9,15	0,20	6,00	100,0	99,8
QRROB/U	0,485	0,990	14,68	10,48	-1,59	8,90	93,9	
QRROB/R	0,485	0,986	14,64	10,44	0,20	6,00	100,0	99,8

[1] Les bornes pour les RAFV sont [0,20, 14] pour $n = 10$ et [0,20, 6] pour $n = 15$.

Tableau 4.4
Résultats de Monte-Carlo lors de l'échantillonnage dans la population MU284

Estimateurs	VAR_M	EQM_M	CV_M	BR_M	MIN	MAX	RAFV [1]	CONV
$n = 30$								
GREG/U	2,833	2,925	20,51	-3,64	-6,83	23,90	89,8	
GREG/R	2,813	2,910	20,46	-3,73	0,20	16,00	100,0	99,2
CHAM	0,645	1,639	15,35	-11,95	-11,80	31,26	77,0	
MODEL	0,709	2,037	17,11	-13,82	-12,06	31,91	68,40	
LEE25/U	0,887	1,877	16,43	-11,93	-11,06	30,93	73,5	
LEE25/R	0,871	1,847	16,30	-11,85	0,20	26,00	100,0	99,2
QRROB/U	0,719	1,532	14,84	-10,81	-9,46	25,84	86,5	
QRROB/R	0,720	1,525	14,81	-10,76	0,20	16,00	100,0	99,2
$n = 60$								
GREG/U	1,473	1,489	14,63	-1,49	-1,19	10,03	90,1	
GREG/R	1,467	1,484	14,61	-1,57	0,20	7,00	100.0	99,7
CHAM	0,357	0,990	11,93	-9,54	-2,53	15,59	69,8	
MODEL	0,380	1,255	13,43	-11,22	-4,93	14,52	58,1	
LEE25/U	0,403	1,201	13,14	-10,72	-4,80	14,20	60,3	
LEE25/R	0,396	1,203	13,16	-10,78	0,20	7,00	100,0	99,7
QRROB/U	0,308	0,976	11,85	-9,80	-2,36	10,99	86,1	
QRROB/R	0,308	0,979	11,87	-9,82	0,20	7,00	100,0	99,7

[1] Les bornes pour les RAFV sont [0,20, 16] pour $n = 30$ et [0,20, 7] pour $n = 60$.

Tableau 4.5
Résultats de Monte-Carlo lors de l'échantillonnage dans la population MU281

Estimateurs	VAR_M	EQM_M	CV_M	BR_M	MIN	MAX	RAFV [1]	CONV
$n = 30$								
GREG/U	17,33	17,35	7,84	-0,26	-38,97	34,56	86,0	
GREG/R	17,40	17,41	7,86	-0,24	0,20	25,00	100,0	99,8
CHAM	13,23	13,26	6,86	-0,33	-47,09	39,08	56,9	
MODEL	11,30	11,91	6,50	1,47	-66,22	41,43	47,9	
LEE25/U	11,21	11,60	6,41	1,17	-59,75	37,03	53,3	
LEE25/R	11,26	11,73	6,45	1,29	0,20	25,00	100,0	99,8
QRROB/U	12,92	13,29	6,86	1,15	-54,14	39,73	70,8	
QRROB/R	12,94	13,34	6,88	1,20	0,20	25,00	100,0	99,8
$n = 60$								
GREG/U	7,57	7,57	5,18	-0,10	-12,77	15,34	86,4	
GREG/R	7,58	7,58	5,18	-0,09	0,20	9,00	100,0	99,9
CHAM	5,85	5,90	4,57	-0,43	-22,97	11,49	51,4	
MODEL	4,53	5,23	4,30	1,57	-24,02	14,58	38,7	
LEE25/U	4,55	5,18	4,28	1,49	-23,74	14,41	41,2	
LEE25/R	4,50	5,21	4,30	1,58	0,20	9,00	100,0	99,9
QRROB/U	5,40	6,16	4,67	1,64	-21,08	21,07	68,6	
QRROB/R	5,39	6,17	4,67	1,66	0,20	9,00	100,0	99,9

[1] Les bornes pour les RAFV sont [0,20, 25] pour $n = 30$ et [0,20, 9] pour $n = 60$.

par exemple 6 et 12, pour voir l'influence de la constante sur l'estimation. Si désiré, les bornes L et U permettent de restreindre les poids, afin qu'ils soient tous positifs par exemple. Nous suggérons d'utiliser de manière générale $L = 0{,}2$, $U = kN/n$, où k est autour de 3.

Notons que le but des estimateurs de calage robustes n'est pas de remplacer le GREG, mais bien d'être utilisé conjointement avec ce dernier. Ainsi, si l'estimateur robuste et l'estimateur GREG sont très différents, alors une analyse plus poussée pourrait permettre d'expliquer le phénomène. Les estimateurs proposés sont potentiellement utiles comme outils de diagnostic.

Il serait intéressant de pousser les études empiriques de la section 4, en étudiant par exemple l'effet du plan d'échantillonnage sur les procédures proposées. L'estimation de la variance est un autre sujet important qui nécessite des développements. Une autre avenue intéressante concerne les sondages «tout usage» (multipurpose).

En effet, dans les applications, nous disposons rarement que d'une seule variable d'intérêt. Les méthodologistes désirent alors disposer d'un unique ensemble de poids utile pour toutes les variables d'intérêt. Dans un contexte de robustesse, une solution a été proposée dans la conclusion de l'article de Gwet et Rivest (1992), où des poids robustes sont calculés pour chaque variable d'intérêt $y^{(i)}$, $i = 1, ..., I$. Pour une unité, le poids final correspond au poids minimal parmi les poids obtenus. Alternativement, afin d'obtenir des estimateurs robustes et calés, nous pourrions calculer les poids implicites robustes pour chaque variable d'intérêt, fournissant un ensemble de $r_k(y^{(i)})$, et poser $r_k = \min r_k(y^{(i)})$, où le minimum est sur $i = 1, ..., I$. On pourrait par la suite transformer ces poids en poids calés. Cette procédure nécessite d'être évaluée plus en détail.

REMERCIEMENTS

Je tiens à remercier Carl-Erik Särndal pour m'avoir initié à la théorie de l'échantillonnage et pour m'avoir suggéré de considérer le problème des valeurs aberrantes en théorie des sondages. Je remercie Roch Roy et Christian Léger pour m'avoir aidé à diverses étapes dans la réalisation de cet article. Je remercie sincèrement le rédacteur associé, le rédacteur adjoint et deux arbitres anonymes pour leurs commentaires constructifs qui ont permis d'améliorer substantiellement le contenu et la présentation de cet article.

ANNEXE A
PREUVE DE LA PROPOSITION 1

Soit $\Delta_h(u;q,r) = r + qu - h(u;q,r)$ et z_k une variable d'intérêt. On suppose les conditions suivantes

C_1. $N^{-1}\sum_s q_k z_k = O_p(1)$.

C_2. $N^{-1}\sum_s \Delta_h(x_k'\lambda_\infty; q_k, r_k)z_k = o_p(n^{-1/2})$,

où λ_∞ est solution de l'équation (2.5).

On remarque que $\sum_s(r_k + q_k x_k'\lambda_1)x_k = T_x$, où $\lambda_1 = -(\sum_s q_k x_k x_k')^{-1}(\hat{T}_{xr} - T_x)$, et aussi que $\sum_s h(x_k'\lambda_\infty; q_k, r_k)x_k = T_x$. Ainsi, on trouve en utilisant C_2 que

$$N^{-1}\sum_s q_k x_k x_k'(\lambda_1 - \lambda_\infty) = o_p(n^{-1/2}),$$

et donc $\lambda_1 - \lambda_\infty = o_p(n^{-1/2})$, en utilisant C_1 avec $z_k = x_k x_k'$. On montre aussi aisément que

$$\begin{aligned} N^{-1}&\left(\hat{T}_{\mathrm{QR}} - \hat{T}_{\mathrm{RQR}}\right) \\ &= N^{-1}\sum_s (r_k + q_k x_k'\lambda_1)y_k - N^{-1}\sum_s h(x_k'\lambda_\infty; q_k, r_k)y_k \\ &= N^{-1}\sum_s q_k x_k' y_k(\lambda_1 - \lambda_\infty) + N^{-1}\sum_s \Delta_h(x_k'\lambda_\infty; q_k, r_k)y_k \\ &= o_p(n^{-1/2}). \end{aligned}$$

ANNEXE B
LISTE DES ABRÉVIATIONS

ADU: Absence de biais asymptotique (Asymptotically Design Unbiased).

BLUP: Meilleur estimateur linéaire sans biais de Royall (1970).

CE: Contraintes d'étalonnage, $\sum_s w_k x_k = T_x$, où $T_x = \sum_U x_k$.

CH: Estimateur robuste proposé par Coakley et Hettmansperger (1993) qui un GM-estimateur en une étape. Il est robuste et efficace.

CHAM: Estimateur robuste de Chambers (1982, 1986).

GM: Estimateurs *M* généralisés, issus de la théorie de la robustesse. Voir par exemple Hampel et coll. (1986).

GREG: Estimateur par la régression généralisée proposé par Cassel et coll. (1976).

HT: Estimateur Horvitz-Thompson $\sum_s d_k y_k$, où $d_k = \pi_k^{-1}$.

QR: Estimateurs de Wright (1983), de la forme $T_x'\hat{B}_q + \sum_s r_k e_k$.

RAFV: Restrictions applicables à la fourchette des valeurs pour les poids w_k, en imposant par exemple que tous les $w_k \in [L, U]$.

RQR: Généralisation des estimateurs de Wright (1983), qui s'obtiennent d'une métrique générale en plus de restrictions sur les poids.

ANNEXE C
LISTE DES PRINCIPALES CONSTANTES

c_k: Facteur pouvant tenir compte de problèmes d'hétéroscédasticité.

d_k: Poids d'échantillonnage.

g_k: Poids-g défini par w_k/d_k.

h_k: Quantité permettant de réduire l'influence d'une information auxiliaire aberrante dans $\hat{B}_q$.

π_k, π_{kl}: Probabilités d'inclusion d'ordre un et deux, respectivement.

q_k, r_k: Quantités définissant un estimateur QR. Les q_k servent à construire les coefficients de régression intervenant dans la premiére partie, $T_x'\hat{B}_q$. Les r_k servent pour la seconde partie, $\sum_s r_k e_k$.

$u_k, \hat{u}_k$: Poids servant à construire $\hat{B}_q$ de maniére robuste.

u_k^*: Poids servant à considérer un facteur de correction $\sum_s r_k e_k$ robuste.

w_k: Poids calé à être apposé à y_k pour former $\sum_s w_k y_k$.

BIBLIOGRAPHIE

BERSHAD, M.A. (1960). Some Observations on Outliers. Note de service non-publiée, Statistical Research Division, U.S. Bureau of the Census.

BOLFARINE, H., et ZACKS, S. (1992). *Prediction Theory for Finite Population*. New York: Springer-Verlag.

BREWER, K.R.W. (1994). Survey sampling inference: Some past perspectives and present prospects. *Pakistan Journal of Statistics* 10, 213-233.

BRUCE, A.G. (1991). Robust Estimation and Diagnostics for Repeated Sample Surveys. Mathematical Statistics Working Paper 1991/1, Statistics New Zealand.

CASSEL, C.M., SÄRNDAL, C.-E., et WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika* 63, 615-620.

CHAMBERS, R.L. (1982). Robust Finite Population Estimation. Thèse de doctorat, Johns Hopkins University, Dept of Biostatistics.

CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

CHAMBERS, R.L., et KOKIC, P.N. (1993). An integrated approach for the treatment of outliers in sub-annual surveys. *Actes de la 49ième Session, Institut International de Statistique.*

COAKLEY, C.W., et HETTMANSPERGER, T.P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association* 88, 872-880.

DALÉN, J. (1987). Practical Estimators of a Population Total Which Reduce the Impact of Large Observations. R & D Report, Statistics Sweden.

DEVILLE, J.-C., et SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DONOHO, D.L., et HUBER, P.J. (1983). The notion of breakdown point. Dans *A Festschrift for Erich Lehmann* (Éds. P.J. Bickel, K.A. Doksum et J.L. Hodges). Belmont, CA: Wadsworth.

FULLER, W.A., LOUGHIN, M.M., et BAKER, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

GAMBINO, J. (1987). Dealing with outliers: A look at some methods used at Statistics Canada. Rapport technique, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

GROSS, W.F., BODE, G., TAYLOR, J.M., et LLOYD-SMITH, C.W. (1986). Some finite population estimators which reduce the contribution of outliers. *Proceedings of the Pacific Statistical Congress*. Aucklaud, New Zealand, 20-24 Mai 1985.

GWET, J.-P., et RIVEST, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., et STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

HIDIROGLOU, M.A., et SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.

HUBER, P.J. (1981). *Robust Statistics*. New York: Wiley.

HULLIGER, B. (1995). Estimateurs Horvitz-Thompson à l'épreuve des valeurs aberrantes. *Techniques d'enquête*, 21, 89-98.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

LEE, H. (1991). Model-based estimators that are robust to outliers. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of Census.

LEE, H. (1995). Outliers in business surveys. Dans *Business Survey Methods*, (Éds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott). New York: Wiley.

LEE, H., GHANGURDE, P.D., MACH, L., et YUNG, W. (1992). Outliers in sample surveys. Direction de la méthodologie document de travail, DMEE-92-008E, Statistique Canada.

MOSTELLER, F., et TUKEY, J.W. (1977). *Data Analysis and Regression, A Second Course in Statistics*. Redding MA: Addison-Wesley.

RIVEST, L.P., et ROUILLARD, E. (1991). Estimateurs-M et estimateurs à l'épreuve des valeurs aberrantes en remplacement de l'estimateur par quotient. *Recueil: Symposium 90, Mesure et amélioration de la qualité des données, Statistique Canada*, 271-285.

ROUSSEEUW, P.J., et LEROY, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

ROYALL, R.M. (1970). On finite population sampling under certain linear regression models. *Biometrika* 57, 377-387.

SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SIMPSON, D.G., et CHANG, Y.-C.I. (1997). Reweighted approximate GM-estimators: asymptotics and residual-based graphics. *Journal of Statistical Planning and Inference*, 57, 273-293.

SINGH, A.C., et MOHL, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.

SINGH, D., et CHAUDHARY, F.S. (1986). *Theory and Analysis of Sample Survey Designs*. New York: Wiley.

STATISTICAL SCIENCES, INC. (1991). *S-PLUS Reference Manual*. Seattle: Statistical Science, Inc.

STUKEL, D.M., HIDIROGLOU, M.A., et SÄRNDAL, C.-E. (1996). Estimation de la variance des estimateurs de calage: comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Technique d'enquêtes*, 22, 117-126.

TAMBAY, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings on the Section on Survey Research Methods, American Statistical Association*, 229-234.

WELSH, A.H., et RONCHETTI, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Séries B*, 60, 413-428.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 61-71
Statistique Canada

Estimation dans des enquêtes par sondage avec des probabilités d'inclusion conditionnelles: enquêtes à plan d'échantillonnage complexe

YVES TILLÉ[1]

RÉSUMÉ

L'auteur étudie une méthode basée sur le plan de sondage permettant d'utiliser une information auxiliaire en vue d'améliorer la précision des estimateurs. L'objectif est de construire un estimateur dont le biais conditionnel est faible grâce à la pondération des valeurs observées par l'inverse des probabilités d'inclusion conditionnelles. L'auteur propose une approximation générale dans les cas où la statistique auxiliaire est un vecteur d'estimateurs de Horvitz-Thompson. Cette approximation est assez proche de l'estimateur optimal décrit par Fuller et Isaki (1981), Montanari (1987, 1997), Deville (1992) et Rao (1994, 1997). Puis, il applique l'estimateur optimal à un plan d'échantillonnage stratifié et montre qu'on peut considérer cet estimateur optimal comme un estimateur de régression généralisé pour lequel les variables indicatrices de la stratification sont également utilisées au stade de l'estimation. Enfin, il discute du domaine d'application de cet estimateur dans le contexte général de l'utilisation de données auxiliaires.

MOTS CLÉS: Estimation conditionnelle; observation pondérée; estimateur de régression généralisé; enquête complexe.

1. INTRODUCTION

Au stade de l'estimation, les praticiens de l'échantionnage d'enquête disposent souvent de données auxiliaires. Il peut s'agir d'un ensemble de moyennes ou de totaux de population. Parfois, les données sont détaillées, par exemple, quand on connaît les valeurs prises par une variable pour les diverses unités d'une population. On peut se servir de ces renseignements pour améliorer la précision des estimateurs.

Notre but est d'examiner le cas des données auxiliaires dans le contexte de conditionnalité. De nombreux articles ont été publiés au sujet de l'inférence conditionnelle. En effet, la question de l'estimateur optimal a été abordée par Fuller et Isaki (1981), Montanari (1987, 1997), Deville (1992) et Rao (1994, 1997). Casady et Valliant (1993), quant à eux, ont étudié les propriétés conditionnelles des estimateurs dans le cas de la stratification a posteriori. Dans un article antérieur (Tillé 1998), nous avons proposé une méthode générale qui permet de construire un estimateur de la moyenne ou du total dont le biais conditionnel est faible dans le cas de l'échantionnage aléatoire simple. Cette méthode se fonde sur l'utilisation des probabilités d'inclusion conditionnelles et permet de tenir compte de données auxiliaires sans devoir s'appuyer sur un modèle de superpopulation.

Dans le présent article, nous généralisons l'utilisation des probabilités d'inclusion conditionnelles à tout plan d'échantillonnage. Nous montrons que cette méthode permet de construire un estimateur fort semblable à l'estimateur optimal décrit par Montanari (1987), Deville (1992) et Rao (1994). Cette famille d'estimateurs fournit une inférence conditionnelle valide et peut aussi être considérée comme l'estimateur linéaire optimal. Puis, nous appliquons ces estimateurs au cas de la stratification et nous les comparons à l'estimateur de régression généralisé (GREG en anglais). L'estimateur GREG est généralement entaché d'un biais conditionnel. Néanmoins, nous montrons que, pour ce qui est de la régression, l'estimateur optimal est un cas particulier de l'estimateur GREG. En effet, quand on utilise de nouveau les variables de stratification comme variables auxiliaires dans l'estimateur GREG, ce dernier est égal à l'estimateur optimal. Ensuite, nous présentons un ensemble de simulations pour illustrer les avantages de l'estimateur optimal dans le cas de la stratification. Le gain de précision peut être très important quand les variables de stratification sont fortement corrélées à la variable étudiée. Enfin, nous discutons du problème général de l'estimation dans le cas d'enquête par sondage et le considérons comme un problème de troisième ordre, où trois ensembles de variables interagissent: les variables de planification, les variables d'étalonnage et les variables étudiées.

L'article est structuré comme suit. À la section 2, nous définissons la notation. À la section 3, nous présentons le problème de l'inférence conditionnelle. À la section 4, nous donnons une approximation de l'estimateur simple pondéré conditionnellement (estimateur SPC) pour les plans d'échantionnage complexes compte tenu d'hypothèses techniques. Nous discutons de ces hypothèses à la section 5. À la section 6, nous comparons l'estimateur optimal et l'estimateur SPC à l'estimateur de régression généralisé GREG dans le contexte de la stratification. Nous montrons que l'estimateur optimal peut être considéré

[1] Yves Tillé, Crest – Ensai, École Nationale de la Statistique et de l'Analyse de l'Information, rue Blaise Pascal, Campus de Ker Lann, 35170 Bruz, France, courrier electronique: tille@ensai.fr.

comme un estimateur GREG pour lequel les variables indicatrices de la stratification sont également utilisées a posteriori. Puis, à la section 7, nous présentons un ensemble de simulations réalisées pour comparer les estimateurs susmentionnés. Enfin, à la section 8, nous examinons le problème de l'interaction entre le plan de sondage et les variables auxiliaires.

2. PROBLÈME ET NOTATION

Considérons une population finie $U = \{1,...,k,...,N\}$ et supposons qu'on sélectionne sans remise à partir de cette population un échantillon aléatoire S conformément à un plan de sondage $p(.)$. La probabilité de sélectionner l'échantillon s est $\Pr(S = s) = p(s)$, pour tout $s \subset U$. Les variables indicatrices I_k prennent la valeur 1 si l'unité k est dans l'échantillon et 0 sinon, pour tout $k \in U$. La probabilité d'inclusion de l'unité k est $\pi_k = E(I_k)$, où le symbole $E(.)$ représente l'espérance mathématique par rapport avec le plan de sondage. La probabilité d'inclusion d'ordre deux des unités k et l est $\pi_{kl} = E(I_k I_l)$. Notons par y_k la valeur de la variable y pour la k-ième unité de la population. Le but est d'estimer la moyenne de population de y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k.$$

Si $\pi_k > 0$, pour tout $k \in U$, l'estimateur d'Horvitz-Thompson (1952) donné par

$$\hat{\bar{y}} = \frac{1}{N}\sum_{k\in S} \frac{y_k}{\pi_k}$$

fournit un estimateur sans biais de $\bar{y}$.

Supposons que T est une statistique. L'objectif est d'estimer $\bar{y}$ avec un biais conditionnel aussi faible que possible par rapport T. Définissons la probabilité d'inclusion conditionnelle de premier ordre comme étant $\pi_{k|T} = E(I_k|T)$ pour tout $k \in U$ et les probabilités d'inclusion conditionnelles d'ordre deux comme étant $\pi_{kl|T} = E(I_k I_l|T)$ pour tout $k \in U, l \in U, k \neq l$. Nous définissons l'estimateur simple pondéré conditionnellement (SPC) comme étant

$$\hat{\bar{y}}_{|T} = \frac{1}{N}\sum_{k\in S} \frac{y_k}{\pi_{k|T}}. \tag{2.1}$$

Cet estimateur n'est pas parfaitement dépourvu de biais conditionnel. En effet, un estimateur sans biais conditionnel existe si, et uniquement si, $\pi_{k|T} > 0$ pour tout $k \in U$. Pour cette raison, il est utile d'élargir la définition de l'absence de biais conditionnel: nous disons qu'un estimateur est virtuellement sans biais conditionnel (VSBC) si le biais conditionnel dépend uniquement des unités dont la probabilité d'inclusion conditionnelle est nulle. L'estimateur SPC est effectivement VSBC puisque:

$$B(\hat{\bar{y}}_{|T}|T) = E(\hat{\bar{y}}_{|T}|T) - \bar{y} = -\frac{1}{N}\sum_{k|\pi_{k|T}=0} y_k.$$

Cet estimateur généralise certains résultats classiques (voir Tillé 1998) comme la stratification a posteriori. Qui plus est, il nous permet de construire un estimateur original pour un tableau de contingence quand on connaît les totaux de population marginaux. Malheureusement, le calcul des $\pi_{k|T}$ devient difficile dans le cas des plans d'échantillonnage complexes. Nous donnerons néanmoins une approximation générale de l'estimateur SPC quand on se sert d'un vecteur d'estimateurs d'Horvitz-Thompson comme statistique auxiliaire.

3. UTILISATION D'UNE STATISTIQUE AUXILIAIRE COMPLEXE

Supposons que les données auxiliaires sont représentées par le vecteur $\mathbf{x}_k = (x_{k1},...,x_{kj},...,x_{kJ})'$ des valeurs prises par les J variables auxiliaires pour la k-ième unité de U. Pour commencer, nous supposons que l'on connaît les $\mathbf{x}_k$ pour chaque unité de la population. Plus tard, nous considérerons le cas plus restreint où l'on ne connaît qu'une fonction des $\mathbf{x}_k$, comme

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{k\in U} \mathbf{x}_k.$$

Considérons aussi l'estimateur d'Horvitz-Thompson de $\bar{\mathbf{x}}$ représenté par

$$\hat{\bar{\mathbf{x}}}_\pi = \frac{1}{N}\sum_{k\in S} \frac{\mathbf{x}_k}{\pi_k}.$$

Si $\pi_k > 0$, pour tout $k \in U$, $\hat{\bar{\mathbf{x}}}_\pi$ est un estimateur non biaisé de $\bar{\mathbf{x}}$, tel que

$$E\left(\hat{\bar{\mathbf{x}}}_\pi\right) = \bar{\mathbf{x}}. \tag{3.2}$$

La variance de $\hat{\bar{\mathbf{x}}}_\pi$ prend la forme

$$\boldsymbol{\Sigma} = \mathrm{Var}\left(\hat{\bar{\mathbf{x}}}_\pi\right) = \frac{1}{N^2}\sum_{l\in U} \frac{\mathbf{x}_l \mathbf{x}_l'}{\pi_l}(1-\pi_l) + \frac{1}{N^2}\sum_{l\in U}\sum_{\substack{m\in U\\ m\neq l}} \frac{\mathbf{x}_l \mathbf{x}_m'}{\pi_l \pi_m}(\pi_{lm} - \pi_l \pi_m). \tag{3.3}$$

Supposons maintenant que le vecteur $(\hat{\bar{y}}_\pi \ \hat{\bar{\mathbf{x}}}_\pi')'$ suit une loi de distribution multinormale. Dans ces conditions, nous pouvons construire un estimateur conditionnel non biaisé (voir, par exemple, Deville 1992). Nous calculons d'abord le biais conditionnel, soit:

$$B\left(\hat{\bar{y}}_\pi|\hat{\bar{\mathbf{x}}}_\pi\right) = E\left(\hat{\bar{y}}_\pi|\hat{\bar{\mathbf{x}}}_\pi\right) - \bar{y} = \left(\hat{\bar{\mathbf{x}}}_\pi - \bar{\mathbf{x}}\right)\mathrm{Var}\left(\hat{\bar{\mathbf{x}}}_\pi\right)^{-1}\mathrm{Cov}\left(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi\right).$$

Si nous disposons d'un estimateur de $B(\hat{\bar{y}}_\pi|\hat{\bar{\mathbf{x}}}_\pi)$, nous pouvons corriger l'estimateur d'Horvitz-Thompson de la façon suivante:

$$\hat{\bar{y}}_c = \hat{\bar{y}}_\pi - \hat{B}\left(\hat{\bar{y}}_\pi | \hat{\bar{\mathbf{x}}}_\pi\right)$$
$$= \hat{\bar{y}}_\pi + (\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}}_\pi)\,\widehat{\mathrm{Var}}\left(\hat{\bar{\mathbf{x}}}_\pi\right)^{-1}\widehat{\mathrm{Cov}}\left(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi\right).$$

Cet estimateur est apparenté à l'estimateur linéaire optimal décrit par Fuller et Isaki (1981), Montanari (1987) et Rao (1994). En effet, Montanari a montré que le meilleur estimateur, au sens de la plus petite erreur quadratique moyenne (EQM), ayant la forme

$$\hat{\bar{y}}_\beta = \hat{\bar{y}}_\pi + \left(\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}}_\pi\right)'\beta \qquad (3.4)$$

s'observe quand β prend la valeur:

$$\beta_{OPT} = \Sigma^{-1}\mathrm{Cov}\left(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi\right).$$

L'estimateur linéaire optimal présenté par Montanari mène donc à un résultat fort semblable à celui de la méthode conditionnelle, même si, au départ, Montanari n'examine pas le problème sous l'angle conditionnel. Selon la méthode de Montanari, on trouve l'estimateur optimal dans une classe d'estimateurs linéaires définie par (3.4) sans aucune référence aux propriétés conditionnelles. Néanmoins, Rao (1994) a fait remarquer que cet estimateur aboutit à une inférence conditionnelle valide. Le problème général de l'estimateur optimal tient au fait qu'on ne connaît pas β_{OPT} et qu'on doit donc l'estimer. Or, les propriétés optimales de l'estimateur se perdent si on estime β_{OPT}.

Afin d'estimer β_{OPT} (ou $B(\hat{\bar{y}}_\pi|\hat{\bar{\mathbf{x}}}_\pi)$), on peut distinguer deux situations. Dans la première, on connaît les valeurs que prennent les variables auxiliaires pour toute les unités de la population. Dans ce cas, on connaît $\sum$ et on peut estimer

$$\mathrm{Cov}\left(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi\right) = \frac{1}{N^2}\sum_{k\in U}\frac{\mathbf{x}_k y_k}{\pi_k}(1-\pi_k) + \frac{1}{N^2}\sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}}\frac{\mathbf{x}_k y_l}{\pi_k\pi_l}(\pi_{kl}-\pi_k\pi_l)$$
$$= \frac{1}{N}\sum_{k\in U}(\bar{\mathbf{x}}_{|k} - \bar{\mathbf{x}})\,y_k$$

sans biais en se servant de

$$\widehat{\mathrm{Cov}}_1\left(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi\right) = \frac{1}{N}\sum_{k\in S}\frac{(\bar{\mathbf{x}}_{|k} - \bar{\mathbf{x}})y_k}{\pi_k} \qquad (3.5)$$

où

$$\bar{\mathbf{x}}_{|k} = E\left(\hat{\bar{\mathbf{x}}}_\pi | k\in S\right) = \frac{1}{N}\sum_{\substack{l\in U\\ l\neq k}}\frac{\mathbf{x}_l\pi_{kl}}{\pi_k\pi_l} + \frac{\mathbf{x}_k}{\pi_k N}. \qquad (3.6)$$

Si l'on se sert de (3.5), on peut construire un premier estimateur asymptotiquement optimal de la forme

$$\hat{\bar{y}}_{AOPT1} = \hat{\bar{y}}_\pi + \left(\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}}_\pi\right)'\Sigma^{-1}\frac{1}{N}\sum_{k\in S}\frac{\bar{\mathbf{x}}_{|k} - \bar{\mathbf{x}}}{\pi_k}y_k. \qquad (3.7)$$

Dans la deuxième situation, on ne connaît que la moyenne de population $\bar{\mathbf{x}}$, on doit estimer Σ et on ne peut estimer $\mathrm{Cov}(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi)$ au moyen de (3.5). Montanari propose qu'on estime Σ et $\mathrm{Cov}(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi)$ à l'aide de l'estimateur classique d'Horvitz-Thompson:

$$\hat{\Sigma} = \frac{1}{N^2}\sum_{k\in S}\frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k^2}(1-\pi_k) + \frac{1}{N^2}\sum_{k\in S}\sum_{\substack{l\in S\\ l\neq k}}\frac{\mathbf{x}_k\mathbf{x}_l'}{\pi_k\pi_l}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}$$

et

$$\widehat{\mathrm{Cov}}_2\left(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi\right) = \frac{1}{N^2}\sum_{k\in S}\frac{\mathbf{x}_k y_k}{\pi_k^2}(1-\pi_k) + \frac{1}{N^2}\sum_{k\in S}\sum_{\substack{l\in S\\ l\neq k}}\frac{\mathbf{x}_k y_l}{\pi_k\pi_l}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}.$$

On estime donc β_{OPT} (en supposant que $\hat{\Sigma}$ n'est pas singulière) par:

$$\hat{\beta}_{OPT} = \hat{\Sigma}^{-1}\,\mathrm{Cov}_2\left(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi\right).$$

En estimant β_{OPT}, nous pouvons construire un autre estimateur asymptotiquement optimal (AOPT) de la forme:

$$\hat{\bar{y}}_{AOPT2} = \hat{\bar{y}}_\pi + \left(\hat{\bar{\mathbf{x}}}_\pi - \bar{\mathbf{x}}\right)'\hat{\beta}_{OPT}. \qquad (3.8)$$

Les estimateurs AOPT1 et AOPT2 se distinguent par la façon selon laquelle nous estimons $\mathrm{Cov}(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi)$ et Σ. Cependant, l'estimateur AOPT1 exige des données auxiliaires plus complètes.

L'estimateur de régression généralisé (GREG) défini par Cassel, Särndal et Wretman (1976), Wright (1983) et Särndal, Swensson et Wretman (1992, p. 225) est également un estimateur de la classe linéaire représentée par l'expression (3.4). Dans le cas de l'estimateur GREG, β est en effet défini par

$$\beta_{GREG} = \left[\sum_{k\in U}\frac{\mathbf{x}_k\mathbf{x}_k'}{c_k}\right]^{-1}\sum_{k\in U}\frac{\mathbf{x}_k y_k}{c_k}$$

et peut être estimé par

$$\hat{\beta}_{GREG} = \left[\sum_{k\in S}\frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k c_k}\right]^{-1}\sum_{k\in S}\frac{\mathbf{x}_k y_k}{\pi_k c_k},$$

où les quantités $c_k > 0, k \in U$, sont les poids définis pour toutes les unités de la population. L'estimateur GREG n'a pas de bonnes propriétés conditionnelles et est généralement entaché d'un biais conditionnel (Rao 1994).

4. APPROXIMATION DE L'ESTIMATEUR SPC

Un autre moyen de construire un estimateur sans biais conditionnel consiste à trouver une approximation de l'estimateur SPC représenté par l'expression (2.1). En effet, cet estimateur a de bonnes propriétés puisqu'il est virtuellement sans biais conditionnel (VSPC). Si nous nous servons de $\hat{\hat{\mathbf{x}}}_\pi$ comme statistique auxiliaire, nous rechercherons une approximation de

$$\pi_{k|T} = E\left(I_k \mid \hat{\hat{\mathbf{x}}}_\pi\right).$$

Si le vecteur aléatoire $\hat{\hat{\mathbf{x}}}_\pi$ prend, par exemple, la valeur z, nous obtenons, en appliquant le théorème de Bayes,

$$E\left(I_k \mid \hat{\hat{\mathbf{x}}}_\pi = \mathbf{z}\right) = \Pr\left(k \in \mid \hat{\hat{\mathbf{x}}}_\pi = \mathbf{z}\right) = \pi_k \frac{\Pr\left(\hat{\hat{\mathbf{x}}}_\pi = \mathbf{z} \mid k \in S\right)}{\Pr\left(\hat{\hat{\mathbf{x}}}_\pi = \mathbf{z}\right)}.$$

Afin de calculer les probabilités d'inclusion conditionnelles, il est donc nécessaire de connaître la loi de distribution de $\hat{\hat{\mathbf{x}}}_\pi$ conditionnellement et non conditionnellement à la présence des unités dans l'échantillon. Sauf dans quelques cas particuliers, cette loi de distribution est très complexe; aussi construirons-nous une approximation.

Nous pouvons dériver la moyenne et la variance de $\hat{\hat{\mathbf{x}}}_\pi$ conditionnellement et non conditionnellement à la présence des unités dans l'échantillon. En effet $E(\hat{\hat{\mathbf{x}}}_\pi)$, Var $(\hat{\hat{\mathbf{x}}}_\pi)$, et $E(\hat{\hat{\mathbf{x}}}_\pi \mid k \in S)$ proviennent respectivement de (3.2), (3.3) et (3.6) et

$$\begin{aligned}\boldsymbol{\Sigma}_k &= \mathrm{Var}\left(\hat{\hat{\mathbf{x}}}_\pi \mid k \in S\right) \\ &= \frac{1}{N^2} \sum_{\substack{l \in U \\ l \neq k}} \frac{\mathbf{x}_l \mathbf{x}_l' \pi_{kl}}{\pi_k \pi_l^2} \left(1 - \frac{\pi_{kl}}{\pi_k}\right) \\ &\quad + \frac{1}{N^2} \sum_{\substack{l \in U \\ l \neq k}} \sum_{\substack{m \in U \\ m \neq l \\ m \neq k}} \frac{\mathbf{x}_l \mathbf{x}_m'}{\pi_k \pi_l \pi_m} (\pi_{klm} - \frac{\pi_{kl}\pi_{km}}{\pi_k}) \qquad (4.9)\end{aligned}$$

où π_{klm} est la probabilité d'inclusion de troisième ordre. Nous supposons que les matrices $\boldsymbol{\Sigma}$ et $\boldsymbol{\Sigma}_k$ ne sont pas singulières.

Comme on ne connaît généralement pas la loi de distribution de $\hat{\hat{\mathbf{x}}}_\pi$ nous émettrons les trois hypothèses qui suivent pour construire une approximation des probabilités d'inclusion conditionnelles.

i) Si la taille n de l'échantillon est grande, $\hat{\hat{\mathbf{x}}}_\pi$ suit une loi de distribution multivariée normale conditionnellement et non conditionnellement à la présence des unités dans l'échantillon.

ii) $\mathbf{R}_k^{-1} - \mathbf{R}^{-1} = O_{J\times J}(n^{-1})$ pour tous les $k \in U$, où $\mathbf{R} = \mathbf{V}^{-1/2} \boldsymbol{\Sigma} \mathbf{V}^{-1/2}$, $\mathbf{R}_k = \mathbf{V}^{-1/2} \boldsymbol{\Sigma}_k \mathbf{V}^{-1/2}$, $\mathbf{V}$ représente une matrice diagonale $J \times J$ ayant pour diagonale les éléments de la diagonale de $\boldsymbol{\Sigma}$ et où $O_{J\times J}(n^{-\alpha})$ représente une matrice de quantités qui, quand on la multiplie par n^α, reste bornée quand $n \to \infty$.

iii) $\gamma_k = \mathbf{V}^{-1/2}(\bar{\mathbf{x}}_{|k} - \bar{\mathbf{x}}) = O_J(n^{-1/2})$ où $O_J(n^{-\alpha})$ représente un vecteur de quantités qui, quand on le multiplie par n^α, reste borné quand $n \to \infty$.

Ces trois hypothèses s'appliquent à la taille de l'échantillon. Nous supposons donc que, quand n croît, N croît au moins aussi rapidement que n. Néanmoins, aucune hypothèse n'est faite en ce qui concerne $f = n/N$. Si nous supposons que les hypothèses énoncées à la section 3 sont vérifiées, le résultat qui suit représente une approximation de l'estimateur SPC:

Résultat 1:

Si les conditions i), ii) et iii) sont remplies et que nous utilisons la statistique auxiliaire $\hat{\hat{\mathbf{x}}}_\pi$, alors

$$\begin{aligned}\hat{\hat{y}}_{|T} &= \hat{\hat{y}}_\pi + (\bar{\mathbf{x}} - \hat{\hat{\mathbf{x}}}_\pi)' \boldsymbol{\Sigma}^{-1} \frac{1}{N} \sum_{k \in S} \frac{\bar{\mathbf{x}}_{|k} - \bar{\mathbf{x}}}{\pi_k} y_k \\ &\quad + O_p(n^{-1}) \approx \hat{\hat{y}}_{\mathrm{AOPT1}}. \qquad (4.10)\end{aligned}$$

où $n \times O_p(n^{-1})$ est une quantité bornée en probabilité. La preuve du résultat 1 est donnée en annexe.

5. DISCUSSION DES HYPOTHÈSES

Les trois hypothèses susmentionnées sont vérifiées dans le cas de l'échantillonnage aléatoire simple sans remise quand on ne dispose que d'une seule variable auxiliaire. En effet, nous avons alors $J = 1$, $\mathbf{x}_k = x_k$, $\bar{\mathbf{x}} = \bar{x}$, $\bar{\mathbf{x}}_{|k} = \bar{x}_{|k}$ $\hat{\hat{\mathbf{x}}}_\pi = \hat{\hat{x}}_\pi$. Nous obtenons

$$\pi_k = \frac{n}{N}, \pi_{kl} = \frac{n}{N}\frac{n-1}{N-1} \text{ et } \pi_{klm} = \frac{n}{N}\frac{n-1}{N-1}\frac{n-2}{N-2}.$$

D'après (3.6), (3.3), (4.9),

$$\bar{\mathbf{x}}_{|k} = \bar{\mathbf{x}} + \frac{N-n}{N-1} \frac{\mathbf{x}_k - \bar{\mathbf{x}}}{n}, \qquad (5.11)$$

$$\mathrm{Var}\,(\hat{\hat{x}}_\pi) = \frac{N-n}{N-1} \frac{\sigma_x^2}{n}, \qquad (5.12)$$

$$\mathrm{Var}\,(\hat{\hat{x}}_\pi \mid k \in S) = \frac{N(N-n)(n-1)}{(N-2)(N-1)n^2} \left\{\sigma_x^2 - \frac{(x_k - \bar{x})^2}{N-1}\right\} \qquad (5.13)$$

où

$$\sigma_x^2 = \frac{1}{N}\sum_{k\in U}(x_k - \bar{x})^2.$$

Maintenant, considérons les trois hypothèses dans ce cas particulier.

