Nº 12-001-XIF au catalogue
ISSN 1712-5685

Nº 12-001-XPB au catalogue
ISSN 0714-0045

Techniques d'enquête

Juin 2013

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Comment accéder à ce produit

Le produit nº 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Ce produit est aussi disponible en version imprimée standard au prix de 30 $CAN l'exemplaire et de 58 $CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	10 $CAN	20 $CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste Statistique Canada
Finances
Immeuble R.-H.-Coats, 6e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2013

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada (http://www.statcan.gc.ca/reference/licence-fra.html).

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans *The ISI Web of knowledge* (*Web of science*), *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*. La revue est également citée par *SCOPUS* sur les bases de données *Elsevier Bibliographic Databases*.

COMITÉ DE DIRECTION

Président J. Kovar

Anciens présidents D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Membres G. Beaudoin
S. Fortier (Gestionnaire de la production)
J. Gambino
M.A. Hidiroglou
H. Mantel

COMITÉ DE RÉDACTION

Rédacteur en chef M.A. Hidiroglou, *Statistique Canada*

Ancien rédacteur en chef J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Rédacteurs associés

J.-F. Beaumont, *Statistique Canada*
J. van den Brakel, *Statistics Netherlands*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
R. Chambers, *Centre for Statistical and Survey Methodology*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
D. Haziza, *Université de Montréal*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
D. Judkins, *Abt Associates*
D. Kasprzyk, *National Opinion Research Center*
J.K. Kim, *Iowa State University*
P.S. Kott, *RTI International*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistique Canada*
P. Lynn, *University of Essex*
D.J. Malec, *National Center for Health Statistics*
J. Opsomer, *Colorado State University*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *Office for National Statistics*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *National Opinion Research Center*
C. Wu, *University of Waterloo*
W. Yung, *Statistique Canada*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints C. Bocci, K. Bosa, G. Dubreuil, C. Leon, H. Mantel, S. Matthews, Z. Patak, S. Rubin-Bleuer et Y. You, *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 en version électronique deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (rte@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca/Techniquesdenquete).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 39, numéro 1, juin 2013

Table des matières

Articles réguliers

Jeremy Strief et Glen Meeden
Pondérations par une méthode bayésienne séquentielle objective dans l'échantillonnage 1

Barry Schouten, Melania Calinescu et Annemieke Luiten
Optimiser la qualité de la réponse au moyen de plans de collecte adaptatifs..... 33

Sander Scholtus
Vérification automatique en présence de vérifications avec rejet et de vérifications avec avertissement..... 67

Jae Kwang Kim et Changbao Wu
Estimation parcimonieuse et efficace de la variance par rééchantillonnage pour les enquêtes complexes.... 105

Anne Massiani
Estimation de la variance d'indicateurs transversaux pour l'enquête SILC en Suisse 139

Iván A. Carrillo et Alan F. Karr
Combiner des cohortes dans les enquêtes longitudinales 169

Pierre Lavallée et Sébastien Labelle-Blanchet
Le sondage indirect appliqué aux populations asymétriques..... 207

Communications brèves

Yong You, J.N.K. Rao et Mike Hidiroglou
De la performance des estimateurs sur petits domaines
autocalés sous le modèle au niveau du domaine de Fay-Herriot..... 243

Peter M. Aronow et Cyrus Samii
Estimation prudente de la variance pour des plans
d'échantillonnage avec probabilités d'inclusion par paire nulles..... 257

Autres revues 269

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 1-32
Statistique Canada, N° 12-001-X au Catalogue

Pondérations par une méthode bayésienne séquentielle objective dans l'échantillonnage

Jeremy Strief et Glen Meeden[1]

Résumé

Bien que l'utilisation de pondérations soit très répandue dans l'échantillonnage, leur justification ultime dans la perspective du plan de sondage pose souvent problème. Ici, nous argumentons en faveur d'une justification bayésienne séquentielle des pondérations qui ne dépend pas explicitement du plan de sondage. Cette approche s'appuie sur le type classique d'information présent dans les variables auxiliaires, mais ne suppose pas qu'un modèle relie les variables auxiliaires aux caractéristiques d'intérêt. La pondération résultante d'une unité de l'échantillon peut être interprétée de la manière habituelle comme étant le nombre d'unités de la population que cette unité représente.

Mots clés : Échantillonage; pondérations; inférence bayésienne.

1 Introduction

Les pondérations jouent un rôle important dans l'approche de l'échantillonnage fondée sur un plan. En théorie, la pondération attribuée à une unité observée dans un échantillon est la réciproque de la probabilité de sélection de cette unité et elle est interprétée comme étant le nombre d'unités de la population que représente l'unité en question. En pratique, après avoir observé un échantillon, on ajuste souvent ses pondérations afin de le rendre plus représentatif de la population. Ces ajustements peuvent être faits pour tenir compte de l'information sur la population non incluse dans le plan de sondage, ainsi que des observations manquantes sur l'échantillon. Bien que ces modifications des pondérations fondées sur le plan de sondage soient sans aucun doute utiles dans certains cas, leur justification théorique n'est, en dernière analyse, pas claire. Selon nous, la confusion nait en partie d'une argumentation sur une base non conditionnelle avant le tirage de l'échantillon; ainsi, l'estimateur de Horvitz-Thompson est sans biais si l'on prend la moyenne sur tous les échantillons possibles, et l'est donc conditionnellement après que l'échantillon a été tiré, en ajustant les pondérations fondées sur le

1. Jeremy Strief, Statisticien principal, Medtronic Energy and Component Center, Brooklyn Center, MN 55430. Courriel : jstrief@gmail.com; Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Courriel : glen@stat.umn.edu.

plan de sondage des unités observées dans l'échantillon. En particulier, accorder une trop grande importance au plan de sondage à la deuxième étape, ou étape conditionnelle, peut compliquer inutilement les choses. Une fois que l'échantillon a été observé, nous pensons qu'une meilleure approche consiste à ignorer formellement le plan d'échantillonnage, mais à utiliser toute l'information disponible, y compris celle intégrée dans le plan, pour trouver un ensemble de pondérations raisonnable. Selon ce mode de pensée, une pondération attribuée à une unité peut encore être interprétée comme le nombre d'unités de la population que l'unité représente, mais elle n'est plus dérivée par ajustement de la probabilité de sélection de cette unité. Comment cela peut-il se faire ?

Dans l'approche bayésienne, l'information au sujet de la population est incorporée dans une loi a priori. En théorie, la loi a priori peut ensuite être utilisée pour sélectionner délibérément un échantillon optimal; cependant, cela ne se fait pratiquement jamais. Une fois que l'échantillon est observé, les inférences sont fondées sur la loi a posteriori des unités non observées dans la population, sachant les valeurs des unités observées dans l'échantillon. Dans la plupart des situations, la loi a posteriori ne dépend pas de la façon dont l'échantillon a été sélectionné, de sorte que le plan de sondage ne joue aucun rôle à l'étape de l'inférence. Les méthodes bayésiennes ont été peu utilisées en pratique, parce qu'il est difficile de trouver des lois a priori qui reflètent les types courants d'information a priori disponible.

Une interprétation bayésienne séquentielle (*stepwise Bayesian*) peut être donnée à de nombreux estimateurs classiques (Ghosh et Meeden 1997). Sous cette approche, étant donné un échantillon, l'inférence est encore fondée sur une loi a posteriori, mais l'ensemble (pour tous les échantillons possibles) des lois a posteriori ne provient pas d'une loi a priori unique, mais d'une famille complète de lois a priori. Dans la situation où l'on pense que les unités observées et les unités non observées sont approximativement interchangeables, la loi a posteriori sous l'approche bayésienne séquentielle est la loi a posteriori de Pólya.

Quand on possède l'information a priori sur les moyennes et sur les quantiles de population des variables auxiliaires, Lazar, Meeden et Nelson (2008) soutiennent que la loi a posteriori de Pólya contrainte, qui est une généralisation de la loi a posteriori de Pólya, est un moyen raisonnable d'intégrer ce genre d'information a priori. Ici, nous montrerons comment la loi a posteriori de Pólya contrainte peut être utilisée pour définir des pondérations pour les unités de l'échantillon. Bien que les pondérations résultantes dépendent des variables auxiliaires, elles ne s'appuient pas explicitement sur le plan de sondage.

À la section 2, nous passons en revue la loi a posteriori de Pólya et, à la section 3, la loi a posteriori de Pólya contrainte. Les deux idées principales qui sous-tendent l'article sont énoncées aux deux sections suivantes. À la section 4, nous montrons comment la loi a posteriori de Pólya contrainte peut être utilisée pour attacher une pondération à chaque unité de l'échantillon de manière que ces pondérations ne dépendent pas directement du plan de sondage. À la section 5, nous introduisons la loi a posteriori de Dirichlet pondérée pour accompagner la loi a posteriori de Pólya contrainte. Elle permet d'utiliser les pondérations définies par la loi a posteriori de Pólya contrainte pour faire des inférences au sujet des paramètres de population au moyen d'une simulation simple. À la section 6, nous comparons les pondérations basées sur la loi a posteriori de Pólya contrainte à celles utilisées dans l'estimateur de Horvitz-Thompson. À la section 7, nous considérons plusieurs exemples afin de voir comment les pondérations résultantes se comportent en pratique et de montrer comment la loi a posteriori de Dirichlet pondérée peut être utilisée pour obtenir une estimation de la variance d'un estimateur sans effectuer d'importants calculs. À la section 8, nous présentons certaines conclusions.

À la première lecture, certains auront l'impression que les méthodes proposées ici sont très bayésiennes, parce que toutes nos inférences sont fondées sur les lois « a posteriori ». Mais comme il est mentionné plus haut, techniquement, nos lois « a posteriori » ne sont pas bayésiennes, mais bayésiennes séquentielles. Cela signifie que, du point de vue de la mise en œuvre, on peut concevoir nos lois a posteriori comme étant construites après que l'échantillon a

été observé. Ces « lois a posteriori » construites ne dépendent pas de l'information a priori subjective ni du plan de sondage, mais s'appuient uniquement sur les valeurs d'échantillons observées et sur l'information objective et publique disponible au sujet des variables auxiliaires. Comme nous le verrons, cela permet de construire des estimateurs des paramètres de population qui sont approximativement sans biais sous divers plans de sondage et qui possèdent de bonnes propriétés fréquentistes. Néanmoins, notre méthode présente des limites importantes. Premièrement, elle n'est applicable qu'à des plans de sondage à une étape et deuxièmement, elle ne permet pas de corriger le biais de sélection.

2 La loi a posteriori de Pólya

Soit s l'ensemble d'étiquettes d'un échantillon de taille n tiré d'une population de taille N. Pour simplifier, nous supposons que les membres de s sont $1, 2, \ldots, n$ et nous supposons aussi que la fraction n / N est très petite. Soit $y = (y_1, y_2, \ldots, y_N)$ la caractéristique d'intérêt et y_s les valeurs d'échantillon observées.

La loi a posteriori de Pólya repose sur l'échantillonnage selon le modèle de l'urne de Pólya dont le principe est le suivant : supposons que les valeurs de n unités observées ou vues sont marquées sur n balles et sont placées dans l'urne 1. Les $N - n$ unités restantes non vues de la population sont représentées par $N - n$ balles non marquées placées dans l'urne 2. De chaque urne on tire une balle avec la même probabilité et on attribue à la balle provenant de l'urne 2 la valeur de la balle provenant de l'urne 1. On retourne ensuite les deux balles dans l'urne 1. Donc, à la deuxième étape de l'échantillonnage de Pólya, l'urne 1 contient $n + 1$ balles et l'urne 2 contient $N - n - 1$ balles. On répète cette procédure jusqu'à ce que l'urne 2 soit vide, moment auquel les N balles contenues dans l'urne 1 constituent une copie simulée complète de la population. Toute quantité de population finie – moyenne, total, quantile, coefficient de régression – peut maintenant être calculée à partir de la copie complète. Nous pouvons simuler K de ces copies complètes et, dans chaque cas, calculer la valeur de la quantité de population

d'intérêt. L'estimation ponctuelle est donnée par la moyenne de ces valeurs simulées et un intervalle de crédibilité bayésien à 95 % approximatif est donné par les quantiles correspondant à 2,5 % et 97,5 % des valeurs.

On peut vérifier que, sous la loi a posteriori de Pólya, l'espérance a posteriori de la moyenne de population est simplement la moyenne d'échantillon et la variance a posteriori est simplement égale à $(n-1)/(n+1)$ fois la variance sous le plan habituelle de la moyenne d'échantillon sous échantillonnage aléatoire simple sans remise. La loi a posteriori de Pólya possède une justification fondée sur la théorie de la décision en raison de sa nature bayésienne séquentielle. En s'appuyant sur ce fait, on peut montrer que de nombreux estimateurs classiques sont admissibles. Des renseignements détaillés figurent dans Ghosh et Meeden (1997). La loi a posteriori de Pólya est le bootstrap bayésien de Rubin (1981) appliqué à l'échantillonnage en population finie. Lo (1988) discute aussi du bootstrap bayésien dans le cas de l'échantillonnage en population finie. Certains des premiers travaux dans ce domaine sont décrits dans Hartley et Rao (1968) et dans Binder (1982).

Pour l'unité échantillonnée i soit p_i la proportion d'unités présentes dans une copie simulée complète de la population qui possédent la valeur y_i. Ghosh et Meeden (1997) ont montré que, sous la loi a posteriori de Pólya, $E(p_i) = 1/n$. Si nous posons que

$$w_i = NE(p_i) = N/n,$$

alors w_i peut être interprété comme la pondération appliquée à l'unité i puisqu'elle est égale au nombre moyen d'unités de la population représentée par l'unité i, sous la loi a posteriori de Pólya. Rappelons que sous échantillonnage aléatoire simple sans remise, la fraction n/N est la probabilité d'inclusion de chaque unité. Donc, dans ce cas, il y a concordance entre la pondération fréquentiste habituelle, qui est la réciproque de la probabilité d'inclusion, et la pondération sous la loi a posteriori de Pólya définie plus haut.

Donc, quand l'information a priori est limitée, la loi a posteriori de Pólya donne des pondérations identiques aux pondérations fréquentistes dérivées du plan d'échantillonnage aléatoire simple sans remise. La justification du choix de la loi a posteriori de Pólya pour ces pondérations ne dépend pas explicitement du plan d'échantillonnage et conviendrait dans toute circonstance où l'échantillonneur considère que les unités observées et non observées de la population sont approximativement interchangeables.

Nous allons maintenant examiner la relation entre la loi a posteriori de Pólya et les méthodes bootstrap habituelles sous échantillonnage en population finie. Les deux approches sont fondées sur une hypothèse d'interchangeabilité. Gross (1980) a présenté l'idée fondamentale du bootstrap. Supposons que l'on procède à un échantillonnage aléatoire simple sans remise et que $N / n = m$ est un entier. Étant donné un échantillon, nous créons une bonne approximation de la population en combinant m répliques de cet échantillon. En tirant des échantillons aléatoires répétés de taille n de cette population créée, nous pouvons étudier le comportement d'un estimateur d'intérêt. Booth, Bulter et Hall (1994) ont étudié les propriétés asymptotiques de ce genre d'estimateurs. Hu, Zhang, Cohen et Salvucci (1997) donnent un exemple d'utilisation de l'échantillon pour construire une population artificielle, suivie du tirage d'échantillons répétés de cette population pour construire une estimation de la variance de leur estimateur et des intervalles de confiance.

Notons que cette situation diffère de celle de la loi a posteriori de Pólya où l'échantillon est considéré fixe et des versions complètes de la population sont générées de manière répétée.

3 La loi a posteriori de Pólya contrainte

Commençons par rappeler une approximation bien connue de la loi a posteriori de Pólya. Si la fraction n / N est petite, sous la loi a posteriori de Pólya, $p = (p_1, \ldots, p_n)$ est alors approximativement une loi de Dirichlet caractérisée par un vecteur de paramètres ne contenant que des valeurs un, c'est-à-dire qu'elle est uniforme sur le simplex de dimension $n - 1$, où

$\sum_{j=1}^{n} p_j = 1$. Il est habituellement plus efficace de produire des copies complètes de la population en utilisant cette approximation plutôt que le modèle de l'urne décrit à la section précédente. En outre, cette approximation sera utile quand nous considérerons la loi a posteriori de Pólya contrainte, une généralisation de la loi a posteriori de Pólya obtenue lorsque l'échantillonneur dispose d'information a priori au sujet des variables auxiliaires.

Dans de nombreux problèmes, en plus de la variable d'intérêt, y, l'échantillonneur possède des variables auxiliaires pour lesquelles de l'information a priori est disponible. Un cas très fréquent est celui où la moyenne de population d'une variable auxiliaire est connue. De manière plus générale, nous supposerons que l'information a priori au sujet de la population peut être exprimée par un ensemble de contraintes linéaires d'égalité et d'inégalité sur une série de variables auxiliaires.

Nous supposons qu'en plus de la caractéristique d'intérêt y, il existe un ensemble de variables auxiliaires $x^1, x^2, \ldots, x^m$. Pour l'unité i, soit

$$(y_i, x_i) = (y_i, x_i^1, x_i^2, \ldots, x_i^m)$$

le vecteur des valeurs de y et des variable auxiliaires. Nous supposons que ce vecteur des valeurs est observé pour toute unité présente dans l'échantillon. Nous supposons aussi que l'information a priori au sujet de la population peut être exprimée au moyen d'un ensemble de contraintes linéaires d'égalité et d'inégalité sur les valeurs de population des variables auxiliaires. Pour l'ensemble de valeurs possibles pour une variable auxiliaire donnée, les coefficients qui définissent une contrainte correspondront aux proportions d'unités dans la population qui prennent ces valeurs. Nous allons maintenant illustrer ceci de manière plus précise en expliquant comment nous transposons cette information a priori au sujet de la population aux valeurs observées dans l'échantillon. Étant donné un échantillon, cela nous permettra de construire des copies simulées de la population compatibles avec l'information a priori.

Étant donné un échantillon s, pour $i = 1, 2, \ldots, n$, soit (y_i, x_i) les valeurs observées que, pour simplifier, nous supposons être distinctes. Soit p_i la proportion d'unités auxquelles est attribuée la valeur (y_i, x_i) dans une copie complète simulée de la population. Toute contrainte linéaire sur la valeur de population d'une variable auxiliaire se traduit de manière évidente par une contrainte linéaire sur ses valeurs observées. Par exemple, si l'on sait que la moyenne de population de x^1 est inférieure ou égale à une certaine valeur, disons b_1, alors pour la population simulée, cette information se traduit en la contrainte

$$\sum_{i=1}^{n} p_i x_i^1 \leq b_1.$$

Si l'on sait que la médiane de population de x^2 est égale à b_2, la contrainte pour la population simulée devient

$$\sum_{i=1}^{n} p_i u_i = 0,5$$

où $u_i = 1$ si $x_i^2 \leq b_2$ et est égal à zéro autrement. Donc, étant donné un ensemble de contraintes de population fondées sur l'information a priori et un échantillon, nous pourrons représenter les contraintes correspondantes sur une valeur simulée de p par deux systèmes d'équations

$$A_{1,s} p = b_1 \tag{3.1}$$

$$A_{2,s} p \leq b_2 \tag{3.2}$$

où $A_{1,s}$ et $A_{2,s}$ sont les matrices de dimensions $m_1 \times n$ et $m_2 \times n$, respectivement, et b_1 et b_2 sont des vecteurs de dimension appropriée.

Soit P le sous-ensemble du simplex de dimension n défini par les équations (3.1) et (3.2). Nous supposons que l'échantillon est tel que P est non vide et donc qu'il s'agit d'un polytope de dimension non pleine. Dans ce cas, la version approximative appropriée de la loi a posteriori de Pólya doit être simplement la loi uniforme sur P. Nous donnons à cette distribution le nom de loi a posteriori de Pólya contrainte (LPPC). S'il était possible de générer des observations indépendantes à partir de la LPPC, on pourrait trouver approximativement l'espérance

a posteriori des paramètres de population d'intérêt et obtenir des intervalles de crédibilité bayésiens séquentiels à 95 % approximatifs. Malheureusement, nous ne savons pas comment le faire. À la place, on peut utiliser des méthodes de Monte Carlo par chaîne de Markov (MCMC) pour trouver approximativement ces estimations. On peut pour cela travailler en R (R Development Core Team 2005) en utilisant le module externe *polypost* qui est disponible dans le CRAN. Des renseignements plus détaillés sur la LPPC et les simulations à partir de cette loi figurent dans Lazar et coll. (2008).

4 Pondérations basées sur la loi a posteriori de Pólya contrainte

Une critique de la loi a posteriori de Pólya et de la loi a posteriori de Pólya contrainte pourrait être que toute copie complète simulée de la population ne contient que les valeurs des caractéristiques qui figurent dans l'échantillon. Cependant, c'est exactement cette propriété qui va nous permettre d'attribuer des pondérations aux membres de l'échantillon.

Nous supposons que nous avons un échantillon fixe pour lequel le sous-ensemble du simplex défini par les équations (3.1) et (3.2) est non vide. Pour $j = 1, \ldots, n$ soit

$$w_j = NE(p_j) = N\mu_j \tag{4.1}$$

où l'espérance est prise par rapport à la LPPC. Notons que la somme des éléments de $w = (w_1, \ldots, w_n)$ est égale à la taille de la population N et que w_j peut être considéré comme la pondération associée au j^{e} membre de l'échantillon. Ces pondérations ne dépendent que des valeurs observées des variables auxiliaires et des contraintes de population connues. Par conséquent, il s'agit d'une méthode bayésienne séquentielle d'attribution des pondérations aux unités de l'échantillon dans laquelle est intégrée l'information a priori présente dans les variables auxiliaires et qui ne dépend pas explicitement du plan de sondage.

Nous supposons ici que la taille de la population N est connue, ce qui n'est pas toujours vrai. Le cas échéant, on pourrait remplacer N par une estimation dans l'équation susmentionnée. Si

l'estimation est bonne, les inférences résultantes pour un total de population devraient être satisfaisantes. Dans le cas de l'estimation d'une moyenne de population, les résultats seraient nettement moins sensibles à la mesure dans laquelle l'estimation est proche de la taille réelle de la population.

De nombreuses données d'enquête utilisées par les chercheurs du domaine de la science sociale sont fournies avec des pondérations appliquées aux unités individuelles. Dans ces situations, les pondérations basées sur la LPPC pourraient être reliées aux unités de la même façon et l'utilisateur n'aurait pas besoin de faire appel à des méthodes MCMC pour calculer les pondérations. Nous allons nous servir des pondérations pour définir la loi a posteriori de Dirichlet pondérée qui peut être utilisée pour trouver les estimations ponctuelles et les estimations des intervalles pour les quantités de population d'intérêt moyennant des calculs relativement modestes. Dans la suite de l'article, nous illustrerons à l'aide d'exemples comment ces poids peuvent être utilisés pour générer des procédures d'inférence ayant de bonnes propriétés fréquentistes.

Mais avant de poursuivre, faisons une simple observation. Supposons que nous disposions de l'échantillon ainsi que d'un ensemble de pondérations. Si N est grand, nous pouvons construire une population dont la proportion d'unités de type (y_i, x_i) est w_i / N pour $i = 1, \ldots, n$. Étant donné l'échantillon et l'ensemble de pondérations, nous pouvons considérer cette population construite comme étant la meilleure approximation de la population inconnue. Alors

$$\bar{y}_{bw} = \sum_{i=1}^{n} \frac{w_i}{N} y_i \text{ et } \sigma_{bw}^2 = \sum_{i=1}^{n} \frac{w_i}{N} (y_i - \bar{y}_{bw})^2 \tag{4.2}$$

sont la moyenne et la variance de cette population construite.

5 La loi a posteriori de Dirichlet pondérée

Il arrive souvent que des pondérations soient attachées à des données dans les fichiers de données à grande diffusion. Ces pondérations sont alors utilisées par les chercheurs pour produire

des estimations ponctuelles et des estimations d'intervalles pour les paramètres de population. Nous allons voir que les pondérations fondées sur la méthode bayésienne séquentielle présentée ici peuvent souvent être utilisées dans les formules fréquentistes classiques pour estimer les paramètres d'intérêt tout comme les pondérations habituelles. Nous utiliserons nos pondérations pour définir la loi a posteriori de Dirichlet pondérée et montrer qu'elle offre un autre moyen de calculer les estimations ponctuelles et les estimations d'intervalles pour diverses quantités de population.

Soit les w_j un ensemble de pondérations défini par l'équation (4.1) avec $\mu_j = w_j / N$. Considérons la distribution de Dirichlet sur le simplex défini par le vecteur $n\mu = (n\mu_1, \ldots, n\mu_n)$ comme une loi a posteriori de rechange pour $p = (p_1, \ldots, p_n)$ en utilisant l'échantillon observé pour produire des copies simulées complètes de la population. Nous donnerons à cette loi a posteriori le nom de loi a posteriori de Dirichlet pondérée (LPDP). Soulignons que la LPDP est une version moins contrainte de la LPPC. Sous la LPPC, chaque copie complète de la population satisfait les contraintes; par contre, sous la LPDP, seule la moyenne des populations simulées satisfait les contraintes. Il est facile de voir que sous la LPDP

$$E\left(\sum_{i=1}^{n} p_i y_i\right) = \sum_{i=1}^{n} \mu_i y_i = \bar{y}_{bw} \tag{5.1}$$

et

$$\begin{aligned} V\left(\sum_{i=1}^{n} p_i y_i\right) &= \sum_{i=1}^{n} y_i^2 V(p_i) + \sum\sum_{i<j} y_i y_j \mathrm{Cov}(p_i, p_j) \\ &= \sum_{i=1}^{n} \frac{n\mu_i(n - n\mu_i) y_i^2}{n^2(n+1)} - 2\sum\sum_{i<j} \frac{n\mu_i n\mu_j y_i y_j}{n^2(n+1)} \\ &= \frac{1}{n+1}\left(\sum_{i=1}^{n} \mu_i(1-\mu_i) y_i^2 + 2\sum\sum_{i<j} \mu_i n\mu_j y_i y_j\right) \\ &= \frac{1}{n+1}\left(\sum_{i=1}^{n} \mu_i y_i^2 - \sum_{i=1}^{n}\sum_{i=1}^{n} \mu_i \mu_j y_i y_j\right) = \frac{1}{n+1}\sigma_{bw}^2 \end{aligned} \tag{5.2}$$

où $\bar{y}_{bw}$ et σ_{bw}^2 sont définies par l'équation (4.2).

Partant de cela, nous voyons que, lorsqu'on estime la moyenne de population, la simulation à partir de la LPDP équivaut à utiliser l'échantillon et ses pondérations pour construire la meilleure approximation possible de la population. En particulier, quand les pondérations sont toutes égales, la LPDP est simplement la loi a posteriori de Pólya.

Deux grandes raisons nous poussent à introduire la LPDP. Premièrement, à mesure que le nombre de contraintes utilisées augmente, les intervalles de crédibilité à 95 % approximatifs fondés sur la LPPC deviennent trop courts et contiennent la valeur réelle du paramètre dans moins de 95 % du temps. Il en est ainsi parce que, quand le nombre de contraintes est grand, la LPPC ne permet pas d'obtenir une variabilité suffisante dans les copies complètes simulées de la population qu'elle produit. Deuxièmement, il est nettement plus facile d'exécuter la simulation à partir de la LPDP qu'à partir de la LPPC. Maintenant, il serait possible d'effectuer la simulation à partir de la LPDP contrainte de manière que toutes les contraintes soient satisfaites, mais cela demanderait autant d'effort que d'effectuer la simulation à partir de la LPPC. En outre, nous pensons que cela donnerait des intervalles de crédibilité à 95 % approximatifs ayant de mauvaises propriétés de couverture fréquentistes parce qu'ils seraient trop courts.

Supposons maintenant que notre ensemble de pondérations est constitué des réciproques des probabilités d'inclusion provenant du plan de sondage. Soit $W = \sum_{i=1}^{n} w_i$. Pour la plupart des échantillons, cette valeur n'est pas égale à N, mais s'en approche souvent. De nouveau, nous pouvons construire notre meilleure approximation de la population en nous basant sur les pondérations. La moyenne et la variance de cette population seront données par

$$\bar{y}_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} y_i \text{ et } \sigma_{dw}^2 = \sum_{i=1}^{n} \frac{w_i}{W} (y_i - \bar{y}_{dw})^2. \tag{5.3}$$

Si nous utilisons $\bar{y}_{dw}$ comme estimation de la moyenne de population inconnue, une estimation sans biais de cette variance dépend des probabilités d'inclusion conjointes des unités dans l'échantillon. Comme il est souvent difficile d'obtenir ces probabilités, il a été recommandé en pratique (Särndal, Swensson et Wretman 1992) de supposer que l'échantillonnage a été fait avec

remise, même si ce n'est pas le cas. Alors, l'estimation approximative résultante de la variance de $\bar{y}_{dw}$ est

$$\begin{aligned}\hat{V}_d(\bar{y}_{dw}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(n\frac{w_i}{W}y_i - \bar{y}_{dw}\right)^2 \\ &= \frac{\sigma_{dw}^2 + \gamma_{dw}}{n-1}\end{aligned} \tag{5.4}$$

où la deuxième ligne découle de simples opérations algébriques et où

$$\gamma_{dw} = \sum_{i=1}^{n}\frac{w_i}{W}y_i^2\left(n\frac{w_i}{W} - 1\right). \tag{5.5}$$

Notons que, sous échantillonnage aléatoire simple avec ou sans remise et $N = nk$, $\gamma_{dw} = 0$. Dans ce cas, l'estimation de la variance donnée par l'équation (5.4) est essentiellement équivalente à celle donnée par l'équation (5.2).

Dans les situations où il est sensé d'utiliser l'estimateur de Horvitz-Thompson, les calculs ont montré que γ_{dw} a tendance à être négatif, ce qui laisse entendre que les intervalles fondés sur la LPDP ont alors tendance à être prudents. Cependant, les calculs montrent aussi que le terme γ_{dw} a tendance à être positif dans les situations où l'estimateur de Horvitz-Thompson n'est pas approprié. Nous verrons que, dans de tels cas, l'approximation habituelle peut donner de médiocres résultats et que les intervalles fondés sur la LPDP peuvent posséder de meilleures propriétés fréquentistes.

6 Pondérations et estimateurs de Horvitz-Thompson

La pondération attribuée à une unité de l'échantillon est habituellement définie comme étant l'inverse de la probabilité d'inclusion de cette unité. On est donc invité à se représenter la pondération de l'unité comme étant le nombre d'unités de la population qu'elle représente. L'estimateur résultant du total de population est l'estimateur de Horvitz-Thompson (HT) qui est sans biais. Comme nous l'avons déjà mentionné, l'estimation sans biais de sa variance dépend

des probabilités de sélection conjointes de toutes les paires d'unités figurant dans l'échantillon. Comme, en pratique, il peut être impossible de calculer ces probabilités, on utilise souvent l'approximation de l'équation (5.4).

L'estimateur HT donne de bons résultats quand y_i est approximativement proportionnel à sa probabilité de sélection. Pour comparer le comportement de cet estimateur à la méthode basée sur la LPDP, nous avons réalisé une petite expérience par simulation. Nous avons construit la variable x en tirant un échantillon aléatoire de 2 000 unités d'une loi de probabilité gamma de paramètre de forme égal à 5 et de paramètre d'échelle égal à 1, et en ajoutant 20 à chaque valeur. Pour générer y, nous avons postulé que la loi conditionnelle de y_i sachant x_i était une loi normale de moyenne $5x_i$ et d'écart-type 20. La corrélation de la population résultante était de 0,49. Nous avons désigné cette population par A. Nous avons créé une deuxième population, B, en utilisant le même vecteur de valeurs de x, mais en ajoutant 400 à chaque valeur de y_i. Dans notre plan d'échantillonnage, nous nous sommes servis de x pour effectuer un échantillonnage proportionnel à la taille, c'est-à-dire $\text{ppt}(x)$. Nous avons utilisé le module externe *sampling* de R, de sorte que les probabilités d'inclusion étaient exactes. Sous ce plan, nous nous attendons à ce que l'estimateur HT donne de bons résultats pour la population A mais qu'il ait de moins bonnes propriétés pour la population B. Nous avons également considéré un troisième estimateur, NHT, obtenu simplement en rééchelonnant les pondérations de l'estimateur HT de manière que leur somme soit égale à la taille N de la population. Nous avons produit 500 échantillons de taille 50. Les résultats sont présentés au tableau 6.1.

Tableau 6.1
Résultats pour les populations A et B basés sur 500 échantillons de taille 50. L'estimateur NHT correspond à l'estimateur HT normalisé de manière que la somme des pondérations soit égale à la taille de la population, $N = 2\,000$. La couverture nominale est égale à 0,95 pour chaque population.

Population	Méthode	Erreur absolue moyenne	Longueur moyenne	Fréquence de couverture
A	HT	4 628	21 898	0,940
B	HT	8 965	43 914	0,960
A et B	LPDP	4 706	24 381	0,960
A	NHT	5 051	21 897	0,896
B	NHT	5 051	43 919	0,998

Bien que cela ne soit pas montré dans le tableau, les estimateurs HT et LPDP sont tous deux sans biais pour les deux populations. Comme prévu, l'estimateur HT est le meilleur pour la population A, mais sa performance baisse considérablement pour la population B. Par ailleurs, les propriétés de l'estimateur LPDP sont exactement les mêmes pour les deux populations. En tant qu'estimateur ponctuel, l'estimateur NHT donne de nettement meilleurs résultats que l'estimateur HT pour la population B, et des résultats moins bons pour la population A. Dans l'ensemble, l'estimateur LPDP est clairement celui qui donne les meilleurs résultats. Comment ces différences s'expliquent-elles ?

Dans la population A, $y_i \propto x_i$ et les calculs montrent que γ_{dw} est presque toujours négatif et que sa valeur absolue est petite comparativement à σ_{dw}. En d'autres termes, quand l'estimateur HT est approprié, c'est la variance de la population construite en se basant sur les pondérations de cet estimateur qui est essentiellement utilisée pour obtenir l'estimation de sa variance.

La seule différence entre les populations A et B est qu'une constante a été ajoutée à la valeur de y de chaque unité. Alors, si les pondérations de l'échantillon nous permettent d'obtenir une bonne approximation de la population dans le premier cas, quel est le problème qui, dans le deuxième cas, fait que l'estimateur HT donne de si mauvais résultats ? Pour le voir, considérons ce qui suit.

Dans l'estimation HT, la somme des pondérations dans l'échantillon n'est presque jamais égale à N, la taille de la population. Étant donné un échantillon de la population B, l'estimation HT est donnée par

$$\sum_{i=1}^{50} w_i y_i = \sum_{i=1}^{50} w_i y_i' + 400 \sum_{i=1}^{50} w_i$$

où y_i' désigne la valeur de l'unité dans la population A et y_i, sa valeur dans la population B. Notons que le deuxième terme de l'équation susmentionnée ajoute une variabilité supplémentaire à l'estimateur HT. Dans la population B, les calculs montrent que le terme γ_{dw} de

l'équation (5.5) est positif et peut être assez grand. Il explique la variabilité excédentaire de l'estimateur HT dans la population B découlant du fait qu'ici, $y_i \propto x_i + 400$ et non x_i.

Notons que Zheng et Little (2003) ont soutenu que, lorsqu'on estime un total de population finie et qu'on utilise un plan d'échantillonnage avec probabilités proportionnelles à la taille, un estimateur fondé sur un modèle non paramétrique à splines pénalisées donne généralement de meilleurs résultats que l'estimateur de Horvitz-Thompson. Zheng et Little (2005) ont élaboré des méthodes pour estimer la variance de leur estimateur. Certains travaux apparentés sont décrits dans Zheng et Little (2004).

Les pondérations basées sur la LPDP ne comprennent que la contrainte voulant que les copies complètes simulées de la population aient la moyenne de population correcte pour x. Cette hypothèse est plus robuste que celle qui sous-tend l'estimateur HT. En toute honnêteté, il faut se rappeler (comme l'a fait remarquer un examinateur) que l'estimateur HT a été élaboré avec l'objectif limité d'obtenir des estimateurs linéaires sans biais du total de population. Toutefois, sa simplicité ne semble plus aussi importante maintenant que des estimateurs plus compliqués et plus efficaces sont devenus plus faciles à calculer. La performance supérieure de la méthode bayésienne séquentielle laisse entendre ici que si l'on pense posséder pour les unités échantillonnées un ensemble de pondérations dont la somme est égale à la taille de la population et qui donne une bonne approximation de la population, il faut utiliser la variance de cette bonne approximation de la population pour construire une estimation de la variance de l'estimation de la moyenne de population au lieu d'utiliser l'équation (5.4). Cela vaut particulièrement pour les grandes enquêtes portant sur plusieurs caractéristiques y d'intérêt. Il serait très surprenant que toutes satisfassent les hypothèses nécessaires pour que l'équation (5.4) soit une bonne estimation de la variance d'une moyenne d'échantillon. Suivant l'observation faite dans Royall et Cumberland (1981) et Royall et Cumberland (1985) voulant que de bons échantillons équilibrés (la moyenne d'échantillon est proche de la moyenne de population) peuvent améliorer la

performance, on devrait fonder les inférences sur les copies complètes simulées de la population auxquelles est intégrée l'information a priori disponible contenue dans les variables auxiliaires.

7 Exemples

Nous pensons que la théorie fondée sur les plans de sondage classiques accorde trop d'importance au rôle que les probabilités de sélection doivent jouer dans les inférences faites après que l'échantillon a été observé. À la présente section, nous présentons des exemples qui montrent comment la LPDP permet d'utiliser de l'information a priori objective après que l'échantillon a été sélectionné.

7.1 Une étude de simulation

Afin de mieux comprendre comment l'utilisation des pondérations obtenues par la méthode bayésienne séquentielle dans la LPDP peut fonctionner, nous avons procédé à une étude en simulation. Nous avons construit une population de 2 000 unités et une variable auxiliaire unique, x. Cette variable consistait en un échantillon aléatoire tiré d'une loi de probabilité gamma dont le paramètre de forme était égal à 5 et le paramètre d'échelle, égal à 1. La loi conditionnelle de y_i sachant x_i était une loi normale de moyenne $100 + (x_i - 8)^2$ et d'écart-type égal à 20. La corrélation pour la population résultante était de -0,38. Nous désignons cette population par quad. Il s'agit clairement d'un petit exemple simple et la forme particulière de la relation entre x et y n'a pas d'importance en ce qui concerne les méthodes basées sur la LPDP, outre le fait que x contient une certaine information au sujet de y. Nous allons maintenant comparer les estimateurs basés sur la LPDP à deux méthodes classiques sous quatre plans de sondage différents.

Pour construire la LPPC, nous avons supposé que les valeurs de x pour la population étaient connues et nous les avons utilisées pour construire trois strates après avoir observé l'échantillon. Ces strates n'ont pas été construites de la manière habituelle. Nous avons procédé ainsi afin de minimiser le rôle habituel du plan de sondage et de mettre l'accent sur la robustesse de notre

approche au choix du plan de sondage. Nous avons tiré un échantillon de taille $n = 60$ et nous avons construit trois strates a posteriori. Soit $x_{[1]} < x_{[2]} < \cdots < x_{[60]}$ la statistique d'ordre des valeurs de x dans l'échantillon. Soit q_{20} et q_{40} les quantiles de population de $x_{[20]}$ et $x_{[40]}$ respectivement. Alors, la LPPC repose sur l'hypothèse que la probabilité totale attribuée aux unités de l'échantillon possédant les 20 plus petites valeurs de x doit être q_{20} et que la probabilité totale attribuée aux unités possédant les 20 plus petites valeurs suivantes doit être $q_{40} - q_{20}$. Autrement dit, nous divisons l'échantillon en trois groupes égaux et utilisons l'information contenue dans les valeurs de x pour obtenir la taille de la population appropriée des strates correspondantes. En outre, la LPPC repose sur l'hypothèse que les probabilités attribuées aux unités de l'échantillon doivent satisfaire la contrainte de la moyenne de population pour x.

La LPDP résultante sera comparée à deux méthodes fréquentistes classiques. La première est l'estimateur poststratifié qui utilise la même information sur les strates que la LPPC. La deuxième est l'estimateur par la régression habituel qui repose sur l'hypothèse que la moyenne de population de x est connue. Bien que l'estimateur par la régression ne soit pas réellement approprié pour la population quad, il a été inclus à titre de comparaison. Les intervalles de confiance à 95 % pour le total de population pour les deux méthodes fréquentistes ont été calculés en supposant un échantillonnage aléatoire simple même si des plans de sondage différents avaient été utilisés. Nous désignons ces deux estimateurs par STR et REG respectivement.

Le premier plan utilisé était un plan d'échantillonnage aléatoire simple sans remise. Pour le deuxième, nous avons produit un ensemble de poids d'échantillonnage en tirant un échantillon aléatoire de 2 000 unités d'une loi gamma dont le paramètre de forme était égal à 5 et le paramètre d'échelle, égal à 1. Puis nous avons ajouté 5 à chaque valeur pour obtenir le vecteur v, disons. Il faut noter que les valeurs de v et de y sont complètement indépendantes. Nous avons ensuite utilisé un plan $\text{ppt}(v)$ approximatif où, à chaque étape, la probabilité qu'une unité soit

sélectionnée est proportionnelle à sa valeur de v et dépend uniquement des unités non sélectionnées qui demeurent dans la population. Nous donnons à ce plan le nom de plan à pondération aléatoire. Pour le troisième plan, nous avons utilisé le plan $\text{ppt}(x)$ approximatif. Pour le quatrième, nous avons trouvé la fonction linéaire, disons l, qui applique le domaine de valeur de y sur l'intervalle $[1, 2]$. Nous avons ensuite utilisé le plan $\text{ppt}(l(y))$ approximatif comme plan de sondage. Nous donnons à ce plan le nom de plan dépendant de y. Dans ce plan, les probabilités de sélection dépendent faiblement des valeurs de y et les unités dont la valeur de y est grande sont plus susceptibles d'être sélectionnées que celles dont la valeur est faible. En particulier, l'unité possédant la valeur de y la plus grande est deux fois plus susceptible d'être sélectionnée que celle possédant la valeur de y la plus petite. De toute évidence, le plan à pondération aléatoire et le plan dépendant de y ne sont pas des plans classiques et ne seraient jamais utilisés en pratique. Nous les avons inclus afin de mettre en relief notre conviction que, dans de nombreux cas, étant donné un échantillon, une bonne estimation ne dépend pas de la façon dont l'échantillon a été sélectionné.

Pour chaque plan, nous avons tiré 500 échantillons de taille 60 et calculé l'estimation ponctuelle, son erreur absolue et la longueur de son intervalle estimé, et nous avons déterminé si celui-ci contenait ou non la valeur réelle du paramètre. Les résultats sont présentés au tableau 7.1.

Rappelons que, dans cet exemple, la LPDP utilise l'information provenant à la fois de la poststratification et du fait que la moyenne de population de x est connue, tandis que l'estimateur STR utilise seulement la première information et l'estimateur REG, seulement la seconde. Sous les plans EAS et à pondération aléatoire, les quatre méthodes donnent à peu près les mêmes résultats. Pour les deux autres plans, la méthode LPDP est celle qui est la meilleure. Sous les quatre plans de sondage, sa fréquence de couverture est celle qui est la plus proche du niveau nominal de 0,95. L'utilisation de la contrainte faisant intervenir la moyenne de population de x permet la correction d'une partie du biais introduit par les plans de sondage, ce que la méthode STR ne peut pas faire. Cependant, cette contrainte a ses limites. Si, sous le plan

dépendant de y, le domaine de valeur de l est $[1, 4]$, l'erreur absolue moyenne de l'estimateur LPDP est 4,5 % meilleure que celle de l'estimateur STR, et la fréquence de couverture sur les intervalles nominaux à 95 % est de 0,86 et 0,80, respectivement. La variable x ne contient tout simplement pas suffisamment d'information pour corriger un biais de sélection de cette importance.

Tableau 7.1
Résultats de la simulation pour la population quad décrite à la section 7.1 pour 500 échantillons aléatoires de taille 60 sous quatre plans de sondage différents. Le total de population réel était égal à 227 923,0. Pour chaque méthode, la couverture nominale était de 0,95.

Méthode	Valeur moyenne	Erreur moyenne	Longueur moyenne	Fréquence de couverture
		EAS		
STR	227 856,1	4 165,0	21 332,1	0,950
REG	227 602,1	4 302,7	21 300,3	0,944
LPDP	227 546,9	4 190,6	23 029,7	0,958
	Les min. et max. moyens des paramètres de la LPDP étaient 0,658 et 1,580.			
		Pondération aléatoire		
STR	227 976,5	4 371,2	21 254,1	0,938
REG	227 715,5	4 462,2	21 305,9	0,934
LPDP	227 721,2	4 420,6	22 901,4	0,950
	Les min. et max. moyens des paramètres de la LPDP étaient 0,651 et 1,583.			
		ppt(x)		
STR	225 295,8	5 228,9	23 008,4	0,916
REG	224 207,2	5 611,2	21 780,3	0,878
LPDP	227 471,1	4 919,2	22 706,6	0,936
	Les min. et max. moyens des paramètres de la LPDP étaient 0,374 et 3,024.			
		Dépendant de y		
STR	231 590,0	5 229,0	21 170,8	0,892
REG	231 424,4	5 143,4	21 127,9	0,902
LPDP	231 139,1	4 967,6	22 867,0	0,938
	Les min. et max. moyens des paramètres de la LPDP étaient 0,660 et 1,643.			

Pour chaque plan de sondage, nous avons inclus la moyenne des valeurs les plus faibles, d'une part, et les plus grandes, d'autre part, des paramètres définissant la LPDP, dont, ici, la somme doit être égale à 60. Nous voyons que c'est pour le plan ppt(x) que l'intervalle est le plus grand.

Nous avons également utilisé la LPDP dans les simulations pour construire des intervalles de crédibilité à 95 % pour la médiane de population de y. Pour les quatre plans de sondage, les fréquences de couverture respectives étaient de 0,956, 0,950, 0,952 et 0,930.

Nous avons effectué une autre étude en simulation où la variable x a été produite de la même façon, mais la loi conditionnelle de y_i sachant x_i était une loi normale de moyenne $60 + x_i$ et d'écart-type $2\sqrt{x_i}$. La corrélation entre x et y était de 0,46. Sous les quatre plans de sondage, les propriétés des estimateurs ponctuels étaient très similaires. Les intervalles avaient tendance à être un peu plus longs pour l'estimateur LPDP que pour les autres, mais, sur les quatre plans de sondage, sa fréquence moyenne de couverture pour le total de population était de 0,949. Sous le plan dépendant de y, sa fréquence de couverture pour le total de population était de 0,934, tandis que pour les estimateurs STR et REG, les couvertures correspondantes étaient de 0,896 et 0,886. Sa fréquence de couverture moyenne pour la médiane de population de y était de 0,942.

Un fréquentiste pourrait soutenir qu'il s'agit d'un exemple injuste, puisque l'estimateur par la régression n'a pas beaucoup de sens pour la population en question et, naturellement, il aurait raison. Si, pour ce problème, on supposait qu'il existe une relation quadratique entre y et x et que les deux premiers moments de population de x étaient connus, alors l'estimateur par la régression résultant donnerait de meilleurs résultats que l'estimateur LPDP. Lazar et coll. (2008) décrivent un exemple de ce genre. En outre, ils montrent qu'inclure une contrainte pour le deuxième moment de la LPPC ne modifie pour ainsi dire pas le comportement des estimations résultantes. Donc, lorsque l'on possède une bonne information a priori au sujet du modèle reliant x et y, elle pourrait être utilisée dans l'analyse. Si ce genre d'information a priori n'est pas disponible, nous pensons que l'estimateur LPDP possède un certain avantage, même s'il ne donne pas nécessairement lieu à une amélioration spectaculaire par rapport aux méthodes classiques. Il s'appuie uniquement sur l'information a priori objective et ne comprend aucune hypothèse de modélisation au sujet des liens entre les caractéristiques d'intérêt et les variables. Il peut faire une correction pour une légère dépendance des probabilités de sélection à l'égard de la caractéristique d'intérêt. Même si le plan de sondage ne joue aucun rôle explicite dans le calcul de cet estimateur, l'information qui y est souvent intégrée peut être reformulée sous forme d'une contrainte et utilisée pour définir la LPPC. Étant donné un échantillon, les inférences fondées sur

la LPDP s'appuient sur de nombreuses copies complètes simulées de la population qui, en moyenne, sont compatibles avec l'information a priori. Par conséquent, il est facile d'estimer d'autres paramètres qu'une moyenne ou un total de population.

7.2 Stratification et estimation de la médiane

Dans de nombreuses applications quelques observations seulement, parfois juste deux, sont tirées de chaque strate. Dans de telles conditions, trouver un bon intervalle de confiance lorsqu'on estime la médiane de population peut être difficile. Nous allons comparer la méthode classique, voir par exemple la section 5.11 de Särndal et coll. (1992), à celle fondée sur la LPDP. Nous émettons l'hypothèse d'un échantillonnage aléatoire simple sans remise dans les strates.

Par souci de définition précise, supposons que nous avons L strates et que la strate j contient N_j unités. Soit $N = \sum_{j=1}^{L} N_j$ la taille totale de la population. Supposons que deux observations sont tirées de chaque strate. Alors, la pondération attribuée à chaque unité échantillonnée est égale à la moitié de la taille de la strate dans laquelle elle a été sélectionnée. Dans la méthode classique, on utilise ces pondérations pour trouver l'intervalle de confiance.

Pour ce scénario, nous appliquons la loi a posteriori de Pólya habituelle dans chaque strate, de manière indépendante d'une strate à l'autre. Ou bien, on peut envisager d'appliquer une loi a posteriori de Pólya contrainte (LPPC) où la quantité de probabilité attribuée aux deux unités échantillonnées dans la strate j doit être égale à N_j / N. Si $p_j = (p_{j,1}, p_{j,2})$ représente la probabilité attribuée aux deux unités échantillonnées dans la strate j, alors sous la LPPC, $E(p_j) = (N_j / (2N), N_j / (2N))$. En reprenant la notation de la section 5, nous voyons que sous la LPDP, la pondération attribuée à chacune des deux unités échantillonnées dans la strate j est $(LN_j) / N$. Rappelons que simuler des copies complètes de la population en utilisant la LPDP signifie qu'il est presque certain que les copies simulées individuelles ne satisferont pas les contraintes, mais que ces dernières seront satisfaites si l'on prend la moyenne sur l'ensemble des copies simulées. À première vue, cela pourrait sembler être une mauvaise idée, mais nous verrons

que, lorsqu'on estime l'intervalle de la médiane de population, les estimations fondées sur la LPDP se comportent mieux que les intervalles classiques, qui sont trop courts. Nous verrons que la variabilité supplémentaire présente dans la LPDP produit des intervalles plus longs ayant de meilleures propriétés fréquentistes.

Les populations stratifiées que nous considérons ont été construites comme il suit. Les tailles de strate correspondaient à un échantillon aléatoire tiré d'une loi de Poisson de paramètre $\lambda = 100$. Les moyennes de strate correspondaient à un échantillon aléatoire tiré d'une loi normale de moyenne $\mu = 150$ et d'écarts-types $\sigma = 10$ ou $\sigma = 20$. Les écarts-types de strate correspondaient à un échantillon aléatoire tiré d'une loi gamma de paramètre d'échelle égal à un et de paramètre de forme $\gamma\sigma$ avec soit $\gamma = 0,10$ ou $\gamma = 0,25$. Nous avons construit deux versions de chacun des quatre types, l'une contenant 20 strates et l'autre, 40 strates. Pour chacune des huit populations, nous avons tiré 500 échantillons, chacun constitué de deux observations sélectionnées au hasard sans remise dans chaque strate. Pour chaque échantillon, nous avons comparé les estimations obtenues par l'approche classique à celles fondées sur la LPDP. Les résultats figurent au tableau 7.2. Nous ne présentons que ceux pour les populations à 20 strates, parce que les résultats pour les populations à 40 strates sont similaires. Les deux méthodes sont approximativement sans biais et l'estimation ponctuelle fondée sur la LPDP paraît être un tout petit peu meilleure, mais les intervalles de confiance produits par la LPDP sont clairement supérieurs. Même si, dans un cas, les intervalles selon la LPDP sont manifestement trop longs, globalement, ces intervalles sont nettement meilleurs que ceux obtenus par la méthode classique.

Quelles sont les causes des mauvaises propriétés des intervalles donnés par la LPDP dans un cas ? Des simulations supplémentaires indiquent que, lorsque les moyennes de strate varient fortement et que les variances de strate ont tendance à être relativement faibles, les intervalles donnés par la LPDP ont tendance à être trop longs. Dans nos simulations, le cas où $\sigma = 20$ et $\gamma = 0,10$ produit une population contenant ce genre de strate. Lorsque la taille d'échantillon a été augmentée pour passer à quatre unités par strate, la différence entre les deux méthodes n'était

plus aussi importante, mais les constatations restaient les mêmes. Les intervalles sous la méthode classique ont tendance à être trop courts et à avoir une couverture trop faible, tandis que les intervalles sous la LPDP sont plus longs et ont tendance à avoir une trop grande couverture.

Tableau 7.2
Résultats de simulation sur 500 échantillons aléatoires stratifiés de taille deux dans chaque strate des populations contenant 20 strates. La couverture nominale est de 0,95 pour chaque méthode.