- Madow (1948) a prouvé l'hypothèse i) dans certaines conditions.
- L'hypothèse ii) devient

$$\frac{\mathrm{Var}\left(\hat{\bar{x}}_\pi \mid k\in S\right)}{\mathrm{Var}(\hat{\bar{x}}_\pi)} - 1 = O(n^{-1}).$$

D'après (5.12) et (5.13), nous obtenons

$$\frac{\mathrm{Var}\left(\hat{\bar{x}}_\pi \mid k\in S\right)}{\mathrm{Var}(\hat{\bar{x}}_\pi)} = \frac{N(n-1)}{(N-2)n}\left\{1 - \frac{(x_k - \bar{x})^2}{(N-1)\sigma_x^2}\right\}$$

$$= 1 - \frac{1}{n}\left\{\frac{N-2n}{(N-2)} + \frac{N(n-1)}{(N-2)}\frac{(x_k-\bar{x})^2}{(N-1)\sigma_x^2}\right\}$$

$$= 1 + O\left(\frac{1}{n}\right).$$

- L'hypothèse iii) devient

$$\frac{\bar{x}_{|k} - \bar{x}}{\sqrt{\mathrm{Var}(\hat{\bar{x}}_\pi)}} = O(n^{1/2}).$$

D'après (5.11) (5.12), nous obtenons

$$\gamma_k = \frac{\bar{x}_{|k} - \bar{x}}{\sqrt{\mathrm{Var}(\hat{\bar{x}}_\pi)}} = \frac{\sqrt{N-n}}{N-1}\frac{x_k - \bar{x}}{\sigma_x\sqrt{n}} = O\left(\frac{1}{\sqrt{n}}\right).$$

Dans le cas de l'échantillonnage aléatoire simple, il est plus facile d'interpréter ces hypothèses. L'hypothèse i) est l'hypothèse classique de normalité qui était également nécessaire pour la construction de l'estimateur optimal. Dans le cas de l'échantillonnage aléatoire simple, il est facile de vérifier que l'hypothèse iii) sous-entend l'hypothèse ii). Ces deux hypothèses techniques impliquent simplement que $|x_k - \bar{x}|$ ne peut pas prendre une valeur beaucoup plus importante pour une unité particulière que pour les autres.

Les trois hypothèses sont donc valides dans le cas de l'échantillonnage aléatoire simple quand on ne dispose que d'une variable. Nous pouvons aussi étendre ce résultat à l'échantillonnage stratifié quand le nombre de strates est fixe et que la taille de l'échantillon dans chaque strate est grande. Dans le cas de l'échantillonnage par grappes, si le nombre de grappes est grand et que les grappes sont sélectionnées selon un plan d'échantillonnage aléatoire simple, ces hypothèses sont encore applicables. L'hypothèse i) a également été prouvée partiellement par Rosén (1972) pour l'échantillonnage avec probabilités inégales. En réalité, la preuve de Rosén est limitée à un plan d'échantillonnage avec rejet.

Les hypothèses proposées sont généralement moins restrictives qu'un modèle de superpopulation. En effet, ce genre de modèle consiste en un ensemble d'hypothèses concernant les variables étudiées, alors que les trois hypothèses présentées ici visent uniquement les variables auxiliaires. Dans le cas d'un modèle de superpopulation, les relations entre la variable étudiée et les variables auxiliaires représentent les contributions les plus importantes du modèle. Dans le cas de la méthode conditionnelle, aucune hypothèse n'est faite en ce qui concerne la variable étudiée. Si l'on peut débattre des hypothèses présentées, il est évident qu'un modèle de superpopulation est un ensemble d'hypothèses beaucoup plus restrictives que celles utilisées dans le cas de la méthode conditionnelle.

6. APPLICATION À L'ÉCHANTILLONNAGE STRATIFIÉ

6.1 Le problème

Lors de la stratification, on se sert de données auxiliaires a priori pour améliorer l'estimation. Dans ce cas, trois ensemble de variables interagissent: les variables de stratification, les variables auxiliaires utilisées a posteriori et la variable étudiée. Supposons que l'on partitionne la population en H strates U_h, $h = 1,...,H$, de taille N_h, $h = 1,...,H$. Soient $\bar{y}_h = N_h^{-1}\sum_{k\in U_h} y_k$ et $\bar{\mathbf{x}}_h = N_h^{-1}\sum_{k\in U_h}\mathbf{x}_k$ les moyennes des strates dans la population. Nous sélectionnons sans remise un échantillon aléatoire simple S_h de taille fixe n_h ($\sum_{h=1}^H n_h = n$) indépendamment dans chaque strate. Conformément à la théorie générale de la stratification (voir, par exemple, Särndal, Swensson et Wretman 1992, p. 100), nous obtenons

$$\hat{\bar{y}}_\pi = \frac{1}{N}\sum_{h=1}^H N_h\hat{\bar{y}}_h \text{ et } \hat{\bar{\mathbf{x}}}_\pi = \frac{1}{N}\sum_{h=1}^H N_h\hat{\bar{\mathbf{x}}}_h$$

où

$$\hat{\bar{y}}_h = \frac{1}{n_h}\sum_{k\in S_h} y_k \text{ et } \hat{\bar{\mathbf{x}}}_h = \frac{1}{n_h}\sum_{k\in S_h}\mathbf{x}_k.$$

En outre, nous obtenons

$$\mathrm{Cov}\,(\hat{\bar{\mathbf{x}}}_\pi, \hat{\bar{y}}_\pi) =$$

$$\frac{1}{N^2}\sum_{h=1}^H N_h^2\frac{1-f_h}{n_h}\frac{1}{N_h-1}\sum_{k\in U_h}(\mathbf{x}_k - \bar{\mathbf{x}}_h)(y_k - \bar{y}_h)'$$

et

$$\Sigma = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h}\frac{1}{N_h-1}\sum_{k\in U_h}(\mathbf{x}_k-\bar{\mathbf{x}}_h)(\mathbf{x}_k-\bar{\mathbf{x}}_h)'$$

où $f_h = n_h/N_h$, $h = 1,\ldots,H$.

6.2 Estimateur AOPT1

Si $k \in U_\alpha$, en étendant l'expression (3.6) à l'échantillonnage stratifié, nous obtenons

$$\bar{\mathbf{x}}_{|k} = E(\hat{\bar{\mathbf{x}}}_\pi | k\in S) = \bar{\mathbf{x}} + \frac{N_\alpha^2}{N(N_\alpha-1)}\frac{1-f_\alpha}{n_\alpha}(\mathbf{x}_k-\bar{\mathbf{x}}_\alpha)$$

et

$$\frac{1}{N}\sum_{k\in S}\frac{(\bar{\mathbf{x}}_{|k}-\bar{\mathbf{x}})y_k}{\pi_k} = \frac{1}{N^2}\sum_{h=1}^{H}\frac{N_h^3}{N_h-1}\frac{1-f_h}{n_h}\frac{1}{n_h}\sum_{k\in S_h}(\mathbf{x}_k-\bar{\mathbf{x}}_h)y_k.$$

D'après l'expression (3.7), nous pouvons dériver l'estimateur AOPT1 sous la forme

$$\hat{\bar{y}}_{\text{AOPT1}} = \hat{\bar{y}}_\pi + (\bar{\mathbf{x}}-\hat{\bar{\mathbf{x}}}_\pi)' \left\{\sum_{h=1}^{H} N_h^2\frac{1-f_h}{n_h}\frac{1}{N_h-1}\sum_{k\in U_h}(\mathbf{x}_k-\bar{\mathbf{x}}_h)(\mathbf{x}_k-\bar{\mathbf{x}}_h)'\right\}^{-1} \times\sum_{h=1}^{H}\frac{N_h^3}{N_h-1}\frac{1-f_h}{n_h}\frac{1}{n_h}\sum_{k\in S_h}(\mathbf{x}_k-\bar{\mathbf{x}}_h)y_k.$$

Pour utiliser cet estimateur, nous devons disposer d'une quantité importante de données auxiliaires. Nous devons connaître la moyenne de population $\bar{\mathbf{x}}_h$ des variables auxiliaires pour chaque strate, ainsi que la taille N_h des strates. De surcroît, nous devons connaître les valeurs que prennent les variables auxiliaires pour chaque unité de U.

Cependant, dans le cas de la stratification, $\hat{\bar{y}}_{\text{AOPT1}}$ présente un inconvénient important: il n'est pas calé sur la taille des strates N_h autrement dit, quand l'objectif consiste à estimer la taille des strates N_h, généralement $\hat{N}_{\text{AOPT1}} \neq N_h$. Nous pouvons facilement remédier à cet inconvénient en centrant la variable étudiée. Nous obtenons alors:

$$\hat{\bar{y}}_{\text{AOPT1C}} = \hat{\bar{y}}_\pi + (\bar{\mathbf{x}}-\hat{\bar{\mathbf{x}}}_\pi)' \left\{\sum_{h=1}^{H} N_h^2\frac{1-f_h}{n_h}\frac{1}{N_h-1}\sum_{k\in U_h}(\mathbf{x}_k-\bar{\mathbf{x}}_h)(\mathbf{x}_k-\bar{\mathbf{x}}_h)'\right\}^{-1} \times\sum_{h=1}^{H}\frac{N_h^3}{N_h-1}\frac{1-f_h}{n_h}\frac{1}{n_h}\sum_{k\in S_h}(\mathbf{x}_k-\bar{\mathbf{x}}_h)(y_k-\hat{\bar{y}}_h).$$

6.3 Estimateur AOPT2

Nous pouvons également utiliser l'estimateur AOPT2 pour la stratification. Dans ce cas, d'après l'expression (3.8), nous obtenons

$$\hat{\bar{y}}_{\text{AOPT2}} = \hat{\bar{y}}_\pi + (\bar{\mathbf{x}}-\hat{\bar{\mathbf{x}}}_\pi)' \left\{\sum_{h=1}^{H} N_h^2\frac{1-f_h}{n_h}\frac{1}{n_h-1}\sum_{k\in S_h}(\mathbf{x}_k-\hat{\bar{\mathbf{x}}}_h)(\mathbf{x}_k-\hat{\bar{\mathbf{x}}}_h)'\right\}^{-1} \times\sum_{h=1}^{H} N_h^2\frac{1-f_h}{n_h}\frac{1}{n_h-1}\sum_{k\in S_h}(\mathbf{x}_k-\hat{\bar{\mathbf{x}}}_h)(y_k-\hat{\bar{y}}_h).$$

Pour utiliser l'estimateur AOPT2, nous devons uniquement connaître le vecteur des moyennes de population $\bar{\mathbf{x}}$ et la taille des strates N_h. L'inconvénient tient ici au fait que, comme les $\bar{\mathbf{x}}_h$ sont estimés, on perd $J\times H$ degrés de liberté. Or, si le nombre de strates est élevé, cette perte de degrés de liberté pourrait augmenter l'instabilité de l'estimateur quand la valeur de $J\times H$ est grande.

6.4 Estimateur GREG

L'estimateur GREG ne tient pas compte des probabilités d'inclusion d'ordre deux. Il s'exprime sous la forme

$$\hat{\bar{y}}_{\text{GREG}} = \hat{\bar{y}}_\pi + (\bar{\mathbf{x}}-\hat{\bar{\mathbf{x}}}_\pi)' \left\{\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h}\frac{\mathbf{x}_k\mathbf{x}_k'}{c_k}\right\}^{-1} \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h}\frac{\mathbf{x}_k y_k}{c_k}.$$

Bien qu'il soit plus stable, cet estimateur est entaché d'un biais conditionnel. En outre, si nous voulons estimer la taille des strates N_h au moyen de l'estimateur GREG, nous n'obtenons pas exactement N_h. En effet, si $y_k = 1$ quand $k\in U_h$ et que $y_k = 0$ quand $k\notin U_h$, alors

$$\hat{N}_{\text{GREG}} = N\hat{\bar{y}}_{\text{GREG}} = \hat{N}_{h\pi} + N(\bar{\mathbf{x}}-\hat{\bar{\mathbf{x}}}_\pi)' \left\{\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h}\frac{\mathbf{x}_k\mathbf{x}_k'}{c_k}\right\}^{-1} \times\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h}\frac{\mathbf{x}_k y_k}{c_k}.$$

Puisque, dans le cas de l'échantillonnage stratifié, $\hat{N}_{h\pi} = N_h$, nous obtenons

$$\hat{N}_{\text{GREG}} = N_h + N(\bar{\mathbf{x}}-\hat{\bar{\mathbf{x}}}_\pi)' \left\{\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h}\frac{\mathbf{x}_k\mathbf{x}_k'}{c_k}\right\}^{-1} \times\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h}\frac{\mathbf{x}_k}{c_k}. \quad (6.14)$$

L'expression (6.14) montre que, en général, $\hat{N}_{\text{GREG}} \neq N_h$. Donc, l'estimateur GREG détruit l'effet de stratification, parce qu'il ne tient pas compte de cette dernière. En effet, la stratification est représentée par les probabilités d'inclusion d'ordre deux. Or, l'estimateur GREG ne dépend que des probabilités d'inclusion de premier ordre. En revanche, il est facile de vérifier que les estimateurs AOPT1 et AOPT2 de N_h sont exactement égaux à N_h. Par conséquent, les estimateurs AOPT sont calés sur N_h.

Nous proposons d'utiliser l'estimateur GREG dans les cas suivants: quand la taille de l'échantillon est petite ou que le nombre de strates est grand et quand la stratification est en information auxiliaire médiocre pour la variable d'intérêt. En effet, dans ce cas, la perte de précision due à la perte de degrés de liberté sera plus importante que le gain de précision qui résulte de l'optimalité de l'estimateur. Montanari (1998) présente aussi une analyse intéressante des avantages de l'estimateur optimal.

6.5 Utilisation de l'estimateur GREG avec les variables de stratification

Une variante de l'utilisation de l'estimateur GREG consiste à utiliser de nouveau les variables de stratification au stade de l'estimation.

Considérons le vecteur-colonne

$$\mathbf{w}_k = (z_{k1}, ..., z_{kh}, ..., z_{k(H)}\, \mathbf{x}_k')'$$

où $z_{kh} = 1$ si $k \in U_h$ et 0 autrement. Ce vecteur est donc composé des valeurs prises par les variables indicatrices de la présence de l'unité k dans la strate H et des valeurs prises par les variables auxiliaires de x.

Si $\bar{\mathbf{w}}$ est la moyenne de population des vecteurs $\mathbf{w}_k$ et $\hat{\bar{\mathbf{w}}}_\pi$ son estimateur d'Horvitz-Thompson, l'estimateur GREG contenant l'information auxiliaire $\bar{\mathbf{w}}_\pi$ prend la forme

$$\hat{\bar{y}}_{\text{GREG}w} = \hat{\bar{y}}_\pi + (\bar{\mathbf{w}} - \hat{\bar{\mathbf{w}}}_\pi)'$$

$$\left\{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in S_h} \frac{\mathbf{w}_k \mathbf{w}_k'}{c_k}\right\}^{-1} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in S_h} \frac{\mathbf{w}_k y_k}{c_k}. \quad (6.15)$$

Nous pouvons simplifier la présentation de l'expression (6.15). En effet, le résultat qui suit a été prouvé par Tillé (1994) et généralisé par Särndal (1996):

Résultat 2:

Quand on utilise de nouveau les variables de stratification au stade de l'estimation et que les c_k sont égaux dans la strate ($c_k = c_h, k \in U_h$) on peut écrire l'estimateur GREG sous la forme

$$\hat{\bar{y}}_{\text{GREG}w} =$$

$$\hat{\bar{y}}_\pi + (\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}}_\pi)' \left\{\sum_{h=1}^{H} \frac{N_h}{n_h c_h} \sum_{k \in S_h} (\mathbf{x}_k - \hat{\bar{\mathbf{x}}}_h)(\mathbf{x}_k - \hat{\bar{\mathbf{x}}}_h)'\right\}^{-1}$$

$$\times \sum_{h=1}^{H} \frac{N_h}{n_h c_h} \sum_{k \in S_h} (\mathbf{x}_k - \hat{\bar{\mathbf{x}}}_h)(y_k - \hat{\bar{y}}_h). \quad (6.16)$$

Une preuve du résultat 2 figure à l'annexe. Il convient de souligner que l'expression (6.16) est égale à l'estimateur AOPT2 quand

$$c_h = C \frac{n_h}{N_h} \frac{1 - 1/n_h}{1 - f_h},$$

pour $h = 1, ..., H$, où $C > 0$ est une constante. Quand les f_h sont petits et que les n_h sont grands et proportionnels aux N_h les deux estimateurs sont équivalents. Ce résultat montre que, dans le cas de l'approche conditionnelle, la méthode d'estimation tient automatiquement compte du fait que le plan d'échantillonnage est stratifié. Par contre, l'estimateur GREG ne tient pas compte de l'effet de stratification et il faut donc introduire de nouveau les variables de stratification au stade de l'estimation, afin de ne pas perdre cet effet.

7. SIMULATIONS

Nous avons réalisé un ensemble de simulations afin de comparer les quatre estimateurs suivants: $\hat{\bar{y}}_\pi$, $\hat{\bar{y}}_{\text{AOPT1C}}$, $\hat{\bar{y}}_{\text{AOPT2}}$, $\hat{\bar{y}}_{\text{GREG}}$. La population comprend 4 strates de 250 unités (N=1000). Nous appliquons un plan d'échantillonnage stratifié avec répartition proportionnelle. Pour chaque simulation, nous sélectionnons 10 000 échantillons de taille n =100 et nous estimons les ratios suivants:

$$M_1 = \text{EQM}(\hat{\bar{y}}_\pi)/\text{EQM}(\hat{\bar{y}}_\pi) = 1,$$

$$M_2 = \text{EQM}(\hat{\bar{y}}_{\text{GREG}})/\text{EQM}(\hat{\bar{y}}_\pi),$$

$$M_3 = \text{EQM}(\hat{\bar{y}}_{\text{AOPT1C}})/\text{EQM}(\hat{\bar{y}}_\pi),$$

$$M_4 = \text{EQM}(\hat{\bar{y}}_{\text{AOPT2}})/\text{EQM}(\hat{\bar{y}}_\pi).$$

Nous générons les populations au moyen des modèles suivants: $\mathcal{P}_1: x_k = a_k, y_k = e_k, k \in U$, (indépendance complète), $\mathcal{P}_2: x_k = a_k, y_k = 3x_k + e_k, k \in U$, (dépendance de x et y), $\mathcal{P}_3: x_k = a_k, y_k = x_k + 2h(k) + e_k, k \in U$, (dépendance de x, y et de la strate), $\mathcal{P}_4: x_k = a_k, y_k = \exp(10 + 2x_k + 10h(k) + e_k), k \in U$, (non-linéarité et dépendance de x, y et de la strate), $\mathcal{P}_5: x_k = a_k, y_k = \exp(e_k + 3x_k) + 3h(k), k \in U$, (non-linéarité et dépendance de x, y et de la strate), $\mathcal{P}_6: x_k = a_k, y_k = 3h(k) + e_k, \ k \in U$, (forte dépendance de y et de la strate), $\mathcal{P}_7: x_k = a_k, y_k = 50h(k) + e_k, k \in U$, (très forte dépendance de y et de la strate) où a_k et e_k sont des variables normales indépendantes dont la moyenne est égale

à 0 et la variance, à 1, et h(k) est le numéro de la strate de l'unité k. Les résultats de la simulation sont présentés au tableau 1.

Tableau 1
Résultats de 10 000 simulations

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$	$\mathcal{P}_6$	$\mathcal{P}_7$
M_1	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
M_2	1,0070	0,0906	0,5180	0,9261	0,9263	1,1047	38,5104
M_3	1,0069	0,0906	0,4835	0,9277	0,9269	1,0015	1,0123
M_4	1,0060	0,0936	0,4850	0,9257	0,9239	1,0006	1,0111

Le tableau 1 montre que l'estimateur GREG produit une bonne estimation quand les variables de stratification ne sont pas corrélées à la variable étudiée. Néanmoins, le gain de précision de $\hat{\bar{y}}_{AOPT1C}$ et de $\hat{\bar{y}}_{AOPT2}$ est d'autant plus prononcé que la dépendance entre la variable de stratification et la variable étudiée est forte. Pour cette taille d'échantillon, la perte de degrés de liberté de l'estimateur optimal ne semble pas avoir d'effet sur la précision. En outre, toujours pour cette taille d'échantillon, le gain dû au fait de connaître la taille des strates dans la population n'est pas significatif. Dans tous les cas étudiés, l'estimateur optimal est donc nettement préférable à l'estimateur GREG.

8. UN PROBLÈME DE TROISIÈME ORDRE

La complexité du calcul des poids conditionnels n'est pas un problème particulier à l'estimateur SPC. Elle tient au problème général que pose l'estimation avec utilisation a posteriori de données auxiliaires quand on a déjà utilisé une variable auxiliaire a priori dans le plan d'échantillonnage. Nous pouvons le résumer à un problème d'interaction de troisième ordre entre

- les variables étudiées;
- le plan d'échantillonnage et, donc, les variables auxiliaires utilisées a priori;
- les variables auxiliaires utilisées a posteriori.

En effet, l'utilisation de variables auxiliaires au stade de l'estimation pose le problème de savoir comment ces variables utilisées a posteriori interagissent avec la variable étudiée au moyen d'un plan d'échantillonnage donné. Le problème étant complexe, nous devons tenir compte des relations entre chaque ensemble de variables susmentionnées, ainsi que des interactions de troisième ordre entre les trois ensembles de variables.

Il est très difficile de trouver un estimateur vraiment opérationnel tenant compte des trois interactions de deuxième ordre et de l'interaction de troisième ordre. Aussi essayons-nous de simplifier le problème. La neutralisation de l'un des aspects du problème facilite considérablement la recherche de l'estimateur. La plupart des simplifications possibles ont déjà été étudiées. Nous en mentionnerons quelques-unes ici.

- Si nous ne nous servons d'aucune donnée auxiliaire a posteriori (sauf la taille de la population N) nous ne pouvons construire que l'estimateur d'Horvitz-Thompson ou le ratio de Hájek (1971).
- La recherche de solutions générales au moyen de données auxiliaires ne pose pas de problème important dans le cas de l'échantillonnage aléatoire simple, puisqu'on n'utilise aucune donnée auxiliaire a priori.
- L'utilisation d'un modèle de superpopulation permet de fixer la relation qui existe entre la variable étudiée et les variables auxiliaires utilisées a posteriori. Dans ce cas, il est possible de déterminer l'estimateur optimal, par rapport au modèle.
- Pour l'estimateur GREG ainsi que pour les méthodes de calage (voir Deville et Särndal 1992), dans le cadre de l'inférence fondée sur le plan de sondage, on ne retient du plan de sondage que les probabilité de premier ordre. Donc, on traite un plan d'échantillonnage aléatoire simple de la même façon qu'un plan stratifié pour lequel les probabilités d'inclusion de premier ordre sont toutes égales. C'est pour cette raison qu'un estimateur de régression appliqué à un plan de sondage stratifié détruit généralement le calage des fréquences dans les states fixé au moment de la stratification a priori. Dans ce cas, la simplification tient au fait qu'on ne peut décrire que par la probabilité de premier ordre la contribution de chaque variable auxiliaire utilisée a priori dans le plan de sondage.
- Enfin, pour l'estimateur linéaire optimal, on suppose implicitement que la relation de dépendance des estimateurs d'Horvitz-Thompson calculés pour les variables $\mathbf{x}$ et y est linéaire. Manifestement, cet estimateur ne tient pas compte de la dépendance non linéaire entre les estimateurs. Néanmoins, il tient compte des probabilités d'inclusion d'ordre deux. Si le plan d'échantillonnage est stratifié, l'estimateur reste calé sur les fréquences dans les strates.

L'estimateur SPC tient compte de cette interaction de troisième ordre. Qui plus est, ici, les variables auxiliaires n'interviennent pas nécessairement de façon linéaire. Les poids dépendent à la fois du plan d'échantillonnage et de la variable auxiliaire. Les poids appliqués aux valeurs prises par la variable étudiée tiennent compte de toutes les interactions entre les trois groupes de variables.

Les méthodes fondées sur les probabilités d'inclusion conditionnelles sont intéressantes pour diverses raisons: elles fournissent un cadre général permettant de rechercher et de concevoir des estimateurs qui s'appuient sur des variables auxiliaires sans devoir se reporter à un modèle de superpopulation et aboutissent à une inférence condition-

nelle valide. Elles font ressortir toute la complexité de l'estimation quand on dispose de données auxiliaires. Selon les données auxiliaires dont nous disposons, nous pouvons obtenir soit des résultats connus (comme, par exemple, la stratification a posteriori) ou des estimateurs très complexes qui ne sont pas vraiment opérationnels. Cependant, une première approximation mène à un résultat connu, c'est-à-dire l'estimateur linéaire optimal.

REMERCIEMENTS

L'auteur exprime sa reconnaissance aux deux examinateurs anonymes et à l'éditeur adjoint pour leurs commentaires constructifs et au professeur Carl Särndal pour les remarques intéressantes qu'il a faites au sujet d'une version antérieure du présent article.

ANNEXE: PREUVE DES RÉSULTATS 1 ET 2

Nous nous servirons du Lemme 1 pour prouver le résultat 1.

Lemme 1: si $\mathbf{R}_k^{-1} - \mathbf{R}^{-1} = O_{J\times J}(n^{-1})$, alors $|\mathbf{R}_k|^{-1}|\mathbf{R}| = 1 + O(n^{-1})$, où on définit $\mathbf{R}$ et $\mathbf{R}_k$ comme dans l'hypothèse ii).

Preuve

$$[\mathbf{R}_k^{-1} - \mathbf{R}^{-1}]\mathbf{R} = O_{J\times J}(n^{-1})\mathbf{R}$$

et donc

$$|\mathbf{R}_k^{-1}\mathbf{R}| = |\mathbf{I} + O_{J\times J}(n^{-1})\mathbf{R}| =$$

$$[1 + O(n^{-1})]^J + O(n^{-1}) = 1 + O(n^{-1})$$

où $\mathbf{I}$ est une matrice unité $J \times J$. Donc,

$$\frac{|\mathbf{R}_k|}{|\mathbf{R}|} = \frac{1}{1 + O(n^{-1})} = 1 + O(n^{-1}).$$

Notons que le lemme 1 est une conséquence de l'hypothèse ii).

Preuve du résultat 1

Si nous définissons

$$d_k = \frac{n}{N\pi_k}, \text{ pour tous } k \in U.$$

par l'hypothèse i), nous obtenons:

$$a_k(\hat{\bar{\mathbf{x}}}_\pi) = \frac{n}{\pi_{k|T}N} = \frac{n\Pr(\hat{\bar{\mathbf{x}}}_\pi)}{N\pi_k \Pr(\hat{\bar{\mathbf{x}}}_\pi \mid k \in S)} = d_k \frac{f(\hat{\bar{\mathbf{x}}}_\pi)}{f_k(\hat{\bar{\mathbf{x}}}_\pi)}$$

où f (resp. f_k) est la fonction de densité d'une variable multivariée normale dont la moyenne est $\bar{\mathbf{x}}$ (resp. $\bar{\mathbf{x}}_{|k}$) et la matrice de variance-covariance $\boldsymbol{\Sigma}$ (resp. $\boldsymbol{\Sigma}_k$). Donc,

$$a_k(\hat{\bar{\mathbf{x}}}_\pi) = d_k \frac{\frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp -\frac{1}{2}(\hat{\bar{\mathbf{x}}}_\pi - \bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\hat{\bar{\mathbf{x}}}_\pi - \bar{\mathbf{x}})}{\frac{1}{|\boldsymbol{\Sigma}_k|^{1/2}} \exp -\frac{1}{2}(\hat{\bar{\mathbf{x}}}_\pi - \bar{\mathbf{x}}_{|k})'\boldsymbol{\Sigma}_k^{-1}(\hat{\bar{\mathbf{x}}}_\pi - \bar{\mathbf{x}}_{|k})}. \quad (8.17)$$

Si nous écrivons aussi

$$\mathbf{R} = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2},$$

$$\mathbf{R}_k = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}_k\mathbf{V}^{-1/2},$$

$$\hat{\bar{\mathbf{x}}}_\pi^c = \mathbf{V}^{-1/2}(\hat{\bar{\mathbf{x}}}_\pi - \bar{\mathbf{x}}),$$

$$\gamma_k = \mathbf{V}^{-1/2}(\bar{\mathbf{x}}_{|k} - \bar{\mathbf{x}}),$$

et

$$c_k = d_k \frac{|\mathbf{R}_k|^{1/2}}{|\mathbf{R}|^{1/2}} \exp \frac{1}{2}\hat{\bar{\mathbf{x}}}_\pi^{c\prime}\left(\mathbf{R}_k^{-1} - \mathbf{R}^{-1}\right)\hat{\bar{\mathbf{x}}}_\pi^c, \quad (8.18)$$

nous obtenons

$$a_k(\hat{\bar{\mathbf{x}}}_\pi) = d_k \frac{|\mathbf{R}|^{-1/2}\exp -\frac{1}{2}\hat{\bar{\mathbf{x}}}_\pi^{c\prime}\mathbf{R}^{-1}\hat{\bar{\mathbf{x}}}_\pi^c}{|\mathbf{R}_k|^{-1/2}\exp -\frac{1}{2}(\hat{\bar{\mathbf{x}}}_\pi^c - \gamma_k)'\mathbf{R}_k^{-1}(\hat{\bar{\mathbf{x}}}_\pi^c - \gamma_k)}$$

$$= c_k \exp \frac{1}{2}\gamma_k\mathbf{R}_k^{-1}(\gamma_k - 2\hat{\bar{\mathbf{x}}}_\pi^c). \quad (8.19)$$

En utilisant un développement en série pour le vecteur γ_k de (8.19), nous obtenons

$$a_k(\hat{\bar{\mathbf{x}}}_\pi) = c_k(1 - \gamma_k\mathbf{R}_k^{-1}\hat{\bar{\mathbf{x}}}_\pi^c) + R(\gamma_k^{(0)}). \quad (8.20)$$

où

$$R(\gamma_k^{(0)}) = c_k\left[\exp \frac{1}{2}\gamma_k^{(0)}\mathbf{R}_k^{-1}(\gamma_k^{(0)} - 2\hat{\bar{\mathbf{x}}}_\pi^c)\right]$$

$$\times \gamma_k'\left\{\left[\mathbf{R}_k^{-1}(\gamma_k^{(0)} - \hat{\bar{\mathbf{x}}}_\pi^c)\right]\left[\mathbf{R}_k^{-1}(\gamma_k^{(0)} - \hat{\bar{\mathbf{x}}}_\pi^c)\right]' - \mathbf{R}_k^{-1}\right\}\gamma_k,$$

et $\gamma_k^{(0)}$ est un vecteur dont les éléments sont compris entre les éléments correspondants de γ_k et 0. D'après l'hypothèse iii), nous obtenons directement

$$R(\gamma_k^{(0)}) = O_p(n^{-1}).$$

Par ailleurs, nous savons d'après l'hypothèse ii), le lemme 1 et l'expression (8.18) que

$$c_k = d_k\left[1 + O(n^{-1})\right]^{1/2}\exp\left\{\frac{1}{2}\hat{\bar{\mathbf{x}}}_\pi^{c\prime}O_{J\times J}(n^{-1})\hat{\bar{\mathbf{x}}}_\pi^c\right\}$$

$$= d_k\left[1 + O_p(n^{-1})\right]. \quad (8.21)$$

D'après (8.20) et (8.21), nous obtenons

$$\begin{aligned} a(\hat{\bar{\mathbf{x}}}_\pi) &= d_k\left[1 + O_p^{-1}(n^{-1})\right] \\ &\quad \left\{1 - \gamma_k\left(\mathbf{R}^{-1} + O_{j\times j}(n^{-1})\right)\hat{\bar{\mathbf{x}}}_\pi^c + O_p(n^{-1})\right\} \\ &= d_k\left\{1 - \gamma_k'\mathbf{R}^{-1}\hat{\bar{\mathbf{x}}}_\pi^c + O_p(n^{-1})\right\} \\ &= d_k\left\{1 + (\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}}_\pi)'\sum^{-1}(\bar{\mathbf{x}}_{|k} - \bar{\mathbf{x}}) + O_p(n_{-1})\right\}. \end{aligned}$$

Enfin, nous obtenons

$$\begin{aligned} \hat{\bar{y}}_{|T} &= \frac{1}{n}\sum_{k\in S} a_k(\hat{\bar{\mathbf{x}}}_\pi)y_k \\ &= \hat{\bar{y}}_\pi + (\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}}_\pi)'\sum^{-1}\frac{1}{N}\sum_{k\in S}\frac{(\bar{\mathbf{x}}_{|k} - \bar{\mathbf{x}})y_k}{\pi_k} + O_p(n^{-1}). \end{aligned}$$

Preuve du résultat 2

D'après Särndal (1980), nous pouvons aussi écrire l'estimateur GREG représenté par (6.15) de la façon suivante:

$$\begin{aligned} \hat{\bar{y}}_{\text{GREG w}} = \hat{\bar{y}}_\pi + N^{-1}&(\mathbf{1}_N'\mathbf{W}_N - \mathbf{1}_S'\boldsymbol{\Pi}_S^{-1}\mathbf{W}_S)' \\ &(\mathbf{W}_S'\mathbf{C}_S^{-1}\boldsymbol{\Pi}_S^{-1}\mathbf{W}_S)^{-1}\mathbf{W}_S'\mathbf{C}_S^{-1}\boldsymbol{\Pi}_S^{-1}\mathbf{y}_S \end{aligned}$$

où $\mathbf{1}_N$ (resp. $\mathbf{1}_S$) est un vecteur-colonne comptant N (resp. n) unités, $\boldsymbol{\Pi}_N$ (resp. $\boldsymbol{\Pi}_S$) est une matrice diagonale dont les éléments diagonalisant les probabilités d'inclusion des unités de la population (resp. de l'échantillon), $\mathbf{C}_S$ est une matrice diagonale dont les éléments diagonalisant les c_k des unités de l'échantillon, $\mathbf{y}_S$ est un vecteur-colonne composé des valeurs prises par la variable étudiée y dans l'échantillon,

$$\mathbf{W}_N = \begin{pmatrix} z_{11} & \cdots & z_{1H} & \mathbf{x}_1' \\ \vdots & & \vdots & \vdots \\ z_{N1} & \cdots & z_{NH} & \mathbf{x}_N' \end{pmatrix}$$

et $\mathbf{W}_S$ est une matrice $n \times (H + J)$ composée des n lignes de $\mathbf{W}_N$ correspondant aux unités sélectionnées dans l'échantillon.

La matrice qu'il faut inverser peut être partitionnée en quatre blocs:

$$(\mathbf{W}_S'\mathbf{C}_S^{-1}\boldsymbol{\Pi}_S^{-1}\mathbf{W}_S) = \begin{bmatrix} \mathbf{A} & \mathbf{D} \\ \mathbf{D}' & \mathbf{B} \end{bmatrix}$$

où A est une matrice $H \times H$ dont la diagonale correspond à N_h/c_h, $h = 1, ..., H$,

$$\mathbf{B} = \sum_{h=1}^{H}\frac{N_h}{n_h c_h}\sum_{k\in S_h}\mathbf{x}_k\mathbf{x}_k'$$

et

$$\mathbf{D}' = \begin{bmatrix} \frac{N_1\hat{\bar{\mathbf{x}}}_1}{c_1} & \cdots & \frac{N_H\hat{\bar{\mathbf{x}}}_H}{c_H} \end{bmatrix}.$$

Par la méthode d'inversion matricielle parties, nous obtenons

$$(\mathbf{W}_S'\mathbf{C}_S^{-1}\boldsymbol{\Pi}_S^{-1}\mathbf{W}_S)^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{D}\mathbf{B}^{-1}\mathbf{D}')^{-1} & -\mathbf{A}^{-1}\mathbf{D}\mathbf{Q} \\ -\mathbf{Q}\mathbf{D}'\mathbf{A}^{-1} & \mathbf{Q} \end{bmatrix}$$

où $\mathbf{Q} = (\mathbf{B} - \mathbf{D}'\mathbf{A}^{-1}\mathbf{D})^{-1}$. Puisque

$$(\mathbf{1}_N'\mathbf{W}_N - \mathbf{1}_s'\boldsymbol{\Pi}_S^{-1}\mathbf{W}_S) = \begin{bmatrix} \mathbf{0}_H \\ \hat{\bar{\mathbf{x}}}_\pi - \bar{\mathbf{x}} \end{bmatrix}$$

où $\mathbf{0}_H$ est un vecteur-colonne composé de H zéros, nous obtenons

$$\begin{aligned} (\mathbf{1}_N'\mathbf{W}_N - \mathbf{1}_s'\boldsymbol{\Pi}_S^{-1}\mathbf{W}_S)'&(\mathbf{W}_S'\mathbf{C}_S^{-1}\boldsymbol{\Pi}_S^{-1}\mathbf{W}_S)^{-1} \\ &= (\hat{\bar{\mathbf{x}}}_\pi - \bar{\mathbf{x}})'\mathbf{Q}[-\mathbf{D}'\mathbf{A}^{-1}\ \mathbf{I}_{(J\times J)}] \end{aligned}$$

où $\mathbf{I}_{(J\times J)}$ est une matrice unité $J \times J$. Puisque

$$\mathbf{Q} = \left\{\sum_{h=1}^{H}\frac{N_h}{n_h c_h}\sum_{k\in S_h}(\mathbf{x}_k - \hat{\bar{\mathbf{x}}}_h)(\mathbf{x}_k - \hat{\bar{\mathbf{x}}}_h)'\right\}^{-1},$$

$$[-\mathbf{D}\mathbf{A}^{-1}\ \mathbf{I}_{(J\times J)}] = [-\hat{\bar{\mathbf{x}}}\ \ldots\ -\hat{\bar{\mathbf{x}}}_H\ \mathbf{J}_{(J\times J)}]$$

et

$$\mathbf{W}_S'\mathbf{C}_S^{-1}\boldsymbol{\Pi}_S^{-1}\mathbf{y}_S = \begin{bmatrix} \frac{N_1\hat{\bar{y}}_1}{c_1} \\ \vdots \\ \frac{N_H\hat{\bar{y}}_H}{c_H} \\ \sum_{h=1}^{H}\frac{N_h}{n_h c_h}\sum_{k\in S_h}x_k y_k \end{bmatrix}, \qquad (8.23)$$

nous obtenons le résultat 2 en multipliant (8.22) par (8.23).

BIBLIOGRAPHIE

CASADY, R.J., et VALLIANT, R. (1993). Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.

CASSEL, C.-M., SÄRNDAL, C.-E., et WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63, 615-620.

DEVILLE J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceeding of the Workshop: Auxiliary Information in Surveys*, Örebro.

DEVILLE J.-C., et SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 376-382.

FULLER, W.A., et ISAKI, C.T. (1981). Survey design under superpopulation models. Dans *Current Topics in Survey Sampling*. (Éds. D. Krewski, R. Platek, J.N.K. Rao, et M.P. Singh). New York: Academic Press, 196-226.

HÁJEK, J. (1971). Comment on a essay of D. Basu. Foundations of statistical inference. (Éds. V.P. Godambe, et D.A. Sprott), Toronto, Holt, Rinehart et Winston, 236.

HORVITZ, D.G., et THOMPSON, D.J. (1952).A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 7, 663-685.

MADOW, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, 19, 535-545.

MONTANARI, G. E. (1987). Post sampling efficient QR-prediction in large sample survey. *Revue Internationale de Statistique*, 55, 191-202.

MONTANARI, G.E. (1997). On conditional properties of finite population mean estimators. *Actes de la 51e Session de l'Institut International de Statistique*, Communication contribuée, 351-352.

MONTANARI, G. E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 71-79.

RAO, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

RAO, J.N.K. (1997). Development in sample survey theory: an appraisal. *La Revue Canadienne de Statistique*, 25, 1-21.

ROSÉN, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. *Annals of Mathematical Statistics*, 43, 373-397.

SÄRNDAL, C.-E. (1980). On π- inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.-E (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

TILLÉ, Y. (1994). Utilisation d'information auxiliaire en théorie des sondages sans référence à un modèle. Thèse de doctorat, Université Libre de Bruxelles, Institut de Statistique.

TILLÉ, Y. (1998). Estimation in surveys using conditional inclusion probabilities: simple random sampling. *Revue Internationale de Statistique*, 66, 303-322.

WRIGHT, R.L. (1983). Finite population with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-883.

Techniques d'enquête, juin 1999
Vol. 25, n° 1; pp. 73-79
Statistique Canada

Estimation régionale robuste au moyen d'un modèle simple à effets aléatoires

N.G.N. PRASAD et J.N.K. RAO[1]

RÉSUMÉ

Les auteurs étudient l'estimation régionale robuste en fonction d'un modèle simple à effets aléatoires constitué d'un modèle de base (ou à effets fixes) et d'un modèle de liaison qui traite les effets fixes comme des réalisations d'une variable aléatoire. À l'aide de ce modèle, on obtient un estimateur modélisé d'une moyenne régionale. Cet estimateur, qui dépend des poids d'enquête, demeure conforme au plan. On obtient également un estimateur à base de modèle de son erreur quadratique moyenne (EQM). Les résultats d'une simulation indiquent que l'estimateur proposé et l'estimateur modélisé de Kott (1989) sont également efficaces, et que l'estimateur de l'EQM proposé est souvent beaucoup plus stable que l'estimateur de l'EQM de Kott, même pour des écarts modérés du modèle de liaison. La méthode s'étend également à des modèles de régression à erreur emboîtée.

MOTS CLÉS: Conforme au plan; modèle de liaison; erreur quadratique moyenne; poids d'enquête.