Méthode	Valeur moyenne	Erreur moyenne	Longueur moyenne	Fréquence de couverture
		$\sigma = 10$ et $\gamma = 0,10$		
Classique	148,40	2,37	8,30	0,808
LPDP	148,39	2,22	12,20	0,95
		$\sigma = 10$ et $\gamma = 0,25$		
Classique	144,28	5,70	20,59	0,834
LPDP	144,18	5,41	28,38	0,950
		$\sigma = 20$ et $\gamma = 0,10$		
Classique	152,75	3,02	10,52	0,828
LPDP	152,61	2,78	22,88	0,996
		$\sigma = 20$ et $\gamma = 0,25$		
Classique	155,94	6,72	23,17	0,826
LPDP	155,89	6,35	34,96	0,962

Clairement, le choix d'une bonne méthode pour construire un intervalle de confiance dépend non seulement de la taille des intervalles qu'elle produit, mais aussi de la probabilité avec laquelle ces intervalles ne contiennent pas la valeur réelle, mais inconnue, du paramètre. Cohen et Strawderman (1973) et Meeden et Vardeman (1985), entre autres, ont étudié la question de l'admissibilité des procédures pour estimer les intervalles de confiance. Bien que les résultats présentés dans ces études ne s'appliquent pas directement ici, le deuxième article montre que, dans certaines situations, certaines procédures bayésiennes peuvent donner des procédures presque admissibles. Ce genre d'arguments et le fait que l'intervalle sous la méthode classique est beaucoup trop court constituent, à notre avis, des preuves circonstancielles que les intervalles sous la LPDP dans le présent exemple ne sont pas outrageusement trop longs. En résumé, nous pensons que, dans le cas particulier important où les tailles d'échantillon sont égales à deux et où

les strates ne diffèrent pas considérablement, les intervalles sous la LPDP semblent être des concurrents sérieux des intervalles sous la méthode classique.

7.3 Série de microdonnées à grande diffusion intégrées

Le Minnesota Population Center (MPC) est un groupe interdépartemental de recherche en démographie à l'Université du Minnesota. Un objectif important du MPC est de créer des bases de données et des outils statistiques qui peuvent être utilisés pour étudier le comportement économique et social. Une base de données d'intérêt est la Integrated Public Use Microdata Series (IPUMS), qui résulte de la consolidation des données des recensements des États-Unis et d'autres enquêtes nationales couvrant la période de 1850 à aujourd'hui (Ruggles, Sobek, Alexander, Fitch, Goeken, Hall, King et Ronnander 2004). Le terme *microdonnées* est utilisé dans ce contexte parce que chaque ligne de l'ensemble de données IPUMS correspond à une personne ou à un ménage; un contraste peut être fait entre ce fin niveau de détail et une publication ou un tableau sommaire en ligne typique du Census Bureau fournissant à l'utilisateur des données une totalisation des microdonnées à un niveau géographique particulier préétabli (qui peut correspondre à l'ensemble du pays, à des États, des comtés, des secteurs de recensement, *etc*.).

Un ensemble de données qui offre un riche tableau de variables numériques est celui de l'American Community Survey (ACS) de 2005. Ce produit du U.S. Census Bureau provient d'une grande enquête sur échantillon, et le Census Bureau ne connaît pas les moyennes de population réelle des variables. Pour procéder à des simulations avec les données de l'ACS de 2005, nous avons attribué à l'échantillon le rôle de la population. Plus précisément, nous avons supposé que la population complète était un ensemble de 3 579 résidents de Minneapolis qui étaient en âge de travailler (de 25 à 75 ans) et dont la rémunération annuelle était comprise entre 20 000 $ et 120 000 $. Pour les besoins de notre étude, les deux variables d'intérêt étaient :

- *inctot* : le revenu total avant impôt en 2004.

- sei : l'indice socioéconomique de Duncan. Créé durant les années 1950, cet indice est une variable numérique visant à évaluer le prestige associé à la profession d'une personne. Le domaine de valeur de cette variable est [1,100].

Pour nos simulations, nous avons pris $y = \log(inctot)$ et $x = sei$. La corrélation entre y et x est de 0,398 et nous supposons que la moyenne de x est connue. Pour estimer la moyenne de population de y nous avons considéré l'estimateur fondé sur la LPDP et l'estimateur par la régression. Nous avons utilisé deux plans de sondage différents : le plan d'échantillonnage aléatoire simple et le plan $\text{ppt}(x)$ approximatif. Dans chaque cas, nous avons tiré 300 échantillons de taille 30. Les résultats sont présentés au tableau 7.3. Nous voyons que, même si les deux méthodes sont comparables, la LPDP donne clairement de meilleurs intervalles.

Tableau 7.3
Résultats de simulation sur 300 échantillons aléatoires de taille 30 provenant de la population IPUMS. La couverture nominale est de 0,95 pour chaque méthode.

Plan	Méthode	Erreur moyenne	Longueur/2 moyenne	Fréquence de couverture
EAS	Régression	0,052	0,128	0,943
	LPDP	0,052	0,138	0,947
$\text{ppt}(x)$	Régression	0,062	0,132	0,897
	LPDP	0,066	0,133	0,937

8 Dernières remarques

Dans le domaine de l'échantillonnage, la construction de pondérations relève souvent davantage de l'art que de la science. Il s'agit de l'une des conclusions que l'on peut tirer de l'article récent de Gelman (2007) et de la discussion qui l'accompagne. Il argumente en faveur d'une approche bayésienne pour construire les pondérations en utilisant des modèles de régression qui relient la caractéristique d'intérêt aux variables auxiliaires. Ici, nous avons présenté des arguments en faveur d'une approche bayésienne séquentielle qui exploite l'information présente dans les variables auxiliaires sans émettre l'hypothèse d'un modèle reliant

la caractéristique d'intérêt à ces variables auxiliaires. La pondération résultante d'une unité de l'échantillon peut être interprétée de la manière habituelle comme étant le nombre d'unités de la population que l'unité en question représente.

Une pondération fréquentiste, disons w_i, est l'inverse de la probabilité d'inclusion, et ce nombre représente le nombre d'unités de la population représentées par une unité particulière dans l'échantillon. Donc, $w_i \geq 1$ pour tout i et $\sum_{i \in s} w_i \approx N$. À la section 6, nous avons vu que, pour l'estimateur de Horvitz-Thompson, la somme des pondérations des unités n'est habituellement pas égale à la taille de la population, ce qui peut donner lieu à un mauvais estimateur, sauf dans des circonstances très particulières. Un autre problème des pondérations fréquentistes est qu'elles sont souvent ajustées – après avoir observé l'échantillon – pour s'assurer que les estimations fréquentistes concordent avec l'information a priori au sujet de la population (Kostanich et Dippo 2002). Après les ajustements, les pondérations peuvent être rééchelonnées de manière que leur somme soit égale à un total de population. Cependant, les pondérations fréquentistes ajustées ne dépendent plus uniquement du plan de sondage et elles ne représentent plus les inverses des probabilités d'inclusion. Les notions intuitives qui sous-tendent les pondérations fréquentistes portent par conséquent quelque peu à confusion. Avant les ajustements, les pondérations fréquentistes sont des fonctions du plan; par contre, après les ajustements, elles sont des fonctions du plan et d'autres informations a priori qui peuvent ou non être reliées au plan.

Les bayésiens pensent que, dans le contexte de l'échantillonnage, l'estimation est un problème de prédiction. Leurs prédictions sont fondées sur un modèle hypothétique qui peut donner lieu à l'attribution de pondération aux unités de l'échantillon. Voir, par exemple, l'article susmentionné de Gelman (2007) et Little (2004). Comme l'ont fait remarquer un certain nombre d'auteurs (Pfeffermann 1993), effectuer une analyse pondérée pour un modèle utilisant les inverses des probabilités d'inclusion peut protéger l'échantillonneur contre l'erreur de spécification du

modèle. En outre, dans certaines situations, les deux approches peuvent produire des résultats similaires.

Récemment, Rao et Wu (2010) ont élaboré des méthodes faisant appel à une approche de pseudo-vraisemblance empirique et fondent leurs inférences sur des lois a posteriori de Dirichlet. Les procédures résultantes, même si elles sont sur le plan formel quelque peu similaires à celles dont il est question ici, s'appuient sur l'information a priori d'une manière différente. Pour ces auteurs, la majorité de l'information a priori doit être filtrée à travers le plan de sondage, alors que nous pensons que l'information a priori qui est souvent incluse dans le plan de sondage peut être utilisée directement pour produire de bonnes lois a posteriori. Pour le meilleur ou pour le pire, nous sommes plus proches du scénario bayésien classique où la loi a posteriori ne dépend pas du plan de sondage.

Ici, nous nous sommes concentrés sur l'utilisation de la loi a posteriori de Pólya contrainte (LPPC) pour produire un ensemble de pondérations fondé sur l'échantillon et sur l'information a priori, puis nous avons fait nos inférences en utilisant la loi a posteriori de Dirichlet pondérée (LPDP) fondée sur ces pondérations. Strief (2007) a considéré des exemples où les pondérations produites par la LPPC étaient utilisées dans les formules fréquentistes appropriées pour obtenir une estimation de la variance et a constaté que les résultats étaient comparables à ceux donnés par les méthodes classiques. On pourrait aussi imaginer de fonder leurs inférences sur la LPDP, mais en utilisant des poids fréquentistes, obtenus disons par des méthodes de calage (Särndal et Lundström 2005). Bien que cette option mérite d'être étudiée plus en profondeur, nous nous attendons à ce que ce genre d'approche donne lieu à des procédures inférencielles ayant de bonnes propriétés fréquentistes.

Dans l'approche fondée sur le plan de sondage, la convergence est une propriété importante que doit posséder un estimateur. Pour un cas particulier important, sous un plan EAS, les estimateurs selon la LPPC sont convergents. Cela est démontré dans Geyer et Meeden (2013).

Tout comme la LPPC, la LPDP possède une justification bayésienne séquentielle. Pour obtenir plus de détails, voir Strief (2007). Les pondérations utilisées dans la LPDP ont une formulation et une interprétation cohérentes. Il s'agit toujours d'une espérance a posteriori et leur somme est toujours égale à la taille de population. Elles représentent le nombre moyen de fois que chaque unité de l'échantillon apparaît dans une copie complète simulée de la population sous la LPPC. Cette moyenne est calculée par rapport à la loi uniforme sur toutes les copies possibles de la population qui contiennent juste les unités comprises dans l'échantillon et qui satisfont les contraintes données. Ces pondérations ne dépendent que du même type d'information a priori objective au sujet de la population que celle souvent utilisée pour définir et ajuster les pondérations fréquentistes. Il est donc possible d'y intégrer l'information a priori sans devoir spécifier explicitement une loi a priori.

Dans la plupart des cas, la pondération attribuée à une unité de l'échantillon dépend des autres unités de l'échantillon. Nous avons soutenu qu'après avoir sélectionné l'échantillon, on devrait raisonner conditionnellement. Autrement dit, étant donné l'échantillon, les pondérations devraient dépendre de toute l'information a priori disponible au sujet de la population, mais non de la façon dont l'échantillon a été sélectionné. (Nous supposons que la personne qui sélectionne l'échantillon et l'analyste ne font qu'un.) Toute procédure élaborée de cette manière devrait donner de bons résultats pour divers plans de sondage. Pour toute procédure, qu'elle soit fréquentiste, bayésienne ou bayésienne séquentielle, il s'agit du critère décisif : elle doit être évaluée d'après la façon dont elle se comporte sous échantillonnage répété conformément au plan de sondage d'intérêt.

Afin de mettre en œuvre les méthodes décrites ici, on doit d'abord utiliser la LPPC pour calculer les pondérations pour l'échantillon observé. Ensuite, on doit utiliser ces pondérations dans la LPDP pour simuler des copies complètes de la population. La première étape est la plus difficile, quoique le module externe *polyapost* la rend relativement simple pour toute personne familiarisée avec R. Une fois que les pondérations sont connues, il est facile d'exécuter la

simulation à partir de la LPDP au moyen de nombreux progiciels. Cela rend notre approche plus pratique pour les ensembles de données d'enquêtes (comme l'IPUMS) qui sont présentés avec les pondérations connexes et sont utilisés par de multiples chercheurs. Une limite plus sérieuse tient au fait que nous n'avons pris en considération que des plans d'échantillonnage simples à un seul degré. Les travaux doivent se poursuivre afin d'étendre ces méthodes à des plans à plusieurs degrés plus compliqués. Si les contraintes sous-jacentes sont sélectionnées judicieusement, les procédures résultantes peuvent avoir de bonnes propriétés fréquentistes pour divers plans de sondage. Les pondérations fondées sur la méthode bayésienne séquentielle peuvent être vues comme notre meilleure approximation de la population inconnue étant donné les unités échantillonnées et l'information a priori dont nous disposons.

Remerciements

L'étude a été financée en partie par la subvention NSF Grant DMS 0406169.

Bibliographie

Binder, D. (1982). Non-parametric Bayesian models for samples from a finite population. *Journal of the Royal Statistical Society*, Séries B, 44, 388-393.

Booth, J.G., Bulter, R.W. et Hall, P. (1994). Bootstrap methods for finite population sampling. *Journal of the American Statistical Association*, 89, 1282-1289.

Cohen, A., et Strawderman, W. (1973). Admissible confidence interval and point estimation for translation of scale parameters. *Annals of Statistics*, 1, 545-550.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (avec discussion). *Statistical Science*, 22, 153-188.

Geyer, C., et Meeden, G. (2013). Asymptotics for constrained Dirichlet distributions. *Bayesian Analysis*, 8, 89-110.

Ghosh, M., et Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman and Hall, Londres.

Gross, S. (1980). Median estimation in survey sampling. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 181-184.

Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 159-167.

Hu, M., Zhang, F., Cohen, M. et Salvucci, S. (1997). On the performance of replication-based variance estimation methods with small number of psus. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Kostanich, D.L., et Dippo, C.S. (2002). Design and methodology: 63rv. Rapport technique, The U.S. Census Bureau et The Department of Labor Statistics.

Lazar, R., Meeden, G. et Nelson, D. (2008). Une approche bayésienne non informative de l'échantillonnage d'une population finie en utilisant des variables auxiliaires. *Techniques d'enquête*, 34, 1, 55-70.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite poplation sampling. *Journal of the American Statistical Association*, 99, 546-556.

Lo, A. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.

Meeden, G., et Vardeman, S. (1985). Bayes and admissible set estimation. *Journal of the American Statistical Association*, 80, 465-471.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

Rao, J.N.K., et Wu, C. (2010). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society*, Séries B, 72, 533-544.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, www.R-project.org.

Royall, R., et Cumberland, W. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 71, 657-664.

Royall, R., et Cumberland, W. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C.A., Goeken, R., Hall, P.K., King, M. et Ronnander, C. (2004). Integrated public use microdata series: Version 3.0 [machine-readable database]. University of Minnesota.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer.

Strief, J. (2007). Bayesian Sampling Weights: Toward a Practical Implementation of the Polya Posterior. Thèse de doctorat, University of Minnesota.

Zheng, H., et Little, R. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zheng, H., et Little, R. (2004). Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés. *Techniques d'enquête*, 30, 2, 233-243.

Zheng, H., et Little, R. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 33-66
Statistique Canada, N° 12-001-X au Catalogue

Optimiser la qualité de la réponse au moyen de plans de collecte adaptatifs

Barry Schouten, Melania Calinescu et Annemieke Luiten[1]

Résumé

Dans la plupart des enquêtes, le même traitement est réservé à toutes les unités échantillonnées et les mêmes caractéristiques de plan de collecte des données s'appliquent à toutes les personnes et à tous les ménages sélectionnés. Le présent article explique comment les plans de collecte des données peuvent être adaptés en vue d'optimiser la qualité, étant donné des contraintes de coûts. Ces types de plans sont appelés plans de collecte adaptatifs. Les éléments fondamentaux de ce genre de plans sont présentés, discutés et illustrés au moyen de divers exemples.

Mots clés : Coûts d'enquête; erreurs d'enquête; non-réponse; plan de collecte des données adaptatif.

1 Introduction

Dans la plupart des enquêtes, le même traitement est réservé à toutes les unités échantillonnées et les mêmes caractéristiques de plan de collecte des données s'appliquent à toutes les personnes et à tous les ménages sélectionnés. Lorsque des données auxiliaires sont disponibles sous forme de données de registre ou d'observations faites par les intervieweurs, les plans de collecte peuvent être adaptés afin d'optimiser les taux de réponse, de réduire la sélectivité de la non-réponse ou, plus généralement, d'améliorer la qualité. Bien qu'une terminologie générale fasse défaut dans la littérature, ce genre de plans sont habituellement appelés plans de collecte adaptatifs.

Le présent article a pour objectif de décrire les éléments fondamentaux des plans de collecte adaptatifs, de systématiser ces plans en fournissant un cadre mathématique, d'illustrer la possibilité qu'ils ont d'améliorer l'efficacité de la collecte des données, et de propager leur utilisation dans la pratique des sondages.

1. Barry Schouten, Statistics Netherlands et Université Utrecht, PO Box 24500, 2490HA La Haye, Pays-Bas. Courriel : jg.schouten@cbs.nl; Melania Calinescu, VU University Amsterdam, Department of Mathematics, De Boelelaan 1081, 1081HV Amsterdam. Courriel : melania.calinescu@vu.nl; Annemieke Luiten, Statistics Netherlands, PO Box 4481, 6401CZ Heerlen, Pays-Bas. Courriel : a.luiten@cbs.nl.

Les plans de collecte adaptatifs reposent sur le principe que différentes personnes ou différents ménages peuvent recevoir un traitement différent. Ces traitements sont définis avant le début de l'enquête, mais peuvent aussi être mis à jour au moyen de renseignements obtenus durant la collecte des données. Autrement dit, l'attribution des traitements se fonde sur des données liées à l'échantillon de l'enquête et sur des paradonnées. Ces dernières sont des données sur le processus de collecte des données, comme les observations des intervieweurs au sujet du quartier, du logement ou des répondants, ou sur le rendement des intervieweurs. Dans le présent article, le terme paradonnées est utilisé dans son sens le plus large pour désigner des données qui sont observées durant la collecte et qui renseignent sur le comportement de réponse des personnes ou des ménages échantillonnés.

Une introduction générale aux plans de collecte adaptatifs est donnée dans Wagner (2008). Ces plans adaptatifs tirent leur origine de la littérature sur les statistiques médicales, où des traitements distincts sont attribués à l'avance à différents groupes de patients, mais dépendent aussi de la réponse des patients, autrement dit de mesures faites durant la collecte des données. Voir, par exemple, Heyd et Carlin (1999), Murphy (2003) et Zajonc (2012).

Un cas particulier de plan de collecte adaptatif est le plan de collecte « réactif » (*responsive survey design*). Les plans de collecte « réactifs » ont été introduits par Groves et Heeringa (2006). Comme les plans de collecte adaptatifs généraux, les plans de collecte « réactifs » peuvent comprendre l'application de caractéristiques de plan de collecte différentielles aux unités échantillonnées. La principale distinction tient au fait que, dans le cas des plans « réactifs », les caractéristiques de plan de collecte ou traitements prometteurs et efficaces sont identifiées durant la collecte des données. Pour cela, on subdivise la collecte en plusieurs phases de conception. Une nouvelle phase emploie les résultats de contrastes randomisés entre unités échantillonnées aux phases antérieures pour faire la distinction entre les traitements efficaces et inefficaces, et pour déterminer les coûts associés aux traitements. Les contrastes randomisés sont les différences de taux de réponse entre les sous-populations pour des caractéristiques de plan de collecte

attribuées aléatoirement. Voir, par exemple, Mohl et Laflamme (2007), Laflamme et Karaganis (2010), Phillips et Tabuchi (2009) et Peytchev, Riley, Rosen, Murphy et Lindblad (2010). L'affectation des caractéristiques de plan de collecte doit se faire de façon que chaque phase atteigne sa capacité de phase, qui est le compromis optimal entre la qualité et les coûts. Le recours à des plans de collecte adaptatifs est motivé par des conditions d'enquête où l'on en sait peu à l'avance sur l'échantillon et où les données historiques fournissent à l'avance peu de renseignements au sujet de l'efficacité des traitements. Dans ces conditions, plusieurs phases sont nécessaires et les plans de collecte « réactifs » sont pratiques. Toutefois, si l'on considère la deuxième phase et les phases de conception de niveau plus élevé des plans de collecte « réactifs », le point de départ est le même que pour les enquêtes où l'on possède une information préalable importante au sujet des unités échantillonnées ou pour une enquête qui est répétée plusieurs fois. La seule distinction est que, durant les phases antérieures du plan de collecte, une partie de l'échantillon a déjà répondu. Dans le présent article, nous supposons que des données historiques sont disponibles, que les traitements efficaces ont été déterminés à l'avance et qu'il est spécifié que des données appariées et des paradonnées seront utilisées pour adapter le plan de collecte.

Quelles sont les nouveautés dans le présent article ? Notre contribution est triple. Premièrement, nous établissons un cadre mathématique général pour optimiser la qualité de la réponse étant donné des contraintes de coûts. Deuxièmement, nous affectons explicitement différentes caractéristiques de plan de collecte à différentes unités échantillonnées à l'intérieur de ce cadre. Troisièmement, nous proposons d'optimiser les indicateurs de qualité pour l'erreur due à la non-réponse. Les deux derniers éléments ne sont, en soi, pas entièrement nouveaux. Des plans de collecte adaptatifs simples ont déjà été appliqués; par exemple, dans l'Enquête sur la population active des Pays-Bas, ni la collecte en ligne ni la collecte par téléphone n'est utilisée pour les grands ménages, et la réponse par personne interposée n'est permise que si elle est fournie par un membre du noyau du ménage. Les efforts en vue d'optimiser les plans d'enquête

afin de tenir compte de l'erreur due à la non-réponse remontent au moins à Hartley et Monroe (1979). En outre, il existe une littérature abondante sur l'optimisation du moment où sont faits les appels et du nombre de prises de contact dans les enquêtes avec intervieweur; par exemple, Kulka et Weeks (1988), Greenberg et Stokes (1990) et Kalsbeek, Botman, Massey et Liu (1994). L'aspect nouveau ici est le regroupement de tous ces éléments dans un cadre mathématique général qui fait abstraction des caractéristiques de plan de collecte uniques et qui permet d'appliquer des indicateurs de qualité généraux. Les principales raisons motivant la proposition de ce genre de cadre sont les pressions importantes en vue de réduire les coûts d'enquête et l'importance croissante d'Internet comme mode de collecte des données d'enquête. La collecte en ligne est caractérisée par un écart qualité-coûts important – elle est bon marché, mais les taux de réponse sont faibles – et elle ne possède pas les mêmes propriétés de mesure que les modes de collecte par intervieweur. Par conséquent, la collecte en ligne pose un défi en ce qui concerne le compromis entre la qualité et les coûts. Alors que les compromis entre les diverses erreurs étudiés au moment de la conception des enquêtes ont fait couler beaucoup d'encre, par exemple Lyberg, Biemer, Collins, De Leeuw, Dippo, Schwarz et Trewin (1997) et Dillman (2007), l'étude et la mise en œuvre de caractéristiques de plan de collecte différentielles en pratique se limitent encore, étonnamment, à quelques cas. Dans le présent article, nous espérons offrir un tremplin pour de futures études et discussions des plans de collecte adaptatifs.

À la section 2, nous décrivons la théorie et les concepts qui sous-tendent les plans de collecte adaptatifs. À la section 3, nous présentons un exemple fondé sur des données d'enquête virtuelles et, à la section 4, nous discutons d'une étude en simulation portant sur des données d'enquête réelles. Enfin, à la section 5, nous concluons par un résumé et une discussion.

2 Que sont les plans de collecte adaptatifs ?

2.1 Plans de collecte adaptifs en général

À la présente section, nous exposons le cadre mathématique pour les plans de collecte

adaptatifs. Aux sections suivantes, nous mettons en relief et élaborons les composantes de ce cadre.

Soit la population composée des unités $k = 1, 2, \ldots, N$. La population d'intérêt peut être constituée de toutes les unités d'une population, mais aussi de tous les membres recrutés d'un panel. À chaque unité sera attribuée une stratégie s provenant de l'ensemble de stratégies considérées $S = \{\varphi, s_1, s_2, \ldots, s_M\}$. L'*ensemble de stratégies d'enquête* S comprend explicitement la stratégie vide φ. La stratégie vide signifie qu'aucune action n'a lieu, c'est-à-dire que l'unité de population n'est pas échantillonnée. Il s'agit du cadre le plus général. En pratique, on sépare souvent le plan d'échantillonnage de la répartition des stratégies et l'on considère l'échantillon comme donné et fixe. Cependant, on peut inclure la décision d'échantillonner une unité explicitement dans l'affectation globale des ressources.

En général, une stratégie s est un ensemble spécifié de caractéristiques de plan de collecte qui peut comprendre une série de traitements, où un traitement donné n'est appliqué que quand tous les traitements antérieurs ont échoué. Certaines de ces caractéristiques peuvent être séquentielles, telles que le type de mode de contact et le type de mode de collecte des données, mais les caractéristiques peuvent aussi décrire divers aspects d'un plan de collecte des données. Voici des exemples de stratégies :

s_1 = (lettre envoyée à l'avance 1, questionnaire en ligne, un appel);

s_2 = (lettre envoyée à l'avance 1, questionnaire en ligne, pas de rappel);

s_3 = (lettre envoyée à l'avance 2, questionnaire administré par ITAO, maximum de 6 tentatives d'appel);

s_4 = (lettre envoyée à l'avance 2, questionnaire administré par ITAO, maximum de 15 tentatives d'appel).

On trouve dans la littérature de nombreuses suggestions et évaluations de caractéristiques de plans de collecte; par exemple, Groves et Couper (1998) et Groves, Dillman, Eltinge et Little (2002). Nous renvoyons le lecteur à De Leeuw (2008) pour une discussion des modes de collecte

des données, à Dillman (2007) et De Leeuw, Callegaro, Hox, Korendijk et Lensvelt-Mulders (2007) pour une élaboration des lettres envoyées à l'avance et des rappels, à Wagner (2008) pour une discussion du protocole de prise de contact, à Barón, Breunig, Cobb-Clark, Gørgens et Sartbayeva (2009) pour une revue des primes d'incitation, à Kersten et Bethlehem (1984), Cobben (2009) et Lynn (2003) pour la recherche sur les questionnaires condensés, à Moore (1988) pour une discussion de la déclaration par personne interposée et à Cobben (2009) pour un exemple de tâches d'intervieweur.

Nous supposons dans le présent article que l'ensemble de stratégies S est connu et fixé une fois que l'attribution des stratégies a débuté. L'ensemble de stratégies peut être déterminé en se basant sur des données d'enquête historiques, l'expérience ou des études pilotes. Voir Schouten, Luiten, Loosveldt, Beullens et Kleven (2010) et Schouten, Shlomo et Skinner (2011) pour des lignes directrices et des exemples sur la façon de construire un ensemble de stratégies.

À chaque unité de population k est associé un vecteur de covariables $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})^T$. X_k contient les caractéristiques connues avant le début de la collecte des données et avant l'attribution des stratégies. Les covariables doivent, par conséquent, être disponibles dans des registres ou des fichiers de données administratives qui peuvent être appariés à la base de sondage, ou directement dans la base de sondage. En plus de ces caractéristiques générales, il peut exister pour l'unité k un deuxième vecteur de covariables $\tilde{X}_k = (\tilde{X}_{k1}, \tilde{X}_{k2}, \ldots, \tilde{X}_{kq})^T$ qui reflète les caractéristiques observées durant la collecte des données pour les unités de population échantillonnées. Ces caractéristiques sont appelées paradonnées ou données sur le processus, parce qu'elles sont recueillies durant le processus de collecte des données par les intervieweurs et par le personnel de la collecte des données. Toutefois, en plus de la définition habituelle des paradonnées comme étant de l'information au sujet du processus, dans le contexte des plans de collecte adaptatifs, $\tilde{X}_k$ comprend des observations au sujet de la personne ou du ménage échantillonné. Le vecteur X_k peut contenir, par exemple, le sexe, l'âge, le type de ménage ou le niveau de scolarité. Le vecteur $\tilde{X}_k$ peut

contenir, par exemple, l'évaluation faite par l'intervieweur de la propension à répondre ou de la propension à être contacté, de l'état du logement ou du quartier, et de la présence d'un intercom. $\tilde{X}_k$ est délibérément limité aux observations au sujet de l'échantillon qui permettent de choisir les caractéristiques de plan de collecte différentes. Il ne contient pas les valeurs des caractéristiques du plan de collecte proprement dites, tel que l'intervieweur qui a été affecté à l'adresse.

La différence importante entre X_k et $\tilde{X}_k$ tient au niveau de disponibilité des données. $\tilde{X}_k$ est connu uniquement pour les unités qui sont échantillonnées et ne peut pas être utilisé pour distinguer des sous-populations a priori. Soit $q(x)$ la distribution de X_k dans la population et $q(\tilde{x}, x)$ la distribution conjointe de X_k et $\tilde{X}_k$ dans l'échantillon. En outre, $q(\tilde{x} \mid x)$ désigne la distribution conditionnelle dans l'échantillon. On suppose que $q(x)$ et $q(\tilde{x}, x)$ sont connus d'avance. Dans les situations où l'on ne peut apparier que peu de données, voire aucune, l'attribution des stratégies doit être fondée entièrement sur les observations faites durant la collecte des données.

Les plans de collecte adaptatifs qui attribuent les stratégies en se basant sur les caractéristiques de la population pour lesquelles des données sont disponibles dans un registre ou dans la base de sondage sont considérés comme étant *statiques*, tandis que les plans de collecte adaptatifs qui attribuent les stratégies en se basant (en outre) sur des paradonnées sont appelés *dynamiques*. Il est important de souligner que les plans tant statiques que dynamiques comprennent un ensemble de stratégies qui est fixé avant le début de la collecte des données. Toutefois, dans le cas des plans dynamiques, on ne sait pas à l'avance quelles stratégies seront attribuées aux unités individuelles, parce que le choix de la stratégie dépend des données qui sont observées durant la collecte.

Soit $\rho(x, s)$ la *propension à répondre* d'une unité présentant la caractéristique $X = x$ et à laquelle est attribuée la stratégie s. On suppose que l'information sur $\rho(x, s)$ est disponible dans les données historiques, c'est-à-dire des données qui proviennent de versions antérieures de la

même enquête, d'enquêtes possédant un thème et un plan similaires, ou de la phase initiale du plan de collecte. Manifestement, la propension à répondre attendue doit être une estimation qui s'approche de la propension réelle. Nous reviendrons sur cette composante essentielle des plans de collecte adaptatifs à la section 2.4.

Le *coût prévu* de l'affectation de la stratégie s à une unité ayant la caractéristique $X = x$ est désigné par $c(x, s)$. Il s'agit d'une composante de coût individuel. Les coûts d'enquête sont décrits dans la littérature comme comprenant de nombreuses composantes dont certaines sont des coûts indirects et d'autres, des coûts individuels; voir, par exemple, Groves (1989). Les fonctions de coût sont discutées à la section 2.3.

Soit $p(s \mid x)$ la *probabilité d'affectation* d'une unité de population ayant la caractéristique x à la stratégie s, et soit $p(s \mid x, \tilde{x})$ la probabilité d'affectation à cette stratégie sachant que des paradonnées $\tilde{x}$ sont également observées. Les conditions suivantes doivent être vérifiées :

$$0 \leq p(s \mid x) \leq 1, \qquad 0 \leq p(s \mid x, \tilde{x}) \leq 1 \tag{2.1}$$

$$\sum_s p(s \mid x) = 1, \qquad \sum_s p(s \mid x, \tilde{x}) = 1, \tag{2.2}$$

c'est-à-dire que toutes les unités sont affectées à une stratégie. En général, les probabilités d'affectation peuvent avoir des valeurs comprises entre 0 et 1. Autrement dit, les sous-population ayant les mêmes scores sur x et $\tilde{x}$ peuvent être affectées (aléatoirement) à des stratégies différentes. Par exemple, on pourrait n'approcher de nouveau au suivi qu'une partie seulement des non-répondants. Permettre que les probabilités d'affectation soient comprises entre 0 et 1 donne plus de souplesse pour atteindre les niveaux de qualité ou respecter les contraintes de coûts. Dans la suite de l'exposé, p désigne la matrice des probabilités d'affectation, c'est-à-dire $p = \{p(s_j \mid x, \tilde{x})\}_{1 \leq j \leq M, x, \tilde{x}}$ et contient les variables décisionnelles pour l'optimisation.

Les propensions à répondre ρ_X peuvent être calculées en se basant sur les propensions à répondre selon la stratégie et sur les probabilités d'affectation par

$$\rho_X(x) = \sum_{s \in S} \sum_{\tilde{x}} q(\tilde{x} \mid x) p(s \mid x, \tilde{x}) \rho(s, x, \tilde{x}). \tag{2.3}$$

Les stratégies, les covariables, les propensions à répondre, les fonctions de coûts et les probabilités d'affectation constituent les éléments fondamentaux des plans de collecte adaptatifs. Partant de ces éléments, on peut formuler le problème d'optimisation du plan de collecte adaptatif. Toutefois, deux éléments manquent encore, à savoir une fonction de qualité et une fonction de coût global. Soit $Q(p)$ un indicateur de qualité et $C(p)$ une évaluation du coût total. La dépendance des deux fonctions à l'égard des probabilités d'affectation est soulignée, car les probabilités sont les variables décisionnelles dans l'optimisation.

Le problème d'optimisation peut maintenant être formulé sous la forme

$$\max_p Q(p) \text{ sachant que } C(p) \leq C_{\max} \tag{2.4}$$

ou sous la forme

$$\min_p C(p) \text{ sachant que } Q(p) \geq Q_{\min}, \tag{2.5}$$

où $C_{\max}$ représente le budget d'une enquête et $Q_{\min}$, les contraintes de qualité minimales. Les problèmes (2.4) et (2.5) sont appelés problèmes d'optimisation double, quoique leurs solutions puissent différer selon les contraintes de qualité et de coût.

Il est important de souligner que l'optimisation de la qualité ou du coût n'est effectuée qu'une seule fois, avant que débute la collecte des données, et qu'elle n'est pas répétée durant la collecte. Donc, c'est la stratégie, et non pas l'optimisation proprement dite, que l'on adapte à l'unité de population et, dans le cas d'un plan dynamique, aux paradonnées au sujet de l'unité en question. L'optimisation est fondée sur des données d'enquête historiques qui comprennent les paradonnées devenues disponibles dans une enquête. La densité de probabilité conjointe $q(x, \tilde{x})$, les probabilités de réponse $\rho(x, \tilde{x}, s)$ et la fonction de coût $c(x, \tilde{x}, s)$ sont toutes estimées d'après les données d'enquête historiques et sont considérées comme étant données. Puisqu'en pratique, les paradonnées ne deviennent disponibles que durant la collecte des données, les stratégies considérées pour les unités comprises dans une même strate x sont les mêmes jusqu'au moment où les paradonnées $\tilde{x}$ deviennent disponibles. Par exemple, on pourrait avoir les quatre

stratégies suivantes : 1) deux tentatives d'appel téléphonique et aucun suivi, 2) deux tentatives d'appel téléphonique et un suivi avec prime d'incitation, 3) trois tentatives d'appel téléphonique et aucun suivi, et 4) trois tentatives d'appel téléphonique et un suivi avec prime d'incitation. La décision de faire deux ou trois tentatives d'appel est fondée sur x, tandis que le suivi est décidé en se servant d'une observation $\tilde{x}$ tirée des paradonnées téléphoniques. Donc, on estime à l'avance combien d'unités seront comprises dans la strate $(x, \tilde{x})$ et combien feront l'objet d'un suivi, mais ce n'est que quand $\tilde{x}$ est mesuré que la stratégie complète est connue pour les unités individuelles.

2.2 Fonctions d'objectif de qualité

Tels qu'ils sont décrits dans la littérature, les plans de collecte adaptatifs sont habituellement axés sur l'erreur due à la non-réponse. À la présente section, nous commençons par présenter une classification générale des fonctions de qualité, puis nous passons aux fonctions de qualité pour l'erreur due à la non-réponse. En général, la concentration sur l'erreur due à la non-réponse donne une perspective trop restreinte, surtout quand le mode de collecte des données est l'une des caractéristiques du plan de collecte prise en considération dans le plan de collecte adaptatif. Ici, nous ne discutons toutefois pas explicitement des autres erreurs d'enquête, mais nous revenons à cette question dans la discussion. Nous renvoyons à Calinescu, Schouten et Bhulai (2012) pour une extension des plans de collecte adaptatifs à l'erreur de mesure et à Beaumont et Haziza (2011) pour une discussion des plans de collecte adaptatifs et de la variance de non-réponse.

2.2.1 Fonctions de qualité fondées sur les covariables et fondées sur l'item

Lorsque l'on optimise la qualité conformément à (2.4), les fonctions de qualité appliquent à l'échantillon pour lequel on dispose de données appariées, de paradonnées et de réponses aux items une valeur unique qui peut être interprétée et optimisée. Lorsque les coûts sont minimisés sous des contraintes de qualité comme dans (2.5), la qualité peut être multidimensionnelle (mais les fonctions de coût doivent être unidimensionnelles).

En général, on distingue deux types de fonctions de qualité, celles qui emploient des covariables provenant de données appariées et de paradonnées seulement, et celles qui emploient également les réponses concernant les variables cibles (ou items) de l'enquête. Nous les désignons comme étant *fondées sur les covariables* et *fondées sur l'item*, respectivement. Une fonction de qualité fondée sur l'item est une fonction de la distribution des réponses à un item de l'enquête et de la distribution attendue, estimée, dans la population complète, étant donné les données appariées et les paradonnées disponibles. La principale distinction entre les fonctions de qualité fondées sur les covariables et celles fondées sur l'item est que les secondes nécessitent des hypothèses. Évidemment, les réponses des non-répondants manquent. Donc, l'évaluation de la qualité doit être fondée sur les relations entre les variables cibles et les covariables telles qu'elles sont observées dans les réponses. Par conséquent, les fonctions de qualité fondées sur l'item comportent un risque découlant directement du phénomène qu'elles essayent de mesurer. Les relations entre les variables cibles et les covariables pourraient être différentes chez les non-répondants et la qualité fondée sur l'item pourrait produire une image incomplète. De surcroît, dans les enquêtes comportant de nombreuses variables cibles, différentes variables peuvent mener à des décisions différentes et des plans de collecte de données optimaux différents. Cependant, contrairement aux fonctions de qualité fondées sur des covariables, les fonctions de qualité fondées sur l'item adaptent les plans de collecte de données spécifiquement aux thèmes de l'enquête. Les fonctions de qualité fondées sur les covariables ne peuvent être reliées qu'au biais de non-réponse des covariables qui sont incluses.

2.2.2 Optimisation de la qualité de la réponse

Pour commencer, nous décrivons brièvement un certain nombre de fonctions de qualité relevées récemment dans la littérature. Ensuite, nous discutons du choix d'une fonction de qualité.

La fonction de qualité fondée sur des covariables la mieux connue pour la non-réponse est le taux de réponse. Il ne s'agit pas d'une vraie fonction de qualité fondée sur des covariables en ce sens qu'il dépend de données appariées ou de paradonnées. Cependant, puisque l'indicateur de réponse 0-1 peut être considéré comme la forme la plus simple de paradonnées, nous le désignons comme une fonction de qualité fondée sur des covariables. Le taux de réponse est représenté comme la propension moyenne à répondre,

Taux de réponse : $$Q(p) = \bar{\rho} = \sum_{x,\tilde{x},s} q(x, \tilde{x}) p(s \mid x, \tilde{x}) \rho(x, \tilde{x}, s) \tag{2.6}$$

Schouten, Cobben et Bethlehem (2009) proposent deux fonctions de qualité fondées sur des covariables, l'indicateur R et une mesure qu'ils appellent le biais maximal ou le pire biais de non-réponse. Le nom donné à ce deuxième indicateur est trompeur, car il s'agit seulement d'un estimateur du biais maximal de la moyenne non corrigée des réponses, et non du biais maximal réel. Le coefficient de variation des propensions à répondre est une meilleure appellation que nous utiliserons ici. Les mesures peuvent s'écrire

Indicateur R : $$Q(p) = R(\rho_Z) = 1 - 2S(\rho_Z) \tag{2.7}$$

Coefficient de variation : $$Q(p) = CV(\rho_Z) = \frac{S(\rho_Z)}{\bar{\rho}}, \tag{2.8}$$

où la représentativité peut être évaluée par rapport aux données appariées seulement, $Z = X$, ou par rapport à un vecteur contenant à la fois les données appariées et les paradonnées, $Z = (X, \tilde{X})^T$. Les écarts-types de propension à répondre, $S(\rho_X)$ et $S(\rho_{X,\tilde{X}})$, peuvent s'écrire en fonction des probabilités d'affectation des stratégies sous la forme

$$S(\rho_X) = \sqrt{\sum_x q(x) \left(\sum_{\tilde{x},s} q(\tilde{x} \mid x) p(s \mid x, \tilde{x}) \hat{\rho}(x, \tilde{x}, s) - \bar{\rho} \right)^2} \tag{2.9a}$$

$$S(\rho_{X,\tilde{X}}) = \sqrt{\sum_{x,\tilde{x}} q(x, \tilde{x}) \left(\sum_s p(s \mid x, \tilde{x}) \hat{\rho}(x, \tilde{x}, s) - \bar{\rho} \right)^2}. \tag{2.9b}$$

Särndal et Lundström (2010) et Särndal (2011a et b) proposent des indicateurs dont la définition et la nature sont très similaires à (2.7) et (2.8). Ces indicateurs ont été dérivés dans la

perspective du calage et de la réponse dite équilibrée, et peuvent être utilisés pour remplacer (2.7) ou (2.8).

Un exemple de fonction de qualité fondée sur l'item pour la non-réponse est présenté par Groves et Heeringa (2006). Pour une variable cible particulière Y, Groves et Heeringa (2006) proposent le biais de non-réponse de la moyenne non corrigée des réponses,

Biais de non-réponse estimé : $$Q(p) = \frac{\mathrm{cov}(Y, \rho_X)}{\bar{\rho}}, \tag{2.10}$$

où $\mathrm{cov}(Y, \rho_X)$ est la covariance de réponse entre la variable cible et les propensions à répondre sachant les covariables X. Elle peut s'écrire sous la forme

$$\mathrm{cov}(Y, \rho_X) = \frac{\sum_x q(x)\rho_X(x)(\rho_X(x) - \bar{\rho}_X)(y(x) - \bar{y}_R)}{\bar{\rho}}, \tag{2.11}$$

avec $\rho_X(x)$ comme dans (2.3), $y(x)$ désignant la valeur moyenne de Y pour $X = x$ et $\bar{y}_R$, la moyenne attendue des réponses. De nouveau, l'expression (2.11) peut être étendue de manière à inclure les paradonnées $\tilde{X}$.

Toutes les fonctions de qualité décrites à la présente section sont définies comme des paramètres de population. En pratique, elles doivent être estimées d'après des données d'enquête. Les propensions à répondre réelles $\rho_X(x)$ doivent être remplacées par les estimateurs $\hat{\rho}_X(x)$, fondés sur une certaine forme de régression, et les sommations sur la population doivent être remplacées par des sommations pondérées par les poids de sondage sur l'échantillon. Nous revenons sur l'estimation des propensions à répondre à la section 2.4.

Maintenant, comment choisir une fonction de qualité ? Toutes les fonctions de qualité mentionnées ici ont pour objectif de mesurer l'effet de la non-réponse au-delà de la simple réduction de la taille de l'échantillon. Elles le font en se basant sur des covariables provenant de données appariées et de paradonnées. La raison qui justifie l'optimisation de ces fonctions de qualité est que des traces plus fortes d'erreur de non-réponse sur ces covariables pourraient

impliquer une erreur de non-réponse la plus importante sur d'autres variables également; les fonctions de qualité sont considérées comme des indicateurs de la qualité du processus plutôt que de la qualité du produit. Quoiqu'intéressante, cette conjecture requiert clairement un soutien empirique. Le choix d'une fonction de qualité pour la non-réponse doit être fondé sur l'ensemble de variables étudiées clés, les paramètres de population d'intérêt et les estimateurs qui seront utilisés. Le taux de réponse et l'indicateur R ne ciblent ni un paramètre de population ni un estimateur particulier. Le coefficient de variation est axé sur les moyennes de population, mais n'est particulier à aucune variable étudiée ni correction de la non-réponse, tandis que le biais de non-réponse estimé naît de la même perspective, mais ne s'applique qu'à une seule variable étudiée. Si une enquête comporte plusieurs variables étudiées clés, il faut éviter d'utiliser une fonction de qualité fondée sur l'item pour la non-réponse, car cela pourrait mener à des problèmes d'optimisation contradictoires. Si une enquête ne comporte qu'une seule variable clé, de bons résultats peuvent être obtenus en utilisant la fonction de qualité fondée sur l'item ou en se limitant aux covariables les plus pertinentes seulement dans des fonctions de qualité fondées sur des covariables. Dans le cas d'une enquête à usage multiple, il est, selon nous, trop contraignant de se concentrer sur un paramètre de population et un estimateur particulier, et nous privilégions l'indicateur R plutôt que toute autre fonction de qualité. Si l'on peut s'attendre à ce que les utilisateurs se concentrent sur les moyennes ou sur les totaux de population, il faut, à notre avis, donner la préférence au coefficient de variation, car il ne suppose pas l'utilisation d'une méthode d'ajustement particulière.

Cependant, l'ensemble de données appariées et de paradonnées qui est entré dans l'indicateur est encore plus important que le choix de l'indicateur proprement dit. Si une enquête ne comporte qu'une variable clé ou quelques-unes seulement, les données appariées et les paradonnées choisies peuvent, et doivent, être en rapport avec ces variables. Toutefois, si une enquête comprend une grande gamme de variables étudiées, on doit nécessairement se limiter aux variables auxiliaires qui distinguent généralement les personnes ou les ménages.

Jusqu'à présent, nous avons ignoré l'effet de la non-réponse sur la précision, alors que des exigences en matière de précision des estimations doivent être énoncées explicitement. Il existe deux options pour inclure la précision dans l'optimisation. Premièrement, on peut ajouter une contrainte supplémentaire. Le choix direct serait une contrainte concernant le nombre minimal de répondants, éventuellement pour un certain nombre de sous-groupes de population, évitant ainsi de spécifier les paramètres de population et les estimateurs. Deuxièmement, la variance due à la non-réponse peut être incluse dans la fonction de qualité proprement dite, c'est-à-dire que l'on considérerait des indicateurs de l'erreur quadratique moyenne. Cette option est proposée et élaborée par Beaumont et Haziza (2011). Sous cette seconde option, il faut de nouveau prendre en considération l'ensemble de variables étudiées clés, le paramètre de population et la stratégie d'estimation, car la précision est propre à un estimateur pour un paramètre de population donné pour une variable étudiée unique.

2.3 Fonctions de coût

Les fonctions de coût sont le pendant des fonctions de qualité. Elles comprennent plusieurs composantes. Il est important de limiter la spécification des coûts aux caractéristiques du plan de collecte que l'on fait varier dans le plan de collecte adaptatif. Par exemple, si c'est la prime d'incitation que l'on fait varier pour différentes sous-populations, il n'est pas nécessaire de spécifier ni de détailler les coûts relatifs aux durées de déplacement des intervieweurs. Quand le protocole concernant le moment de la prise de contact est la caractéristique du plan de collecte que l'on adapte, les durées et les coûts de déplacement jouent manifestement un rôle prépondérant.

Si un grand nombre de caractéristiques du plan de collecte sont optionnelles, les fonctions de coût prennent des formes complexes comportant de nombreuses composantes de coûts indirects et de coûts variables. En général, les coûts variables dépendent de la taille de l'échantillon, mais non les coûts indirects. Les composantes de coûts indirects peuvent être associées au personnel de

collecte des données, à l'échantillonnage et au traitement des échantillons. Les coûts variables découlent, par exemple, de la formation des intervieweurs, de l'impression et de l'envoi par la poste des questionnaires et des rappels, du traitement des questionnaires papier, du taux de rémunération horaire et des déplacements des intervieweurs, des primes d'incitation et de l'appariement des numéros de téléphone, de l'utilisation du téléphone et des serveurs informatiques.

Dans le problème d'optimisation, on peut préciser deux composantes de coût, l'une fixe et l'autre variable. La composante variable dépend de l'affectation des unités de population aux stratégies, tandis que la composante fixe comprend tous les autres coûts. Il convient de souligner que la composante fixe diffère selon la caractéristique de collecte sur laquelle est axé le plan de collecte des données adaptatif. La fonction de coût $C(p)$ est la somme de deux composantes

$$C(p) = C_F + C_V(p), \tag{2.12}$$

dont seulement la seconde, la composante variable, dépend des probabilités d'affectation.

En général, pour une stratégie de collecte s, les coûts $c(\tilde{x}, x, s)$ sont associés aux unités de population provenant du groupe $(\tilde{x}, x)$. La fonction de coût individuelle peut être une fonction des propensions à répondre, voire même plus spécifiquement, des propensions à la prise de contact et à la participation. Par exemple, les coûts d'intervieweur dans différents protocoles concernant le moment de la prise de contact dépendent des taux de prise de contact et des sous-population sélectionnées. La fonction de coût $c(\tilde{x}, x, s)$ est une fonction de coût relative car elle décrit uniquement la contribution de la stratégie à la composante de coût variable $C_V(p)$

$$C_V(p) = \sum_{\tilde{x}, x, s} q(\tilde{x}, x) p(s \mid x, \tilde{x}) c(\tilde{x}, x, s). \tag{2.13}$$

Trois remarques s'imposent. Premièrement, l'établissement des composantes de coût fixe et variable est compliqué quand un organisme d'enquête exécute de nombreuses enquêtes en parallèle. D'une part, l'interaction entre les enquêtes rend difficile la ventilation des coûts par

enquête, surtout quand les stratégies sont taillées sur mesure. Par ailleurs, quand plusieurs enquêtes sont réalisées, certaines composantes de coût variable peuvent être désignées comme étant fixes. Ainsi, si un nombre relativement faible seulement d'unités de population sont attribuées au mode d'interview sur place, on peut supposer que les frais de déplacement restent constants, car les adresses sont regroupées avec celles d'autres enquêtes. La deuxième remarque concerne l'aspect multidimensionnel des coûts. Mis à part le budget global, on peut demander que le taux d'activité des intervieweurs soit proche de 1 au cours du temps ou qu'aucun intervieweur ne doive faire plus d'heures supplémentaires qu'un nombre fixe déterminé. La fonction de coût devient alors un vecteur et la contrainte, un vecteur de contraintes. La troisième remarque a trait à la validité des fonctions de coût. Puisque ces dernières sont difficiles à construire en pratique, l'optimisation peut parfois être trop optimiste. Il est important de surveiller de près la collecte des données et d'intégrer des indicateurs dans les stratégies.

2.4 Estimation des probabilités de réponse

Après les paramètres de coût et les fonctions de qualité vient un autre élément important des plans de collecte adaptatifs, à savoir l'ensemble de propensions à répondre pour les diverses stratégies. Les données sur ces propensions à répondre doivent provenir d'enquêtes antérieures, de préférence la même enquête ou une enquête similaire. Ou bien, comme le proposent Groves et Heeringa (2006), on peut s'appuyer sur les premières phases de la collecte de données pour dériver ces propensions à répondre. Cela se fait aux dépens de l'efficacité, puisqu'une partie de l'enquête est déjà réalisée. Néanmoins, l'information recueillie directement est réinjectée dans l'enquête courante.

La littérature sur les enquêtes-ménages fournit une longue liste de modèles de réponse qui comprennent les caractéristiques du plan de collecte. Le dénominateur commun à tous ces modèles est que les propensions à répondre sont estimées en s'appuyant sur un certain nombre d'hypothèses quant à la nature réelle du mécanisme de formation des données manquantes dues à

la non-réponse. En général, ces modèles sont des simplifications. Conséquemment, les propensions à répondre attendues $\rho(x, s)$ présentent une erreur-type, et peuvent même être elles-mêmes biaisées lorsqu'elles sont fondées sur des enquêtes similaires, mais différentes. Dans le problème d'optimisation, cette incertitude peut être prise en compte en permettant aux propensions à répondre d'être des variables aléatoires plutôt que des quantités fixes. Le caractère aléatoire demande des analyses de sensibilité, ainsi que des évaluations de la robustesse de l'optimisation permettant de mieux comprendre la variation de la qualité des coûts quand l'enquête est réalisée plusieurs fois (dans les mêmes circonstances).

2.5 Le problème d'optimisation

L'optimisation de (2.3) et de (2.4) peut se faire selon deux approches : une approche par tâtonnement ou une optimisation mathématique. Dans le présent article, nous nous concentrons sur un cadre et une optimisation mathématiques, mais on peut aussi être plus modeste et introduire les plans de collecte adaptatifs progressivement au moyen d'études pilotes et d'essais sur le terrain.

Les fonctions de qualité (2.6), (2.7), (2.8) et (2.10) sont toutes des fonctions des probabilités d'affectation de la stratégie p. Le taux de réponse est une fonction linéaire des probabilités d'affectation, de sorte qu'il est relativement facile de l'optimiser en se servant d'un logiciel d'optimisation standard (par exemple, le progiciel *linprog* en R ou tout autre logiciel permettant de résoudre des problèmes de programmation linéaire). Néanmoins, autant que nous sachions, en raison de la forte dimensionnalité de p, il n'existe aucune solution explicite de (2.4), même pour les problèmes linéaires. Cependant, en général, les fonctions de qualité sont des fonctions non linéaires non convexes en ce qui concerne les probabilités d'affectation et elles ne peuvent pas être optimisées sans recourir à des méthodes numériques ou de Monte Carlo. La complexité du problème augmente rapidement avec le nombre de stratégies considérées et le nombre de sous-groupes fondés sur les données appariées et les paradonnées.

Les logiciels statistiques courants contiennent des procédures ou des modules externes qui permettent de traiter les problèmes d'optimisation non linéaires, comme *nlm* ou *nlminb* en R ou *proc optmodel* en SAS. Cependant, les problèmes non linéaires non convexes nécessitent parfois de longs temps de calcul ou peuvent converger vers des optima locaux. Pour cette raison, il est recommandé d'utiliser des logiciels d'optimisation spécialisés, tels que Xpress, Baron ou AMPL.

Dans les exemples des sections 3 et 4, nous exécutons un certain nombre d'optimisations. Le problème d'optimisation de la section 3 est relativement simple; la fonction d'objectif de qualité est l'indicateur R qui est évalué sur deux sous-groupes de population. Pour les deux sous-groupes, l'optimisation peut être réécrite sous forme d'un problème de programmation linéaire. Dans le cas de l'exemple de la section 4, nous avons pu construire un algorithme qui converge vers la solution optimale en un petit nombre d'étapes. Toutes les optimisations ont été programmées en R et le code peut être obtenu sur demande.

3 Un plan de collecte adaptatif dynamique : réaffectation des intervieweurs dans une enquête de suivi

À la présente section, nous donnons un exemple de plan adaptatif dynamique ayant trait à la réaffectation des intervieweurs en s'appuyant sur des observations de la propension à coopérer. L'exemple est fondé sur des propensions à répondre et des fonctions de coût hypothétiques. Les intervieweurs sont affectés aux cas qui ont refusés de répondre une fois, d'après une évaluation de la propension à répondre faite durant une première phase de l'enquête. L'évaluation est faite pour les répondants et pour les personnes qui refusent de répondre, mais elle n'est pas disponible pour les unités échantillonnées avec lesquelles un contact n'a pas été pris durant la première phase. Elle fournit un jugement de la propension qu'auront les unités échantillonnées à participer à l'enquête lorsque l'on prendra de nouveau contact avec elles. L'évaluation est faite sur une échelle de trois points : *forte, moyenne, faible*. Une propension forte signifie qu'il existe une probabilité élevée que l'unité échantillonnée réponde s'il est de nouveau pris contact avec elle.

Après une première phase de collecte des données, les résultats intermédiaires de la collecte sont évalués et les unités échantillonnées sont réparties en trois catégories, à savoir répondants, cas de refus et cas de non-contact. Un traitement différent est appliqué aux cas de refus. Les intervieweurs sont évalués en se basant sur leur rendement historique et groupés en *bons* et *moins bons* intervieweurs. Les cas de refus sont réaffectés à l'un des deux groupes d'intervieweurs. Comme il n'existe pas d'évaluation disponible pour les cas de non-contact, le traitement de ce groupe n'est pas modifié.

Nous utilisons l'indicateur R donné par (2.7) comme fonction d'objectif de qualité. Nous répartissons l'échantillon en utilisant $X = (\text{âge})$ en deux groupes, nommés *personnes jeunes* et *personnes âgées*. L'objectif à la deuxième phase est d'affecter les cas de refus aux deux groupes d'intervieweurs de manière que l'indicateur R en ce qui concerne l'âge soit maximisé.

Soit n la taille de l'échantillon de l'enquête. Les proportions de population des deux sous-populations, *personnes jeunes* et *personnes âgées*, sont désignées par $q(1)$ et $q(2)$. Soit $q(\tilde{x} \mid x)$ la probabilité conditionnelle qu'une unité échantillonnée provenant de la sous-population d'âge x soit de type $\tilde{x}$, où $\tilde{x} \in \{\text{forte, moyenne, faible}\}$. En outre, soit $\lambda(x, \tilde{x})$ la probabilité qu'une unité échantillonnée de type $\tilde{x}$ provenant de la sous-population d'âge x soit un cas de refus. Si une personne n'est pas un cas de refus, alors $\mu(x, \tilde{x})$ est la probabilité que cette personne était un répondant après la première phase ou qu'elle devienne un répondant si elle était un cas de non-contact après la première phase.

Le nombre total d'intervieweurs est M et $p_s M$ représente le nombre d'intervieweurs ayant la compétence $s \in S = \{\text{bon, moins bon}\}$, $0 \leq p_s \leq 1$ et $p_{\text{bon}} + p_{\text{moins bon}} = 1$. L'ensemble S forme l'ensemble de stratégies, c'est-à-dire que nous voulons attribuer chaque cas de refus à un bon intervieweur ou à un moins bon intervieweur. Nous supposons que chaque intervieweur peut traiter au plus c cas de refus à la deuxième phase de l'enquête. La probabilité qu'un cas de refus de type $\tilde{x}$ provenant de la sous-population x réponde s'il est contacté par un intervieweur ayant

la compétence s est désignée par $\rho(s, x, \tilde{x})$ et nous supposons de nouveau qu'elle est connue d'après les enquêtes antérieures.

Soit $\{p(s \mid x, \tilde{x})\}_{x,\tilde{x}}$ l'ensemble de variables décisionnelles, où $p(s \mid x, \tilde{x})$ représente la probabilité qu'une unité échantillonnée de type $\tilde{x}$ soit attribuée à un intervieweur de compétence s sachant qu'elle appartient à la sous-population x. Autrement dit, nous permettons l'affectation aléatoire des unités échantillonnées aux deux groupes d'intervieweurs.

Dans cet exemple, nous exprimons les coûts en fonction des taux globaux d'activité des intervieweurs. Puisque les intervieweurs peuvent traiter au plus c cas, il existe deux contraintes,

$$n\sum_{x,\tilde{x}} q(x)q(\tilde{x} \mid x)p(s \mid x, \tilde{x})\,\lambda(x, \tilde{x}) \le Mp_s c, \quad \forall s \in S.$$

Autrement dit, le nombre total de cas de refus qui peuvent être attribués aux intervieweurs de compétence s est limité à la charge de travail maximale possible pour le groupe ayant cette compétence.

La propension à répondre d'une unité provenant de la sous-population x peut maintenant être dérivée sous la forme

$$\sum_{\tilde{x}} q(\tilde{x} \mid x)\left[\,(1-\lambda(x, \tilde{x}))\mu(x, \tilde{x}) + \lambda(x, \tilde{x})\sum\nolimits_s p(s \mid x, \tilde{x})\rho(s, x, \tilde{x})\,\right],$$

et représente la valeur entrée dans l'indicateur R.

Considérons maintenant les données d'entrée suivantes pour l'exemple : une taille d'échantillon de $n = 2\,000$, un total de 80 intervieweurs, $M = 80$, une charge de travail maximale de 30 cas par intervieweur, $c = 30$, une distribution selon l'âge égale à $q(1) = q(2) =$ 0,5, les distributions conditionnelles de type de refus $q(\tilde{x} \mid 1) = (0{,}2; 0{,}3; 0{,}5)'$ et $q(\tilde{x} \mid 2) = (1/3\,; 1/3\,; 1/3)'$ et la classification de 25 % des intervieweurs comme étant bons, $p_1 = 0{,}25 = 1 - p_2$.

Les tableaux 3.1 et 3.2 donnent les probabilités de réponse hypothétique $\rho(s, x, \tilde{x})$ pour les deux sous-groupes quand on applique la conversion des cas de refus, de même que les probabilités de coopération $\mu(x, \tilde{x})$ et les probabilités de refus $\lambda(x, \tilde{x})$.

Nous optimisons l'indicateur R par rapport aux deux groupes d'âge. Dans le cas de deux strates, on peut montrer que l'indicateur R est maximal quand la distance absolue entre les propensions à répondre dans les deux strates est minimale. La valeur optimale de l'indicateur R s'avère être de 0,827. Le tableau 3.3 donne les valeurs optimales des variables décisionnelles; toutes ces variables sauf une, $p(s \mid x, \tilde{x})$, ont une valeur de 0 ou de 1, c'est-à-dire que les réaffectations sont principalement non probabilistes. Fait exception la sous-population des personnes jeunes pour lesquelles la propension à répondre est évaluée comme étant moyenne.

Tableau 3.1
Probabilités de réponse quand la conversion des refus est appliquée aux personnes jeunes et âgées qui refusent de répondre sachant l'évaluation de la propension à répondre

	Refus d'une personne jeune					
	Bon intervieweur			**Moins bon intervieweur**		
	Forte	**Moyenne**	**Faible**	**Forte**	**Moyenne**	**Faible**
$\rho(s, 1, \tilde{x})$	0,8	0,6	0,4	0,7	0,5	0,3
	Refus d'une personne âgée					
	Bon intervieweur			**Moins bon intervieweur**		
	Forte	**Moyenne**	**Faible**	**Forte**	**Moyenne**	**Faible**
$\rho(s, 2, \tilde{x})$	0,9	0,7	0,5	0,8	0,6	0,4

Tableau 3.2
Probabilités de refus et de coopération à la première phase de collecte des données

	Personne jeune			**Personne âgée**		
	Forte	**Moyenne**	**Faible**	**Forte**	**Moyenne**	**Faible**
$\lambda(x, \tilde{x})$	0,5	0,6	0,7	0,2	0,3	0,4
$\mu(x, \tilde{x})$	0,85	0,8	0,76	0,95	0,93	0,91

Tableau 3.3
Affectation optimale des cas aux intervieweurs

	Personne jeune			**Personne âgée**		
	Forte	**Moyenne**	**Faible**	**Forte**	**Moyenne**	**Faible**
Bon intervieweur	1	0,83	1	0	0	0
Moins bon intervieweur	0	0,17	0	1	1	1

Il est utile de comparer l'affectation optimale à une affectation aléatoire des intervieweurs afin d'évaluer le gain d'efficacité. Si nous affectons aléatoirement les cas de refus aux intervieweurs, la valeur de l'indicateur R est égale à 0,749. Donc, l'affectation optimale donne lieu à une augmentation considérable de la valeur de l'indicateur R. Les taux de réponse sont, respectivement, de 72,0 % et de 70,1 % pour les affectations optimale et aléatoire.

Si nous augmentons le nombre d'intervieweurs tout en fixant le nombre maximal de cas par intervieweur ainsi que les autres paramètres, pour tout nombre d'intervieweurs supérieur à $M = 84$, l'indicateur R ne s'améliore pas. Les deux groupes d'intervieweurs sont suffisamment grands pour traiter tous l'échantillon et la contrainte de coût n'est plus vraiment une contrainte. L'indicateur R pour $M = 84$ est égal à 0,830 et le taux de réponse est de 72,1 %. Si nous maximisons le taux de réponse plutôt que l'indicateur R, l'affectation des intervieweurs convergera vers l'affectation de tous les cas aux *bons* intervieweurs seulement.

4 Un plan de collecte adaptatif statique : affectation des intervieweurs travaillant par téléphone

À la présente section, nous présentons une étude en simulation où la caractéristique du plan de collecte d'intérêt est l'affectation des intervieweurs travaillant par téléphone. Les probabilités de réponse utilisées dans l'exemple sont estimées après les données d'une enquête téléphonique réelle.

L'Enquête sur la satisfaction des consommateurs (ESC) des Pays-Bas est une enquête téléphonique mensuelle sur les sentiments des ménages au sujet de leur situation économique et de leurs dépenses. L'enquête donne un aperçu du développement économique à court terme et des indicateurs précoces des différences dans les tendances de consommation. Chaque mois, 1 500 ménages sont échantillonnés. Les deux causes les plus influentes de non-réponse dans l'enquête sont la non-prise de contact et le refus de participer. De l'échantillon sélectionné, 95 % sont contactés et des ménages contactés, 71 % participent. Le taux de réponse est de 67 %.

L'intervieweur est l'un des facteurs qui influent le plus sur la participation. Le rendement des intervieweurs peut varier considérablement quand il s'agit d'obtenir une réponse. En tout, 60 intervieweurs ont travaillé à l'ESC en 2005. Cela signifie qu'un intervieweur a pris contact avec 280 ménages en moyenne. Les taux de participation par intervieweur variaient de 50 % à 79 %. Le taux le plus faible, soit 50 %, était toutefois exceptionnel, car l'avant-dernier taux de participation était de 61 %. Le taux de participation moyen par intervieweur était de 67 %. Les ménages ont été affectés aléatoirement aux intervieweurs dans le système de gestion de l'ITAO. Donc, en ce qui concerne l'intervieweur, les données ont été randomisées (ou entremêlées). Dans la suite de l'exposé, l'intervieweur sera la caractéristique d'intérêt du plan de collecte. L'ensemble de stratégies de collecte S comprend soixante stratégies, $S = \{s_1, s_2, \ldots, s_{60}\}$.

Partant des variables auxiliaires disponibles, nous avons sélectionné un vecteur X contenant l'ethnicité, la composition du ménage en ce qui concerne le sexe (masculin, féminin ou mélange), l'âge moyen du noyau du ménage par tranche de cinq ans, le type de ménage, le degré d'urbanisation du quartier de résidence et la valeur moyenne des logements dans le quartier. L'âge, la valeur moyenne des logements et le type de ménage sont surtout reliés aux statistiques clés déduites de l'ESC. Aucune paradonnée n'était disponible dans la présente étude. Par conséquent, le plan de collecte adaptatif est statique. Dans le problème d'optimisation, les probabilités d'affectation $p(s_k \mid x)$ doivent être choisies, c'est-à-dire qu'il faut décider à quel intervieweur sont affectées les sous-populations en se basant sur X $\left(\text{de sorte que } \sum_k p(s_k \mid x) = 1\right)$.

Le coefficient de variation des propensions à répondre ρ_X défini par (2.8) est sélectionné comme étant la fonction de qualité cible. Pour estimer les propensions à répondre $\rho(s_k, x)$ pour les intervieweurs, nous utilisons un modèle multiniveaux comportant une fonction de lien identité, c'est-à-dire une régression linéaire comprenant deux niveaux. Les intervieweurs forment le premier niveau du modèle et les ménages, le second. Le modèle multiniveaux est utilisé pour séparer les propensions à répondre individuelles et les propensions à répondre selon

l'intervieweur. La logique est qu'en faisant cette distinction, l'effet d'intervieweur peut être isolé et l'affectation des intervieweurs peut être optimisée. Nous choisissons un modèle linéaire car il est facile à optimiser. Puisque les propensions ne sont jamais proches de 0 ni de 1, le modèle linéaire produit presque les mêmes estimations qu'un modèle logit ou probit.

En ce qui concerne l'effet d'intervieweur, nous avons d'abord cherché à déterminer s'il était suffisant d'utiliser un modèle multiniveaux à pente fixe, c'est-à-dire un modèle dans lequel l'intervieweur est ajouté comme un effet principal seulement et dans lequel il n'y a aucune interaction avec les variables auxiliaires. Toutes les covariables présélectionnées ont contribué de manière significative au modèle multiniveaux, mais aucune des interactions avec l'intervieweur n'était significative au seuil de signification de 5 %. Par conséquent, nous nous sommes limités au modèle d'effet principal suivant

$$\rho(s_k, x_i) = \beta_0 + \beta_X x_i + \beta_k \tag{4.1}$$

où x_i est le vecteur de covariables du ménage $i, 1 \leq i \leq n, n$ est la taille d'échantillon, β_k est l'effet (fixe) d'intervieweur pour l'intervieweur k, β_0 est le terme constant ou ordonnée à l'origine, et β_X est le paramètre de pente. Soit $\rho(x_i) = \sum_k p(s_k \mid x_i)\, \rho(s_k, x_i)$ la propension à répondre de l'unité échantillonnée i.

Le modèle (4.1) a été ajusté au jeu de données de l'ESC. Puis, nous nous sommes servis de l'effet d'intervieweur estimé β_k pour optimiser le coefficient de variation, sous deux contraintes de coût : la durée totale de l'interview ainsi que le nombre individuel d'appels faits par chaque intervieweur devaient être les mêmes que dans le plan original. Puisque le système de gestion des appels téléphoniques traite les appels, la durée de l'interview est la composante dominante du coût. Si nous fixons la durée totale de l'interview, nous contraignons le coût d'être le même que pour l'ESC ordinaire. Puisque les intervieweurs ne peuvent traiter qu'un certain nombre d'appels, nous devons aussi fixer le nombre d'appels qui leur sont affectés. La première contrainte implique que nous fixons le taux de réponse, car la durée totale d'interview est égale au produit

de la durée individuelle moyenne d'interview et du nombre de répondants. Le questionnaire de l'ESC est simple et ne contient aucun groupe d'items emboîtés. Par conséquent, la durée individuelle d'interview ne varie pour ainsi dire pas sur les sous-groupes de population. La deuxième contrainte est égale à

$$\sum_i p(s_k \mid x_i) = n_k, \tag{4.2}$$

où n_k est le nombre spécifié préalablement d'appels pour l'intervieweur k et $\sum_k n_k = n$.

Nous optimisons le coefficient de variation en répartissant les β_k entre les ménages. Étant donné la nature additive du modèle, il est facile de montrer que toute permutation des intervieweurs attribués aux cas mène à la même propension moyenne à répondre et, donc, aux mêmes durée et coût d'interview. La propension moyenne à répondre est donnée par

$$\bar{\rho} = \frac{1}{n}\sum_i \rho(x_i) = \frac{1}{n}\sum_{i,k} p(s_k \mid x_i)(\beta_0 + \beta_X x_i + \beta_k) = \beta_0 + \frac{1}{n}\sum_i \beta_X x_i + \frac{1}{n}\sum_k n_k \beta_k,$$

qui ne dépend pas de l'ensemble de probabilités d'affectation $p(s_k \mid x)$. Par conséquent, optimiser le coefficient de variation revient à optimiser la variance des propensions à répondre $S^2(\rho_X)$.

Si nous nous limitons à des variables décisionnelles de type 0-1, c'est-à-dire $p(s_k \mid x) \in \{0, 1\}, \forall x, k,$ alors il est relativement facile de montrer que l'affectation optimale consiste à unir les meilleurs intervieweurs aux unités échantillonnées les plus difficiles et inversement. Autrement dit, les unités échantillonnées sont triées en classant les propensions à répondre individuelles sans l'effet d'intervieweur, $\beta_0 + \beta_X x_i$, par ordre croissant, et les intervieweurs sont triés par ordre décroissant en se basant sur l'effet d'intervieweur, β_k. Si deux unités échantillonnées i et j sont affectées à deux intervieweurs différents, disons k et l, et que $\beta_X x_i < \beta_X x_j, \beta_k < \beta_l$ et $p(s_k \mid x_i) = p(s_l \mid x_j) = 1$, alors il est optimal de permuter les deux intervieweurs, c'est-à-dire $p(s_l \mid x_i) = p(s_k \mid x_j) = 1$. Nous pouvons le montrer comme il suit. La différence de variance $S^2(\rho_X)$ est proportionnelle à

$$\begin{aligned}\Delta S^2(\rho_X) &= (\beta_0 + \beta_X x_i + \beta_k - \overline{\rho})^2 \\ &\quad + (\beta_0 + \beta_X x_j + \beta_l - \overline{\rho})^2 - (\beta_0 + \beta_X x_i + \beta_l - \overline{\rho})^2 - (\beta_0 + \beta_X x_j + \beta_l - \overline{\rho})^2 \\ &= 2(\beta_l - \beta_k)(\beta_0 + \beta_X x_j - \overline{\rho}) - 2(\beta_l - \beta_k)(\beta_0 + \beta_X x_i - \overline{\rho}) \\ &= 2(\beta_l - \beta_k)(\beta_X x_j - \beta_X x_i) > 0.\end{aligned} \tag{4.3}$$

Partant de (4.3), nous pouvons conclure à une diminution de la variance et, donc, du coefficient de variation si nous permutons les deux intervieweurs pour les cas i et j. De cet argument, il découle facilement que la solution optimale est celle proposée. D'une manière similaire, mais qui nécessite plus d'opérations algébriques, on peut montrer que la solution optimale pour les affectations probabilistes, $p(s_k \mid x) \in [0, 1]$, est la même.

Les deux premières lignes du tableau 4.1 contiennent la propension moyenne à répondre et le coefficient de variation avant et après la réaffectation des intervieweurs. Le coefficient de variation est passé de 0,117 à 0,035. Afin d'avoir une idée du degré de signification de la variation de la fonction de qualité, nous avons calculé les erreurs-types par la méthode du bootstrap. Pour chaque bootstrap, nous avons effectué la réaffectation des intervieweurs. Les erreurs sont données au tableau 4.1.

Tableau 4.1
Propension moyenne à répondre et coefficient de variation pour l'ESC ordinaire, l'ESC après réaffectation des intervieweurs avec et sans correction pour la durée de l'interview (les erreurs-types bootstrap sont entre parenthèses)

ESC	Correction pour la durée de l'interview ?	$\overline{\rho}$	$Q(p)$
Ordinaire	-	70,8 %	0,117 (0,005)
Réaffectation	Non	70,8 %	0,035 (0,003)
Réaffectation	Oui	70,8 %	0,034 (0,003)

Le lecteur aura peut-être remarqué que des nombres fixes de cas par intervieweur n'impliquent pas des nombres fixes d'interview par intervieweur. En fait, en réarrangeant les intervieweurs, ceux qui sont bons feront moins d'interviews car ils obtiennent les cas difficiles, alors que ceux qui sont moins bons feront plus d'interviews. Par conséquent, les bons

intervieweurs travailleront un plus petit nombre d'heures qu'ils ne le feraient dans le cadre de l'ESC ordinaire et les moins bons intervieweurs travailleront plus. Cela serait un effet secondaire indésirable qui peut, toutefois, être corrigé assez facilement. En partant de la solution optimale et en triant de nouveau les unités échantillonnées en fonction de leur propension à répondre individuelle sans tenir compte de l'effet d'intervieweur, nous pouvons déplacer les cas avoisinants des intervieweurs moins bons aux bons intervieweurs. Nous effectuons ce déplacement de telle façon que la durée totale d'interview par intervieweur ne dépasse pas celle de l'ESC ordinaire. On peut de nouveau prouver que cette procédure donne lieu à une nouvelle solution optimale où la contrainte du nombre fixe de cas présente dans (4.2) est remplacée par la contrainte du nombre fixe d'interviews

$$\sum_i p(s_k \mid x_i)\, \rho(s_k, x_i) = r_k, \tag{4.4}$$

où r_k est le nombre d'interviews spécifié préalablement. Le tableau 4.1 donne le coefficient de variation pour la solution optimale étant donné (4.4). Le taux de réponse demeure fixe et le coefficient de variation est légèrement plus faible.

En 2009, l'ESC a été utilisée comme instrument pour tester un plan de collecte adaptatif statique. Nous renvoyons le lecteur à Luiten et Wetzels (2009) et à Luiten et Schouten (2013) pour des renseignements détaillés. L'affectation des intervieweurs était l'une des principales caractéristiques du plan de collecte qui a été adaptée. Les autres caractéristiques du plan étaient le mode de collecte des données et le protocole de prise de contact. Outre l'interview par téléphone, l'interview en ligne a également été choisie comme mode possible de collecte. Les unités échantillonnées dont la probabilité de contact estimée était faible ont été affectées à des protocoles de prise de contact plus intensifs et ont été classées par ordre de priorité. Les probabilités de prise de contact et les probabilités de réponse ont été estimées en se basant sur les données historiques de l'ESC. L'essai pilote a montré qu'il était possible d'améliorer significativement le coefficient de variation, tout en fixant le taux de réponse et le budget.

À la présente section, nous avons présenté une étude en simulation dans laquelle les bons intervieweurs par téléphone se voient attribuer les cas les plus difficiles. En pratique, cela pourrait irriter ces intervieweurs. Lorsqu'un plan de ce genre est mis en œuvre, il convient d'en informer les intervieweurs à l'avance. Durant l'essai pilote de l'ESC de 2009, cela n'a entraîné aucun commentaire négatif de la part des intervieweurs. Dans les enquêtes sur place, une réaffectation des intervieweurs ne peut pas avoir lieu aussi facilement, car les frais de déplacement peuvent varier considérablement. Néanmoins, dans les régions à forte densité de population, la réaffectation des intervieweurs peut être une option.

5 Discussion

Le présent article décrit les plans de collecte de données d'enquête dans lesquels différents traitements ou stratégies de collecte sont appliqués à différentes unités de la population. Les différences entre les unités de population sont reflétées par des covariables pour lesquelles on dispose de données de registre appariées ou de paradonnées. Les stratégies de collecte des données sont définies comme étant différentes spécifications des caractéristiques du plan de collecte. Les plans de ce genre sont appelés plans de collecte adaptatifs car ils adaptent (taillent sur mesure) la collecte des données à la population d'intérêt. Les éléments fondamentaux des plans de collecte adaptatifs sont les stratégies de collecte, les covariables de population, les propensions à répondre, les fonctions de coût et de qualité, et les probabilités d'affectation des stratégies. Les plans de collecte adaptatifs ont pour objectif d'optimiser la qualité des réponses en affectant différentes stratégies à différentes unités de population. Les probabilités d'affectation des stratégies représentent les variables décisionnelles dans l'optimisation.

Selon nous, le présent article contribue à la littérature de trois façons : il présente un cadre général, il donne explicitement l'option de choisir parmi un ensemble de stratégies en se basant sur un compromis qualité-coût, et il se concentre sur des indicateurs de l'erreur de non-réponse. Les deux dernières composantes sont déjà décrites dans la littérature sur les enquêtes; c'est la

généralisation à plusieurs caractéristiques du plan de collecte et à l'erreur de non-réponse qui est nouvelle. Sous leur forme la plus modeste, les plans de collecte adaptatifs représentent une affectation des stratégies d'enquête stratifiée sur différents sous-groupes de population. Dans leur forme la plus ambitieuse, les plans sont des extensions des plans d'échantillonnage à des stratégies multiples en mettant l'accent sur l'erreur de non-réponse. Cependant, même sous leur forme la plus modeste, les plans de collecte adaptatifs peuvent englober les modes de collecte des données, les primes d'incitation, les rappels, la durée du travail sur le terrain dans les enquêtes avec interview sur place, les tâches des intervieweurs et le type de déclaration.

Les plans de collecte adaptatifs se prêtent le mieux aux circonstances dans lesquelles les enquêtes sont exécutées de manière répétée pendant une longue période. C'est dans ces conditions que les données historiques sont les plus puissantes. Les plans adaptatifs conviennent aussi pour les organismes qui réalisent de nombreuses enquêtes dont les thèmes et le budget sont relativement similaires. Les nouvelles enquêtes et les enquêtes ponctuelles demandent de la modestie et de la prudence. Cependant, cela est vrai aussi pour les plans à stratégie unique. Les plans de collecte adaptatifs peuvent tenir compte du manque de données historiques puissantes en permettant une incertitude dans les propensions à répondre et d'autres paramètres, et en introduisant une période d'apprentissage ou une phase de conception initiale.

À notre avis, la concentration sur l'erreur de non-réponse est un élément important du cadre. Dans le présent article, notre objectif est la représentativité de la réponse. Cet objectif vient de notre conviction que la non-réponse est toujours de type ne manquant pas au hasard (*not-missing-at-random*). Nous observons les écarts plus importants par rapport au mécanisme de formation de données manquantes entièrement au hasard (*missing-completely-at-random*) pour les variables auxiliaires pertinentes comme des indices d'une non-réponse produite au hasard plus forte sur les variables étudiées étant donné ces variables auxiliaires. Théoriquement, il ne doit pas nécessairement en être ainsi. Considérons une simple question à réponse binaire oui-non et un taux de non-réponse de 50 %. Ces cas extrêmes se produiraient si les non-répondants disaient

tous soit oui soit non. Ils peuvent le faire quel que soit le choix des variables auxiliaires et par conséquent le biais de non-réponse maximal sur cette question est le même indépendamment du choix des variables auxiliaires. Donc, la recherche fournira un soutien empirique pour la concentration sur les mesures indirectes de l'erreur de non-réponse.

Il est également nécessaire de poursuivre les travaux de recherche portant sur les plans de collecte adaptatifs et « réactifs » dans le contexte d'autres questions. La recherche devrait viser à étendre les plans à plusieurs erreurs d'enquête et à étudier la robustesse des plans à l'erreur de spécification des modèles de propension à répondre. Jusqu'à présent, les plans de collecte adaptatifs et « réactifs » ont été axés sur l'erreur de non-réponse, en ignorant l'erreur de réponse ou de mesure. Or, il est bien connu que certaines caractéristiques du plan de collecte des données, par exemple le mode de collecte des données ou les intervieweurs, peuvent avoir un effet important sur l'erreur de réponse et, par conséquent, sur l'erreur totale d'enquête. Les plans de collecte adaptatifs doivent, par conséquent, tenir compte de l'erreur de mesure également, quand on peut s'attendre à ce que les caractéristiques du plan auront un effet différentiel important sur l'erreur de réponse. L'optimisation tenant compte d'erreurs multiples représente un important domaine de recherche future.

Les plans de collecte des données adaptatifs devraient dans tous les cas être modestes en ce qui concerne le nombre de stratégies employées, afin d'éviter un processus de collecte exagérément complexe et l'optimisation sur des propensions et des fonctions de coût qui sont sujettes à incertitude. Néanmoins, il convient toujours de privilégier une méthode structurée d'examen; les plans de collecte adaptatifs fournissent ce genre de cadre et permettent une recherche structurée de plans de collecte des données améliorés.

Remerciements

Les auteurs remercient James Wagner, Mick Couper, Fannie Cobben et Mariëtte Vosmer de leurs commentaires utiles au sujet de la version provisoire du présent article. Les auteurs

remercient aussi le rédacteur adjoint et les examinateurs de leurs remarques qui leur ont permis d'améliorer considérablement l'article.

Bibliographie

Barón, J.D., Breunig, R.V., Cobb-Clark, D., Gørgens, T. et Sartbayeva, A. (2009). Does the effect of incentive payments on survey response rates differ by income support history. *Journal of Official Statistics*, 25, 483-507.

Beaumont, J.-F., et Haziza, D. (2011). A theoretical framework for adaptive collection designs. Document présenté au 5e atelier de l'International Total Survey Error, 21 au 23 juin, Québec, Canada.

Calinescu, M., Schouten, B. et Bhulai, S. (2012). Adaptive survey designs that minimize nonresponse and measurement risk, document de travail, Statistics Netherlands. Disponible au www.cbs.nl.

Cobben, F. (2009). Nonresponse in sample surveys. Methods for analysis and adjustment, Thèse de doctorat, University of Amsterdam.

De Leeuw, E.D. (2008). Choosing the Method of Data Collection. Dans *International Handbook of Survey Methodology*, (Éds., E.D. De Leeuw, J.J. Hox et D.A. Dillman), Larence Erlbaum Associates, New York, États-Unis, 113-135.

De Leeuw, E., Callegaro, M., Hox, J., Korendijk, E. et Lensvelt-Mulders, G. (2007). The influence of advance letters on response in telephone surveys. *Public Opinion Quarterly*, 71, 413-443.

Dillman, D.A. (2007). Mail and internet surveys: The tailored design method, Wiley, Hoboken, New Jersey, États-Unis.

Greenberg, B.S., et Stokes, S.L. (1990). Developing an optimal call scheduling strategy for a telephone survey. *Journal of Official Statistics*, 6, 421-435.

Groves, R.M. (1989). Survey errors and survey costs, Wiley, New York, États-Unis.

Groves, R.M., et Couper, M.P. (1998). Nonresponse in Household Interview Surveys, Wiley series in probability and statistics, Survey methodology section.

Groves, R.M., Dillman, D., Eltinge, J. et Little, R. (2002). Survey Nonresponse, New York : Wiley Series in Probability and Statistics.

Groves, R.M., et Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society*, Séries A, 169, 439-457.

Hartley, H.O., et Monroe, H. (1979). Interviewer assignments which minimize the effects of nonsampling errors. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 216-220.

Heyd, J.M., et Carlin, B.P. (1999). Adaptive design improvements in the continual reassessment method for Phase I studies. *Statistics in Medicine*, 18, 1307-1321.

Kalsbeek, W.D., Botman, S.L., Massey, J.T. et Liu, P.W. (1994). Cost-efficiency and the number of allowable call attempts in the National Health Interview Survey. *Journal of Official Statistics*, 10, 133-152.

Kersten, H.M.P., et Bethlehem, J.G. (1984). Exploring and reducing the non-response bias by asking the basic question, BPA 627-84-M1, CBS, Voorburg.

Kulka, R.A., et Weeks, M.F. (1988). Towards the development of optimal calling protocols for telephone surveys: A conditional probabilities approach. *Journal of Official Statistics*, 4, 319-322.

Laflamme, F., et Karaganis, M. (2010). Implementation of responsive collection design for CATI surveys at Statistics Canada, document présenté à la Q2010, 3 au 6 mai, Helsinki, Finland.

Luiten, A., et Schouten, B. (2013). Adaptive fieldwork design to increase representative household survey response. *Journal of the Royal Statistical Society*, Séries A, 176, 1, 169-190.

Luiten, A., et Wetzels, W. (2009). Indicators and data collection control. Work plan and preliminary findings pilot Statistics Netherlands, RISQ deliverable 8.2, www.risq-project.eu.

Lyberg, L.E., Biemer, P., Collins, M., de Leeuw, E.D., Dippo, C., Schwarz, N. et Trewin, D. (1997). Survey measurement and process quality, Wiley Series in Probability and Statistics, Wiley, Hoboken, New Jersey, États-Unis.

Lynn, P. (2003). PEDAKSI: Methodology for Collecting Data about Survey Non-Respondents. *Quality and Quantity*, Vol. 37, 3, 239-261.

Mohl, C., et Laflamme, F. (2007). Research and responsive design options for survey data collection at Statistics Canada. Proceedings of ASA Joint Statistical Meeting, Section 293, 29 juillet au 2 août, Salt Lake City, États-Unis.

Moore, J. (1988). Self/Proxy response status and survey response quality. A review of literature. *Journal of Official Statistics*, 4, 2, 155-172.

Murphy, S.A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society*, Séries B, 65, 331-355.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. et Lindblad, M. (2010). Reduction of Nonresponse Bias in Surveys through Case Prioritization, Survey Research Methods, 4, 1, 21-29.

Phillips, O., et Tabuchi, T. (2009). Un plan de sondage réactif pour l'Enquête sur la dynamique du travail et du revenu. *Recueil : Symposium 2009, Les enquêtes longitudinales : de la conception à l'analyse*, Statistique Canada.

Särndal, C.-E. (2011a). Three factors to signal non-response bias with applications to categorical auxiliary variables. *Revue Internationale de Statistique*, 79, 2, 233-254.

Särndal, C.-E. (2011b). The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27, 1, 1-21.

Särndal, C.-E., et Lundström, S. (2010). Plan d'estimation : détermination de vecteurs auxiliaires en vue de réduire le biais de non-réponse. *Techniques d'enquête*, 36, 2, 141-156.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 107-121.

Schouten, B., Luiten, A., Loosveldt, G., Beullens, K. et Kleven, Ø. (2010). Monitoring and changing data collection through R-indicators and partial R-indicators, document de travail du RISQ. Disponible à l'adresse www.risq-project.eu.

Schouten, J.G., Shlomo, N. et Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 2, 231-253.

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias, Thèse de doctorat, University of Michigan, États-Unis.

Zajonc, T. (2012). Bayesian inference for dynamic treatment regimes: Mobility, equity and efficiency in student tracking. *Journal of the American Statistical Association*, 107, 80-92.

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 67-103
Statistique Canada, N° 12-001-X au Catalogue

Vérification automatique en présence de vérifications avec rejet et de vérifications avec avertissement

Sander Scholtus[1]

Résumé

Une limite importante des méthodes actuelles de vérification automatique des données tient au fait que toutes les vérifications sont traitées comme des contraintes fermes. Autrement dit, un rejet à la vérification suppose systématiquement une erreur dans les données. Par contre, dans le cas de la vérification manuelle, les spécialistes du domaine recourent aussi à de très nombreuses vérifications avec avertissement, c'est-à-dire des contraintes destinées à repérer les valeurs et combinaisons de valeurs qui sont suspectes mais pas forcément incorrectes. L'incapacité des méthodes de vérification automatique à traiter les vérifications avec avertissement explique partiellement pourquoi, en pratique, de nombreuses différences sont constatées entre les données vérifiées manuellement et celles vérifiées automatiquement. L'objet du présent article est de présenter une nouvelle formulation du problème de localisation des erreurs qui permet de faire la distinction entre les vérifications avec rejet (*hard edits*) et les vérifications avec avertissement (*soft edits*). En outre, il montre comment ce problème peut être résolu grâce à une extension de l'algorithme de localisation des erreurs de De Waal et Quere (2003).

Mots clés : Localisation automatique des erreurs; paradigme de Fellegi-Holt; algorithme *branch-and-bound* ou algorithme de séparation et évaluation; données numériques; données catégoriques; données mixtes.

1 Introduction

La *vérification des données*, c'est-à-dire la détection et la correction des erreurs ainsi que des valeurs manquantes dans les données recueillies, est un élément important de tout processus statistique. Les instituts nationaux de statistique recourent habituellement à la vérification manuelle, durant laquelle des spécialistes du domaine vérifient les données et les corrigent au besoin. Malheureusement, cette approche est lente et coûteuse. D'autres méthodes ont par conséquent été élaborées afin de rendre le processus de vérification plus efficace, dont la *vérification sélective* et la *vérification automatique*. Le présent article porte sur cette dernière approche. Pour une discussion de la vérification sélective et d'autres formes de vérification

1. Sander Scholtus, Statistics Netherlands, P.O. Box 24500, 2490 HA La Haye, Pays-Bas. Courriel : sshs@cbs.nl.

statistique des données, consulter De Waal, Pannekoek et Scholtus (2011) et les références mentionnées dans leur article.

L'objectif de la vérification automatique est de déceler et de corriger exactement les erreurs et les valeurs manquantes dans un fichier de données de manière entièrement automatisée, c'est-à-dire sans intervention humaine. À condition qu'elle produise des données de qualité suffisante, la vérification automatique peut être utilisée pour remplacer partiellement la vérification manuelle. En pratique, la vérification automatique signifie que l'on oblige les données à satisfaire un ensemble de contraintes, appelées *vérifications* ou *contrôles*. En voici des exemples :

$$\text{Profit} = \text{Chiffre d'affaires total} - \text{Coût total}; \tag{1.1}$$

et

$$\text{Profit} \leq 0{,}6 \times \text{Chiffre d'affaires total}. \tag{1.2}$$

La plupart des méthodes de vérification automatique comprennent la résolution de deux problèmes distincts : d'abord le *problème de localisation des erreurs*, c'est-à-dire la détermination des variables dont les valeurs sont incorrectes ou manquantes, et ensuite le *problème d'imputation cohérente*, c'est-à-dire la détermination de nouvelles valeurs pour ces variables qui satisfont toutes les vérifications. Le présent article porte sur le problème de localisation des erreurs.

En ce qui concerne les deux exemples de vérification susmentionnés, il est intéressant de noter ce qui les distingue conceptuellement. La vérification (1.1) est un exemple de vérification qui doit être satisfaite par définition, de sorte que toute combinaison de valeurs qui est rejetée à cette vérification contient forcément une erreur. Les vérifications de ce type sont appelées *vérifications avec rejet*, *vérifications de dépistage d'erreur fatale* ou *vérifications logiques*. La vérification (1.2), quant à elle, est un exemple de vérification permettant de repérer des combinaisons de valeurs peu plausibles, mais pas forcément incorrectes. Dans cet exemple, les enregistrements

pour lesquels la valeur de *Profit* dépasse 60 % du *chiffre d'affaires total* sont considérés comme étant suspects. Toutefois, il est concevable qu'une telle combinaison de valeurs puisse être correcte à l'occasion. Les vérifications de ce type, qui ne détectent pas les erreurs avec certitude, sont appelées *vérifications avec avertissement* ou *vérifications d'interrogation*.

Une limite importante des algorithmes de vérification automatique existants tient au fait qu'ils traitent toutes les vérifications comme des vérifications avec rejet. Autrement dit, un rejet dans la vérification suppose systématiquement une erreur dans les données. Dans le cas de la vérification manuelle, par contre, des spécialistes du domaine procèdent aussi à de nombreuses vérifications avec avertissement. Durant la vérification automatique, ces vérifications avec avertissement ne sont pas utilisées du tout ou sont interprétées comme des vérifications avec rejet. Ces deux solutions sont insatisfaisantes : dans le premier cas, certaines erreurs pourraient être manquées durant la vérification automatique, tandis que dans le second, certaines valeurs correctes pourraient être désignées à tort comme étant incorrectes. En fait, l'incapacité des méthodes de vérification automatique à traiter les vérifications avec avertissement explique partiellement pourquoi, en pratique, de nombreuses différences sont constatées entre les données vérifiées manuellement et celles vérifiées automatiquement.

L'objet du présent article est d'offrir une nouvelle formulation du problème de localisation automatique des erreurs qui permet de faire la distinction entre les vérifications avec rejet et les vérifications avec avertissement. En outre, l'article montre comment l'algorithme de localisation des erreurs de SLICE – le progiciel de vérification automatique développé par Statistics Netherlands – peut être adapté afin de résoudre ce nouveau problème de localisation des erreurs.

La présentation de la suite de l'article est la suivante. La section 2 résume brièvement les méthodes existantes de résolution du problème de localisation des erreurs. Une distinction entre les vérifications avec rejet et les vérifications avec avertissement est introduite dans le problème de localisation des erreurs à la section 3. À la section 4, la théorie qui sous-tend l'algorithme de SLICE est étendue au cas où les vérifications ne doivent pas être toutes satisfaites. Sur la base de

ces résultats théoriques, un algorithme qui résout le problème de localisation des erreurs pour les vérifications avec rejet et les vérifications avec investissements est décrit à la section 5. À la section 6, le nouvel algorithme est illustré au moyen d'un petit exemple. La section 7 décrit les premières expériences de mise en œuvre pratique du nouvel algorithme. Enfin, certaines conclusions sont présentées à la section 8.

2 Contexte

2.1 Vérifications

Le problème discuté dans le présent article comporte, sous sa forme la plus générale, la détection de valeurs incorrectes ou manquantes dans un enregistrement contenant à la fois des variables catégoriques $(v_1, \ldots, v_m)$ et des variables numériques $(x_1, \ldots, x_p)$. Ces variables sont supposées satisfaire un ensemble de contraintes (vérifications), qui peuvent chacune s'écrire sous l'une des formes suivantes :

$$\begin{aligned} \psi^k : \ & \text{SI } (v_1, \ldots, v_m) \in F_1^k \times \cdots \times F_m^k \\ & \text{ALORS } (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k \geq 0\} \end{aligned} \tag{2.1}$$

ou

$$\begin{aligned} \psi^k : \ & \text{SI } (v_1, \ldots, v_m) \in F_1^k \times \cdots \times F_m^k \\ & \text{ALORS } (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k = 0\}. \end{aligned} \tag{2.2}$$

Dans ces expressions, F_j^k est un sous-ensemble de D_j, le domaine des valeurs observées pour la variable catégorique v_j, et a_{kj} et b_k sont des constantes numériques connues. L'indice k est utilisé pour numéroter les vérifications. Notons que D_j est censé contenir toutes les valeurs de v_j qui peuvent être observées en pratique, y compris les valeurs incorrectes. Pour simplifier, nous limitons le problème aux vérifications comprenant des contraintes numériques linéaires. Cette catégorie de vérifications s'avère suffisamment puissante pour la plupart des applications pratiques (voir De Waal 2005).

On dit qu'un enregistrement $(v_1^0, \ldots, v_m^0, x_1^0, \ldots, x_p^0)$ est *rejeté* à la vérification si la condition SI catégorique est vraie (c'est-à-dire si $v_j^0 \in F_j^k$ pour tout $j = 1, \ldots, m$), mais que la condition ALORS numérique est fausse (c'est-à-dire si $a_{k1}x_1 + \cdots + a_{kp}x_p + b_k < 0$ ou $a_{k1}x_1 + \cdots + a_{kp}x_p + b_k \neq 0$, selon la forme de la vérification). Sinon, nous disons que l'enregistrement *satisfait* à la vérification. Il convient de souligner que tout enregistrement pour lequel la condition SI catégorique est fausse satisfait toujours à la vérification, quelle que soit la situation pour la condition ALORS numérique. Un enregistrement qui satisfait à toutes les vérifications est dit *cohérent*.

On dit qu'une variable catégorique v_j *intervient* dans une vérification si, et uniquement si, $F_j^k \neq D_j$, puisque toute vérification pour laquelle $F_j^k = D_j$ donne lieu à un rejet ou à une acceptation indépendamment de la valeur de v_j. De même, nous disons qu'une variable numérique x_j intervient dans une vérification si, et uniquement si, $a_{kj} \neq 0$. Nous pouvons supposer que $F_j^k \neq \varnothing$, où $\varnothing$ désigne l'ensemble vide. Manifestement, une vérification dégénérée pour laquelle $F_j^k = \varnothing$ peut être écartée sans perte d'information, puisqu'elle ne donne jamais lieu à un rejet. Il en est de même de toute vérification pour laquelle une condition ALORS numérique est toujours vraie.

Deux cas particuliers importants de (2.1) et (2.2) sont les vérifications qui ne font intervenir que des variables catégoriques ou que des variables numériques. Une vérification purement catégorique prend la forme suivante :

$$\psi^k : \text{SI } (v_1, \ldots, v_m) \in F_1^k \times \cdots \times F_m^k \text{ ALORS } \varnothing. \tag{2.3}$$

La vérification (2.3) donne lieu à un rejet pour chaque enregistrement pour lequel la condition catégorique est vraie. Une vérification purement numérique peut s'écrire sous la forme :

$$\psi^k : (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k \geq 0\} \tag{2.4}$$

ou

$$\psi^k : (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k = 0\}. \tag{2.5}$$

Les vérifications (2.4) et (2.5) donnent lieu au rejet de chaque enregistrement pour lequel les conditions numériques sont fausses.

Les vérifications (1.1) et (1.2) susmentionnées sont des exemples de vérifications purement numériques. Voici un simple exemple de vérifications purement catégoriques :

SI (Âge, État matrimonial) $\in$ {« <16 »} $\times$ {« Marié »} ALORS $\varnothing$.

Cette vérification énonce que les personnes de moins de 16 ans ne peuvent pas être mariées. Enfin, voici un exemple de vérification mixte :

SI Âge $\in$ {« <12 »} ALORS Revenu = 0.

Selon cette vérification, les personnes de moins de 12 ans n'ont pas de revenu positif.

2.2 Le problème de localisation des erreurs

Pour un enregistrement donné $(v_1^0, \ldots, v_m^0, x_1^0, \ldots, x_p^0)$ et une série de vérifications, il est simple de vérifier quelles valeurs de l'enregistrement manquent et si certaines vérifications donnent lieu à un rejet. Par contre, sachant que certaines vérifications donnent lieu à un rejet, résoudre le problème de localisation des erreurs est une tâche nettement moins facile. D'une part, la majorité des vérifications font intervenir plus d'une variable, et d'autre part, la majorité des variables interviennent dans plus d'une vérification.

Afin de résoudre automatiquement le problème de localisation des erreurs, on doit adopter une stratégie formelle pour découvrir les valeurs incorrectes. La stratégie utilisée le plus fréquemment est fondée sur le paradigme de Fellegi et Holt (1976) voulant que l'on rende l'enregistrement cohérent en modifiant le plus petit nombre possible de valeurs originales. D'autres stratégies ont également été proposées. Par exemple, Little et Smith (1987) ont suggéré un critère fondé sur la détection des valeurs aberrantes (sans vérification), et Casado Valero, Del Castillo Cuervo-Arango, Mateo Ayerra et De Santos Ballesteros (1996) ont formulé la localisation des erreurs

sous la forme d'un problème de minimisation quadratique. Nous nous limiterons ici à examiner le paradigme de Fellegi-Holt, parce que son usage est fréquent en statistique officielle.

Le paradigme de Fellegi-Holt original est facile à généraliser afin de permettre une distinction entre les variables suspectes a priori et les variables moins suspectes. Pour cela, on associe un *poids de confiance* à chaque variable. Selon le paradigme de Fellegi-Holt généralisé, on doit rechercher un sous-ensemble de variables qui *i*) peut être imputé de manière que l'enregistrement imputé satisfasse à toutes les vérifications et *ii*) minimise la fonction cible suivante :

$$D_{\mathrm{FH}} = \sum_{j=1}^{m} w_j^C y_j^C + \sum_{j=1}^{p} w_j^N y_j^N . \tag{2.6}$$

Ici, w_j^C et w_j^N désignent les poids de confiance des variables catégoriques et numériques, respectivement. Les variables cibles binaires y_j^C et y_j^N décrivent la structure de la solution : $y_j^C = 1$ si v_j doit être imputée et $y_j^C = 0$ autrement, et de même $y_j^N = 1$ si x_j doit être imputée et $y_j^N = 0$ autrement. Puisque les variables pour lesquelles des valeurs manquent doivent être imputées avec certitude, nous fixons $y_j^C = 1$ ou $y_j^N = 1$ quand v_j^0 ou x_j^0 manque.

Fellegi et Holt (1976) ont également présenté une méthode pour résoudre le problème de localisation des erreurs sous ce paradigme. Cette méthode consiste à dériver d'abord un ensemble bien défini de vérifications logiques implicites à partir de l'ensemble original de vérifications, pour obtenir ce qu'ils appellent un *ensemble complet de vérifications*. Ensuite, le problème de localisation des erreurs peut être formulé pour tout enregistrement comme un simple problème de couverture d'ensemble (Fellegi et Holt 1976; Boskovitz, Goré et Wong 2005). Malheureusement, surtout pour les données numériques, l'ensemble complet de vérifications peut être très grand en pratique, de sorte que la méthode de Fellegi et Holt comporte parfois des calculs infaisables.

De nombreux algorithmes de rechange ont été élaborés pour localiser les erreurs conformément au paradigme de Fellegi-Holt. Outre les améliorations de la méthode originale de Fellegi et Holt (Garfinkel, Kunnathur et Liepins 1986; Winkler 1995), la liste comprend des

formules fondées sur la génération de sommets (Sande 1978; Kovar et Whitridge 1990; Todaro 1999; De Waal 2003a), sur des plans sécants (Garfinkel et coll. 1986; Garfinkel, Kunnathur et Liepins 1988; Ragsdale et McKeown 1996), et sur la programmation en nombres entiers mixtes (Schaffer 1987; Riera-Ledesma et Salazar-González 2003) et la programmation en nombres entiers (Bruni 2004 et 2005); voir aussi De Waal et coll. (2011) pour une vue d'ensemble. Ici, nous nous concentrons sur un algorithme *branch-and-bound* que l'on doit à De Waal et Quere (2003) et qui, contrairement à certaines approches susmentionnées, est capable de traiter un mélange de données catégoriques et numériques. Cet algorithme a été implémenté dans le progiciel SLICE à Statistics Netherlands et les calculs se sont avérés résolubles en pratique.

2.3 L'algorithme *branch-and-bound* de SLICE

L'algorithme de localisation des erreurs implémenté dans SLICE est décrit en détail dans De Waal et Quere (2003), De Waal (2003b) et De Waal et coll. (2011). Ici, nous mentionnons uniquement deux aspects de l'algorithme dont nous aurons besoin plus tard. Pour une introduction générale aux algorithmes *branch-and-bound*, voir par exemple Nemhauser et Wolsey (1988).

Pour chaque enregistrement, l'algorithme SLICE crée un arbre binaire tel que celui illustré à la figure 2.1. Dans le nœud racine de l'arbre, nous débutons avec l'ensemble original de vérifications et nous choisissons l'une des variables. Partant d'une racine, deux branches sont ajoutées à l'arbre. Sur la première branche, la valeur originale de la variable sélectionnée dans l'enregistrement est supposée correcte et sur la seconde branche, elle est supposée incorrecte. Chacune des hypothèses correspond à une transformation de l'ensemble de vérifications qui sera décrite plus bas, après laquelle la variable sélectionnée n'intervient plus dans les vérifications : la variable sélectionnée a été *traitée*. Ensuite, l'une des variables restantes est sélectionnée et l'opération est répétée.

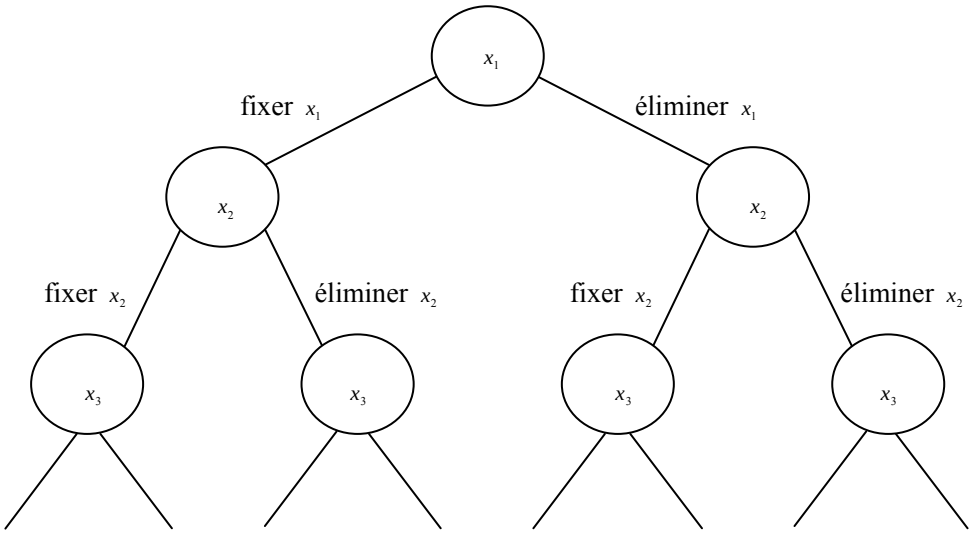


Figure 2.1 Illustration de l'algorithme *branch-and-bound* sous forme d'un arbre binaire

Lorsque toutes les variables ont été traitées, l'algorithme atteint un nœud final de l'arbre. On voit qu'ensemble, les nœuds finaux de l'arbre binaire énumèrent tous les choix possibles de sous-ensembles erronés de variables. L'ensemble de vérifications transformé correspondant à un nœud final ne fait intervenir aucune des variables, de sorte qu'il doit être vide ou être constitué de relations élémentaires telles que « $1 \geq 0$ » (une tautologie) et « $-1 \geq 0$ » (une déclaration autocontradictoire). Comme nous en discuterons plus bas, il est possible de satisfaire aux vérifications originales en imputant uniquement les variables qui ont été considérées comme étant incorrectes sur la branche menant à un nœud final, si et uniquement si l'ensemble de vérifications transformé pour ce nœud final ne contient aucune déclaration autocontradictoire. En utilisant cette propriété, on peut identifier toutes les solutions possibles du problème de localisation des erreurs. En outre, comme nous nous intéressons seulement aux solutions possibles qui minimisent la fonction cible (2.6), nous pouvons élaguer une branche de l'arbre aussitôt que nous constatons qu'elle mène uniquement à des nœuds finaux correspondant à des solutions impossibles ou sous-optimales.