1. INTRODUCTION

On utilise souvent, pour l'estimation régionale, des modèles à effets aléatoires au niveau de l'unité pour obtenir des estimateurs efficaces à base de modèle pour les moyennes régionales. Typiquement, ce genre d'estimateur ne fait pas appel aux poids d'enquête (par exemple Ghosh et Meeden 1986; Battese, Harter et Fuller 1988; Prasad et Rao 1990). Par conséquent, les estimateurs ne sont pas conformes au plan à moins que le plan d'échantillonnage ne soit autopondéré au sein des régions. On trouvera dans Ghosh et Rao (1994) une évaluation des méthodes d'estimation régionale.

Kott (1989) a recommandé l'utilisation d'estimateurs à base de modèles conformes au plan (par exemple des estimateurs modélisés) parce que de tels estimateurs offrent une protection contre la défaillance du modèle à mesure que la taille de l'échantillon des petites régions augmente. Il a calculé un estimateur conforme au plan pour une moyenne régionale à l'aide d'un modèle simple à effets aléatoires. Ce modèle englobe deux composantes: le modèle de base (ou à effets fixes) et le modèle de liaison. Le modèle de base est donné d'après

$$y_{ij} = \theta_i + e_{ij}, j = 1, 2, ..., N_i; i = 1, 2, ..., m \qquad (1)$$

où les y_{ij} sont les observations de l'échantillon et les e_{ij} sont des erreurs aléatoires non corrélées de moyenne zéro et de variance σ_i^2 pour chaque petite région i (= 1, 2, ..., m). Par souci de simplicité, nous notons θ_i la moyenne régionale $\bar{Y}_i = \sum_j y_{ij}/N_i$, où N_i est le nombre d'unités de la population de la i-ième région. À noter que $\bar{Y}_i = \theta_i + \bar{E}_i$ et $\bar{E}_i = \sum_j e_{ij}/N_i \approx 0$ si N_i est grand.

Le modèle de liaison suppose que θ_i est une réalisation d'une variable aléatoire qui satisfait au modèle

$$\theta_i = \mu + v_i \qquad (2)$$

où les v_i sont des variables aléatoires non corrélées de moyenne zéro et de variance σ_v^2. De plus, $\{v_i\}$ et $\{e_{ij}\}$ sont supposés non corrélés.

En supposant que le modèle (1) reste valable pour l'échantillon $\{y_{ij}, j = 1, 2, ..., n_i; i = 1, 2, ..., m\}$ et en combinant le modèle échantillon et le modèle de liaison, Kott (1989) a obtenu le modèle familier à effets aléatoires au niveau de l'unité

$$y_{ij} = \mu + v_i + e_{ij}, j = 1, 2, ..., n_i; i = 1, 2, ..., m, \qquad (3)$$

appelé également modèle des composantes de la variance. On suppose habituellement des variances égales $\sigma_i^2 = \sigma^2$, même si le cas des variances d'erreur aléatoires a également été étudié (Kleffe et Rao 1992; Arora et Lahiri 1997).

En supposant $\sigma_i^2 = \sigma^2$, Kott (1989) a calculé un estimateur efficace $\hat{\theta}_{iK}$ de θ_i qui est à la fois non biaisé pour le modèle d'après (3) et conforme au plan. Il a également proposé un estimateur de son erreur quadratique moyenne (EQM) qui est non biaisé pour le modèle de base (1), tout en étant conforme au plan. Toutefois, cet estimateur de l'EQM peut être assez instable et peut même avoir des valeurs négatives, comme l'a remarqué Kott (1989) dans son exemple empirique. Kott (1989) a utilisé ses estimateurs de l'EQM surtout pour comparer la réduction globale de l'EQM lorsqu'on utilise $\hat{\theta}_{iK}$ au lieu d'un estimateur direct à base de plan $\bar{y}_{iw}$ donné d'après (4) ci-dessous. Il a fait remarquer qu'il faut des estimateurs de l'EQM plus stables.

Le présent exposé est axé surtout sur l'obtention d'un estimateur de meilleure prédiction linéaire non biaisée (MPLNB) pseudo-empirique de θ_i qui dépend des poids

[1] N.G.N. Prasad, Department of Mathematical Sciences, University of Alberta, Edmonton, (Alberta), T6G 2G1; J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa (Ontario), K1S 5B6.

d'enquête et qui est conforme au plan (section 2). Un estimateur de l'EQM stable à base de modèle est également obtenu (section 3). Les résultats d'une étude de simulation à la section 4 montrent que l'estimateur de l'EQM proposé est souvent beaucoup plus stable que l'estimateur de l'EQM de Kott, mesuré selon le coefficient de variation, même pour des écarts modérés du modèle de liaison (2). Les résultats pour le modèle simple (3) s'étendent également à un modèle de régression à erreur emboîtée (section 5).

2. ESTIMATEUR DE MPLNB PSEUDO-EMPIRIQUE

Soit $\tilde{w}_{ij}$ le poids du plan de base attribué à la j-ième unité d'échantillonnage ($j = 1, 2, ..., n_i$) de la i-ième région ($i = 1, 2, ..., m$). Un estimateur direct à base de plan de θ_i est alors donné d'après l'estimateur par quotient

$$\bar{y}_{iw} = \sum_j \tilde{w}_{ij} y_{ij} \Big/ \sum_j \tilde{w}_{ij} = \sum_j w_{ij} y_{ij} \tag{4}$$

où $w_{ij} = \tilde{w}_{ij} / \sum_j \tilde{w}_{ij}$. L'estimateur direct $\bar{y}_{iw}$ est conforme au plan, mais il n'est pas renforcé par les autres régions.

Afin d'obtenir un estimateur plus efficace, nous examinons le modèle réduit ci-dessous obtenu du modèle combiné (3) avec $\sigma_i^2 = \sigma^2$:

$$\begin{aligned} \bar{y}_{iw} &= \sum_j w_{ij}(\mu + v_i + e_{ij}) \\ &= \mu + v_i + \bar{e}_{iw}, \end{aligned} \tag{5}$$

où les $\bar{e}_{iw}$ sont des variables aléatoires non corrélées de moyenne zéro et de variance $\delta_i = \sigma^2 \sum_j w_{ij}^2$. Le modèle réduit (5) est un modèle au niveau de la région qui est semblable au modèle bien connu de Fay-Herriot (Fay et Herriot 1979). Il s'ensuit d'après la théorie standard de la meilleure prédiction linéaire non biaisée (MPLNB) (voir Prasad et Rao 1990) que l'estimateur de MPLNB de $\theta_i = \mu + v_i$ pour le modèle réduit (5) est donné d'après

$$\tilde{\theta}_i = \tilde{\mu}_w + \tilde{v}_i, \tag{6}$$

où

$$\tilde{v}_i = \gamma_{iw}(\bar{y}_{iw} - \tilde{\mu}_w)$$

avec $\tilde{\mu}_w = \sum_i \gamma_{iw} \bar{y}_{iw} / \sum_i \gamma_{iw}$ et $\gamma_{iw} = \sigma_v^2/(\sigma_v^2 + \delta_i)$. À noter que $\tilde{\theta}_i$ est différent de l'estimateur de MPLNB pour le modèle intégral (3). Soit donc $\tilde{\theta}_i$ un estimateur pseudo-MPLNB. L'estimateur (6) s'écrit également sous forme de combinaison convexe de l'estimateur direct $\bar{y}_{iw}$ et $\tilde{\mu}_w$:

$$\tilde{\theta} = \gamma_{iw} \bar{y}_{iw} + (1 - \gamma_{iw}) \tilde{\mu}_w. \tag{7}$$

L'estimateur $\tilde{\theta}_i$ dépend des paramètres σ_v^2 et σ^2 qui, concrètement, ne sont généralement pas connus. Par conséquent, nous remplaçons σ_v^2 et σ^2 dans l'équation (7) par des estimateurs conformes aux modèles $\hat{\sigma}_v^2$ et $\hat{\sigma}^2$ dans le cadre du modèle original au niveau de l'unité (3) de façon à obtenir l'estimateur

$$\hat{\theta}_i = \hat{\gamma}_{iw} \bar{y}_{iw} + (1 - \hat{\gamma}_{iw}) \hat{\mu}_w, \tag{8}$$

où

$$\hat{\gamma}_{iw} = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}^2 \sum_j w_{ij}^2 \right)$$

et

$$\hat{\mu}_w = \sum_i \hat{\gamma}_{iw} \bar{y}_{iw} \Big/ \sum_i \hat{\gamma}_{iw}.$$

Nous désignons l'estimateur $\hat{\theta}_i$ l'estimateur de MPLNB pseudo-empirique. Nous utilisons des estimateurs standard de σ_v^2 et de σ^2 en fonction des sommes des carrés intrarégion

$$Q_w = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

et des sommes des carrés interrégion

$$Q_b = \sum_i n_i (\bar{y}_i - \bar{y})^2,$$

où $\bar{y} = \sum_i n_i \bar{y}_i / \sum_i n_i$ est la moyenne globale de l'échantillon. Nous avons

$$\hat{\sigma}^2 = Q_w \Big/ \left(\sum_i n_i - m \right)$$

et $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$ où

$$\tilde{\sigma}_v^2 = [Q_b - (m - 1)\hat{\sigma}^2]/n^*$$

avec

$$n^* = \sum_i n_i - \sum_i n_i^2 \Big/ \sum_i n_i.$$

Il est à noter que σ_v^2 et σ^2 ne peuvent pas être estimés ou sont estimés de façon imparfaite à partir du modèle réduit (5) à cause de problèmes d'identification. Suivant Kackar et Harville (1984), on peut montrer que l'estimateur de MPLNB pseudo-empirique $\hat{\theta}_i$ n'est pas biaisé selon le modèle pour θ_i dans le cadre du modèle original (3) pour des erreurs $\{v_i\}$ et $\{e_{ij}\}$, réparties symétriquement, pas nécessairement normales. Il est aussi conforme au plan, à supposer que $n_i \sum_j w_{ij}^2$ soit limité à mesure que n_i augmente, parce que $\hat{\gamma}_{iw}$ converge en probabilité vers 1 selon que $n_i \to \infty$ peu importe la validité du modèle (3), à supposer que $\hat{\sigma}_v^2$ et $\hat{\sigma}^2$ convergent en probabilité, par exemple vers $\hat{\sigma}_v^{*2}$ et σ^{*2}.

On obtient l'estimateur à base de modèle de Kott (1989) de θ_i grâce à une combinaison pondérée de $\bar{y}_{iw}$ et de $\sum_{l \neq i} c_l^{(i)} \bar{y}_l$, c'est-à-dire,

$$f_i(\alpha_i, c^{(i)}) = (1 - \alpha_i) \bar{y}_{iw} + \alpha_i \sum_{l \neq i} c_l^{(i)} \bar{y}_l,$$

et en minimisant ensuite l'erreur quadratique moyenne (EQM) du modèle de $f_i(\alpha_i, c^{(i)})$ relativement à α_i et à $c_l^{(i)}$ sous réserve de l'absence de biais pour le modèle: $\sum_{l \neq i} c_l^{(i)} = 1$. Cela donne

$$\hat{\theta}_{iK} = f_i(\hat{\alpha}_i, \hat{c}^{(i)}) \tag{9}$$

avec

$$\hat{\alpha}_i = \sum_j \left[w_{ij}^2 \Big/ \left\{ \sum_j w_{ij}^2 + \sum_{l \neq i} \hat{c}_l^{(i)2} \Big/ n_i + \left(1 + \sum_{l \neq i} \hat{c}_l^{(i)2} \right) \left(\hat{\sigma}_v^2 / \hat{\sigma}^2 \right) \right\} \right]$$

et

$$\hat{c}_l^{(i)} = \left[(\hat{\sigma}_v^2 / \hat{\sigma}^2) + n_l^{-1} \right] \Big/ \sum_{h \neq i} \left[(\hat{\sigma}_v^2 / \hat{\sigma}^2) + n_h^{-1} \right].$$

L'estimateur $\hat{\theta}_{iK}$ est également non biaisé pour le modèle et conforme au plan. Dans une version antérieure du présent exposé, nous avons proposé un estimateur semblable à (9). Il fait appel aux meilleurs estimateurs de μ dans le cadre du modèle au niveau de l'unité, d'après les moyennes non pondérées $\bar{y}_i$, plutôt que $\hat{\mu}_w$, le meilleur estimateur de μ dans le cadre du modèle réduit (4), d'après les moyennes $\bar{y}_{iw}$ pondérées en fonction de l'enquête.

3. ESTIMATEURS DE L'EQM

Le calcul de l'EQM de l'estimateur pseudo-MPLNB $\tilde{\theta}_i$ est simple dans le cadre du modèle au niveau de l'unité (3). Nous avons

$$\mathrm{EQM}(\tilde{\theta}_i) = E(\tilde{\theta}_i - \theta_i)^2 = g_{1i}(\sigma_v^2, \sigma^2) + g_{2i}(\sigma_v^2, \sigma^2) \tag{10}$$

avec

$$g_{1i}(\sigma_v^2, \sigma^2) = (1 - \gamma_{iw})\sigma_v^2$$

et

$$g_{2i}(\sigma_v^2, \sigma^2) = \sigma_v^2 (1 - \gamma_{iw})^2 \Big/ \sum_i \gamma_{iw}.$$

Le terme dominant $g_{1i}(\sigma_v^2, \sigma^2)$ est d'ordre $O(1)$, tandis que le deuxième terme $g_{2i}(\sigma_v^2, \sigma^2)$, à cause de l'estimation de μ, est d'ordre $O(m^{-1})$ pour un grand m.

On obtient un estimateur de l'EQM naïf de l'estimateur de MPLNB pseudo-empirique $\hat{\theta}_i$ en estimant l'EQM$(\hat{\theta}_i)$ donnée d'après (10):

$$\mathrm{eqm}_N(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2). \tag{11}$$

Toutefois, (11) pourrait entraîner une nette sous-estimation de l'EQM$(\hat{\theta}_i)$ en laissant de côté l'incertitude associée à $\hat{\sigma}_v^2$ et à $\hat{\sigma}^2$. À noter que

$$\mathrm{EQM}(\hat{\theta}_i) = \mathrm{EQM}(\tilde{\theta}_i) + E(\hat{\theta}_i - \tilde{\theta}_i)^2 \tag{12}$$

en cas de normalité des erreurs $\{v_i\}$ et $\{e_{ij}\}$, de sorte que l'EQM$(\tilde{\theta}_i)$ est toujours plus petite que l'EQM$(\hat{\theta}_i)$; (voir Kackar et Harville 1984).

Afin d'obtenir un estimateur «correct» de l'EQM$(\hat{\theta}_i)$, nous établissons une approximation du terme de second ordre $E(\hat{\theta}_i - \tilde{\theta}_i)^2$ dans (12) pour un grand m, en supposant que $\{v_i\}$ et $\{e_{ij}\}$, sont distribués normalement. Suivant Prasad et Rao (1990), nous avons

$$E(\hat{\theta}_i - \tilde{\theta}_i)^2 \approx g_{3i}(\sigma_v^2, \sigma^2) \tag{13}$$

où les termes laissés de côté sont d'ordre inférieur à m^{-1}, et

$$g_{3i}(\sigma_v^2, \sigma^2) = \gamma_{iw}(1 - \gamma_{iw})^2 \sigma_v^{-2} \{ V(\tilde{\sigma}_v^2) - 2(\sigma_v^2/\sigma^2)\mathrm{Cov}(\tilde{\sigma}_v^2, \hat{\sigma}^2) + (\sigma_v^2/\sigma^2)^2 \mathrm{Var}(\hat{\sigma}^2) \}; \tag{14}$$

(voir l'annexe 1). On trouvera aussi à l'annexe 1 les variances et covariances de $\hat{\sigma}_v^2$ et $\hat{\sigma}^2$. Il est possible de montrer que $g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ est à peu près non biaisé pour $g_{1i}(\sigma_v^2, \sigma^2)$ du fait que son biais est d'ordre inférieur à m^{-1} (voir l'annexe 2). De même, $g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ et $g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ sont à peu près non biaisés pour $g_{2i}(\sigma_v^2, \sigma^2)$ et $g_{3i}(\sigma_v^2, \sigma^2)$, respectivement. Il s'ensuit qu'un estimateur à peu près non biaisé pour le modèle de l'EQM$(\hat{\theta}_i)$ est donné d'après

$$\mathrm{eqm}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2). \tag{15}$$

Pour l'estimateur $\hat{\theta}_{iK}$ donné d'après (9), Kott (1989) a proposé comme estimateur de l'EQM

$$\mathrm{eqm}(\hat{\theta}_{iK}) = (1 - 2\hat{\alpha}_i) v^*(\bar{y}_{iw}) + \hat{\alpha}_i^2 \left(\bar{y}_{iw} - \sum_{l \neq i} c_l^{(i)} \bar{y}_l \right)^2, \tag{16}$$

où $v^*(\bar{y}_{iw})$ est à la fois un estimateur conforme au plan de l'EQM pour le plan de $\bar{y}_{iw}$ et un estimateur non biaisé pour le modèle de la variance-modèle de $\bar{y}_{iw}$ dans le cadre du modèle de base (1). Puisque $\hat{\alpha}_i$ converge en probabilité vers zéro selon que $n_i \to \infty$, il s'ensuit d'après (16) que l'eqm$(\hat{\theta}_{iK})$ est aussi à la fois conforme au plan et non biaisée pour le modèle si l'on considère uniquement le modèle de base (1). Toutefois, l'eqm$(\hat{\theta}_{iK})$ est instable et peut même avoir une valeur négative lorsque $\hat{\alpha}_i$ dépasse 0,5, comme l'a fait remarquer Kott (1989).

À noter que notre estimateur de l'EQM$(\hat{\theta}_i)$, se fonde sur le modèle intégral (3) obtenu en combinant le modèle de base (1) et le modèle de liaison (2). Toutefois, les résultats de notre simulation de la section 4 indiquent qu'il peut fonctionner convenablement même en présence d'écarts modérés relativement au modèle de liaison.

4. ÉTUDE DE SIMULATION

Nous avons mené une étude de simulation limitée afin d'évaluer le fonctionnement de l'estimateur proposé $\hat{\theta}_i$, donné d'après (8), et de son estimateur de l'EQM, donné d'après (15), relativement à l'estimateur de Kott $\hat{\theta}_{iK}$, donné d'après (9), et à son estimateur de l'EQM, donné d'après (16). Nous avons mené l'étude en fonction de deux stratégies: i) pour chaque simulation, on établit une population finie de $m = 30$ petites régions avec $N_i = 200$ unités de population dans chaque région, en fonction du modèle proposé au niveau de l'unité, puis on tire indépendamment un échantillon avec PPT (probabilité proportionnelle à la taille) au sein de chaque petite région en fixant de $n_i = 20$; ii) on établit d'abord une population finie fixe en fonction du modèle proposé au niveau de l'unité, puis pour chaque simulation on tire indépendamment un échantillon avec PPT au sein de chaque petite région à l'aide de la population finie fixe. La première stratégie (i) se rapporte au plan aussi bien qu'au modèle de liaison, tandis que la deuxième stratégie (ii) se fonde sur le plan en ce sens qu'elle se rapporte uniquement au plan. Nous supposons que les erreurs $\{v_i\}$ et $\{e_{ij}\}$ sont distribuées normalement au moment de l'établissement des populations finies $\{y_{ij}, i = 1,2,...,30; j = 1,2,...,200\}$. Nous avons considéré deux cas: 1) le modèle de liaison (2) est vrai pour $\mu = 50$; 2) on enfreint le modèle de liaison en laissant μ varier d'une région à l'autre: $\mu_i = 50$, $i = 1,2,...,10$; $\mu_i = 55$, $i = 11,12,...,20$; $\mu_i = 60$, $i = 21,22,...,30$. Pour la mise en oeuvre de l'échantillonnage avec PPT dans chaque région, des mesures de la taille ($i = 1,2,...,30; j = 1,2,...,200$) ont été établies à partir d'une distribution exponentielle de moyenne 200. À l'aide de ces valeurs z, nous avons calculé les probabilités de sélection $p_{ij} = z_{ij} / \sum_j z_{ij}$ pour chaque région i, puis nous les avons utilisées pour la sélection des échantillons de la taille $n_i = n$, avec PPT et avec remise, en prenant $n = 20$, et nous avons observé les valeurs d'échantillon associées $\{y_{ij}\}$.

Les poids de base du plan sont donnés d'après $\tilde{w}_{ij} = n^{-1} p_{ij}^{-1}$ de sorte que $w_{ij} = p_{ij}^{-1} / \sum_j p_{ij}^{-1}$. À l'aide de ces poids et des valeurs d'échantillon associées y_{ij} nous avons calculé les estimations $\hat{\theta}_i$ et $\hat{\theta}_{iK}$ et les estimations associées de l'EQM, de même que l'estimation par quotient $\bar{y}_{iw}$ pour chaque simulation; on trouvera à l'annexe 3 la formule pour $v^*(\bar{y}_{iw})$ dans le cadre de l'échantillonnage avec PPT. Ce processus a été répété $R = 10\,000$ fois de façon à obtenir pour chaque exécution les estimations $r(=1,2,...,R)$ $\hat{\theta}_i(r)$ et $\hat{\theta}_{iK}(r)$ et les estimations associées de l'$\text{eqm}_i(\hat{\theta}_i(r))$ et l'$\text{eqm}_i(\hat{\theta}_{iK}(r))$ de même que l'estimation directe $\bar{y}_{iw}(r)$. À l'aide de ces valeurs, nous avons calculé l'efficacité relative (ER) empirique de $\hat{\theta}_i$ et de $\hat{\theta}_{iK}$ pour $\bar{y}_{iw}$ sous la forme

$$\text{ER}(\hat{\theta}_i) = \text{EQM}_*(\bar{y}_{iw}) / \text{EQM}_*(\hat{\theta}_i)$$

et

$$\text{ER}(\hat{\theta}_{iK}) = \text{EQM}_*(\bar{y}_{iw}) / \text{EQM}_*(\hat{\theta}_{iK})$$

où EQM_* désigne l'EQM pour $R = 10\,000$ exécutions. Ainsi, $\text{EQM}_*(\hat{\theta}_i) = \sum_r [\hat{\theta}_i(r) - \bar{Y}_i(r)]^2 / R$, où $\bar{Y}_i(r)$ est la i-ième moyenne de population aréolaire pour la r-ième exécution. À noter que $\bar{Y}_i(r)$ demeure identique pour les exécutions r dans le cadre de la stratégie fondée sur le plan puisque la population finie reste fixe pour les simulations.

De même, nous avons calculé le biais relatif des estimateurs de l'EQM sous la forme

$$\text{BR}[\text{eqm}(\hat{\theta}_i)] = [\text{EQM}_*(\hat{\theta}_i) - E_*\text{eqm}(\hat{\theta}_i)] / \text{EQM}_*(\hat{\theta}_i)$$

et

$$\text{BR}[\text{eqm}(\hat{\theta}_{iK})] = [\text{EQM}_*(\hat{\theta}_{iK}) - E_*\text{eqm}(\hat{\theta}_{iK})] / \text{EQM}_*(\hat{\theta}_{iK}),$$

où E_* désigne l'espérance pour $R = 10\,000$ exécutions. Ainsi, $E_*\text{eqm}(\hat{\theta}_i) = \sum_r \text{eqm}(\hat{\theta}_i(r)) / R$. Enfin, nous avons calculé le coefficient de variation (cv) empirique des estimateurs de l'EQM sous la forme

$$\text{CV}[\text{eqm}(\hat{\theta}_i)] = [\text{EQM}_*\{\text{eqm}(\hat{\theta}_i)\}]^{1/2} / \text{EQM}_*(\hat{\theta}_i)$$

et

$$\text{CV}[\text{eqm}(\hat{\theta}_{iK})] = [\text{EQM}_*\{\text{eqm}(\hat{\theta}_{iK})\}]^{1/2} / \text{EQM}_*(\hat{\theta}_{iK}).$$

À noter que $\text{EQM}_*[\text{eqm}(\hat{\theta}_i)] = \sum_r [\text{eqm}(\hat{\theta}_i(r)) - \text{EQM}_*(\hat{\theta}_i)]^2 / R$ avec une expression semblable pour $\text{EQM}_*[\text{eqm}(\hat{\theta}_{iK})]$.

Le tableau 1 présente une mesure sommaire des valeurs ER en pourcentage, |BR| et cv pour les cas (1) et (2) d'après la première stratégie. On trouvera au tableau 2 des mesures sommaires pour la deuxième stratégie. Les mesures sommaires considérées sont la moyenne et la médiane (med) pour les petites régions $i = 1, 2, ..., 30$.

Il est clair, d'après les tableaux 1 et 2, que $\hat{\theta}_{iK}$ et $\hat{\theta}_i$ fonctionnent de façon semblable relativement à une ER qui diminue à mesure que σ_v / σ augmente. Dans le cadre de la deuxième stratégie, l'ER est grande à la fois pour le premier et le deuxième cas lorsque $\sigma_v / \sigma \leq 0{,}4$, tandis qu'elle diminue nettement dans le cadre de la première stratégie si le modèle de liaison est enfreint (deuxième cas); l'estimateur direct $\bar{y}_{iw}$ est plutôt instable dans le cadre de la deuxième stratégie.

Pour ce qui est du rendement des estimateurs de l'EQM dans le cadre de la première stratégie, le tableau 1 indique que le |BR| de l'$\text{EQM}(\hat{\theta}_i)$ est négligeable (< 4%) lorsque le modèle de liaison reste vrai (premier cas) et qu'il est petit (< 10%) bien qu'il augmente même lorsque le modèle de liaison est enfreint. L'estimateur de l'$\text{eqm}(\hat{\theta}_{iK})$ comporte un |BR| plus grand mais inférieur à 15%. Le cv de l'$\text{eqm}(\hat{\theta}_i)$ est beaucoup plus petit que le cv de l'$\text{eqm}(\hat{\theta}_{iK})$ tant pour le premier que pour le deuxième cas. Ainsi, lorsque le modèle est vrai (premier cas), le cv médian est de 25% pour l'$\text{eqm}(\hat{\theta}_i)$ comparativement à 148% pour l'$\text{eqm}(\hat{\theta}_{iK})$ lorsque $\sigma_v = 1$; le cv médian tombe à 8% pour l'$\text{eqm}(\hat{\theta}_i)$ comparativement à 48% pour l'$\text{eqm}(\hat{\theta}_{iK})$ lorsque $\sigma_v = 2$. Ce régime se maintient lorsque le modèle est

enfreint (deuxième cas). À noter que la probabilité d'une valeur négative de l'eqm($\hat{\theta}_{iK}$) est assez grande (>0,3) lorsque $\sigma_v / \sigma \leq 0,4$.

Tableau 1
Efficacité relative (ER) des estimateurs, du biais relatif (|BRI) absolu et du coefficient de variation (CV) des estimateurs de l'EQM ($\sigma = 5,0$, $n = 20$): première stratégie

σ_v		ER%		|BRI|%		CV%	
		$\hat{\theta}_{iK}$	$\hat{\theta}_i$	eqm($\hat{\theta}_{iK}$)	eqm($\hat{\theta}_i$)	eqm($\hat{\theta}_{iK}$)	eqm($\hat{\theta}_i$)
				Premier cas			
1	Moyenne	190	177	15,3	3,5	148	25
	Médiane	190	182	14,8	2,6	148	25
2	Moyenne	126	123	5,1	3,2	48	8
	Médiane	127	124	5,6	2,9	48	8
3	Moyenne	113	111	3,5	2,7	35	6
	Médiane	112	111	3,2	3,0	35	6
				Deuxième cas			
1	Moyenne	108	103	10,4	7,9	39	6
	Médiane	108	104	11,1	7,7	38	5
2	Moyenne	108	104	13,3	8,9	39	6
	Médiane	108	104	13,6	7,9	37	6
3	Moyenne	104	103	11,5	7,2	37	5
	Médiane	105	105	13,1	8,0	36	6

Cas 1: μ_i=50, i=1,2,...,30; Cas 2: μ_i=50, i=1,2,...,10; μ_i=55, i=11,12,...,20; μ_i=60, i=21,22,...,30.

Tableau 2
Efficacité relative (ER) des estimateurs, du biais relatif (|BRI) absolu et du coefficient de variation (CV) des estimateurs de l'EQM (σ=5,0, n=20): deuxième stratégie

σ_v		ER%		|BRI|%		CV%	
		$\hat{\theta}_{iK}$	$\hat{\theta}_i$	eqm($\hat{\theta}_{ik}$)	eqm($\hat{\theta}_i$)	eqm($\hat{\theta}_{iK}$)	eqm($\hat{\theta}_i$)
				Premier cas			
1	Moyenne	283	281	14,2	25,4	289	39
	Médiane	275	279	15,0	24,7	295	38
2	Moyenne	180	182	7,3	19,2	115	24
	Médiane	177	181	6,9	18,7	122	23
3	Moyenne	129	129	4,8	14,8	68	24
	Médiane	129	128	4,2	13,9	65	24
				Deuxième cas			
1	Moyenne	278	276	15,7	26,8	291	41
	Médiane	271	275	16,6	26,2	297	40
2	Moyenne	175	177	8,8	20,7	117	26
	Médiane	173	177	8,5	20,3	124	25
3	Moyenne	124	124	6,3	16,2	70	25
	Médiane	125	124	6,8	15,5	67	26

Cas 1: μ_i=50, i=1,2,...,30; Cas 2: μ_i=55, i=1,2,...,10; μ_i=55, i=11,12,...,20; μ_i =60, i=21,22,...,30.

Pour ce qui est de la deuxième stratégie, le tableau 2 indique que le |BRI de l'eqm($\hat{\theta}_i$) est plus grand que la valeur relevant de la première stratégie et qu'il varie entre 15% et 25%. Par contre, le |BRI de l'eqm($\hat{\theta}_{iK}$) est plus petit et varie entre 4% et 15%. Le cv de l'eqm(θ_{iK}), par contre, est beaucoup plus grand que pour la première stratégie. Ainsi, le cv médian pour le premier cas est de 295% comparativement à 38% pour l'eqm($\hat{\theta}_i$) lorsque σ_v = 1 avec diminution à 122% comparativement à 23% lorsque σ_v = 2. Un régime semblable est observé pour le deuxième cas, la population finie fixe étant établie en fonction du modèle avec des moyennes qui varient.

Afin de réduire le |RB| de l'eqm($\hat{\theta}_i$) dans le cadre de la deuxième stratégie, on pourrait le combiner à l'eqm($\hat{\theta}_{iK}$) en établissant une moyenne pondérée, mais le choix des poids appropriés semble difficile. La moyenne pondérée sera plus stable que l'eqm($\hat{\theta}_{iK}$).

5. MODÈLE DE RÉGRESSION À ERREUR EMBOÎTÉE

On peut étendre les résultats des sections 2 et 3 aux modèles de régression à erreur emboîtée

$$y_{ij} = x'_{ij}\beta + v_i + e_{ij},\ j = 1, 2, ..., n_i;\ i = 1, 2, ..., m \qquad (17)$$

en utilisant les résultats de Prasad et Rao (1990), où x_{ij} est un vecteur p des variables auxiliaires de moyenne de population connue $\bar{X}_i$ et liées à y_{ij}, et β est le vecteur p des coefficients de régression. Le modèle réduit est donné d'après

$$\bar{y}_{iw} = \bar{x}'_{iw}\beta + v_i + \bar{e}_{iw} \qquad (18)$$

avec $\bar{x}'_{iw} = \sum_j w_{ij} x_{ij}$. On obtient des estimations conformes au modèle $\hat{\sigma}^2_v$ et $\hat{\sigma}^2$ du modèle au niveau de l'unité (17), à l'aide de la méthode de l'ajustement des constantes (Prasad et Rao 1990) ou encore de l'estimation PMR (probabilité maximale restreinte) (Datta et Lahiri 1997).

La MPLNB pseudo-empirique de $\theta_i = \bar{X}'_i\beta + v_i$ est donnée d'après

$$\hat{\theta}_i = \hat{\gamma}_{iw}\bar{y}_{iw} + (1 - \hat{\gamma}_{iw})\bar{X}'_i\hat{\beta}_w, \qquad (19)$$

où

$$\hat{\beta}_w = \left(\sum_i \hat{\gamma}_{iw}\bar{x}_{iw}\bar{x}'_{iw}\right)^{-1}\left(\sum_i \hat{\gamma}_{iw}\bar{x}_{iw}\bar{y}_{iw}\right).$$

Un estimateur approximatif non biaisé pour le modèle de l'EQM($\hat{\theta}_i$) est donné d'après (15) avec

$$g_{1i}(\hat{\sigma}^2_v, \hat{\sigma}^2) = (1 - \hat{\gamma}_{iw})\hat{\sigma}^2_v$$

comme auparavant,

$$g_{2i}(\hat{\sigma}^2_v, \hat{\sigma}^2) = \hat{\sigma}^2_v(\bar{X}_i - \hat{\gamma}_{iw}\bar{x}_{iw})'\left(\sum_i \hat{\gamma}_{iw}\bar{x}_{iw}\bar{x}'_{iw}\right)^{-1}(\bar{X}_i - \hat{\gamma}_{iw}x_{iw}\bar{x}_{iw})$$

et $g_{3i}(\hat{\sigma}^2_v, \hat{\sigma}^2)$, que l'on obtient de (14), comportant les variances et covariances estimatives de $\tilde{\sigma}^2_v$ et de $\hat{\sigma}^2$. On peut obtenir ce dernier élément de Prasad et Rao (1990) pour la méthode de l'ajustement des constantes et de Datta et Lahiri (1997) pour la probabilité maximale restreinte.

6. CONCLUSION

Nous avons proposé un estimateur modélisé d'une moyenne régionale dans le cadre d'un modèle simple à effets aléatoires au niveau de l'unité. Cet estimateur dépend des poids d'enquête et il est conforme au plan. Nous avons également obtenu un estimateur à base de modèle de l'EQM. Les résultats de notre étude de simulation ont indiqué que l'estimateur de l'EQM proposé fonctionne bien, même pour des écarts modérés du modèle de liaison. La stratégie proposée s'étend aussi à un modèle de régression à erreur emboîtée.

REMERCIEMENTS

Les auteurs ont reçu des subventions du Conseil de recherches en sciences naturelles et en génie du Canada. Nous remercions le rédacteur adjoint et l'examinateur de leurs remarques et suggestions constructives.

ANNEXE 1

Preuve de (13):

D'après les résultats généraux (Prasad et Rao 1990), nous avons

$$E(\hat{\theta}_i - \tilde{\theta}_i)^2 \approx tr\left[A_i(\sigma_v^2, \sigma^2) B_i(\sigma_v^2, \sigma^2)\right],$$

où $B_i(\sigma_v^2, \sigma^2)$ est la matrice de covariance 2×2 de $\tilde{\sigma}_v^2$ et de $\hat{\sigma}^2$, et $A_i(\sigma_v^2, \sigma^2)$ est la matrice de covariance 2×2 de

$$\left(\frac{\partial \theta_i^*}{\partial \sigma_v^2}, \frac{\partial \theta_i^*}{\partial \sigma^2}\right).$$

Or en notant que

$$\frac{\partial \theta_i^*}{\partial \sigma_v^2} = \frac{\partial \gamma_{iw}}{\partial \sigma_v^2} \bar{y}_{iw} = \left[\frac{\gamma_{iw}(1 - \gamma_{iw})}{\sigma_v^2}\right] \bar{y}_{iw},$$

$$\frac{\partial \theta_i^*}{\partial \sigma^2} = \frac{\partial \gamma_{iw}}{\partial \sigma^2} \bar{y}_{iw} = -\left[\frac{\gamma_{iw}(1 - \gamma_{iw})}{\sigma^2}\right] \bar{y}_{iw},$$

et $V(\bar{y}_{iw}) = \sigma_v^2 + \delta_i = \sigma_v^2 / \gamma_{iw}$, nous avons

$$A_i(\sigma_v^2, \sigma^2) = \left[\gamma_{iw}(1 - \gamma_{iw})^2 \sigma_v^{-2}\right] \begin{bmatrix} 1 & -\sigma_v^2/\sigma^2 \\ -\sigma_v^2/\sigma^2 & (\sigma_v^2/\sigma)^2 \end{bmatrix},$$

et par conséquent le résultat (14).

Matrice de covariance de $\tilde{\sigma}_v^2$ et de $\hat{\sigma}^2$:

En cas de normalité, nous avons

$$V(\hat{\sigma}^2) = 2\sigma^4 / \left(\sum_i n_i - m\right),$$

$$V(\tilde{\sigma}_v^2) = 2n_*^{-2}$$

$$\left[\sigma^4 (m-1)\left(\sum n_i - 1\right)\left(\sum n_i - m\right)^{-1} + 2n_* \sigma^2 \sigma_v^2 + n_{**} \sigma_v^4\right]$$

et

$$\mathrm{Cov}(\hat{\sigma}^2, \tilde{\sigma}_v^2) = -(m-1) n_*^{-1} V(\hat{\sigma}^2),$$

où

$$n_{**} = \sum n_i^2 - 2\sum n_i^3 \Big/ \sum n_i + \left(\sum n_i^2\right)^2 \Big/ \left(\sum n_i\right)^2;$$

(voir Searle, Cassella et McCulloch 1992, p. 428).

ANNEXE 2

Preuve de $E[g_{1i}(\tilde{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2)] \approx g_{1i}(\sigma_v^2, \sigma^2)$:

Grâce à un développement de Taylor de $g_{1i}(\tilde{\sigma}_v^2, \hat{\sigma}^2)$ autour de (σ_v^2, σ^2) au second ordre et en notant que $E(\hat{\sigma}^2 - \sigma^2) = 0$ et $E(\tilde{\sigma}_v^2 - \sigma_v^2) \approx 0$, nous obtenons

$$E\left[g_{1i}(\tilde{\sigma}_v^2, \hat{\sigma}^2) - g_{1i}(\sigma_v^2, \sigma^2)\right] \approx \frac{1}{2} tr\left[D_i(\sigma_v^2, \sigma^2) B_i(\sigma_v^2, \sigma^2)\right],$$

où $D_i(\sigma_v^2, \sigma^2)$ est la matrice 2×2 des dérivées secondes de $g_{1i}(\sigma_v^2, \sigma^2)$ relativement à σ_v^2 et à σ^2. Il est facile de vérifier que

$$\frac{1}{2}\ tr\left[D_i(\sigma_v^2, \sigma^2) B_i(\sigma_v^2, \sigma^2)\right] = g_{3i}(\sigma_v^2, \sigma^2).$$

Or en notant que $E[g_{3i}(\tilde{\sigma}_v^2, \hat{\sigma}^2)] \approx g_{3i}(\sigma_v^2, \sigma^2)$ nous obtenons le résultat souhaité.

ANNEXE 3

L'estimateur à base de plan de la variance de $\bar{y}_{iw}$ dans le cadre d'un échantillonnage avec PPT est donné d'après

$$v(\bar{y}_{iw}) = \frac{m}{m-1} \sum_j w_{ij}^2 (y_{ij} - \bar{y}_{iw})^2.$$

L'estimateur modélisé de la variance de Kott (1989) est

$$v^*(\bar{y}_{iw}) = \left\{V(\bar{y}_{iw}) / Ev(\bar{y}_{iw})\right\} v(\bar{y}_{iw}) = \frac{\left(\sum_j w_{ij}^2\right) \sum_j w_{ij}^2 (y_{ij} - \bar{y}_{iw})^2}{\sum_j w_{ij}^2 \left(1 - 2w_{ij} + \sum_j w_{ij}^2\right)},$$

où E et V désignent l'espérance et la variance pour ce qui est du modèle au niveau de l'unité (3).

BIBLIOGRAPHIE

ARORA,V., et LAHIRI, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

BATTESE, G.E., HARTER, R., et FULLER, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

DATTA, G.S., et LAHIRI, P. (1997). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictor in Small-Area Estimation Problems. Rapport technique, University of Nebraska-Lincoln.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GHOSH, M., et MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 81, 1058-1069.

GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

KACKAR, R.N., et HARVILLE, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects models. *Journal of the American Statistical Association*, 79, 853-862.

KLEFFE, J., et RAO, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *Journal of Multivariate Analysis*, 43, 1-15.

KOTT, P. (1989). Estimation robuste pour petits domaines à l'aide du modèle des effets aléatoires. *Techniques d'enquête*, 15, 3-13.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

SEARLE, S.R., CASELLA, G., et McCULLOCH, C.E. (1992). *Variance Components*. New York: John Wiley and Sons.

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 81-89
Statistique Canada

Production d'estimations régionales à partir de modèles multiniveau

FERNANDO A.S. MOURA et DAVID HOLT[1]

RÉSUMÉ

Dans cet article, nous utilisons un modèle général à niveaux multiples pour obtenir des estimations régionales à partir des données d'enquête. Ce type de modèle permet de tenir compte des variations entre les régions, en raison: (i) des différences dans la distribution des variables unitaires entre les régions; (ii) des différences dans la distribution des variables régionales entre les régions et (iii) des composantes de la variance propres à la région, qui permettent de tenir compte de la variation locale additionnelle qui ne peut être expliquée par les covariables unitaires ou régionales. Des estimateurs régionaux sont calculés pour ce modèle multiniveau et nous fournissons une approximation de l'erreur quadratique moyenne (EQM) de chaque estimation régionale, pour cette catégorie générale de modèles mixtes, ainsi qu'un estimateur de cette EQM. L'approximation de la EQM et l'estimateur de la EQM tiennent compte de trois sources de variation: (i) la prévision de l'erreur quadratique moyenne en présumant que les termes fixes et les composantes de la variance dans le modèle à niveaux multiples sont connus; (ii) la composante additionnelle du fait qu'il faille estimer les coefficients fixes et (iii) la composante additionnelle du fait que les composantes de la variance dans le modèle doivent être estimées. Les estimations par les méthodes proposées sont obtenues à partir d'un grand ensemble de données qui sert de base aux études numériques. Les résultats obtenus confirment que les composantes additionnelles de la variance incluses dans les modèles à niveaux multiples, de même que les covariables régionales, peuvent améliorer les estimations régionales et que l'approximation de la EQM et l'estimateur de la EQM sont satisfaisants.