Nous allons maintenant décrire les transformations de l'ensemble de vérifications qui ont lieu, selon que la valeur d'une variable est supposée correcte ou incorrecte. Une variable dont la valeur est supposée correcte est enlevée des vérifications en la remplaçant simplement par la valeur originale provenant de l'enregistrement dans les vérifications. On dit alors que la variable est *fixée* à sa valeur originale. Une variable dont la valeur est supposée incorrecte est éliminée des vérifications par une opération plus complexe, appelée *élimination*. Les variables numériques et les variables catégoriques sont éliminées par deux méthodes différentes, mais équivalentes.

Pour éliminer une variable numérique, disons x_g, d'un ensemble de vérifications ayant les formes générales (2.1) et (2.2), nous générons des vérifications logiques implicites en considérant toutes les paires de vérifications ψ^s et ψ^t dans lesquelles intervient x_g. Nous commençons par vérifier si $F_j^s \cap F_j^t \neq \emptyset$ pour tout $j = 1, \ldots, m$; si l'une de ces intersections donne l'ensemble vide, alors la paire ψ^s et ψ^t ne génère pas de vérification implicite. Si la condition ALORS numérique de l'une des vérifications, disons ψ^s, est une égalité, cette dernière peut être résolue pour trouver x_g. En remplaçant x_g par l'expression résultante dans la condition ALORS de ψ^t, nous obtenons la condition ALORS numérique de la vérification implicite. La condition SI catégorique de la vérification implicite s'obtient en prenant les intersections non vides $F_j^* = F_j^s \cap F_j^t$ pour $j = 1, \ldots, m$.

Si les conditions ALORS numériques de ψ^s et ψ^t sont toutes deux des inégalités, l'algorithme suit une technique appelée *élimination de Fourier-Motzkin* (voir par exemple Williams 1986) pour générer une vérification implicite. Une paire de vérifications n'est pertinente pour cette méthode d'élimination que si les coefficients de x_g sont de signes opposés, de sorte que nous pouvons supposer sans perte de généralité que $a_{sg} < 0$ et $a_{tg} > 0$. La vérification implicite générée à partir de ψ^s et ψ^t peut alors s'écrire sous la forme (voir De Waal et Quere 2003) :

$$\begin{aligned} \psi^* : \text{SI } (v_1, \ldots, v_m) &\in F_1^* \times \cdots \times F_m^* \\ \text{ALORS } (x_1, \ldots, x_p) &\in \{a_1^* x_1 + \cdots + a_p^* x_p + b^* \geq 0\} \end{aligned} \tag{2.7}$$

avec $a_j^* = a_{tg}a_{sj} - a_{sg}a_{tj}, b^* = a_{tg}b_s - a_{sg}b_t$, et $F_j^* = F_j^s \cap F_j^t$ comme plus haut. Cette vérification ne fait pas intervenir x_g, puisque $a_g^* = 0$. De cette façon, des vérifications implicites sont générées en considérant toutes les paires de vérifications qui font intervenir x_g. Ces vérifications sont ajoutées à l'ensemble de vérifications original ne faisant pas intervenir x_g pour trouver l'ensemble transformé de vérifications obtenu en éliminant x_g.

Pour l'élimination des variables catégoriques, De Waal et Quere (2003) émettent pour simplifier l'hypothèse que ces variables ne sont sélectionnées que si toutes les variables numériques ont été traitées. Cette hypothèse implique que les variables catégoriques sont toujours éliminées des vérifications purement catégoriques de la forme (2.3). Pour éliminer une variable catégorique, disons v_g, d'un ensemble de vérifications de la forme (2.3), nous utilisons une technique qui a été décrite pour la première fois par Fellegi et Holt (1976).

Considérons tous les ensembles minimaux de vérifications T ayant les propriétés suivantes :

$$F_g^*(T) = \bigcup_{k \in T} F_g^k = D_g \tag{2.8}$$

et

$$F_j^*(T) = \bigcap_{k \in T} F_j^k \neq \emptyset, \text{ pour } j = 1, \ldots, g-1, g+1, \ldots, m. \tag{2.9}$$

Ici, par « minimal » nous entendons que la propriété (2.8) n'est pas vérifiée pour tout ensemble $T' \subset T$. Chacun de ces ensembles minimaux T génère une vérification implicite :

$$\psi^* : \text{SI } (v_1, \ldots, v_m) \in F_1^*(T) \times \cdots \times F_m^*(T) \text{ ALORS } \emptyset, \tag{2.10}$$

qui ne fait pas intervenir v_g à cause de la propriété (2.8). Ces vérifications implicites sont ajoutées à l'ensemble de vérifications original ne faisant pas intervenir v_g pour trouver l'ensemble transformé de vérifications obtenu en éliminant v_g.

Il est clair que le travail de calcul de l'algorithme correspond principalement à l'étape d'élimination. En particulier, on sait que, sous élimination de Fourier-Motzkin, le nombre de

vérifications implicites peut être exponentiel en le nombre de variables éliminées (Schrijver 1986).

Le résultat qui suit illustre une propriété fondamentale des deux techniques d'élimination, pour les variables numériques et catégoriques. Considérons un système de vérifications implicites Ψ_1 obtenus en éliminant x_g ou v_g d'un système de vérifications Ψ_0. Alors, les valeurs originales des variables non traitées satisfont à toutes les vérifications dans Ψ_1 si, et uniquement si, il existe une valeur de x_g ou de v_g qui, avec ces valeurs originales, satisfait à toutes les vérifications dans Ψ_0. Pour une preuve, voir le théorème 8.1 dans De Waal (2003b) ou le théorème 4.3 dans De Waal et coll. (2011). La correspondance susmentionnée entre les nœuds finaux sans relations élémentaires autocontradictoires et les solutions possibles du problème de localisation des erreurs découle de l'application répétée de cette propriété fondamentale.

3 Un problème de localisation des erreurs comprenant des vérifications avec rejet et avec avertissement

Dans la formulation du problème de localisation des erreurs donnée à la section 2.2, qui est fondée sur le paradigme de Fellegi-Holt, il est supposé tacitement que toutes les vérifications sont de type avec rejet. Par conséquent, les seuls sous-ensembles de variables considérés comme des solutions possibles de ce problème sont ceux qui peuvent être imputés afin de rendre l'enregistrement cohérent pour toutes les vérifications. Comme il est mentionné dans l'introduction, le fait d'interpréter toutes les vérifications comme étant des vérifications avec rejet peut donner lieu à des différences systématiques entre la vérification automatique et la vérification manuelle, parce que cela exclut l'usage judicieux de vérifications avec avertissement. Dans la présente section, nous proposons une nouvelle formulation du problème de localisation des erreurs qui fait la distinction entre les vérifications avec rejet (dans la notation, H pour *Hard edits*) et les vérifications avec avertissement (S pour *Soft edits*).

Soit Ψ l'ensemble de vérifications à utiliser dans le problème de localisation des erreurs. Nous supposons que cet ensemble peut être partitionné en deux sous-ensembles disjoints : $\Psi = \Psi_H \cup \Psi_S$. Les vérifications comprises dans Ψ_H sont des vérifications avec rejet; celles comprises dans Ψ_S sont des vérifications avec avertissement. À partir de maintenant, un sous-ensemble des variables est considéré comme une solution possible du problème de localisation des erreurs s'il peut être imputé de manière à satisfaire à toutes les vérifications comprises dans Ψ_H. En outre, nous voulons utiliser l'état de l'enregistrement imputé en regard des vérifications comprises dans Ψ_S comme information auxiliaire dans le choix d'une solution optimale. Nous pouvons pour cela ajouter un autre terme à (2.6).

Plus précisément, l'objectif du nouveau problème de localisation des erreurs consiste à trouver un sous-ensemble des variables qui *i*) peut être imputé de manière que l'enregistrement imputé satisfasse à toutes les vérifications comprises dans Ψ_H, et *ii*) minimise la fonction cible suivante :

$$D = \lambda D_{\mathrm{FH}} + (1 - \lambda) D_{\mathrm{soft}}, \tag{3.1}$$

où D_{soft} représente le coût associé aux vérifications non satisfaites dans Ψ_S. Le paramètre $\lambda \in [0, 1]$ détermine la contribution relative des deux termes dans (3.1). On retrouve le paradigme de Fellegi-Holt original en tant que cas particulier en choisissant $\lambda = 1$. Donc, le nouveau problème de localisation des erreurs peut être considéré comme une généralisation de l'ancien.

Afin d'utiliser (3.1) en pratique, on doit choisir une expression pour D_{soft}. Le moyen sans doute le plus facile d'attribuer des coûts aux vérifications avec avertissement non satisfaites consiste à associer un *poids de non-satisfaction* fixe s_k à chaque vérification comprise dans Ψ_S, et à définir D_{soft} comme étant la somme des poids de non-satisfaction des vérifications avec avertissement qui demeurent non satisfaites :

$$D_{\text{soft}} = \sum_{k=1}^{K_S} s_k z_k , \tag{3.2}$$

avec K_S désignant le nombre de vérifications dans Ψ_S et z_k, une variable binaire telle que $z_k = 1$ si la k^{e} vérification avec avertissement n'est pas satisfaite et $z_k = 0$ autrement. Les poids de non-satisfaction peuvent être choisis par les spécialistes du domaine, de la même façon que les poids de confiance, pour exprimer l'importance qui est accordée aux diverses vérifications avec avertissement dans la perspective du domaine spécialisé. Autrement, les poids de non-satisfaction peuvent être fondés, pour chaque vérification avec avertissement, sur la proportion d'enregistrements non satisfaisants dans un ensemble de données historiques qui a été vérifié manuellement.

Un inconvénient de l'utilisation de poids de non-satisfaction fixes est qu'ils ne tiennent pas compte de la mesure dans laquelle la vérification n'est pas satisfaite : chaque enregistrement qui ne satisfait pas à une vérification avec avertissement particulière se voit attribuer la même contribution à D_{soft}, à savoir s_k. Par contre, dans le cas de la vérification manuelle, le vérificateur considère la non-satisfaction d'une vérification avec avertissement comme une indication qu'une combinaison observée de valeurs est suspecte, et le degré de suspicion dépend de la mesure dans laquelle la vérification n'est pas satisfaite : un petit écart sera ignoré plus facilement qu'un grand. Donc, il semble intéressant de tenir compte dans D_{soft} de la mesure dans laquelle les vérifications ne sont pas satisfaites. Cet élément sera examiné à la section 8, puisqu'il entraîne certaines difficultés supplémentaires. Pour le moment, nous supposons que l'on utilise l'expression (3.2).

Il convient de mentionner que tenir compte des contraintes souples, ou molles, en ajoutant un terme approprié à la fonction cible est une méthode bien connue en optimisation mathématique. L'idée est apparentée à d'autres techniques d'optimisation, telles que la relaxation lagrangienne (voir par exemple, Nemhauser et Wolsey 1988). Ledit problème d'étalonnage des comptes nationaux est un exemple d'applications pratiques de contraintes souples (Magnus, Van Tongeren

et De Vos 2000). Autant que nous sachions, l'application dans le contexte du problème de localisation des erreurs est nouvelle.

Il convient également de souligner que l'expression (3.1) est, à certains égards, semblable au critère de minimisation de la méthode d'imputation par le plus proche voisin (NIM) élaborée par Statistique Canada pour la vérification des données démographiques du recensement (Bankier, Lachance et Poirier 2000; Bankier et Crowe 2009). En particulier, la NIM s'écarte aussi du paradigme de Fellegi-Holt en minimisant une combinaison convexe de deux termes, le premier mesurant la quantité d'imputation et le second, la plausibilité de l'enregistrement imputé.

4 Une brève théorie des rejets à la vérification

4.1 Données numériques

Maintenant que nous avons formulé un nouveau problème de localisation des erreurs, nous allons montrer comment il peut être résolu en se servant d'une version adaptée de l'algorithme *branch-and-bound* de De Waal et Quere (2003). Pour cela, nous devons d'abord étendre la propriété fondamentale mentionnée à la fin de la section 2.3 au cas où certaines vérifications donnent lieu à un rejet. Pour simplifier, nous allons examiner le cas de données purement numériques. À la sous-section suivante, nous examinons le cas des données purement catégoriques et des données mixtes.

Dans le cas de données purement numériques, toutes les vérifications prennent la forme (2.4) ou (2.5). En outre, la vérification implicite (2.7) se réduit à sa partie numérique. La propriété fondamentale donnée à la fin de la section 2.3 implique en particulier ce qui suit : si un ensemble donné de valeurs pour $x_1, \ldots, x_{g-1}, x_{g+1}, \ldots, x_p$ ne satisfait pas à la vérification implicite (2.7), il est impossible de trouver pour x_g une valeur qui satisfait ψ^s et ψ^t simultanément. Cependant, il est encore possible dans ce cas de trouver une valeur de x_g qui satisfait à l'une des vérifications

ψ^s ou ψ^t. Cette observation, qui est plus ou moins triviale, sert de fondement à la preuve du théorème 1 présenté plus loin.

Supposons qu'à un moment donné durant une exécution de l'algorithme *branch-and-bound* de De Waal et Quere (2003), q variables numériques ont été traitées (c'est-à-dire éliminées ou fixées). Nous désignons l'ensemble courant de vérifications par Ψ_q, et les vérifications qu'il contient par ψ_q^k. Par définition, $\Psi_0 \equiv \Psi$, l'ensemble original de vérifications. Il est possible d'associer à chaque vérification existante ψ_q^k un ensemble d'indices B_q^k, qui contient les indices de toutes les vérifications originales qui ont été utilisées, directement ou indirectement, pour dériver cette vérification. En fait, B_q^k peut être défini de manière récursive comme il suit :

- Pour une vérification originale ψ_0^k, nous définissons $B_0^k := \{k\}$.
- Pour une vérification ψ_q^k qui est dérivée d'une autre vérification ψ_{q-1}^l, en fixant une variable à sa valeur originale ou en copiant simplement la valeur de vérification, nous définissons $B_q^k := B_{q-1}^l$.
- Pour une vérification ψ_q^k qui est dérivée en éliminant une variable d'un ensemble de vérifications ψ_{q-1}^t $(t \in T)$, nous définissons $B_q^k := \bigcup_{t \in T} B_{q-1}^t$.

Notons que, pour des données numériques, l'ensemble T dans le dernier élément contient toujours exactement deux vérifications. De plus grands ensembles de vérifications peuvent être observés dans le cas des données catégoriques considérées plus loin.

Un ensemble B est appelé *ensemble représentatif* d'un jeu d'ensembles $B_q^{k_1}, \ldots, B_q^{k_r}$ s'il contient au moins un élément de chacun des ensembles $B_q^{k_1}, \ldots, B_q^{k_r}$; voir, par exemple, Mirsky (1971, page 25). Il convient de mentionner que, dans notre cas, un ensemble représentatif B identifie un sous-ensemble de Ψ_0, l'ensemble de vérifications original. Nous pouvons maintenant formuler le théorème qui suit.

Théorème 1. Supposons que q variables numériques ont été traitées et que l'ensemble existant de vérifications numériques peut être partitionné comme $\Psi_q = \Psi_q^{(1)} \cup \Psi_q^{(2)}$, où les vérifications comprises dans $\Psi_q^{(1)}$ sont satisfaites par les valeurs originales des $p - q$ variables restantes, et

les vérifications comprises dans $\Psi_q^{(2)}$ ne sont pas satisfaites. Soit B un ensemble représentatif des ensembles d'indices B_q^k pour tout $\psi_q^k \in \Psi_q^{(2)}$. Alors, il existe des valeurs pour les variables éliminées qui, de concert avec les valeurs originales des autres variables, satisfont à toutes les vérifications originales sauf celles comprises dans B.

Preuve. La preuve de ce théorème est donnée à l'annexe A.1.

Exemple : Supposons qu'il existe trois variables numériques (x_1, x_2, x_3) qui doivent satisfaire aux huit vérifications suivantes :

$$\begin{aligned}
\psi_0^1 &: x_1 + x_2 + x_3 = 20 \\
\psi_0^2 &: x_1 - x_2 \geq 3 \\
\psi_0^3 &: -x_1 + x_2 \geq -6 \\
\psi_0^4 &: -x_1 + x_3 \geq 5 \\
\psi_0^5 &: x_1 - x_3 \geq -10 \\
\psi_0^6 &: x_1 \geq 0 \\
\psi_0^7 &: x_2 \geq 0 \\
\psi_0^8 &: x_3 \geq 0.
\end{aligned}$$

L'enregistrement $(x_1^0, x_2^0, x_3^0) = (10, 1, -3)$ est incohérent en ce qui concerne ces vérifications. En éliminant x_1 de l'ensemble original de vérifications, nous trouvons l'ensemble de vérifications mis à jour suivant :

$$\begin{aligned}
\psi_1^1 &: -2x_2 - x_3 \geq -17 \quad (B_1^1 = \{1, 2\}) \\
\psi_1^2 &: 2x_2 + x_3 \geq 14 \quad (B_1^2 = \{1, 3\}) \\
\psi_1^3 &: x_2 + 2x_3 \geq 25 \quad (B_1^3 = \{1, 4\}) \\
\psi_1^4 &: -x_2 - 2x_3 \geq -30 \quad (B_1^4 = \{1, 5\}) \\
\psi_1^5 &: -x_2 - x_3 \geq -20 \quad (B_1^5 = \{1, 6\}) \\
\psi_1^6 &: x_2 \geq 0 \quad (B_1^6 = \{7\}) \\
\psi_1^7 &: x_3 \geq 0 \quad (B_1^7 = \{8\}) \\
\psi_1^8 &: 0 \geq -3 \quad (B_1^8 = \{2, 3\}) \\
\psi_1^9 &: -x_2 + x_3 \geq 8 \quad (B_1^9 = \{2, 4\}) \\
\psi_1^{10} &: x_2 - x_3 \geq -16 \quad (B_1^{10} = \{3, 5\}) \\
\psi_1^{11} &: 0 \geq -5 \quad (B_1^{11} = \{4, 5\}) \\
\psi_1^{12} &: x_2 \geq -6 \quad (B_1^{12} = \{3, 6\}) \\
\psi_1^{13} &: x_3 \geq 5 \quad (B_1^{13} = \{4, 6\}).
\end{aligned}$$

L'ensemble d'indices B_1^k est inscrit entre parenthèses à côté de chaque vérification.

En remplaçant x_2 et x_3 par leur valeur originale dans l'ensemble courant de vérifications, nous voyons que $\psi_1^2, \psi_1^3, \psi_1^7, \psi_1^9,$ et ψ_1^{13} ne sont pas satisfaites. L'ensemble $B = \{1, 4, 8\}$ est un ensemble représentatif des ensembles d'indices associés B_1^k. Selon le théorème 1, il existe une valeur de x_1 qui, de concert avec les valeurs originales de x_2 et x_3, satisfait aux vérifications originales excepté ψ_0^1, ψ_0^4 et ψ_0^8. On peut voir que cette affirmation est correcte en substituant $x_2^0 = 1$ et $x_3^0 = -3$ dans l'ensemble original de vérifications; en fait, toute valeur $x_1 \in [4, 7]$ fait l'affaire.

L'importance du théorème 1 tient au fait qu'il permet de déterminer, à chaque nœud de l'algorithme *branch-and-bound*, les combinaisons des vérifications originales qui pourraient être satisfaites en imputant les variables qui ont été éliminées jusque-là, ainsi que les vérifications qui demeureraient non satisfaites. En particulier, si nous faisons la distinction entre les vérifications originales avec rejet et avec avertissement, ce résultat permet d'utiliser l'algorithme *branch-and-bound* pour trouver toutes les solutions possibles du nouveau problème de localisation des erreurs données à la section 3, et de déterminer pour chaque solution possible, lesquelles des vérifications avec avertissement demeurent non satisfaites, donc de déterminer la valeur de D_{soft}. Cette idée sera exposée plus en détail à la section 5.

4.2 Données catégoriques et mixtes

Nous allons maintenant établir un résultat semblable au théorème 1 pour le cas des données purement catégoriques. À la fin de la présente section, nous combinerons les deux résultats pour qu'ils puissent également être appliqués à des données mixtes.

Dans le cas de données purement catégoriques, toutes les vérifications prennent la forme (2.3). Considérons la méthode d'élimination décrite pour les variables catégoriques à la section 2.3. Si un ensemble donné de valeurs pour $v_1, \ldots, v_{g-1}, v_{g+1}, \ldots, v_m$ ne satisfait pas à la vérification implicite (2.10), il est impossible de trouver une valeur de v_g qui, de concert avec les autres

valeurs, satisfait à toutes les vérifications comprises dans ψ^k avec $k \in T$ simultanément. Il en est ainsi parce que, en vertu de la propriété (2.9), $F_j^*(T) \subseteq F_j^k$ pour tout $j \neq g$ et tout $k \in T$. Donc, si (2.10) n'est pas satisfaite par $v_1, \ldots, v_{g-1}, v_{g+1}, \ldots, v_m$, alors l'insertion de ces valeurs dans une vérification originale avec $k \in T$ produit une vérification univariée non dégénérée pour v_g. En outre, chaque valeur possible de v_g ne satisfait pas au moins une de ces vérifications univariées, en raison de la propriété (2.8). Fait intéressant, il demeure possible dans ce cas de toujours trouver une valeur de v_g qui satisfait toutes les vérifications dans T sauf une. Cela découle de la propriété (2.8) et du fait que T est un ensemble minimal ayant cette propriété : pour chaque $k \in T$, F_g^k doit contenir au moins une valeur provenant de D_g qui n'est couverte par aucun autre F_g^l avec $l \in T$.

Nous présentons maintenant l'analogue du théorème 1 pour les données catégoriques, en utilisant la même notation que pour les données numériques. En particulier, la définition récursive de B_q^k est exactement la même qu'à la section 4.1.

Théorème 2. Supposons que q variables catégoriques ont été traitées et que l'ensemble courant de vérifications catégoriques peut être partitionné comme $\Psi_q = \Psi_q^{(1)} \cup \Psi_q^{(2)}$, où les vérifications comprises dans $\Psi_q^{(1)}$ sont satisfaites par les valeurs originales des $m - q$ variables restantes, et les vérifications comprises dans $\Psi_q^{(2)}$ ne sont pas satisfaites. Soit B un ensemble représentatif des ensembles d'indices B_q^k pour tout $\psi_q^k \in \Psi_q^{(2)}$. Alors, il existe des valeurs pour les variables éliminées qui, de concert avec les valeurs originales des autres variables, satisfont à toutes les vérifications originales, sauf celles comprises dans B.

Preuve. La preuve de ce théorème est donnée à l'annexe A.2.

Pour un exemple illustrant l'utilisation de ce théorème, voir Scholtus (2011).

Enfin, nous remarquons que les théorèmes 1 et 2 peuvent être utilisés ensemble quand les données comprennent un mélange de variables catégoriques et numériques. Cela découle de la structure de l'algorithme *branch-and-bound* de De Waal et Quere (2003), où les variables

catégoriques ne sont traitées qu'après avoir éliminé toutes les variables numériques ou fixé leur valeur. D'où, les deux résultats peuvent être appliqués consécutivement. La procédure d'élimination des variables numériques présente une légère différence, à savoir que les vérifications implicites ne sont générées qu'en partant de paires de vérifications présentant une condition SI chevauchante; voir la section 2.3. Toutefois, cela n'a pas d'incidence sur la justesse du théorème 1.

5 Résolution du problème de localisation des erreurs en présence de vérifications avec rejet et avec avertissement

Nous allons maintenant décrire une version adaptée de l'algorithme *branch-and-bound* de De Waal et Quere qui peut être utilisée pour résoudre le problème de localisation des erreurs défini à la section 3. La configuration fondamentale de l'algorithme est la même qu'à la section 2.3. En particulier, les procédures suivies pour éliminer les variables ou pour fixer leurs valeurs sont exécutées de la même façon que dans l'algorithme original.

La principale différence tient au fait que, maintenant, dans chaque nœud, l'ensemble courant de vérification Ψ_q est partitionné en un ensemble courant de vérifications avec rejet Ψ_{qH} et d'un ensemble courant de vérifications avec avertissement Ψ_{qS}. Pour le nœud racine, la partition suit simplement celle de l'ensemble original de vérifications, c'est-à-dire $\Psi_{0H} = \Psi_H$ et $\Psi_{0S} = \Psi_S$. Pour tous les autres nœuds, la partition peut se résumer comme il suit : si une vérification est générée uniquement à partir de vérifications avec rejet, il s'agit alors d'une vérification avec rejet; si une vérification avec avertissement intervient dans sa création, il s'agit alors d'une vérification avec avertissement. En outre, pour chaque vérification avec avertissement $\psi_{qS}^k \in \Psi_{qS}$, nous construisons un ensemble d'indices B_{qS}^k – analogue à B_q^k à la section 4 – qui contient les indices pour toutes les vérifications *avec avertissement* originales ψ_{0S}^k qui sont intervenues, directement ou indirectement, dans sa création.

Après avoir généré Ψ_{qH} et Ψ_{qS} pour un nœud particulier, nous pouvons introduire les valeurs originales des variables qui n'ont pas encore été traitées, pour déterminer lesquelles de ces vérifications n'ont pas été satisfaites. Dans l'ancien algorithme, cette étape pouvait avoir deux résultats possibles : soit un plus grand nombre de variables devait être éliminé (au moins l'une des vérifications n'était pas satisfaite), soit une solution possible était trouvée (toutes les vérifications étaient satisfaites). Dans le nouvel algorithme, trois situations peuvent se produire.

Avant tout, si au moins une vérification dans Ψ_{qH} n'est pas satisfaite, les variables qui ont été éliminées jusque-là ne peuvent pas être imputées pour satisfaire les vérifications avec rejet originales. Donc, d'autres variables doivent être éliminées. Dans ce cas, nous continuons à générer des branches à partir du nœud courant.

Une deuxième possibilité est qu'aucune des vérifications dans Ψ_{qH} et Ψ_{qS} ne donne lieu à un rejet ou à un avertissement. Cela signifie que les variables qui ont été éliminées jusque-là peuvent être imputées pour satisfaire à toutes les vérifications originales, tant avec rejet qu'avec avertissement. Donc, une solution possible a été trouvée, pour laquelle la valeur de la fonction cible (3.1) est égale à $D = \lambda D_{\mathrm{FH}}$. Si cette valeur est inférieure ou égale à la valeur de (3.1) pour la meilleure solution trouvée jusque-là, disons $D_{\min}$, la nouvelle solution est mémorisée. Sinon, elle est écartée. De l'une ou l'autre façon, il n'est pas utile de poursuivre l'exécution de l'algorithme à partir du nœud courant, car si d'autres variables sont éliminées, la valeur de D ne peut qu'augmenter. Donc, nous retournons à la dernière branche antérieure qui n'a pas encore fait l'objet d'une recherche exhaustive et nous poursuivons l'exécution de l'algorithme à partir de là.

La dernière possibilité est que les vérifications comprises dans Ψ_{qH} sont satisfaites, mais qu'au moins une vérification comprise dans Ψ_{qS} ne l'est pas. Dans ce cas, les variables qui ont été éliminées jusque-là peuvent être imputées pour satisfaire les vérifications avec rejet originales, mais pas toutes les vérifications avec avertissement originales. Par conséquent, une solution possible du problème de localisation des erreurs a été trouvée, mais la contribution de D_{soft} à D n'est pas nulle. Selon le théorème 1 ou le théorème 2, il est possible de satisfaire à

toutes les vérifications avec avertissement originales, sauf celles contenues dans un ensemble représentatif B des ensembles d'indices B_{qS}^{k} pour toutes les vérifications non satisfaites dans Ψ_{qS}. Puisque cette propriété est commune à tous les ensembles représentatifs, nous sommes libres de choisir B de manière que D_{soft} soit minimisé, sachant le choix des variables à imputer. Si l'expression (3.2) est utilisée pour D_{soft}, alors le choix optimal de B s'obtient en résolvant le problème de minimisation suivant :

$$\begin{aligned} &\min \sum_{k=1}^{K_S} s_k z_k, \text{ sous la contrainte :} \\ &\sum_{k \in B_{qS}^{l}} z_k \geq 1, \text{ pour toutes les } \psi_{qS}^{l} \in \Psi_{qS} \text{ non satisfaites,} \\ &z_k \in \{0, 1\}, k = 1, \ldots, K_S. \end{aligned} \tag{5.1}$$

Il s'agit d'un problème d'optimisation linéaire binaire classique pour lequel des algorithmes existent (voir, par exemple, Nemhauser et Wolsey 1988). La solution consiste en un vecteur $(z_1^*, \ldots, z_{K_S}^*)$ de zéros et de uns. L'ensemble représentatif optimal associé est $B^* = \{k : z_k^* = 1\}$ et la contribution associée de D_{soft} à D est précisément la valeur minimale du problème (5.1), disons

$$D_{\text{soft}}^* = \sum_{k=1}^{K_S} s_k z_k^* = \sum_{k \in B^*} s_k.$$

Comme dans le cas précédent, la valeur $D = \lambda D_{\text{FH}} + (1 - \lambda) D_{\text{soft}}^*$ est comparée à $D_{\min}$. Si $D \leq D_{\min}$, la solution courante est mémorisée; sinon, elle est écartée. De l'une ou l'autre façon, il est utile dans ce cas de poursuivre l'exécution de l'algorithme à partir du nœud courant, parce que l'élimination d'un plus grand nombre de variables peut mener à une valeur plus faible de la fonction cible. Cela peut se produire parce qu'une solution comprenant l'imputation d'un plus grand nombre de variables aboutit habituellement à la non-satisfaction d'un plus petit nombre de vérifications avec avertissement. Par conséquent, nous continuons à générer des branches à partir du nœud courant.

La justesse de cet algorithme découle de la justesse de l'algorithme original de De Waal et Quere (2003) et de la théorie présentée à la section 4. Les ensembles d'indices B_q^k ne doivent être calculés que pour les vérifications avec avertissement, parce qu'un sous-ensemble des variables n'est jamais considéré comme une solution possible du problème de localisation des erreurs quant au moins l'une des vérifications avec rejet demeure non satisfaite. Cela signifie que, dans chaque application du théorème 1 ou du théorème 2, toutes les vérifications implicites contenues dans Ψ_{qH} doivent être contenues dans $\Psi_q^{(1)}$. Enfin, nous notons que le nouvel algorithme se réduit à l'algorithme original de De Waal et Quere (2003) dans le cas particulier où aucune vérification avec avertissement n'a été spécifiée.

6 Exemple

Afin d'illustrer l'algorithme de la section 5, nous allons l'appliquer à un petit exemple comprenant des données numériques. Il s'agit essentiellement d'un exemple tiré de De Waal (2003b) auquel nous avons ajouté une distinction entre les vérifications avec rejet et les vérifications avec avertissement. Pour un exemple un peu plus important comprenant un mélange de variables catégoriques et numériques, voir Scholtus (2011).

Une enquête-entreprise fictive comporte quatre variables numériques : *chiffre d'affaires total* (T), *profit* (P), *coût total* (C), et *nombre d'employés* (N). Les vérifications avec rejet et les vérifications avec avertissement qui suivent sont définies :

$$
\begin{aligned}
\psi_{0H}^1 &: \quad T - C - P = 0 \\
\psi_{0H}^2 &: \quad T \geq 0 \\
\psi_{0H}^3 &: \quad C \geq 0 \\
\psi_{0H}^4 &: \quad N \geq 0 \\
\psi_{0H}^5 &: \quad 550N - T \geq 0 \\
\psi_{0S}^1 &: \quad 0,5T - P \geq 0 \quad (B_{0S}^1 = \{1\}) \\
\psi_{0S}^2 &: \quad P + 0,1T \geq 0 \quad (B_{0S}^2 = \{2\}).
\end{aligned}
$$

Considérons l'enregistrement non vérifié suivant : $(T^0, P^0, C^0, N^0) = (100, 40\,000, 60\,000, 5)$. Cet enregistrement est rejeté à la première vérification avec rejet et ne satisfait pas la première vérification avec avertissement. Les poids de confiance des variables sont $(w_T, w_P, w_C, w_N) = (2, 1, 1, 3)$. Nous choisissons les poids de non-satisfaction des deux vérifications avec avertissement comme étant $s_1 = s_2 = 2$. Enfin, nous choisissons $\lambda = 1/2$ dans l'expression (3.1).

Supposons que l'on sélectionne la variable P pour commencer. Sur la branche où P est éliminée des vérifications originales, nous obtenons le nouvel ensemble de vérifications suivant :

$$
\begin{array}{lrll}
\psi^1_{1H}: & T \geq 0 & & (\psi^2_{0H}) \\
\psi^2_{1H}: & C \geq 0 & & (\psi^3_{0H}) \\
\psi^3_{1H}: & N \geq 0 & & (\psi^4_{0H}) \\
\psi^4_{1H}: & 550N - T \geq 0 & & (\psi^5_{0H}) \\
\psi^1_{1S}: & -0,5T + C \geq 0 & (B^1_{1S} = \{1\}) & (\psi^1_{0H}, \psi^1_{0S}) \\
\psi^2_{1S}: & 1,1T - C \geq 0 & (B^2_{1S} = \{2\}) & (\psi^1_{0H}, \psi^2_{0S}) \\
\psi^3_{1S}: & 0,6T \geq 0 & (B^3_{1S} = \{1, 2\}) & (\psi^1_{0S}, \psi^2_{0S}).
\end{array}
$$

Nous avons indiqué entre parenthèses de quelle vérification antérieure chaque nouvelle vérification est dérivée. La troisième vérification avec avertissement ψ^3_{1S} est en fait équivalente à la première vérification avec rejet ψ^1_{1H}, ce qui signifie qu'elle peut être écartée.

Lorsque l'on substitue les valeurs originales $(T^0, C^0, N^0) = (100, 60\,000, 5)$ dans les vérifications courantes, on voit que toutes les vérifications sont satisfaites sauf ψ^2_{1S}. Puisque toutes les vérifications avec rejet sont satisfaites, identifier seulement la valeur originale de P comme étant incorrecte est une solution possible du problème de localisation des erreurs. En outre, puisque $B = \{2\}$ est (trivialement) un ensemble représentatif minimal de B^2_{1S}, il est possible d'imputer pour P une valeur qui satisfait à toutes les vérifications originales, sauf ψ^2_{0S}. D'où, la valeur de la fonction cible (3.1) pour cette solution est $(w_P + s_2) / 2 = 3 / 2$.

L'élimination d'une autre variable, disons C, de l'ensemble courant de vérifications pourrait peut-être améliorer la solution. On obtient :

$$
\begin{array}{lrll}
\psi^1_{2H}: & T \geq 0 & & (\psi^1_{1H}) \\
\psi^2_{2H}: & N \geq 0 & & (\psi^3_{1H}) \\
\psi^3_{2H}: & 550N - T \geq 0 & & (\psi^4_{1H}) \\
\psi^1_{2S}: & 1,1T \geq 0 & (B^1_{2S} = \{2\}) & (\psi^2_{1H}, \psi^2_{1S}) \\
\psi^2_{2S}: & 0,6T \geq 0 & (B^2_{2S} = \{1, 2\}) & (\psi^1_{1S}, \psi^2_{1S}).
\end{array}
$$

Chacune des deux nouvelles vérifications avec avertissement est redondante, parce que toutes deux sont équivalentes à la vérification avec rejet ψ^1_{2H}. En fait, les valeurs originales restantes $(T^0, N^0) = (100, 5)$ satisfont à toutes les vérifications courantes. Cela signifie que les variables P et C peuvent être imputées pour satisfaire à toutes les vérifications originales, avec rejet ainsi qu'avec avertissement. La valeur de la fonction cible (3.1) de cette solution est égale à $(w_P + w_C) / 2 = 1$. Donc, la nouvelle solution est meilleure que la précédente. De surcroît, cette solution ne peut pas être améliorée davantage en éliminant d'autres variables sur la branche courante de l'arbre binaire.

Si l'on explore le reste de l'arbre binaire, on finit par constater que la meilleure solution découverte jusqu'à présent (imputer P et C) est aussi la solution optimale. Un enregistrement cohérent possible obtenu par imputation de P et C est : $(T, P, C, N) = (100, 40, 60, 5)$. Cette solution mène à la belle interprétation que les valeurs originales des variables *profit* et *coût total* ont été surestimées d'un facteur 1 000. Il est intéressant de souligner que, si l'on utilise seulement des vérifications avec rejet dans cet exemple, la première solution découverte plus haut (imputer uniquement P) est la solution optimale. Dans ce cas, il n'existe qu'un seul moyen d'obtenir un enregistrement cohérent : $(T, P, C, N) = (100, -59\,900, 60\,000, 5)$. Cela montre que, du moins dans le présent exemple, les vérifications avec avertissement sont importantes pour trouver des imputations qui non seulement satisfont aux vérifications avec rejet, mais sont également plausibles.

7 Application

Afin de tester le nouvel algorithme de localisation des erreurs en pratique, une implémentation prototype a été écrite en langage de programmation R. Ce prototype s'inspire fortement de la

fonctionnalité de localisation des erreurs de R qui a été rendue disponible dans le progiciel `editrules` (De Jonge et Van der Loo 2011; Van der Loo et De Jonge 2011).

Pour tester le prototype, un ensemble de données artificielles a été construit en sélectionnant 12 variables numériques $(x_1, \ldots, x_{12})$ parmi les statistiques structurelles sur les entreprises de 2007 aux Pays-Bas pour le secteur du commerce de gros. Nous avons sélectionné tous les enregistrements se rapportant à des moyennes entreprises (comptant de 10 à 100 employés) qui avaient été vérifiés manuellement durant la production ordinaire, et nous les avons répartis en deux ensembles de données de 728 enregistrements chacun. Ces ensembles de données originaux étaient tous deux considérés comme exempts d'erreur. Nous avons introduit aléatoirement un nombre important d'erreurs dans l'un des ensembles de données en appliquant la procédure suivante :

- dans 4 % des valeurs non nulles originales, deux chiffres ont été permutés;
- dans 4 % des valeurs non nulles originales, un chiffre a été ajouté aléatoirement;
- dans 4 % des valeurs non nulles originales, un chiffre a été omis aléatoirement;
- dans 4 % des valeurs non nulles originales, un chiffre a été remplacé aléatoirement par un autre;
- 4 % des valeurs non nulles originales ont été multipliées par 25;
- 4 % des valeurs non nulles originales ont été divisées par 25 et le résultat a été arrondi à l'entier le plus proche;
- 6 % des valeurs non nulles originales ont été remplacées par zéro;
- 5 % des valeurs nulles originale ont été remplacées par des entiers aléatoires tirés de {1, …, 1 000};
- 10 % des valeurs originales de x_{11} et de x_{12} ont été multipliées par –1.

Cette procédure a été exécutée de manière qu'un changement tout au plus ait lieu dans chaque valeur. Le deuxième ensemble de données a été maintenu exempt d'erreurs et a servi de référence.

Le tableau 7.1 montre les vérifications avec rejet et avec avertissement qui ont été appliquées aux données de test. Les vérifications avec rejet ont été copiées du système de production

ordinaire. Les vérifications avec avertissement ont été déterminées en examinant un certain nombre de distributions univariées et bivariées dans les données de référence.

Tableau 7.1
Vérifications utilisées dans l'application de test

Vérifications avec rejet :	$x_1 + x_2 = x_3$ $x_2 = x_4$ $x_5 + x_6 + x_7 = x_8$ $x_3 + x_8 = x_9$ $x_9 - x_{10} = x_{11}$ $x_j \geq 0 \quad (j = 1, \ldots, 10 \text{ et } j = 12)$
Vérifications avec avertissement :	$x_2 \geq 0,5x_3$ $x_3 \geq 0,9x_9$ $x_5 + x_6 \geq x_7$ $x_9 \geq 50x_{12}$ $x_9 \leq 5\ 000x_{12}$ $x_{11} \leq 0,4x_9$ $x_{11} \geq -0,1x_9$ $x_{12} \geq 1$ $x_{12} \geq 5$ $x_{12} \leq 100$

L'algorithme de localisation des erreurs a été appliqué à l'ensemble de données contenant des erreurs artificielles sous diverses conditions. Tout au long du test, tous les poids de confiance w_j^N ont été fixés à 1 et le paramètre λ dans (3.1) a été choisi égal à $1 / 2$. Nous avons considéré les approches suivantes :

A. Le premier test ne comprenait que les vérifications avec rejet du tableau 7.1.
B. Le deuxième test comprenait chacune des vérifications du tableau 7.1, toutes interprétées chacune comme étant des vérifications avec rejet.
C. Le troisième test comprenait chacune des vérifications du tableau 7.1, en faisant une distinction entre les vérifications avec rejet et les vérifications avec avertissement.

Chaque vérification avec avertissement a reçu le même poids de non-satisfaction fixe $s_k = 1$.

D. Le quatrième test était semblable au troisième, sauf que les poids de non-satisfaction fixes différaient entre les vérifications avec avertissement. Pour chacune de celles-ci, s_k a été calculé comme étant la fraction des enregistrements dans l'ensemble de données de référence qui satisfaisaient à la vérification. Donc, un poids de non-satisfaction plus faible a été attribué aux vérifications avec avertissement pour lesquelles les cas de non-satisfaction étaient plus fréquents dans l'ensemble de données de référence, et inversement. La raison logique est que tous les cas de non-satisfaction des vérifications avec avertissement survenant dans l'ensemble de données de référence étaient causés par des combinaisons inhabituelles, mais correctes des valeurs. En associant de faibles poids aux vérifications avec avertissement qui sont souvent non satisfaites dans les données de référence, nous faisons en sorte que ces vérifications puissent aussi être non satisfaites plus facilement lors de la vérification des données de test.

Puisque la distribution des erreurs dans nos ensembles de données de test était connue, nous avons pu évaluer directement la performance de chaque approche de localisation automatique des erreurs. Pour cela, nous avons utilisé plusieurs indicateurs de qualité. Considérons la table de contingence 2×2 qui suit :

		décelé : erreur	**pas d'erreur**
vrai :	erreur	*VP*	*FN*
	pas d'erreur	*FP*	*VN*

Le premier indicateur de qualité mesure la proportion d'erreurs vraies qui ont été manquées par l'algorithme (proportion de faux négatifs) :

$$\alpha = \frac{FN}{VP + FN}.$$

Le deuxième indicateur de qualité mesure la proportion de valeurs correctes qui ont été prises à tort pour des erreurs par l'algorithme (proportion de faux positifs) :

$$\beta = \frac{FP}{FP + VN}.$$

Le troisième indicateur de qualité mesure la proportion globale de décisions incorrectes faites par l'algorithme :

$$\gamma = \frac{FN + FP}{VP + FN + FP + VN}.$$

Ces trois indicateurs évaluent la performance de l'algorithme en ce qui concerne la détermination des valeurs individuelles comme étant correctes ou incorrectes. Ils ont été utilisés dans des études d'évaluation antérieures; voir, par exemple, Pannekoek et De Waal (2005).

Afin d'évaluer la performance de l'algorithme sous un angle légèrement différent, nous avons également calculé le pourcentage d'enregistrements pour lesquels l'algorithme a trouvé exactement la bonne solution – c'est-à-dire la solution qui identifie comme étant incorrectes toutes les valeurs incorrectes, et uniquement celles-là. Cet indicateur est désigné par δ. Une bonne approche de vérification devrait donner de faibles scores pour α, β et γ, mais un score élevé pour δ.

Le tableau 7.2 montre les valeurs des indicateurs de qualité pour les approches de vérification A, B, C et D. On voit que l'approche B est surpassée par les autres pour toutes les mesures, sauf la proportion d'erreurs manquées. Donc, utiliser les vérifications avec avertissement comme s'il s'agissait de vérifications avec rejet ne donne pas de bons résultats pour cet ensemble de données. En fait, de meilleurs résultats sont obtenus avec l'approche A, qui ne comporte aucune vérification avec avertissement. On peut également constater que les approches C et D, qui utilisent le nouvel algorithme pour tenir compte des vérifications avec avertissement, donnent de meilleurs résultats que les approches A et B, qui s'appuient sur l'ancien algorithme. Dans l'ensemble, l'approche D semble donner les meilleurs résultats dans la présente expérience. Comparativement à l'approche A, l'approche D détecte en fait correctement un plus grand nombre d'erreurs *et* un plus grand nombre de valeurs correctes.

Il convient de souligner que, sous l'ancienne définition du problème de localisation des erreurs, les approches A et B représentent les deux options extrêmes existant pour l'utilisation de vérifications avec avertissement : soit ne pas les utiliser, ou les utiliser toutes comme des vérifications avec rejet. En guise de compromis entre ces deux options, on pourrait aussi décider d'utiliser uniquement un sous-ensemble de vérifications avec avertissement comme vérifications avec rejet et écarter les autres. Nous n'avons pas testé cette approche durant l'expérience. On pourrait s'attendre à ce qu'elles produisent des scores pour les mesures α, β, γ et δ comprises entre celles données par les approches A et B.

Tableau 7.2
Résultats de la localisation automatique des erreurs pour les données artificielles

	Indicateurs de qualité			
Approche	**α**	**β**	**γ**	**δ**
A	0,364	0,047	0,115	40 %
B	0,232	0,131	0,153	37 %
C	0,227	0,060	0,096	47 %
D	0,253	0,037	0,083	52 %

8 Conclusion

Dans le présent article, nous avons proposé une nouvelle formulation du problème de localisation des erreurs qui permet de faire la distinction entre les vérifications avec rejet et les vérifications avec avertissement. En outre, nous avons montré qu'une version modifiée de l'algorithme *branch-and-bound* de De Waal et Quere (2003) peut être utilisée pour résoudre ce nouveau problème de localisation des erreurs. Nous avons suggéré que ce nouvel algorithme pourrait être utilisé pour améliorer la vérification automatique. Les résultats empiriques présentés à la section 7 confirment la validité de cette suggestion, mais il faut souligner que ces résultats ont été obtenus au moyen de données contenant des erreurs synthétiques. Une application du nouvel algorithme de localisation des erreurs à des données réelles est à l'étude.

Un problème encore non résolu est celui de savoir quel est le meilleur moyen de modéliser le coût de la non-satisfaction des vérifications avec avertissement, c'est-à-dire la façon dont le terme D_{soft} dans (3.1) doit être défini. Les résultats différents obtenus en suivant les approches C et D à la section 7 montrent que la qualité de la localisation automatique des erreurs peut être améliorée grâce à un choix convenable des poids de non-satisfaction. Il sera intéressant de déterminer dans quelle mesure on pourrait encore améliorer la qualité de la vérification automatique en expérimentant avec diverses combinaisons de poids de non-satisfaction s_k, de poids de confiance w_j et du paramètre d'équilibrage λ dans (3.1).

D'autres formes de D_{soft} que celle donnée par (3.2) pourraient également être envisagées, y compris des formes qui dépendent du degré de non-satisfaction des vérifications avec avertissement. Comme il est mentionné à la section 3, il est intuitivement intéressant de tenir compte de la mesure dans laquelle les vérifications avec avertissement ne sont pas satisfaites dans le problème de localisation des erreurs, afin que les cas de non-satisfaction plus importante donnent des valeurs plus élevées de D_{soft}. Un choix intéressant pour D_{soft} pourrait être la distance de Mahalanobis des cas de non-satisfaction des vérifications avec avertissement, comme l'a proposé Hedlin (2003) dans un contexte différent. Il convient de souligner que l'algorithme mentionné à la section 6 peut être utilisé pour résoudre le problème de localisation des erreurs pour tous les choix de D_{soft} qui peuvent être exprimés sous forme de fonctions (se comportant raisonnablement bien) de $z_1, \ldots, z_{K_S}$. On utilise simplement l'expression appropriée pour D_{soft} comme fonction cible dans le problème (5.1). Par ailleurs, si D_{soft} dépend explicitement de la mesure dans laquelle les vérifications avec avertissement ne sont pas satisfaites, nous devons recourir à une approche plus complexe. En particulier, l'information fournie par les théorèmes 1 et 2 ne suffit plus, parce que nous devons maintenant savoir non seulement quelles vérifications avec avertissement ne seront pas satisfaites après l'imputation, mais aussi la mesure dans laquelle elles ne seront pas satisfaites. Une approche en vue de résoudre le problème de localisation des erreurs dans cette situation plus compliquée figure dans Scholtus (2011).

En résumé, il reste à voir comment les résultats théoriques exposés dans le présent article doivent être appliqués pour obtenir les meilleurs résultats possibles en pratique. Néanmoins, étant donné que les spécialistes du domaine se servent de la différence conceptuelle entre les vérifications avec rejet et les vérifications avec avertissement durant la vérification manuelle, il semble évident que le nouvel algorithme de localisation des erreurs offre la possibilité d'accroître la qualité de la vérification automatique.

Remerciements

Les opinions exprimées dans le présent article sont celles de l'auteur et ne reflètent pas forcément les politiques de Statistics Netherlands. L'auteur tient à remercier Jeroen Pannekoek et Ton de Waal pour leurs suggestions judicieuses, Edwin de Jonge et Mark van der Loo pour la programmation de l'algorithme *branch-and-bound* en R, Sevinç Göksen pour avoir recueilli les résultats d'évaluation présentés à la section 7 et enfin, le rédacteur associé et deux examinateurs anonymes pour leurs commentaires constructifs.

Annexe A – Preuves

A.1 Preuve du théorème 1

Afin de prouver le théorème 1, il est commode de prouver d'abord un lemme auxiliaire. Supposons que Ψ_q est obtenu à partir de Ψ_{q-1} en éliminant x_g. Nous définissons, pour chaque vérification ψ_q^k, l'ensemble d'indices A_q^k des vérifications comprises dans l'ensemble Ψ_{q-1} à partir duquel elle a été dérivée. Autrement dit, nous définissons $A_q^k := \{l\}$ si ψ_q^k est obtenue en copiant la vérification ψ_{q-1}^l et nous définissons $A_q^k := \{s, t\}$ si ψ_q^k est obtenue en éliminant une variable de la paire de vérifications $(\psi_{q-1}^s, \psi_{q-1}^t)$.

Lemme 1. Considérons la situation du théorème 1 pour $q \geq 1$, et supposons que x_g a été éliminée pour obtenir Ψ_q à partir de Ψ_{q-1}. Soit A un ensemble représentatif des ensembles d'indices A_q^k appartenant à toutes les vérifications $\psi_q^k \in \Psi_q^{(2)}$. Alors, il existe une valeur de x_g

qui, de concert avec les valeurs originales des variables qui interviennent dans Ψ_q, satisfait toutes les vérifications comprises dans Ψ_{q-1} sauf celles comprises dans A.

Preuve (du lemme 1). Par construction, A contient tous les indices des vérifications non satisfaites provenant de Ψ_{q-1} qui ne font pas intervenir x_g. Donc, le lemme ne peut être faux que s'il existe deux vérifications qui font intervenir x_g, disons ψ_{q-1}^s et ψ_{q-1}^t, avec $s \notin A$ et $t \notin A$, de sorte qu'il soit impossible de trouver une valeur de x_g qui satisfait les deux vérifications simultanément. Dans ce cas, une vérification implicite dans Ψ_q est générée en éliminant x_g de ψ_{q-1}^s et ψ_{q-1}^t. De surcroît, en vertu de la propriété fondamentale énoncée à la fin de la section 2.3, cette vérification implicite doit ne pas être satisfaite par les valeurs originales des autres variables, autrement dit la vérification implicite doit être un élément de $\Psi_q^{(2)}$. Cependant, cela contredirait l'hypothèse voulant que A est un ensemble représentatif de A_q^k pour toute $\psi_q^k \in \Psi_q^{(2)}$. Donc, il est impossible de trouver une telle paire de vérifications, et le lemme s'ensuit.

La preuve du théorème 1 procède maintenant par induction sur le nombre de variables traitées q. Pour $q = 0$, la déclaration est triviale. Pour $q = 1$, le théorème suit un cas particulier du lemme 1; notons que $B_1^k \equiv A_1^k$. Nous supposons par conséquent que la déclaration a été prouvée pour tout $q \in \{0, 1, \ldots, Q-1\}$, et nous considérons le cas $q = Q$, avec $Q \geq 2$.

Si Ψ_Q est obtenu à partir de Ψ_{Q-1} en fixant une variable à sa valeur originale, et que B est un ensemble représentatif des ensembles B_Q^k pour les vérifications non satisfaites dans Ψ_Q, alors par construction B est également un ensemble représentatif des ensembles B_{Q-1}^k pour les vérifications non satisfaites provenant de Ψ_{Q-1}. D'où, dans ce cas, la déclaration pour $q = Q$ découle directement de l'hypothèse d'induction.