MOTS CLÉS: Estimation régionale; modèles mixtes; modèles multiniveau; meilleur estimateur linéaire empirique sans biais.

1. INTRODUCTION

Le besoin d'établir des estimations régionales (et des estimations par petit domaine) à partir des données d'enquête n'est pas nouveau. La production de telles estimations est toutefois difficile du fait que, pour la plupart voire la totalité des petites régions, l'échantillon obtenu dans le cadre d'une enquête nationale est de trop petite taille pour permettre l'établissement d'estimations directes qui soient d'un degré de précision acceptable. Les premières tentatives visant à résoudre ce problème au moyen de méthodes telles que l'estimation synthétique (Gonzalez 1973) étaient basées sur l'utilisation d'informations auxiliaires et la mise en commun de l'information recueillie dans diverses régions. Ghosh et Rao (1994) présentent une excellente rétrospective et bibliographie sur cette question.

Les études empiriques ont révélé que de telles méthodes ne tenaient pas suffisamment compte de la variation locale et donc que les estimations régionales ainsi obtenues étaient trop réduites vers une moyenne prévue. Des méthodes plus récentes (par exemple Battese et Fuller 1981 et Battese, Harter et Fuller 1988) reposent sur l'utilisation de certaines composantes du modèle de la variance, ou l'équivalent, pour tenir compte de la variation locale; des études empiriques ont démontré la supériorité de cette dernière approche (par exemple Prasad et Rao 1990).

Nous proposons ici un modèle général à niveaux multiples pour la production des estimations régionales, lequel offre la possibilité d'utiliser des données auxiliaires, à la fois au niveau unitaire et régional. Il permet également de considérer que tout paramètre de régression, plutôt que seulement l'ordonnée à l'origine comme le proposent Battese et Fuller (1981), varie de façon aléatoire entre les régions. On tient compte de la variation locale en utilisant les différences entre les moyennes des variables auxiliaires unitaires, des variables régionales et diverses composantes de la variance qui permettent de tenir compte des variations entre les régions.

À partir de ce modèle général, nous obtenons le prédicteur régional. De plus, une approximation de l'erreur quadratique moyenne de chaque prévision régionale distincte ainsi qu'un estimateur de cette erreur sont élaborés.

L'étude numérique, basée sur un grand fichier obtenu du Brésil, montre que de tels modèles peuvent être utiles pour prévoir les estimations régionales. Nous examinons également la robustesse de l'approche proposée en regard de l'erreur de spécification de la matrice variance-covariance des effets aléatoires régionaux et de l'erreur de spécification des covariables régionales. D'autres résultats numériques viennent confirmer l'efficacité de l'approximation de l'erreur quadratique moyenne et de son estimateur.

2. MODÈLE À NIVEAUX MULTIPLES

2.1 Introduction

Nous examinons ici le modèle à niveaux multiples suivant pour la prévision des moyennes régionales:

[1] Fernando A.S. Moura, Instituto de Matemática, UFRJ, Rio de Janeiro, Brazil, CP: 68530, CEP: 21941-590, courrier électronique: fmoura@dme.ufrj.br; David Holt, Office for National Statistics, 1 Drummond Gate, London, SW1P 2QQ, courrier électronique: tholt@ons.gov.uk.

$$Y_i = X_i\beta_i + \varepsilon_i$$
$$\beta_i = Z_i\gamma + v_i \quad i = 1,...,m \tag{2.1}$$

où Y_i est le vecteur de longueur n_i pour la caractéristique à l'étude pour les unités d'échantillonnage dans la i-ième région, $i = 1, ..., m$; X_i est la matrice des variables explicatives au niveau de l'unité d'échantillonnage; Z_i est la matrice de plan des variables régionales; γ est le vecteur de longueur q des coefficients fixes et $v_i = (v_{i0},...,v_{ip})^T$ est le vecteur de longueur $(p + 1)$ des effets aléatoires pour la i-ième région. Nous formulons les hypothèses suivantes au sujet de la distribution des vecteurs aléatoires: (a) les v_i sont indépendants entre les régions et ont une distribution dépendant de plusieurs variables à l'intérieur de chaque région avec $E(v_i) = 0$ et $V(v_i) = \Omega$, (b) les ε_i's et v_i's sont indépendants et $V(\varepsilon_i) = \sigma^2 I$.

Pour l'ensemble de la population, l'équation (2.1) s'applique en remplaçant n_i par N_i, la taille de la population de la région.

La série de m équations en (2.1) peut être représentée de façon concise, comme suit:

$$Y = XZ_\gamma + Xv + \varepsilon. \tag{2.2}$$

Il convient de souligner que le modèle avec ordonnée à l'origine aléatoire (voir section 2.3) peut être considéré comme un cas spécial du modèle (2.1) où Z_i est égal à la matrice unité pour chaque région et où Ω contient des termes qui sont tous contraints à zéro, sauf celui qui correspond à la variance du terme de l'ordonnée à l'origine. Il existe d'autres modèles intermédiaires, par exemple celui où Ω est diagonal de sorte que les coefficients de régression régionaux sont aléatoires mais non corrélés entre les covariables.

Holt (dans Ghosh 1994, page 82) fait remarquer que l'avantage du modèle (2.1) par rapport à ses concurrents vient de ce que le premier permet d'intégrer efficacement l'utilisation de covariables de niveaux unitaire et régional dans un même modèle. De plus, l'utilisation d'effets aléatoires additionnels pour les coefficients de régression permet une plus grande latitude dans les cas où il ne convient pas de présumer que les mêmes pentes s'appliquent à toutes les régions.

2.2 Estimation des paramètres fixes et des paramètres des composantes de la variance

Dans le modèle (2.1), les paramètres fixes et les paramètres des composantes de la variance sont représentés respectivement par γ et $\theta = ([\text{Vech}(\Omega)]^T, \sigma^2)^T$. Il existe plusieurs méthodes pour estimer ces paramètres avec un modèle linéaire mixte général; la plupart d'entre elles, qui sont basées sur des algorithmes itératifs, produisent l'estimateur du maximum de vraisemblance (EMV) ou l'estimateur restreint du maximum de vraisemblance dans certaines conditions de régularité.

Goldstein (1986) montre comment obtenir des estimateurs cohérents par l'application de méthodes itératives par les moindres carrés généralisés (MCG). Il démontre également l'équivalence de cette méthode à l'estimateur du maximum de vraisemblance, dans des conditions de normalité. Goldstein (1989) a par la suite proposé une légère modification de son algorithme (à savoir la méthode itérative restreinte par les moindres carrés généralisés), laquelle équivaut à l'estimateur restreint du maximum de vraisemblance dans des conditions de normalité. Contrairement à la méthode itérative par les moindres carrés généralisés, la méthode d'estimation itérative restreinte par les moindres carrés généralisés fournit des estimations sans biais des paramètres des composantes de la variance, en tenant compte de la diminution des degrés de liberté due à l'estimation des paramètres fixes.

Les présents travaux se limitent à la méthode itérative restreinte des MCG, comme dans Goldstein (1989), laquelle méthode est décrite en détail à l'annexe A.

2.3 Estimateur de la moyenne régionale

En partant du modèle (2.1) et en présumant que la taille de la population, N_i, dans la i-ième région est grande, nous pouvons représenter la moyenne pour la i-ième région par l'équation suivante:

$$\mu_i = \bar{X}_i^T Z_i \gamma + \bar{X}_i^T v_i \tag{2.3}$$

où $\bar{X}_i$ est le vecteur $(p + 1)$ de la moyenne de la population pour la i-ième région.

On peut obtenir un estimateur de μ_i en incorporant les estimateurs de γ et θ, obtenus par la méthode itérative restreinte des MCG, dans les termes respectifs de l'équation (2.3), où le prédicteur de l'effet aléatoire v_i pour la i-ième région est représenté par $\hat{v}_i = \hat{\Omega} X_i^T \hat{V}_i^{-1} (Y_i - X_i Z_i \hat{\gamma})$, où $\hat{V}_i^{-1} = \hat{\sigma}^{-2} I - \hat{\sigma}^{-4} X_i \hat{\Omega}\, \hat{G}_i^{-1} X_i^T$ et $\hat{G}_i^{-1} = (I + \hat{\sigma}^{-2} X_i^T X_i \hat{\Omega})^{-1}$.

Cet estimateur de μ_i est connu sous le nom du meilleur estimateur linéaire empirique sans biais

$$\hat{\mu}_i = \bar{X}_i^T Z_i \hat{\gamma} + \bar{X}_i^T \hat{v}_i. \tag{2.4}$$

Battese et coll. (1981, 1988) proposent et utilisent un modèle avec ordonnée à l'origine aléatoire pour calculer les estimations régionales. Dans ce dernier cas, le meilleur estimateur empirique linéaire sans biais est représenté par

$$\hat{\mu}_{i(\text{RI})} = \bar{X}_i^T \hat{\beta} + \hat{v}_{i0}.$$

Dans le modèle avec ordonnée à l'origine aléatoire, seule l'ordonnée à l'origine de chaque région est aléatoire et les autres composantes de β demeurent fixes.

2.4 Approximation de l'erreur quadratique moyenne (EQM)

Kackar et Harville (1984) montrent que, si $\hat{\theta}$ est un estimateur invariant sous la translation de θ et que les termes aléatoires sont distribués normalement, alors l'erreur

quadratique moyenne d'un prédicteur d'une combinaison linéaire des effets fixes et aléatoires peut être décomposée en deux termes. Le premier est dû à la variabilité dans l'estimation des paramètres fixes lorsque les composantes de la variance sont connues et le deuxième vient de l'estimation des composantes de la variance.

Comme dans les conditions de normalité l'estimateur par la méthode itérative restreinte des MCG équivaut à l'estimateur du maximum de vraisemblance restreint (EMVR) et que l'EMVR est invariant sous la translation, les résultats de Kackar et Harville (1984) peuvent être appliqués aux estimateurs des moyennes régionales $\hat{\mu}_i$, $i=1, ..., m$:

$$\text{EQM}(\hat{\mu}_i) = E[\hat{\mu}_i - \mu_i]^2 = E[\tilde{\mu}_i - \mu_i]^2 + E[\hat{\mu}_i - \tilde{\mu}_i]^2 \quad (2.5)$$

où $\tilde{\mu}_i$ est le meilleur estimateur linéaire sans biais de μ_i.

Le premier terme de (2.5), c'est-à-dire EQM$[\tilde{\mu}_i]$, peut être calculé directement, comme suit

$$\text{EQM}(\tilde{\mu}_i) = \bar{X}_i^T (G_i^{-1})^T \Omega \bar{X}_i +$$

$$\sigma^2 \bar{X}_i (G_i^{-1})^T Z_i \left(\sum_{i=1}^{m} Z_i^T G_i^{-1} X_i^T X_i Z_i \right)^{-1} Z_i^T G_i^{-1} \bar{X}_i \quad (2.6)$$

où $G_i = I + \sigma^2 X_i^T X_i \Omega$. Kackar et Harville (1984) font remarquer que le deuxième terme de (2.6) est impossible à résoudre, sauf dans des cas spéciaux, et ils en proposent donc une approximation. Prasad et Rao (1990) proposent eux aussi une approximation du deuxième terme et fournissent des précisions sur cette approximation, pour trois cas particuliers: le modèle avec ordonnée à l'origine aléatoire, le modèle des coefficients de régression aléatoires et le modèle de Fay-Herriot. Ils définissent également certaines conditions de régularité pour que leur approximation soit de deuxième ordre et prouvent que leur approximation de l'erreur quadratique moyenne pour le modèle de Fay-Herriot est de second ordre. Il semble toutefois plus difficile d'énoncer des conditions générales pour des modèles plus complexes comme celui proposé en (2.1).

Une approximation du deuxième terme de (2.5) est présentée à l'annexe B, à partir de l'approche de Prasad et Rao.

Il convient de souligner que l'approximation de l'erreur quadratique moyenne (EQM$\hat{\mu}_i$) peut être décomposée en trois termes, comme suit:

$$\text{EQM}(\hat{\mu}_i) \approx T_1 + T_2 + T_3 \quad (2.7)$$

où T_1 et T_2 représentent respectivement les premier et deuxième termes de l'équation (2.6) et T_3 est décrit à l'annexe B.

Le terme T_1 est la variabilité de $\hat{\mu}_i$ lorsque tous les paramètres sont connus; le deuxième terme, T_2, est dû à l'estimation des effets fixes et le troisième terme, T_3, vient de l'estimation des composantes de la variance.

Lorsque les taux d'échantillonnage ne sont pas négligeables, les estimateurs des moyennes régionales peuvent être définis en s'inspirant du modèle de la population finie, en faisant des prévisions spécifiquement pour les unités non échantillonnées:

$$\hat{\mu}_i^F = f_i \bar{y}_i + (\bar{X}_i - f_i \bar{x}_i)^T (Z_i \hat{\gamma} + \hat{v}_i) \quad (2.8)$$

où l'exposant F indique qu'une correction pour le taux d'échantillonnage de la population finie f_i a été utilisée; $\bar{x}_i$ est le vecteur $(p+1)$ des moyennes de l'échantillon.

La EQM$(\hat{\mu}_i^F)$ peut être calculée à partir de:

$$\hat{\mu}_i^F - \bar{Y}_i = (1 - f_i) \left[\left(\bar{X}_i^C \right)^T \left(Z_i (\hat{\gamma} - \gamma) + \hat{v}_i - v_i - \bar{\varepsilon}_i^C \right) \right]$$

où $\bar{X}_i^C = (1 - f_i)^{-1} (\bar{X}_i - f_i \bar{x}_i)$ et $\bar{\varepsilon}_i^C$ est la moyenne de ε_{ij} pour les unités non échantillonnées dans la i-ième région. Par conséquent,

$$\text{EQM}(\hat{\mu}_i^F) = (1 - f_i)^2 \left[\text{EQM}^*(\hat{\mu}_i) + N_i^{-1} (1 - f_i)^{-1} \sigma^2 \right] \quad (2.9)$$

où EQM$^*(\hat{\mu}_i)$ est l'équation (2.7) dans laquelle $\bar{X}_i$ est remplacé par $\bar{X}_i^C$.

2.5 Estimation de l'erreur quadratique moyenne

Il est courant d'estimer l'erreur quadratique moyenne d'une combinaison linéaire des effets fixes et aléatoires d'un modèle mixte comme celui présenté en (2.1), en remplaçant les estimations des composantes de la variance dans l'expression respective de la EQM. Cet estimateur ne tient pas compte de l'effet de l'estimation des paramètres des composantes de la variance sur l'erreur quadratique moyenne. Selon plusieurs études (voir par exemple Singh, Stukel et Pfeffermann 1998 ou Harville et Jeske 1992), une telle procédure tend à sous-estimer l'erreur quadratique moyenne. Prasad et Rao (1990) font état d'une étude de simulation qui indique que l'utilisation de cet estimateur «naïf» introduit un biais par défaut important. Ils montrent également que, pour le modèle de Fay-Herriot (cas spécial du modèle (2.1)), en utilisant les estimations «tronquées de Henderson» pour les composantes de la variance, l'on obtient

$$E(\hat{T}_1) = T_1 - T_3 + o(m^{-1}); E(\hat{T}_2) = T_2 + o(m^{-1});$$

$$E(\hat{T}_3) = T_3 + o(m^{-1}).$$

Harville et Jeske (1992) définissent certaines conditions pour l'élimination du biais dans l'estimateur de l'erreur quadratique moyenne de Prasad et Rao. Cependant, avec le modèle plus général (2.1), il semble là encore plus difficile de définir les conditions générales pour lesquelles l'ordre du biais de l'estimateur de Prasad et Rao est de $o(m^{-1})$, en particulier si les estimations des paramètres sont calculées à partir de méthodes itératives comme la méthode itérative restreinte des MCG.

Malgré tout, nous appuyant sur les résultats de l'étude de simulation résumés à la section 3.4 et sur la vaste étude de simulation décrite par Moura (1994), nous proposons d'utiliser un estimateur similaire à celui de Prasad et Rao pour calculer EQM$(\hat{\mu}_i)$:

$$\hat{EQM} = \hat{T}_1 + \hat{T}_2 + 2\hat{T}_3 \qquad (2.10)$$

où $\hat{T}_i$ sont calculés à partir de (2.5), en remplaçant σ^2 et Ω par leurs estimateurs respectifs obtenus par la méthode itérative restreinte des MCG.

À partir de l'équation (2.9), nous pouvons également calculer un estimateur pour $EQM(\hat{\mu}_i^F)$ comme suit:

$$\hat{EQM}(\hat{\mu}_i^F) = (1-f_i)^2[\hat{EQM}^*(\hat{\mu}_i) + N_i^{-1}(1-f_i)^{-1}\hat{\sigma}^2] \qquad (2.11)$$

où $\hat{EQM}^*(\hat{\mu}_i)$ est l'équation (2.10) dans laquelle $\bar{X}_i$ a été remplacé par $\bar{X}_i^C$.

3. ÉTUDE NUMÉRIQUE BASÉE SUR MODÈLE

3.1 Comparaison des estimateurs

Afin d'examiner les propriétés de divers estimateurs, nous avons utilisé les données de 38 740 ménages des districts de recensement d'un comté, au Brésil. Le revenu du chef du ménage a été utilisé comme variable dépendante. Deux variables indépendantes de niveau unitaire ont été définies, soit le niveau de scolarité du chef du ménage (échelle ordinale de 0 à 5) et le nombre de pièces dans le domicile (1-11+).

Le modèle provisoire se lit comme suit:

$$Y_{ij} = \beta_{i0} + \beta_{i1}x_{1ij} + \beta_{i2}x_{2ij} + \varepsilon_{ij} \quad i = 1,...,m; j = 1,...,N_i$$

$$\beta_{i0} = \gamma_{00} + \nu_{i0};\ \beta_{i1} = \gamma_{10} + \nu_{i1};\ \beta_{i2} = \gamma_{20} + \nu_{i2} \qquad (3.1)$$

où x_1 et x_2 représentent respectivement le nombre de pièces et le niveau de scolarité du chef du ménage (centrés autour de leur moyenne de population respective).

Les valeurs du paramètre pour le modèle d'ajustement et leurs erreurs-types respectives sont:

$\gamma_{00} = 8{,}456(0{,}108)$ $\gamma_{10} = 1{,}223(0{,}046)$ $\gamma_{20} = 2{,}596(0{,}086)$
$\sigma_{00} = 1{,}385(0{,}194)$ $\sigma_{01} = 0{,}354(0{,}66)$ $\sigma_{02} = 0{,}492(0{,}117)$
$\sigma_{12} = 0{,}333(0{,}054)$ $\sigma_{11} = 0{,}234(0{,}35)$ $\sigma_{22} = 0{,}926(0{,}124)$
$\sigma^2 = 47{,}74(0{,}345)$

Pour procéder aux études numériques à l'intérieur du cadre basé sur modèle, une simulation a été faite en maintenant constants les identificateurs du district de recensement et les valeurs des deux variables explicatives (X). Au départ, les moyennes de la population régionale $\bar{X}_{1i}$ et $\bar{X}_{2i}$ ont été calculées pour l'ensemble du fichier de données et un sous-échantillon choisi au hasard et correspondant à 10 % des enregistrements de chaque région a été défini. Le même sous-ensemble a été conservé pour toutes les simulations (sous-ensemble de simulation).

La production de données pour les simulations a été faite en deux étapes, à partir d'un des modèles de génération de données suivants – modèle général (G), modèle diagonal (D) ou modèle avec ordonnée à l'origine aléatoire (A) – selon les besoins. Dans le premier cas, les valeurs des paramètres ont été obtenues à partir des estimations précitées. Dans le deuxième cas, les termes hors-diagonal ont été fixés à zéro et, dans le troisième cas, seul $\sigma_{00} = 1{,}385$ était différent de zéro.

La première étape du processus de simulation des données consistait à générer les termes aléatoires de niveau 2 (c'est-à-dire les éléments différents de zéro de ν_{i0}, ν_{i1} et ν_{i2}) selon le modèle de génération de données choisi. Ces termes aléatoires étaient distribués normalement (distribution normale multi-dimensionnelle dans le cas du modèle général de génération de données et du modèle diagonal). À ce stade-ci, la valeur prévue de la moyenne pour la i^e région, sur la base des effets aléatoires régionaux obtenus à partir du modèle m_1 = G, D, A dans la r-ième simulation, pourrait être représentée par

$$\mu_{im_1}^{(r)} = \beta_{0i}^{(r)} + \beta_{1i}^{(r)}\bar{X}_{1i} + \beta_{2i}^{(r)}\bar{X}_{2i}.$$

À la deuxième étape du processus de simulation des données, des valeurs unitaires (Y_{ij}) ont été créées pour chacun des modèles de génération de données. Après avoir généré les données pour le sous-ensemble de simulation à partir d'un des trois modèles de génération, les trois modèles d'estimation (G, D et A) pourraient être ajustés en fonction des données simulées, pour obtenir les estimations des paramètres et les prédicteurs pour les moyennes régionales.

Pour chaque modèle de génération de données m_1=G, D, A, l'ensemble du processus de simulation a été répété R=5 000 fois pour obtenir une série de moyennes régionales $\mu_{im_1}^{(r)}$ et de moyennes prévues $\hat{\mu}_{im_1,m_2}^{(r)}$, $r = 1,...,R$ pour chaque région, i, $i = 1,...,m$ et pour les trois modèles d'estimation: m_2 = G, D, A. Pour chaque région et pour les données générées en vertu du modèle m_1 = G, D, A, l'erreur quadratique moyenne (EQM) du processus de prévision pour chaque modèle d'estimation m_2 peut être définie par

$$EQM[\hat{\mu}_{m_1,m_2}] = R^{-1}\sum_{r=1}^{R}\left(\hat{\mu}_{im_1,m_2}^{(r)} - \mu_{im_1}^{(r)}\right)^2$$

et l'erreur relative absolue (ERA) par:

$$ERA[\hat{\mu}_{m_1,m_2}] = R^{-1}\sum_{r=1}^{R}|\hat{\mu}_{im_1,m_2}^{(r)} - \mu_{im_1}^{(r)}|/\mu_{im_1}^{(r)}.$$

Pour fins de comparaison, nous comparons les propriétés de chaque estimateur à celles de l'estimateur qui est le même que le modèle de génération de données. Nous définissons ainsi le ratio de l'erreur quadratique moyenne (REQM):

$$\text{REQM}_{m_2,m_1} = \left\{\sum_{i=1}^{m} \text{EQM}[\hat{\mu}_{im_1,m_2}]\right\} \Big/ \left\{\sum_{i=1}^{m} \text{EQM}[\hat{\mu}_{im_1,m_1}]\right\} \times 100$$

et le ratio de l'erreur relative absolue (RERA):

$$\text{RERA}_{m_2,m_1} = \left\{\sum_{i=1}^{m} \text{ERA}[\hat{\mu}_{im_1,m_2}]\right\} \Big/ \left\{\sum_{i=1}^{m} \text{ERA}[\hat{\mu}_{im_1,m_1}]\right\} \times 100.$$

On remarquera par ailleurs que, lorsque les données sont générées à partir d'un modèle plus simple (par exemple le modèle avec ordonnée à l'origine aléatoire), on ne constate pas de dégradation appréciable, au plan de l'efficacité ou du biais, dans les procédures d'estimation plus complexes. Par contre, lorsque les données sont obtenues à partir d'un modèle plus complexe, les estimateurs sont de qualité inférieure. On observe cependant une différence beaucoup moindre entre les estimateurs obtenus par les modèles général et diagonal qu'entre ces estimateurs et celui obtenu par le modèle avec ordonnée à l'origine aléatoire. À partir des données du tableau 1, on peut en conclure qu'il est avantageux d'introduire des coefficients aléatoires additionnels, qui vont au-delà du simple modèle avec ordonnée à l'origine aléatoire sans devoir nécessairement aller jusqu'au modèle général intégral.

Tableau 1
Ratios des erreurs quadratiques moyennes et ratios des erreurs relatives absolues (entre parenthèses) pour les trois estimateurs et les trois modèles de génération de données

Estimateur	Modèle de génération de données		
	G	D	A
Général (G)	100,0 (100,0)	101,8 (100,9)	101,2 (100,6)
Diagonal (D)	108,8 (82,6)	100,0 (100,0)	100,2 (100,1)
Ordonnée à l'origine aléatoire (A)	131,9 (176,9)	109,1 (105,6)	100,0 (100,0)

Les mesures sommaires présentées au tableau 1 sont les propriétés moyennes pour l'ensemble des régions. Une analyse approfondie de l'erreur quadratique moyenne (EQM) des estimateurs pour chaque région indique que la EQM pour l'estimateur diagonal est légèrement plus élevée que celle de l'estimateur général, et ce pour toutes les régions, alors que pour l'estimateur obtenu par le modèle avec ordonnée à l'origine aléatoire, un nombre relativement faible de régions présentent une hausse substantielle de la EQM. Une différence similaire ressort entre l'estimateur par le modèle diagonal et celui par le modèle avec ordonnée à l'origine aléatoire, lorsque c'est le modèle de génération de données diagonal qui est utilisé.

3.2 Introduction d'une covariable régionale

Nous tentons ici d'examiner l'incidence de l'introduction d'une covariable régionale Z sur les estimations régionales. Il a malheureusement été impossible, avec le fichier de données utilisé, de définir une seule covariable contextuelle ayant un effet substantiel sur les modèles multiniveau. Malgré tout, le nombre de véhicules automobiles par ménage dans chaque région s'est avéré une covariable utile pour les coefficients aléatoires des pentes aléatoires individuelles pour «Room» et «Edu», mais non pour le terme avec ordonnée à l'origine aléatoire. Nous en sommes arrivés à cette conclusion après une analyse préliminaire de l'ajustement du modèle sur des données réelles. Bien que le «nombre de voitures» ait été la meilleure covariable au niveau régional pour expliquer les variations entre les régions, cette covariable n'était pas aussi puissante, au niveau individuel, que «Room» et «Edu» – les covariables individuelles qui avaient été choisies.

Le modèle qui précède avec la covariable régionale Z peut être représenté par

$$Y_{ij} = \beta_{i0} + \beta_{i1} x_{1ij} + \beta_{i2} x_{2ij} + \varepsilon_{ij} \quad i = 1, \ldots, m; j = 1, \ldots, N_i$$

$$\beta_{i0} = \gamma_{00} + \nu_{i0};\ \beta_{i1} = \gamma_{10} + \gamma_{11} z_i + \nu_{i1};\ \beta_{i2} = \gamma_{20} + \gamma_{21} z_i + \nu_{i2} \quad (3.2)$$

Nous avons présumé que les effets aléatoires régionaux n'étaient pas corrélés, pour éviter un défaut de convergence dans l'étude de simulation.

Le tableau 2 présente les estimations des paramètres et leurs erreurs-types respectives, obtenus en ajustant le modèle diagonal, avec (3.2) et sans (2.1) la covariable Z. Il convient de souligner la réduction significative de toutes les composantes des estimations de la variance, à l'exception de $\hat{\sigma}_{00}$ et $\hat{\sigma}^2$, après l'introduction de la covariable explicative régionale Z.

Tableau 2
Estimations des paramètres et erreurs-types selon le modèle général avec la covariable régionale: Données démographiques

Paramètre	Modèle diagonal avec Z	Modèle diagonal
γ_{00}	8,442(0,112)	8,688(0,136)
γ_{10}	0,451(0,179)	1,321(0,085)
γ_{20}	0,744(0,272)	2,636(0,134)
γ_{11}	3,779(0,507)	-
γ_{22}	1,659(0,323)	-
σ_{00}	0,745(0,308)	0,637(0,303)
σ_{11}	0,237(0,083)	0,471(0,116)
σ_{22}	0,700(0,197)	1,472(0,295)
σ^2	44,00(1,05)	44,01(1,05)

Afin d'étudier l'effet de l'erreur de spécification de la variable Z, la simulation basée sur modèle décrite à la section 3.1 a été appliquée aux deux modèles précités, la génération des données étant faite selon les paramètres

présentés au tableau 2. Le tableau 3 résume les résultats de la simulation.

Tableau 3
Ratios des erreurs quadratiques moyennes et ratios des erreurs relatives absolues (entre parenthèses) pour les estimateurs par le modèle diagonal et le modèle diagonal avec Z, selon les deux modèles respectifs de génération de données

Estimateur	Modèle de génération de données	
	Diagonal	Diagonal avec Z
Diagonal	100,0 (100,0)	110,3 (125,4)
Diagonal avec Z	126,2 (107,5)	100,0 (100,0)

Il convient de souligner que, dans les deux cas, l'utilisation d'un estimateur non approprié entraîne une perte significative d'efficacité. On constate également, à partir d'une analyse individuelle de l'erreur quadratique moyenne pour chaque région, que l'on réalise un gain significatif d'efficacité par l'introduction de la covariable régionale Z dans le modèle diagonal. Pour bon nombre de régions, l'erreur quadratique moyenne dans le modèle diagonal avec Z est nettement inférieure à l'erreur quadratique moyenne de l'estimateur correspondant sans Z. Même pour les quelques régions dans lesquelles l'erreur quadratique moyenne par le modèle diagonal avec Z demeure inchangée ou augmente légèrement avec l'introduction de Z, la différence n'est pas appréciable.

3.3 Comparaisons avec l'estimateur de régression

Un des principaux avantages du modèle multiniveau sur le modèle de régression vient de ce qu'il reconnaît que les groupes (en l'occurrence ici, les régions) ont des caractéristiques en commun; ces groupes ne sont pas complètement indépendants, comme on pourrait s'y attendre, par exemple en utilisant un modèle de régression linéaire distinct pour chaque région. Cependant, compte tenu de la corrélation intraclasse relativement faible observée pour l'ensemble de données utilisé et du fait que chaque région compte en moyenne 28 unités, on pourrait croire que l'utilisation du modèle multiniveau dans le cas présent ne permettrait pas d'améliorer sensiblement les estimateurs régionaux. Il est toutefois intéressant de constater que, même dans ces circonstances, l'estimateur régional par le modèle multiniveau donne dans l'ensemble de meilleurs résultats que l'estimateur de régression synthétique distincte, que ce soit avec le modèle multiniveau ou même avec le modèle de régression. Les résultats présentés au tableau 4 viennent corroborer cette allégation.

Le modèle multiniveau de génération de données est le modèle général avec les paramètres définis à la section 3.1. Les paramètres utilisés dans le modèle de génération par régression ont été obtenus en ajustant une régression distincte pour chaque région.

À la lecture des résultats présentés au tableau 4, on constate une perte d'efficacité substantielle avec l'estimateur par régression distincte par comparaison à l'estimateur par le modèle général, car le premier ne tient pas compte de la différence entre les régions résultant des effets aléatoires régionaux.

Tableau 4
Ratios des erreurs quadratiques moyennes et ratios des erreurs relatives absolues (entre parenthèses) pour les estimateurs par le modèle général et par régression distincte, selon les deux modèles respectifs de génération de données

Estimateur	Modèle de génération de données	
	Général	Régression distincte
Général	100,0 (100,0)	88,1 (83,1)
Régression distincte	247,6 (154,7)	100,0 (100,0)

La figure 1 illustre ces résultats, en représentant graphiquement le ratio entre l'erreur quadratique moyenne de l'estimateur par modèle général et de l'estimateur par régression distincte, pour chaque région. Pour illustrer l'effet de la taille de l'échantillon de la région sur l'efficacité, nous représentons graphiquement l'erreur quadratique moyenne en fonction de la taille de l'échantillon, pour chaque région. Il ressort clairement de la figure 1 que le gain d'efficacité tend à diminuer à mesure que la taille de l'échantillon augmente.

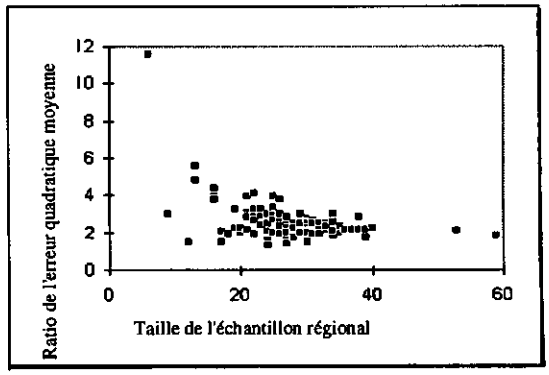


Figure 1. Comparaison entre l'efficacité basée sur modèle de l'estimateur par modèle général et de l'estimateur par régression distincte, pour chaque petite région

3.4 Évaluation de l'approximation de la EQM et de l'estimateur de la EQM

À partir des résultats de la simulation, nous pouvons étudier les propriétés de l'approximation de l'erreur quadratique moyenne (2.7). Dans le cas de l'estimateur par modèle général avec utilisation du modèle général de génération de données, l'approximation de l'erreur quadratique moyenne

semble très bonne. La sous-estimation moyenne de l'approximation a été de 0,31 % de la valeur de EQM, la fourchette variant d'une sous-estimation maximale de 5,4 % de la valeur de EQM à une surestimation maximale de 4,8 % de cette valeur. Dans le cas examiné ici, T_1 explique en moyenne 94,6 % de la variation totale et T_3, un autre 4,3 %. Compte tenu de la forte composante de la variance due à σ^2, ces résultats sont prévisibles. Dans chaque région, la composante T_1 varie entre 87,4 % et 99,1 % du total et T_3, entre 0,7 % et 10,5 % du total. Dans les régions spécifiques, la composante T_2 n'a-t-elle contribué à plus de 2,2 % de l'erreur quadratique moyenne.

Nous avons également évalué la performance de l'estimateur de l'erreur quadratique moyenne, représenté par l'équation (2.10), à celle de l'estimateur «naïf» de l'erreur quadratique moyenne qui ne tient pas compte du dernier terme de l'équation (2.10). La racine carrés de l'EQM de l'estimateur de l'erreur quadratique moyenne proposé varie entre 4,7 % et 32,3 %, pour une moyenne de 17,5 %, tandis qu'elle est en moyenne de 20,9% (5,2 % à 47,5 %) pour l'estimateur naïf. L'estimateur de l'erreur quadratique moyenne est généralement sans biais, tandis que l'estimateur naïf sous-estime la EQM en moyenne de 9,1 %, son biais relatif variant entre -23,5 % et -0,9 %. Nos résultats viennent corroborer ceux d'autres chercheurs; voir Singh, Stukel et Pfeffermann (1998) ainsi que Prasad et Rao (1990), qui montrent que l'estimateur naïf peut présenter un biais important.

4. DISCUSSION

Prasad et Rao (1990) et Battese et coll. (1981, 1988) ont démontré que les modèles qui incluent des composantes de la variance spécifiques des régions peuvent produire des estimateurs régionaux nettement améliorés. Certains des résultats numériques présentés dans ce document montrent que, à l'intérieur du cadre de simulation basé sur modèle, il est possible d'obtenir des estimateurs encore meilleurs en rendant aléatoires les pentes et l'ordonnée à l'origine.

Les conclusions générales qui se dégagent de notre étude basée sur cette série de paramètres s'énoncent comme suit: il est bénéfique d'utiliser une composante du modèle de la variance qui soit plus complexe que l'estimateur par le modèle avec ordonnée à l'origine aléatoire; la surspécification du modèle (par exemple en utilisant l'estimateur général avec les données obtenues par le modèle avec ordonnée à l'origine aléatoire) ne réduit pas de façon significative l'efficacité; l'utilisation de covariables régionales peut améliorer les estimations régionales; et il faudrait privilégier les modèles multiniveau au modèle de régression distincte. L'étude de simulation vient confirmer que l'approximation de l'erreur quadratique moyenne semble précise et que l'estimation de cette erreur est à peu près sans biais, celle-ci reflétant la variation dans l'erreur quadratique moyenne entre les régions; il y aurait lieu toutefois de pousser plus loin les études théoriques afin de déterminer l'ordre exact d'approximation.

Il ne fait aucun doute que l'ajustement du modèle et les diagnostics sont essentiels. Si nous appliquons un modèle mixte général alors qu'il y a mauvais ajustement avec les données, les résultats pourraient être décevants. Des études beaucoup plus poussées devront être menées pour en arriver à comprendre quelles caractéristiques de régions spécifiques sont susceptibles de procurer des gains d'efficacité lorsque des modèles mixtes généraux sont utilisés plutôt que des modèles plus simples.

REMERCIEMENTS

Nous aimerions remercier les examinateurs externes et le rédacteur en chef, pour leurs commentaires utiles au sujet d'une version antérieure du présent document.

ANNEXE A: MÉTHODE GÉNÉRALISÉE ITÉRATIVE RESTREINTE

L'estimateur par les moindres carrés généralisés de γ dans le modèle (2.1) est représenté par

$$\tilde{\gamma} = (Z^T X^T V^{-1} XZ)^{-1} (Z^T X^T V^{-1} Y) = \left(\sum_{i=1}^{m} Z_i^T X_i^T V_i^{-1} X_i Z_i \right)^{-1} \left(\sum_{i=1}^{m} Z_i^T X_i^T V_i^{-1} Y_i \right) \quad \text{(A.1)}$$

où $V = \text{Diag}(V_1, ..., V_m)$ et $V_i = \sigma^2 I + X^T \Omega X$ est la matrice des covariances de Y_i, $i = 1, ..., m$.

On présume toutefois que V est fonction de paramètres inconnus, de sorte qu'on ne peut estimer γ à partir de (A.1). Par contre, si la valeur de γ est connue, alors

$$Y^* = \text{vech}\,[(Y - XZ\gamma)(Y - XZ\gamma)^T] \quad \text{(A.2)}$$

est un estimateur sans biais de vech(V). Qui plus est, vech(V) est une fonction linéaire de θ. Nous pouvons donc considérer le modèle linéaire suivant:

$$Y^* = F\theta + \xi. \quad \text{(A.3)}$$

Où $F = \partial \text{vech}(V)/\partial\theta$ et ξ est une variable aléatoire dont la moyenne est représentée par $O = (0, ..., 0)$ et la covariance de ξ est représentée par $V_\xi = 2\varphi_n(V \otimes V)\varphi_n^T$. La matrice φ_n représente toute transformation linéaire de vec(A) en vech(A), et A est toute matrice $n \times n$ de sorte que vech(A) = $\varphi_n$vec(A) (voir Fuller, 1987 pour plus de détails à ce sujet). Si l'on présume maintenant que F est de rang complet et que V_ξ est connu et non singulier, il peut être démontré que l'estimateur par les moindres carrés généralisés de θ est représenté par

$$\tilde{\theta}_a = \text{cov}(\tilde{\theta}_a)\left(\frac{\partial \text{vec}(V)}{\partial \theta}\right)^T\left(\frac{1}{2}V^{-1}\otimes V^{-1}\right)\text{vec}(\tilde{Y}\tilde{Y}^T) \quad \text{(A.4)}$$

où

$$\text{cov}(\tilde{\theta}_a) = \left[\left(\frac{\partial \text{vec}(V)}{\partial \theta}\right)^T\left(\frac{1}{2}V^{-1}\otimes V^{-1}\right)\left(\frac{\partial \text{vec}(V)}{\partial \theta}\right)\right]^{-1}$$

et

$$\tilde{Y} = Y - XZ\gamma.$$

À noter que $\tilde{\theta}_a$ dépend de θ et γ, de sorte que tous deux peuvent être estimés de façon itérative. Avec la méthode itérative par les moindres carrés généralisés, on commence avec une estimation initiale de V (c'est-à-dire en fixant les valeurs initiales de θ), ce qui donne une estimation de γ. Donc, en remplaçant l'estimation initiale de V avec l'estimation de γ dans (A.1), on obtient une estimation améliorée de θ. Dans la plupart des cas, il y a convergence après quelques itérations entre les équations (A.1) et (A.4), bien que ceci ne soit pas toujours garanti.

La méthode itérative restreinte par les moindres carrés généralisés s'appuie sur le fait que, si γ est estimé par les moindres carrés généralisés et que V est connu, alors

$$E\left[(Y-XZ\hat{\gamma})(Y-XZ\hat{\gamma})^T\right] = V-XZ(Z^TX^TV^{-1}XZ)^{-1}Z^TX^T.$$

L'équation qui précède suggère que nous utilisions

$$(Y-XZ\gamma)(Y-XZ\gamma)^T+XZ(Z^TX^TV^{-1}XZ)^{-1}Z^TX^T \quad \text{(A.5)}$$

à la place de $(Y-XZ\gamma)(Y-XZ\gamma)^T$ à chaque cycle d'itération décrit précédemment, de manière à obtenir un estimateur approximativement sans biais de V et, conséquemment, de θ.