Donc, il nous reste le cas où Ψ_Q est obtenu à partir de Ψ_{Q-1} en éliminant une variable, disons x_g. Nous définissons, pour chaque vérification $\psi_Q^k \in \Psi_Q^{(2)}$, l'ensemble d'indices A_Q^k des vérifications provenant de Ψ_{Q-1} duquel ψ_Q^k est dérivée, comme plus haut. Ensuite, nous utilisons

B pour construire un ensemble A, en appliquant la procédure suivante à chaque vérification $\psi_Q^k \in \Psi_Q^{(2)}$:

- Si ψ_Q^k est obtenue en copiant ψ_{Q-1}^l (de sorte que $A_Q^k = \{l\}$ et $B_Q^k = B_{Q-1}^l$), alors nous ajoutons l à A.
- Si ψ_Q^k est obtenue en éliminant x_g de ψ_{Q-1}^s et ψ_{Q-1}^t (de sorte que $A_Q^k = \{s, t\}$ et $B_Q^k = B_{Q-1}^s \cup B_{Q-1}^t$), alors nous ajoutons s à A si B contient un élément de B_{Q-1}^s, et nous ajoutons t à A autrement.

Il est facile de voir que cette procédure produit un ensemble représentatif A des ensembles d'indices A_Q^k pour toute $\psi_Q^k \in \Psi_Q^{(2)}$.

Selon le lemme 1, il existe une valeur de x_g qui, de concert avec les valeurs originales des $p - q$ variables qui n'ont pas été traitées, satisfait les vérifications comprises dans Ψ_{Q-1} sauf celles comprises dans A. Autrement dit, Ψ_{Q-1} peut être partitionné de la même façon que Ψ_Q comme $\Psi_{Q-1} = \Psi_{Q-1}^{(1)} \cup \Psi_{Q-1}^{(2)}$, où $\Psi_{Q-1}^{(2)}$ contient les vérifications dont les indices figurent dans A. En outre, il n'est pas difficile de voir que la procédure susmentionnée implique que B est un ensemble représentatif des ensembles d'indices B_{Q-1}^k pour toute $\psi_{Q-1}^k \in \Psi_{Q-1}^{(2)}$. Donc, l'hypothèse d'induction établit que, sachant les valeurs originales des variables qui n'ont pas été éliminées *et* sachant la valeur choisie pour x_g, il existe des valeurs pour les autres variables éliminées qui satisfont toutes les vérifications originales sauf celles comprises dans B. Cela montre que la déclaration est vérifiée pour $q = Q$ et achève la preuve du théorème 1.

A.2 Preuve du théorème 2

Pour prouver le théorème 2, nous partons de nouveau d'un lemme auxiliaire. De manière analogue au cas numérique, quand Ψ_q est obtenu à partir de Ψ_{q-1} en éliminant v_g, nous définissons l'ensemble d'indices A_q^k des vérifications comprises dans Ψ_{q-1} à partir duquel la vérification $\psi_q^k \in \Psi_q$ est dérivée. Pour être précis, nous définissons $A_q^k := \{l\}$ si ψ_q^k est obtenue en copiant la vérification ψ_{q-1}^l, et nous définissons $A_q^k := T$ si ψ_q^k est obtenue en éliminant une variable de l'ensemble de vérifications $\psi_{q-1}^t (t \in T)$.

Lemme 2. Considérons la situation du théorème 2 pour $q \geq 1$, et supposons que v_g a été éliminée pour obtenir Ψ_q à partir de Ψ_{q-1}. Soit A un ensemble représentatif des ensembles d'indices A_q^k appartenant à toute vérification $\psi_q^k \in \Psi_q^{(2)}$. Alors, il existe une valeur de v_g qui, de concert avec les valeurs originales des variables qui interviennent dans Ψ_q, satisfait toutes les vérifications comprises dans Ψ_{q-1} sauf celles comprises dans A.

Preuve (du lemme 2). Par construction, A contient tous les indices des vérifications non satisfaites provenant de Ψ_{q-1} qui ne font pas intervenir v_g. Donc, le lemme ne peut être faux que s'il existe des vérifications qui font intervenir v_g, disons $\psi_{q-1}^{t_1}, \ldots, \psi_{q-1}^{t_r}$, avec $A \cap \{t_1, \ldots, t_r\} = \varnothing$, de sorte qu'il soit impossible de trouver une valeur de v_g qui satisfait ces vérifications simultanément, sachant les valeurs des autres variables. Clairement, cela ne pourrait se produire que si $F_g^{t_1} \cup \cdots \cup F_g^{t_r} = D_g$, puisqu'autrement, toute valeur de v_g en dehors de $F_g^{t_1} \cup \cdots \cup F_g^{t_r}$ conviendrait. Nous pouvons supposer sans perte de généralité que $T' = \{t_1, \ldots, t_r\}$ est un ensemble minimal ayant cette propriété. En outre, il doit être vérifié dans ce cas que, pour toutes les variables intervenant dans Ψ_q, la valeur originale de v_j est contenue dans tous les ensembles $F_j^{t_1}, \ldots, F_j^{t_r}$. Autrement dit, T' doit satisfaire les propriétés (2.8) et (2.9). Cela signifie que T' créerait une vérification implicite dans Ψ_q qui, en vertu de la propriété fondamentale énoncée à la fin de la section 2.3, doit ne pas être satisfaite par les valeurs originales des variables restantes. Cependant, cela contredirait l'hypothèse voulant que A est un ensemble représentatif de A_q^k pour toute $\psi_q^k \in \Psi_q^{(2)}$. Cela achève la preuve du lemme 2.

La preuve du théorème 2 est maintenant clairement analogue à celle du théorème 1, avec le lemme 2 jouant le rôle du lemme 1.

Bibliographie

Bankier, M., Lachance, M. et Poirier, P. (2000). *2001 Canadian Census Minimum Change Donor Imputation Methodology*. Document de travail, UN/ECE Work Session on Statistical Data Editing, Cardiff.

Bankier, M., et Crowe, S. (2009). *Enhancements to the 2011 Canadian Census E&I System*. Document de travail, UN/ECE Work Session on Statistical Data Editing, Neuchâtel.

Boskovitz, A., Goré, R. et Wong, P. (2005). *Data Editing and Logic*. Document de travail, UN/ECE Work Session on Statistical Data Editing, Ottawa.

Bruni, R. (2004). Discrete models for data imputation. *Discrete Applied Mathematics*, 144, 59-69.

Bruni, R. (2005). Error correction for massive datasets. *Optimization Methods and Software*, 20, 297-316.

Casado Valero, C., Del Castillo Cuervo-Arango, F., Mateo Ayerra, J. et De Santos Ballesteros, A. (1996). *Quantitative Data Editing*: *Quadratic Programming Method*. Présenté à la conférence COMPSTAT 1996, Barcelone.

De Jonge, E., et Van der Loo, M. (2011). *Manipulation of Linear Edits and Error Localization with the Editrules Package*. Document de travail 201120, Statistics Netherlands, La Haye.

De Waal, T. (2003a). Résolution du problème de localisation des erreurs par la génération de sommets. *Techniques d'enquête*, 29, 1, 81-90.

De Waal, T. (2003b). *Processing of Erroneous and Unsafe Data*. Thèse de doctorat, Erasmus University, Rotterdam.

De Waal, T. (2005). *SLICE 1.5: a Software Framework for Automatic Edit and Imputation*. Document de travail, UN/ECE Work Session on Statistical Data Editing, Ottawa.

De Waal, T., Pannekoek, J. et Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. New Jersey : John Wiley & Sons, Inc.

De Waal, T., et Quere, R. (2003). A fast and simple algorithm for automatic editing of mixed data. *Journal of Official Statistics*, 19, 383-402.

Fellegi, I.P., et Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

Garfinkel, R.S., Kunnathur, A.S. et Liepins, G.E. (1986). Optimal imputation of erroneous data: Categorical data, general edits. *Operations Research*, 34, 744-751.

Garfinkel, R.S., Kunnathur, A.S. et Liepins, G.E. (1988). Error localization for erroneous data: Continuous data, linear constraints. *SIAM Journal on Scientific and Statistical Computing*, 9, 922-931.

Hedlin, D. (2003). Score functions to reduce business survey editing at the U.K. Office for National Statistics. *Journal of Official Statistics*, 19, 177-199.

Kovar, J., et Whitridge, P. (1990). Generalized edit and imputation system; Overview and applications. *Revista Brasileira de Estadistica*, 51, 85-100.

Little, R.J.A., et Smith, P.J. (1987). Editing and imputation of quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.

Magnus, J.R., Van Tongeren, J.W. et De Vos, A.F. (2000). National accounts estimation using indicator ratios. *Review of Income and Wealth*, 46, 329-350.

Mirsky, L. (1971). *Transversal Theory*. New York : Academic Press, Inc.

Nemhauser, G.L., et Wolsey, L.A. (1988). *Integer and Combinatorial Optimization*. New York : John Wiley & Sons, Inc.

Pannekoek, J., et De Waal, T. (2005). Automatic edit and imputation for business surveys: The Dutch contribution to the EUREDIT project. *Journal of Official Statistics*, 21, 257-286.

Ragsdale, C.T., et McKeown, P.G. (1996). On solving the continuous data editing problem. *Computers & Operations Research*, 23, 263-273.

Riera-Ledesma, J., et Salazar-González, J.J. (2003). *New Algorithms for the Editing and Imputation Problem*. Document de travail, UN/ECE Work Session on Statistical Data Editing, Madrid.

Sande, G. (1978). An Algorithm for the Fields to Impute Problems of Numerical and Coded Data. Rapport technique, Statistique Canada.

Schaffer, J. (1987). Procedure for solving the data-editing problem with both continuous and discrete data types. *Naval Research Logistics*, 34, 879-890.

Scholtus, S. (2011). *Automatic Editing with Soft Edits*. Document de travail 201130, Statistics Netherlands, La Haye.

Schrijver, A. (1986). *Theory of Linear and Integer Programming*. New York : John Wiley & Sons, Inc.

Todaro, T.A. (1999). *Overview and Evaluation of the AGGIES Automated Edit and Imputation System*. Document de travail, UN/ECE Work Session on Statistical Data Editing, Rome.

Van der Loo, M., et De Jonge, E. (2011). *Manipulation of Categorical Data Edits and Error Localization with the Editrules Package*. Document de travail 201129, Statistics Netherlands, La Haye.

Williams, H.P. (1986). Fourier's method of linear programming and its dual. *The American Mathematical Monthly*, 93, 681-695.

Winkler, W.E. (1995). *Editing Discrete Data*. Document de travail, UN/ECE Work Session on Statistical Data Editing, Athènes.

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 105-137
Statistique Canada, N° 12-001-X au Catalogue

Estimation parcimonieuse et efficace de la variance par rééchantillonnage pour les enquêtes complexes

Jae Kwang Kim et Changbao Wu[1]

Résumé

Il est courant que les organismes d'enquête fournissent des poids de rééchantillonnage dans les fichiers de données d'enquête. Ces poids de rééchantillonnage servent à produire de manière simple et systématique des estimations valides et efficaces de la variance pour divers estimateurs. Cependant, la plupart des méthodes existantes de construction de poids de rééchantillonnage ne sont valides que pour des plans d'échantillonnage particuliers et nécessitent habituellement un très grand nombre de répliques. Dans le présent article, nous montrons d'abord comment produire les poids de rééchantillonnage en se basant sur la méthode décrite dans Fay (1984) de manière que l'estimateur de la variance par rééchantillonnage résultant soit algébriquement équivalent à l'estimateur de la variance par linéarisation entièrement efficace pour tout plan d'échantillonnage donné. Puis, nous proposons une nouvelle méthode de calage des poids afin que l'estimation soit simultanément efficace et parcimonieuse au sens où un petit nombre de jeux de poids de rééchantillonnage peuvent produire des estimateurs de la variance par rééchantillonnage valides et efficaces pour les paramètres de population importants. La méthode que nous proposons peut être conjuguée aux méthodes de rééchantillonnage existantes pour les enquêtes complexes à grande échelle. Nous discutons également de la validité des méthodes proposées et de leur extension à certains plans d'échantillonnage équilibrés. Les résultats de simulations montrent que les estimateurs de variance que nous proposons suivent très bien les probabilités de couverture des intervalles de confiance. Les stratégies que nous proposons auront vraisemblablement des répercussions sur la façon de produire les fichiers de données d'enquête à grande diffusion et d'analyser ces ensembles de données.

Mots clés : Bootstrap; calage; jackknife; méthode de linéarisation; poids de rééchantillonnage; plan d'échantillonnage; décomposition spectrale.

1 Introduction

L'estimation de la variance est un problème important en pratique des sondages. Outre l'usage analytique des variances, par exemple pour tester des hypothèses statistiques et construire des intervalles de confiance, l'estimation des variances peut servir à fournir des mesures descriptives de l'exactitude des estimations d'après des données d'enquête et de l'efficacité du plan d'échantillonnage donné. Deux types de techniques sont utilisés fréquemment pour estimer la

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames IA 50011-1210. Courriel : jkim@iastate.edu ; Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (ON) N2L 3G1. Courriel : cbwu@uwaterloo.ca.

variance dans le cadre de l'estimation fondée sur le plan de sondage. La première, appelée méthode de linéarisation, consiste à appliquer la formule classique de la variance directement à l'estimateur si le paramètre est un total de population ou au développement en série de Taylor en une étape linéarisée de l'estimateur si le paramètre est une fonction non linéaire d'un ou de plusieurs totaux de population. La deuxième, appelée méthode de rééchantillonnage ou méthode par répliques, consiste à construire des estimateurs de variance d'une manière systématique simple en utilisant plusieurs jeux de poids de rééchantillonnage ainsi que l'ensemble de données d'enquête original.

Les méthodes d'estimation de la variance par rééchantillonnage sont devenues très répandues pour les inférences sous le plan de sondage en utilisant des données d'enquête complexes. Certaines des premières utilisations pratiques des poids de rééchantillonnage remontent aux années 1970 au U.S. Bureau of the Census, au Bureau of Labor Statistics et à Westat (Dippo, Fay et Morganstein 1984). Il est maintenant pratique courante pour les organismes d'enquête de fournir des poids de rééchantillonnage avec les données d'enquête. La caractéristique la plus intéressante de cette approche tient au fait qu'elle fonctionne de la même façon quelle que soit la complexité du paramètre. Pour les paramètres qui sont des fonctions lisses des moyennes ou des totaux de population, l'étape de « linéarisation » a été intégrée automatiquement dans le processus d'estimation de sorte qu'il n'est pas nécessaire de calculer les dérivées partielles requises dans le développement en série de Taylor. L'approche est très conviviale pour les analyses de données à objectifs multiples une fois que l'ensemble de données d'enquête diffusé est assorti des poids de rééchantillonnage. En outre, l'utilisation de méthodes de rééchantillonnage réduit les préoccupations concernant les questions de confidentialité puisque l'on ne diffuse pas de renseignements détaillés sur le plan de sondage, tels que les identificateurs de strate ou de grappe (Lu et Sitter 2008).

Les poids de rééchantillonnage sont habituellement construits par la méthode du bootstrap, du jackknife ou des répliques répétées équilibrées (BRR pour *balanced repeated replication*). Rust

et Rao (1996), Shao (1996, 2003) et Wolter (2007) ont donné d'excellents aperçus du sujet. Les trois grands problèmes de la construction des poids de rééchantillonnage sont la *validité*, l'*efficacité* et la *parcimonie*. Par validité, on entend l'absence asymptotique de biais des estimateurs de la variance par rééchantillonnage sous le plan d'échantillonnage donné. L'absence asymptotique de biais d'un estimateur est généralement un concept plus faible que celui d'estimateur convergent. Si le coefficient de variation de l'estimateur de la variance tend vers zéro, l'estimateur asymptotiquement sans biais de la variance est également convergent. L'efficacité est mesurée par la performance relative de l'estimateur de la variance par rééchantillonnage comparativement à l'estimateur classique de la variance par linéarisation qui est considéré comme étant entièrement efficace. La parcimonie fait référence au nombre de jeux de poids de rééchantillonnage nécessaires pour obtenir une estimation de la variance entièrement efficace.

La validité des estimateurs de la variance par rééchantillonnage a été discutée par Krewski et Rao (1981), Shao et Tu (1995) et Fuller (2009a), entre autres. L'efficacité et la stabilité des estimateurs de la variance par rééchantillonnage ont été discutées par Rust et Kalton (1987) et par Jang et Eltinge (2009). En ce qui concerne la parcimonie, Kott (2001) a considéré l'utilisation du jackknife avec suppression d'un groupe pour atteindre la parcimonie sous certains plans, et Lu, Brick et Sitter (2006) ont également discuté du groupement de strates pour l'estimation parcimonieuse de la variance par rééchantillonnage.

La plupart des méthodes de rééchantillonnage sur lesquelles portent la littérature ne sont valides que pour certains plans d'échantillonnage. Par exemple, la méthode du jackknife est habituellement utilisée pour l'échantillonnage aléatoire stratifié (Krewski et Rao 1981). La méthode du bootstrap possède plusieurs procédures très utilisées, y compris la méthode du bootstrap sans remise (Gross 1980; McCarthy et Snowden 1985), la méthode du bootstrap avec rééchelonnement des poids (*re-scaling bootstrap*) (Rao et Wu 1988; Preston 2009) et la méthode

du bootstrap à concordance-miroir (*mirror-match bootstrap*) (Sitter 1992). Toutefois, ces procédures ne s'appliquent qu'à certains types de plans d'échantillonnage.

La parcimonie d'une méthode de rééchantillonnage dépend de la façon dont sont construits les poids de rééchantillonnage. Le nombre de jeux de poids de rééchantillonnage jackknife est relié au nombre d'unités dans l'échantillon et peut être très élevé si l'échantillon est de grande taille. Les méthodes du bootstrap nécessitent habituellement au moins 1 000 jeux de poids de rééchantillonnage pour atteindre le niveau souhaité d'efficacité. En guise de compromis, la plupart des organismes d'enquête fournissent 500 jeux de poids bootstrap ainsi que les données sur les principales variables étudiées. Les ensembles de données résultant demeurent néanmoins trop grands pour que les utilisateurs puissent les vérifier visuellement et peuvent être très lourds à manipuler en pratique.

Le présent article expose des méthodes de construction de poids de rééchantillonnage efficaces et parcimonieux pour l'estimation de la variance dans le cadre d'estimation fondée sur le plan de sondage. En maintenant l'efficacité complète de l'estimateur de variance résultant pour les variables clés à l'aide d'un petit nombre de jeux de poids de rééchantillonnage, nos méthodes apportent une solution à l'une des tâches principales à l'étape de la préparation des fichiers de données et peuvent être appliquées facilement par les réalisateurs d'enquêtes afin de réduire le fardeau que représente pour les utilisateurs des données le traitement de fichiers de données excessivement grands. Une limite importante de la méthode que nous proposons tient au fait qu'elle ne traite pas directement les situations où les poids de sondage sont corrigés de la non-réponse ou calés sur des données de population auxiliaires connues.

À la section 2, nous présentons une procédure générale de construction de poids de rééchantillonnage fondée sur la méthode de Fay (1984) ainsi que de Fay et Dippo (1989), qui fournit des poids de rééchantillonnage entièrement efficaces pour des plans d'échantillonnage arbitraires. À la section 3, nous discutons de deux stratégies, l'échantillonnage aléatoire et la pondération par calage, pour construire des poids de rééchantillonnage parcimonieux. En utilisant

une nouvelle application de la méthode de calage, les méthodes que nous proposons permettent d'utiliser un petit nombre de jeux de poids de rééchantillonnage tout en obtenant des estimateurs de variance par rééchantillonnage efficaces. À la section 4, nous présentons une théorie asymptotique pour la validité de l'estimateur de la variance par rééchantillonnage. À la section 5, nous discutons des extensions de trois méthodes à certains plans d'échantillonnage équilibrés. À la section 6, nous présentons les résultats d'une étude par simulation réalisée en utilisant des données réelles de l'Enquête sur les dépenses des familles réalisée par Statistique Canada, pour évaluer l'efficacité des stratégies proposées d'estimation de la variance par rééchantillonnage. Enfin, à la section 7, nous présentons certaines conclusions.

2 Une procédure générale de construction de poids de rééchantillonnage entièrement efficace

En principe, nous pouvons construire des poids de rééchantillonnage pour tout plan d'échantillonnage mesurable en utilisant la méthode décrite dans Fay (1984) et dans Fay et Dippo (1989), de manière que les estimateurs de la variance par rééchantillonnage résultants soient algébriquement équivalents aux estimateurs de la variance par linéarisation classiques.

Soit $\mathcal{U} = \{1, 2, \ldots, N\}$ l'ensemble de N unités dans la population finie et $\mathcal{S} = \{1, 2, \ldots, n\}$ l'ensemble de n unités dans l'échantillon, sélectionné selon un plan d'échantillonnage probabiliste. Soit $w_i = 1 / \pi_i$ les poids de sondage de base, où $\pi_i = P(i \in \mathcal{S})$ est la probabilité d'inclusion de premier ordre de l'unité i.

Soit y_i la valeur de la variable étudiée y pour l'unité i et $t_y = \sum_{i=1}^{N} y_i$ le total de population d'intérêt. L'estimateur de Horvitz-Thompson de t_y est donné par

$$\hat{t}_y = \sum_{i \in \mathcal{S}} w_i y_i. \tag{2.1}$$

L'estimateur $\hat{t}_y$ donné en (2.1) est également appelé estimateur à facteur d'extension les poids de sondage de base w_i indiquant le nombre d'unités de la population représentées par l'unité i dans l'échantillon. L'estimateur classique de la variance de $\hat{t}_y$ peut s'écrire

$$v = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \Omega_{ij} y_i y_j, \tag{2.2}$$

où $\Omega_{ij} = (\pi_{ij} - \pi_i \pi_j) / (\pi_{ij} \pi_i \pi_j)$ et $\pi_{ij} = P(i, j \in \mathcal{S})$ est la probabilité d'inclusion conjointe de deuxième ordre des unités (ij). On suppose que $\pi_{ij} > 0$ pour toutes les (ij). Notons que $\pi_{ii} = \pi_i$. L'estimateur classique de la variance v est souvent considéré comme entièrement efficace puisqu'il s'agit de l'estimateur de Horvitz-Thompson de la variance sous le plan $V(\hat{t}_y)$.

Soit $\boldsymbol{\Delta} = (\Omega_{ij})$ une matrice de dimensions $n \times n$. Nous pouvons réécrire (2.2) sous la forme $v = \boldsymbol{y}' \boldsymbol{\Delta} \boldsymbol{y}$, où $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ est le vecteur des y_i échantillonnés. La matrice $\boldsymbol{\Delta}$ est définie non négative et peut être décomposée comme il suit

$$\boldsymbol{\Delta} = \sum_{k=1}^{p} \lambda_k \boldsymbol{\delta}_k \boldsymbol{\delta}_k' \tag{2.3}$$

pour des valeurs de $\lambda_k > 0$ et des vecteurs n-dimensionnels $\boldsymbol{\delta}_k, k = 1, 2, \ldots, p$. La décomposition (2.3) la mieux connue est donnée par la décomposition spectrale où $\boldsymbol{\delta}_k$ est le vecteur propre associé à la valeur propre λ_k. En pratique, on ignore souvent les très petites valeurs propres pour des raisons de calcul. Pour l'échantillonnage stratifié, la matrice $\boldsymbol{\Delta}$ est diagonale par bloc, ce qui allège les calculs. Cependant, nous ne limitons pas (2.3) à la décomposition spectrale. Toute décomposition qui satisfait l'expression (2.3) peut être utilisée.

Supposons que nous voulions exprimer l'estimateur de la variance entièrement efficace v donné par (2.2) comme un estimateur de la variance par rééchantillonnage de la forme

$$v_R = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2, \tag{2.4}$$

où $\hat{t}_y^{(k)} = \sum_{i \in \mathcal{S}} w_i^{(k)} y_i, \boldsymbol{w}^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})'$ est le k^{e} jeu de poids de rééchantillonnage, $c_k > 0$ est un facteur associé avec le k^{e} jeu de poids de rééchantillonnage et L est le nombre total de répliques; voir Kim, Navarro et Fuller (2006) pour une discussion plus approfondie.

La forme donnée par (2.4) n'inclut pas tous les estimateurs de variance par rééchantillonnage. Par exemple, Campbell (1980) a donné un estimateur de variance par le jackknife où les pseudo

valeurs sont obtenues en se basant sur l'approximation de von Mises du paramètre d'intérêt. Néanmoins, la plupart des estimateurs de la variance par rééchantillonnage peuvent être mis sous cette forme.

Nous obtenons le résultat qui suit sur la construction de $\boldsymbol{w}^{(k)}$ pour v_R fondé sur la décomposition (2.3).

Théorème 1. L'estimateur de la variance entièrement efficace v et l'estimateur de la variance par rééchantillonnage v_R sont algébriquement identiques si nous posons que $L = p$ et $\boldsymbol{w}^{(k)} = \boldsymbol{w} + (\lambda_k / c_k)^{1/2}\boldsymbol{\delta}_k$, où $\boldsymbol{w} = (w_1, \ldots, w_n)'$ est le jeu de poids de sondage de base original.

Preuve. La preuve découle directement du fait que $v = \boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y} = \sum_{k=1}^{p} \lambda_k (\boldsymbol{\delta}_k'\boldsymbol{y})^2$ et que $\hat{t}_y^{(k)} - \hat{t}_y = (\boldsymbol{w}^{(k)} - \boldsymbol{w})'\boldsymbol{y} = (\lambda_k / c_k)^{1/2}\boldsymbol{\delta}_k'\boldsymbol{y}$.

Le choix des c_k peut être arbitraire et n'a aucune incidence sur la validité ni sur l'efficacité des estimateurs de la variance par rééchantillonnage. Cependant, certains choix de c_k produiront des poids de rééchantillonnage négatifs, ce qui n'est pas souhaitable, car cela peut produire des répliques négatives pour des paramètres qui sont toujours positifs. Dans les situations pratiques, on peut toujours choisir une valeur de c_k relativement grande pour éviter d'obtenir des valeurs négatives des poids de rééchantillonnage. Dans notre étude par simulation (*cas I*) présentée à la section 5, le problème des poids de rééchantillonnage négatifs peut être éliminé en choisissant $c_k = 1$.

L'estimateur de la variance par rééchantillonnage $v_R = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2$ avec $L = p$ et les poids de rééchantillonnage $\boldsymbol{w}^{(k)} = \boldsymbol{w} + (\lambda_k / c_k)^{1/2}\boldsymbol{\delta}_k$ est entièrement efficace pour une variable arbitraire y. Il fournit aussi un estimateur de la variance entièrement efficace pour $\hat{\theta}$ quand θ est une fonction lisse des moyennes ou des totaux de population. La mise en œuvre pratique de la méthode dépend de manière cruciale de deux problèmes apparentés : i) la faisabilité de la décomposition de la matrice $\boldsymbol{\Delta}$ de dimensions $n \times n$ spécifiée dans (2.3), et ii) le

nombre de jeux de poids de rééchantillonnage requis pour obtenir l'efficacité complète déterminée par $p = \text{rank}(\boldsymbol{\Delta})$.

En ce qui concerne le premier problème, les progrès contemporains en matière de puissance de calcul et l'amélioration de la performance des progiciels disponibles rendent possible la décomposition spectrale pour une valeur de n relativement grande. Par exemple, sur une machine unix à 12 CPU avec 96 gigaoctets de mémoire, la fonction R *eigen*() peut traiter des matrices de taille au moins aussi grande que $n = 4\,000$. Notons que la tâche de calcul dont il est question ici concerne les réalisateurs d'enquêtes à l'étape de la préparation des données et non les utilisateurs des fichiers de données. En ce qui concerne le deuxième problème, la valeur de p est reliée au plan d'échantillonnage donné. Pour un échantillonnage aléatoire simple sans remise, nous avons

$$\boldsymbol{\Delta} = N^2(1 - n / N)(n(n-1))^{-1}(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n' / n),$$

où $\boldsymbol{I}_n$ est la matrice identité de dimensions $n \times n$ et $\mathbf{1}_n = (1,1,\ldots,1)'$ est le vecteur de 1 de dimension $n \times 1$. Il s'ensuit que $p = \text{rank}(\Delta) = \text{trace}(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n' / n) = n - 1$. Il en est habituellement ainsi pour les plans d'échantillonnage avec probabilités inégales à un seul degré. Pour l'échantillonnage aléatoire simple stratifié, nous avons $p = n - H$, où H est le nombre total de strates.

Il convient de souligner que $p \le n$ pour tout plan d'échantillonnage et que la valeur exacte de p n'est pas requise pour mettre en œuvre la procédure proposée. Cependant, puisque les valeurs de p et de n sont du même ordre de grandeur, la méthode proposée nécessite un grand nombre de répliques quand n est grand. Dans le contexte des pratiques courantes en sondage, les poids de rééchantillonnage entièrement efficaces décrits plus haut deviennent directement applicables si $p \le 500$ et que les probabilités d'inclusion de deuxième ordre π_{ij} sont disponibles. Quand p est grand, une procédure en deux étapes, qui sera décrite à la section 3, peut être suivie pour produire un petit nombre L_0 de jeux de poids de rééchantillonnage pour les fichiers de données à grande diffusion.

Dans certains cas, on peut éviter la décomposition spectrale (2.3). Par exemple, Deville (1999) soutient que l'estimateur de la variance de $\hat{t}_y$ sous échantillonnage avec probabilités inégales lorsque la taille d'échantillon est fixe peut être approximé par

$$v \doteq c\sum_{i\in\mathcal{S}}(1-\pi_i)\left(\frac{y_i}{\pi_i}-\tilde{t}_y\right)^2 \tag{2.5}$$

où $c = \left(1-\sum_{i\in\mathcal{S}} b_i^2\right)^{-1}, b_i = (1-\pi_i)\big/\sum_{k\in\mathcal{S}}(1-\pi_k)$ et $\tilde{t}_y = \sum_{i\in\mathcal{S}} b_i(y_i / \pi_i)$. De manière plus générale, nous considérons la forme suivante de matrice $\boldsymbol{\Delta}$ dans $v = \boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y}$, où

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}_0 - \boldsymbol{\Delta}_0\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Delta}_0 \tag{2.6}$$

où $\boldsymbol{\Delta}_0 = \text{diag}\{\lambda_1,\ldots,\lambda_n\}, \lambda_i > 0$ pour tout $i = 1,2,\ldots,n, \boldsymbol{X}' = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ et $\boldsymbol{x}_i$ est un vecteur de variables de plan et de variables auxiliaires. De nombreux plans d'échantillonnage à un seul degré élémentaire prennent la forme (2.6) pour l'estimation de la variance. En particulier, la formule de Deville (2.5) peut être exprimée comme $v \doteq \boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y}$ avec $\boldsymbol{\Delta}$ donnée par (2.6), où $\lambda_i = c\pi_i^{-2}(1-\pi_i)$ dans $\boldsymbol{\Delta}_0$ et $\boldsymbol{x}_i = \pi_i$. Le plan d'échantillonnage de Poisson conditionnel dont nous discuterons à la section 5 prend aussi la forme (2.6) où les $\boldsymbol{x}_i$ sont les variables de plan dans la contrainte de plan $\sum_{i\in\mathcal{S}}\pi_i^{-1}\boldsymbol{x}_i = \sum_{i=1}^N \boldsymbol{x}_i$.

Pour la matrice donnée par (2.6), on peut montrer que

$$\boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y} = (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Delta}_0(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

où $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{y}$. Donc, nous avons

$$\boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y} = \sum_{k=1}^n \lambda_k(y_k - \boldsymbol{x}_k'\hat{\boldsymbol{\beta}})^2, \tag{2.7}$$

qui est utile pour obtenir une expression pour l'estimateur de la variance par rééchantillonnage de la forme donnée par (2.4). L'estimateur de la variance entièrement efficace v dans (2.7) et l'estimateur de la variance par rééchantillonnage v_R dans (2.4) sont algébriquement identiques si nous prenons $L = n$ et $\boldsymbol{w}^{(k)} = \boldsymbol{w} + (\lambda_k / c_k)^{1/2}\boldsymbol{\delta}_k$, où $\boldsymbol{w} = (w_1,\ldots,w_n)'$ est le jeu de poids de sondage de base original et $\boldsymbol{\delta}_k = (\delta_{1k},\ldots,\delta_{nk})'$ avec

$$\delta_{ik} = \begin{cases} -1 + \boldsymbol{x}_k'(\boldsymbol{X}'\Delta_0\boldsymbol{X})^{-1}\boldsymbol{x}_i\lambda_i & \text{si } i = k \\ \boldsymbol{x}_k'(\boldsymbol{X}'\Delta_0\boldsymbol{X})^{-1}\boldsymbol{x}_i\lambda_i & \text{autrement.} \end{cases}$$

La preuve découle directement du fait que $\delta_k'\boldsymbol{y} = -y_k + \boldsymbol{x}_k'\hat{\boldsymbol{\beta}}, \boldsymbol{y}'\Delta\boldsymbol{y} = \sum_{k=1}^{n}\lambda_k(\delta_k'\boldsymbol{y})^2$ et que $\hat{t}_y^{(k)} - \hat{t}_y = (\boldsymbol{w}^{(k)} - \boldsymbol{w})'\boldsymbol{y} = (\lambda_k / c_k)^{1/2}\delta_k'\boldsymbol{y}$.

3 Poids de rééchantillonnage parcimonieux et efficaces

Les enquêtes complexes à grande échelle sont habituellement réalisées auprès d'un échantillon d'assez grande taille variant de quelques centaines à plusieurs milliers d'unités. Les poids de rééchantillonnage entièrement efficaces décrits à la section 2 ou les poids de rééchantillonnage construits par certaines méthodes existantes, telles que les méthodes du jackknife ou du bootstrap, nécessiteraient un très grand nombre de jeux de poids. Bien qu'il soit très commode pour les utilisateurs des données d'enquête, qui ne sont pas nécessairement les réalisateurs des enquêtes, de disposer de poids de rééchantillonnage valides, la manipulation d'un ensemble de données avec des centaines, voire des milliers de poids de rééchantillonnage peut représenter une tâche énorme. Par conséquent, la question de savoir comment obtenir une estimation efficace de la variance par rééchantillonnage à l'aide d'un nombre assez petit de poids de rééchantillonnage est importante en théorie ainsi qu'en pratique.

Nous proposons deux stratégies pour construire des poids de rééchantillonnage parcimonieux et efficaces. Nous partons d'un grand nombre L de jeux de poids de rééchantillonnage. Ces poids initiaux peuvent être produits par la méthode générale décrite à la section 2 ou par des méthodes existantes. Supposons qu'ils peuvent être considérés comme entièrement efficaces. La première stratégie consiste à sélectionner un petit nombre L_0 de jeux de poids à partir du grand nombre initial L de jeux de poids en utilisant une méthode d'échantillonnage probabiliste. Le petit nombre L_0 satisfait la propriété souhaitée de *parcimonie*, tandis que la méthode de sélection aléatoire garantit la *validité* des estimateurs de la variance résultants. La deuxième stratégie

consiste à atteindre le niveau souhaité d'*efficacité* au moyen d'une nouvelle méthode de calage des poids. Les L_0 jeux de poids de rééchantillonnage calés fournissent des estimateurs de la variance entièrement efficaces pour les variables utilisées dans le calage et des estimateurs de la variance hautement efficaces pour les variables reliées aux variables de calage.

3.1 Réalisation de la parcimonie et de l'efficacité par échantillonnage aléatoire

Supposons que l'estimateur de la variance par rééchantillonnage entièrement efficace soit donné par $v_R = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2$, avec les poids de rééchantillonnage construits en utilisant le théorème 1. Observons que v_R peut être considéré comme un total de population finie. Si nous voulons utiliser $L_0 (< L)$ jeux de poids de rééchantillonnage pour fournir une inférence valide pour l'estimation de la variance, nous pouvons utiliser la simple stratégie suivante. Premièrement, sélectionner L_0 jeux de poids parmi les L jeux de poids originaux par échantillonnage aléatoire simple sans remise. Pour simplifier la notation et sans perte de généralité, nous désignons les jeux de poids sélectionnés par $\boldsymbol{w}^{(j)}, j = 1, 2, \ldots, L_0$. Puis, calculer l'estimateur de la variance par échantillonnage de $\hat{t}_y$ fondé sur les L_0 jeux de poids comme étant

$$v_R^{(1)} = \frac{L}{L_0} \sum_{j=1}^{L_0} c_j (\hat{t}_y^{(j)} - \hat{t}_y)^2. \tag{3.1}$$

L'estimateur de la variance $v_R^{(1)}$ est encore sans biais pour une variable arbitraire y, puisque $E^*(v_R^{(1)}) = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2 = v_R$, où $E^*(\cdot)$ désigne l'espérance sous la sélection aléatoire de L_0 jeux de poids.

Une forme de rechange de l'estimateur de la variance par rééchantillonnage fondé sur les L_0 jeux de poids peut être obtenue comme il suit. En notant que $\hat{t}_y^{(k)} - \hat{t}_y = (\lambda_k / c_k)^{1/2} \boldsymbol{\delta}_k' \boldsymbol{y}$, nous pouvons réécrire l'estimateur de la variance entièrement efficace sous la forme

$$v_R = m \left\{ \sum_{k=1}^{L} \lambda_k (\boldsymbol{\delta}_k' \boldsymbol{y})^2 \Big/ \sum_{k=1}^{L} \lambda_k \right\},$$

où $m = \sum_{k=1}^{L} \lambda_k$. Les δ_k sont des vecteurs propres orthogonaux satisfaisant la contrainte $|\delta_k| = 1$ sous la décomposition spectrale et les $\delta'_k \boldsymbol{y}$ sont des projections de $\boldsymbol{y}$ sur la sphère unitaire n-dimensionnelle. Il est très naturel d'utiliser la version pondérée pour l'estimateur de la variance de $\hat{t}_y$ fondé sur les L_0 jeux de poids sélectionnés aléatoirement :

$$v_R^{(2)} = m \left\{ \sum_{j=1}^{L_0} \lambda_j (\delta'_j \boldsymbol{y})^2 \Big/ \sum_{j=1}^{L_0} \lambda_j \right\} = \frac{m}{m_0} \sum_{j=1}^{L_0} c_j (\hat{t}_y^{(j)} - \hat{t}_y)^2, \tag{3.2}$$

où $m_0 = \sum_{j=1}^{L_0} \lambda_j$. Notons que $v_R^{(2)}$ est un estimateur par le ratio de v_R, qui est habituellement plus efficace que $v_R^{(1)}$.

Une troisième version de l'estimateur de la variance par rééchantillonnage peut être construite en commençant par sélectionner L_0 jeux de poids avec probabilités inégales, puis en utilisant un estimateur de Horvitz-Thompson de v_R. Mentionnons que nous pouvons considérer $v_R = \sum_{k=1}^{L} \lambda_k (\delta'_k \boldsymbol{y})^2$ comme un total de population et λ_k comme une variable de taille. Nous sélectionnons L_0 parmi les L jeux de poids originaux avec les probabilités d'inclusion η_k proportionnelles à λ_k. L'estimateur par la variance résultant de $\hat{t}_y$ est donné par

$$v_R^{(3)} = \sum_{j=1}^{L_0} \frac{c_j}{\eta_j} (\hat{t}_y^{(j)} - \hat{t}_y)^2, \tag{3.3}$$

où $\eta_j = L_0 \lambda_j \Big/ \sum_{k=1}^{L} \lambda_k$. Il s'avère que la grandeur des valeurs propres λ_k diffère considérablement et que l'utilisation de λ_k comme mesure de taille aboutit à la sélection d'une forte proportion des L jeux de poids avec une probabilité de un. Dans l'étude par simulation décrite à la section 5, nous avons également inclus une quatrième version de l'estimateur de la variance par rééchantillonnage, désignée $v_R^{(4)}$, avec $\lambda_k^{1/2}$ comme mesure de taille et $\eta_j = L_0 \lambda_j^{1/2} \Big/ \sum_{k=1}^{L} \lambda_k^{1/2}$.

Une autre version possible de l'estimateur de la variance par rééchantillonnage consiste à sélectionner simplement L_0 jeux de poids correspondant aux L_0 valeurs les plus grandes de λ_k,

puis à utiliser $v_R^{(2)}$. Les résultats de simulation, qui ne sont pas présentés ici, montrent que l'estimateur de la variance résultant est gravement biaisé et ne doit pas être utilisé en pratique.

3.2 Réalisation de la parcimonie et de l'efficacité par calage des poids

Nous discutons maintenant d'une nouvelle approche en vue d'arriver à la parcimonie sans perte d'efficacité des estimateurs de la variance pour certaines variable clés. Supposons que L_0 est la taille de réplique souhaitée, qui est beaucoup plus faible que la taille de l'échantillon n. Par exemple, le programme de la Natural Resources Inventory Survey (parrainé par le US department of Agriculture) a utilisé $L_0 = 29$, alors que la taille de l'échantillon peut être aussi grande que $n = 300\,000$ des UPÉ. Nous présentons une méthode de calage des poids qui permet non seulement d'utiliser une petite taille L_0, mais aussi de fournir des estimateurs de la variance entièrement efficaces pour les paramètres de population importants. La stratégie que nous proposons pour construire le plus petit nombre L_0 de jeux de poids de rééchantillonnage comprend les quatre étapes suivantes :

Étape 1. Choisir un ensemble de variables clés pour lesquelles on souhaite un estimateur de la variance entièrement efficace. Soit $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})'$ le vecteur de variables clés pour la i^{e} unité incluse dans le fichier de données d'enquête, où $m \leq L_0$. Les variables clés peuvent comprendre des variables auxiliaires importantes et les variables étudiées, ainsi que des variables de plan. Soit $\hat{t}_{\boldsymbol{z}} = \sum_{i \in \mathcal{S}} w_i \boldsymbol{z}_i$. Soit $v_1(\hat{t}_{\boldsymbol{z}})$ une matrice de variance-covariance de dimensions $m \times m$ estimée pour $\hat{t}_{\boldsymbol{z}}$ calculée par la méthode classique de linéarisation ou par une méthode de rééchantillonnage qui est entièrement efficace.

Étape 2. Construire un nombre initial L_0 de jeux de poids de rééchantillonnage qui produisent un estimateur de la variance asymptotiquement sans biais. Ces répliques initiales peuvent être obtenue par une méthode du bootstrap avec L_0 répliques ou par la méthode du jackknife avec suppression d'un groupe de Kott (2001) ou par la méthode d'échantillonnage décrite à la

section 3.1. Soit $\boldsymbol{w}_0^{(k)} = (w_{10}^{(k)}, \ldots, w_{n0}^{(k)})', k = 1, \ldots, L_0$ les ensembles initiaux de poids. Désignons $\hat{t}_{y0}^{(k)} = \sum_{i \in \mathcal{S}} w_{i0}^{(k)} y_i$ et soit

$$v_0 = \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{y0}^{(k)} - \hat{t}_y)^2 \tag{3.4}$$

l'estimateur de la variance par rééchantillonnage fondé sur les L_0 jeux de poids.

Nous pouvons appliquer la formule de la variance (3.4) au vecteur de variables clés $\boldsymbol{z}$ pour obtenir $v_0(\hat{t}_{\boldsymbol{z}}) = \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{\boldsymbol{z}0}^{(k)} - \hat{t}_{\boldsymbol{z}})(\hat{t}_{\boldsymbol{z}0}^{(k)} - \hat{t}_{\boldsymbol{z}})'$, où $\hat{t}_{\boldsymbol{z}0}^{(k)} = \sum_{i \in \mathcal{S}} w_{i0}^{(k)} \boldsymbol{z}_i$. Soulignons que $v_0(\hat{t}_{\boldsymbol{z}})$ n'est pas aussi efficace que $v_1(\hat{t}_{\boldsymbol{z}})$ obtenu à l'*étape* 1.

Étape 3. Décomposer la matrice de variance-covariance définie non négative $v_1(\hat{t}_{\boldsymbol{z}})$ comme

$$v_1(\hat{t}_{\boldsymbol{z}}) = \sum_{k=1}^{m} \alpha_k \boldsymbol{q}_k \boldsymbol{q}_k' \tag{3.5}$$

en utilisant la décomposition spectrale ou toute autre méthode appropriée. Poser que $\alpha_k = 0$ pour $k = m + 1, \ldots, L_0$ et définir

$$\hat{t}_{\boldsymbol{z}}^{(k)} = \hat{t}_{\boldsymbol{z}} + (\alpha_k / c_{k0})^{1/2} \boldsymbol{q}_k, k = 1, 2, \ldots, L_0.$$

Il s'ensuit que les L_0 pseudo répliques $\hat{t}_{\boldsymbol{z}}^{(k)}$ définie ci-dessus satisfont

$$\sum_{k=1}^{L_0} c_{k0} (\hat{t}_{\boldsymbol{z}}^{(k)} - \hat{t}_{\boldsymbol{z}})(\hat{t}_{\boldsymbol{z}}^{(k)} - \hat{t}_{\boldsymbol{z}})' = v_1(\hat{t}_{\boldsymbol{z}}), \tag{3.6}$$

en raison de la décomposition de $v_1(\hat{t}_{\boldsymbol{z}})$ donnée par l'expression (3.5). Il convient de souligner qu'il n'existe aucun lien entre l'expression (3.5) et la décomposition de $\boldsymbol{\Delta}$ décrite à la section 2 et que la condition $m \leq L_0$ est requise pour que (3.6) soit possible.

Étape 4. Améliorer l'efficacité de v_0 calculé d'après (3.4) pour une variable y arbitraire par une méthode de calage des poids. Pour le k^{e} jeu de poids initiaux $\boldsymbol{w}_0^{(k)} = (w_{10}^{(k)}, \ldots, w_{n0}^{(k)})'$, les poids calés $\boldsymbol{w}_c^{(k)} = (w_{1c}^{(k)}, \ldots, w_{nc}^{(k)})'$ minimisent la mesure de la distance du khi-carré

$$\Phi(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)}) = \sum_{i \in \mathcal{S}_k} \tau_i (w_{ic}^{(k)} - w_{i0}^{(k)})^2 \ / \ w_{i0}^{(k)} \tag{3.7}$$

sous la contrainte

$$\sum_{i \in \mathcal{S}} w_{ic}^{(k)} \boldsymbol{z}_i = \hat{t}_{\boldsymbol{z}}^{(k)}, \tag{3.8}$$

où $\mathcal{S}_k = \left\{ i \middle| i \in \mathcal{S}; w_{i0}^{(k)} > 0 \right\}$, les τ_i sont des constantes connues et $\hat{t}_{\boldsymbol{z}}^{(k)}$ est la k^{e} pseudo réplique de $\hat{t}_{\boldsymbol{z}}$ définie à l'*étape* 3.

Les poids calés $\boldsymbol{w}_c^{(k)} = (w_{1c}^{(k)}, \ldots, w_{nc}^{(k)})', k = 1, 2, \ldots, L_0$ sont utilisés dans (3.4) pour calculer l'estimateur de la variance par rééchantillonnage final $v_C(\hat{t}_y) = \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{yc}^{(k)} - \hat{t}_y)^2$, où $\hat{t}_{yc}^{(k)} = \sum_{i \in \mathcal{S}} w_{ic}^{(k)} y_i$.

La méthode de calage des poids proposée fait en sorte que l'estimateur par rééchantillonnage $v_C(\hat{t}_{\boldsymbol{z}})$ fondé sur les L_0 jeux de poids calés correspond exactement à l'estimateur entièrement efficace $v_1(\hat{t}_{\boldsymbol{z}})$, en raison des contraintes de calage (3.8) et de l'équation (3.6). De surcroît, les poids de rééchantillonnage calés fournissent des estimateurs plus efficaces de la variance pour une variable y arbitraire qui est reliée à $\boldsymbol{z}$. Pour le montrer, nous réécrivons $\hat{t}_{y0}^{(k)} = \sum_{i \in \mathcal{S}} w_{i0}^{(k)} y_i$ sous la forme

$$\hat{t}_{y0}^{(k)} = \hat{t}_{e0}^{(k)} + (\hat{t}_{\boldsymbol{z}0}^{(k)})' \hat{\boldsymbol{\beta}}, \tag{3.9}$$

où $\hat{t}_{e0}^{(k)} = \sum_{i \in \mathcal{S}} w_{i0}^{(k)} \hat{e}_i, \hat{e}_i = y_i - \boldsymbol{z}_i' \hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}} = \left\{ \sum_{i \in \mathcal{S}} w_i \boldsymbol{z}_i \boldsymbol{z}_i' \big/ \tau_i \right\}^{-1} \sum_{i \in \mathcal{S}} w_i \boldsymbol{z}_i y_i \big/ \tau_i$. Soit $\hat{t}_e = \sum_{i \in \mathcal{S}} w_i \hat{e}_i$. L'estimateur de la variance de $\hat{t}_y$ fondé sur les L_0 jeux de poids initiaux peut s'exprimer

$$\begin{aligned} v_0(\hat{t}_y) &= \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{y0}^{(k)} - \hat{t}_y)^2 \\ &= \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{e0}^{(k)} - \hat{t}_e)^2 + \sum_{k=1}^{L_0} c_{k0} ((\hat{t}_{\boldsymbol{z}0}^{(k)})' \hat{\boldsymbol{\beta}} - (\hat{t}_{\boldsymbol{z}})' \hat{\boldsymbol{\beta}})^2 + 2 \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{e0}^{(k)} - \hat{t}_e)((\hat{t}_{\boldsymbol{z}0}^{(k)})' \hat{\boldsymbol{\beta}} - (\hat{t}_{\boldsymbol{z}})' \hat{\boldsymbol{\beta}}) \\ &= v_0(\hat{t}_e) + \hat{\boldsymbol{\beta}}' v_0(\hat{t}_{\boldsymbol{z}}) \hat{\boldsymbol{\beta}} + 2 \hat{\boldsymbol{\beta}}' \operatorname{cov}_0(\hat{t}_e, \hat{t}_{\boldsymbol{z}}), \end{aligned}$$

où $\mathrm{cov}_0(\hat{t}_e, \hat{t}_{\boldsymbol{z}})$ est la covariance estimée entre $\hat{t}_e$ et $\hat{t}_{\boldsymbol{z}}$ fondée sur les L_0 jeux de poids de rééchantillonnage initiaux. Dans de nombreux plans, nous pouvons choisir un τ_i approprié pour que $\mathrm{Cov}(\hat{t}_e, \hat{t}_{\boldsymbol{z}}) \doteq \mathbf{0}$. Il en est ainsi, par exemple, si l'on choisit $\tau_i = (w_i - 1)^{-1}$ ou $\tau_i = w_i^{-1}$ sous l'échantillonnage de Poisson. Fuller (1998) a discuté des conditions requises dans le contexte de l'échantillonnage à deux phases. Il s'ensuit que

$$v_0(\hat{t}_y) \doteq v_0(\hat{t}_e) + \hat{\boldsymbol{\beta}}' v_0(\hat{t}_{\boldsymbol{z}}) \hat{\boldsymbol{\beta}}. \tag{3.10}$$

En utilisant un argument comparable, on peut montrer que l'estimateur de la variance fondé sur les L_0 jeux de poids calés satisfait

$$v_C(\hat{t}_y) \doteq v_0(\hat{t}_e) + \hat{\boldsymbol{\beta}}' v_1(\hat{t}_{\boldsymbol{z}}) \hat{\boldsymbol{\beta}}. \tag{3.11}$$

L'estimateur de la variance $v_C(\hat{t}_y)$ donné par (3.11) est généralement plus efficace que $v_0(\hat{t}_y)$ donné par (3.10), en raison de l'utilisation de $v_1(\hat{t}_{\boldsymbol{z}})$ plutôt que $v_0(\hat{t}_{\boldsymbol{z}})$. Le gain d'efficacité dépend de la grandeur relative de $\hat{\boldsymbol{\beta}}' v_1(\hat{t}_{\boldsymbol{z}}) \hat{\boldsymbol{\beta}}$ par rapport à $v_0(\hat{t}_e)$. Si la variable y est fortement corrélée avec $\hat{y} = \boldsymbol{z}' \hat{\boldsymbol{\beta}}$, la variance du terme résiduel $v_0(\hat{t}_e)$ sera relativement faible. Dans ce cas, $v_C(\hat{t}_y)$ sera très efficace. Par ailleurs, s'il n'y a aucune corrélation entre y et $\hat{y}$, aucune amélioration ne sera obtenue en utilisant les poids calés $\boldsymbol{w}_c^{(k)}$; voir aussi le théorème 3 à la section 4.

L'un des inconvénients de la distance du khi-carré $\Phi(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)})$ à l'*étape* 4 est que certains poids calés résultants pourraient prendre des valeurs négatives. Pour éviter les poids négatifs, nous proposons de remplacer la distance du khi-carré dans (3.7) par la « distance d'entropie » minimale suivante

$$D(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)}) = -\sum_{i \in \mathcal{S}_k} \tau_i^{-1} \left\{ w_{i0}^{(k)} \log\left(\frac{w_{ic}^{(k)}}{w_{i0}^{(k)}} \right) - w_{ic}^{(k)} + w_{i0}^{(k)} \right\} \tag{3.12}$$

pour deux raisons. Premièrement, il est garanti que les poids calés $w_{ic}^{(k)}$ seront positifs. Deuxièmement, il existe un algorithme de calcul se comportant bien pour ce problème de

minimisation sous contraintes. On peut montrer que les $\boldsymbol{w}_c^{(k)}$ qui minimisent $D(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)})$ sous les contraintes (3.8) sont donnés par

$$w_{ic}^{(k)} = \frac{w_{i0}^{(k)} \ / \ \tau_i}{1 + \boldsymbol{\lambda}'\mathbf{z}_i}, \tag{3.13}$$

où le multiplicateur de Lagrange $\boldsymbol{\lambda}$ est la solution de

$$g(\boldsymbol{\lambda}) = \sum_{i \in \mathcal{S}} \frac{w_{i0}^{(k)} \mathbf{z}_i \ / \ \tau_i}{1 + \boldsymbol{\lambda}'\mathbf{z}_i} - \hat{t}_{\mathbf{z}}^{(k)} = \mathbf{0}. \tag{3.14}$$

Un algorithme de calcul efficace pour trouver la solution $\boldsymbol{\lambda}$ de (3.14) se trouve dans Wu (2004) et une fonction R apparentée peut être obtenue par une modification mineure de la fonction R présentée dans Wu (2005).