Comme le souligne Goldstein (1986, 1989), si l'on utilise au départ une estimation cohérente de γ, disons l'estimateur par les moindres carrés ordinaires, alors les estimations finales seront elles aussi cohérentes, à la condition qu'il existe des moments finis du quatrième ordre.

Il convient de souligner qu'il est possible d'obtenir des estimations négatives de la variance à partir de la méthode décrite précédemment. On peut toutefois éviter ce problème en imposant des contraintes à chaque itération. Pour plus d'information à ce sujet, voir Goldstein (1986).

ANNEXE B: APPROXIMATION DE $E[\hat{\mu}_i - \tilde{\mu}_i]^2$

S'inspirant de Kachar et Harville (1984), Prasad et Rao (1990) ont élaboré une approximation du second ordre pour le deuxième terme de (2.5), dans certaines conditions de régularité.

$$E[\hat{\mu}_i - \tilde{\mu}_i]^2 \approx T_3 = tr\left[\left(\frac{\partial d_i}{\partial \theta}\right)V\left(\frac{\partial d_i}{\partial \theta}\right)^T E\left[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T\right]\right] \quad \text{(B.1)}$$

où, pour le modèle (2.1), $d_i = \bar{X}_i^T K_i (I\otimes\Omega)X^TV^{-1}$, $K_i = [0,...,I,...0]$, est la matrice $(p+1)\times(p+1)m$, avec la matrice unité I d'ordre $p+1$ dans la i-ième position, 0 est la matrice nulle d'ordre $p+1$ et $\hat{\theta}$ désigne tout estimateur invariant sous la translation de $\theta = (\theta_1,...,\theta_s)$ où $\theta_s = \sigma^2$ et θ_k; $k = 1,...,s-1$ sont les éléments distincts de Ω. Goldstein (1989) prouve que, dans les conditions de normalité des termes aléatoires du modèle (2.1), l'estimateur de θ par la méthode itérative restreinte des MCG équivaut à l'estimateur par le maximum de vraisemblance restreint, lequel est invariant sous la translation.

Supposons que $E\left[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T\right]$ s'approche de la matrice des covariances asymptotiques de l'estimateur par le maximum de vraisemblance restreint (B). Le jk^e élément de B^{-1} est représenté par (voir Harville 1977)

$$b_{jk}^* = Tr\left(\sum_{i=1}^m P_i\frac{\partial V}{\partial \theta_j}P_i\frac{\partial V}{\partial \theta_K}\right)$$

pour j et $k = 1,...,s$ où $P_i = V_i^{-1} - V_i^{-1}X_iZ_i(\sum_{i=1}^m Z_i^TX_i^TV_i^{-1}X_iZ_i)\,Z_i^TX_i^TV_i^{-1}$. Supposons que $b_{j,k}$ est le jk-ième élément de B. Après quelques opérations d'algèbre matricielle, il peut être démontré que

$$T_3 = \bar{X}_i^T(G_i^{-1})^T\left(\sum_{j=1}^{s-1}\sum_{k=1}^{s-1}b_{jk}\Delta_jC_i\Delta_k^T\right)G_i^{-1}\bar{X}_i -$$
$$2\bar{X}_i^T(G_i^{-1})^T\left(\sum_{j=1}^{s-1}b_{j,s}\Delta_j\right)R_i\Omega\bar{X}_i + b_{ss}\bar{X}_i^T\Omega S_i\Omega\bar{X}_i \quad \text{(B.2)}$$

où $C_i = \sigma^{-2}G_i^{-1}X_i^TX_i$; $R_i = \sigma^{-4}G_i^{-2}X_i^TX_i$; $S_i = \sigma^{-6}G_i^{-3}X_i^TX_i$; et

$$\Delta_k = \frac{\partial\Omega}{\partial\Omega_k}\ k = 1,...,s-1$$

est la matrice $s-1$ de la dérivée carrée par rapport à θ_k; $k = 1,...,s-1$.

BIBLIOGRAPHIE

BATTESE, G.E., et FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

FULLER, W.A. (1987). *Measurement Error Models*. Chichester: John Wiley.

GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares estimation. *Biometrika*, 73, 43-56.

GOLDSTEIN, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika*, 76, 622-623.

GONZALES, M.E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

HARVILLE, D.A. (1977). Maximum likelihood approach to variance component estimation and related problems. *Journal of the American Statistical Association*, 72, 320-340.

HARVILLE, D.A., et JESKE, D.R. (1992). Mean squared error of estimation on prediction under a general linear model. *Journal of the American Statistical Association*, 87, 724-731.

HENDERSON, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.

HOLT, D., et MOURA, F. (1993a). Mixed models for making small area estimates. Dans *Small Area Statistics and Survey Design*, (Éds., G. Kalton, J. Kordos et R. Platek,). 1, 221-231. Warsaw: Central Statistical Office.

HOLT, D., et MOURA, F. (1993b). Small area estimation using multilevel models. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 21-31.

KACKAR, R.N., et HARVILLE, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853-862.

LONGFORD, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrika*, 79, 817-827.

MOURA, F.A.S. (1994). Small Area Estimation Using Multilevel Models. University of Southampton. Thèse de doctorat (non publiée).

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

SINGH, A.C., STUKEL, D., et PFEFFERMANN, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, Séries B, 377-396.

Techniques d'enquête, juin 1999
Vol. 25 n° 1, pp. 91-97
Statistique Canada

Estimation composite de la prévalence des drogues pour des zones infraétats

MANAS CHATTOPADHYAY, PARTHA LAHIRI, MICHAEL LARSEN et JOHN REIMNITZ[1]

RÉSUMÉ

The Gallup Organization a mené des enquêtes-ménages pour étudier la prévalence à l'échelle des États de la consommation d'alcool et de drogues (par exemple, cocaïne, marijuana, etc.). Les estimations traditionnelles dans le cadre d'enquêtes fondées sur le plan d'échantillonnage relativement à l'utilisation et à la dépendance pour des comtés et des groupes démographiques choisis ont des erreurs-types inacceptables parce que la taille des échantillons dans des groupes infra-états est trop petite. L'estimation synthétique incorpore des renseignements démographiques et des indicateurs sociaux dans les estimations de la prévalence grâce à un modèle de régression implicite. Les estimations synthétiques ont tendance à produire des variances plus petites que les estimations fondées sur le plan d'échantillonnage, mais elles peuvent être très homogènes d'un comté à l'autre lorsque les variables auxiliaires sont homogènes. Les estimations composites pour les données régionales sont des moyennes pondérées des estimations synthétiques et des estimations d'enquêtes fondées sur le plan d'échantillonnage. Un deuxième problème que l'on ne rencontre pas généralement au niveau de l'État, mais qui est présent pour des zones infra-états et des groupes concerne l'estimation des erreurs-types des prévalences estimées qui sont près de zéro. Cette difficulté touche non seulement les estimations des enquêtes-ménages téléphoniques, mais aussi les estimations composites. Un modèle hiérarchique est proposé pour régler ce problème. Des estimateurs composites empiriques de Bayes, qui incorporent des facteurs de pondération des enquêtes, des prévalences et des estimateurs selon la technique du jackknife de leurs erreurs quadratiques moyennes sont présentés et illustrés.

MOTS CLÉS: Abus d'alcool; abus de drogues; méthode empirique de Bayes; technique du jackknife; erreur quadratique moyenne; estimation de données régionales; estimation synthétique.

1. INTRODUCTION

The Gallup Organization a mené une série d'enquêtes-ménages pour différents États afin d'étudier la prévalence à l'échelle des États de la consommation d'alcool et de drogues (par exemple, cocaïne, marijuana) chez les adultes et les adolescents civils, hors-établissement. L'objectif commun de ces enquêtes est d'estimer la prévalence en ce qui concerne la consommation et la dépendance à l'égard de l'alcool et des drogues et, sur cette base, de projeter les besoins en matière de traitement pour les consommateurs en situation de dépendance. Pour ce qui est de la planification et de l'affectation des ressources, les États ont besoin d'estimations précises sur la prévalence pour certains sous-groupes de la population visée. Par exemple, il est intéressant d'estimer la prévalence pour des comtés et des régions de planification au niveau infra-état dans des sous-populations démographiques (par exemple, hommes blancs âgés).

Les procédures traditionnelles fondées sur le plan d'échantillonnage pour estimer la consommation et la dépendance de sous-populations ont deux inconvénients. Tout d'abord, si l'estimation traditionnelle de l'enquête fondée sur le plan d'échantillonnage en ce qui concerne un sous-groupe est positive, mais la taille de l'échantillon est faible, alors l'erreur-type correspondant est inacceptablement élevée. Deuxièmement, étant donné que le problème est d'évaluer la proportion d'un événement rare, il est possible que la procédure fondée sur le plan d'échantillonnage produise une estimation nulle et des formules d'estimation des erreurs-types pour un sous-groupe donné, le cas échéant, donneraient une fausse impression de la véritable variabilité sous-jacente.

Dans le but d'améliorer les estimateurs traditionnels fondés sur le plan d'échantillonnage, on peut utiliser certains renseignements supplémentaires que l'on obtient habituellement dans des dossiers administratifs, de concert avec des données obtenues lors d'enquêtes téléphoniques. On y parvient habituellement en utilisant des modèles implicites ou explicites qui «empruntent de la force» ou qui incorporent une information additionnelle qui met en rapport les divers groupes, comtés et régions de planification l'un envers l'autre. La méthode proposée dans le présent document combine l'information à l'échelle des comtés afin de composer avec le problème des estimations nulles dans certains comtés. On l'obtient à partir d'un modèle qui lie les proportions loin de 1, ce qui est raisonnable dans une application pour laquelle on s'attend à ce que les proportions soient très faibles, et estime les paramètres au moyen de méthodes empiriques de Bayes. La procédure incorpore également les facteurs de pondération de l'échantillonnage de l'enquête dans l'estimation.

Pour un compte rendu détaillé des méthodes d'estimation des données régionales, voir Ghosh et Rao (1994). On peut

[1] Manas Chattopadhyay, The Gallup Organization; Partha Lahiri, University of Nebraska/ Lincoln; Michael Larsen, Harvard University, Department of Statistics, Science Center, One Oxford Street, Cambridge, MA 02138, U.S.A.; John Reimnitz, The Gallup Organization.

également trouver des références récentes dans Farrell MacGibbon et Tomberlin (1997) et dans Malec, Sedransk, Moriarity et Leclere (1997). Farrell et coll. (1997) proposent d'estimer des proportions de données régionales à l'aide de procédures empiriques de Bayes. Ils modélisent les proportions à l'aide d'une régression logistique qui met en rapport des proportions attendues avec des variables de l'échantillon et incluent des effets aléatoires pour les données régionales. Malec et coll. (1997) utilisent des modèles hiérarchiques de Bayes. Ils utilisent des modèles de régression logistique pour mettre en rapport des caractéristiques individuelles avec des probabilités d'un résultat et alors utilisent un modèle de régression linéaire pour mettre en rapport des coefficients entre des données régionales. La plupart des méthodes actuelles, y compris celles de Farrell et coll. (1997) et de Malec et coll. (1997) n'ont pas recours directement à des facteurs de pondération d'échantillonnage d'enquêtes dans l'estimation.

Le plan d'enquête utilisé par Gallup est décrit à la section 2. Dans la section 3, on présente les notations utilisées dans le document. Un estimateur direct fondé sur le plan d'échantillonnage et deux estimateurs synthétiques sont présentés à la section 4. En ce qui concerne la section 5, on y donne plusieurs estimateurs composites de prévalence de la consommation d'alcool et de drogues et de la dépendance à leur égard. Dans cette section, on propose certains estimateurs empiriques de Bayes et des estimateurs selon la technique du jackknife de leurs erreurs quadratiques moyennes. Dans la section 6, les estimateurs présentés dans les sections 4 et 5 sont appliquées à un ensemble de données provenant d'un État précis. Dans la présente étude, l'analyse se concentrait sur des estimations au niveau des comtés. À l'origine, des aspects liés à la planification de la taille des échantillons portaient sur des zones de planification d'infra-états plus importantes.

2. ENQUÊTE

À des fins d'échantillonnage, on divise l'État en quelques régions de planification et les échantillons sont prélevés de façon indépendante pour chaque région de planification en utilisant la méthode tronquée de composition aléatoire (CA) stratifiée de Casady et Lepkowski (1993). Ce plan d'échantillonnage stratifie la structure de Bellcore (BCR) en deux strates: une strate de haute densité composée de blocs de 100 ayant un ou plusieurs numéros résidentiels publiés et une strate de faible densité se composant de tous les autres numéros dans la structure BCR. Environ 52 percent des numéros dans la strate de haute densité sont estimés être des numéros résidentiels actifs tandis que dans la strate de faible densité, le pourcentage correspondant n'est que de 2 percent. La procédure de Cassidy-Lepkowski exploite la différence importante dans le coût de l'échantillonnage entre les deux strates en déterminant de façon optimale la taille de l'échantillon dans chaque strate. Dans la version tronquée de la procédure, l'échantillonnage se fait uniquement à partir de la strate de haute densité.

On a déterminé la taille de l'échantillon dans l'étude initiale afin d'évaluer la prévalence à l'échelle de l'État avec un degré souhaité d'exactitude. Les tailles des échantillons ont été attribuées aux régions de planification grâce à un mécanisme d'attribution optimale. On a utilisé les données sur les admissions à une désintoxication pour la population adulte dans chaque comté pour calculer l'indice de prévalence (le taux des admissions) en pourcentage dans chaque région de planification. On a alors utilisé ces indices pour calculer la taille optimale de l'échantillon pour chaque région de planification. Suite à une attribution optimale, on a attribué des échantillons de tailles relativement importantes à des régions de planification ayant des indices de prévalence plus élevés. L'attribution optimale minimise également la variance des estimateurs. Gallup a également suréchantillonné le groupe d'âge 18-45 par région de planification parce qu'il s'agit du groupe d'âge ayant des taux relativement élevés de consommation de drogues illicites. En raison de l'attribution optimale (qui peut être disproportionnelle), du suréchantillonnage du groupe d'âge et de la conception complexe, il a été nécessaire de recourir à la pondération pour calculer les estimations provenant des données de l'échantillon. Les facteurs de pondération nécessaires, que l'on appelle couramment les facteurs de pondération, ont été calculés à l'aide des estimations actuelles de la population fondées sur les données du recensement.

En raison de contraintes budgétaires, il n'est pas possible d'accroître la taille de l'échantillon pour toutes les régions infra-états et tous les groupes afin d'obtenir l'exactitude souhaitée. Pour estimer la prévalence de l'alcool et des drogues, nous prenons en compte les procédures empiriques de Bayes (voir Efron et Morris 1973, Fay et Herriot 1979, Ghosh et Lahiri 1987, entre autres) pour améliorer les estimations habituelles fondées sur le plan d'échantillonnage de la prévalence des drogues en tirant profit de mesures démographiques et de données d'indicateurs sociaux.

D'autres variables qui sont peut-être reliées à la prévalence de la consommation et de la dépendance et que l'on peut obtenir du recensement incluent le pourcentage de la population qui a plus de 65 ans, qui a moins de 30 ans, qui est blanche, de sexe masculin, marié et à loyer. Les administrations locales peuvent fournir des données selon le comté relativement aux indicateurs sociaux, tels le taux de conduite en état d'ébriété, le taux de mortalité, les permis d'alcool par habitant, et les taux d'admission à des centres de désintoxication (alcool et drogues). Plus les variables auxiliaires sont étroitement reliées à la prévalence de la consommation et de la dépendance, plus il est probable que les méthodes qui «empruntent de la force» dans les régions et les groupes, telles les méthodes empiriques de Bayes présentées dans le présent document, peuvent être utilisées pour satisfaire aux niveaux souhaités d'exactitude pour les régions infra-états.

3. NOTATIONS

Soit n_i la taille de l'échantillon attribué à la i-ième région de planification, $i = 1, ..., I$ ($n = \sum_{i=1}^{I} n_i$). Des échantillons sont prélevés de façon indépendante dans chaque région de planification à l'aide d'enquêtes téléphoniques de CA. Une fois qu'on a observé l'échantillon, on suppose que chaque région fait l'objet d'une stratification a posteriori en groupes démographiques K. Ces groupes sont formés en recoupant le sexe (homme, femme) et l'âge (18-24, 25-44, 45-64, 65+), ce qui donne $K = 2 \times 4 = 8$ groupes. On suppose qu'il y a J_i comtés dans la i-ième région de planification ($i = 1, ..., I$) et n_{ijk} observations au sein du k-ième groupe démographique dans le j-ième comté appartenant à la i-ième région de planification ($i = 1, ..., I$; $j = 1, ..., J_i$; $k = 1, ..., K$). Étant donné que de façon générale n_{ijk} est petit, il y a une bonne chance qu'un certain nombre des groupes démographiques k ne soient pas représentés dans un comté donné. Soit S_{ij} l'ensemble de groupes démographiques dans le j-ième comté au sein de la i-ième strate ($i = 1, ..., I$; $j = 1, ..., J_i$) pour laquelle des personnes ont terminé des enquêtes.

On suppose que y_{ijkl} soit la l-ième observation (0 ou 1) pour le k-ième groupe démographique dans le j-ième comté appartenant à i-ième région de planification ($i = 1, ..., I$; $j = 1, ..., J_i$; $k \in S_{ij}$; $l = 1, ..., n_{ijk}$). Soit w_{ijkl} le facteur de pondération de l'échantillonnage correspondant disponible pour l'enquête. L'objectif est d'estimer π_{ij}, la véritable prévalence de consommation de substances ou de la dépendance à l'égard de substances pour le j-ième comté au sein de la i-ième zone de planification ($i = 1, ..., I$; $j = 1, ..., J_i$).

4. ESTIMATEUR DIRECT D'ENQUÊTE ET ESTIMATEURS SYNTHÉTIQUES

L'estimateur d'enquête direct de π_{ij} est défini par

$$\widehat{\pi}_{ij}^{D} = \frac{\sum_{k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl} y_{ijkl}}{\sum_{k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}}.$$

La taille de l'échantillon disponible dans un comté pourrait être très petite (parfois aussi petite que 3 ou 4). Ainsi, l'estimateur est nettement non fiable. D'autres estimateurs d'enquête directs sont définis de façon semblable. Par exemple, l'estimateur d'enquête direct de π_{ik}, la véritable prévalence dans le k-ième groupe démographique dans la i-ième région de planification est

$$\widehat{\pi}_{ik}^{D} = \frac{\sum_{j: k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl} y_{ijkl}}{\sum_{j: k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}},$$

où la notation $j: k \in S_{ij}$ signifie que la totalisation se fait pour les comtés j dans lesquels le groupe démographique k est observé.

D'autres problèmes surgissent lors de l'estimation de la proportion d'événements rares. Il est fort probable que toutes les observations dans un comté puissent être nulles, ce qui donne lieu à une estimation nulle pour un comté. Si des estimations habituelles des erreurs-types s'appliquaient, une erreur-type estimative nulle de l'estimation donnerait une fausse impression quant à l'incertitude de l'estimation. Ainsi, il est très important d'améliorer l'estimateur de l'enquête direct.

Les estimateurs synthétiques tirent leur puissance de comtés connexes grâce à la modélisation implicite de données supplémentaires provenant du U.S. Census Bureau ainsi que des données de l'enquête téléphonique. Un estimateur synthétique, que l'on a utilisé dans le passé pour estimer la prévalence de l'alcool au niveau du comté, est donné par

$$\widehat{\pi}_{ij}^{S1} = \sum_{k=1}^{K} a_{ijk} \widehat{\pi}_{k}^{D},$$

où $\widehat{\pi}_{k}^{D}$ est l'estimateur d'enquête direct à l'échelle de l'État de la prévalence de l'alcool pour le k-ième groupe démographique et a_{ijk} est la proportion de personnes appartenant au k-ième groupe démographique dans le j-ième comté au sein de la i-ième zone de planification ($i = 1, ..., I$; $j = 1, ..., J_i$; $k = 1, ..., K$). La valeur a_{ijk} provient des estimations actuelles du recensement. Dans le cas de l'enquête-ménage don't il est fait mention dans le présent document, les valeurs a_{ijk} proviennent de fournisseurs de bases de données tels Claritas Data Services d'Ithaca, New York. D'après les plus récentes données disponibles du recensement, on estime habituellement les valeurs a_{ijk} à l'aide de modèles de projection. Cependant, dans la pratique, les valeurs a_{ijk} ne sont pas des proportions véritables, mais sont des estimations actuelles du recensement de précision raisonnable. Des valeurs a_{ijk} périmées ou inexactes ont pour conséquence de rendre biaisés les estimateurs qui les utilisent. Si on a recours aux projections démographiques pour calculer les ajustements de pondération de stratification a posteriori, l'estimateur d'enquête direct est également influencé par cette source de biais. Il ne relève pas de la portée du présent document d'étudier l'impact de projections démographiques de rechange. En proposant $\widehat{\pi}_{ij}^{S1}$, on suppose de façon implicite que les prévalences à l'égard de la consommation d'alcool et de drogues pour le k-ième groupe dans tous les comtés sont les mêmes (ou pratiquement les mêmes).

Un estimateur synthétique moins restrictif de la prévalence de la consommation d'alcool et de drogues est donné par

$$\widehat{\pi}_{ij}^{S2} = \sum_{k=1}^{K} a_{ijk} \widehat{\pi}_{ik}^{D},$$

où $\widehat{\pi_{ik}}^{D}$ est un estimateur d'enquête direct de π_{ik}, la prévalence pour la consommation d'alcool ou de drogues en ce qui concerne le k-ième groupe démographique dans la i-ième région de planification. On suppose de façon implicite que les prévalences dans le k-ième groupe sont les mêmes (ou pratiquement les mêmes) pour tous les comtés dans une région de planification. Cette hypothèse est plus «régionale» ou moins restrictive que celle faite en proposant $\widehat{\pi_{ij}}^{S1}$. Un estimateur semblable d'enquête direct $\widehat{\pi_{ijk}}^{D}$ pour le k-ième groupe démographique au sein d'un comté précis j dans une région i peut être défini en limitant l'échantillon au comté j seulement. Comparativement à $\widehat{\pi_{ijk}}^{D}$, l'estimateur $\widehat{\pi_{ik}}^{D}$ aura un écart relativement plus faible bien qu'il puisse faire l'objet d'un certain biais étant donné qu'il ne fait pas la distinction entre les comtés. Par contre, $\widehat{\pi_{ijk}}^{D}$ peut se fonder sur un échantillon de très faible taille et par conséquent être significativement moins fiable pour ce qui est de sa variabilité.

Les estimateurs synthétiques susmentionnés obtiennent des réductions au niveau des variances mais ont pour conséquence une augmentation du biais. Les estimateurs synthétiques font la distinction entre les comtés uniquement par l'intermédiaire d'une variable indirecte a_{ijk} obtenue du recensement, tandis que l'estimateur direct traite chaque comté de façon distincte.

5. ESTIMATEURS COMPOSITES DE π_{ij} UTILISANT DES DONNÉES DU RECENSEMENT ET D'ENQUÊTES TÉLÉPHONIQUES

Un compromis entre un estimateur d'enquête direct et un estimateur synthétique est un estimateur composite. On propose un certain nombre d'estimateurs composites différents dans le présent document fondés sur l'identité suivante:

$$\pi_{ij} = \sum_{k \in S_{ij}} a_{ijk}\pi_{ijk} + \sum_{k \notin S_{ij}} a_{ijk}\pi_{ijk},$$

où π_{ijk} est la prévalence pour la consommation d'alcool et de drogues et a_{ijk}, tel qu'il est défini ci-dessus, est la proportion de personnes appartenant au k-ième groupe démographique dans le j-ième comté au sein de la i-ième zone de planification ($i = 1, ..., I$; $j = 1, ..., J_i$; $k = 1, ..., K$).

On obtient un estimateur composite simple de π_{ij} lorsque, pour $k \in S_{ij}$, π_{ijk} est estimé par $\widehat{\pi_{ijk}}^{D}$, l'estimateur d'enquête direct de π_{ijk}, et pour $k \notin S_{ij}$, π_{ijk} est estimé par $\widehat{\pi_{ik}}^{D}$. L'estimateur est alors donné par

$$\widehat{\pi_{ij}}^{C} = \sum_{k \in S_{ij}} a_{ijk}\widehat{\pi_{ijk}}^{D} + \sum_{k \notin S_{ij}} a_{ijk}\widehat{\pi_{ik}}^{D}.$$

Dans la formule ci-dessus, π_{ijk} ($k \in S_{ij}$) est estimé à l'aide d'un petit échantillon et il y a donc la possibilité d'améliorer $\widehat{\pi_{ijk}}^{D}$ (et, par conséquent, $\widehat{\pi_{ij}}^{C}$) en empruntant de la force des ressources pertinentes. À cette fin, une estimation empirique de Bayes de π_{ij} est proposée, fondée sur le modèle suivant.

Modèle

1. Étant donnés les π_{ijk}, les y_{ijkl} sont sans corrélation l'un envers l'autre avec $\mathrm{E}(y_{ijkl}|\pi_{ijk}) = \pi_{ijk}$ et $\mathrm{Var}(y_{ijkl}|\pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk})$ pour $i = 1, ..., I$; $j = 1, ..., J_i$; $k = 1, ..., K$; $l = 1, ..., n_{ijk}$.
2. Les π_{ijk} sont sans corrélation avec $\mathrm{E}(\pi_{ijk}) = \mu_{ik}$; $\mathrm{Var}(\pi_{ijk}) = d\mu_{ik}^2$ ($i = 1, ..., I$; $j = 1, ..., J_i$; $k = 1, ..., K$).

Si $\pi_{ijk} \sim \text{Uniforme}(0, 2\mu_{ik})$, alors dans l'énoncé (2) $d = 1/3$. Ainsi, à la différence de l'hypothèse implicite émise dans l'estimateur synthétique $\widehat{\pi_{ij}}^{S2}$ (c'est-à-dire $\pi_{ijk} = u_{ik}$), une certaine variabilité des proportions entre les comtés au sein d'une région pour un groupe démographique donné est permise.

La première hypothèse du modèle sous-entend que, étant donné π_{ijk}, les $\widehat{\pi_{ijk}}^{D}$ sont sans corrélation l'un avec l'autre, $\mathrm{E}(\widehat{\pi_{ijk}}^{D}|\pi_{ijk}) = \pi_{ijk}$, et $\mathrm{Var}(\widehat{\pi_{ijk}}^{D}|\pi_{ijk}) = c_{ijk}\pi_{ijk}(1 - \pi_{ijk})$, où $c_{ijk} = \sum_{l=1}^{n_{ijk}} w_{ijkl}^2 / (\sum_{l=1}^{n_{ijk}} w_{ijkl})^2$ pour $i = 1, ..., I$; $j = 1, ..., J_i$; $k = 1, ..., K$. L'estimateur linéaire de Bayes de π_{ij}, dans le modèle et la fonction quadratique de perte, est donné par

$$\widehat{\pi_{ij}}^{B} = \sum_{k \in S_{ij}} a_{ijk}\left(B_{ijk}\widehat{\pi_{ijk}}^{D} + (1 - B_{ijk})\mu_{ik}\right) + \sum_{k \notin S_{ij}} a_{ijk}\mu_{ik},$$

où $B_{ijk} = d\mu_{ik}^2 / \left(d\mu_{ik}^2 + c_{ijk}\left(u_{ik} - (d+1)u_{ik}^2\right)\right)$.

Étant donné que l'estimateur de Bayes met en cause le paramètre inconnu μ_{ik}, on ne peut l'utiliser dans la pratique. On obtient l'estimateur empirique de Bayes suivant de π_{ij} lorsque μ_{ik} est remplacé par un estimateur, disons $\widehat{\mu_{ik}}$, de μ_{ik}:

$$\widehat{\pi_{ij}}^{EB} = \sum_{k \in S_{ij}} a_{ijk}\left(\widehat{B_{ijk}}\widehat{\pi_{ijk}}^{D} + (1 - \widehat{B_{ijk}})\widehat{\mu_{ik}}\right) + \sum_{k \notin S_{ij}} a_{ijk}\widehat{\mu_{ik}},$$

où $\widehat{B_{ijk}} = d\widehat{\mu_{ik}}^2 / (d\widehat{\mu_{ik}}^2 + c_{ijk}(\widehat{\mu_{ik}} - (d+1)\widehat{\mu_{ik}}^2))$. Le facteur de pondération ou le facteur de retrait $\widehat{B_{ijk}}$ est un ratio de la variance de π_{ijk} dans le modèle par rapport à la variance (inconditionnelle) de $\widehat{\pi_{ijk}}$. L'estimateur de μ_{ik} est censé être $\widehat{\mu_{ik}} = \widehat{\pi_{ik}}^{D}$.

Erreurs quadratiques moyennes

L'erreur quadratique moyenne de l'estimateur de Bayes $\widehat{\pi_{ij}}^{B}$ est défini comme l'erreur quadratique moyenne $\mathrm{EQM}(\widehat{\pi_{ij}}^{B}) = \mathrm{E}(\widehat{\pi_{ij}}^{B} - \pi_{ij})^2$, où l'espérance (inconditionnelle) est prise par rapport au modèle. On peut vérifier que

$$\begin{aligned}\mathrm{EQM}(\widehat{\pi_{ij}}^{B}) &= \mathrm{Var}\,(\widehat{\pi_{ij}}^{B} - \pi_{ij}) \\ &= \mathrm{Var}\,(\widehat{\pi_{ij}}^{B}) + \mathrm{Var}(\pi_{ij}) - 2\mathrm{Cov}(\widehat{\pi_{ij}}^{B}, \pi_{ij}) \\ &= \mathrm{Var}\,(\pi_{ij}) - \mathrm{Var}\,(\widehat{\pi_{ij}}^{B}) \\ &= d\left(\sum_{k\in S_{ij}} a_{ijk}^2 (1 - B_{ijk})\,\mu_{ik}^2 + \sum_{k\notin S_{ij}} a_{ijk}^2 \mu_{ik}^2 \right).\end{aligned}$$

Il est de règle de prendre $\mathrm{EQM}(\widehat{\pi_{ij}}^{B})$ comme l'erreur quadratique moyenne de l'estimateur empirique de Bayes $\widehat{\pi_{ij}}^{EB}$. Cependant, $\mathrm{EQM}(\widehat{\pi_{ij}}^{B})$ sous-estimera l'erreur quadratique moyenne de $\widehat{\pi_{ij}}^{EB}$ étant donné qu'elle n'incorpore pas l'incertitude attribuable à l'estimation du paramètre μ_{ik}. Voir Prasad et Rao (1990) et Lahiri et Rao (1995) dans ce contexte. À l'aide d'un argument standard de Bayes, on peut démontrer que

$$\mathrm{EQM}(\widehat{\pi_{ij}}^{EB}) = \mathrm{EQM}(\widehat{\pi_{ij}}^{B}) + \mathrm{E}(\widehat{\pi_{ij}}^{EB} - \widehat{\pi_{ij}}^{B})^2.$$

Il est nécessaire d'estimer $\mathrm{EQM}(\widehat{\pi_{ij}}^{EB})$ étant donné qu'elle contient le paramètre inconnu μ_{ik}. Le premier terme $\mathrm{EQM}(\widehat{\pi_{ij}}^{B})$ peut être estimé à l'aide de

$$\mathrm{eqm}_J(\widehat{\pi_{ij}}^{B}) = \mathrm{eqm}(\widehat{\pi_{ij}}^{B}) - \frac{J_i - 1}{J_i} \sum_{u=1}^{J_i} \left(\mathrm{eqm}_{(-u)}(\widehat{\pi_{ij}}^{B}) - \mathrm{eqm}(\widehat{\pi_{ij}}^{B}) \right),$$

où

$$\mathrm{eqm}(\widehat{\pi_{ij}}^{B}) = d\left(\sum_{k\in S_{ij}} a_{ijk}^2 (1 - \widehat{B_{ijk}})\,\widehat{\mu_{ik}}^{2} + \sum_{k\notin S_{ij}} a_{ijk}^2 \widehat{\mu_{ik}}^{2} \right)$$

et

$$\mathrm{eqm}_{(-u)}(\widehat{\pi_{ij}}^{B}) = d\left(\sum_{k\in S_{ij}} a_{ijk}^2 (1 - \widehat{B_{ijk(-u)}})\,\widehat{\mu_{ik(-u)}}^{2} + \sum_{k\notin S_{ij}} a_{ijk}^2 \widehat{\mu_{ik(-u)}}^{2} \right),$$

avec

$$\widehat{\mu_{ik(-u)}} = \sum_{j\neq u}^{J_i} \sum_{l=1}^{n_{ijk}} w_{ijkl}\, y_{ijkl} \Big/ \sum_{j\neq u}^{J_i} \sum_{l=1}^{n_{ijk}} w_{ijkl},$$

et

$$\widehat{B_{ijk(-u)}} = d\widehat{\mu_{ik(-u)}}^{2} \Big/ \left(d\widehat{\mu_{ik(-u)}}^{2} + c_{ijk}(\widehat{\mu_{ik(-u)}} - (d+1)\widehat{\mu_{ik(-u)}}^{2}) \right).$$

Voir Jiang, Lahiri et Wan (1998) pour des observations à l'égard de ces estimateurs. Le deuxième terme $\mathrm{E}(\widehat{\pi_{ij}}^{EB} - \widehat{\pi_{ij}}^{B})^2$ peut être estimé à l'aide de l'estimateur suivant de la technique du jackknife:

$$\mathrm{E}_J(\widehat{\pi_{ij}}^{EB} - \widehat{\pi_{ij}}^{B})^2 = \frac{J_i - 1}{J_i} \sum_{u=1}^{J_i} (\widehat{\pi_{ij(-u)}}^{EB} - \widehat{\pi_{ij}}^{EB})^2,$$

où

$$\widehat{\pi_{ij(-u)}}^{EB} = \sum_{k\in S_{ij}} a_{ijk} (\widehat{B_{ijk(-u)}}\,\widehat{\pi_{ijk}}^{D} + (1 - \widehat{B_{ijk(-u)}})\,\widehat{\mu_{ik(-u)}}) + \sum_{k\notin S_{ij}} a_{ijk}\,\widehat{\mu_{ik(-u)}}.$$

Ainsi, $\mathrm{EQM}(\widehat{\pi_{ij}}^{EB})$ est estimé par

$$\mathrm{eqm}(\widehat{\pi_{ij}}^{EB}) = \mathrm{eqm}_J(\widehat{\pi_{ij}}^{B}) + \mathrm{E}_J(\widehat{\pi_{ij}}^{EB} - \widehat{\pi_{ij}}^{B})^2.$$

Les méthodes jackknife sont examinées dans le récent texte de Shao et Tu (1995).

6. UN EXEMPLE

Dans la présente étude, l'objectif premier est de fournir des renseignements au sujet du besoin de traitement. Quiconque satisfait aux critères d'abus ou de dépendance à vie tel que les définissent les critères DSM-III-R du National Technical Center est considéré comme un membre du groupe de répondants qui peuvent avoir eu besoin de traitements au cours de la dernière année. Plusieurs variables d'indicateurs ont été créées dans l'ensemble de données pour identifier les répondants ayant un diagnostic de dépendance à l'égard de substances ou de consommations d'alcool ou de drogues. Aux fins de calcul numérique, ces variables d'indicateurs avec 0 et 1 comme valeurs possibles ont été traitées comme des variables de réponse (y_{ijkl}).

Dans le but d'économiser de l'espace, les résultats sont présentés uniquement pour la variable de résultat Dépendance à l'égard de l'alcool. On peut obtenir auprès des auteurs les résultats relatifs aux autres variables de réponse. Afin de protéger la confidentialité, on donne les résultats pour uniquement 40 comtés, identifiés de 1 à 40. Le tableau 1 renferme cinq estimations différentes de prévalence pour la dépendance à l'égard de l'alcool. De façon générale, les estimations directes sont hautement variables et sont souvent nulles. Le premier estimateur synthétique (S1) est le plus stable, produisant aucune estimation nulle et des estimations ayant peu de variabilité. Le deuxième estimateur synthétique (S2) est semblable à S1, mais n'est pas aussi restrictif. Les premières estimations synthétiques sont très homogènes tandis que les deuxièmes sont homogènes au sein des quatre régions de planification. Les estimations produites par l'estimateur composite sont plus variables que les autres estimations. L'estimateur empirique de Bayes produit des estimations très semblables à celles de S2. Dans le modèle menant à l'estimateur empirique de Bayes, d a été choisi pour correspondre à un tiers.

Tableau 1

Cinq estimateurs de la prévalence de la Dépendance à l'égard de l'alcool exprimée en pourcentage pour quarante comtés. Les erreurs-types estimées pour l'estimateur direct (Est. e.t.) et la racine carrée de l'erreur quadratique moyenne estimée pour les estimations empiriques de Bayes ($\sqrt{\text{Est.eqm}}$) entre parenthèses, également sous forme de pourcentages

	Estimateur								
	Direct		Synthétique 1	Synthétique 2	Composite	Empirique de Bayes		Taille de	Nombre de groupes observés
Comté	$\widehat{\pi}_{ij}^{D}$	(Est. e.t.)	$\widehat{\pi}_{ij}^{S1}$	$\widehat{\pi}_{ij}^{S2}$	$\widehat{\pi}_{ij}^{C}$	$\widehat{\pi}_{ij}^{EB}$	($\sqrt{\text{Est.eqm}}$)	l'échantillon	dans le comté
1	1,7	(2,4)	3,4	1,6	0,9	1,6	(0,33)	30	8
2	4,4	(2,0)	3,8	1,8	7,2	2,1	(0,35)	111	8
3	0,0	(0,0)	3,6	3,3	0,0	3,0	(0,85)	36	8
4	0,0	(0,0)	3,3	5,6	1,6	5,3	(1,79)	6	5
5	9,4	(4,8)	3,3	5,6	14,1	6,9	(1,78)	37	8
6	1,6	(1,1)	3,4	3,0	1,7	2,7	(0,67)	136	8
7	9,3	(5,8)	3,4	3,1	9,9	3,1	(0,81)	25	6
8	0,0	(0,0)	3,6	3,2	0,4	3,1	(0,84)	20	7
9	0,0	(0,0)	3,4	5,8	5,6	5,8	(1,93)	3	3
10	1,5	(1,3)	3,4	2,1	0,7	1,9	(0,54)	81	8
11	0,0	(0,0)	3,3	1,6	0,0	1,5	(0,33)	58	8
12	7,0	(6,8)	3,5	1,7	5,0	1,8	(0,35)	14	6
13	5,7	(3,8)	3,3	5,5	12,9	6,4	(1,75)	37	8
14	0,0	(0,0)	3,5	1,7	0,8	1,6	(0,33)	12	4
15	2,4	(1,4)	3,3	5,6	2,0	4,4	(1,56)	120	8
16	4,1	(3,5)	3,3	3,0	2,5	3,0	(0,77)	32	7
17	2,8	(2,4)	3,8	1,8	1,3	1,8	(0,37)	48	8
18	3,9	(1,1)	3,4	3,0	3,2	3,2	(0,60)	316	8
19	0,0	(0,0)	3,4	5,7	3,7	5,7	(1,95)	19	5
20	3,1	(3,9)	3,6	3,2	14,9	3,2	(0,82)	20	6
21	2,7	(1,6)	3,3	5,6	4,1	5,8	(1,50)	102	8
22	4,2	(1,8)	3,3	2,1	1,8	2,2	(0,42)	124	8
23	9,7	(2,7)	4,3	8,0	11,8	8,8	(2,11)	121	8
24	0,0	(0,0)	3,3	2,0	0,2	1,9	(0,54)	22	6
25	7,8	(4,7)	3,3	1,6	2,8	1,8	(0,33)	32	6
26	0,0	(0,0)	3,5	1,7	0,0	1,6	(0,37)	28	7
27	2,2	(1,8)	3,2	5,6	1,6	4,9	(1,74)	63	8
28	10,5	(13,7)	3,4	1,6	14,2	1,7	(0,35)	5	5
29	0,0	(0,0)	3,5	3,1	1,8	3,0	(0,81)	12	5
30	0,0	(0,0)	3,2	1,5	0,0	1,5	(0,33)	11	6
31	4,6	(3,2)	3,5	5,9	17,0	5,8	(1,87)	44	8
32	8,4	(3,8)	3,7	3,4	8,4	4,1	(0,84)	52	8
33	2,5	(1,3)	3,4	2,2	2,5	2,1	(0,50)	144	8
34	2,9	(2,4)	3,6	1,7	1,3	1,7	(0,35)	49	7
35	0,0	(0,0)	3,3	3,0	0,0	2,8	(0,77)	22	8
36	0,0	(0,0)	3,4	3,1	0,3	2,9	(0,82)	17	6
37	4,2	(4,0)	3,0	2,0	3,4	2,1	(0,54)	26	6
38	0,0	(0,0)	3,4	5,8	3,7	5,7	(1,97)	16	6
39	0,0	(0,0)	3,5	3,1	0,6	3,0	(0,81)	10	6
40	5,3	(1,9)	3,4	3,1	2,9	3,5	(0,69)	144	8

Tableau 2

Résumé des cinq estimateurs de la prévalence de la Dépendance à l'égard de l'alcool pour tous les comtés. Les résultats sont exprimés en pourcentage

Estimateur	minimum	1er quartile	median	3^{e} quartile	maximum	moyenne	erreurs-type
Direct	0,0	0,0	2,2	4,3	10,5	2,8	3,2
Synthétique 1	3,0	3,3	3,4	3,5	4,3	3,5	0,2
Synthétique 2	1,5	1,8	3,0	4,4	8,0	3,2	1,7
Composite	0,0	0,4	1,7	4,6	17,5	3,7	4,8
Empirique de Bayes	,15	1,8	2,8	4,2	8,8	3,2	4,8

Le tableau 1 illustre également les erreurs-types estimées des estimations directes et la racine carrée des erreurs quadratiques moyennes estimées (voir la section 4) des estimations empiriques de Bayes. Les erreurs-types des estimations directes, qui sont calculées comme

$$\sqrt{\hat{\pi}_{ij}^{D}(1-\hat{\pi}_{ij}^{D})/n_{ij}},$$

sont souvent estimées (incorrectement) comme nuls et sont passablement variables. Les racines carrées de l'erreur quadratique moyenne estimée des estimations empiriques de Bayes sont relativement stables et toujours inférieures à 0,025.

Le tableau 2 résume les estimations relatives à la dépendance à l'égard de l'alcool dans le tableau précédent pour tous les comptés de l'État. Les moyennes des estimations synthétiques et composites sont supérieures à la moyenne des estimations directes parce qu'il y a un plus petit nombre d'estimations nulles et que les moyennes dans les tableaux sommaires sont non pondérées.

7. CONCLUSION

Nous avons proposé des estimateurs empiriques simples de Bayes pour estimer les prévalences au niveau du comté. Les estimateurs empiriques de Bayes sont très efficaces lorsque la taille des échantillons pour les comtés est petite et lorsque les prévalences sont extrêmement petites. Nous avons introduit une mesure d'incertitude de l'estimateur empirique proposé de Bayes fondée sur la méthode de la technique du jackknife. La mesure proposée incorpore des sources additionnelles de variabilité attribuables à l'estimation de divers paramètres de modèles. Dans notre modèle, présenté dans le présent document, nous avons implicitement supposé que les probabilités de sélection ne sont pas reliées à y_{ijkl}. Dans l'étude-ménage dont il est fait mention dans le présent document, les probabilités de sélection étaient inégales et dépendaient de plusieurs facteurs tels le nombre de lignes téléphoniques et le nombre d'adultes dans le ménage. Aucune de ces variables n'était reliée à y_{ijkl}. Cependant, la répartition de l'échantillon dans différentes régions s'est faite en fonction du nombre «d'admissions à des centres de désintoxication» dans chaque région. Donc, les probabilités de sélection pourraient être indirectement reliées à y_{ijkl}. Dans le présent document, nous n'avons pas abordé la question du biais de sélection de l'échantillon, que l'on peut traiter de façon appropriée en suivant les procédures dont il est question dans Pfeffermann (1993).

Dans le présent document, nous n'avons pas tenu compte de l'utilisation des variables auxiliaires dans le modèle pour mettre en rapport les données régionales l'une avec l'autre et pour faciliter l'estimation améliorée. L'utilisation de données auxiliaires disponibles provenant du recensement des États-Unis et d'autres documents administratifs peut constituer une utilisation sensée de ressources que l'on peut utiliser pour améliorer la planification des traitements de désintoxication. Nous prévoyons approfondir le travail dans ce domaine à l'aide d'un exemple réel dans un prochain document.

REMERCIEMENTS

Nous aimerions remercier un examinateur anonyme qui nous a donné plusieurs suggestions pour l'amélioration du document, et aussi le Gallup Organization pour un soutien partiel. De plus, Partha Lahiri aimerait reconnaître le soutien partiel du U.S. National Science Foundation Grant SBR-9705574.

BIBLIOGRAPHIE

CASADY, R.J., et LEPKOWSKI, J.M. (1993). Plans d'enquête téléphonique stratifiée. *Techniques d'enquête*, 19, 115-125.

EFRON, B., et MORRIS, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *Journal of the American Statistical Association*, 68, 117-130.

FARRELL, P.J., MacGIBBON, B., et TOMBERLIN, T.J. (1997). Empirical Bayes estimators of small area proportions in multistage designs. *Statistica Sinica*, 7, 1065-1083.

FAY, R., et HERRIOT, R. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GHOSH, M., et LAHIRI, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 82, 1153-1162.

GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

JIANG, J., LAHIRI, P., et WAN, S. (1998). Jackknifing Mean Squared Error of Empirical Best Predictor. Manuscript non-publié.

LAHIRI, P., et RAO, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.

MALEC, D., SEDRANSK, J., MORIARITY, C.L., et LECLERE, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

SHAO, J., et TU, D. (1995). *The Jackknife and the Bootstrap*. New York: Springer-Verlag.

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 99-111
Statistique Canada

Étude de certains problèmes posés par l'estimation de la dynamique du revenu

SUSANA RUBIN BLEUER et MILORAD KOVAČEVIĆ[1]

RÉSUMÉ

Les auteurs étudient deux estimateurs fondés sur le plan de sondage des flux et des taux de transition. L'un, appelé estimateur mixte, s'appuie sur les échantillons transversaux pour l'estimation des bornes des catégories de revenu pour chaque période de référence et sur l'échantillon longitudinal pour l'estimation du nombre d'unités dans la population longitudinale (dénombrement longitudinal). L'autre, appelé estimateur longitudinal, se fonde sur l'échantillon longitudinal aussi bien pour l'estimation des bornes des catégories de revenu que pour les dénombrements longitudinaux. Par une méthode de simulation, les auteurs comparent les deux estimateurs lorsque les taux d'érosion sont élevés. Ils constatent que, si le modèle de correction pour l'érosion est imparfait, l'estimateur mixte est ordinairement plus sensible que l'estimateur longitudinal au biais lié au modèle. Ils observent aussi que, dans le cas de l'estimateur mixte, ce biais annule le faible gain de précision obtenu comparativement à l'estimateur longitudinal. Enfin, ils se servent des données de l'Enquête sur la dynamique du travail et du revenu et de la base de données longitudinales administratives de Statistique Canada pour illustrer les résultats.

MOTS CLÉS: Érosion; flux brut; taux de transition; pondération longitudinale; pondération transversale; estimateur bootstrap de la variance.

1. INTRODUCTION

Les flux bruts s'entendent du dénombrement des transitions entre un certain nombre d'états vécues, d'un point dans le temps à un autre, par les membres d'une population. Les paramètres connexes sont les proportions longitudinales et les taux de transition. Les proportions longitudinales correspondent aux flux bruts relatifs, et les taux de transition, aux flux bruts relatifs compte tenu de l'état de transition initial. L'estimation de ces paramètres pour les transitions entre diverses catégories de revenu, qui est essentielle à l'étude de la dynamique du revenu, peut se faire d'après des données d'enquête longitudinale. Les bornes des catégories de revenu doivent souvent être estimées, elles aussi, d'après les données de l'enquête. La mesure de faible revenu, définie comme étant la moitié du revenu médian, après correction pour tenir compte de la taille de la famille, en est un exemple. Par conséquent, l'estimation du nombre de transitions vers l'«état de faible revenu» et en provenance de ce dernier nécessite l'estimation de la médiane de revenu durant la période de référence.

Puisque les bornes des catégories de revenu ont habituellement trait aux populations transversales respectives, on doit les estimer d'après les échantillons transversaux pour obtenir des estimations non biaisées. Si les changements d'un cycle à l'autre autrement dit, le nombre d'unités «ajoutées» ou «disparues») est négligeable, l'échantillon longitudinal est représentatif de la population aux deux points de référence dans le temps et nous pouvons estimer les bornes des catégories de revenu d'après cet échantillon. Sinon, l'estimation des bornes d'après l'échantillon longitudinal pourrait produire des estimations biaisées. Par «unité disparue», nous entendons le décès ou l'émigration d'une personne; pareillement, par «unité ajoutée», nous entendons la naissance ou l'immigration d'une personne.

Nous étudions deux méthodes fondées sur le plan de sondage pour estimer les paramètres longitudinaux ayant trait à deux cycles. L'une s'appuie sur les échantillons transversaux pour l'estimation des bornes des catégories de revenu pour chaque période de référence et sur l'échantillon longitudinal pour l'estimation du nombre d'unités dans la population longitudinale (dénombrement longitudinal). Nous obtenons ainsi un estimateur que nous baptisons estimateur mixte. L'autre méthode repose sur un estimateur fondé sur l'échantillon longitudinal, aussi bien pour les bornes des catégories de revenu que pour les dénombrements longitudinaux. Nous l'appelons estimateur longitudinal. L'étude vise avant tout à comparer l'efficacité des deux méthodes quand on applique divers modèles de correction pour tenir compte de l'érosion de l'échantillon.

Pour faire cette comparaison, nous examinons deux problèmes connexes, à savoir l'effet de l'érosion sur les estimateurs étudiés et l'estimation de leur variance. Par érosion, nous entendons le type de non-réponse qui survient à partir d'un cycle donné jusqu'à la fin de la période d'observation. Le problème que pose réellement l'érosion tient au fait que les non-répondants s'accumulent au fil du temps, si bien que la non-réponse est d'autant plus importante que l'étude dure longtemps. Pour certaines enquêtes, comme la Survey of Income Program Participation (SIPP), l'érosion avait atteint 20% au moment du troisième cycle (Rizzo, Kalton et Brick 1996). Même si

[1] Susana Rubin Bleuer, Division des méthodes des enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario) K1A 0T6, courrier électronique: rubisus@statcan.ca; Milorad S. Kovačević, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario) K1A 0T6, courrier électronique: kovamil@statcan.ca.

l'on met minutieusement au point les corrections pour combler les lacunes de données, les estimateurs pourraient demeurer sensibles à un modèle de correction imparfait. Nous étudions empiriquement la sensibilité à l'érosion des estimateurs proposés.

L'estimation de la variance est également problématique, parce que les paramètres étudiés sont des fonctions non linéaires des observations qui dépendent des bornes des catégories de revenu. L'estimation de la variance des proportions de faibles revenus et d'autres mesures de l'inégalité du revenu d'après des échantillons transversaux complexes a été étudiée, entre autre, par Shao et Rao (1993), Binder et Kovačević (1995) et Kovacevic et Yung (1997). Dans le cas d'un échantillon longitudinal, les variations démographiques au fil du temps obligent à combiner divers échantillons et divers systèmes de pondération, situation qui complique l'estimation de la variance. Les unités qui constituent, en dernière analyse, un échantillon longitudinal pourraient appartenir à des unités primaires d'échantillonnage (UPE) différentes selon le cycle, certaines UPE présentes dans l'échantillon durant un cycle particulier pourraient ne plus y figurer lors d'un autre cycle, etc. Nous développons ici un estimateur bootstrap de la variance approprié aux estimateurs de la dynamique du revenu et au plan de sondage complexe utilisé dans l'exemple.

Les données utilisées à titre d'illustration proviennent de l'Enquête sur la dynamique du travail et du revenu (EDTR) et de la base de données longitudinales administratives (BDLA) de Statistique Canada. L'une et l'autre fournissent des données longitudinales sur le revenu assez précises extraites des déclarations de revenus. Au moment de l'étude, on ne disposait que des données de deux cycles de l'EDTR et le taux d'érosion était d'environ 10%.

À la section 2, nous décrivons les hypothèses générales concernant la population étudiée et le plan de sondage. À la section 3, nous examinons le problème de l'estimation des proportions longitudinales de faibles revenus et de l'effet de l'érosion sur cette estimation au moyen d'une petite étude en simulation portant sur un ensemble de données artificielles créé en supposant que la distribution est lognormale. À la section 4, nous décrivons l'estimation bootstrap de la variance dans le cas d'enquêtes longitudinales complexes. À la section 5, nous présentons une simulation plus poussée de divers modèles d'érosion, s'appuyant sur différentes méthodes de correction et sur des données provenant d'un plan de sondage complexe. Enfin, à la section 6, nous présentons les résultats et une discussion de ces derniers.

2. POPULATION, ÉCHANTILLON ET POIDS LONGITUDINAUX

Représentons la population au temps 0 par U_0 et la population au temps 1 par U_1. Nous considérons ici les paramètres ayant trait à deux périodes de référence et, par conséquent, nous définissons la population longitudinale en fonction de deux cycles par U_L, où $U_L = U_0 \cap U_1$.

Les «disparitions» et les «ajouts» de personnes qui surviennent d'une période de référence à l'autre causent une variation de la population. Si nous représentons par U_D l'ensemble de personnes qui appartiennent à la population U_0 au temps $t = 0$ mais qui n'appartiennent pas à la population U_1 au $t = 1$ à cause d'une «disparition», et par U_B l'ensemble d'«ajouts» survenus entre les temps 0 et 1, alors nous pouvons représenter la population longitudinale par l'expression $U_L = U_0 \setminus U_D = U_1 \setminus U_B$.

De la même façon, nous représentons par s_0 un échantillon représentatif de U_0, par s_1, un échantillon représentatif de U_1 et par s_d et s_b, les sous-échantillons respectifs d'unités de s_0 «disparues» entre les temps $t = 0$ et $t = 1$, et des unités de s_1 «ajoutées» entre le temps $t = 0$ et $t = 1$. Par conséquent, l'échantillon longitudinal, représentatif U_L, est défini par $s_L = s_0 \cap s_1 = s_0 \setminus s_d = s_1 \setminus s_b$.

Il existe des non-répondants lors du premier cycle de l'enquête, au temps $t = 0$, mais ils sont assez peu nombreux comparativement aux cycles ultérieurs. Par souci de simplicité, nous supposerons que s_0 représente l'échantillon dont sont exclus les non-répondants du cycle initial et dont la pondération connexe a été corrigée pour tenir compte de cette situation. On représentera l'érosion subie par l'échantillon du cycle 0 au cycle 1 par un sous-ensemble d'unités de s_0 appelé s_a. Par conséquent, nous pouvons représenter l'échantillon longitudinal touché par l'érosion par l'expression $s_L^A = s_0 \setminus (s_d \cup s_a)$.

Il convient de souligner que, pour certains paramètres étudiés, s_d devrait demeurer dans l'échantillon longitudinal aux fins de la pondération (Tambay, Schiopu-Kratina, Mayda, Stukel et Nadon 1997). Les paramètres examinés ici ont trait à la population longitudinale, U_L, et les unités qui «disparaissent» d'un cycle à l'autre sont donc hors du champ d'observation.

Les grandes enquêtes s'appuient souvent sur un plan de sondage stratifié à plusieurs degrés comptant de nombreuses strates dans chacune desquelles sont sélectionnées un assez petit nombre d'unités primaires d'échantillonnage (UPE). Les UPE sélectionnées sont sous-échantillonnées en une ou plusieurs étapes jusqu'à l'obtention des unités finales. Ici, nous supposons que les nombres de strates et de grappes dans les strates ne varient pas d'un cycle à l'autre de l'enquête.

Supposons que les échantillons transversaux s_t comprennent n_h UPE échantillonnées avec remise dans la strate h et m_{hi}^t unités échantillonnées dans la i-ième UPE dans la strate h, pour $t = 0, 1$, $h = 1, ..., H$ et $i = 1, ..., n_h$. Posons que $\{w_{hij}^t\}$, $j = 1, 2, ..., m_{hi}^t$ est l'ensemble de poids de sondage correspondant à l'échantillon transversal s_t. Nous supposons que les poids de sondage produisant des estimateurs approximativement non biaisés des totaux de population, si bien que $E_p(\sum_{s_t} w_{hij}^t) \approx N^t$, où N^t est la taille de U_t pour $t = 0, 1$. Ici, E_p représente l'espérance mathématique en regard du plan de sondage $p(s)$. Si l'ensemble s_a d'unités

disparues est grand, il faut corriger les poids de départ w_{hij}^0 pour tenir compte des unités manquantes et la somme de ces poids corrigés devrait correspondre, en moyenne, à la taille N_L de la population longitudinale:

$$E_p E_m\left(\sum_{s_0\setminus(s_d\cup s_a)} w_{hij}^L\right) \approx N_L.$$

Ici, E_m est l'espérance mathématique compte tenu du modèle hypothétique m de la probabilité de réponse.

Exemples:

1. Dans le cas de l'Enquête sur la dynamique du travail et du revenu (EDTR), chaque cycle comprend une composante supplémentaire formée des «cohabitants», c'est-à-dire les personnes qui vivent dans les ménages des membres de l'échantillon longitudinal (Lavallée et Hunter 1992). Le plan de sondage de l'EDTR est un plan stratifié à deux degrés contenant approximativement $H = 400$ strates lors de chaque cycle. Le nombre de grappes sélectionnées dans la strate h peut varier en cas de croissance de cette strate. On sélectionne ordinairement deux ou trois grappes. Si l'on sélectionne un nouveau panel ou qu'on en remplace un ancien entre le temps $t=0$ et le temps $t=1$, alors le nombre de grappes échantillonnées par strate peut varier.
2. La base de données longitudinales administrative (BDLA) de Statistique Canada est un échantillon longitudinal sélectionné d'après les fichiers de données administratives et représentatif de la population de personnes qui remplissent une déclaration de revenus une année donnée. La BDLA constitue une collection nombreuse de panels, puisqu'un panel est «ajouté» à chaque cycle (année). Ici, la non-réponse est de l'ordre de 5% de l'effectif de l'échantillon transversal de chaque année. Les échantillons longitudinaux de données administratives comme la BDLA ne subissent pas d'érosion, mais sont sujets à la non-réponse durant le cycle d'enquête (ordinairement causée par la production tardive de la déclaration de revenus, Rubin Bleuer 1996). Le plan de sondage de la BDLA est un plan non stratifié à un seul degré. Nous utilisons la BDLA comme point de départ pour simuler l'érosion de l'échantillon, parce que les données qu'elle contient sont représentatives de la population canadienne de revenus lors de chaque cycle.

3. ESTIMATION DANS DES CONDITIONS D'ÉROSION

Sans perte de généralité, nous allons maintenant définir et expliquer l'application des deux méthodes à l'estimation des proportions longitudinales de faibles revenus. À la section 6, nous présentons les résultats concernant les effets de l'érosion pour d'autres paramètres ayant trait à deux cycles, comme les flux bruts et les taux de transition. Nous supposons aussi que les nombres d'«ajouts» et de «disparitions» dans la population finie sont négligeables comparativement au taux d'érosion. En fait, nous supposons que $U_0 = U_1$ et que, même si les unités restent les mêmes aux deux points dans le temps, les revenus associés à ces unités peuvent varier.

Supposons que y_{hij}^t est la valeur de la caractéristique étudiée (revenu familial corrigé pour tenir compte de la taille de la famille) pour la j-ième unité finale dans la i-ième UPE de la strate $h, j=1,\ldots,M_{hi}^t\, i=1,\ldots,N_h^t, h=1,\ldots,H$ et $t=0,1$. Alors, la proportion longitudinale de personnes dont le revenu est inférieur ou égal à x au temps $t=0$ et dont le revenu est inférieur ou égal à y au temps $t=1$ est représentée par

$$F(x,y) = \frac{1}{N_L}\sum_{h=1}^{H}\sum_{i=1}^{N_h^0}\sum_{j=1}^{M_{hi}^0} I\left(y_{hij}^0 \le x\right) I\left(y_{hij}^1 \le y\right) \quad (3.1)$$

où, puisque les deux population coïncident, $M_{hi}^0 = M_{hi}^1$, et où

$$N_L = \sum_{h=1}^{H}\sum_{i=1}^{N_h^0} M_{hi}^0$$

coïncide avec la taille de la population originale U_0. I est la fonction indicateur des revenus inférieurs ou égaux à x et à y, respectivement. $F(x, y)$ est la fonction de distribution bivariée des revenus aux temps 0 et 1. Représentons maintenant par $M_0/2$ la moitié du revenu médian au temps $t=0$ et par $M_1/2$, la moitié du revenu médian au temps $t=1$. On peut alors définir la proportion longitudinale de faibles revenus comme

$$\theta = F(M_0/2, M_1/2). \quad (3.2)$$

Si l'on suppose que la réponse est complète et que $U_0 = U_1, \theta$ est la version bivariée de la proportion transversale de faibles revenus étudiée, entre autres, par Shao et Rao (1993). Dans un cadre de référence permettant le développement d'une théorie asymptotique dans l'espace du plan, et en assujettissant le plan de sondage et la distributions des revenus à certaines conditions de régularité, Shao et Rao ont prouvé que l'estimateur de la proportion transversale de faibles revenus, $\hat{F}_t(\hat{M}_t/2)$, est convergent quand le nombre d'UPE, N_{UPE}, devient infiniment grand, pour des plans de sondage généraux stratifiés à plusieurs degrés où les UPE sont échantillonnés avec remise. Le cadre de référence s'appuie sur les hypothèses selon lesquelles il existe i) une série de populations finies contenant soit un nombre croissant, N_{UPE}, d'UPE soit un nombre croissant d'unités indépendantes si la population ne comprend pas de grappes et ii) une série correspondante de plans de sondage probabilistes pour lesquels la taille n_{UPE} de l'échantillon de premier degré devient infiniment grande quand $N_{\text{UPE}} \to \infty$.

Il est facile d'extrapoler ce résultat à la situation longitudinale pour l'estimateur

$$\hat{\theta} = \hat{F}(\hat{M}_0/2, \hat{M}_1/2) \qquad (3.3)$$

si on émet les hypothèses que la population ne varie pas du temps $t = 0$ au temps $t = 1$ et que l'échantillon ne subit aucune érosion.

Représentons par $\hat{M}_t$ l'estimateur du revenu médian au temps t fondé sur l'échantillon transversal s_t et les poids transversaux correspondants $\{w_{hij}^t\}$, pour $t = 0, 1$

$$\hat{M}_t = \inf\left\{y_{hij}^t \in s_t \mid \hat{F}_t(y_{hij}^t) \geq 1/2\right\},$$

où

$$\hat{F}_t(y) = \sum_{s_t} w_{hij}^t I(y_{hij}^t \leq y) \Big/ \sum_{s_t} w_{hij}^t,$$

et représentons par $\tilde{M}_t$ l'estimateur du revenu médian au temps t fondé sur l'échantillon longitudinal s_L et sur les poids longitudinaux $\{w_{hij}^L\}$:

$$\tilde{M}_t = \inf\left\{y_{hij}^t \in s_L \mid \tilde{F}_t(y_{hij}^t) \geq 1/2\right\},$$

où

$$\tilde{F}_t(y) = \sum_{s_L} w_{hij}^L I(y_{hij}^t \leq y) \Big/ \sum_{s_L} w_{hij}^L.$$

Il existe alors deux moyens d'exprimer le paramètre longitudinal (3.2), à savoir:

$$\hat{\theta}_{\text{mixte}} = \tilde{F}(\hat{M}_0/2, \tilde{M}_1/2) = \sum_{s_L} w_{hij}^L I\left(y_{hij}^0 \leq \hat{M}_0/2\right) I\left(y_{hij}^1 \leq \tilde{M}_1/2\right) \Big/ \sum_{s_L} w_{hij}^L, \qquad (3.4)$$

et

$$\hat{\theta}_{\text{long}} = \tilde{F}(\tilde{M}_0/2, \tilde{M}_1/2) = \sum_{s_L} w_{hij}^L I\left(y_{hij}^0 \leq \tilde{M}_0/2\right) I\left(y_{hij}^1 \leq \tilde{M}_1/2\right) \Big/ \sum_{s_L} w_{hij}^L. \qquad (3.5)$$

Le premier estimateur est qualifié de «mixte» parce qu'il regroupe les échantillons longitudinaux et transversaux. Le deuxième se fonde uniquement sur l'échantillon longitudinal. On notera que, s'il n'y a aucun «ajout» ni «disparition» d'unités d'un cycle à l'autre de l'enquête, on ne peut estimer la médiane au temps $t=1$ que d'après l'échantillon longitudinal et, par conséquent, nous utilisons $\tilde{M}_1$ pour définir l'estimateur mixte.

En cas d'érosion de l'échantillon, la plupart des données manquantes pourraient correspondre à des personnes dont les caractéristiques diffèrent de celles du reste de la population et les estimations produites pourraient être biaisées si l'on omet de tenir compte de cette situation. Par conséquent, nous corrigeons les poids pour tenir compte des renseignements manquants conformément à un modèle. La sensibilité des estimations à la spécification erronée du modèle de correction se prononcera à mesure que l'érosion augmentera. Donc, il est souhaitable de choisir des estimateurs robustes en regard du modèle adopté pour la correction pour la non-réponse.

Afin de comparer la robustesse des estimateurs (3.4) et (3.5) en regard du modèle de correction pour la non-réponse, nous avons exécuté une étude en simulation simple qui nous a permis d'estimer empiriquement les valeurs probables (compte tenu du plan de sondage et du modèle d'érosion) de $\hat{\theta}_{\text{mixte}}$ et $\hat{\theta}_{\text{long}}$, si le modèle de correction est choisi correctement ou incorrectement.

Comme nous l'avons déjà fait remarquer, si la population ne varie pas d'un cycle à l'autre et qu'on ne sélectionne donc aucun nouvel échantillon au deuxième cycle pour tenir compte de la variation, l'estimation de la médiane au moment du deuxième cycle ne peut se fonder que sur l'échantillon longitudinal. Par conséquent, le seul écart entre les deux estimateurs tient à l'estimation de la mesure de faible revenu durant le premier cycle. Donc, sans perte de généralité, nous considérons le paramètre $\theta = F(M_0/2, \infty)$ pour notre simulation. Dans ce cas, θ coïncide avec la portion transversale de faibles revenus et l'estimateur de θ, si l'on suppose que la réponse est complète, $\hat{\theta}_{\text{cross}} = \hat{F}_0(\hat{M}_0/2)$, est convergent (et donc asymptotiquement non biaisé) quand N_{UPE} tend vers l'infini.

La simulation, décrite en détail à l'annexe A, consistait à simuler 1 000 échantillons d'effectif égal à 1 000 à partir d'une population de revenus lognormale similaire à la population canadienne de revenus. Pour commencer, nous avons sélectionné un échantillon aléatoire simple sans remise (EASSR) à partir d'une grande population finie de revenus, puis nous avons simulé l'érosion de cet échantillon. Ici, nous considérons un modèle d'érosion avec unités manquants au hasard, que nous appelons MAR pour «missing at random», si la probabilité de non-réponse durant le deuxième cycle est constante à l'intérieur des catégories de réponses, et nous considérons un modèle d'érosion à unités manquants complètement au hasard, que nous appelons MCAR pour «missing completely at random», si la probabilité de non-réponse durant le deuxième cycle est uniforme pour l'ensemble de la population. Pour commencer, nous avons simulé l'érosion selon un modèle à unités manquants au hasard (MAR), où la non-réponse était provoquée dans une catégorie de faibles revenus. Nous avons fait coïncider la borne de la catégorie de faibles revenus avec le premier quintile de la population finie, connu a priori. Pour chaque échantillon, nous avons calculé $\hat{\theta}_{\text{mixte}}$, $\hat{\theta}_{\text{long}}$ et $\hat{\theta}_{\text{cross}}$ en faisant les corrections conformément au modèle d'érosion correct (MAR) et à un modèle d'érosion MCAR. La moyenne arithmétique des estimations s'approche de la double espérance (en regard du modèle et du plan de sondage) des deux premiers estimateurs et s'approche de l'espérance en regard du plan de sondage pour $\hat{\theta}_{\text{cross}}$. Cette dernière espérance s'approche, à son tour, de la valeur du paramètre θ, puisque $\hat{\theta}_{\text{cross}}$ est asymptotiquement non biaisé quand $n_{\text{UPE}} \to \infty$. Si

l'on calcule la correction des poids conformément au modèle d'érosion MCAR (incorrect), on obtient empiriquement la relation suivante:

$$|E_p E_m(\hat{\theta}_{\text{mixte}}) - \theta| > |E_p E_m(\hat{\theta}_{\text{long}}) - \theta|,$$

où m désigne le modèle d'érosion simulé et p, le plan de sondage dans les conditions d'échantillonnage aléatoire simple sans remise. Notons que l'érosion des groupes de personnes à faible revenu entachera toujours l'estimateur de la médiane d'une erreur systématique par excès et que nous obtiendrons donc systématiquement $\hat{\theta}_{\text{mixte}} \leq \hat{\theta}_{\text{long}}$.

Le résultat quelque peu curieux ici est que l'estimateur qui se fonde sur le moins de renseignements est, en moyenne, celui dont la valeur est la plus proche de la valeur réelle. Autrement dit, l'intégration d'un grand nombre de renseignements n'améliore pas l'estimateur si ces renseignements ne sont pas utilisés correctement. Pareillement, si θ_+ représente la proportion de revenus supérieurs à la borne de catégorie de revenus estimée pour l'échantillon et que l'érosion est plus forte pour les catégories inférieures de revenus, nous observons systématiquement l'inégalité

$$\hat{\theta}_{\text{long},+} \leq \hat{\theta}_{\text{mixte},+},$$

et, comme pour la proportion de faibles revenus θ, l'estimateur de θ_+ qui se fonde sur moins d'information est, en moyenne, celui dont la valeur est la plus proche de la valeur réelle. La description de l'étude en simulation et les résultats numériques figurent à l'annexe A.

La question qui se pose maintenant est celle de savoir si l'erreur systématique causée par la spécification erronée du modèle est plus grande que l'augmentation de la variance due à l'érosion. Aux sections 5 et 6, nous examinons le problème en simulant l'érosion à l'aide de données de l'EDTR et de la BDLA, et en calculant θ, selon l'équation (3.3), $\hat{\theta}_{\text{mixte}}$ et $\hat{\theta}_{\text{long}}$, ainsi que la variance des estimateurs due au plan de sondage.

4. ESTIMATION BOOTSTRAP DE LA VARIANCE POUR LES ÉCHANTILLONS LONGITUDINAUX

Afin d'évaluer les deux méthodes d'estimation, nous devons les comparer sous l'angle de la variance et du biais pour divers scénarios d'érosion. Les estimateurs $\hat{\theta}_{\text{mixte}}$ et $\hat{\theta}_{\text{long}}$ définis à la section 3 sont des fonctions non linéaires des observations; en outre, les données sur le revenu proviennent d'enquêtes complexes. Nous ne pouvons donc exprimer la variance de ces estimateurs en termes simples et devons recourir aux méthodes de calcul de la variance approximative. Nous recherchons une méthode facile à appliquer à un grand nombre de paramètres complexes dans le cas de divers plans de sondage. Nous aimerions évaluer les deux méthodes d'estimation pour tout paramètre, en utilisant les mêmes critères et une méthode uniforme d'estimation de la variance. Nous nous concentrons sur le développement d'un estimateur bootstrap de la variance que l'on peut appliquer à un plan de sondage stratifié à plusieurs degrés longitudinal. Il importe de souligner que ce sont les unités primaires d'échantillonnage seulement qui sont rééchantillonnées et non les unités qu'elles contiennent.

Kovačević et Yung (1997) ont comparé plusieurs méthodes de rééchantillonnage et la méthode de linéarisation proposée par Taylor pour l'estimation de la variance d'estimateurs transversaux de l'inegalité du revenu dans le cas d'un plan de sondage complexe. Une étude en simulation les a porté à conclure que la méthode de linéarisation de Taylor au moyen de l'équation d'estimation est la meilleure (en ce qui concerne le biais relatif, les propriétés de couverture, la stabilité, la robustesse en regard des hypothèses, etc.) et que la méthode bootstrap se classe directement après.

Le calcul du nombre de personnes qui appartiennent à une catégorie de revenu au temps 0 et à une autre au temps 1 vise les unités de l'échantillon longitudinal s_L et le schéma d'échantillonnage bootstrap doit assurer que des unités soient sélectionnées dans s_L. Cependant, si nous limitons le rééchantillonnage aux unités de s_L, la variabilité ne sera pas suffisante pour permettre l'estimation convergente de la variance des estimateurs quintiles transversaux $\hat{M}_0$ et $\hat{M}_1$. Donc, l'échantillon bootstrap devrait contenir aussi des éléments de $s_0 \backslash s_1$, s_L et $s_1 \backslash s_0$ à chaque itération.

Supposons que le plan de sondage est un plan stratifié à deux degrés et que les unités primaires d'échantillonnage (UPE) sont exactement les mêmes aux temps $t=0$ et $t=1$, autrement dit qu'il n'y a ni «disparition» ni «ajout» d'UPE d'un cycle à l'autre.

Les premier et deuxième cycles de l'EDTR répondent à cette condition. Les «nouvelles unités» et les cohabitants qui apparaissent lors du deuxième cycle vivent dans les logements de personnes qui ont été sélectionnées durant le premier cycle. Chaque unité u dans s_0 ou s_1 est affectée à une UPE qui a été sélectionnée lors du premier cycle:

si $u \in s_0 \cap s_1$, alors nous affectons u à l'UPE qui correspond à son logement de départ, au temps $t=0$;

si $u \in s_0 \backslash s_1$, alors nous affectons u à l'UPE à laquelle elle appartenait au temps $t=0$;

si et $u \in s_1 \backslash s_0$ que u vit avec $v \in s_0 \cap s_1$, alors nous affectons u à l'UPE de v.

De cette façon, nous réduisons le problème à une situation transversale. Puis, nous procédons aux étapes qui suivent.

Supposons que les poids attribués au départ à une personne sont w_{hij}^0, w_{hij}^1, w_{hij}^L.

1. Nous sélectionnons un échantillon aléatoire simple avec remise (EASAR) d'UPE de taille $n_h - 1$, indépendamment dans chaque strate. Nous représentons par s_{boot}^* une union d'échantillon de ce genre. Elle contient un sous-échantillon $s_{0\,\text{boot}}^*$ d'unités provenant de s_0 qui

ne figurent pas dans s_1, un sous-échantillon $s_{1\,\text{boot}}^{*}$ d'unités provenant de s_1 qui ne figurent pas dans s_0 et un sous-échantillon $s_{L\text{boot}}^{*}$ d'unités qui figurent à la fois dans s_0 et s_1.

2. Représentons par m_{hi}^{*} le nombre de fois que la hi-ième UPE est sélectionnée; les modifications bootstrap des poids sont:

$$w_{hij}^{*(0)} = \frac{n_h}{n_h - 1}\, m_{hi}^{*}\, w_{hij}^{0};$$

$$w_{hij}^{*(1)} = \frac{n_h}{n_h - 1}\, m_{hi}^{*}\, w_{hij}^{1};$$

$$w_{hij}^{*(L)} = \frac{n_h}{n_h - 1}\, m_{hi}^{*}\, w_{hij}^{L}.$$

Nous obtenons les poids $w_{hij}^{(1)}$ et $w_{hij}^{(L)}$ d'après les poids de départ en les multipliant par un facteur de correction pour tenir compte de l'érosion. Les facteurs de correction sont égaux aux inverses des probabilités de réponse (que l'on suppose différentes pour chaque catégorie de réponse). Nous estimons ces probabilités d'après l'ensemble de données de départ. La méthode de rééchantillonnage bootstrap imite le processus d'estimation des probabilités de réponse et de corrections subséquente: les corrections sont recalculées pour chaque échantillon bootstrap s_{boot}^{*}, afin de produire de nouvelles valeurs de $w_{hij}^{(1)}$ et $w_{hij}^{(L)}$.

Alors, l'estimateur $\hat{\theta}_{\text{mixte}}$ calculé d'après un échantillon bootstrap s_{boot}^{*} prend la forme

$$\hat{\theta}_{mb}^{*} = \sum_{s_{L\text{boot}}^{*}} w_{hij}^{*(L)} I\left\{y_{hij}^{0} \le \hat{M}_0^{*}/2\right\} I\left\{y_{hij}^{1} \le \hat{M}_1^{*}/2\right\} \Big/ \sum_{s_{L\text{boot}}^{*}} w_{hij}^{*(L)},$$

où

$$\hat{M}_t^{*} := \inf\left\{y_{hij}^{t} \in s_{t\text{boot}}^{*} \cup s_{L\text{boot}}^{*} \mid \hat{F}_t^{*}(y_{hij}^{t}) \ge 1/2\right\}$$

et

$$\hat{F}_t^{*}(y) := \sum_{s_{t\text{boot}}^{*} \cup s_{L\text{boot}}^{*}} w_{hij}^{*(t)} I(y_{hij}^{t} \le y) \Big/ \sum w_{hij}^{*(t)}, \; t = 0, 1.$$

L'estimateur $\hat{\theta}_{\text{long}}$ calculé d'après $s_{L\text{boot}}^{*}$ prend la forme

$$\hat{\theta}_{lb}^{*} = \sum_{s_{L\text{boot}}^{*}} w_{hij}^{*(L)} I\left\{y_{hij}^{0} \le \tilde{M}_0^{*}/2\right\} I\left\{y_{hij}^{1} \le \tilde{M}_1^{*}/2\right\} \Big/ \sum_{s_{L\text{boot}}^{*}} w_{hij}^{*(L)},$$

où les médianes $\tilde{M}_{*}^{t}$ sont estimées d'après l'échantillon longitudinal $s_{L\text{boot}}^{*}$.

3. Nous répétons les étapes 1 et 2 un grand nombre de fois, disons B. Nous obtenons une estimation de Monte Carlo de la variance de la forme

$$v_B(\hat{\theta}_{\text{mixte}}) = \frac{1}{B} \sum_b \left(\hat{\theta}_{mb}^{*} - \bar{\hat{\theta}}_{\text{mixte}}\right)^2$$

et

$$v_B(\hat{\theta}_{\text{long}}) = \frac{1}{B} \sum_b \left(\hat{\theta}_{lb}^{*} - \bar{\hat{\theta}}_{\text{long}}\right)^2,$$

où

$$\bar{\hat{\theta}}_{\text{mixte}} = \sum_b \hat{\theta}_{\text{mb}}^{*} \Big/ B \text{ et } \bar{\hat{\theta}}_{\text{long}} = \sum_b \hat{\theta}_{lb}^{*} \Big/ B.$$

Grâce au rééchantillonnage de l'UPE nous réduisons l'estimation de la variance dans le contexte longitudinal à un cadre transversal. Cette méthode est une extension de l'estimateur bootstrap de la variance mis au point par Rao et Wu (1988) et, plus tard, par Kovačević et Yung (1997) pour estimer la variance des mesures transversales de l'inégalité du revenu d'après un plan de sondage stratifié à plusieurs degrés.

Afin de tenir compte de l'érosion de l'échantillon, nous traitons l'ensemble de données de départ comme s'il s'agissait d'un ensemble d'enregistrements longitudinaux. Nous pouvons alors considérer l'érosion comme un cas de non-réponse à une question et, conséquemment, la correction des poids pour tenir compte de l'érosion comme une imputation par la méthode des quotients (ratio de Hajek).

En effet, rappelons-nous que $s_0 = s_L \cup s_a$, et représentons par $y_{hij}^{(1)*}$ $(hij \in s_a)$ la valeur des données du cycle 1 imputées selon la méthode des quotients d'après les données observées pour l'échantillon longitudinal et représentons par $y_{hij}^{(0)*}$ $(hij \in s_a)$, la valeur des données du cycle 0 imputées selon la méthode des quotients d'après l'échantillon longitudinal (s_L). Remarquons que les valeurs $y_{hij}^{(0)}$ $(hij \in s_a)$ ne sont pas manquantes, mais que nous devons nous servir de $y_{hij}^{(0)*}$ au lieu de la correction de poids pour l'échantillon longitudinal.

L'estimation de θ si l'on corrige les poids pour tenir compte de l'érosion équivaut à l'estimation calculée d'après un ensemble de données «complètement» imputées, pondérées par $w_{hij}^{(0)}$, $(hij \in s_0)$. On se sert de l'ensemble

$$\mathcal{D}_{\text{mixte}} = \left\{(y_{hij}^{(0)}, y_{hij}^{(1)}),\, hij \in s_L,\, (y_{hij}^{(0)}, y_{hij}^{(1)*}),\, hij \in s_a\right\},$$

pour calculer $\hat{\theta}_{\text{mixte}}$ et de l'ensemble

$$\mathcal{D}_{\text{long}} = \left\{(y_{hij}^{(0)}, y_{hij}^{(1)}), hij \in s_L, (y_{hij}^{(0)*}, y_{hij}^{(1)*}), hij \in s_a\right\},$$

pour calculer $\hat{\theta}_{\text{long}}$.