4 Validité

À la présente section, nous présentons une discussion générale de la validité de l'estimateur de la variance par rééchantillonnage. Soit $\theta = f(t_y)$ un paramètre de population finie, qui est une fonction lisse du total de population $t_y = \sum_{i=1}^{N} y_i$. Nous supposons que $\hat{\theta} = f(\hat{t}_y)$ est utilisé pour estimer θ, où $\hat{t}_y$ est l'estimateur de Horvitz-Thompson de t_y défini en (2.1). L'estimateur de la variance par rééchantillonnage de $\hat{\theta}$ est formé par

$$v_R(\hat{\theta}) = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2, \tag{4.1}$$

où $\hat{\theta}^{(k)} = f(\hat{t}_y^{(k)})$ et $\hat{t}_y^{(k)}$ est la k^{e} réplique de $\hat{t}_y$.

Pour explorer les propriétés asymptotiques de l'estimateur de la variance par rééchantillonnage (4.1), nous émettons l'hypothèse d'une série de populations finies et d'échantillons, telle que décrite dans Isaki et Fuller (1982). Les populations finies et les plans d'échantillonnage satisfont les conditions de régularité qui suivent.

C1. Pour toute caractéristique de la population $\boldsymbol{u}_i$ avec moment d'ordre deux borné,

$$\sum_{i \in \mathcal{S}} w_i \boldsymbol{u}_i \boldsymbol{u}_i' - \sum_{i=1}^{N} \boldsymbol{u}_i \boldsymbol{u}_i' = O_p(n^{-1/2} N).$$

C2. Les poids de sondage sont uniformément bornés. Autrement dit, $K_1 < N^{-1} n w_i < K_2$ pour tout i et tout n, où K_1 et K_2 sont des constantes fixes.

C3. $nV(N^{-1}\hat{t}_y)$ est borné.

C4. Pour tout y avec moment d'ordre quatre borné, l'estimateur de la variance par rééchantillonnage $v_R(\hat{t}_y) = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2$ satisfait

$$E[\{c_k(\hat{t}_y^{(k)} - \hat{t}_y)^2\}^2] < KL^{-2}\{V(\hat{t}_y)\}^2 \tag{4.2}$$

pour un certain K, uniformément en $k = 1, \ldots, L$,

$$\max_k c_k^{-1} = O(L), \tag{4.3}$$

et

$$E\left[\left\{\frac{v_R(\hat{t}_y)}{V(\hat{t}_y)} - 1\right\}^2\right] = o(1). \tag{4.4}$$

La condition (4.2) fait en sorte qu'aucune réplique particulière ne domine les autres. La condition (4.3) contrôle l'ordre des facteurs c_k. La condition (4.4) implique que $v_R(\hat{t}_y)$ est un estimateur convergent de $V(\hat{t}_y)$. Les conditions (4.2) à (4.4) ont également été utilisées dans Kim, Navarro et Fuller (2006).

En utilisant les conditions de régularité susmentionnées, le théorème suivant prouve la convergence de l'estimateur de la variance par rééchantillonnage de la forme (4.1).

Théorème 2. Soit $\theta = f(t_y)$ le paramètre d'intérêt et $\hat{\theta} = f(\hat{t}_y)$, où $f(\cdot)$ est une fonction lisse avec une dérivée continue à t_y. Sous les conditions de régularité susmentionnées, l'estimateur de la variance $v_R(\hat{\theta})$ dans (4.1) satisfait

$$\frac{v_R(\hat{\theta})}{V(\hat{\theta})} = 1 + o_p(1). \tag{4.5}$$

Preuve. Voir l'annexe A.

Nous prouvons maintenant la validité de l'estimateur amélioré de la variance $v_C(\hat{t}_y)$ proposé à la section 3.2. Pour simplifier, nous supposons que $v_1(\hat{t}_y)$ est un estimateur entièrement efficace de la variance $V(\hat{t}_y)$ pour $\hat{t}_y = \sum_{i \in \mathcal{S}} w_i y_i$. Nous supposons aussi que $v_0(\hat{t}_y)$, défini en (3.4), satisfait

$$E^*\{v_0(\hat{t}_y)\} = v_1(\hat{t}_y), \tag{4.6}$$

où $E^*(\cdot)$ désigne l'espérance sous la sélection aléatoire des L_0 répliques pour les L jeux de poids de rééchantillonnage entièrement efficaces, comme il est discuté à la section 3.1. Si $v_1(\hat{t}_y)$ est asymptotiquement sans biais, alors $v_0(\hat{t}_y)$ est également asymptotiquement sans biais en vertu de (4.6). Pour le jackknife avec suppression d'un groupe, la condition (4.6) peut être interprétée comme $E\{v_0(\hat{t}_y)\} = E\{v_1(\hat{t}_y)\}$ et $V\{v_0(\hat{t}_y)\} \geq V\{v_1(\hat{t}_y)\}$.

Théorème 3. Supposons que l'estimateur de la variance initial $v_0(\hat{t}_y)$ défini en (3.4) satisfait (4.6). Supposons que l'estimateur amélioré de la variance $v_C(\hat{t}_y) = \sum_{k=1}^{L_0} c_{k0}(\hat{t}_{yc}^{(k)} - \hat{t}_y)^2$ est calculé en utilisant les poids de rééchantillonnage calés comme il est décrit à la section 3.2, avec le choix de τ_i satisfaisant $\mathrm{Cov}(\hat{t}_e, \hat{t}_{\mathbf{z}}) \doteq \mathbf{0}$. En ignorant les termes d'ordre faible, nous avons

$$E\{v_C(\hat{t}_y)\} = E\{v_1(\hat{t}_y)\} \tag{4.7}$$

et

$$V\{v_1(\hat{t}_y)\} \leq V\{v_C(\hat{t}_y)\} \leq V\{v_0(\hat{t}_y)\}. \tag{4.8}$$

Preuve. Voir l'annexe B.

Pour un paramètre général $\theta = f(t_y)$, nous posons que $\hat{\theta}_c^{(k)} = f(\hat{t}_{yc}^{(k)})$ et calculons $v_C(\hat{\theta}) = \sum_{k=1}^{L_0} c_{k0}(\hat{\theta}_c^{(k)} - \hat{\theta})^2$. La validité de $v_C(\hat{\theta})$ peut être établie en combinant les résultats des théorèmes 2 et 3.

5 Extension à certains plans d'échantillonnage équilibrés

Nous considérons maintenant des plans d'échantillonnage qui sont équilibrés en $\boldsymbol{x}_i$ au sens que $\hat{t}_{\boldsymbol{x}} = \sum_{i\in\mathcal{S}} \boldsymbol{x}_i / \pi_i = t_{\boldsymbol{x}}$ est vérifié exactement ou quasi exactement, où $\boldsymbol{x}_i$ est un vecteur de dimension q et $t_{\boldsymbol{x}} = \sum_{i\in\mathcal{U}} \boldsymbol{x}_i$ est connu. Nous supposons que le premier élément de $\boldsymbol{x}_i$ est égal à π_i, ce qui suppose implicitement qu'il s'agit d'un plan de sondage avec taille d'échantillon fixe. Tillé (2006) donne un exposé complet des plans d'échantillonnage équilibrés.

Deville et Tillé (2005) soutiennent que $\hat{t}_y = \sum_{i\in\mathcal{S}} y_i / \pi_i$ sous échantillonnage équilibré possède une variance qui peut être approximée par sa variance sous échantillonnage de Poisson conditionnel. En utilisant la même approximation, Breidt et Chauvet (2011) dérivent

$$v(\hat{t}_y) = \frac{n}{n-q} \sum_{i\in\mathcal{S}} (1-\pi_i) \left(\frac{y_i}{\pi_i} - \frac{\tilde{y}_i}{\pi_i} \right)^2, \tag{5.1}$$

où $\tilde{y}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_P$ et $\hat{\boldsymbol{\beta}}_P = \left\{ \sum_{i\in\mathcal{S}} (1-\pi_i)\pi_i^{-2} \boldsymbol{x}_i \boldsymbol{x}_i' \right\}^{-1} \sum_{i\in\mathcal{S}} (1-\pi_i)\pi_i^{-2} \boldsymbol{x}_i y_i$. Approximativement parlant, la formule de variance (5.1) peut être interprétée comme l'approximation de $\hat{t}_y$ sous le plan d'échantillonnage équilibré par un estimateur par la régression généralisée sous échantillonnage de Poisson. Autrement dit, $V(\hat{t}_y) \doteq V(\hat{t}_P)$, où $\hat{t}_P = \hat{t}_y + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}})' \hat{\boldsymbol{\beta}}_P$. Pour une justification formelle de cette approximation, voir Fuller (2009b).

La formule de variance (5.1) peut être utilisée pour obtenir les poids de rééchantillonnage. Pour le montrer, nous réexprimons (5.1) sous forme d'un estimateur de la variance par rééchantillonnage jackknife

$$v_J(\hat{t}_y) = \sum_{k=1}^{n} c_k (\tilde{t}_y^{(k)} - \hat{t}_y)^2, \tag{5.2}$$

où $\tilde{t}_y^{(k)} = \hat{t}_y^{(k)} + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}}^{(k)})' \hat{\boldsymbol{\beta}}_P^{(k)}, (\hat{t}_{\boldsymbol{x}}^{(k)}, \hat{t}_y^{(k)}) = \sum_{i\in\mathcal{S}^{(k)}} \pi_i^{-1} (\boldsymbol{x}_i', y_i)$,

$$\hat{\boldsymbol{\beta}}_P^{(k)} = \left\{ \sum_{i\in\mathcal{S}^{(k)}} (1-\pi_i)\pi_i^{-2} \boldsymbol{x}_i \boldsymbol{x}_i' \right\}^{-1} \sum_{i\in\mathcal{S}^{(k)}} (1-\pi_i)\pi_i^{-2} \boldsymbol{x}_i y_i,$$

$c_k = (1 - \pi_k)n / (n - q)$ et $\mathcal{S}^{(k)} = \mathcal{S} \setminus \{k\}$. Afin de montrer l'équivalence asymptotique entre (5.1) et (5.2), nous commençons par noter que

$$\tilde{t}_y^{(k)} - \hat{t}_y = (\hat{t}_y^{(k)} - \hat{t}_y) + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}}^{(k)})'\hat{\boldsymbol{\beta}}_P + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}}^{(k)})'(\hat{\boldsymbol{\beta}}_P^{(k)} - \hat{\boldsymbol{\beta}}_P).$$

Sous certaines conditions de régularité, nous avons $\hat{\boldsymbol{\beta}}_P^{(k)} = \hat{\boldsymbol{\beta}}_P + O_p(n^{-1})$ et $t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}}^{(k)} = \boldsymbol{x}_k / \pi_k = O_p(n^{-1}N)$. Ici, nous avons utilisé la condition $t_{\boldsymbol{x}} = \hat{t}_{\boldsymbol{x}}$ sous le plan d'échantillonnage équilibré. Il s'ensuit que $\tilde{t}_y^{(k)} - \hat{t}_y = -\pi_k^{-1}(y_k - \boldsymbol{x}_k'\hat{\boldsymbol{\beta}}_P) + O_p(n^{-2}N)$, et $v_J(\hat{t}_y)$ dans (5.2) est asymptotiquement équivalent à $\sum_{k=1}^n c_k \pi_k^{-2}(y_k - \tilde{y}_k)^2$, qui est égal à $v(\hat{t}_y)$ donné par (5.1). La formule de variance (5.2) est assez utile, parce qu'elle rend la construction des poids de rééchantillonnage relativement simple pour les plans d'échantillonnage équilibrés. Quand n est grand, le nombre de répliques peut être réduit en utilisant la méthode de calage des poids décrite à la section 3.2. Les résultats des simulations fondées sur l'échantillonnage de Poisson réjectif de Fuller (2009b), non présentés ici faute d'espace, montrent que l'estimateur de la variance par rééchantillonnage proposé donne de très bons résultats.

6 Étude par simulation

À la présente section, nous présentons les résultats d'une étude par simulation. Nous considérons une population finie synthétique de taille $N = 2\,248$ familles obtenue en utilisant un ensemble de données réelles provenant de l'Enquête sur les dépenses des familles de 2000 de Statistique Canada pour la province de l'Ontario. Pour la i^{e} famille sélectionnée, l'ensemble de données contient des observations sur plusieurs variables, dont x_{i1} : le nombre de personnes dans la famille; x_{i2} : le nombre d'enfants (âge < 15 ans); x_{i3} : le nombre de jeunes (âge de 15 à 24 ans); x_{i4} : le revenu annuel total après impôt; y_{i1} : les dépenses annuelles totales; y_{i2} : les dépenses annuelles en habillement; y_{i3} : les dépenses annuelles en meubles et équipement.

Nous considérons trois paramètres de population pour comparer les différentes versions de l'estimateur de variance par rééchantillonnage. Le premier est le total de population des dépenses annuelles globales, c'est-à-dire $\theta_1 = t_{y1} = \sum_{i=1}^N y_{i1}$. Le deuxième est le ratio des totaux de

population des dépenses en habillement aux dépenses en meubles et en équipement, c'est-à-dire $\theta_2 = t_{y2} / t_{y3} = \left(\sum_{i=1}^{N} y_{i2}\right) \Big/ \left(\sum_{i=1}^{N} y_{i3}\right)$. Mentionnons que $\theta_2 = \mu_{y2} / \mu_{y3}$. Le troisième est le coefficient de corrélation de population $\theta_3 = \rho(y_1, y_2)$ entre les dépenses annuelles globales (y_1) et les dépenses annuelles en habillement (y_2). Pour chaque paramètre, nous avons évalué par simulation plusieurs estimateurs de la variance par rééchantillonnage.

Nous examinons deux approches d'estimation de la variance par rééchantillonnage. Pour la première, les L jeux initiaux de poids de rééchantillonnage sont construits en utilisant la méthode générale décrite à la section 2. Pour la seconde, les $L = n$ jeux de poids de rééchantillonnage jackknife avec suppression d'un groupe classiques sont utilisés pour produire des estimateurs de la variance entièrement efficaces.

Cas I. Des échantillons sont sélectionnés avec probabilités inégales selon la méthode d'échantillonnage PPT de Rao-Sampford (Rao 1965; Sampford 1967), avec les probabilités d'inclusion π_i proportionnelles au revenu annuel total x_{i4}. L'une des caractéristiques intéressantes de la méthode de Rao-Sampford est que les probabilités d'inclusion d'ordre deux π_{ij} peuvent être calculées exactement. La méthode générale décrite à la section 2 est utilisée pour créer $L = n$ jeux de poids de rééchantillonnage entièrement efficaces, et l'estimateur de la variance correspondant est désigné par v_R. Ces poids sont utilisés comme fondement pour calculer et comparer diverses versions des estimateurs de variance $v_R^{(l)}, l = 1, 2, 3, 4$ décrits à la section 3.1, en se basant sur un plus petit nombre L_0 de jeux de poids. Nous contraignons L_0 d'être égal à 25 ou à 50.

L'estimateur de la variance par rééchantillonnage calé décrit à la section 3.2 est désigné par v_C. Les L_0 jeux de poids initiaux sont sélectionnés parmi les L jeux de poids originaux par échantillonnage aléatoire simple, et les variables $\boldsymbol{z}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, y_{i2}, y_{i3})'$ sont utilisées comme variables de calage. Dans ces conditions, le premier paramètre θ_1 n'est pas directement relié à $\boldsymbol{z}$, mais le deuxième paramètre θ_2 est défini comme une fonction non linéaire mais lisse

de $t_{\boldsymbol{z}}$. Le troisième paramètre $\rho(y_1, y_2)$ est plus complexe et fait intervenir des quantités de populations non incluses dans $t_{\boldsymbol{z}}$.

Cas II. La population de $N = 2\,248$ unités est d'abord reproduite dix fois, pour créer une population plus grande de 22 480 unités. Des échantillons aléatoires simples de taille $n = 100$, 200 ou 400 sont sélectionnés dans la population. Les fractions d'échantillonnage sont inférieures à 2 %. Sous de tels scénarios, les n jeux de poids jackknife avec suppression d'un groupe classiques fournissent un estimateur de la variance entièrement efficace v_J. Soit $v_J^{(1)}$ l'estimateur de la variance utilisant L_0 jeux de poids, sélectionnés aléatoirement parmi les n jeux de poids jackknife. Soit $v_J^{(C)}$ l'estimateur de la variance utilisant les L_0 jeux de poids ainsi que le calage sur les $\boldsymbol{z}$ variables.

Pour chaque échantillon simulé de taille n et un paramètre de population particulier θ, nous calculons l'estimateur fondé sur le plan de sondage $\hat{\theta}$ et diverses versions des estimateurs de la variance. Le processus est répété B fois, indépendamment, avec $B = 5\,000$ pour le *cas I* et $B =$ 10 000 pour le *cas II*. La vraie variance $V = V(\hat{\theta})$ est approximée par $V \doteq B^{-1}\sum_{b=1}^{B}(\hat{\theta}_b - \theta)^2$, où $\hat{\theta}_b$ est calculé d'après le b^{e} échantillon simulé, en utilisant un autre jeu indépendant de B échantillons simulés. Les résultats de la simulation montrent que le biais de $\hat{\theta}$ est négligeable pour les trois paramètres. Les propriétés d'un estimateur de la variance v sont mesurées par la probabilité de couverture simulée de l'intervalle de confiance à 95 % sous la loi normale, calculée comme $\text{PC} = B^{-1}\sum_{b=1}^{B} I[\hat{\theta}_b - 1,96(v_b)^{1/2} \le \theta \le \hat{\theta}_b + 1,96(v_b)^{1/2}]$, la longueur moyenne de l'intervalle $\text{LM} = B^{-1}\sum_{b=1}^{B} 2 \times 1,96(v_b)^{1/2}$ et la racine carrée relative de l'erreur quadratique moyenne (RCREQM), calculée comme $\text{RCREQM} = \{\text{EQM}(v)\}^{1/2} / V$, où v_b est l'estimateur de la variance v calculé pour le b^{e} échantillon simulé, et $\text{EQM}(v) = B^{-1}\sum_{b=1}^{B}(v_b - V)^2$.

Les probabilités de couverture simulées sont présentées aux tableaux 6.1 et 6.2. Les estimateurs de la variance entièrement efficaces v_R et v_J donnent une bonne couverture pour

tous les scénarios sauf pour $\rho(y_1, y_2)$ dans le *cas I*, où la couverture est un peu faible. Les estimateurs de la variance $v_R^{(l)}, l = 1, 2, 3, 4$ et $v_J^{(1)}$ fondés sur les L_0 jeux de poids semblent bien fonctionner pour θ_1, dans une certaine mesure pour θ_2 également, mais aucunement pour $\theta_3 = \rho(y_1, y_2)$. L'estimateur calé v_C a une couverture satisfaisante pour tous les scénarios pour le *cas I*. Comme pour l'estimateur calé $v_J^{(C)}$ avec le *cas II*, les résultats sont très bons pour θ_1 et θ_2, mais aucun estimateur ne fonctionne bien pour $\theta_3 = \rho(y_1, y_2)$.

Il convient de mentionner que la définition de $\rho(y_1, y_2)$ comprend les moyennes de population sur trois variables dérivées y_1^2, y_2^2 et $y_1 y_2$. Lorsque ces trois variables sont également incluses à l'étape du calage, en plus de $\boldsymbol{z}$, nous désignons l'estimateur de la variance résultant par $v_J^{(C+)}$ pour le *cas II*. Il se fait que $v_J^{(C+)}$ donne de nettement meilleurs résultats pour θ_3 et améliore également les résultats pour θ_1 et θ_2.

Est également présentée dans les tableaux 6.1 et 6.2 la longueur moyenne des intervalles de confiance en utilisant $v_C, v_J^{(C)}$ et $v_J^{(C+)}$. Les résultats (LM, entre parenthèses) sont donnés par rapport à la longueur de l'intervalle en utilisant v_R (tableau 6.1) ou v_J (tableau 6.2) avec une valeur (disons) de 1,05 indiquant un accroissement de 5 % de la longueur. On peut voir que les estimateurs de la variance calés produisent des intervalles de confiance qui sont soit de longueur comparable à celle des intervalles utilisant $v_R(v_J)$ ou légèrement plus larges, selon le paramètre et (ou) les tailles d'échantillon.

La racine carrée relative de l'erreur quadratique moyenne (RCREQM) des estimateurs de la variance est présentée aux tableaux 6.3 et 6.4. Les résultats semblent dépendre non seulement du paramètre et de son estimateur, mais aussi du plan d'échantillonnage et de la méthode de rééchantillonnage. Pour le *cas I*, l'estimateur de la variance v_C, qui est de tout premier intérêt, est plus stable que v_R pour θ_1, ou presque aussi stable pour θ_2, et moins stable pour θ_3. Comme y_{i1} est bien expliqué par $\boldsymbol{z}_i, v_C$ est assez efficace pour estimer la variance de $\hat{\theta}_1 = \hat{t}_{y1}$. Pour le *cas II*, $v_J^{(C)}$ et $v_J^{(C+)}$ sont comparables l'un à l'autre, mais sont tous deux sont moins stables que v_J.

Tableau 6.1
Probabilités de couverture des intervalles de confiance à 95 % (*cas I*)

θ	L_0	n	v_R	$v_R^{(1)}$	$v_R^{(2)}$	$v_R^{(3)}$	$v_R^{(4)}$	v_C (LM)
t_{y1}	25	50	93,9	92,9	93,1	92,4	93,1	94,3 (1,03)
		100	94,4	92,0	92,4	91,9	93,0	93,4 (1,01)
		150	95,1	91,5	91,9	92,1	93,2	93,7 (0,99)
	50	100	94,5	93,2	93,2	93,4	93,6	94,1 (1,01)
		150	95,1	93,0	93,3	93,5	93,8	94,5 (0,99)
μ_{y2} / μ_{y3}	25	50	92,6	91,0	91,2	90,6	91,0	92,9 (1,02)
		100	93,7	91,1	91,2	89,6	90,8	93,7 (1,01)
		150	94,3	91,1	90,7	89,5	90,8	94,3 (1,00)
	50	100	93,6	92,6	92,5	91,9	92,5	93,8 (1,01)
		150	94,2	92,7	92,6	91,9	92,9	94,3 (1,00)
$\rho(y_1, y_2)$	25	50	89,0	85,7	85,6	79,3	81,9	91,3 (1,14)
		100	90,5	85,4	85,3	78,6	81,5	92,1 (1,17)
		150	90,7	84,6	84,5	76,9	81,6	91,9 (1,17)
	50	100	90,4	88,2	88,2	83,2	85,8	92,9 (1,18)
		150	90,7	87,5	87,6	81,7	84,8	93,4 (1,18)

v_R : estimateur de la variance par rééchantillonnage entièrement efficace (section 2); $v_R^{(l)}$, $l = 1, 2, 3, 4$: estimateurs de la variance par rééchantillonnage fondés sur L_0 jeux de poids (section 3.1); v_C : estimateur de la variance par rééchantillonnage fondé sur L_0 jeux de poids calés (section 3.2); LM : longueur moyenne de l'intervalle de confiance comparativement à celui utilisant v_R.

Tableau 6.2
Probabilités de couverture des intervalles de confiance à 95 % (*cas II*)

θ	L_0	n	v_J	$v_J^{(1)}$	$v_J^{(C)}$ (LM)	$v_J^{(C+)}$ (LM)
t_{y1}	25	100	94,4	92,0	94,4 (1,02)	94,9 (1,07)
		200	95,0	92,4	95,0 (1,01)	95,2 (1,03)
		400	95,3	92,5	95,1 (0,99)	95,3 (1,01)
	50	100	94,1	93,1	94,2 (1,02)	94,8 (1,07)
		200	94,7	93,3	94,8 (1,01)	95,0 (1,04)
		400	94,7	93,4	94,5 (0,99)	94,8 (1,02)
μ_{y2} / μ_{y3}	25	100	92,6	87,3	92,6 (1,05)	93,3 (1,11)
		200	93,6	86,8	93,3 (1,02)	93,7 (1,07)
		400	94,1	86,8	93,8 (0,99)	94,1 (1,04)
	50	100	92,8	90,3	92,9 (1,06)	94,2 (1,11)
		200	93,9	89,8	93,8 (1,03)	94,3 (1,08)
		400	94,1	89,6	93,8 (1,00)	94,1 (1,05)
$\rho(y_1, y_2)$	25	100	92,5	78,0	89,4 (1,06)	91,7 (1,09)
		200	92,7	72,3	86,3 (1,00)	91,4 (1,05)
		400	93,2	71,2	84,5 (0,95)	92,1 (1,04)
	50	100	92,2	84,5	92,2 (1,09)	92,5 (1,11)
		200	92,8	80,5	90,3 (1,05)	92,2 (1,08)
		400	93,1	77,4	88,1 (1,00)	92,6 (1,06)

v_J : estimateur de la variance jackknife avec suppression d'un groupe; $v_J^{(1)}$: estimateur de la variance par rééchantillonnage fondé sur L_0 jeux de poids jackknife; $v_J^{(C)}$: estimateur de la variance par rééchantillonnage fondé sur L_0 jeux de poids jackknife calés; $v_J^{(C+)}$: estimateur de la variance par rééchantillonnage fondé sur L_0 jeux de poids jackknife calés, avec variables ajoutées pour le calage des poids; LM : longueur moyenne de l'intervalle de confiance comparativement à celui utilisant v_J.

Tableau 6.3
Racine carrée relative de l'erreur quadratique moyenne (RCREQM, *cas I*)

θ	L_0	n	v_R	$v_R^{(1)}$	$v_R^{(2)}$	$v_R^{(3)}$	$v_R^{(4)}$	v_C
t_{y1}	25	50	1,84	2,76	2,24	1,99	1,86	1,43
		100	1,32	2,34	1,67	1,89	1,40	0,83
		150	1,19	1,99	1,34	1,37	1,46	0,87
	50	100	1,32	1,91	1,69	1,63	1,35	0,92
		150	1,19	1,81	1,50	1,62	1,24	0,73
μ_{y2} / μ_{y3}	25	50	0,72	0,89	0,88	1,07	0,89	0,74
		100	0,45	0,78	0,77	0,99	0,72	0,46
		150	0,41	0,93	0,87	1,01	0,74	0,41
	50	100	0,46	0,60	0,60	0,77	0,56	0,46
		150	0,41	0,70	0,68	0,70	0,54	0,41
$\rho(y_1, y_2)$	25	50	0,65	0,79	0,83	1,45	1,26	0,96
		100	0,65	1,12	1,16	2,20	1,37	1,24
		150	0,59	1,29	1,34	2,27	1,43	1,50
	50	100	0,65	0,84	0,88	1,63	0,95	1,03
		150	0,59	0,88	0,94	1,48	1,05	1,12

v_R : estimateur de la variance par rééchantillonnage entièrement efficace (section 2); $v_R^{(l)}, l = 1, 2, 3, 4$: estimateurs de la variance par rééchantillonnage fondés sur L_0 jeux de poids (section 3.1); v_C : estimateur de la variance par rééchantillonnage fondé sur L_0 jeux de poids calés (section 3.2).

Tableau 6.4
Racine carrée relative de l'erreur quadratique moyenne (RCREQM, *cas II*)

θ	L_0	n	v_J	$v_J^{(1)}$	$v_J^{(C)}$	$v_J^{(C+)}$
t_{y1}	25	100	0,29	0,56	0,56	0,66
		200	0,21	0,57	0,47	0,53
		400	0,15	0,56	0,20	0,41
	50	100	0,29	0,41	0,50	0,57
		200	0,21	0,41	0,39	0,44
		400	0,15	0,41	0,17	0,35
μ_{y2} / μ_{y3}	25	100	0,78	1,58	1,90	1,98
		200	0,56	1,44	1,41	1,57
		400	0,39	1,54	0,87	1,40
	50	100	0,81	1,12	1,61	1,67
		200	0,57	1,10	1,22	1,32
		400	0,38	1,04	0,72	1,00
$\rho(y_1, y_2)$	25	100	1,02	2,43	2,71	2,72
		200	0,74	2,44	2,64	2,57
		400	0,47	2,51	2,52	2,42
	50	100	1,04	1,64	1,97	2,12
		200	0,71	1,75	2,01	1,96
		400	0,48	1,76	1,83	1,73

v_J : estimateur de la variance jackknife avec suppression d'un groupe; $v_J^{(1)}$: estimateur de la variance par rééchantillonnage fondé sur L_0 jeux de poids jackknife; $v_J^{(C)}$: estimateur de la variance par rééchantillonnage fondé sur L_0 jeux de poids jackknife calés; $v_J^{(C+)}$: estimateur de la variance par rééchantillonnage fondé sur L_0 jeux de poids jackknife calés, avec variables ajoutées pour le calage des poids.

7 Certaines conclusions

Les méthodes de rééchantillonnage offrent une alternative asymptotiquement équivalente aux méthodes de linéarisation, mais sont plus commodes et plus souples sur le plan opérationnel. Nous nous sommes concentrés sur des paramètres de population qui sont des fonctions lisses des moyennes ou des totaux. Nos résultats théoriques et nos études par simulation limitées montrent que les stratégies proposées pour construire des poids de rééchantillonnage parcimonieux et efficaces donnent de bons résultats pour l'estimation de la variance et les intervalles de confiance. Néanmoins, un certain nombre de problèmes doivent être étudiés plus en profondeur. Premièrement, pour des paramètres complexes tels que les coefficients de corrélation de population, les estimateurs de la variance par rééchantillonnage parcimonieux ne sont pas très stables. Deuxièmement, d'autres preuves de l'efficacité des stratégies proposées pour les grandes enquêtes complexes conjuguées à l'utilisation de poids bootstrap ou jackknife généraux sont nécessaires. Troisièmement, il n'est pas certain que les poids de rééchantillonnage parcimonieux seront efficaces pour des paramètres qui sont des fonctions non lisses des moyennes ou des totaux, tels que les quantiles de population, pour lesquels on sait que les intervalles de confiance de la théorie normale sont inefficaces (Sitter et Wu 2001).

Une autre question importante est celle de l'application éventuelle des méthodes proposées à des paramètres et à des estimateurs définis au moyen d'équations d'estimation. Soit $\boldsymbol{\theta}$ défini comme étant la solution de

$$U_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} u_i(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{7.1}$$

Soit $\hat{\boldsymbol{\theta}}$ obtenu en résolvant une version fondée sur l'échantillon de (7.1) donnée par

$$U_n(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} w_i u_i(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{7.2}$$

Les analyses par la régression ou par la régression logistique en utilisant des données d'enquête complexes peuvent être considérées comme des cas particuliers de forme générale donnés par (7.1) et (7.2). La variance de type sandwich habituelle de $\hat{\boldsymbol{\theta}}$ est donnée par

$$V(\hat{\boldsymbol{\theta}}) \doteq \left\{ \frac{\partial U_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{-1} V\{U_n(\boldsymbol{\theta})\} \left\{ \frac{\partial U_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{-1} \tag{7.3}$$

Nous pouvons maintenant obtenir un estimateur de la variance $v(\hat{\boldsymbol{\theta}})$ si nous remplaçons $\partial U_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ par $\partial U_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ et estimons $V\{U_n(\boldsymbol{\theta})\}$ en appliquant la méthode d'estimation de la variance par rééchantillonnage à $\hat{U}_n = \sum_{i \in \mathcal{S}} w_i \boldsymbol{u}_i$ avec $\boldsymbol{u}_i = u_i(y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}})$. Pour des discussions détaillées des équations d'estimation et de l'échantillonnage, consulter entre autres, Binder (1983), Skinner (1989), et Godambe et Thompson (2009).

Arriver à une estimation efficace de la variance en utilisant un nombre limité de jeux de poids de rééchantillonnage est un problème de recherche important du point de vue tant théorique que pratique. Les poids de rééchantillonnage entièrement efficaces conçus en suivant la procédure décrite à la section 2 peuvent être traités comme des jeux de poids initiaux si la taille de l'échantillon n est grande. En principe, les stratégies que nous proposons à la section 3 pour produire des poids de rééchantillonnage parcimonieux et efficaces peuvent être combinées avec d'autres jeux de poids de rééchantillonnage initiaux, y compris les poids bootstrap (Shao 1996) ou les poids jackknife avec suppression d'un groupe (Kott 2001). On devrait également inclure autant de variables pertinentes que possible dans l'étape de calage, afin que les poids de rééchantillonnage calés finaux ne soient pas seulement parcimonieux, mais également efficaces pour l'obtention d'estimateurs de la variance pour une grande classe d'estimateurs. Des extensions de la méthode proposée afin de traiter les poids calés ou les corrections de la non-réponse sont en cours d'étude.

Remerciements

Nous remercions deux examinateurs anonymes et le rédacteur associé de leurs commentaires très utiles. Les présents travaux ont pour origine les discussions initiales entre le premier auteur J.K. Kim et le professeur Randy Sitter de la Simon Fraser University qui a disparu en mer

tragiquement durant une expédition en kayak en 2007. Les auteurs souhaitent dédier le présent article à la mémoire du professeur Sitter qui était également le superviseur de la thèse de doctorat du deuxième auteur C. Wu. Les travaux de recherche de J.K. Kim ont été financés en partie par une entente de coopération entre le Natural Resources Conservation Service du US Department of Agriculture et la Iowa State University. Les travaux de recherche de C. Wu ont été financés par des subventions du Conseil de recherches en sciences naturelles et en génie du Canada et du réseau des mathématiques et des technologies de l'information et des systèmes complexes (MITACS).

Annexe

A Preuve du théorème 2

En vertu de l'hypothèse (4.2), nous avons

$$\max_{1\le k\le L} c_k(\hat{t}_y^{(k)} - \hat{t}_y)^2 = O_p(L^{-1}n^{-1}N^2),$$

qui, combiné à (4.3), implique que

$$\max_{1\le k\le L}(\hat{\mu}_y^{(k)} - \hat{\mu}_y) = o_p(1), \tag{A.1}$$

où $\hat{\mu}_y^{(k)} = N^{-1}\hat{t}_y^{(k)}$ et $\hat{\mu}_y = N^{-1}\hat{t}_y$. Soit $g(\mu_y) = f(N\mu_y)$. Nous pouvons écrire

$$\hat{\theta}^{(k)} - \hat{\theta} = g(\hat{\mu}_y^{(k)}) - g(\hat{\mu}_y) = \dot{g}(\hat{\mu}_y)(\hat{\mu}_y^{(k)} - \hat{\mu}_y) + Q_{nk}(\hat{\mu}_y^{(k)} - \hat{\mu}_y),$$

où $\dot{g}(\mu) = \partial g(\mu) / \partial\mu, Q_{nk} = \dot{g}(\mu_k^*) - \dot{g}(\hat{\mu}_y)$ et μ_k^* est un point intérieur sur le segment de droite compris entre $\hat{\mu}^{(k)}$ et $\hat{\mu}$. En vertu de (A.1), nous avons

$$\max_{1\le k\le L}(\mu_k^* - \hat{\mu}_y) = o_p(1). \tag{A.2}$$

Définissons

$$D_\delta = \left\{\mu \,\middle|\, \max_k \left\|\mu_k^* - \mu\right\| < \delta \text{ et } \max_k \left\|\dot{g}(\mu_k^*) - \dot{g}(\mu)\right\| > \epsilon\right\}.$$

Par construction, nous avons, pour tout $\epsilon > 0$ et $\delta > 0$,

$$P\left\{\max_k \left\|\dot{g}(\mu_k^*) - \dot{g}(\hat{\mu}_y)\right\| > \epsilon\right\} \leq P(\hat{\mu}_y \in D_\delta) + P\left(\max_k \left\|\mu_k^* - \hat{\mu}_y\right\| \geq \delta\right).$$

En vertu de la continuité de $\dot{g}(\mu)$ à $\mu = \mu_y$ et du fait que $\hat{\mu}_y = \mu_y + o_p(1)$, nous avons que, pour tout $\epsilon > 0$, il existe un $\delta = \delta(\epsilon) > 0$ tel que $P(\hat{\mu}_y \in D_\delta) = o(1)$. Cela, conjugué à (A.2), implique que

$$\max_k \left\|\dot{g}(\mu_k^*) - \dot{g}(\hat{\mu}_y)\right\| = o_p(1). \tag{A.3}$$

Maintenant, nous avons

$$\sum_{k=1}^{L} c_k(\hat{\theta}^{(k)} - \hat{\theta})^2 = A_n + B_n + 2C_n, \tag{A.4}$$

où

$$\begin{aligned} A_n &= \sum\nolimits_{k=1}^{L} c_k\{\dot{g}(\hat{\mu}_y)(\hat{\mu}_y^{(k)} - \hat{\mu}_y)\}^2, \\ B_n &= \sum\nolimits_{k=1}^{L} c_k\{Q_{nk}(\hat{\mu}_y^{(k)} - \hat{\mu}_y)\}^2 \quad \text{et} \\ C_n &= \sum\nolimits_{k=1}^{L} c_k\dot{g}(\hat{\mu}_y)(\hat{\mu}_y^{(k)} - \hat{\mu}_y)^2 Q_{nk}. \end{aligned}$$

Notons que (4.4) implique que

$$\sum_{k=1}^{L} c_k(\hat{\mu}_y^{(k)} - \hat{\mu}_y)^2 \,/\, V(\hat{\mu}_y) = 1 + o_p(1). \tag{A.5}$$

En vertu des arguments de linéarisation classique, nous avons $A_n \,/\, V(\hat{\theta}) \to 1$ en probabilité. En outre, en vertu de (A.3) et (A.5), nous avons $B_n \,/\, V(\hat{\theta}) = o_p(1)$ et $C_n \,/\, V(\hat{\theta}) = o_p(1)$. Cela établit (4.5).

B Preuve du théorème 3

En combinant (3.10) et (3.11) et en ignorant les termes d'ordre plus faible, nous avons

$$v_0(\hat{t}_y) - v_C(\hat{t}_y) \doteq \hat{\boldsymbol{\beta}}' v_0(\hat{t}_{\mathbf{z}})\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}' v_1(\hat{t}_{\mathbf{z}})\hat{\boldsymbol{\beta}} \doteq \boldsymbol{\beta}' v_0(\hat{t}_{\mathbf{z}})\boldsymbol{\beta} - \boldsymbol{\beta}' v_1(\hat{t}_{\mathbf{z}})\boldsymbol{\beta}.$$

où $\boldsymbol{\beta}$ est la limite de probabilité de $\hat{\boldsymbol{\beta}}$. En vertu de (4.6), nous avons

$$E^*\{v_0(\hat{t}_{\mathbf{z}})\} = v_1(\hat{t}_{\mathbf{z}}), \tag{B.1}$$

où $E^*(\cdot)$ désigne l'espérance sous la sélection aléatoire de L_0 jeux de poids conditionnellement aux L jeux de poids. De même, en vertu de (3.11), nous avons

$$v_1(\hat{t}_y) - v_C(\hat{t}_y) \doteq v_1(\hat{t}_e) - v_0(\hat{t}_e).$$

En vertu de (4.6) de nouveau, nous avons

$$E^*\{v_0(\hat{t}_e)\} = v_1(\hat{t}_e). \tag{B.2}$$

Soit $\hat{d}_1 = v_C(\hat{t}_y) - v_1(\hat{t}_y)$, nous avons $E(\hat{d}_1) = 0$ en vertu de (B.2), ce qui prouve (4.7). En outre, de nouveau en vertu de (B.2), nous avons $\mathrm{Cov}\{\hat{d}_1, v_1(\hat{t}_y)\} = 0$. Donc, nous avons

$$V\{v_C(\hat{t}_y)\} = V\{v_1(\hat{t}_y)\} + V(\hat{d}_1) \geq V\{v_1(\hat{t}_y)\}. \tag{B.3}$$

De la même façon, nous pouvons également prouver que $V\{v_0(\hat{t}_y)\} \geq V\{v_C(\hat{t}_y)\}$.

Bibliographie

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Breidt, F.J., et Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 411-425.

Campbell, C. (1980). A different view of the finite population estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 319-324.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 2, 219-230.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 411-425.

Dippo, C.S., Fay, R.E. et Morganstein, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 489-494.

Fay, R.E. (1984). Some properties of estimators of variance based on replication methods. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 495-500.

Fay, R.E., et Dippo, C.S. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 212-217.

Fuller, W.A. (1998). Replication variance estimation for two phase samples. *Statistica Sinica*, 8, 1153-1164.

Fuller, W.A. (2009a). *Sampling Statistics*. Hoboken, New Jersey : John Wiley & Sons, Inc.

Fuller, W.A. (2009b). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Godambe, V.P., et Thompson, M.E. (2009). Estimating functions and survey sampling. Dans *Handbook of Statistics*, (Éds., D. Pfeffermann et C.R. Rao), Sample Surveys: Inference and Analysis, Hollande du Nord, Vol. 29B, 83-101.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 181-184.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Jang, D., et Eltinge, J.L. (2009). Utilisation des variances à l'intérieur des unités primaires d'échantillonnage pour évaluer la stabilité d'un estimateur classique de variance fondé sur le plan de sondage. *Techniques d'enquête*, 35, 2, 255-266.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Lu, W.W., Brick, J.M. et Sitter, R.R. (2006). Algorithms for constructing combining strata variance estimators. *Journal of the American Statistical Association*, 101, 1680-1692.

Lu, W.W., et Sitter, R.R. (2008). Disclosure risk and replication-based variance estimation. *Statistica Sinica*, 18, 1669-1687.

McCarthy, P.J., et Snowden, C.B. (1985). *The Bootstrap and Finite Population Sampling*. Vital and Health Statistics, Sér. 2, No. 95, Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington, DC.

Preston, J. (2009). Bootstrap rééchelonné pour l'échantillonnage stratifié à plusieurs degrés. *Techniques d'enquête*, 35, 2, 247-254.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rust, K.F., et Kalton, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*, 3, 69-81.

Rust, K.F., et Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Shao, J. (1996). Resampling methods in sample surveys (avec discussion). *Statistics*, 27, 203-254.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical Science*, 18, 191-198.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York : Springer.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Sitter, R.R., et Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters*, 52, 353-358.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys*, (Éds., C.J. Skinner, D. Holt et T.M. Smith), New York : John Wiley & Sons, Inc., 59-88.

Tillé, Y. (2006). *Sampling Algorithms*. Springer Science + Business Media, Inc.

Wolter, K.M. (2007). *Introduction to Variance Estimation* (2e Édition). New York : Springer-Verlag.

Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.

Wu, C. (2005). Algorithmes et codes R pour la méthode de la pseudo-vraisemblance empirique dans les sondages. *Techniques d'enquête*, 31, 2, 261-266.

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 139-167
Statistique Canada, N° 12-001-X au Catalogue

Estimation de la variance d'indicateurs transversaux pour l'enquête SILC en Suisse

Anne Massiani[1]

Résumé

L'enquête SILC (*Statistics on Income and Living Conditions*) est une enquête européenne annuelle visant à mesurer la répartition des revenus, la pauvreté et les conditions de vie de la population. Elle est réalisée en Suisse depuis 2007 selon un schéma rotatif de quatre panels qui permet de produire à la fois des estimations transversales et des estimations longitudinales. Dans cet article, nous abordons le problème de l'estimation de la variance des indicateurs transversaux sur la pauvreté et l'exclusion sociale retenus par Eurostat. Nos calculs tiennent compte de la non-linéarité des estimateurs, de la non-réponse totale à différentes phases d'enquête, du sondage indirect et du calage. Nous adaptons la méthode d'estimation de variance en cas de non-réponse après un partage des poids proposée par Lavallée (2002) et obtenons un estimateur de variance asymptotiquement sans biais et très simple à programmer.

Mots clés : Enquête SILC; panel rotatif; indices d'inégalité; estimation de variance; méthode du partage des poids.

1 Introduction

L'enquête SILC (*Statistics on Income and Living Conditions*) est une enquête européenne annuelle qui permet d'obtenir des indicateurs comparables d'un pays à un autre sur la pauvreté, l'exclusion sociale et les conditions de vie de la population. Une présentation détaillée de cette enquête peut être trouvée dans Clemenceau et Museux (2006). Conformément aux recommandations d'Eurostat (Eurostat 2003), l'enquête est réalisée en Suisse selon un schéma rotatif de quatre panels, le premier panel ayant été interrogé en 2007. La durée de chaque panel est de quatre ans et, tous les ans, un panel est remplacé. En rythme de croisière, l'échantillon sélectionné pour un panel comporte environ 3 600 ménages. Nous nous intéressons dans cet article à l'estimation transversale, pour la population présente lors d'une année t, des indicateurs sur la pauvreté et sur les conditions de vie retenus par Eurostat tels que le taux de risque de

1. Anne Massiani, Institut de Statistique, Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, Suisse & Office Fédéral de la Statistique, Espace de l'Europe 10, 2010 Neuchâtel, Suisse. Courriel : anne.massiani@unine.ch.

pauvreté et le rapport interquintile de revenu. Nous renvoyons à Osier (2009) pour une présentation de ces indicateurs et à Ardilly et Lavallée (2007) pour une description de l'approche transversale dans le cadre de l'enquête SILC. Dans cette vision transversale, l'évolution de la composition des ménages des panels sélectionnés est prise en compte grâce à l'utilisation de la théorie du sondage indirect développée par Lavallée (2002).

Nous présentons dans cette étude des formules d'estimation de la variance des indicateurs qui tiennent compte de la grande complexité de l'enquête SILC-Suisse. Les résultats que nous obtenons, bien que développés spécifiquement pour l'enquête réalisée en Suisse, sont susceptibles d'intéresser les autres pays participants, puisque leurs méthodologies sont voisines. Les principaux éléments pris en compte dans l'estimation de la variance des indicateurs sont la non-linéarité des estimateurs, la non-réponse totale à différentes phases d'enquête, le sondage indirect et le calage. Une solution pour estimer la variance d'indicateurs non-linéaires est de recourir à des techniques de linéarisation (voir Deville 1999). Des formules spécifiquement adaptées aux indicateurs retenus par Eurostat ont été développées par Osier (2009). Une formule alternative pour le rapport interquintile de revenu est disponible dans Langel et Tillé (2011). Une fois que les techniques de linéarisation sont appliquées, subsiste encore le problème de l'estimation de la variance d'un total dans un schéma d'enquête complexe. Une des difficultés tient à la présence de non-réponse après un partage des poids. Lavallée (2002) propose un estimateur de la variance d'un total qui en tient compte. Cet estimateur n'est cependant généralement pas sans biais, même lorsque les probabilités de réponse sont connues. Nous proposons ici une adaptation qui corrige ce biais.

Nous présentons brièvement dans la section 2 le déroulement de l'enquête et son plan de sondage. Dans la section 3, nous exposons la pondération utilisée. La section 4 décrit l'application des techniques de linéarisation à ces estimateurs pour obtenir une approximation de leur variance. La section 5 est consacrée au problème de l'estimation de la variance d'un total en présence de non-réponse après un partage des poids. La formule finale utilisée pour l'estimation

de la variance est donnée dans la section 6. Enfin, nous proposons une application numérique dans la section 7, avant de tirer les conclusions de notre étude.

2 Plan de sondage et déroulement de l'enquête

L'enquête est réalisée en Suisse selon un schéma rotatif de quatre panels. Chaque année t, un nouvel échantillon de ménages $s_m^{A_1,t}$ est sélectionné pour remplacer un panel sortant. L'indice m est utilisé ici pour les échantillons de ménages, l'indice p sera réservé pour les échantillons de personnes. En accord avec les notations de Lavallée (2002), les exposants A_i se réfèrent à des échantillons sélectionnés de manière directe, tandis que l'exposant B sera utilisé dans la suite du document pour les échantillons sélectionnés de manière indirecte. Le nouveau panel sélectionné sera suivi durant quatre ans selon le scénario suivant :

- Lors de la première année t, on sollicite tout d'abord les ménages de $s_m^{A_1,t}$ pour remplir un questionnaire préliminaire appelé « grille ». Les variables relevées dans la grille sont communes à tout le ménage et concernent principalement sa composition. On désigne par $s_m^{A_2,t}$ le sous-échantillon de $s_m^{A_1,t}$ qui répond à la grille. L'échantillon des individus appartenant aux ménages de $s_m^{A_2,t}$ est noté $s_p^{A_2,t}$ et ces individus sont appelés des longitudinaux. Les ménages de $s_m^{A_2,t}$ atteints lors de cette première année d'enquête t sont ensuite invités à remplir un questionnaire commun à tout le ménage, appelé questionnaire ménage.
- Lors des trois années suivantes $t + i$, pour $i = 1$ à 3, on tente de recontacter les longitudinaux de $s_p^{A_2,t}$ âgés de 16 ans et plus dans le but d'enquêter leurs ménages tels qu'ils sont composés lors de l'année $t + i$, les ménages pouvant en effet se modifier au cours du temps. Tous les individus de ces ménages sont intégrés à l'enquête de l'année $t + i$ et sont appelés individus transversaux. Les ménages de l'année $t + i$ atteints par le biais des longitudinaux, donc de manière indirecte, sont appelés à remplir à nouveau la grille, puis le questionnaire ménage. On ne parvient cependant pas à recontacter tous les longitudinaux, notamment à cause des déménagements, ce qui constitue une cause significative de non-réponse et d'érosion de l'échantillon.

Nous nous fixons maintenant une année d'enquête t donnée et adoptons une approche transversale, c'est-à-dire que nous nous intéressons aux estimations que l'on peut produire pour

la population présente lors de l'année t. L'échantillon interrogé durant l'année d'enquête t est composé de quatre panels contactés respectivement pour la première, deuxième, troisième ou quatrième fois. On désigne par $s_m^{B,\tau}$ l'échantillon de ménages répondant au questionnaire ménage en τ^{e} interrogation (ou vague), pour $\tau = 1, \ldots, 4$. Comme le montre la figure 2.1, l'échantillon $s_m^{B,\tau}$ a été contacté par l'intermédiaire des longitudinaux sélectionnés lors de l'année $t_\tau = t - \tau + 1$. Les parties grisées à droite représentent les échantillons participant à l'enquête lors de l'année t (sondage indirect), tandis que les échantillons de gauche contiennent les ménages initiaux des longitudinaux ayant permis de les contacter (sondage direct).

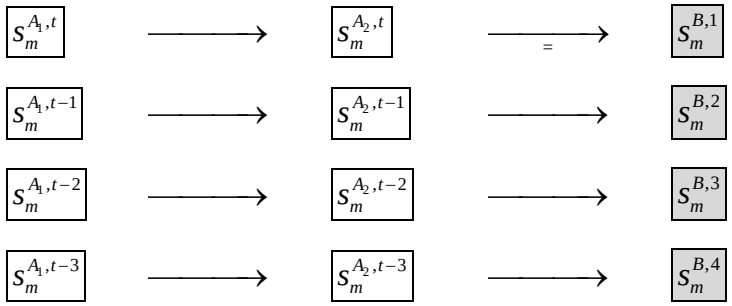


Figure 2.1 Panels composant l'échantillon enquêté lors de l'année *t*

Nous décrivons ci-dessous de façon plus détaillée le plan de sondage de $s_m^{A_2,t_\tau}$, pour $\tau = 1, \ldots, 4$. La composition des ménages pouvant évoluer au fil du temps, nous indicerons dans la suite du document par k_1 les ménages de la première année d'interrogation t_τ et par k ceux de l'année d'enquête t afin de les distinguer. Chaque échantillon $s_m^{A_2,t_\tau}$ provient de deux phases de sélection.

- **Phase 1 :** on sélectionne un échantillon de ménages $s_m^{A_1,t_\tau}$ selon un plan stratifié par grande région, qui sont en Suisse au nombre de sept. Au sein de chaque strate, les tirages sont réalisés selon un plan simple. On note $\pi_{k_1}^{A_1}$ la probabilité d'inclusion d'ordre 1 du ménage k_1 et $\pi_{k_1 k_1'}^{A_1}$ la probabilité d'inclusion d'ordre 2 des ménages k_1 et k_1'.