Nous avons fait remarquer plus haut que la correction pour l'érosion équivaut à l'imputation selon la méthode des quotients pour tenir compte de la non-réponse à une question. Donc, l'estimateur de la variance proposé ici a les mêmes propriétés que les estimateurs transversaux de la variance utilisés pour les données d'enquête imputées: la convergente découle maintenant de la convergente de la

variance bootstrap pour les données d'enquête imputées (Shao et Sitter 1996), ainsi que des propriétés de bonne couverture et faible biais relatif décrites par Kovačević et Yung (1997), Kovar, Rao et Wu (1988) ont montré empiriquement que, si le nombre d'UPE par strate est faible (deux ou trois dans le cas de l'EDTR), l'estimation bootstrap de la variance n'excède pas de plus de 10% la valeur réelle de la variance.

5. ÉTUDE EMPIRIQUE

Afin de comparer les deux méthodes d'estimation, nous examinons maintenant deux enquêtes longitudinales réelles, à savoir l'EDTR et la base de données longitudinales administrative (BDLA), et nous simulons divers scénarios d'érosion de l'échantillon. Nous prenons pour point de départ une enquête à deux cycles avec échantillon complet de répondants à chaque cycle. Nous représentons le schéma de réponse des personnes formant cet échantillon comme suit:

RÉPONDANTS	$t=0$	$t=1$
Réponse lors des deux cycles	X	X

Ici, X indique que la personne a répondu au cycle en question et 0 signifie le contraire. Nous avons omis les «ajouts» et les «disparitions» d'unités aussi bien dans le cas de l'EDTR que de la BDLA.

Nous répartissons les personnes qui ont répondu lors du premier cycle en deux catégories de réponse, à savoir celles qui ont déclaré un faible revenu et les autres. La limite est donnée par $\hat{M}_0/2$, où $\hat{M}_0$ est l'estimation du revenu médian au temps $t=0$, compte tenu de tous les répondants du premier cycle. L'effectif de l'échantillon de l'EDTR (Ontario) est de l'ordre de 10 000 personnes, dont 2 000 ayant un faible revenu.

Nous simulons trois scénarios distincts d'érosion de l'échantillon. Dans chacun d'eux, nous sélectionnons un sous-échantillon de répondants ayant répondu complètement, selon le schéma XX, et nous le convertissons au schéma suivant:

RÉPONDANTS	$t=0$	$t=1$
Réponse au premier cycle, mais pas au deuxième	X	0

1) Érosion de 10%. Nous sélectionnons un sous-échantillon aléatoire de personnes ayant un faible revenu $\{y_{hij}^0 \le \hat{M}_0/2\}$ au temps $t=0$, et nous les considérons comme des non-répondants. Afin que 10% de l'échantillon global aient disparu au temps $t=1$, nous convertissons 50% du groupe de personnes à faible revenu en non-répondants au moment du deuxième cycle.

2) Érosion de 20%. Nous sélectionnons au hasard 70% de personnes à partir de $\{y_{hij}^0 \le \hat{M}_0/2\}$ et 7,5% à partir de $\{y_{hij}^0 > \hat{M}_0/2\}$ et nous les convertissons en non-répondants au moment du deuxième cycle. Nous obtenons ainsi un taux d'érosion global de 20% pour l'EDTR.

3) Érosion de 30%. Nous sélectionnons au hasard 80% de personnes à partir de $\{y_{hij}^0 \le \hat{M}_0/2\}$ et 17,5% à partir de $\{y_{hij}^0 > \hat{M}_0/2\}$ et nous les convertissons en non- répondants au moment du deuxième cycle. Nous obtenons ainsi un taux d'érosion de 30% pour l'ETDR.

Puis, nous envisageons deux modèles de correction distincts pour chaque scénario:

Modèle 1: Les non-répondants manquent complètement au hasard (MCAR). Ce modèle est le pire que nous puissions utiliser étant donné que l'érosion se manifeste ordinairement dans le groupe des personnes à faible revenu.

Modèle 2: Les non-répondants sont des unités qui manquent au hasard de la catégorie de faibles revenus. Nous permettons une légère augmentation de la borne supérieure de la catégorie de faibles revenus, de sorte que les catégories de réponses soient définies par $\{y_{hij}^0 \le \hat{M}_0/2 + \hat{M}_0/10\}$ et $\{y_{hij}^0 > \hat{M}_0/2 + \hat{M}_0/10\}$. Nous supposons que ce modèle est l'un des meilleurs qui soit compte tenu de nos paramètres, puisqu'il reconnaît les catégories de réponses comme étant séparées par la borne de faibles revenus.

Tableau 5.1
Certains paramètres longitudinaux évalués lors de l'étude empirique

Proportions	
Proportion de personnes dont le revenu (revenu familial corrigé) est inférieur à l_t au temps $t=0,1$	$p(y_0 \le l_0, y_1 \le l_1)$
Proportion de personnes dont le revenu est inférieur à l_0 au temps $t=0$ et supérieur à $l_1^* = (11/10)l_1$ au temps $t=1$; on utilise le facteur 11/10 afin de pouvoir déceler les transitions réelles d'un état à l'autre	$p(y_0 \le l_0, y_1 \ge l_1^*)$
Proportion de personnes dont le revenu est supérieur à $l_0^* = (11/10)l_0$ au temps $t=0$ et inférieur à l_1 au temps $t=1$	$p(y_0 \ge l_0^*, y_1 \le l_1)$
Proportion de personnes dont le revenu est supérieur à l_0^* et à l_1^* aux temps 0 et 1, respectivement	$p(y_0 \ge l_0^*, y_1 \ge l_1^*)$
Taux conditionnels	
Probabilité qu'une personne ait un faible revenu au deuxième cycle (deuxième année), étant donné qu'elle avait un faible revenu au premier cycle	$p(y_1 \le l_1 \mid y_0 \le l_0)$
Probabilité qu'une personne n'ait pas un faible revenu au deuxième cycle, étant donné qu'elle avait un faible revenu au premier cycle	$p(y_1 > l_1^* \mid y_0 \le l_0)$

Nous choisissons ces deux modèles de correction parce-qu'il représentent les deux situations extrêmes. En pratique, il se peut que nous ne puissions en choisir qu'un seul.

Représentons par l_t la mesure de faible revenu définie comme étant la moitié de la médiane: $l_t = M_t/2, t = 0,1$. On a étudié plusieurs paramètres longitudinaux dont certains apparaissent au tableau 5.1. Les valeurs des estimations sont présentées dans les tableaux B1 à B3. Nous avons déterminé les erreurs-types selon la méthode bootstrap décrite à la section 4, en supposant que les poids corrigés correspondants sont connus a priori et ne varient pas pour chaque échantillon bootstrap.

6. RÉSULTATS ET DISCUSSION

L'étude empirique montre que l'érosion de l'échantillon nuit aux estimations, mais que les résultats diffèrent selon que le paramètre étudié est transversal ou longitudinal. Si l'on estime les paramètres transversaux de cycles ultérieurs au cycle de départ, les estimateurs fondés sur l'échantillon longitudinal réel sont plus biaisés que ceux fondés sur l'échantillon transversal, que le modèle de correction soit valide ou non (voir, à titre d'exemple, les estimations de la médiane au temps $t = 0$ dans les tableaux B.1 à B.3.) Cependant, si l'on estime les paramètre longitudinaux, les estimateurs longitudinaux (fondés entièrement sur l'échantillon longitudinal) sont moins biaisés que les estimateurs mixtes (fondés sur les trois échantillons).

Les tableaux 6.1a et 6.1b présentent les flux bruts estimés d'après les données de l'EDTR et de la BDLA, respectivement. Nous avons calculé les estimations pour l'ensemble complet de données et après simulation d'une non-réponse de 20% lors du deuxième cycle. Pour l'ensemble complet de données, les estimateurs longitudinal et mixte coïncident (il s'agit du scénario sans érosion). Comme nous l'avons expliqué à la section précédente, nous avons simulé une non-réponse de 20% en éliminant les réponses de 70% de personnes dont le revenu était faible et de 7,5% de personnes dont le revenu était supérieur à l_0 lors du premier cycle. La correction pour la non-réponse a été effectuée en supposant que les personnes manquaient complètement au hasard.

Le modèle de correction choisi signifie que l'on a corrigé les poids de sondage de départ en leur appliquant un facteur de 1,25 (représentant une érosion de 20%) dans tout l'échantillon, alors qu'il aurait fallu, pour que la correction soit correcte, appliquer un facteur de 3,33 (représentant une érosion de 70%) au domaine des personnes ayant un faible revenu au moment du premier cycle (1993) et un facteur de 1,08 (représentant une érosion de 7,5%) à celui des personnes dont le revenu était supérieur à la mesure de faible revenu (MFR) en 1993. Donc, si nous nous servons d'un modèle de correction incorrect, nous commettons une erreur beaucoup plus importante lors de l'estimation d'un domaine ($y_o \le M_0/2$) que de l'autre ($y_o > M_0/2$).

Nous constatons, d'après ces tableaux, que tant l'estimateur mixte que l'estimateur longitudinal mène à une sous-estimation grave du paramètre étudié dans la première colonne et à une surestimation de celui étudié dans la deuxième colonne, si l'on suppose que les calculs fondés sur l'ensemble complet de données produisent des estimations raisonnables, de qualité acceptable. Il est manifeste que le choix incorrect du modèle de correction influe plus sur les estimations mixtes, que sur les autres, observation qui confirme les inégalités énoncées à la section 3.

Tableau 6.1
Flux bruts estimés d'après les données de l'EDTR et de la BDLA
Érosion de 20% (70% d'unités à faible revenu manquantes)

a. EDTR, Ontario

1994	1993		
	$y_o \le M_0/2$	$y_o > M_0/2$	
$y_1 \le M_1/2$	1 602 000	113 000	Pas d'érosion
	425 000	152 000	Mixte
	710 000	125 700	Longitudinal
$y_1 > 1,1 M_1/2$	70 000	8 080 000	Pas d'érosion
	15 000	8 975 000	Mixte
	30 000	8 870 000	Longitudinal

b. BDLA, sous-région de Toronto

1992	1991		
	$y_o \le M_0/2$	$y_o > 1,1 M_0/2$	
$y_1 \le M_1/2$	2 700	640	Pas d'érosion
	1 100	800	Mixte
	1 500	740	Longitudinal
$y_1 > 1,1 M_1/2$	580	10 420	Pas d'érosion
	190	12 150	Mixte
	380	11 650	Longitudinal

Les tableaux B.1 à B.3, qui sont présentés à l'annexe B, montrent les résultats de l'étude dans le cas de l'EDTR pour trois niveaux distincts d'érosion: 10%, 20% et 30%. Pour chaque paramètre, nous avons calculé les estimations et les erreurs-types bootstrap correspondantes au moyen des estimateurs longitudinal et mixte. Nous les avons d'abord calculées pour l'échantillon longitudinal idéal «sans érosion», puis, pour l'échantillon réduit, après correction selon les deux modèles de non-réponse décrits à la section 5. Nous présentons l'estimation du biais du au modèle sous la forme de la différence $\hat{\Delta}$ entre l'estimation obtenue en appliquant le modèle étudié et celle correspondant au scénario «sans érosion».

Les chiffres des tableaux B.1 à B.3 révèlent la même tendance que celle observée pour les flux bruts dans les tableaux 6.1a et 6.1b, à savoir que les estimations calculées grâce à l'estimateur longitudinal sont «moins sensibles» au choix du modèle de correction de la non-réponse. Parallèlement, nous n'observons pratiquement aucune différence entre les erreurs-types correspondantes (données entre parenthèses). Dans l'ensemble, nous constatons que l'estimation de l'erreur-type est un peu plus faible pour $\hat{\theta}_{\text{mixte}}$ que pour $\hat{\theta}_{\text{long}}$, et que cet écart négligeable en faveur de $\hat{\theta}_{\text{mixte}}$ n'est pas suffisant pour compenser le biais du à l'application d'un modèle incorrect de correction pour la non-réponse plus important pour $\hat{\theta}_{\text{mixte}}$.

Nous n'observons aucune différence entre $\hat{\theta}_{\text{mixte}}$ et $\hat{\theta}_{\text{long}}$ quand nous appliquons la deuxième méthode (par ordre

décroissant d'efficacité) de correction pour la non-réponse, et cela pour la plupart des paramètres, sauf les taux conditionnels. Par exemple, au tableau B.1 (érosion de 10%), le biais empirique de l'estimation mixte du taux conditionnel de personnes qui continuent à avoir un faible revenu en 1994 est statistiquement significatif, tandis que les biais empiriques des estimations longitudinales des deux taux conditionnels ne le sont pas (voir l'annexe C). Naturellement, le modèle «parfait» de correction (non présenté ici) produit exactement les mêmes chiffres pour les deux méthodes d'estimation, et $\hat{\theta}_{mixte}$ est approximativement égal à $\hat{\theta}_{long}$ pour tout paramètre étudié, mais les variances diffèrent.

Nous introduisons une «mesure de sensibilité» unique qui combine des renseignements sur l'erreur-type d'échantillonnage et sur le biais du au modèle choisi pour la correction pour l'érosion:

$$s(\hat{\theta}^A_{mixte}) = \left| \sqrt{\frac{(\hat{\theta}^A_{mixte} - \hat{\theta}_0)^2 + e.t.^2(\hat{\theta}^A_{mixte})}{e.t.^2(\hat{\theta}_0)}} - 1 \right|. \qquad (6.1)$$

Ici, $\hat{\theta}_0$ et $\hat{\theta}^A_{mixte}$ représentent les estimations obtenues dans le cas du scénario «sans érosion» et de l'application d'un modèle de correction pour l'érosion (A= MCAR ou MAR), respectivement et e.t.(.) représente l'erreur-type d'échantillonnage. De la même façon, nous définissons $s(\hat{\theta}^A_{long})$. Si une correction particulière pour l'érosion modifie peu la valeur d'un estimateur et de son erreur-type (comparativement à l'estimation obtenue pour une autre correction pour l'érosion), nous disons que l'estimateur est relativement insensible au modèle de correction appliqué. Nous exprimons les ratios des deux mesures de sensibilité pour les deux modèles appliqués par

$$ratio_{mixte} = s\left(\hat{\theta}^{MCAR}_{mixte}\right) \Big/ s\left(\hat{\theta}^{MAR}_{mixte}\right) \text{ et}$$

$$ratio_{long} = s\left(\hat{\theta}^{MCAR}_{long}\right) \Big/ s\left(\hat{\theta}^{MAR}_{long}\right) \qquad (6.2)$$

Les valeurs du ratio pour divers scénarios d'érosion sont présentées aux graphiques B.1 à B.3 (à l'annexe B).

Selon les graphiques B.1 à B.3, il est manifeste que le ratio des mesures de sensibilité (6.2) est systématiquement plus faible pour l'estimateur longitudinal, donc, que les mesures de sensibilité de l'estimateur longitudinal, étant donné les modèles de correction appliqués, sont plus semblables que celles obtenues pour l'estimateur mixte. L'estimateur longitudinal semble être moins sensible au modèle de correction appliqué. Nous utilisons le terme «robustesse» pour désigner cette caractéristique. En ce qui concerne les taux d'érosion simulés, les graphiques montrent que les mesures de la sensibilité tendent vers 1 quand le taux d'érosion augmente. Autrement dit, les deux modèles de correction donnent des résultats comparables pour les taux élevés d'érosion et le choix de l'estimateur devient alors plus important que celui du modèle utilisé pour les corrections.

Nous résumons nos observations comme suit:

1) Si l'on estime les paramètres transversaux de cycles ultérieurs au cycle de départ, les estimateurs fondés sur l'échantillon longitudinal courant sont plus biaisés que les estimateurs fondés sur l'échantillon transversal.
2) Si l'on estime les paramètres longitudinaux, les estimateurs mixte et longitudinal sont tous deux biaisés considérablement si l'on choisit le modèle inapproprié de correction pour l'érosion.
3) L'estimateur longitudinal est plus robuste en regard de la correction inappropriée pour tenir compte de l'érosion que l'estimateur mixte. Dans le cas du modèle de correction parfait, les deux estimateurs donnent les mêmes résultats.
4) En général, la variance d'échantillonnage de l'estimateur mixte est plus faible que celle de l'estimateur longitudinal. Cette relation se vérifie pour divers taux d'érosion et divers modèles de correction.
5) Étant donné son importance, le biais du à la correction incorrecte pour la non-réponse qui entache l'estimateur mixte annule le faible gain de précision obtenu comparativement à l'estimateur longitudinal.
6) Divers modèles de correction donnent des résultats comparables si le taux d'érosion est élevé. Le cas échéant, le choix de l'estimateur est plus important que les efforts en vue d'améliorer le modèle de correction.

REMERCIEMENTS

Les auteurs aimeraient remercier le rédacteur associé et les examinateurs pour leurs commentaires utiles et les suggestions.

ANNEXE A

Description de la simulation présentée à la section 3

La simulation comprend les étapes qui suivent.

1. Supposons que X est une variable aléatoire lognormale, $X \sim \exp\{N(\mu = 10{,}3, \sigma 2 = 0.64)\}$. Ces paramètres correspondent à la même médiane que celle de l'estimation de l'EDTR pour 1992 et à la même étendue que celle de la population du Canada. La borne de faible revenu coïncide avec le premier quintile de la population de revenu. Nous avons estimé ce premier quintile, q_1, d'après un échantillon simulé d'effectif égal à 50 000. La valeur obtenue est q_1 =14 901.
2. À partir de cette population infinie, nous avons sélectionné 1 000 échantillons aléatoires indépendants d'effectif égal à 1 000.
3. Pour simuler l'érosion, dans chaque échantillon, nous avons sélectionné au hasard 50% des unités situées au-dessous de q_1 et les avons supprimées de l'échantillon pour les calculs ayant trait au deuxième cycle. Nous avons ainsi simulé une érosion de 10% de la catégorie de faibles revenus (MAR modèle).
4. Pour chaque échantillon i, nous avons estimé les estimateurs de la proportion de faibles revenus

$\hat{\theta}_{mixte}(i)$ et $\hat{\theta}_{long}(i)$ en corrigeant les poids par application du modèle d'érosion approprié (MAR) ainsi que du modèle inapproprié supposant que les unités manquent complètement au hasard (MCAR).

5. En outre, pour chaque échantillon i, nous avons calculé l'estimateur transversal dans le cas où la réponse est complète, $\hat{\theta}_{mixte}(i)$ et $\hat{\theta}_{long}(i)$, qui se fonde entièrement sur l'échantillon avant érosion.
6. Pour chaque type de correction, nous avons estimé l'espérance mathématique de $\hat{\theta}_{mixte}$, $\hat{\theta}_{long}$ et $\hat{\theta}_{cross}$ en regard du plan de sondage et du modèle d'érosion, en calculant les moyennes mathématiques respectives sur 1 000 échantillons pour le modèle de correction inapproprié et sur 539 échantillons pour le modèle de correction approprié.

Résultat de la simulation

Les deux tableaux qui suivent montrent que, pour le modèle inapproprié de correction, l'estimateur longitudinal est entaché d'un biais plus faible que l'estimateur mixte et que, dans le cas du modèle approprié de correction, les deux estimateurs sont non biaisés (en ce qui concerne le modèle et le plan de sondage).

La moyenne arithmétique de $\hat{\theta}_{cross}$ sur 1 000 valeurs différentes est égale à 0,193 et l'erreur-type des 1 000 valeurs est égal à 0,012. Ces résultats sous-entendent que, pour un échantillon aléatoire simple sans remise dont l'effectif est égal à 1 000, l'estimation $\hat{\theta}_{cross}$ est assez stable et que l'on peut se servir de sa valeur probable pour remplacer θ. Les valeurs probables de $\hat{\theta}_{mixte}$ et $\hat{\theta}_{long}$ sont, selon les estimations, égales à 0,109 et 0,145, et leur erreur-type, à 0,011 et 0,013, respectivement.

Valeurs probables pour le modèle approprié de correction de l'érosion

Estimateur	Valeurs probables	Erreur-type	Nombre d'échantillons
$\hat{\theta}_{mixte}$	0,109	0,011	1 000
$\hat{\theta}_{long}$	0,145	0,013	1 000
$\hat{\theta}_{cross}$	0,193	0,012	1 000

Valeurs probables pour le modèle approprié de correction de l'érosion

Estimateur	Valeurs probables	Erreur-type	Nombre d'échantillons
$\hat{\theta}_{mixte}$	0,193	0,016	539
$\hat{\theta}_{long}$	0,193	0,014	539
$\hat{\theta}_{cross}$	0,194	0,012	539

Nous constatons d'après le tableau qui précède que les deux estimateurs s'approchent de la valeur réelle du paramètre si le modèle de correction est approprié.

ANNEXE B

Tableau B.1

Estimations de diverses caractéristiques de la population et de leur erreur-type calculées d'après l'ensemble complet de données et en supposant une érosion de 10% au deuxième cycle

Paramètre	Type d'estimateur	Pas d'érosion		Modèle de correction pour l'érosion					
				MCAR			MAR		
		$\hat{\theta}$	e.t.	$\hat{\theta}$	e.t.	$\hat{\Delta}$	$\hat{\theta}$	e.t.	$\hat{\Delta}$
Quintiles									
M_0 (médiane, t=0)	M	29 300	(1 000)	29 300	(1 000)	0	29 300	(1 100)	0
	L	29 300	(1 000)	30 900	(1 000)	-1 600	29 300	(1 200)	0
M_1 (médiane, t=1)	M	28 600	(1 100)	30 400	(900)	-1 800	28 600	(1 100)	0
	L	28 600	(1 000)	30 400	(1 000)	-1 800	28 600	(1 200)	0
Proportions									
$p(y_0 \le l_0, y_1 \le l_1)$	M	0,156	(0,010)	0,092	(0,008)	0,064	0,124	(0,009)	0,032
	L	0,156	(0,010)	0,111	(0,008)	0,045	0,124	(0,009)	0,032
$p(y_0 \le l_0, y_1 \ge l_1^*)$	M	0,007	(0,002)	0,003	(0,001)	0,004	0,005	(0,002)	0,002
	L	0,007	(0,002)	0,004	(0,003)	0,003	0,005	(0,002)	0,002
$p(y_0 \ge l_0^*, y_1 \le l_1)$	M	0,011	(0,002)	0,013	(0,003)	-0,002	0,012	(0,002)	0,001
	L	0,011	(0,002)	0,011	(0,003)	0,000	0,012	(0,002)	-0,001
$p(y_0 \ge l_0^*, y_1 \ge l_1^*)$	M	0,790	(0,040)	0,840	(0,040)	-0,050	0,804	(0,042)	-0,005
	L	0,790	(0,040)	0,831	(0,042)	-0,041	0,804	(0,043)	-0,005
Taux conditionnels									
$p(y_1 \le l_1 \mid y_0 \le l_0)$	M	0,923	(0,023)	0,546	(0,025)	0,377	0,734	(0,033)	0,189
	L	0,923	(0,023)	0,926	(0,030)	-0,003	0,921	(0,031)	0,002
$p(y_1 > l_1^* \mid y_0 \le l_0)$	M	0,040	(0,010)	0,018	(0,006)	0,022	0,030	(0,009)	0,010
	L	0,040	(0,010)	0,036	(0,010)	0,004	0,038	(0,012)	0,002

$l_i^* = (11/10)l_i$; l_i^* est utilisé pour repérer les transitions réelles d'un cycle à l'autre.

M représente les estimations calculées au moyen de l'estimateur mixte ($\hat{\theta}$) et L, les estimations calculées au moyen de l'estimateur longitudinal ($\tilde{\theta}$).

$\hat{\theta}$ est l'estimation, e.t. représente l'erreur-type de l'estimation et $\hat{\Delta}$ est l'écart entre les estimations obtenues en appliquant le modèle de correction pour l'érosion et en supposant qu'il n'y a pas d'érosion, respectivement.

Tableau B.2
Estimations de diverses caractéristiques de la population et de leur erreur-type calculées d'après l'ensemble complet de données et en supposant une érosion de 20% au deuxième cycle

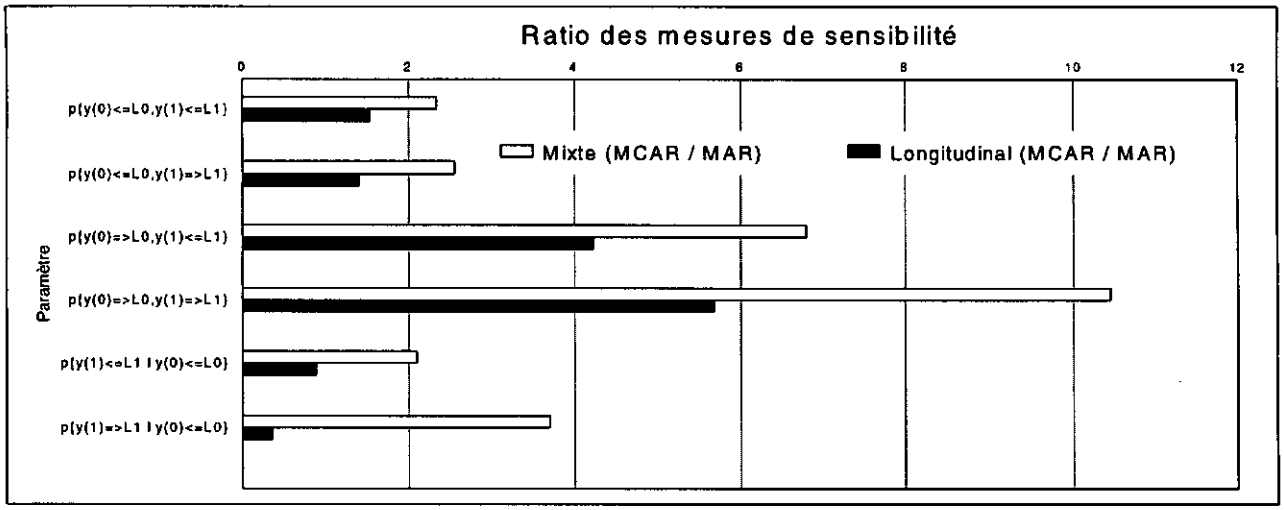
Paramètre	Type d'estimateur	Pas d'érosion		Modèle de correction pour l'érosion					
				MCAR			MAR		
		$\hat{\theta}$	e.t.	$\hat{\theta}$	e.t.	$\hat{\Delta}$	$\hat{\theta}$	e.t.	$\hat{\Delta}$
				Quintiles					
M_0 (médiane, t=0)	M	29 300	(1 000)	29 300	(1 000)	0	29 300	(1 000)	0
	L	29 300	(1 000)	31 800	(1 100)	-2 500	29 300	(1 000)	0
M_1 (médiane, t=1)	M	28 600	(1 100)	31 100	(800)	-2 500	28 600	(1 000)	0
	L	28 600	(1 000)	31 100	(900)	-2 500	28 600	(1 000)	0
				Proportions					
$p(y_0 \le l_0, y_1 \le l_1)$	M	0,156	(0,010)	0,055	(0,008)	0,101	0,096	(0,012)	0,060
	L	0,156	(0,010)	0,080	(0,006)	0,076	0,096	(0,013)	0,060
$p(y_0 \le l_0, y_1 \ge l_1^*)$	M	0,007	(0,002)	0,001	(0,001)	0,006	0,004	(0,002)	0,003
	L	0,007	(0,002)	0,001	(0,001)	0,004	0,004	(0,002)	0,003
$p(y_0 \ge l_0^*, y_1 \le l_1)$	M	0,011	(0.002)	0,014	(0,003)	-0,003	0,012	(0,002)	-0,001
	L	0,011	(0,002)	0,012	(0,002)	-0,002	0,012	(0,002)	-0,001
$p(y_0 \ge l_0^*, y_1 \ge l_1^*)$	M	0,790	(0,040)	0,864	(0,041)	-0,074	0,820	(0,049)	-0,030
	L	0,790	(0,040)	0,855	(0,041)	-0,065	0,820	(0,048)	-0,030
				Taux conditionnels					
$p(y_1 \le l_1 \mid y_0 \le l_0)$	M	0,923	(0,023)	0,323	(0,051)	0,600	0,570	(0,057)	0,353
	L	0,923	(0,023)	0,914	(0,040)	0,009	0,928	(0,058)	-0,005
$p(y_1 > l_1^* \mid y_0 \le l_0)$	M	0,040	(0,010)	0,007	(0,005)	0,033	0,026	(0,010)	0,014
	L	0,040	(0,010)	0,030	(0,012)	0,010	0,042	(0,014)	-0,002

Tableau B.3
Estimations de diverses caractéristiques de la population et de leur erreur-type calculées d'après l'ensemble complet de données et en supposant une érosion de 30% au deuxième cycle

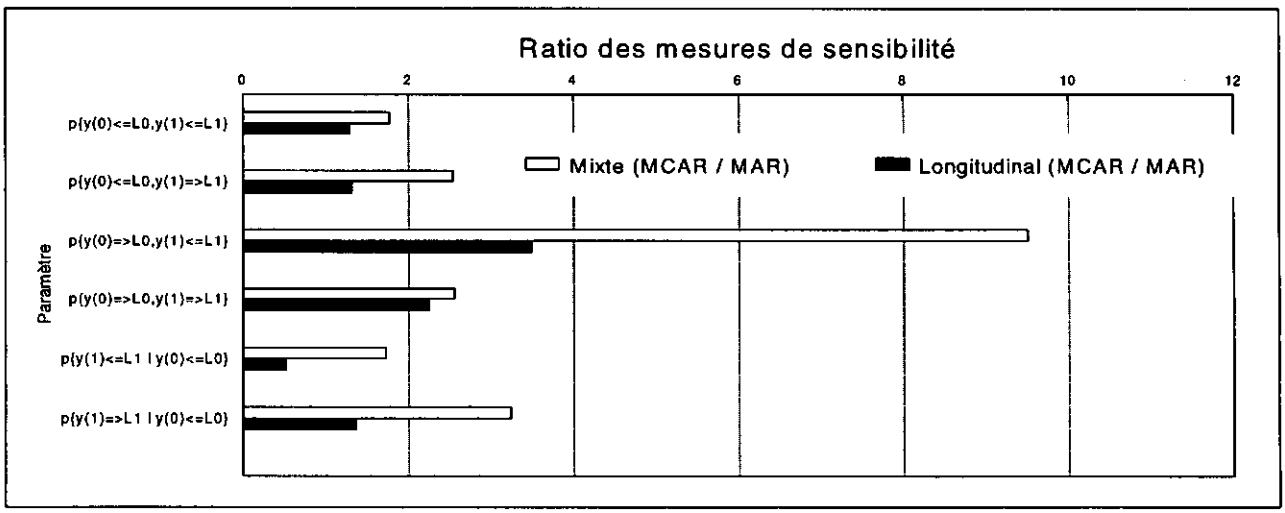
Paramètre	Type d'estimateur	Pas d'érosion		Modèle de correction pour l'érosion					
				MCAR			MAR		
		$\hat{\theta}$	e.t.	$\hat{\theta}$	e.t.	$\hat{\Delta}$	$\hat{\theta}$	e.t.	$\hat{\Delta}$
				Quintiles					
M_0 (médiane, t=0)	M	29 300	(1 000)	29 300	(1 000)	0	29 300	(1 000)	0
	L	29 300	(1 000)	32 000	(900)	-2 700	29 300	(1 000)	0
M_1 (médiane, t=1)	M	28 600	(1 100)	31 200	(800)	-2 600	28 600	(1000)	0
	L	28 600	(1 000)	31 300	(900)	-2 700	28 600	(1100)	0
				Proportions					
$p(y_0 \le l_0, y_1 \le l_1)$	M	0,156	(0,01)	0,04	(0,011)	0,116	0,080	(0,014)	0,076
	L	0,156	(0,01)	0,07	(0,006)	0,086	0,080	(0,016)	0,076
$p(y_0 \le l_0, y_1 \ge l_1^*)$	M	0,007	(0,002)	0,001	(0,001)	0,006	0,004	(0,002)	0,003
	L	0,011	(0,002)	0,003	(0,001)	0,004	0,004	(0,002)	0,003
$p(y_0 \ge l_0^*, y_1 \le l_1)$	M	0,011	(0.002)	0,015	(0,003)	-0,005	0,013	(0,002)	-0,002
	L	0,011	(0,002)	0,012	(0,003)	-0,001	0,013	(0,002)	-0,002
$p(y_0 \ge l_0^*, y_1 \ge l_1^*)$	M	0,790	(0,04)	0,874	(0,037)	-0,084	0,829	(0,049)	-0,039
	L	0,790	(0,04)	0,864	(0,038)	-0,074	0,828	(0,051)	-0,038
				Taux conditionnels					
$p(y_1 \le l_1 \mid y_0 \le l_0)$	M	0,923	(0,023)	0,245	(0,045)	0,678	0,465	(0,071)	0,458
	L	0,923	(0,023)	0,885	(0,045)	0,038	0,930	(0,099)	-0,007
$p(y_1 > l_1^* \mid y_0 \le l_0)$	M	0,040	(0,010)	0,008	(0,006)	0,032	0,022	(0,012)	0,018
	L	0,040	(0,010)	0,037	(0,016)	0,003	0,044	(0,017)	-0,004

Graphique B.
Ratio des mesures de sensibilité des estimateurs mixte et longitudinal

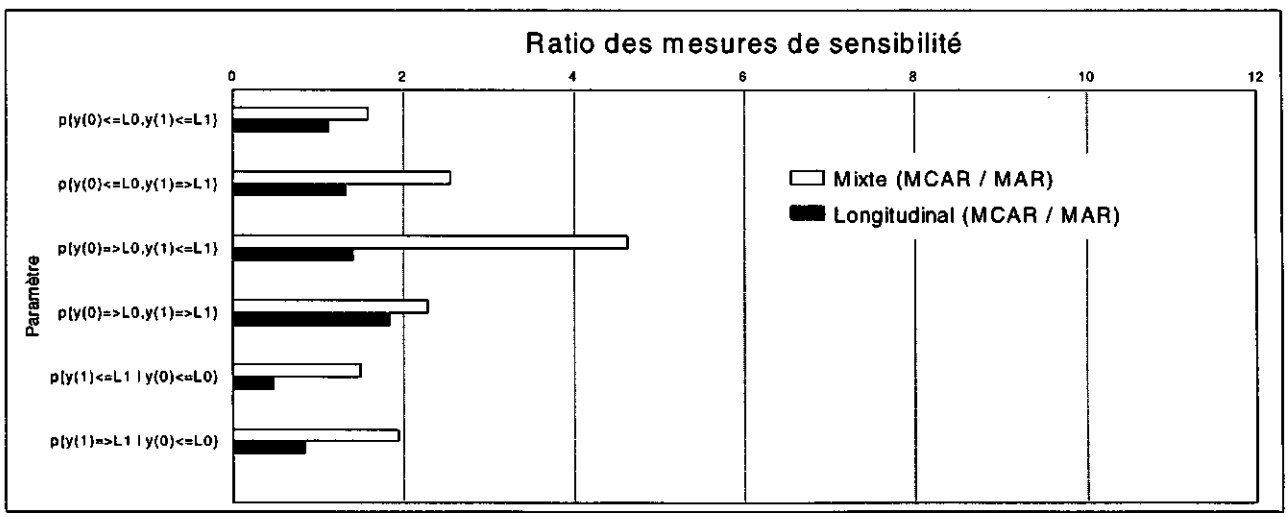
1. Taux d'érosion de 10% au deuxième cycle



2. Taux d'érosion de 20% au deuxième cycle



3. Taux d'érosion de 30% au deuxième cycle



ANNEXE C

Nous avons conclu que le biais empirique qui entâche les estimations mixtes (quand on applique un modèle de correction MAR) du taux conditionnel de personnes qui continuent d'avoir un faible revenu en 1994, étant donné que leur revenu était faible en 1993 était statistiquement significative après avoir exécuté un test prudent de la forme

$$\left(\hat{\theta}_0 - \hat{\theta}_{\text{mixte}}\right) \Big/ \sqrt{\text{var}(\hat{\theta}_{\text{mixte}}) + \text{var}(\hat{\theta}_0)}.$$

Nous avons conclu que le biais empirique qui entâche les estimations mixtes du taux conditionnel de personnes qui auront un revenu supérieur à la MFR en 1994 étant donné que leur revenu était faible en 1993 (avec application d'un modèle de correction MAR) n'était pas significatif après avoir exécuté un test «radical» de la forme

$$\left(\hat{\theta}_0 - \hat{\theta}_{\text{mixte}}\right) \Big/ \text{e.t.}\left(\hat{\theta}_{\text{mixte}}\right).$$

Pareillement, nous avons conclu que le biais empirique qui entâche les estimations longitudinales (quand on applique un modèle de correction MAR) des deux taux conditionnels n'était pas significatif après avoir exécuté le même type de test «radical» que celui susmentionné, c'est-à-dire en supposant que, dans le cas où il n'y a pas d'érosion, l'estimation est non stochastique.

Ces résultats restent vrais si le taux d'érosion est de 10%, 20% ou 30%.

BIBLIOGRAPHIE

BINDER, D., et KOVAČEVIĆ, M.S (1995). Estimation de l'inégalité du revenu d'après les données d'enquête: application de la méthode des équations d'estimation. *Techniques d'enquête*, 21, 151-159.

KOVAČEVIĆ, M.S., et YUNG, W. (1997). Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu - Étude empirique. *Techniques d'enquête*, 23, 47-59.

KOVAR, J.G., RAO, J.N.K., et WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *La Revue Canadienne de Statistique*, 16, 25-45.

LAVALLÉE, P., et HUNTER, L. (1992). Weighting for the survey of labour and income dynamics. Dans *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales,* Statistique Canada.

LEPKOWSKI, J.M. (1989). Non-response adjustments for wave non-response. Dans *Panel Surveys*, (Éds. D. Kasprzyk, et coll.). New York: John Wiley and Sons.

RAO, J.N.K., et WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of American Statistical Association*, 83, 231-241.

RIZZO, L., KALTON, G., et BRICK, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

RUBIN BLEUER, S. (1996). Gross flows estimation from longitudinal administrative data with missing waves. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* II, 681-686.

SHAO, J., et RAO, J.N.K. (1993). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyā*, B, 55, 393-414.

SHAO, J., et SITTER, R. (1996). Bootstrap for imputed data, *Journal of the American Statistical Association*, 91, 435, 1278-1288.

TAMBAY, J-L., SCHIOPU-KRATINA, I., MAYDA, J., STUKEL, D., et NADON, S. (1997). Traitement de la non-réponse du cycle deux de l'enquête nationale sur la santé de la population. *Techniques d'enquête*, 24, 159-169.

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 113-118
Statistique Canada

Utilisation de données reliées longitudinalement de l'Enquête britannique sur la population active

PAM F. TATE[1]

RÉSUMÉ

L'Enquête britannique sur la population active (EPA) utilise un plan d'échantillonnage avec renouvellement, chaque ménage de l'échantillon étant conservé pendant cinq trimestres consécutifs. Le fait de relier ensemble les renseignements sur les mêmes personnes d'un trimestre à l'autre produit une source potentiellement très riche des données longitudinales. Cependant, il y a de graves risques de distorsion dans les résultats du fait de cette liaison longitudinale, surtout en raison de l'érosion de l'échantillon, et des erreurs de réponse, ce qui peut produire de faux flux entre les situations au regard de l'activité économique. Le présent document décrit les premiers résultats d'investigations menées par l'Office for National Statistics (ONS) quant à la nature et à l'ampleur des problèmes.

MOTS CLÉS: Données longitudinales; enquête sur la population active; activité économique; biais relatif à l'érosion; erreur de réponse.

1. INTRODUCTION

L'Enquête britannique sur la population active (EPA) est une enquête-ménage qui permet de recueillir des renseignements sur un grand éventail de caractéristiques sur la population active ainsi que sur des sujets connexes. Depuis 1992, elle est réalisée trimestriellement, chaque ménage de l'échantillon étant conservé pendant cinq trimestres consécutifs, et un cinquième de l'échantillon étant remplacé à chaque trimestre.

L'enquête est conçue de façon à produire des données transversales, mais au cours des dernières années on a reconnu que le fait de relier ensemble des données sur chaque personne d'un trimestre à l'autre pouvait produire une riche source de données longitudinales, dont on pouvait se servir notamment pour l'estimation de flux bruts de la population active.

Le processus qui permet de relier des renseignements sur la même personne à partir de trimestres différents dans l'EPA est relativement simple. Cependant, il y a des problèmes d'ordre méthodologique qui posent de sérieux risques de distorsion dans les résultats provenant de cette nouvelle utilisation jusqu'à présent non vérifiée de données de l'EPA. Des problèmes semblables ont été identifiés dans des enquêtes sur la population active effectuées dans d'autres pays, mais il n'y a encore pour l'instant aucune méthode généralement acceptée pour les régler. L' Office for National Statistics (ONS) a par conséquent entrepris un programme de travail visant à aborder cette question.

Le présent document décrit les résultats jusqu'à présent des investigations sur la nature et l'ampleur des problèmes, ainsi que les méthodes proposées pour les régler. Les questions appartiennent à deux grands groupes: les biais découlant de l'érosion de l'échantillon et des facteurs connexes; et les biais découlant des erreurs de réponse, en particulier leurs répercussions du fait qu'elles produisent de faux flux entre les situations au regard de l'activité économique. On examine ces aspects l'un après l'autre.

2. ÉROSION DE L'ÉCHANTILLON ET SES RÉPERCUSSIONS EN MATIÈRE DE BIAIS

Certains membres de l'échantillon sont perdus dès le premier stade en raison de la non-réponse à la première interview, soit parce qu'il n'a pas été possible pour eux d'être contactés lors de la brève fenêtre disponible, soit parce qu'ils ont refusé d'être interviewés. Par la suite, d'autres membres de l'échantillon sont perdus chaque interview trimestrielle successive, soit parce qu'ils ont changé de logement (l'unité d'échantillonnage de base pour cette enquête étant le logement), soit parce qu'il s'avère impossible de les joindre ou qu'ils refusent de poursuivre. Tous ces groupes de personnes sont, de diverses façons, atypiques de la population dans son ensemble, de sorte que leur départ de l'échantillon peut introduire des biais.

Un certain nombre de ces biais sont compensés au moment de l'application de la procédure normale de pondération de l'EPA, qui produit des estimations au niveau de la population conformes aux totaux de contrôles fondés sur le recensement selon le sexe, le groupe d'âge et la région. Ce processus compensera les biais résultant, à toutes les étapes de l'enquête, de l'érosion différentielle selon le sexe, l'âge et la région. Cependant, des biais dans d'autres caractéristiques qui ne sont pas elles-mêmes utilisées dans la procédure de pondération ne seront pas compensés (et peuvent même être augmentés) dans ce processus, sauf lorsqu'ils ont trait à l'âge, au sexe ou à la région, de telle façon que le biais est causé entièrement par la sous-représentation ou la surreprésentation de catégories particulières d'âge, de sexe ou de région.

[1] Pam F. Tate, Office for National Statistics, Room RG/11, 1 Drummond Gate, London SW1V 2QQ, United Kingdom.

Par conséquent, le travail à cet égard a examiné d'abord les caractéristiques qui sont plus ou moins représentées dans l'échantillon de l'EPA que dans la population entière, et dans différents volets de l'échantillon de l'EPA. (Chaque trimestre, l'échantillon est formé de cinq volets, les personnes du premier volet ayant leur première interview, celles du deuxième volet ayant leur deuxième interview, et ainsi de suite.) On a alors examiné si ces caractéristiques étaient reliées l'une à l'autre et, le cas échéant, dans quelle mesure, et s'il était possible de définir un ensemble de variables qui caractérisent ces personnes susceptibles d'être sous- représentées.

3. CARACTÉRISTIQUES DES NON-RÉPONDANTS

Une analyse des proportions qui ne pouvaient être reliées au trimestre suivant, selon le volet, pour des variables démographiques et économiques clés (le tableau 1 donne une illustration de grands groupes d'âge/sexe), indiquait que, de façon uniforme d'un volet à l'autre, il y a une plus grande propension à la sous-représentation dans le cas des jeunes personnes âgées de 18 à 29 ans (et en particulier pour celles de 18 à 24 ans), les célibataires, les personnes vivant à Londres, les personnes à loyer (en particulier des loyers privés), les personnes en chômage, et les personnes occupant un emploi temporaire. On s'est rendu compte (Foster 1994 dans une étude qui reliait des données provenant du recensement de 1991 à des ménages non répondants de l'échantillon de l'EPA) que la plupart de ces caractéristiques étaient associées à une non-réponse élevée à la première interview, en particulier chez les jeunes adultes, les célibataires, les ménages à une personne et les personnes vivant à Londres.

Tableau 1
Pourcentage de cas non reliés selon le groupe d'âge et de sexe par volet

Variable et catégorie	Pourcentage non relié	
	1er volet	4^{ième} volet
Toutes les personnes	8,6	4,8
AGE et SEXE		
Homme	8,6	4,9
15-17	6,8	4,9
18-29	13,8	10,2
30-44	6,8	4,1
45-64	7,2	2,5
Femme	8,7	4,7
15-17	5,9	3,0
18-29	13,9	10,9
30-44	6,3	2,8
45-59	7,5	2,2

Note: On peut obtenir des analyses plus détaillées auprès de l'auteur.

Plusieurs des caractéristiques des personnes qu'a perdues l'échantillon semblent susceptibles d'être reliées, et l'on a enquêté sur cet aspect en tout premier lieu en utilisant la régression logistique. On a constaté que les variables identifiées comme étant associées avec le fait que les cas soient ou non perdus étaient très constantes pour les quatre volets. Dans chaque cas, elles incluaient le groupe d'âge, l'état matrimonial, le mode d'occupation, (c'est-à-dire si le logement était en propriété, en location d'un propriétaire privé, ou en location d'une autorité locale ou d'une association de logements), le niveau de qualifications et une variable définissant les grandes catégories d'activité économique (employé, en chômage ou inactif), et, dans le cas des personnes employées, le statut professionnel, à temps partiel/à temps plein et temporaire/ permanent. On a constaté que la région était associée de façon indépendante dans seulement deux des quatre volets, et le sexe dans aucun.

Pour ce qui est des cinq variables qui revenaient constamment pour tous les volets, il y avait un bon degré de constance en ce qui concerne les catégories associées à l'érosion de l'échantillon. Le tableau 2 donne les coefficients de multiplication pour les risques relatifs en ce qui concerne toutes les catégories ayant une association constante avec une érosion croissante. Le fait d'être dans les groupes d'âge les plus jeunes, entre 18 et 29 ans (et particulièrement entre 18 et 24 ans), a une incidence particulièrement forte, tout comme le fait d'être dans des logements loués dans le secteur privé. Le fait d'être célibataire (c'est-à-dire n'avoir jamais été marié et ne pas cohabiter) a une association modérée. Il n'y a aucune association constante avec des catégories particulières d'activité économique ou de niveau de qualification, sauf une légère avec les employés temporaires à plein temps. L'incidence de la région n'est pas constante, même pour les deux volets dans lesquels elle est présente.

Tableau 2
Coefficients de multiplication pour les risques relatifs - catégories associées à une érosion élevée

		Coefficient de multiplication pour les risques relatifs			
Variable	Catégorie	Volet 1	Volet 2	Volet 3	Volet 4
Groupe d'âge	18-19	1,89	2,56	2,86	1,92
	20-24	1,79	2,08	2,10	2,83
	25-29	1,17	1,30	1,44	1,55
Mode d'occupation	Location privée	2,12	1,52	1,86	2,29
	Célibataire	1,25	(1,12)	1,27	1,49
Activité économique/ statut	Employé, à plein temps, temporaire	(1,12)	(1,36)	(1,13)	1,75

Note: ( ) indique que le coefficient n'est pas significatif à 5%.

L'analyse de régression logistique exécutée ne permettait pas d'interaction entre les variables; pour investiguer cette possibilité, une analyse plus approfondie a été réalisée, à l'aide du module CHAID du SPSS pour produire une segmentation de l'ensemble de données en groupes qui ont un écart aussi grand que possible pour ce qui est de la proportion de cas non reliés. Cependant, les résultats de cette analyse ressemblaient beaucoup à ceux de l'analyse de régression logistique. Globalement, les principales caractéristiques associées de façon indépendante à une proportion élevée de perte de l'échantillon étaient les groupes d'âge des jeunes adultes (18 à 29 ans, en particulier de 18 à 24 ans) et les personnes vivant dans des logements loués du secteur privé, avec quelques répercussions additionnelles relativement mineures du fait d'occuper un emploi temporaire pour les groupes d'âge les plus jeunes. Les analyses distinctes des caractéristiques des membres de l'échantillon qui ont été perdus en raison d'un déménagement, et de ceux qui ont été perdus du fait qu'on n'a pu les contacter (ou, plus rarement, en raison d'un refus) ont produit des résultats semblables.

4. COMPENSER LE BIAIS DE L'ÉROSION

Jusqu'à présent l'analyse a porté sur l'effet de biais de l'érosion de l'échantillon sur les caractéristiques transversales de l'échantillon longitudinal, et a permis d'identifier les caractéristiques associées de façon indépendante à une non-réponse plus grande. Une méthode possible pour compenser le biais qui en découle est d'incorporer le mode d'occupation ainsi que l'âge dans la procédure de pondération pour les données longitudinales. On examine cette possibilité en se servant d'une méthode d'étalonnage avec le logiciel CALMAR, et en incluant des poids antérieurs obtenus des travaux décrits ci-dessus pour compenser la non-réponse différentielle selon le mode d'occupation.

Cependant, il peut y avoir un problème qui limiterait l'efficacité de cette méthode. La propension à ne pas répondre peut être directement dépendante du statut non observé au sein de la population active de la personne, et peut-être indépendamment de ses caractéristiques observées. Une non-réponse de ce genre est un exemple de non-réponse dont on a à tenir compte (Rubin 1976), et sa présence signifierait que les estimations des mesures importantes des flux bruts de la population active seraient biaisées même après l'application d'un processus de pondération du type envisagé.

Il y a deux méthodes indirectes qui donnent une certaine indication quant à savoir si une non-réponse dont on a à tenir compte pourrait constituer un problème pour les données longitudinales de l'EPA. On doit investiguer pour savoir si la proportion des flux bruts dans l'échantillon qui sont des transitions entre les différentes situations d'activité économique qui diminuent (ou augmentent) de façon systématique du volet 1-2 au volet 4-5 – le cas échéant, cela laisserait entendre que les gens qui passent d'un état à un autre sont plus (ou moins) susceptibles de ne pas répondre que ceux qui sont dans une situation stable. Cependant, le tableau 3 illustre qu'il n'y a aucun modèle systématique constant d'un volet à l'autre – quoique cela n'exclut pas la possibilité d'autres motifs de non-réponse différentielle selon la catégorie des flux de population active.

Tableau 3
Pourcentage de transitions entre les situations différentes au regard de l'activité économique selon le volet pour des paires de trimestres adjacents

Ensemble de données	Pourcentages			
	Volet 1-2	Volet 2-3	Volet 3-4	Volet 4-5
Été/automne 1995	8,0	7,3	7,3	7,1
Automne 1995/ hiver 1995-1996	7,2	6,7	6,5	6,5
Été/automne 1996	7,6	7,0	7,3	7,5
Automne 1996/ hiver 1996-1997	6,8	6,5	6,2	6,5

Une autre possibilité serait que des gens changent d'adresse (ils sont par conséquent perdus pour l'échantillon de l'EPA) peuvent avoir un motif différent de flux de la population active que le reste de la population. Nous ne disposons d'aucun renseignement sur les personnes qui ont déménagé, mais nous savons quelque chose au sujet des personnes qui sont passées d'une adresse ailleurs à une adresse au sein de l'échantillon. Ces nouveaux entrants peuvent raisonnablement être choisis pour représenter ceux qui ont déménagé à l'extérieur de l'échantillon, étant donné qu'ils sont également des échantillons provenant de la même population de personnes qui déménagent (sans tenir compte des effets possibles de la petite proportion de personnes qui déménagent à l'échelle internationale). Le tableau 4 illustre la répartition de l'échantillon relié (tous les adultes dont les dossiers correspondaient) et les nouveaux entrants identifiables pour une paire de trimestres adjacents en 1995. (Cependant, il est à remarquer que les catégories de flux ne sont pas strictement comparables, étant donné que la situation au regard de l'activité économique antérieure pour les personnes qui déménagent à l'intérieur de l'échantillon s'obtient par déclaration rétrospective.) Il est évident que l'échantillon de personnes ayant déménagé est différent, une plus petite proportion étant stable dans la catégorie des inactifs et une proportion plus élevée dans toutes les autres catégories de flux, et en particulier une plus grande proportion de gens changeant leur situation au regard de l'activité économique; mais que les personnes ayant déménagé prennent dans l'ensemble une si petite proportion que l'effet sur tout l'échantillon est négligeable.

Tableau 4
Flux bruts des personnes ayant déménagé par rapport à l'échantillon relié

Situations au regard de l'activité	Échantillon relié (%)	Personnes ayant déménagé (%)	Relié + personnes ayant déménagé (%)
EE	55,1	56,9	55,1
EU	0,8	1,5	0,8
EN	1,1	1,6	1,1
UE	1,0	1,7	1,0
UU	2,9	6,5	3,0
UN	0,7	1,1	0,7
NE	1,2	2,5	1,3
NU	1,0	2,3	1,1
NN	36,2	26,0	35,9
Toutes les transitions	5,9	10,6	6,0
TOTAL (n^{bre})	80 664	1 790	82 454

Note: E représente ayant un emploi
U représente sans emploi selon l'OIT
N représente économiquement inactif
donc EE représente ayant un emploi dans les deux trimestres
EU représente ayant un emploi, puis étant sans emploi au sens de l'OIT, etc.

Ces méthodes indirectes n'indiquent pas une très forte incidence de non-réponse dont on a à tenir compte, mais elles ne l'éliminent pas. Cette possibilité est par conséquent examinée dans le cadre de travaux faisant intervenir la modélisation de la non-réponse dans l'EPA.

5. L'ERREUR DE RÉPONSE ET SES RÉPERCUSSIONS EN MATIÈRE DE BIAIS

De façon générale, toutes les enquêtes, et en particulier les enquêtes-ménages, sont sujettes à l'erreur de réponse lorsque les renseignements donnés par le répondant ne sont pas un reflet exact de l'actualité. Cela peut se produire pour diverses raisons – le répondant peut mal comprendre la question; l'intervieweur peut mal comprendre ou mal enregistrer la réponse; le répondant peut ne pas connaître la bonne réponse ou ne pas s'en souvenir; ou le répondant peut sciemment donner une réponse incorrecte pour des motifs d'embarras, de prestige, de crainte de divulgation de renseignements confidentiels ou parce qu'il veut fournir la réponse «prévue».

Dans le domaine des enquêtes sur la population active, on a constaté de façon générale (pour avoir un aperçu, voir Lemaître 1994) que, pour les données transversales, il n'y a aucune tendance particulière à ce que les erreurs soient systématiques de sorte qu'en moyenne elles tendent à s'annuler. Cependant, dans le cas des données longitudinales produites par le fait de relier ensemble des données recueillies sur la même personne à des points différents dans le temps, cette annulation ne peut survenir.

En particulier, ce sera probablement le cas des données sur les flux bruts entre les situations au regard de l'activité économique. Les nombres de personnes qui passent d'une situation (ayant un emploi, en chômage, économiquement inactif) à une autre au cours de la période relativement brève habituellement envisagée (un mois, un trimestre, ou peut-être une année) sont faibles par rapport aux nombres de personnes qui demeurent dans la même situation. Une erreur de réponse à un point dans le temps est beaucoup plus susceptible de mener à un changement apparent de situation lorsque la véritable situation est stable, plutôt que l'inverse. Ainsi, des erreurs de réponse sont susceptibles d'avoir une incidence très disproportionnée en biaisant à la hausse les flux entre les situations signalées. En ce qui concerne l'EPA, elles peuvent découler de l'utilisation d'enquêtés-substituts, alors qu'une personne répond aux questions au nom d'une autre personne dans le même ménage; et d'erreurs de répondant. Nous allons examiner ces aspects l'un après l'autre.

6. L'INCIDENCE DES ENQUÊTÉS-SUBSTITUTS

Pour investiguer l'incidence des enquêtés-substituts, nous avons besoin d'examiner la répartition des situations au regard de l'activité pendant les deux trimestres en fonction du fait que l'interview du premier trimestre était sur place ou par personne interposée, et de savoir si l'interview du deuxième trimestre était sur place ou par personne interposée. De très jeunes adultes, moins de 20 ans, sont à la fois exceptionnellement susceptibles d'être représentés par des enquêtés-substituts et aussi susceptibles d'être particulièrement volatils en ce qui concerne leur catégorie d'activité économique; ils peuvent donc entraîner une distorsion de la relation entre ces deux facteurs. Par conséquent, le tableau 5 illustre la répartition des situations au regard de l'activité pendant les deux trimestres, pour les hommes de 20 à 64 ans et les femmes de 20 à 59 ans. Il y a une proportion plus élevée de transitions dans le cas des interviews sur place suivies des interviews par personne interposée que dans le cas des interviews sur place pendant les deux trimestres, mais les interviews par personne interposée suivies des interviews sur place illustrent seulement une proportion très légèrement plus forte que pour les interviews sur place pendant les deux trimestres. Ainsi, le fait de passer d'une interview sur place à une interview par personne interposée n'indique pas une proportion constamment plus grande de transitions. Les cas pour lesquels les deux interviews sont par personne interposée ont la plus faible proportion de transitions de toutes, et l'inclusion de ces dernières amène la proportion globale à un niveau conforme à celui des interviews sur place pendant les deux trimestres. Ainsi, il semble y avoir des différences entre les diverses combinaisons d'interviews qui méritent

une analyse plus approfondie; mais dans l'EPA, le recours à des enquêtés-substituts ne produit pas en soi une estimation exagérée de flux bruts.

Tableau 5
Pourcentage de transitions selon le type d'interview

Type d'interview	Hommes (20-64)		Femmes (20-59)	
	N^bre dans l'échantillon	% trans.	N^bre dans l'échantillon	% trans
Sur place/ sur place	14 527	5,2	19 582	7,3
Sur place/ personne interposée	2 044	7,0	1 597	8,3
Personne interposée/ sur place	2 214	5,4	1 632	7,6
Personne interposée/ personne interposée	8 602	4,7	4 206	5,4
Toutes	27 387	5,2	27 017	7,1

7. ERREURS DE RÉPONDANT

De par leur nature, les erreurs de répondant sont impossibles à identifier directement (sauf peut-être à la suite d'une autre interview, et même là il pourrait y avoir des doutes quant à ce qui constitue la bonne réponse). Cependant, il est parfois possible d'identifier des incohérences internes dans les données d'enquête, ce qui peut indiquer une erreur de réponse. Dans l'EPA, les répondants qui ont un emploi, et les répondants qui sont en chômage, se font demander combien de temps ils sont dans cette situation. Si la période est supérieure à trois mois, mais qu'ils ont dit dans le trimestre précédent qu'ils étaient dans une situation différente, il y a une incohérence qui peut signaler une fausse transition entre les situations au regard de l'activité économique.

Le tableau 6 illustre le pourcentage d'incohérences pour divers types de transitions – il s'agit de transitions élevées partout. Les transitions venant de l'inactivité économique produisent les plus forts pourcentages, en particulier lorsque la transition est vers le chômage. (Il n'y a aucune différence importante ou constante entre les différentes sous-catégories de l'inactivité.) Le fait de séparer les personnes ayant un emploi à temps partiel et à plein temps indique qu'il y a un motif constant d'une plus grande proportion d'incohérences pour les emplois à temps partiel, et on a relevé des résultats semblables mais moins prononcés dans le cas des travailleurs autonomes.

Il est possible que les incohérences puissent être dues à des erreurs dans la durée signalée pour ce qui est de la situation au regard de l'activité économique au deuxième trimestre plutôt que dans la situation initiale au premier trimestre. La répartition de la durée n'illustre toutefois pas une accumulation à environ quatre à cinq mois (comme on pourrait s'y attendre dans le cas d'erreurs dans les données relatives à la durée). De même, on a constaté que les données relatives à la durée signalées dans des trimestres consécutifs pour des personnes dans une situation stable étaient très cohérentes. Ces constatations tendent à indiquer, quoique les preuves soient indirectes et aucunement définitives, que les erreurs sont plus susceptibles d'être dues au signalement de l'activité économique à l'une ou l'autre des interviews. Il ne s'agit pas de la seule possibilité – par exemple, il se pourrait que l'on ait pour certains répondants les bonnes données relatives à la transition, mais les mauvaises données relatives à la durée en raison d'une interprétation de leur activité économique antérieure qui n'est pas cohérente avec les définitions standard appliquées au signalement de leur situation actuelle – mais les constatations indiquent jusqu'à présent que c'est probablement la plus courante.

Tableau 6
Pourcentage d'incohérences selon le type de transition

Type de transition	Pourcentage d'incohérences		
	Tous (%)	Temps plein (%)	Temps partiel (%)
Chômage à employé	8,7	7,8	12,2
Inactif à employé	26,2	18,1	30,4
Employé à chômage	18,7	14,7	23,3
Inactif à chômage	49,5		
Tous	23,9		

On pourrait peut-être déterminer laquelle parmi les catégories incohérentes est la bonne en examinant la tendance des réponses sur trois interviews. Le tableau 7 illustre les proportions de chaque groupe de transitions incohérentes d'un trimestre au suivant qui sont suivies par chaque catégorie d'activité économique au troisième trimestre. (Tous les volets pertinents sont combinés dans le but d'obtenir des échantillons d'une taille raisonnable.) Il est évident que dans le cas des transitions vers l'emploi au deuxième trimestre, la grande majorité demeure dans cette catégorie au troisième trimestre. Les transitions vers le chômage illustrent une tendance beaucoup plus mélangée, un peu plus de la moitié restant en chômage, mais un groupe important d'environ 30 à 40% revenant à la situation signalée au premier trimestre. Il convient de noter que pratiquement aucune des transitions du deuxième au troisième trimestres pour ce groupe ne comportait d'incohérences répétées entre les données relatives à la transition et les données signalées relatives à la durée. Jusqu'à maintenant, les résultats indiquent que dans le cas d'une

transition incohérente vers l'emploi, c'est probablement la bonne situation, mais des recherches plus poussées sont nécessaires pour clarifier le tout.

Tableau 7
Pourcentages de transitions incohérentes selon l'activité économique au trimestre suivant

Type de transition	Total des incohérences	Situation au regard de l'activité au trimestre suivant		
		Emploi (%)	Chômage (%)	Inactif (%)
Chômage à employé	60	90	7	3
Inactif à employé	159	79	4	17
Employé à chômage	87	39	53	8
Inactif à chômage	229	17	55	28

8. AJUSTEMENT EN FONCTION DU BIAIS ATTRIBUABLE À L'ERREUR DE RÉPONSE

Il est évident selon ce qui précède qu'il est probable qu'un niveau important d'erreurs de réponse influe sur les données brutes relatives aux flux bruts. Les travaux pour ajuster de telles erreurs se sont en grande partie limités au Canada et aux États-Unis. Flaim et Hogue (1985) font un examen des trois méthodes proposées en ce qui concerne les données aux États-Unis, et Singh et Rao (1995) ont par la suite fait une proposition en ce qui concerne les données canadiennes. Cependant, jusqu'à maintenant, à la connaissance de l'auteur, aucune donnée officielle ajustée des flux bruts n'est publiée, quoique plusieurs pays publient des données non corrigées tout en attirant l'attention sur leurs limites. Les méthodes d'ajustement proposées jusqu'ici reposent toutes sur des hypothèses en rapport avec la nature des erreurs qui semblent improbables en pratique – soit une indépendance complète des erreurs de classification ou des écarts très limités de cette hypothèse. (Voir Lemaître 1994)pour une révision des problèmes associés à ces méthodes d'ajustement.)

Il semble intéressant d'explorer des avenues différentes pour la mise au point de méthodes d'ajustement ou de compensation du biais attribuable à l'erreur de réponse. Dans une première étape, les travaux se poursuivent sur l'examen des caractéristiques et circonstances de cas d'incohérence, et d'autres façons possibles d'identifier les fausses transitions. On propose également d'investiguer les circonstances concernant les personnes qui donnent des réponses incohérentes du type que l'on a analysé précédemment, au moyen d'interviews de suivi plus détaillées. On obtiendrait ainsi de meilleures indications de la mesure dans laquelle les incohérences représentent une erreur de réponse, et on pourrait en retirer des résultats utiles pour réduire et ajuster l'erreur de réponse. Ces deux voies fourniront des données à un troisième élément du programme, dans lequel il est proposé de mettre au point des modèles d'erreur de classification dans le signalement de l'activité économique.

REMERCIEMENTS

L'auteur aimerait remercier le rédacteur ainsi que les deux arbitres poour leurs commentaires.

BIBLIOGRAPHIE

FLAIM, P. O., et HOGUE, C. R. (1985). Measuring labor force flows: a special conference examines the problems. *Monthly Labor Review*, Juillet 1985, U.S. Bureau of Labor Statistics.

FOSTER, K. (1994). The Labour Force Survey – Report of the 1991 Census-linked Study of Survey Non-respondents. Office of Population Censuses and Surveys.

LEMAÎTRE, G. (1994). Data on Labour Force Dynamics from Labour Force Surveys. L'Organisation de coopération et de développement économique.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

SINGH, A.C., et RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90, 478-488.

Techniques d'enquête, juin 1999
Vol. 25, n° 1, pp. 119-120
Statistique Canada

Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe

SIEGFRIED GABLER, SABINE HAEDER et PARTHA LAHIRI[1]

RÉSUMÉ

Les auteurs de la présente note montrent que la formule bien connue de l'effet de plan de sondage, proposée intuitivement par Kish, comporte une justification à base de modèle. La formule peut être interprétée comme une valeur prudente de l'effet réel de plan de sondage.

MOTS CLÉS: Taille des grappes; coefficient de corrélation intraclasse; probabilités de sélection.

1. INTRODUCTION

Nous considérons des plans de sondage en grappes à plusieurs degrés pour lesquels chaque observation appartient à une classe de pondération. Les grappes sont, par exemple, des îlots qui sont choisis proportionnellement au nombre de ménages qu'ils comportent. À l'intérieur de chaque îlot, le même nombre de ménages est choisi avec égale probabilité. Une personne du ménage choisie au hasard doit être interviewée. Ensuite, la taille des ménages détermine la classe de pondération. Kish (1987) a proposé la formule suivante afin de déterminer l'effet de plan de sondage de manière à intégrer les effets tant de la pondération pour tenir compte des probabilités de sélection inégales que de la sélection en grappes:

$$\text{deff}_{\text{Kish}} = m \frac{\sum_{i=1}^{I} w_i^2 m_i}{\left(\sum_{i=1}^{I} w_i m_i\right)^2} [1 + (\bar{b} - 1)\rho],$$

où m_i et w_i désignent respectivement le nombre d'observations et le poids attribué à la i-ième classe de pondération $(i = 1, ..., I)$, $m = \sum_{i=1}^{I} m_i$, la taille totale de l'échantillon, $\bar{b}$ la taille moyenne des grappes et ρ le coefficient de corrélation intraclasse. La formule de Kish est très intuitive et originale, mais Kish a affirmé que son «traitement était peut-être incomplet et imparfait».

La formule de Kish est maintenant utilisée par de nombreux échantillonneurs. La formule ci-dessus sera utilisée dans la détermination de la taille de l'échantillon des enquêtes sociales européennes qui doivent être menées par les pays membres. La présente note fournit une justification à base de modèle de l'utilisation de la formule de Kish.

2. JUSTIFICATION À BASE DE MODÈLE DE LA FORMULE DE KISH

Soit m_{ic}, le nombre d'observations de la c-ième grappe de référence appartenant à la i-ième classe de pondération $(i = 1, ..., I; c = 1, ..., C)$. Soit $m_i = \sum_{c=1}^{C} m_{ic}$, le nombre d'observations de la i-ième classe de pondération. Soit $b_c = \sum_{i=1}^{I} m_{ic}$, le nombre d'observations de la c-ième grappe $(i = 1, ..., I; c = 1, ..., C)$ de manière que $\bar{b} = C^{-1} \sum_{c=1}^{C} b_c$. Soit y_{cj} et w_{cj}, l'observation et le poids respectivement de la j-ième unité d'échantillonnage de la c-ième grappe $(c = 1, ..., C; j = 1, ..., b_c)$. L'estimateur habituel à base de plan de sondage pour la moyenne de la population se définit comme suit:

$$\bar{y}_w = \frac{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj} y_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}}.$$

Afin de justifier la formule de Kish, nous supposons le modèle suivant:

$$\text{Var}(y_{cj}) = \sigma^2 \quad \text{pour } c = 1, ..., C; j = 1, ..., b_c$$

$$\text{Cov}(y_{cj}, y_{c'j'}) = \begin{cases} \rho\sigma^2 & \text{si } c = c'; j \neq j' \\ 0 & \text{autrement.} \end{cases} \qquad (1)$$

Le modèle ci-dessus permet de tenir compte de l'effet de grappe; il a été utilisé antérieurement par d'autres auteurs (voir par exemple Skinner, Holt et Smith 1989). Nous définissons donc l'effet de plan de sondage comme suit: $\text{deff} = \text{Var}_1(\bar{y}_w)/\text{Var}_2(\bar{y})$, où $\text{Var}_1(\bar{y}_w)$ est la variance de $\bar{y}_w$ dans le cadre du modèle (1) et $\text{Var}_2(\bar{y})$ est la variance de la moyenne globale de l'échantillon $\bar{y}$, définie comme $\sum_{c=1}^{C} \sum_{j=1}^{b_c} y_{cj}/m$, que l'on calcule en fonction du modèle suivant:

$$\text{Var}(y_{cj}) = \sigma^2 \quad \text{pour } c = 1, ..., C; j = 1, ..., b_c$$

$$\text{Cov}(y_{cj}, y_{c'j'}) = 0 \quad \text{pour tous les } (c, j) \neq (c', j'). \qquad (2)$$

À noter que le modèle (2) est approprié dans le cadre de l'échantillonnage aléatoire simple et qu'il fournit la formule habituelle σ^2/m pour $\text{Var}_2(\bar{y})$.

[1] Siegfried Gabler et Sabine Haeder, ZUMA, B 2,1, D-68159 Mannheim; Partha Lahiri, University of Nebraska-Lincoln, NE 68588-0323.

Nous considérons maintenant $\mathrm{Var}_1(\bar{y}_w)$, en remarquant d'abord que

$$\mathrm{Var}_1\left(\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}y_{cj}\right)$$

$$=\sum_{c=1}^{C}\mathrm{Var}\left(\sum_{j=1}^{b_c} w_{cj}y_{cj}\right)$$

$$=\sum_{c=1}^{C}\left\{\sum_{j=1}^{b_c} w_{cj}^2\mathrm{Var}(y_{cj})+\sum_{j\neq j'}^{b_c} w_{cj}w_{cj'}\mathrm{Cov}(y_{cj},y_{cj'})\right\}$$

$$=\sigma^2\sum_{c=1}^{C}\left\{\sum_{j=1}^{b_c} w_{cj}^2+\rho\sum_{j\neq j'}^{b_c} w_{cj}w_{cj'}\right\}$$

$$=\sigma^2\left\{\sum_{i=1}^{I} w_i^2m_i+\rho\sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_im_{ic}\right)^2-\rho\sum_{i=1}^{I} w_i^2m_i\right\} \qquad (3)$$

puisque $\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2=\sum_{i=1}^{I} w_i^2m_i$ et

$$\sum_{c=1}^{C}\sum_{j\neq j'}^{b_c} w_{cj}w_{cj'}=\sum_{c=1}^{C}\left\{\left(\sum_{j=1}^{b_c} w_{cj}\right)^2-\sum_{j=1}^{b_c} w_{cj}^2\right\}$$

$$=\sum_{c=1}^{C}\left\{\left(\sum_{i=1}^{I} w_im_{ic}\right)^2-\sum_{i=1}^{I} w_i^2m_{ic}\right\}$$

$$=\sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_im_{ic}\right)^2-\sum_{i=1}^{I} w_i^2m_i.$$

Étant donné que $\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}=\sum_{i=1}^{I} w_im_i$, nous avons

$$\mathrm{Var}_1(\bar{y}_w)$$

$$=\frac{\mathrm{Var}_1\left(\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}y_{cj}\right)}{\left(\sum_{i=1}^{I} m_iw_i\right)^2}$$

$$=\frac{\sigma^2\left\{\sum_{i=1}^{I} w_i^2m_i+\rho\sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_im_{ic}\right)^2-\rho\sum_{i=1}^{I} w_i^2m_i\right\}}{\left(\sum_{i=1}^{I} w_im_i\right)^2}$$

de sorte que

$$\mathrm{deff}=m\frac{\sum_{i=1}^{I} w_i^2m_i}{\left(\sum_{i=1}^{I} w_im_i\right)^2}[1+(b^*-1)\rho], \qquad (4)$$

où $b^*=\sum_{c=1}^{C}(\sum_{i=1}^{I} w_im_{ic})^2/\sum_{i=1}^{I} w_i^2m_i$.

À l'aide de l'inégalité de Cauchy-Schwarz, nous obtenons

$$\left(\sum_{i=1}^{I} w_im_{ic}\right)^2=b_c^2\left(\sum_{i=1}^{I} w_i\frac{m_{ic}}{b_c}\right)^2\leq b_c^2\sum_{i=1}^{I} w_i^2\frac{m_{ic}}{b_c}$$

$$=b_c\sum_{i=1}^{I} w_i^2m_{ic}$$

de sorte que

$$b^*\leq\frac{\sum_{c=1}^{C} b_c\sum_{i=1}^{I} w_i^2m_{ic}}{\sum_{c=1}^{C}\sum_{i=1}^{I} w_i^2m_{ic}}=\bar{b}_w, \text{ par exemple.} \qquad (5)$$

Ainsi (4) et (5) donnent

$$\mathrm{deff}\leq m\frac{\sum_{i=1}^{I} w_i^2m_i}{\left(\sum_{i=1}^{I} w_im_i\right)^2}[1+(\bar{b}_w-1)\rho]. \qquad (6)$$

À noter que $\bar{b}_w$ peut être interprété comme une taille moyenne (pondérée) des grappes. Si $\bar{b}_w$ est égal à $\bar{b}$, par exemple si tous les b_c sont égaux, la limite supérieure de deff est simplement la formule de Kish. Ainsi, la formule de Kish sert de valeur prudente pour l'effet de plan de sondage.

REMERCIEMENTS

Les auteurs aimeraient remercier l'éditeur et les arbitres pour leurs commentaires qui ont contribué à améliorer ce document. Ce travail a été terminé pendant que le dernier auteur était professeur invité à ZUMA, le centre de méthodologie et de recherche sur les enquêtes de Mannheim (Allemagne).

BIBLIOGRAPHIE

KISH, L. (1987). Weighting in Deft^2. *The Survey Statistician*, Juin 1987.

SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (Éds.) (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board

Contents
Volume 14, Number 4, 1998

Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data
Steven E. Fienberg and Leon C.R.J. Willenborg 337

A Database System Prototype for Remote Access to Information Based on Confidential Data
Sallie Keller-McNulty and Elizabeth A. Unger 347

Estimating the Re-identification Risk Per Record in Microdata
C.J. Skinner and D.J. Holmes 361

A Bayesian Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment
Stephen M. Samuels 373

Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data
Stephen E. Fienberg and Udi E. Makov 385

Synthetic and Combined Estimators in Statistical Disclosure Control
Jeroen Pannekoek and Ton de Waal 399

Balancing Disclosure Risk Against the Loss of Nonpublication
Alan M. Zaslavsky and Nicolas J. Horton 411

Optimal Local Suppression in Microdata
A.G.de Waal and L.C.R.J. Willenborg 421

Models and Methods for the Microdata Protection Problem
C.A.J. Hurkens and S.R. Tiourine 437

Masking Microdata Using Micro - Aggregation
D. Defays and M.N. Anwar 449

Post Randomisation for Statistical Disclosure Control: Theory and Implementation
J.M. Gouweleeuw, P. Kooiman, L.C.R.J. Willenborg and P.-P. de Wolf 463

Comment
G. Sande 479

Disclosure Limitation Using Perturbation and Related Methods for Categorical Data
Stephen E. Fienberg, Udi E. Makov, and Russell J. Steel 485

Comment
Peter Kooiman 503

Rejoinder
Stephen E. Fienberg, Udi E. Makov, and Russell J. Steel 509

Comparision of Systems Implementing Automated Cell Suppression for Economic Statistics
N: Kirkendall and G. Sande 513

Using Noise for Disclosure Limitation Establishment Tabular Data
Timothy Evans, Laura Zayatz, and John Slanta 537

Experiments with Controlled Rounding for Statistical Disclosure Control in Tabular Data with Linear Constraints
Matteo Fischetti and Juan-José Salazar-González 553

Editorial Collaborators 567

Index to Volume 14, 1998 571

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

CONTENTS TABLE DES MATIÈRES

Volume 27, No. 1, March/mars 1999

Christian GENEST
Editor's report /Rapport du rédacteur en Chef

Mousumi BANERJEE, Michelle CAPOZZOLI, Laura MCSWEENEY and Debajyoti SINHA
Beyond kappa: A review of interrater agreement measures

L.MANCHESTER, C. A. FIELD and A. MCDOUGALL
Regression for overdetermined systems: A fisheries example

E.MONGA and S. TARDIF
Asymptotic optimality of a class of rank tests for replicated Latin square designs

Guohua PAN and Winson TAAM
Distribution-free subset selection for incompletely ranked data

Gary SNEDDON
Smoothing in an underdetermined linear model with random explanatory variables

John R.COLLINS
Robust *M*-estimators of scale: Minimax bias versus maximal variance

Dongchu SUN and Keying YE
Reference priors for a product of normal means when variances are unknown

Sonia PETRONE
Bayesian density estimation using Bernstein polynomials

M. PENSKY and R. S. SINGH
Empirical Bayes estimation of reliability characteristics for an exponential family

M. C. FINOCCHIARO and D. SACCHETTI
The variance function of the natural exponential family generated by a measure on n integers

John E. KOLASSA
Confidence intervals for parameters lying in a random polygon

Nabendu PAL and Wooi K. LIM
Second order properties of intraclass correlation estimators for a symmetric normal distribution

Andrew HEARD and Tim SWARTZ
Extended voting measures

Michael BARON
Convergence rates of change-point estimators and tail probabilities of the first-passage-time process

Célestin C. KOKONENDJI
Le problème d'Anscombe pour les lois binomiales négatives généralisées

Helge BLAKER
A class of shrinkage estimators in linear regression

Acknowledgement of referees' services / Remerciements aux membres des jurys

Forthcoming papers/Articles à paraître

CONTENTS TABLE DES MATIÈRES

Volume 27, No. 2, June/juin 1999

Andrey FEUERVERGER, John ROBINSON and Augustine WONG I
On the relative accuracy of certain bootstrap procedures

Biao ZHANG
Bootstrapping with auxiliary information

Steven N. MACEACHERN, Merlise CLYDE and Jun S. LIU
Sequential importance sampling for nonparametric Bayes models: The next generation

George D. PAPANDONATOS and Seymour GEISSER
Bayesian interim analysis of lifetime data

Paul DAMIEN and Stephen WALKER
A full Bayesian analysis of circular data using the von Mises distribution

Craig A. COOLEY and Steven N. MACEACHERN
Prior elicitation in the classification problem

Schultz CHAN and Malay GHOSH
A geometric optimality of Cox's partial likelihood

Y. LEE and John A. NELDER
The robustness of the quasilikelihood estimator

Hui CHEN and Joseph P. ROMANO
An invariance principle for triangular arrays of dependent variables with application to autocovariance Estimation

Kanchan MUKHERJEE
The asymptotic behavior of a class of L-estimators under long-range dependence

Ross H. TAPLIN
Robust F-tests for linear models

Ping ZHANG
The optimal prediction of cross-sectional proportions in categorical panel-data analysis

Dan NETTLETON
Order-restricted hypothesis testing in a variation of the normal mixture model

K. KRISHNAMOORTHY and Maruthy K. PANNALA
Confidence estimation of a normal mean vector with incomplete data

Marten H. WEGKAMP
Quasi-universal bandwidth selection for kernel density estimators

Christian GENEST
Probability and statistics: A tale of two worlds?

Forthcoming Papers / Articles à paraître

BLACKWELL Journals

Journal of the Royal Statistical Society
Series D (The Statistician)

Edited by N. R. J. Fieller and U. T. Moorthy

Covering a broad range of topics of interest to professional statisticians, *The Statistician* includes applied papers on education, business, sport, industry and agriculture, statistical computing and professional affairs and obituaries of eminent statisticians.

Recent and forthcoming highlights:

Some Statistical Heresies (with Discussion), *J. K. Lindsey*

Can Takeover Targets be Identified by Statistical Techniques? Some UK Evidence, *P. Barnes*

A Fractional Factorial Design for Bench-mark Testing of a Bayesian Method for Multilocation Audits, *V. Barnett and J. Haworth*

Using Maximum Entropy to Double One's Expected Winnings in the UK National Lottery, *S. J. Cox, G. J. Daniell and D. A. Nicole*

Demonstrating the Durbin-Watson Statistic, *R. Champion, C. T. Lenard and T. M. Mills*

Tiers, Structure Formulae and the Analysis of Complicated Experiments, *C. J. Brien and R. W. Payne*

Journal of the Royal Statistical Society: Series D (The Statistician)
ISSN 0039-0526 Published in March, June, September and December

Subscription Rates, Vol. 48/1999:
Europe £80, N. America $149, Rest of World £90.

To subscribe to *The Statistician* please use the order form on the Blackwell website: http://www.blackwellpublishers.co.uk, send an email to jnlinfo@blackwellpublishers.co.uk, or contact either of the following:

Blackwell Publishers Journals, PO Box 805, 108 Cowley Road, Oxford OX4 1FH, UK. Tel: +44 (0)1865 244083, fax +44 (0)1865 381381

- **Journals Marketing (RSSD), Blackwell Publishers, 350 Main Street, Malden, MA 02148, USA. Tel. +1 (781) 388 8200, fax +1 (781) 388 8210**

For further information or to request a sample copy please visit our website

http://www.blackwellpublishers.co.uk

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ x 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.