- **Phase 2 :** la deuxième phase de sélection provient de la non-réponse à la grille lors de la première année d'interrogation t_τ. Cette non-réponse est modélisée par un plan de Poisson sur les ménages et on note, pour tout ménage k_1 :

$$\mathrm{P}(k_1 \in s_m^{A_2,t_\tau} \mid k_1 \in s_m^{A_1,t_\tau}) = q_{k_1}^a. \tag{2.1}$$

Pour tous ménages k_1 et k_1', on définit alors :

$$\pi_{k_1}^{A_2} = \mathrm{P}(k_1 \in s_m^{A_2,t_\tau}) = \pi_{k_1}^{A_1} q_{k_1}^a \tag{2.2}$$

$$\pi_{k_1k_1'}^{A_2} = \mathrm{P}(k_1, k_1' \in s_m^{A_2,t_\tau}) = \begin{cases} \pi_{k_1}^{A_1} q_{k_1}^a & \text{si } k_1 = k_1' \\ \pi_{k_1k_1'}^{A_1} q_{k_1}^a q_{k_1'}^a & \text{si } k_1 \neq k_1'. \end{cases} \tag{2.3}$$

Enfin, pour tous longitudinaux j et j' appartenant respectivement aux ménages k_1 et k_1', on pose :

$$\pi_j^{A_2} = \mathrm{P}(j \in s_p^{A_2,t_\tau}) = \pi_{k_1}^{A_2} \quad \pi_{jj'}^{A_2} = \mathrm{P}(j, j' \in s_p^{A_2,t_\tau}) = \pi_{k_1k_1'}^{A_2}. \tag{2.4}$$

Pour des raisons pratiques, les échantillons $s_m^{A_1,t_\tau}$, pour $\tau = 1, \ldots, 4$ sont tirés de façon à être disjoints, ce qui implique que les échantillons $s_m^{A_2,t_\tau}$ le sont également. En théorie, les quatre échantillons $s_m^{B,\tau}$ ne sont en revanche pas nécessairement disjoints puisqu'il n'est pas exclu que deux longitudinaux sélectionnés lors de deux vagues différentes se trouvent dans le même ménage lors de l'année d'enquête. Compte tenu de la faible taille des échantillons, cela est extrêmement peu probable et n'a jamais été observé en pratique jusqu'à présent. En effet, en rythme de croisière, un panel comporte environ 3 600 ménages tirés avec des probabilités de sélection $\pi_{k_1}^{A_1}$ dont la valeur la plus élevée est de 0,0011. Soulignons toutefois que la méthodologie utilisée permettrait de traiter des échantillons $s_m^{B,\tau}$ non-disjoints. On note :

$$s_m^B = \bigcup_{\tau=1}^{4} s_m^{B,\tau}$$

et un poids w_k, appelé poids transversal ménage, est calculé pour chaque ménage k de s_m^B. On désigne par s_p^B les individus des ménages de s_m^B et chaque individu j d'un ménage k de s_m^B reçoit le poids $w_j = w_k$.

L'information utilisée pour calculer les indices de pauvreté et d'exclusion sociale se base sur le revenu de chaque ménage k de s_m^B. Certaines composantes du revenu du ménage sont relevées dans le questionnaire ménage rempli par tous les ménages de s_m^B. D'autres composantes du revenu, comme les salaires, nécessitent un relevé supplémentaire et sont obtenues à partir de questionnaires individuels soumis aux membres des ménages de s_m^B. En cas de non-réponse complète ou partielle aux questionnaires individuels ou bien en cas de non-réponse partielle au questionnaire ménage, certaines composantes du revenu des ménages de s_m^B sont manquantes. Dans un certain nombre de cas, notamment en cas d'absence d'information sur les salaires, les valeurs manquantes peuvent être obtenues par d'autres sources provenant de données administratives. Dans les cas restants, elles sont imputées. Dans le présent article, nous ne tenons pas compte de l'effet des imputations. Dans ces conditions, le revenu de chaque ménage k de s_m^B est connu et les non-réponses observées lors des différentes phases de l'enquête sont des non-réponses totale de ménage : la non-réponse aux grilles successives et la non-réponse totale au questionnaire ménage. Comme les indices de pauvreté et d'exclusion sociale sont calculés au niveau des individus et non des ménages, on divise le revenu de chaque ménage par un facteur qui dépend du nombre de personnes dans le ménage et de leur âge. Le résultat de ce calcul, appelé revenu équivalent, est attribué à chaque individu j de s_p^B. Les indices de pauvreté et d'exclusion sociale sont calculés à partir du revenu équivalent et des poids w_j.

3 Pondération transversale ménage

La pondération a été réalisée par Graf (2008). Elle fait appel à des techniques de non-réponse et de calage (*cf.* Deville et Särndal 1992, pour plus de détails sur le calage). Elle utilise également la méthode du partage des poids (Lavallée 2002) afin d'attribuer un poids aux individus atteints de manière indirecte. Cette pondération est effectuée en cinq grandes étapes. L'objectif des quatre premières étapes est de calculer un poids pour les ménages de chaque panel $s_m^{B,\tau}$, pour

$\tau = 1, \ldots, 4$. Ces poids se calculent séparément pour chacun des panels. La dernière étape permet ensuite de combiner ces panels. Nous rappelons ci-dessous le principe de chaque étape dans le cas où les probabilités de réponse intervenant dans les différentes phases de l'enquête sont connues. Des explications sur la façon dont ces paramètres sont estimés sont fournies dans l'application numérique de la section 7.

1. *Calcul du poids initial*
   Soit $\tau \in \{1, \ldots, 4\}$ fixé. Pour tout longitudinal $j \in s_p^{A_2, t_\tau}$, on calcule le poids $w_j^{A_2}$ qui tient compte de la non-réponse à la grille en vague 1 :

$$w_j^{A_2} = \frac{1}{\pi_j^{A_2}}. \tag{3.1}$$

2. *Correction de la non-réponse à la grille lors de l'année d'enquête*
   On observe une érosion de l'échantillon entre la première interrogation et l'année d'enquête. D'une part, on ne parvient pas à recontacter tous les longitudinaux de $s_p^{A_2, t_\tau}$. D'autre part, les ménages que l'on parvient à recontacter par l'intermédiaire des longitudinaux n'acceptent pas tous de remplir à nouveau la grille. On désigne par q_k^b la probabilité de réponse à la grille du ménage k lors de l'année d'enquête, conditionnellement à $s_p^{A_2, t_\tau}$. Pour le panel répondant en première interrogation, les q_k^b sont tous égaux à 1 puisqu'il n'y a pas encore eu d'érosion. Pour tout longitudinal j appartenant au ménage k, on pose $q_j^b = q_k^b$ puis on calcule le poids :

$$w_j^g = w_j^{A_2} \frac{1}{q_j^b}. \tag{3.2}$$

   Pour des raisons qui seront explicitées dans la section 7, l'estimation des probabilités de réponse q_j^b est un problème délicat. Compte tenu des informations disponibles, ces estimations ne sont pas toutes égales pour les longitudinaux d'un même ménage k, ce qui n'est pas en adéquation avec le mécanisme de réponse. Cette difficulté, ainsi que la façon dont nous la traitons dans le cadre de l'estimation de la variance des indicateurs, seront abordées dans la section 7.
3. *Partage des poids*
   En suivant la méthodologie de Lavallée (2002, chapitre 6), nous introduisons la définition de cohabitant initialement présent : il s'agit d'un individu qui n'a pas été sélectionné lors de la vague 1, mais qui faisait partie de la population cible lors de la

sélection en vague 1. Les nouveaux-nés ou les immigrants sont appelés des cohabitants initialement absents. On désigne par L_k le nombre de longitudinaux et par P_k le nombre de cohabitants initialement présents dans un ménage k de l'année d'enquête. Le poids d'un ménage k répondant à la grille lors de l'année d'enquête est calculé grâce à un partage des poids de la façon suivante :

$$w_k^p = \frac{1}{L_k + P_k} \sum_{j=1}^{L_k} w_j^g . \tag{3.3}$$

4. *Correction de la non-réponse au questionnaire ménage*

On modélise la non-réponse observée entre la grille de l'année d'enquête et le questionnaire ménage par un plan de Poisson sur les ménages. On note q_k^c la probabilité pour que le ménage k réponde au questionnaire ménage, sachant qu'il a répondu à la grille de l'année d'enquête. Pour tout ménage $k \in s_m^{B,\tau}$, on calcule alors le poids :

$$w_k^{nr} = w_k^p \frac{1}{q_k^c} . \tag{3.4}$$

5. *Combinaison des panels puis calage*

- Combinaison des panels

  On souhaite obtenir une pondération adaptée à la réunion s_m^B des quatre échantillons $s_m^{B,\tau}$, pour $\tau = 1, \ldots, 4$. A cet effet, le poids w_k^{nr} calculé à l'étape précédente doit être divisé, en ordre de grandeur, par un facteur 4. On pose, pour tout ménage $k \in s_m^B$:

$$w_k^u = \phi_k^\tau \cdot w_k^{nr} , \tag{3.5}$$

  où les ϕ_k^τ sont des facteurs d'allocation proche de $1/4$ et devant être optimisés. Merkouris (2001) a utilisé des critères de minimisation de variance pour calculer les valeurs optimales des ϕ_k^τ. Ces valeurs dépendent des variances, donc des effets de plan, associés à chaque panel. Ces effets de plan étant inconnus, on suppose qu'au sein de chaque grande région, ils sont identiques pour chacun des panels. Cela conduit à calculer, au sein de chaque grande région, des facteurs ϕ_k^τ proportionnels à la taille de l'échantillon de ménages répondant en τ^e interrogation.

- *Calage*

  On cale ensuite les poids w_k^u sur des marges connues sur la population des ménages de l'année d'enquête. On note w_k le poids calé ainsi obtenu pour le ménage k.

- *Attribution des poids aux individus*
  Pour tout individu j appartenant à un ménage k de s_m^B, on pose :

$$w_j = w_k .$$

Remarque 1 : Les différentes étapes de la pondération ont étés présentées dans l'ordre où elles sont réalisées en pratique. Nous exposons ci-dessous une autre façon d'aboutir exactement aux mêmes poids finaux lorsque les probabilités de réponse q_j^b sont connues. Dans ce cas, comme $q_j^b = q_k^b$ pour tous les longitudinaux j d'un même ménage k de l'année d'enquête, il est aisé de constater que le poids w_k^{nr} donné par la formule (3.4) peut se réécrire de la façon suivante :

$$w_k^{nr} = \frac{1}{L_k + P_k} \left(\sum_{j=1}^{L_k} w_j^{A_2} \frac{1}{q_j^b} \right) \frac{1}{q_k^c} = w_k^{p^*} \frac{1}{q_k^b q_k^c} , \tag{3.6}$$

où

$$w_k^{p^*} = \frac{1}{L_k + P_k} \sum_{j=1}^{L_k} w_j^{A_2} . \tag{3.7}$$

La formule (3.6) montre que l'on peut aboutir au même poids w_k^{nr} que Graf (2008), donc aux mêmes poids finaux, en procédant comme suit :

- attribuer au ménage k de l'année d'enquête le poids partagé $w_k^{p^*}$,
- corriger dans une même étape la non-réponse à la grille et au questionnaire ménage de l'année d'enquête en la modélisant par un plan de Poisson sur les ménages de paramètre $q_k^b q_k^c$, et en appliquant le facteur de correction 1 / ($q_k^b q_k^c$).

Cette constatation sera utile pour les calculs de variance dans la section 5.

4 Linéarisation et approximation de la variance

On souhaite estimer la variance $\mathrm{var}(\hat{\Theta})$ d'un estimateur $\hat{\Theta}$ calculé sur l'échantillon d'individus transversaux s_p^B affectés des poids w_j. Lavallée (2002, pages 122-123) a développé un cadre asymptotique pour une population sondée de manière indirecte. Ce cadre permet

d'utiliser des techniques de linéarisation (*cf.* Deville 1999) pour obtenir une approximation de la variance d'un estimateur complexe calculé sur une population sondée de manière indirecte. Si $\hat{\Theta}$ est l'estimateur d'un des indices d'inégalité retenus par Eurostat, on emploie des techniques de linéarisation de façon à ramener l'estimation de la variance de $\hat{\Theta}$ à l'estimation de la variance d'un total. Les macros de Osier (2009) peuvent être utilisées à cet effet. Ses formules de linéarisation sont rappelées dans l'annexe A dans le cas des quatre indicateurs considérés dans l'application numérique de la section 7. On note ℓ_j les linéarisées de $\hat{\Theta}$. On a alors :

$$\mathrm{var}(\hat{\Theta}) \simeq \mathrm{var}\left(\sum_{j \in s_p^B} w_j \ell_j\right). \tag{4.1}$$

L'utilisation des résidus de la régression de la variable d'intérêt par rapport aux variables de calage permet de tenir compte de l'effet du calage dans les calculs de variance (*cf.* Deville et Särndal 1992). Comme les variables de calage x_k sont définies au niveau du ménage, on calcule tout d'abord, pour tout ménage k de l'année d'enquête :

$$\ell_k = \sum_{j \in m_k} \ell_j,$$

où m_k désigne l'ensemble des membres du ménage k, puis on définit $e_k = \ell_k - x_k^T \beta$, le paramètre β étant calculé ici à partir de l'ensemble des ménages présents dans la population. On a alors :

$$\mathrm{var}(\hat{\Theta}) \simeq \mathrm{var}\left(\sum_{j \in s_p^B} w_j \ell_j\right) = \mathrm{var}\left(\sum_{k \in s_m^B} w_k \ell_k\right) \simeq \mathrm{var}\left(\sum_{k \in s_m^B} w_k^u e_k\right). \tag{4.2}$$

Remarque 2 : Les linéarisées ℓ_j font intervenir des quantités calculées sur toute la population, comme le montre par exemple la formule (A.6) de l'annexe A. Conformément à la pratique habituelle, les linéarisées ℓ_j seront remplacées *in fine* par des estimations $\hat{\ell}_j$. De même, les quantités e_k étant inconnues, elles seront remplacées par les estimations

$$\hat{e}_k = \hat{\ell}_k - x_k^T \hat{\beta}, \tag{4.3}$$

où

$$\hat{\ell}_k = \sum_{j \in m_k} \hat{\ell}_j$$

et

$$\hat{\beta} = \left(\sum_{k \in s_m^B} w_k x_k x_k^T \right)^{-1} \left(\sum_{k \in s_m^B} w_k x_k \hat{\ell}_k \right).$$

Enfin, les quatre échantillons $s_m^{B,\tau}$, pour $\tau = 1, \ldots, 4$, qui composent s_m^B étant atteints par l'intermédiaire d'échantillons $s_m^{A_1,t_\tau}$ disjoints, ils ne sont en toute rigueur pas indépendants. Cependant, nous faisons l'approximation que ces quatre échantillons sont indépendants car les probabilités de sélection $\pi_k^{A_1}$ sont très faibles. Nous supposons également que les facteurs d'allocation ϕ_k^τ intervenant dans la formule (3.5) ne sont pas aléatoires. En posant, pour tout ménage k :

$$e'_k = \phi_k^\tau \cdot e_k \tag{4.4}$$

et en rappelant (3.5), on peut réécrire la somme intervenant dans le dernier membre de (4.2) sous la forme suivante :

$$\sum_{k \in s_m^B} w_k^u e_k = \sum_{\tau=1}^{4} \sum_{k \in s_m^{B,\tau}} w_k^{nr} (\phi_k^\tau e_k) = \sum_{\tau=1}^{4} \left(\sum_{k \in s_m^{B,\tau}} w_k^{nr} e'_k \right). \tag{4.5}$$

Cela permet d'obtenir l'approximation suivante de la variance :

$$\mathrm{var}(\hat{\Theta}) \simeq \sum_{\tau=1}^{4} \mathrm{var}(\hat{T}_\tau), \tag{4.6}$$

où

$$\hat{T}_\tau = \sum_{k \in s_m^{B,\tau}} w_k^{nr} e'_k. \tag{4.7}$$

Les quatre composantes de variance intervenant dans la formule (4.6) peuvent être calculées et estimées de la même manière. Nous donnons dans la section suivante un estimateur de la variance de $\hat{T}_\tau$, pour τ quelconque.

5 Estimation de la variance et partage des poids

Les calculs présentés dans cette section sont des adaptations des techniques développées par Lavallée (2002, chapitre 8.5) pour le traitement de la non-réponse de grappe (NRG) dans le cadre du sondage indirect. Les résultats sont donnés dans le cas fictif où les probabilités de réponse intervenant aux différentes phases de l'enquête sont connues. Les quantités e_k ainsi que les quantités e'_k définies par (4.4) sont également supposées connues. Toutes ces quantités seront remplacées par des estimations dans la section 7. On désigne par $\mathbf{1}_{\{k \in s_m^{B,\tau}\}}$ la variable indicatrice qui vaut 1 si le ménage k présent lors de l'année d'enquête t fait partie de l'échantillon $s_m^{B,\tau}$ répondant en τ^{e} vague. Conditionnellement au fait que l'on ait tenté de le contacter par l'intermédiaire des longitudinaux qu'il contient, le ménage k appartient à $s_m^{B,\tau}$ si il a répondu à la grille puis au questionnaire lors de l'année d'enquête t. Par conséquent, on a :

$$E(\mathbf{1}_{\{k \in s_m^{B,\tau}\}} \mid s_p^{A_2,t_\tau}) = \mathrm{P}(k \in s_m^{B,\tau} \mid s_p^{A_2,t_\tau}) = q_k^b q_k^c. \tag{5.1}$$

En utilisant le théorème 8.1 de Lavallée (2002, page 151) et la remarque 1, on vérifie aisément que l'estimateur (4.7) peut se réécrire sous la forme :

$$\hat{T}_\tau = \sum_{j \in s_p^{A_2,t_\tau}} w_j^{A_2} \hat{Z}_j = \sum_{j \in s_p^{A_2,t_\tau}} \frac{1}{\pi_j^{A_2}} \hat{Z}_j, \tag{5.2}$$

où l'on a noté, pour tout longitudinal $j \in s_p^{A_2,t_\tau}$ faisant partie du ménage k lors de l'année d'enquête :

$$\hat{Z}_j = \frac{e'_k}{L_k + P_k} \frac{1}{q_k^b} \frac{1}{q_k^c} \mathbf{1}_{\{k \in s_m^{B,\tau}\}}. \tag{5.3}$$

L'estimateur $\hat{T}_\tau$ est ainsi ramené à une somme sur des individus sélectionnés de manière directe. On décompose maintenant la variance de $\hat{T}_\tau$ de façon standard en conditionnant par rapport à $s_p^{A_2,t_\tau}$:

$$\mathrm{var}(\hat{T}_\tau) = \mathrm{var}_{s_p^{A_2,t_\tau}} \left[\mathrm{E}(\hat{T}_\tau \mid s_p^{A_2,t_\tau}) \right] + \mathrm{E}_{s_p^{A_2,t_\tau}} \left[\mathrm{var}(\hat{T}_\tau \mid s_p^{A_2,t_\tau}) \right]. \tag{5.4}$$

Pour tout individu j présent dans la population lors de l'année du tirage de $s_p^{A_2,t_\tau}$ et appartenant au ménage k lors de l'année d'enquête, on pose :

$$Z_j = \frac{e'_k}{L_k + P_k}. \tag{5.5}$$

En utilisant (5.1), on vérifie que pour tous longitudinaux j et j' appartenant à $s_p^{A_2,t_\tau}$ et faisant partie respectivement des ménages k et k' lors de l'année d'enquête, on a :

$$\mathrm{E}(\hat{Z}_j \mid s_p^{A_2,t_\tau}) = Z_j \tag{5.6}$$

et

$$\mathrm{cov}_{jj'} = \mathrm{cov}\left[(\hat{Z}_j, \hat{Z}_{j'}) \mid s_p^{A_2,t_\tau} \right] = \begin{cases} 0 & \text{si } k \neq k' \\ \left(\dfrac{e'_k}{L_k + P_k} \right)^2 \dfrac{1 - q_k^b q_k^c}{q_k^b q_k^c} & \text{si } k = k'. \end{cases} \tag{5.7}$$

La formule (5.4) devient alors :

$$\mathrm{var}(\hat{T}_\tau) = \underbrace{\sum_{j=1}^{J_\tau}\sum_{j'=1}^{J_\tau} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} Z_j Z_{j'}}_{V_{A_2}^\tau} + \underbrace{\mathrm{E}_{s_p^{A_2,t_\tau}} \left\{ \sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{1}{\pi_j^{A_2}} \frac{1}{\pi_{j'}^{A_2}} \mathrm{cov}_{jj'} \right\}}_{V_{\mathrm{NRG}}^\tau}, \tag{5.8}$$

où J_τ désigne le nombre de personnes présentes dans la population lors de l'année du tirage de $s_p^{A_2,t_\tau}$. Le premier terme $V_{A_2}^\tau$ est la part de variance due au mécanisme de sélection des longitudinaux de $s_p^{A_2,t_\tau}$, tandis que le deuxième terme V_{NRG}^τ est celle due à la non-réponse des ménages à la grille puis au questionnaire ménage lors de l'année t, qui constituent une non-réponse de grappe.

Pour obtenir un estimateur de la variance de $\hat{T}_\tau$, nous adaptons la formule d'estimation de variance (8.37) de Lavallée (2002, page 154). Les différences suivantes sont à noter. Tout d'abord, nous négligeons le fait, qu'en pratique, les probabilités de réponse devront être estimées alors que Lavallée (2002) en tient compte. D'autre part, la méthode d'estimation proposée par

Lavallée (2002) fournit des estimations biaisées, même lorsqu'elle est appliquée dans le cas où les probabilités de réponse sont connues. Par conséquent, nous adaptons sa méthode de façon à obtenir un estimateur sans biais de la variance. Afin de justifier notre propos, nous explicitons tout d'abord ci-dessous le biais obtenu en appliquant la méthode de Lavallée (2002) dans le cas où les probabilités de réponse sont connues. Sa méthode consiste à estimer V_{NRG}^{τ} par un estimateur sans biais $\hat{V}_{\mathrm{NRG}}^{\tau}$ puis à estimer $V_{A_2}^{\tau}$ par $\hat{V}_1^{\tau} = \tilde{V}_{A_2}^{\tau}(\hat{Z}_1, \hat{Z}_2, \ldots)$ où :

$$\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots) = \sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} Z_j Z_{j'} \tag{5.9}$$

est l'estimateur de Horvitz-Thompson de la variance de

$$\tilde{T}_\tau = \sum_{j \in s_p^{A_2,t_\tau}} \frac{1}{\pi_j^{A_2}} Z_j. \tag{5.10}$$

Ceci conduit à l'estimateur de variance :

$$\widehat{\mathrm{var}}_L(\hat{T}_\tau) = \hat{V}_1^{\tau} + \hat{V}_{\mathrm{NRG}}^{\tau}. \tag{5.11}$$

L'utilisation de $\hat{V}_1^{\tau}$, qui est construit en remplaçant les Z_j intervenant dans (5.9) par les $\hat{Z}_j$, est motivée par le fait que les Z_j ne sont pas connus pour tous les longitudinaux j de $s_p^{A_2,t_\tau}$, mais seulement pour les longitudinaux qui font partie lors de l'année d'enquête d'un ménage k ayant répondu au questionnaire, c'est-à-dire un ménage k appartenant à $s_m^{B,\tau}$. Le recours aux $\hat{Z}_j$ permet d'attribuer plus de poids aux longitudinaux de $s_p^{A_2,t_\tau}$ pour lesquels les Z_j sont connus. Le problème est que l'estimateur $\hat{V}_1^{\tau}$ ainsi construit n'estime pas sans biais $V_{A_2}^{\tau}$. Ceci se constate aisément en observant ce qui se produit pour les termes diagonaux : pour tout longitudinal j faisant partie du ménage k lors de l'année d'enquête, la quantité Z_j^2 intervenant dans (5.9) est remplacée par $(\hat{Z}_j)^2 = Z_j^2 / (q_k^b q_k^c)^2 \mathbf{1}_{\{k \in s_m^{B,\tau}\}}$ alors qu'une augmentation de poids d'un facteur $1 / (q_k^b q_k^c)$ seulement est sans doute un meilleur choix. Le même type de problème se pose pour le produit $Z_j Z_{j'}$ lorsque les longitudinaux j et j' appartiennent au même ménage lors de

l'année d'enquête. Plus précisément, on a, pour tous longitudinaux j et j' appartenant à $s_p^{A_2,t_\tau}$:

$$\mathrm{E}(\hat{Z}_j \hat{Z}_{j'} \mid s_p^{A_2,t_\tau}) = Z_j Z_{j'} + \mathrm{cov}_{jj'} \tag{5.12}$$

ce qui implique

$$\mathrm{E}[\hat{V}_1^\tau] - V_{A_2}^\tau = \mathrm{E}_{s_p^{A_2,t_\tau}} \left\{ \mathrm{E}\left[\tilde{V}_{A_2}^\tau(\hat{Z}_1, \hat{Z}_2, \ldots) \mid s_p^{A_2,t_\tau} \right] \right\} - V_{A_2}^\tau \tag{5.13}$$

$$= \mathrm{E}_{s_p^{A_2,t_\tau}} \left[\sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} \mathrm{cov}_{jj'} \right]. \tag{5.14}$$

Comme en revanche $\hat{V}_{\mathrm{NRG}}^\tau$ estime sans biais V_{NRG}^τ, il vient :

$$\mathrm{B}^\tau = E\left[\widehat{\mathrm{var}}_L(\hat{T}_\tau) \right] - \mathrm{var}(\hat{T}_\tau) \tag{5.15}$$

$$= \mathrm{E}_{s_p^{A_2,t_\tau}} \left[\sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} \mathrm{cov}_{jj'} \right]. \tag{5.16}$$

Le biais B^τ dépend notamment du terme $\mathrm{cov}_{jj'}$ défini par (5.7), et donc des probabilités de réponse q_k^b et q_k^c à la grille et au questionnaire lors de l'année d'enquête. Plaçons-nous dans le cas simple du panel répondant en première interrogation, $\tau = 1$, dans lequel la composition des ménages n'a pas encore pu évoluer. La quantité $\mathrm{cov}_{jj'}$ définie par (5.7) est positive si les longitudinaux j et j' appartiennent au même ménage k, et nulle sinon. De plus, pour tous longitudinaux j et j' appartenant au même ménage k, on a d'après la relation (2.4) :

$$\frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_{jj'}^{A_2}} = 1 - \pi_k^{A_2}. \tag{5.17}$$

Cette dernière quantité étant également positive, il découle de l'expression du biais donné par la formule (5.16) que B^1 est positif. D'autre part, les probabilités d'inclusion $\pi_k^{A_2}$ sont très faibles, de sorte que $1 - \pi_k^{A_2} \simeq 1$. En se servant de (5.16) et (5.17), ceci permet de faire l'approximation suivante :

$$\mathrm{B}^1 \simeq \mathrm{E}_{s_p^{A_2,t_1}} \left[\sum_{j \in s_p^{A_2,t_1}} \sum_{j' \in s_p^{A_2,t_1}} \frac{1}{\pi_j^{A_2}} \frac{1}{\pi_{j'}^{A_2}} \mathrm{cov}_{jj'} \right] = V_{\mathrm{NRG}}^1, \tag{5.18}$$

où on rappelle que V_{NRG}^{τ} est défini par (5.8) et correspond en première vague à la part de variance due à la non-réponse observée entre la grille et le questionnaire. Par conséquent, l'estimateur $\widehat{\mathrm{var}}_L(\hat{T}_1)$ surestime la variance de $\hat{T}_1$ et l'erreur commise est de l'ordre de grandeur de V_{NRG}^{1}. Le biais peut-être relativement important si les probabilités de réponse au questionnaire sont faibles. En ce qui concerne les autres vagues, la quantité $(\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}) / \pi_{jj'}^{A_2}$ intervenant dans (5.16) dépend des ménages k_1 et k_1' auxquels les longitudinaux j et j' appartenaient lors de l'année de leur sélection et obtenir un ordre de grandeur du biais B^{τ} n'est plus chose aisée.

Nous introduisons un terme correcteur du biais B^{τ} et donnons notre formule d'estimation de la variance dans la proposition 1 ci-dessous. On rappelle que m_k désigne l'ensemble des membres qui composent le ménage k lors de l'année d'enquête t (*cf.* page 11) et on note $\tilde{m}_k$ l'ensemble, de cardinal L_k, constitué des longitudinaux j appartenant à m_k.

Proposition 1 : On peut estimer sans biais la variance de $\hat{T}_{\tau}$ par

$$\widehat{\mathrm{var}}(\hat{T}_{\tau}) = \hat{V}_1^{\tau} + \hat{V}_2^{\tau}, \tag{5.19}$$

où

$$\hat{V}_1^{\tau} = \tilde{V}_{A_2}^{\tau}(\hat{Z}_1, \hat{Z}_2, \ldots) = \sum_{j \in s_p^{A_2,t_{\tau}}} \sum_{j' \in s_p^{A_2,t_{\tau}}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} \hat{Z}_j \hat{Z}_{j'}$$

et

$$\hat{V}_2^{\tau} = \sum_{k \in s_m^{B,\tau}} \sum_{j,j' \in \tilde{m}_k} \frac{1}{\pi_{jj'}^{A_2}} \left(\frac{e_k'}{L_k + P_k} \right)^2 \frac{1 - q_k^b q_k^c}{(q_k^b q_k^c)^2}.$$

La démonstration de la proposition 1 est donnée dans l'annexe B.

Remarque 3 : L'estimateur $\widehat{\mathrm{var}}(\hat{T}_{\tau}) = \hat{V}_1^{\tau} + \hat{V}_2^{\tau}$ est la somme de deux estimateurs biaisés dont les biais s'équilibrent par construction, de sorte que $\widehat{\mathrm{var}}(\hat{T}_{\tau})$ estime sans biais la variance de $\hat{T}_{\tau}$.

Remarque 4 : La proposition 1 est énoncée sous l'hypothèse que les e_k et les probabilités de réponse sont connues, ce qui permet de conclure à l'absence de biais pour l'estimateur $\widehat{\mathrm{var}}(\hat{T}_{\tau})$

donné par la formule (5.19). En pratique, ces quantités doivent être estimées. Ceci a pour conséquence que l'estimateur de variance ainsi obtenu n'est plus sans biais mais seulement asymptotiquement sans biais, sous réserve que les modèles de non-réponse puissent être considérés comme justes et que leurs paramètres soient estimés par une méthode appropriée.

6 Formule finale

Le terme $\hat{V}_2^{\tau}$ qui intervient dans la proposition 1 comporte une somme double, mais celle-ci ne pose pas de problème sur le plan opérationnel. En effet, elle ne compte que très peu de termes puisqu'elle ne porte que sur les individus d'un même ménage. L'expression de $\hat{V}_1^{\tau} = \tilde{V}_{A_2}^{\tau}(\hat{Z}_1, \hat{Z}_2, \ldots)$ doit en revanche être transformée pour être calculable plus facilement. Nous commençons donc par donner une autre expression du terme $\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots)$ défini par la formule (5.9). On observe pour cela que :

$$\tilde{T}_{\tau} = \sum_{j \in s_p^{A_2, t_\tau}} \frac{1}{\pi_j^{A_2}} Z_j = \sum_{k_1 \in s_m^{A_2, t_\tau}} \frac{1}{\pi_{k_1}^{A_2}} T_{k_1}, \tag{6.1}$$

où

$$T_{k_1} = \sum_{j \in m_{k_1}} Z_j. \tag{6.2}$$

Rappelons que la sélection de $s_m^{A_2, t_\tau}$ résulte de la succession d'un plan stratifié par grandes régions et d'une phase de non-réponse modélisée par un plan de Poisson sur les ménages de la vague 1 (*cf.* section 2). Donner une expression simple de l'estimateur de variance de Horvitz-Thompson d'un total pour ce plan très classique est un problème déjà largement étudié, notamment dans le cadre du logiciel de calcul de précision POULPE (*cf.* Caron, Deville et Sautory 1998, page 13). Afin de donner une telle expression, nous introduisons les notations suivantes. Pour chacune des sept strates h, on désigne par N_h le nombre de ménages qu'elle contient et par n_h le nombre de ménages sélectionnés. Pour tout ménage $k_1 \in s_m^{A_1, t_\tau}$, on pose :

$$T_{k_1}^* = \begin{cases} \dfrac{T_{k_1}}{q_{k_1}^a} & \text{si } k_1 \in s_m^{A_2,t_\tau} \\ 0 & \text{sinon.} \end{cases} \tag{6.3}$$

On pose également, pour tout h :

$$(s_h^*)^2 = \frac{1}{n_h - 1} \sum_{k_1 \in \{s_m^{A_1,t_\tau} \cap h\}} (T_{k_1}^* - \overline{T}_h^*)^2 = \frac{1}{n_h - 1} \left[\sum_{k_1 \in \{s_m^{A_2,t_\tau} \cap h\}} \left(\frac{T_{k_1}}{q_{k_1}^a} \right)^2 \right] - \frac{n_h}{n_h - 1} (\overline{T}_h^*)^2 \tag{6.4}$$

où

$$\overline{T}_h^* = \frac{1}{n_h} \sum_{k_1 \in \{s_m^{A_1,t_\tau} \cap h\}} T_{k_1}^* = \frac{1}{n_h} \sum_{k_1 \in \{s_m^{A_2,t_\tau} \cap h\}} \frac{T_{k_1}}{q_{k_1}^a} .$$

D'après Caron et coll. (1998, page 13), le terme $\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots)$ peut s'écrire ici [voir également la formule (11.12) de Särndal et Lundström 2005) :

$$\begin{aligned} \tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots) = & \sum_{k_1 \in s_m^{A_1,t_\tau}} \sum_{k_1' \in s_m^{A_1,t_\tau}} \frac{\pi_{k_1 k_1'}^{A_1} - \pi_{k_1}^{A_1} \pi_{k_1'}^{A_1}}{\pi_{k_1}^{A_1} \pi_{k_1'}^{A_1}} \frac{1}{\pi_{k_1 k_1'}^{A_1}} T_{k_1}^* T_{k_1'}^* - \sum_{k_1 \in s_m^{A_2,t_\tau}} \frac{1 - \pi_{k_1}^{A_1}}{(\pi_{k_1}^{A_1})^2} \frac{1 - q_{k_1}^a}{(q_{k_1}^a)^2} T_{k_1}^2 \\ & + \sum_{k_1 \in s_m^{A_2,t_\tau}} \frac{1 - q_{k_1}^a}{(q_{k_1}^a)^2} \left(\frac{T_{k_1}}{\pi_{k_1}^{A_1}} \right)^2 . \end{aligned} \tag{6.5}$$

En regroupant les deux derniers termes et en utilisant le fait que le plan de sondage de $s_m^{A_1,t_\tau}$ est un plan stratifié, on aboutit à l'expression simple suivante pour $\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots)$:

$$\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots) = \sum_{h=1}^{7} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h} \right) (s_h^*)^2 + \sum_{k_1 \in s_m^{A_2,t_\tau}} \frac{1 - q_{k_1}^a}{(q_{k_1}^a)^2} \frac{1}{\pi_{k_1}^{A_1}} T_{k_1}^2 . \tag{6.6}$$

On note $\hat{T}_{k_1}$ et $\hat{s}_h^*$ les estimateurs obtenus en remplaçant les variables Z_j par les $\hat{Z}_j$ dans les formules (6.2) et (6.4). La relation (6.6), combinée à la formule (4.6) et à la proposition 1, permet d'obtenir la formule finale ci-dessous pour l'estimation de la variance de l'estimateur complexe $\hat{\Theta}$:

$$\widehat{\mathrm{var}}(\hat{\Theta}) \simeq \sum_{\tau=1}^{4}\left[\hat{V}_1^{\tau} + \hat{V}_2^{\tau}\right] = \underbrace{\sum_{\tau=1}^{4}\hat{V}_{1,1}^{\tau}}_{\hat{V}_{1,1}} + \underbrace{\sum_{\tau=1}^{4}\hat{V}_{1,2}^{\tau}}_{\hat{V}_{1,2}} + \underbrace{\sum_{\tau=1}^{4}\hat{V}_2^{\tau}}_{\hat{V}_2}, \tag{6.7}$$

où

$$\hat{V}_{1,1}^{\tau} = \sum_{h=1}^{7}\frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)(\hat{s}_h^*)^2 \tag{6.8}$$

$$\hat{V}_{1,2}^{\tau} = \sum_{k_1 \in s_m^{A_2,t_\tau}} \frac{1-q_{k_1}^a}{(q_{k_1}^a)^2}\frac{1}{\pi_{k_1}^{A_1}}(\hat{T}_{k_1})^2 \tag{6.9}$$

$$\hat{V}_2^{\tau} = \sum_{k \in s_m^{B,\tau}} \sum_{j,j' \in \tilde{m}_k} \frac{1}{\pi_{jj'}^{A_2}}\left(\frac{\hat{e}'_k}{L_k + P_k}\right)^2 \frac{1-q_k^b q_k^c}{(q_k^b q_k^c)^2}. \tag{6.10}$$

Remarque 5 : La formule d'estimation de variance (6.7) fournit toujours des estimations positives. De plus, les trois termes qui la composent peuvent être programmés très aisément.

7 Application numérique et discussion

Nous présentons maintenant une application numérique permettant d'illustrer les résultats développés dans les sections précédentes. Jusqu'à présent, nous avons supposé que les probabilités de réponse intervenant dans les différentes phases de l'enquête sont connues, mais en pratique, elles doivent être estimées. En ce qui concerne la non-réponse à la grille lors de la première interrogation, l'estimation des probabilités de réponse $q_{k_1}^a$ définies par (2.1) utilise des informations sur les ménages non-répondants collectées à partir d'une enquête de non-réponse. L'échantillon $s_m^{A_1,t_\tau}$ des ménages contactés en première interrogation lors de l'année t_τ est divisé en groupes de réponse homogènes au sein desquels des estimations $\hat{q}_{k_1}^a$ des probabilités de réponse $q_{k_1}^a$ sont calculées. Les facteurs q_k^c intervenant dans (3.4) et qui représentent les probabilités de réponse entre la grille et le questionnaire lors de l'année d'enquête sont également estimés au sein de groupes de réponse homogènes. Ce sont cette fois les ménages ayant répondu à la grille lors de l'année d'enquête qui sont répartis en groupes de réponse homogènes. On note

$\hat{q}_k^c$ l'estimation de la probabilité de réponse q_k^c ainsi obtenue. L'estimation des facteurs q_j^b intervenant dans (3.2) et qui représentent les probabilités de réponse entre la grille de la première interrogation et celle de l'année d'enquête est un peu plus problématique. Les deux options suivantes sont envisageables. La première option est de constituer des groupes de réponse homogènes au sein de l'ensemble des ménages que l'on a tenté de contacter lors de l'année d'enquête. On est alors confronté au problème que très peu d'informations sont disponibles pour constituer ces groupes. On ne connaît en effet pas la nouvelle composition des ménages que l'on n'a pas réussi à recontacter ou qui refusent de remplir à nouveau la grille. La deuxième option est d'estimer les q_j^b au sein de groupes de réponse homogènes définis sur les longitudinaux de $s_p^{A_2,t_\tau}$. Cela n'est pas en totale adéquation avec le mécanisme de réponse mais permet d'utiliser toutes les informations qui ont été collectées au sujet des longitudinaux lors des vagues précédentes. Depuis la première exploitation officielle de l'enquête, le choix s'est porté sur la deuxième option (voir Graf 2008). Compte tenu de cette méthode d'estimation, les quantités $\hat{q}_j^b$ ne sont pas toutes égales au sein d'un même ménage k de l'année d'enquête, ce qui pose problème pour l'application des formules d'estimation de variance développées dans les sections précédentes. On contourne cette difficulté grâce à la constatation suivante. On désigne par $\tilde{w}_k^p$ le poids partagé obtenu à partir de (3.3) et de l'estimation des probabilités de réponse $q_{k_1}^a$ et q_j^b :

$$\tilde{w}_k^p = \frac{1}{L_k + P_k} \sum_{j=1}^{L_k} \left(\tilde{w}_j^{A_2} \frac{1}{\hat{q}_j^b} \right), \tag{7.1}$$

où $\tilde{w}_j^{A_2} = 1 / (\pi_{k_1}^{A_1} \hat{q}_{k_1}^a)$ pour tout longitudinal j appartenant au ménage k_1 en première interrogation. On peut réécrire le poids partagé $\tilde{w}_k^p$ du ménage k de la façon suivante :

$$\tilde{w}_k^p = \frac{1}{L_k + P_k} \sum_{j=1}^{L_k} \left(\tilde{w}_j^{A_2} \frac{1}{\hat{q}_j^b} \right) = \frac{1}{L_k + P_k} \sum_{j=1}^{L_k} \left(\tilde{w}_j^{A_2} \frac{1}{\tilde{q}_j^b} \right) \tag{7.2}$$

où, pour tous les longitudinaux j du ménage k, les quantités $\tilde{q}_j^b$ sont égales à une valeur commune $\tilde{q}_k^b$ définie par

$$\tilde{q}_k^b = \frac{\sum_{j=1}^{L_k} \tilde{w}_j^{A_2}}{\sum_{j=1}^{L_k} \left(\tilde{w}_j^{A_2} \frac{1}{\hat{q}_j^b} \right)}. \tag{7.3}$$

Par conséquent, que l'on estime les quantités q_j^b par $\hat{q}_j^b$ ou par $\tilde{q}_j^b$, on obtient exactement le même poids partagé $\tilde{w}_k^p$, et donc le même poids final et les mêmes estimations des indices de pauvretés. En se basant sur cette constatation, nous obtenons une approximation de la précision de $\hat{\Theta}$ en remplaçant dans les formules d'estimation de variance les quantités inconnues $q_{k_1}^a, q_k^b, q_k^c$ par $\hat{q}_{k_1}^a, \tilde{q}_k^b, \hat{q}_k^c$. Les e_k sont également remplacés par les $\hat{e}_k$ donnés par (4.3). Afin de ne pas compliquer inutilement les notations, on note dans ce qui suit de la même manière les estimateurs de variance calculés à partir des paramètres inconnus et ceux calculés à partir de l'estimation de ceux-ci. Si les estimateurs des probabilités de réponse sont convergents, l'estimateur $\widehat{\text{var}}(\hat{\Theta})$ que nous proposons dans la formule (6.7) est asymptotiquement sans biais. Le caractère asymptotique de cette propriété est dû à l'effet de l'estimation des probabilités de réponse ainsi qu'aux approximations faites lors de la linéarisation, de la prise en compte du calage, et de l'obtention de la formule (4.6).

Les calculs présentés dans cette section ont été réalisés à partir de l'enquête SILC-Suisse 2009. L'enquête SILC n'a débuté en Suisse qu'en 2007. En 2009, l'enquête n'a donc pas encore atteint son rythme de croisière et n'est composée que de trois panels au lieu de quatre. Toutefois cela ne change rien à la méthodologie présentée dans les sections précédentes. En 2009, 7 372 ménages, soit 17 561 personnes, ont accepté de participer à l'enquête. Nous donnons dans le tableau 7.1 les taux de réponse obtenus aux différentes phases de l'enquête pour chacun des panels.

Tableau 7.1
Taux de réponse aux différentes phases de l'enquête

Taux de réponse	Panel 07	Panel 08	Panel 09
A la grille en vague 1	0,688	0,694	0,687
Entre la grille de la vague 1 et celle de la vague courante	0,803	0,834	1,000
Entre la grille et le questionnaire ménage de la vague courante	0,966	0,951	0,942

Sur la base de l'enquête de 2009, les taux de pauvreté de différents sous-groupes de population ont été calculés, analysés et commentés par l'Office fédéral de la statistique (OFS 2010). Nous présentons dans le tableau 7.2 différentes mesures de précision pour quelques uns des indices d'inégalité utilisés le plus fréquemment : taux de risque de pauvreté, rapport interquintile de revenu, indice de Gini, écart relatif médian du risque de pauvreté. Pour chaque indicateur $\hat{\Theta}$ considéré, le tableau 7.2 contient les mesures suivantes :

$$\mathrm{ET} = \sqrt{\widehat{\mathrm{var}}(\hat{\Theta})}, \quad \mathrm{CV} = \frac{\mathrm{ET}}{\hat{\Theta}}, \quad \mathrm{ET}_{1,1} = \sqrt{\hat{V}_{1,1}}, \quad \Delta_1 = \frac{\mathrm{ET}_{1,1}}{\mathrm{ET}} - 1,$$

où les variances $\widehat{\mathrm{var}}(\hat{\Theta})$ et $\hat{V}_{1,1}$ sont données par la formule (6.7). On désigne par $\widehat{\mathrm{var}}_L(\hat{\Theta})$ l'estimateur de variance obtenu par la méthode de Lavallée (2002) :

$$\widehat{\mathrm{var}}_L(\hat{\Theta}) = \sum_{\tau=1}^{3} \widehat{\mathrm{var}}_L(\hat{T}_\tau), \tag{7.4}$$

où $\widehat{\mathrm{var}}_L(\hat{T}_\tau)$ est défini par l'équation (5.11). Les mesures :

$$\mathrm{ET}_L = \sqrt{\widehat{\mathrm{var}}_L(\hat{\Theta})}, \quad \mathrm{CV}_L = \frac{\mathrm{ET}_L}{\hat{\Theta}}, \quad \Delta_2 = \frac{\mathrm{ET}_L}{\mathrm{ET}} - 1,$$

figurent également dans le tableau 7.2.

Tableau 7.2
Précisions estimées à partir des données 2009 pour différents indicateurs

Indicateur	$\hat{\Theta}$	**ET**	**CV**	$\mathbf{ET_{1,1}}$	Δ_1	$\mathbf{ET}_L$	$\mathbf{CV}_L$	Δ_2
Taux de risque de pauvreté								
Population totale	14,6 %	0,581	4,0 %	0,580	-0,61 %	0,632	4,3 %	8,9 %
0-17 ans	18,3 %	1,310	7,2 %	1,310	-0,60 %	1,426	7,9 %	8,8 %
18-24 ans	11,9 %	1,302	10,9 %	1,301	-0,61 %	1,388	11,6 %	6,6 %
25-49 ans	10,7 %	0,585	5,5 %	0,585	-0,62 %	0,645	6,1 %	10,2 %
50-64 ans	9,9 %	0,752	7,6 %	0,752	-0,63 %	0,828	8,4 %	10,1 %
65 ans et plus	26,4 %	1,291	4,8 %	1,291	-0,62 %	1,442	5,4 %	11,6 %
Rapport interquintile de revenu	4,4	0,109	2,4 %	0,109	-0,55 %	0,118	2,7 %	8,9 %
Indice de Gini	***[a]	***	1,7 %	***	-0,56 %	***	1,8 %	8,0 %
ERMRP	***[a]	***	5,8 %	***	-0,60 %	***	6,4 %	9,7 %

[a] Nous ne communiquons pas la valeur de ces indicateurs car ils ne sont pas encore publiés officiellement par l'OFS.

Les comparaisons entre les différentes mesures de précision que permet le tableau 7.2 conduisent aux constatations suivantes. Tout d'abord, pour les indicateurs considérés, la quasi totalité de l'estimation de l'écart-type ET de l'indicateur $\hat{\Theta}$ est contenue dans le terme $\mathrm{ET}_{1,1}$. La différence relative Δ_1 entre ET et $\mathrm{ET}_{1,1}$ varie en effet de -0,55 % pour le rapport interquintile de revenu à -0,63 % pour le taux de risque de pauvreté des 50-64 ans. En ne calculant que le terme dominant $\mathrm{ET}_{1,1}$, on obtient donc généralement une excellente approximation de l'estimation de l'écart-type de $\hat{\Theta}$. Ce terme est extrêmement simple à programmer. D'autre part, la méthode d'estimation de la variance de $\hat{\Theta}$ proposée par Lavallée (2002), qui, comme le montre la formule (5.14), n'est pas sans biais même lorsque les probabilités de réponse sont connues, conduit ici à des estimations de précisions légèrement plus élevées qu'avec la méthode d'estimation que nous proposons. La différence relative Δ_2 entre ET et ET_L varie en effet de 6,6 % pour le taux de risque de pauvreté des 18-24 ans à 11,6 % pour celui des 65 ans et plus. Il découle des formules (5.7) et (5.16) que l'estimateur ET_L est asymptotiquement sans biais et égal à ET lorsque les probabilités de réponse à la grille et au questionnaire ménage q_k^b et q_k^c lors de l'année d'enquête sont égales à 1. Bien que les taux de réponse à la grille et au questionnaire ménage lors de l'année d'enquête soient assez élevés (*cf.* tableau 7.1), les différences entre ET et ET_L sont non négligeables.

Conclusion

Nous avons proposé un estimateur de la variance pour les indicateurs sur la pauvreté et l'exclusion sociale qui tient compte de la non-linéarité des estimateurs, de la non-réponse totale à différentes phases de l'enquête, du sondage indirect et du calage. Idéalement, les effets des imputations devraient également être pris en compte. Notons toutefois que notre démarche est compatible avec les exigences de Eurostat (2010) à qui de nombreux pays européens ne fournissent qu'une approximation de la variance due à l'échantillonnage et à la non-réponse

totale, ainsi que des indications minimales sur les imputations telles que le pourcentage de valeurs imputées. Nous avons modifié la méthode d'estimation de la variance en présence de non-réponse après un partage des poids proposée par Lavallée (2002) de façon à corriger le biais qu'elle induit. Notre estimateur est toujours positif dans le cas de l'enquête SILC-Suisse et est composé de trois termes très simples à programmer. De plus, en ne calculant que le premier de ces trois termes, on obtient une excellente approximation de la variance.

Remerciements

Je tiens à remercier Yves Tillé et Lionel Qualité de l'université de Neuchâtel pour les discussions fructueuses, leurs conseils et leur relecture attentive. Un grand merci également à Eric Graf, Johan Pea, Thomas Christin et Stéphane Fleury de l'Office fédéral de la statistique qui ont fourni les informations nécessaires pour le bon déroulement du projet. Je souhaite aussi remercier Philippe Eichenberger et Monique Graf de l'Office fédéral de la statistique pour leur soutien. Enfin, je suis très reconnaissante envers les arbitres et le rédacteur associé d'avoir pris le temps de relire attentivement ce travail. Leurs commentaires ont été une grande source d'améliorations.

Annexe A

Formules de linéarisation

Nous rappelons ci-dessous les formules de linéarisation de Osier (2009) dans le cas des quatre indicateurs de pauvreté considérés dans la section 7. On note $\mathcal{U}$ la population des individus présents lors de l'année d'enquête. Conformément aux notations à la section 5, J_1 désigne la taille de $\mathcal{U}$, puisque $\mathcal{U}$ correspond à la population cible du panel répondant en vague 1. Cependant, afin d'alléger les notations de cette annexe, la taille de $\mathcal{U}$ est simplement notée J. On désigne par R_j le revenu équivalent de l'individu $j \in \mathcal{U}$ utilisé pour calculer les indicateurs de pauvreté et on note pour tout $x \in \mathbb{R}$:

$$F(x) = \frac{1}{J} \sum_{j \in \mathcal{U}} \mathbf{1}_{\{R_j \leq x\}}, \tag{A.1}$$

où $\mathbf{1}_{\{R_j \leq x\}}$ est une variable indicatrice qui vaut 1 si le revenu R_j de l'individu j est inférieur ou égal à x, et 0 sinon. Les formules de linéarisation des indicateurs de pauvreté sont obtenues dans un premier temps en supposant que F est dérivable et que F' est non nulle. Comme cela n'est pas le cas, on contourne cette difficulté en approchant F par la fonction F_K définie, pour tout $x \in \mathbb{R}$, par :

$$F_K(x) = \int F(z) K(x, z) dz, \tag{A.2}$$

où

$$K(x, z) = \frac{1}{h\sqrt{2\pi}} \exp\left[-\frac{(x - z)^2}{2h^2} \right], \tag{A.3}$$

le paramètre h étant un paramètre de lissage.

1. Taux de risque de pauvreté

   On désigne par SRP seuil de risque de pauvreté calculé à partir du revenu médian MED de la population $\mathcal{U}$:

   $$\text{SRP} = 0{,}6 \times \text{MED}. \tag{A.4}$$

   Le taux de risque de pauvreté, noté TRP, est défini par :

   $$\text{TRP} = \frac{1}{J} \sum_{j \in \mathcal{U}} \mathbf{1}_{\{R_j \leq \text{SRP}\}}. \tag{A.5}$$

   La linéarisée ℓ_j intervenant dans l'approximation de la variance de l'estimateur du taux de risque de pauvreté donnée dans la formule (4.1) s'écrit, pour chaque individu j :

   $$\ell_j = \frac{1}{J}\left[\mathbf{1}_{\{R_j \leq \text{SRP}\}} - \text{TRP}\right] - \frac{0{,}6}{J} \cdot \frac{F'_K(\text{SRP})}{F'_K(\text{MED})} \cdot [\mathbf{1}_{\{R_j \leq \text{MED}\}} - 0{,}5]. \tag{A.6}$$

   Dans la section 7, le taux de pauvreté a également été estimé au sein de différentes sous-populations. La généralisation de la formule (A.6) dans le cas de sous-populations s'obtient facilement et n'est pas rappelée ici.

2. Rapport interquintile de revenu

   On désigne par γ_+ le quantile d'ordre 0,8 et par γ_- le quantile d'ordre 0,2, puis on pose :

$$R = \sum_{j \in \mathcal{U}} R_j \quad \text{et} \quad S(x) = \sum_{j \in \mathcal{U}} R_j \mathbf{1}_{\{R_j \leq x\}}. \tag{A.7}$$

Le rapport interquintile de revenu, noté RIR, est défini par :

$$\text{RIR} = \frac{R - S(\gamma_+)}{S(\gamma_-)}. \tag{A.8}$$

De façon analogue à ce qui a été appliqué pour la fonction F, on approche la fonction S par la fonction dérivable S_K définie par :

$$S_K(x) = \int S(z) K(x, z) dz. \tag{A.9}$$

La linéarisée ℓ_j intervenant dans l'approximation de la variance de l'estimateur du *rapport interquintile de revenu* donnée dans la formule (4.1) s'écrit, pour chaque individu j :

$$\ell_j = \frac{S(\gamma_-)\left[R_j - Q_j(\gamma_+)\right] - \left[R - S(\gamma_+)\right] Q_j(\gamma_-)}{\left[S(\gamma_-)\right]^2}, \tag{A.10}$$

où la quantité $Q_j(\gamma)$ est définie, pour tout quantile γ d'ordre α, par

$$Q_j(\gamma) = R_j \mathbf{1}_{\{R_j \leq \gamma\}} - \frac{S'_K(\gamma)}{JF_K(\gamma)} [\mathbf{1}_{\{R_j \leq \gamma\}} - \alpha]. \tag{A.11}$$

3. Indice de Gini

L'indice de Gini, noté G, est défini par :

$$\text{G} = \frac{2M - R}{JR} - 1, \tag{A.12}$$

où

$$M = \sum_{j \in \mathcal{U}} R_j \left(\sum_{j' \in \mathcal{U}} \mathbf{1}_{\{R_{j'} \leq R_j\}} \right) \tag{A.13}$$

et R est défini par (A.7). La linéarisée ℓ_j intervenant dans l'approximation de la variance de l'estimateur de l'indice de Gini donnée dans la formule (4.1) s'écrit, pour chaque individu j :

$$\ell_j = \frac{JR(2U_j - R_j) - (2M - R)(R + JR_j)}{(JR)^2}, \tag{A.14}$$

où la quantité U_j est définie par

$$U_j = \sum_{j' \in \mathcal{U}} R_{j'} \mathbf{1}_{\{R_j \le R_{j'}\}} + R_j \sum_{j' \in \mathcal{U}} \mathbf{1}_{\{R_{j'} \le R_j\}}. \tag{A.15}$$

4. Équart relatif médian du risque de pauvreté

   L'équart relatif médian du risque de pauvreté, noté ERMRP, est défini par :

$$\text{ERMRP} = \frac{\text{SRP} - \text{MED}^p}{\text{SRP}}, \tag{A.16}$$

   où on rappelle que SRP désigne le seuil de risque de pauvreté et où MED^p est le revenu médian calculé parmi les personnes ayant un revenu inférieur au seuil de pauvreté. La linéarisée ℓ_j intervenant dans l'approximation de la variance de l'estimateur du ERMRP donnée dans la formule (4.1) s'écrit, pour chaque individu j :

$$\ell_j = -\frac{\text{SRP} \cdot Y_j - \text{MED}^p \cdot W_j}{(\text{SRP})^2}, \tag{A.17}$$

   où W_j est défini par

$$W_j = -\frac{0,6}{F'_K(\text{MED})} \frac{1}{J} \left[\mathbf{1}_{\{R_j \le \text{MED}\}} - 0,5 \right] \tag{A.18}$$

   et Y_j vérifie l'équation

$$\begin{aligned} F'_K(\text{MED}^p) Y_j = {} & \frac{1}{2} \left\{ \frac{1}{J} \left[\mathbf{1}_{\{R_j \le \text{SRP}\}} - F(\text{SRP}) \right] + F'_K(\text{SRP}) \cdot W_j \right\} \\ & - \frac{1}{J} \left[\mathbf{1}_{\{R_j \le \text{MED}^p\}} - F(\text{MED}^p) \right]. \end{aligned} \tag{A.19}$$

Annexe B

Démonstration de la proposition 1

La preuve est composée de quatre parties.

1. Définition de $\widehat{\text{cov}}_{jj'}$

   Pour tous longitudinaux j et j' appartenant $s_p^{A_2, t_\tau}$ et faisant partie respectivement des ménages k et k' lors de l'année d'enquête, on pose :

$$\widehat{\mathrm{cov}}_{jj'} = \begin{cases} 0 & \text{si } k \neq k' \\ \dfrac{1}{(L_k + P_k)^2} (e'_k)^2 \dfrac{1 - q_k^b q_k^c}{(q_k^b q_k^c)^2} \mathbf{1}_{\{k \in s_m^{B,\tau}\}} & \text{si } k = k', \end{cases} \tag{B.1}$$

de sorte que

$$\mathrm{E}\left[\widehat{\mathrm{cov}}_{jj'} \mid s_p^{A_2,t_\tau} \right] = \mathrm{cov}_{jj'}. \tag{B.2}$$

2. Estimation de $V_{A_2}^\tau$

   Si les Z_j étaient connus sur $s_p^{A_2,t_\tau}$, on pourrait estimer sans biais $V_{A_2}^\tau$ par $\tilde{V}_{A_2}^\tau(Z_1, Z_2, \ldots)$. Comme ce n'est pas le cas, on estime sans biais $V_{A_2}^\tau$ par :

$$\hat{V}_{A_2}^\tau = \sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} \left[\hat{Z}_j \hat{Z}_{j'} - \widehat{\mathrm{cov}}_{jj'} \right].$$

   En effet, pour tous longitudinaux j et j' appartenant à $s_p^{A_2,t_\tau}$, on a :

$$\mathrm{E}\left[\hat{Z}_j \hat{Z}_{j'} - \widehat{\mathrm{cov}}_{jj'} \mid s_p^{A_2,t_\tau} \right] = Z_j Z_{j'},$$

   ce qui implique

$$\mathrm{E}(\hat{V}_{A_2}^\tau) = \mathrm{E}_{s_p^{A_2,t_\tau}} \left[\mathrm{E}\left(\hat{V}_{A_2}^\tau \mid s_p^{A_2,t_\tau} \right) \right] = \mathrm{E}_{s_p^{A_2,t_\tau}} \left[\tilde{V}_{A_2}^\tau(Z_1, Z_2, \ldots) \right] = V_{A_2}^\tau.$$

3. Estimation de V_{NRG}^τ

   On peut estimer V_{NRG}^τ par :

$$\hat{V}_{\mathrm{NRG}}^\tau = \sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{1}{\pi_j^{A_2}} \frac{1}{\pi_{j'}^{A_2}} \widehat{\mathrm{cov}}_{jj'}.$$

4. Formule finale

   On estime $\mathrm{var}(\hat{T}_\tau)$ par :

$$\widehat{\mathrm{var}}(\hat{T}_\tau) = \hat{V}_{A_2}^\tau + \hat{V}_{\mathrm{NRG}}^\tau.$$

   Après quelques simplifications, il vient :

$$\widehat{\mathrm{var}}(\hat{T}_\tau) = \tilde{V}_{A_2}^\tau(\hat{Z}_1, \hat{Z}_2, \ldots) + \sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{1}{\pi_{jj'}^{A_2}} \widehat{\mathrm{cov}}_{jj'}. \tag{B.3}$$

   En utilisant (B.1) on peut simplifier le deuxième terme du membre de droite, de façon à obtenir la formule (5.19) donnée dans la proposition 1.

Bibliographie

Ardilly, P., et Lavallée, P. (2007). Pondération dans les échantillons rotatifs : le cas de l'Enquête SILC en France. *Techniques d'enquête*, 33, 2, 149-156.

Caron, N., Deville, J. et Sautory, O. (1998). Estimation de données issues d'enquêtes : document méthodologique sur le logiciel POULPE. Rapport technique 9806, INSEE, Paris.

Clemenceau, A., et Museux, J. (2006). EU-SILC (Community Statistics on Income and Living Conditions): General presentation of the instrument. Dans *Proceedings of the EU-SILC Conference* (Helsinki, 6 au 8 novembre 2006).

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 2, 219-230.

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

Eurostat (2003). *Règlement (CE) No 1177/2003 du parlement européen et du conseil du 16 juin 2003 relatif aux statistiques communautaires sur le revenu et les conditions de vie (EU-SILC)*. Luxembourg, Journal officiel de l'Union européenne.

Eurostat (2010). 2008 comparative EU intermediate quality report. Rapport technique, European Commission, Luxembourg. http://circa.europa.eu/Public/irc/dsis/eusilc/library?l=/quality_assessment/comparative_quality_1/intermediate_version/_EN_1.0_&a=d.

Graf, E. (2008). Pondérations du SILC pilote, SILC i vague 2, SILC ii vague 1, SILC i et SILC ii combinés. Rapport technique 338-0051, OFS, Neuchâtel. http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/methodenberichte.Document.105401.pdf.

Langel, M., et Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, 141, 2976-2985.

Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*. Paris : Ellipses.

Merkouris, T. (2001). Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples. *Techniques d'enquête*, 27, 2, 189-200.

OFS (2010). *Les conditions de vie en Suisse en 2009, Résultats de l'enquête sur les revenus et les conditions de vie* (*SILC*). No de commande 1192-0900. http://www.bfs.admin.ch/bfs/portal/fr/index/news/publikationen.Document.138922.pdf.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 167-195.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley & Sons, Inc.

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 169-206
Statistique Canada, N° 12-001-X au Catalogue

Combiner des cohortes dans les enquêtes longitudinales

Iván A. Carrillo et Alan F. Karr[1]

Résumé

Une question fréquente concernant les enquêtes longitudinales est celle de savoir comment combiner les différentes cohortes. Dans le présent article, nous présentons une nouvelle méthode qui permet de combiner différentes cohortes et d'utiliser toutes les données à notre disposition dans une enquête longitudinale pour estimer les paramètres d'un modèle semi-paramétrique qui relie la variable réponse à un jeu de covariables. La procédure s'appuie sur la méthode des équations d'estimation généralisées pondérées pour traiter les données manquantes pour certaines vagues dans les enquêtes longitudinales. Notre méthode s'appuie, pour l'estimation des paramètres du modèle, sur un cadre de randomisation conjointe qui tient compte à la fois du modèle de superpopulation et de la sélection aléatoire selon le plan de sondage. Nous proposons aussi une méthode d'estimation de la variance sous le plan et sous randomisation conjointe. Pour illustrer la méthode, nous l'appliquons à l'enquête Survey of Doctorate Recipients réalisée par la National Science Foundation des États-Unis.

Mots clés : Paramètres de superpopulation; inférence sous randomisation conjointe; estimation de variance par rééchantillonnage; enquêtes à panel rotatif; enquête longitudinale à cohortes multiples; équations d'estimation généralisées pondérées.

1 Introduction

L'enquête Survey of Doctorate Recipients (SDR) est une enquête longitudinale menée par la National Science Foundation (NSF) dont le plan de sondage comprend des caractéristiques des panels répétés ainsi que des panels rotatifs. L'objectif de l'enquête est d'étudier les titulaires d'un doctorat en sciences, en génie ou en sciences de la santé aux États-Unis. Elle est réalisée tous les deux ans environ. Une description détaillée de la SDR peut être consultée dans NSF (2012). Dans le présent article, nous nous concentrons sur les données recueillies de 1995 à 2008 (7 vagues).

Une nouvelle cohorte est sélectionnée à l'occasion de chacune des vagues de l'enquête. La nouvelle cohorte est composée d'un échantillon de personnes récemment diplômées (au cours des deux années précédentes) tiré du Doctorate Records File, une base de données construite principalement d'après les données de l'enquête Survey of Earned Doctorates

1. Iván A. Carrillo et Alan F. Karr, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709, États-Unis. Courriel : ivan@niss.org et karr@niss.org.

(http://www.nsf.gov/statistics/srvydoctorates/). Les personnes sélectionnées sont gardées dans l'échantillon, c'est-à-dire interviewées tous les deux ans, jusqu'à l'âge de 75 ans, à condition qu'elles vivent aux États-Unis durant la semaine de référence de l'enquête et qu'elles ne soient pas placées en établissement. Cependant, les diplômés échantillonnés qui satisfont à ces critères ne sont pas *tous* gardés indéfiniment. Certaines personnes, plutôt que des cohortes entières, sont supprimées de l'échantillon afin a) d'inclure les nouveaux diplômés dans de nouvelles cohortes et b) de maintenir une taille d'échantillon relativement constante d'une vague à l'autre. À la section 2.2, nous décrivons comment sont sélectionnées les personnes qui sont supprimées de l'échantillon.

Des poids de sondage existent déjà pour les analyses transversales des données de la SDR, mais non pour les analyses longitudinales. Au lieu de nécessiter un *nouveau* poids longitudinal pour *toutes* les données, la méthode proposée ici permet d'utiliser les poids transversaux existants pour les analyses longitudinales en n'ignorant aucune donnée. Nous nous concentrons sur l'estimation des paramètres des modèles statistiques de l'effet des covariables sur une réponse d'intérêt, mais la méthode peut également être utilisée pour estimer des quantités de population finie (Carrillo et Karr 2012). Notre analyse est axée sur la SDR, mais notre méthode est applicable à n'importe quelle enquête à panel fixe, à panel fixe plus « nouvelles unités », à panel répété, à panel rotatif, à panel divisé ou à renouvellement de l'échantillon, à condition qu'il existe pour chaque vague un poids transversal pour représenter la population d'intérêt de la vague en question. Voir Smith, Lynn et Elliot (2009), Hirano, Imbens, Ridder et Rubin (2001), et Nevo (2003) pour les définitions de tous ces types de plans de sondage longitudinaux.

La SDR est un hybride de plan à panel répété et de plan à panel rotatif. Il ne s'agit pas purement d'un plan à panel répété, à cause de la suppression de certains sujets à chaque vague. Il ne s'agit pas purement d'un plan à panel rotatif, parce que l'on supprime des personnes et *non* des panels (ou cohortes) complets; en outre, la composition de la population finie d'intérêt évolue au cours du temps, ce qui n'est pas le cas d'une enquête à panel rotatif.

Diggle, Heagerty, Liang et Zeger (2002) et Hedeker et Gibbons (2006) font remarquer que dans les enquêtes longitudinales, contrairement aux études transversales, il est possible d'isoler l'effet de l'âge (changement réel chez les sujets au cours du temps) et l'effet de cohorte (différences entre les unités au début de la période d'étude).

Hedeker et Gibbons (2006) laissent aussi entendre que, puisque les études longitudinales permettent de mesurer des variables explicatives (covariables) variant au cours du temps, les inférences statistiques au sujet de la relation dynamique entre le résultat d'intérêt (réponse) et ses covariables sont beaucoup plus puissantes que celles fondées sur des études transversales.

Si nous nous intéressons à la moyenne marginale d'une variable, éventuellement en conditionnant sur certaines covariables, plutôt qu'à la mesure du changement, une étude longitudinale n'est pas nécessaire; une étude transversale suffit. Cependant, même dans ce cas, une étude longitudinale a tendance à être plus puissante, parce que chaque sujet peut être utilisé comme son propre contrôle pour toute caractéristique non mesurée (Diggle et coll. 2002).

Notre approche diffère des solutions décrites dans la littérature, lesquelles présentent certaines limites pour l'analyse de ce genre de données, et en particulier pour l'application à la SDR. Par exemple, dans Berger (2004a) et Berger (2004b), l'estimation du changement est examinée en détail en utilisant des échantillons rotatifs, mais en posant que la composition de la population finie ne varie pas au cours du temps, ce qui n'est pas le cas de la SDR. Cette hypothèse ne tient pas non plus dans de nombreuses autres enquêtes à grande échelle. En outre, la méthodologie proposée par Berger n'est pas facilement généralisable à plus de deux vagues d'enquête. Similairement, Qualité et Tillé (2008) supposent que la population finie est fixe au cours du temps. Hirano et coll. (2001) et Nevo (2003) présentent diverses méthodes d'estimation en supposant que le plan est à panel fixe avec rafraîchissement pour tenir compte de l'attrition, mais émettent aussi l'hypothèse que la composition de la population finie est fixe au cours du temps.

McLaren et Steel (2000), et Steel et McLaren (2007) utilisent une approche fondée sur des séries chronologiques pour estimer la variation et la tendance dans les données d'enquête. Bien

que leur approche permette d'intégrer l'association intrasujet dans les estimations ponctuelles, ils ne considèrent pas de covariables dans leur modèle (sauf les covariables temporelles implicites). En outre, ils discutent uniquement de l'estimation du changement pour les variables continues.

Une autre option pour analyser des données longitudinales consiste à considérer la population finie d'intérêt comme étant fixe, sauf peut-être en ce qui concerne les décès, qui pourraient être permis. Les études de ce genre sont celles pour lesquelles des données n'existent que pour une seule cohorte. Par exemple, Vieira et Skinner (2008), Carrillo, Chen et Wu (2010), et Carrillo, Chen et Wu (2011) illustrent certaines options de modélisation en se basant sur des données d'enquête recueillies auprès d'une seule cohorte. Cependant, pour procéder à ce genre d'analyse sur les données d'enquête à plusieurs cohortes, on doit ignorer certaines (ou de nombreuses) données existantes, par exemple celles recueillies auprès des sujets qui ne sont pas présents à toutes les vagues. Un exemple de procédure de pondération de ce genre est décrit dans Ardilly et Lavallée (2007).

Enfin, l'approche de Larsen, Qing, Zhou et Foulkes (2011) est séduisante en principe, parce qu'il s'agit de la façon dont procèdent généralement les praticiens des sondages. Un poids initial est ajusté, entre autres par calage sur des totaux connus, ici sur des totaux par vague d'enquête. Néanmoins, pour les panels rotatifs, cette méthode en est encore à ses balbutiements; la manière d'exécuter certains de ses éléments n'est pas entièrement claire. Ainsi, le choix du poids initial n'est pas évident : un poids constant ?, le premier poids disponible ?, la moyenne des poids disponibles pour chaque cas ?, ou le dernier poids disponible ? En outre, en cas de décrochages, comme il en existe dans la SDR, les auteurs ne précisent pas comment procéder à un ajustement pour la non-réponse. De surcroît, on s'explique mal pourquoi un ajustement pour la non-réponse des décrocheurs, disons, à la vague 4 devrait avoir une influence sur les observations à la vague 3, comme le permet cette méthode, puisqu'elle comporte un poids unique pour chaque sujet. De plus, les auteurs mentionnent qu'ils ont estimé les erreurs-types, mais ils n'indiquent pas comment tenir compte de toutes les caractéristiques du plan de sondage, telles que les

modifications apportées au cours du temps à la stratification et aux classes pour l'ajustement des pondérations de la SDR. En revanche, notre méthode utilise uniquement des pondérations et des méthodes d'estimation de variance transversales, qui ont été étudiées en profondeur dans la documentation et auxquelles on a facilement accès pour la SDR.

La présentation du reste de l'article est la suivante. À la section 2, nous décrivons le plan de sondage de la SDR. À la section 3, nous proposons une nouvelle approche pour l'analyse longitudinale des modèles de moyenne marginale dans le cas d'enquêtes à plusieurs cohortes. À la section 4, nous présentons l'application de la méthode à la SDR. Enfin, à la section 5, nous offrons quelques points de discussion.

2 Le plan de sondage de la SDR

2.1 Population finie

La population finie d'intérêt de la SDR peut être représentée comme au tableau 2.1. À la vague 1, c'est-à-dire la première période d'intérêt, il existe un ensemble fini, $U_{1(1)} = U_1$, de $N_{1(1)} = N_1$ titulaires d'un doctorat, obtenu récemment ou non, qui satisfont aux exigences de la SDR.

Tableau 2.1
Population finie de la SDR

j:	1		2		3		…		$J-1$		J
	$U_{1(1)}$	$\supseteq$	$U_{2(1)}$	$\supseteq$	$U_{3(1)}$	$\supseteq$	$\cdots$	$\supseteq$	$U_{J-1(1)}$	$\supseteq$	$U_{J(1)}$
	$N_{1(1)}$	$\geq$	$N_{2(1)}$	$\geq$	$N_{3(1)}$	$\geq$	$\cdots$	$\geq$	$N_{J-1(1)}$	$\geq$	$N_{J(1)}$
			$U_{2(2)}$	$\supseteq$	$U_{3(2)}$	$\supseteq$	$\cdots$	$\supseteq$	$U_{J-1(2)}$	$\supseteq$	$U_{J(2)}$
			$N_{2(2)}$	$\geq$	$N_{3(2)}$	$\geq$	$\cdots$	$\geq$	$N_{J-1(2)}$	$\geq$	$N_{J(2)}$
						$\ddots$			$\vdots$		$\vdots$
									$U_{J-1(J-1)}$	$\supseteq$	$U_{J(J-1)}$
									$N_{J-1(J-1)}$	$\geq$	$N_{J(J-1)}$
											$U_{J(J)}$
											$N_{J(J)}$
	U_1		U_2		U_3		$\cdots$		U_{J-1}		U_J
	N_1		N_2		N_3		$\cdots$		N_{J-1}		N_J

À la vague 2, un sous-ensemble seulement des sujets compris dans $U_{1(1)}$ satisfont encore aux exigences de la SDR; nous appelons ce sous-ensemble de $N_{2(1)}$ sujets, $U_{2(1)}$. En outre, il existe un ensemble de nouveaux titulaires d'un doctorat, qui ont obtenu leur diplôme depuis la vague 1, et qui satisfont aussi aux autres exigences de l'enquête. Cet ensemble de nouveaux diplômés dans le champ de l'enquête est appelé $U_{2(2)}$ et est de taille $N_{2(2)}$. Par conséquent, à la vague 2, il y a un total de $N_2 = N_{2(1)} + N_{2(2)}$ sujets dans la population d'intérêt $U_2 = U_{2(1)} \cup U_{2(2)}$.

À la vague suivante, la vague 3, le même processus a lieu. Certains sujets compris dans $U_{2(1)}$ ont quitté la population d'intérêt et il n'en reste que $N_{3(1)}$ dans $U_{3(1)}$. La même chose se produit avec l'ensemble $U_{2(2)}$; seulement un sous-ensemble $U_{3(2)}$ de $N_{3(2)}$ sujets satisfera encore aux exigences de la SDR. En outre, $N_{3(3)}$ diplômés récents entrent dans la population d'intérêt; cet ensemble est appelé $U_{3(3)}$. En tout, la population finie d'intérêt à la vague 3 est $U_3 = U_{3(1)} \cup U_{3(2)} \cup U_{3(3)}$, avec $N_3 = N_{3(1)} + N_{3(2)} + N_{3(3)}$ sujets.

Cette procédure de réduction des cohortes anciennes et d'ajout de nouvelles cohortes se poursuit jusqu'à la dernière vague d'intérêt, la vague J. Nous constatons que la population finie d'intérêt change à chaque vague, principalement pour deux raisons. Premièrement, certains sujets appartenant aux anciennes cohortes ne sont plus dans le champ de la vague courante et ne font pas partie de la population cible courante. Deuxièmement, des diplômés récents sont ajoutés à la population cible de la vague courante. Nous désignons par $j = 1, 2, \ldots, J$ la vague d'intérêt (hors des parenthèses) et par $j' = 1, 2, \ldots, J$ la cohorte à laquelle le sujet appartient (entre parenthèses), et par conséquent $U_{j(j')} = U_{\text{vague(cohorte)}}$.

2.2 Échantillonnage

Le plan d'échantillonnage de la SDR possède une structure similaire à celle de la population finie et est illustré au tableau 2.2. À la vague 1, un échantillon (complexe) $s_{1(1)} = s_1$ de $n_{1(1)} = n_1$ sujets est sélectionné parmi les N_1 éléments de U_1. Chaque élément i dans s_1 est interviewé et les données qu'il fournit sont recueillies; en outre, il existe un poids de sondage

$w_{i1} = 1 / \pi_{i1}$ associé à l'élément, qui est l'inverse de la probabilité d'inclusion de ce dernier dans l'échantillon à la vague 1.

Tableau 2.2
Échantillon de la SDR

j :	1		2		3		$\cdots$		$J-1$		J
	$s_{1(1)}$	$\supseteq$	$s_{2(1)}$	$\supseteq$	$s_{3(1)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(1)}$	$\supseteq$	$s_{J(1)}$
	$n_{1(1)}$	$\geq$	$n_{2(1)}$	$\geq$	$n_{3(1)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(1)}$	$\geq$	$n_{J(1)}$
			$s_{2(2)}$	$\supseteq$	$s_{3(2)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(2)}$	$\supseteq$	$s_{J(2)}$
			$n_{2(2)}$	$\geq$	$n_{3(2)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(2)}$	$\geq$	$n_{J(2)}$
					$s_{3(3)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(3)}$	$\supseteq$	$s_{J(3)}$
					$n_{3(3)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(3)}$	$\geq$	$n_{J(3)}$
							$\ddots$		$\vdots$		$\vdots$
									$s_{J-1(J-1)}$	$\supseteq$	$s_{J(J-1)}$
									$n_{J-1(J-1)}$	$\geq$	$n_{J(J-1)}$
											$s_{J(J)}$
											$n_{J(J)}$
	s_1		s_2		s_3		$\cdots$		s_{J-1}		s_J
	n_1		n_2		n_3		$\cdots$		n_{J-1}		n_J

À la deuxième vague, les éléments compris dans $s_{1(1)}$ qui ne sont plus dans le champ de l'enquête sont simplement supprimés de la base de sondage (mais les observations les concernant faites à la vague 1 sont gardées), et un sous-échantillon $s_{2(1)}$, de taille $n_{2(1)}$, des sujets encore dans le champ de l'enquête est sélectionné. Les membres de $s_{1(1)}$ qui sont encore dans le champ de l'enquête à la vague 2 ne sont pas tous gardés dans l'échantillon, et ce pour permettre d'ajouter l'échantillon de nouveaux titulaires d'un doctorat tout en maintenant plus ou moins la même taille d'échantillon qu'à la vague 1. Un échantillon $s_{2(2)}$ de taille $n_{2(2)}$ est tiré de $U_{2(2)}$; les sujets compris dans $s_{2(2)}$ forment la deuxième cohorte. L'échantillon total à la vague 2 est $s_2 = s_{2(1)} \cup s_{2(2)}$, dont la taille est $n_2 = n_{2(1)} + n_{2(2)}$, qui est approximativement égale à n_1. Tous les sujets compris dans s_2 sont interviewés à la vague 2. Les poids de sondage à la vague 2, $w_{i2} = 1 / \pi_{i2}$, sont tels que l'échantillon s_2 représente la population d'intérêt à la vague 2, à savoir U_2.

La même procédure est répétée à chaque vague jusqu'à la dernière (J), où un sous-échantillon des sujets restants en provenance de chacune des $J-1$ cohortes antérieures est sélectionné, et un nouvel échantillon (nouvelle cohorte) $s_{J(J)}$ de diplômés récents est tiré de $U_{J(J)}$. À la dernière vague, tous les sujets compris dans $s_J = \bigcup_{j'=1}^{J} s_{J(j')}$ sont interviewés et un poids de sondage $w_{iJ} = 1 / \pi_{iJ}$ est créé pour chacun, de sorte que s_J représente la population finie U_J.

En ce qui concerne la façon dont sont sélectionnés les sujets supprimées de l'échantillon, selon NSF (2012), en 2008 par exemple, le sous-échantillon $s_{08} \setminus s_{08(08)}$ a été sélectionné en stratifiant s_{06} « en 150 strates en fonction de 3 variables : groupe démographique, domaine du diplôme et sexe. » Le rapport explique aussi que :

- l'ancienne pratique consistant à sélectionner l'échantillon avec probabilité proportionnelle à la taille s'est poursuivie, la mesure de taille étant le poids de base associé au cycle précédent de l'enquête. Pour chaque strate, l'algorithme d'échantillonnage a commencé par repérer et supprimer les cas autoreprésentatifs selon une procédure itérative. Ensuite, dans chaque strate, les cas non autoreprésentatifs ont été triés en fonction de la citoyenneté, de l'état d'incapacité, du domaine du diplôme et de l'année d'obtention du diplôme de doctorat. Enfin, le solde de l'échantillon (c'est-à-dire le total attribué à la strate moins le nombre de cas autoreprésentatifs) a été sélectionné dans chaque strate systématiquement avec probabilité proportionnelle à la taille.

Il convient de mentionner que, jusqu'à 1989, la cohorte (ou plus précisément l'année d'obtention du diplôme) faisait partie des variables de stratification (et des cellules d'ajustement des poids), mais qu'à partir de 1991, elle ne l'a plus été; elle a été remplacée par l'état d'incapacité. Pour des renseignements plus détaillés sur la procédure de sous-échantillonnage, y compris la description de la répartition de l'échantillon, voir NSF (2012) ou Cox, Grigorian, Wang et Harter (2010).

La description qui précède montre clairement que la SDR n'est pas réalisée selon un plan à panel rotatif. Outre le fait que la composition de la population finie d'intérêt évolue avec le

temps, un plan à panel rotatif donnerait lieu à la sélection, au temps j, d'une nouvelle cohorte à partir de U_j, et non à partir de $U_j \setminus U_{j-1}$ comme cela est le cas dans la SDR.

Une autre particularité de la SDR est qu'à chaque vague j, une base de sondage de diplômés récents $U_{j(j)}$ existe, de laquelle peut être tirée directement la nouvelle cohorte $s_{j(j)}$. Cependant, dans d'autres applications, le coût de la création d'une telle base de sondage, par exemple une liste de nouveaux membres, peut être excessif (particulièrement lorsqu'il est cumulé sur l'ensemble des vagues), et la nouvelle cohorte doit parfois être sélectionnée à partir de U_j (par opposition à $U_{j(j)}$). La méthode proposée dans le présent article peut également être appliquée à ce genre de cas, à condition que l'on puisse créer pour l'échantillon total à la vague j, s_j, un poids transversal pour représenter U_j. Nous discutons de cet aspect plus en détail à la section 3.2.

Soulignons que, dans la notation $s_{j(j')}$, la quantité j représente la vague à laquelle l'échantillon se rapporte, et j' désigne la cohorte de l'échantillon, c'est-à-dire la vague à laquelle l'échantillon a été sélectionné initialement. La notation pour les pondérations est w_{ij}, où le premier indice inférieur désigne le sujet et le second, la vague d'intérêt, quelle que soit la période où le sujet a été sélectionné initialement.

3 Méthodologie

3.1 Motivation

Supposons que (hors du contexte d'une enquête) l'on s'intéresse au paramètre vectoriel $\boldsymbol{\beta}$ de dimension $p \times 1$ dans le modèle suivant :

$$\xi : \begin{cases} E[Y_{ij} \mid X_{ij}] = \mu_{ij} = g^{-1}(X'_{ij}\boldsymbol{\beta}), & j = 1,2,\ldots,J, i = 1,2,\ldots \\ \mathrm{Var}[Y_{ij} \mid X_{ij}] = \phi\nu(\mu_{ij}), & j = 1,2,\ldots,J, i = 1,2,\ldots \\ \mathrm{Cov}[Y_i \mid X_i] = \Sigma_i, & i = 1,2,\ldots \\ Y_k \perp Y_l \mid X_k, X_l, & k \neq l = 1,2,\ldots; \end{cases} \tag{3.1}$$

où Y_{ij} est la variable réponse pour le sujet i à la vague j, X_{ij} est un vecteur de covariables de dimension $p \times 1, Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$, $X_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})$ est une matrice de dimensions $p \times J$, $g(\cdot)$ est une « fonction de lien » monotone de type un à un différenciable, $\nu(\cdot)$ est la « fonction de variance » de forme connue, et $\phi > 0$ est le « paramètre de dispersion ». Puisqu'en général, la matrice de covariance Σ_i de dimensions $J \times J$ est difficile à spécifier, nous la modélisons sous la forme $\mathrm{Cov}[Y_i \mid X_i] = V_i = A_i^{1/2}\mathbf{R}(\alpha)A_i^{1/2}$, une matrice de covariance « de travail », où $A_i = \mathrm{diag}[\phi\nu(\mu_{i1}), \phi\nu(\mu_{i2}), \ldots, \phi\nu(\mu_{iJ})]$ et $\mathbf{R}(\alpha)$ est une matrice de corrélation « de travail », toutes deux de dimensions $J \times J$, et α est un vecteur qui caractérise entièrement $\mathbf{R}(\alpha)$ (voir Liang et Zeger 1986).

Pour estimer $\boldsymbol{\beta}$, nous sélectionnons un échantillon (cohorte unique) de n éléments à partir du modèle ξ et nous mesurons (avons l'intention de mesurer) chacun d'eux à J occasions. Si tous les éléments de l'échantillon répondent à chaque occasion j, la tâche peut être achevée en appliquant la méthode aux équations d'estimation généralisées (EEG) habituelles de Liang et Zeger (1986). Cependant, dans toute étude, il est rare que tous les sujets répondent à toutes les vagues. Il est plus fréquent que certains éléments de l'échantillon décrochent de l'étude.

Dans ces conditions, et en supposant que les réponses manquantes peuvent être considérées comme manquant au hasard ou MAR (pour *missing at random*) (voir Rubin 1976), en particulier que le décrochage durant une vague donnée ne dépend pas de la valeur courante (non observée), Robins, Rotnitzky et Zhao (1995) ont proposé d'estimer $\boldsymbol{\beta}$ en résolvant les équations d'estimation $\sum_{i=1}^{n}(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})V_i^{-1}\hat{\Delta}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, où $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iJ})'$, $\hat{\Delta}_i = \mathrm{diag}[R_{i1}\hat{q}_{i1}^{-1}, R_{i2}\hat{q}_{i2}^{-1}, \ldots, R_{iJ}\hat{q}_{iJ}^{-1}]$, R_{ij} est l'indicateur de réponse pour le sujet i à la vague j, et $\hat{q}_{ij}$ est une estimation de la probabilité que le sujet i soit observé durant la vague j.

Pour les applications d'enquête, on utiliserait l'équation d'estimation $\sum_{i \in s}[w_i(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})\, V_i^{-1}\hat{\Delta}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i)] = \mathbf{0}$, où w_i est le poids de sondage du sujet i. Une autre façon d'écrire cette équation est $\sum_{i \in s}(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})\, V_i^{-1}\hat{\Delta}_{wi}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, avec $\hat{\Delta}_{wi} = \mathrm{diag}[w_iR_{i1}\hat{q}_{i1}^{-1}, w_iR_{i2}\hat{q}_{i2}^{-1}, \ldots, w_iR_{iJ}\hat{q}_{iJ}^{-1}]$.

Nous constatons que les éléments de la diagonale de $\hat{\Delta}_{wi}$ sont simplement égaux aux poids de sondage propres à la vague non corrigés de la non-réponse quand le sujet est observé et sont égaux à zéro quand le sujet est manquant. Cette caractéristique suggère en soi une solution au problème des cohortes multiples, qui sera présentée à la section suivante.

3.2 Une nouvelle approche pour combiner les cohortes dans les enquêtes longitudinales

Compte tenu de la discussion de la section précédente, si nous avons une enquête à panel fixe, à panel fixe plus des « unités nouvelles », à panel répété, à panel rotatif, à panel divisé ou à renouvellement de l'échantillon, nous proposons d'estimer le paramètre de superpopulation $\boldsymbol{\beta}$ dans le modèle ξ par la solution des équations d'estimation :

$$\Psi_s(\boldsymbol{\beta}) = \sum_{i \in s} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} W_i (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3.2}$$

où la sommation est faite sur l'échantillon s, c'est-à-dire sur les éléments sélectionnés (pour la première fois) dans l'un des échantillons $s_{1(1)}, s_{2(2)}, \ldots, s_{J(J)}$. La matrice diagonale W_i est $W_i = \text{diag}[I_i(U_1)w_{i1}, I_i(U_2)w_{i2}, \ldots, I_i(U_J)w_{iJ}]$, où w_{ij} est le poids transversal (corrigé pour la non-réponse) pour le sujet i à la vague j (à condition que le sujet i fasse partie de l'échantillon s_j) et $I_i(U_j)$ est l'indicateur signalant si le sujet i appartient ou non à la population finie U_j. À la section 3.2.1, nous présentons des arguments justifiant qu'il s'agit d'une procédure d'estimation raisonnable, et à la section 3.2.2, nous discutons du problème des valeurs manquantes.

Les poids transversaux w_{ij}, dans W_i, sont tels que l'échantillon s_j représente U_j lorsqu'il est utilisé avec lesdits poids. Cela signifie que, pour chaque observation i dans l'échantillon s_j, il existe un poids de sondage w_{ij}, qui pourrait être considéré comme le nombre d'unités que cette observation représente dans U_j. Cependant, rappelons que l'échantillon s_j est composé de différents ensembles de sujets, ou différents sous-échantillons (les différentes cohortes), et que

l'intégration de ces sous-échantillons en une variable de pondération transversale unique w_{ij} n'est pas forcément une tâche facile.

Pour la SDR, la construction d'un poids transversal pour la vague j n'est pas trop compliquée, parce que les diverses cohortes sont sélectionnées indépendamment les unes des autres, à partir de populations non chevauchantes. Dans ces conditions, le poids de base est facile à calculer, et tout ce qu'il reste à faire est la conversion pour tenir compte d'aspects tels que l'attrition et le calage sur des totaux connus de la population U_j.

Par ailleurs, dans d'autres situations, par exemple lorsqu'il n'existe pas de liste des *nouveaux* membres, la nouvelle cohorte doit parfois être sélectionnée dans la population globale au moment de la vague en question, ou en se servant d'une base de sondage contenant les nouveaux membres *plus* certains anciens membres, ou à partir de bases de sondage multiples. Le cas échéant, la construction de poids transversaux n'est pas nécessairement simple, et la théorie des bases de sondage multiples peut devoir être appliquée. Nous renvoyons le lecteur aux travaux de Lohr (2007) et de Rao et Wu (2010), ainsi qu'aux références mentionnées dans ces articles, pour les cas de ce genre.

L'expression (3.2) est une généralisation de l'équation (2.25) donnée dans Vieira (2009). Cette dernière n'est applicable que si le nombre d'observations est le même pour tous les sujets ou que toute réponse manquante peut être considérée comme manquant entièrement au hasard ou MCAR (pour *missing completely at random*) (voir Rubin 1976). Comme il est discuté dans Robins et coll. (1995), l'utilisation d'une telle équation lorsque les réponses manquantes ne sont pas de type MCAR produit des estimateurs non convergents; par conséquent, sous un schéma de rotation tel que celui de la SDR, où les sujets ne sont pas tous supprimés (ou gardés) avec les mêmes probabilités, son utilisation ne serait pas appropriée. La question de l'adéquation de l'équation (3.2) dans ce cas et quand des réponses manquent est abordée aux sections 3.2.1 et 3.2.2, respectivement. Si tous les sujets possèdent des poids transversaux qui ne varient pas au

cours du temps (ou qu'ils possèdent un seul poids longitudinal), l'équation (3.2) se réduit à l'équation (2.25) donnée dans Vieira (2009).

3.2.1 Absence de biais

La propriété d'absence de biais de la fonction d'estimation est importante, parce que, comme le soutient Song (2007, section 5.4), il s'agit de l'hypothèse la plus cruciale en vue d'obtenir un estimateur convergent.

Définissons $\boldsymbol{\beta}_N$, ledit « estimateur par recensement », comme étant la solution de l'équation d'estimation en population finie suivante :

$$\Psi_U(\boldsymbol{\beta}_N) = \sum_{i \in U} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_N} V_i^{-1} \mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)) = \mathbf{0}, \tag{3.3}$$

où la somme est calculée sur U, c'est-à-dire sur tous les éléments qui sont devenus membres de la population cible dans l'une des $U_{1(1)}, U_{2(2)}, \ldots, U_{J(J)}$, et $\mathrm{I}_i(\mathrm{U}) = \mathrm{diag}[I_i(U_1), I_i(U_2), \ldots I_i(U_J)]$. Afin de montrer l'absence de biais sous le plan de la fonction d'estimation $\Psi_s(\boldsymbol{\beta})$, nous devons montrer que son espérance sous le plan est $\Psi_U(\boldsymbol{\beta})$ pour tout $\boldsymbol{\beta}$.

Les caractéristiques du plan d'échantillonnage d'une enquête longitudinale peuvent être vues comme celles d'un échantillon à plusieurs phases tel que l'ont montré Särndal, Swensson et Wretman (1992, section 9.9). Par conséquent, nous utilisons la méthodologie d'échantillonnage à plusieurs phases pour les calculs. Nous supposons, sans perte de généralité, que l'enquête ne comprend que trois vagues; les calculs pour trois vagues seulement montrent les tendances pour J, général, en ce qui concerne l'absence de biais et la variance.

Comme nous l'avons mentionné plus haut, nous supposons que w_{ij} est le poids transversal pour le sujet i à la vague j, si ce sujet appartient à s_j, et zéro autrement. Partant de la théorie de l'échantillonnage à plusieurs phases, nous avons que, pour $i \in s_{1(1)}$, $w_{i1} = \pi_{i1}^{-1}$, $w_{i2} = \pi_{i1}^{-1}\pi_{i2|s_{1(1)}}^{-1}$ et $w_{i3} = \pi_{i1}^{-1}\pi_{i2|s_{1(1)}}^{-1}\pi_{i3|s_{2(1)}}^{-1}$, pour $i \in s_{2(2)}$, $w_{i2} = \pi_{i2}^{-1}$ et $w_{i3} = \pi_{i2}^{-1}\pi_{i3|s_{2(2)}}^{-1}$, et pour $i \in s_{3(3)}$,

$w_{i3} = \pi_{i3}^{-1}$, où π_{ij} est la probabilité d'inclusion du sujet i dans l'échantillon $s_{j(j)}$ et $\pi_{ij|s_{j-1(j')}}$ est la probabilité d'inclusion conditionnelle du sujet i dans l'échantillon $s_{j(j')}$ sachant $s_{j-1(j')}$.

En utilisant $E_p(\cdot)$ pour désigner l'espérance par rapport au plan d'échantillonnage, nous avons :

$$E_p\left[\sum_{i\in s}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\right] = E_p\left[\sum_{j=1}^{3}\sum_{i\in s_{j(j)}}B_iW_i\boldsymbol{e}_i\right], \tag{3.4}$$

où $B_i = (\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})V_i^{-1}$ et $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i$. Par exemple, pour $\sum_{i\in s_{2(2)}}B_iW_i\boldsymbol{e}_i$, nous obtenons :

$$\begin{aligned} E_p\left[\sum_{i\in s_{2(2)}}B_iW_i\boldsymbol{e}_i\right] &= E\left\{E\left[\sum_{i\in U_{2(2)}}B_iD_i\boldsymbol{e}_i \mid s_{2(2)}\right]\right\} = E\left\{\sum_{i\in U_{2(2)}}B_iD_i^*\boldsymbol{e}_i\right\} \\ &= \sum_{i\in U_{2(2)}}B_iD_i^{**}\boldsymbol{e}_i \overset{\text{def}}{=} \sum_{i\in U_{2(2)}}B_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i, \end{aligned} \tag{3.5}$$

où $D_i = \mathrm{diag}[0, I_i(U_2)\, w_{i2}I_i(s_{2(2)}), I_i(U_3)\, w_{i3}I_i(s_{3(2)})\, I_i(s_{2(2)})]$, $D_i^* = \mathrm{diag}[0, (I_i(U_2)\, w_{i2} \times I_i(s_{2(2)})), (I_i(U_3)\pi_{i3|s_{2(2)}}I_i(s_{2(2)})) / (\pi_{i2}\pi_{i3|s_{2(2)}})]$, et $D_i^{**} = \mathrm{diag}[0, (I_i(U_2)\pi_{i2}) / \pi_{i2}, (I_i(U_3) \times \pi_{i2}) / \pi_{i2}]$; similairement, nous pouvons montrer que $E_p\left[\sum_{i\in s_{1(1)}}B_iW_i\boldsymbol{e}_i\right] = \sum_{i\in U_{1(1)}}B_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$ et $E_p\left[\sum_{i\in s_{3(3)}}B_iW_i\boldsymbol{e}_i\right] = \sum_{i\in U_{3(3)}}B_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$. De ces expressions et de l'équation (3.4), nous concluons que $E_p[\Psi_s(\boldsymbol{\beta})] = \Psi_U(\boldsymbol{\beta})$ pour tout $\boldsymbol{\beta}$, ce qui signifie que la fonction d'estimation $\Psi_s(\boldsymbol{\beta})$ est sans biais sous le plan pour la fonction d'estimation en population finie.

En outre, comme la cible de l'inférence est le paramètre de superpopulation, nous devons garantir que le modèle pour μ_{ij} est tel que l'expression $E_\xi(Y_{ij} - \mu_{ij}) = 0$ est satisfaite, où $E_\xi(\cdot)$ représente l'espérance sous le modèle ξ car, si cela est le cas, nous avons :

$$E_{\xi p}[\Psi_s(\boldsymbol{\beta})] \overset{\text{def}}{=} E_\xi E_p[\Psi_s(\boldsymbol{\beta})] = E_\xi[\Psi_U(\boldsymbol{\beta})] = \sum_{i\in U}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\mathrm{I}_i(\mathrm{U})E_\xi(\boldsymbol{y}_i-\boldsymbol{\mu}_i) = \mathbf{0},$$

de sorte que la fonction d'estimation $\Psi_s(\boldsymbol{\beta})$ est sans biais par rapport au modèle et au plan. La contrainte $E_\xi(Y_{ij} - \mu_{ij}) = 0$ signifie que le modèle de la moyenne doit être spécifié correctement; par conséquent, il faut faire attention aux tests diagnostiques sur les résidus pour le modèle particulier qui est ajusté.

3.2.2 Une remarque concernant la non-réponse

Dans le cas de la SDR, comme dans celui de toute autre enquête (longitudinale), il y a de la non-réponse. Certains sujets échantillonnés choisissent de ne pas participer du tout, tandis que certains participent à certaines vagues, mais pas à d'autres. Dans le cas de la SDR, pour remédier à cette situation, les poids de sondage transversaux sont corrigés pour tenir compte de la non-réponse.

Supposons que la correction pour la non-réponse à la vague j est une multiplication par l'inverse de la probabilité estimée de réponse à la vague j, $\hat{\pi}_{rij}$. Par exemple, le poids corrigé de la non-réponse pour une personne qui *a répondu* à la vague 3 (et qui avait été sélectionnée initialement à la vague 2), c'est-à-dire pour $i \in r_{3(2)}$, serait $w_{ri3} = \pi_{i2}^{-1}\pi_{i3|s_{2(2)}}^{-1}\hat{\pi}_{ri3}^{-1}$.

Nous devons redéfinir l'équation d'estimation afin d'inclure uniquement les répondants comme étant $\Psi_r(\boldsymbol{\beta}) = \sum_{i\in r}(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})V_i^{-1}W_{ri}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, où la somme est calculée sur l'ensemble des répondants r, c'est-à-dire sur tous les éléments qui appartenaient pour la première fois à n'importe lequel des ensembles de répondants $r_{1(1)}, r_{2(2)}, \ldots, r_{J(J)}$, et la matrice W_{ri} est $W_{ri} = \text{diag}[I_i(U_1)w_{ri1}, I_i(U_2)w_{ri2}, \ldots, I_i(U_J)w_{riJ}]$. En outre, désignons par $r_{j(j')}$ l'ensemble de répondants de la cohorte j' à la vague j. Manifestement, $w_{rij} = 0$ si $i \notin r_j = \bigcup_{j'=1}^{j} r_{j(j')}$.

Si, de surcroît, on peut supposer que le mécanisme de réponse (R) est de type MAR, nous avons alors, par exemple pour $\sum_{i\in r_{2(2)}} B_i W_{ri}\boldsymbol{e}_i$:

$$E_R\left\{\sum_{i\in r_{2(2)}} B_i W_{ri}\boldsymbol{e}_i\right\} = E_R\left\{\sum_{i\in s_{2(2)}} B_i D_i \boldsymbol{e}_i\right\} = \sum_{i\in s_{2(2)}} B_i D_i^{*}\boldsymbol{e}_i = \sum_{i\in s_{2(2)}} B_i D_i^{**}\boldsymbol{e}_i \overset{\text{def}}{=} \sum_{i\in s_{2(2)}} B_i W_i \boldsymbol{e}_i, \quad (3.6)$$

où $D_i = \text{diag}[0, I_i(U_2)w_{ri2}I_i(r_{2(2)}), I_i(U_3)w_{ri3}I_i(r_{3(2)})]$, $D_i^{*} = \text{diag}[0, (I_i(U_2)\pi_{ri2}) / (\pi_{i2}\hat{\pi}_{ri2}), (I_i(U_3)\pi_{ri3}) / (\pi_{i2}\pi_{i3|s_{2(2)}}\hat{\pi}_{ri3})]$, et $D_i^{**} = \text{diag}[0, I_i(U_2)w_{i2}, I_i(U_3)w_{i3}]$. La troisième égalité dans (3.6) requiert que le modèle de non-réponse utilisé pour $\hat{\pi}_{rij}$ satisfasse $E_R[I_i(r_{j(j')})] \overset{\text{def}}{=} \pi_{rij} = \hat{\pi}_{rij}$. Cela signifie que, dans le modèle pour $\hat{\pi}_{rij}$, nous devons inclure le

plus possible d'information possible considérée comme ayant une influence sur la propension à répondre, pour que cette hypothèse (c'est-à-dire l'hypothèse MAR) tienne. Par exemple, si l'on pense que la non-réponse est indépendante d'une vague à l'autre, on doit inclure dans le modèle pour $\hat{\pi}_{rij}$ autant de variables que possible provenant de la vague correspondante. Si, par ailleurs, il est raisonnable de supposer que la propension à répondre à une vague donnée dépend des réponses précédentes (et éventuellement de l'historique des réponses), ces réponses doivent être incluses dans le modèle de réponse, et ainsi de suite.

L'absence de biais par rapport au plan ainsi que l'absence de biais par rapport au modèle et au plan découle directement de (3.6) ainsi que de la section précédente. Par conséquent, dans la suite de l'exposé, nous ignorons la question de la non-réponse pour simplifier la notation.

3.3 Variance et estimation de la variance

Nous développons maintenant une linéarisation (développement en série de Taylor) pour la variance de l'estimateur proposé. La technique de base a été élaborée par Binder (1983). Pour simplifier les calculs et la notation, nous divisons tous les termes par N; nous redéfinissons

$$\Psi_s(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \text{ et } \Psi_U(\boldsymbol{\beta}) = N^{-1}\sum_{i\in U}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i - \boldsymbol{\mu}_i),$$

où $N = \sum_{j=1}^{J} N_j$. Soit $\hat{\boldsymbol{\beta}}$ notre estimateur, qui satisfait $\Psi_s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, et soit $\boldsymbol{\beta}_N$ l'« estimateur par recensement », qui satisfait $\Psi_U(\boldsymbol{\beta}_N) = \mathbf{0}$. Supposons que $\boldsymbol{\beta}_N - \boldsymbol{\beta} = O_P(1/\sqrt{N_m})$ et $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = O_P(1/\sqrt{n_m})$, avec $N_m = \min\{N_1, N_2, \ldots, N_J\}$ et $n_m = \min\{n_1, n_2, \ldots, n_J\}$. Nous pouvons écrire l'erreur totale de $\hat{\boldsymbol{\beta}}$ sous la forme $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) + (\boldsymbol{\beta}_N - \boldsymbol{\beta})$ = erreur d'échantillonnage + erreur du modèle. Après certains calculs simples, la variance totale, ou plus précisément l'EQM totale, peut être décomposée comme il suit :

$$V_{\text{Tot}} = E_{\xi p}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' = V_{\text{Éch}} + 2 \otimes C_{\text{Éch}-\text{Mod}} + o(1/n_m), \tag{3.7}$$

où $2 \otimes A = A + A'$ pour toute matrice $A, V_{\text{Éch}} = E_\xi V_p$ est la composante de « variance d'échantillonnage », $2 \otimes C_{\text{Éch}-\text{Mod}}$ est la composante croisée « variance d'échantillonnage-

modèle », $V_p = E_p[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)']$, $C_{\text{Éch}-\text{Mod}} = E_p C_\xi$, et $C_\xi = E_\xi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{\beta}_N - \boldsymbol{\beta})'$. En outre, par développements en série de Taylor, nous pouvons obtenir les approximations suivantes : $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = [H(\boldsymbol{\beta}_N)]^{-1}\Psi_s(\boldsymbol{\beta}_N) + o_P(1/\sqrt{n_m})$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = [\hat{H}(\boldsymbol{\beta})]^{-1}\Psi_s(\boldsymbol{\beta}) + o_P(1/\sqrt{n_m})$ et $\boldsymbol{\beta}_N - \boldsymbol{\beta} = [H(\boldsymbol{\beta})]^{-1}\Psi_U(\boldsymbol{\beta}) + o_P(1/\sqrt{N_m})$, où nous définissons $H(\boldsymbol{\beta}) = N^{-1}\sum_{i\in U}(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})V_i^{-1}\mathrm{I}_i(\mathrm{U})(\partial\boldsymbol{\mu}_i / \partial\boldsymbol{\beta})$ et $\hat{H}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})V_i^{-1}W_i(\partial\boldsymbol{\mu}_i / \partial\boldsymbol{\beta})$.

Nous obtenons alors, pour V_p et C_ξ dans (3.7),

$$V_p = [H(\boldsymbol{\beta}_N)]^{-1}\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1} + o_P(1/n_m), \tag{3.8}$$

$$\begin{aligned} C_\xi &= [\hat{H}(\boldsymbol{\beta})]^{-1}E_\xi[\Psi_s(\boldsymbol{\beta})\Psi_U'(\boldsymbol{\beta})][H(\boldsymbol{\beta})]^{-1} + o_P(1/n_m) \\ &= N^{-1}[\hat{H}(\boldsymbol{\beta})]^{-1}\hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1} + o_P(1/n_m), \end{aligned} \tag{3.9}$$

où $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] = E_p[\Psi_s(\boldsymbol{\beta}_N)\Psi_s'(\boldsymbol{\beta}_N)]$ et $\hat{H}_{\Sigma V}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}[(\partial\boldsymbol{\mu}_i' / \partial\boldsymbol{\beta})V_i^{-1}W_i\Sigma_i \times V_i^{-1}(\partial\boldsymbol{\mu}_i / \partial\boldsymbol{\beta})]$; la dérivation de l'expression (3.9) est donnée en annexe.

En conclusion, jusqu'à présent, nous avons trouvé que :

$$\begin{aligned} V_{\text{Tot}} &= E_\xi V_p + 2 \otimes E_p C_\xi + o(1/n_m) \\ &= E_\xi\left\{[H(\boldsymbol{\beta}_N)]^{-1}\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1}\right\} \\ &\quad + 2 \otimes N^{-1}E_p\left\{[\hat{H}(\boldsymbol{\beta})]^{-1}\hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1}\right\} + o(1/n_m). \end{aligned} \tag{3.10}$$

Dans (3.10), tous les termes peuvent être estimés en « insérant » l'estimation $\hat{\boldsymbol{\beta}}$, sauf pour le terme $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$; celui-ci est le sujet de la section suivante.

Si la fraction d'échantillonnage est faible, c'est-à-dire que $n \ll N$, le premier terme de l'expression (3.10) est une bonne approximation de la variance totale; autrement, l'expression pour V_{Tot} est simplement $E_\xi V_p$ (et les termes d'ordre inférieur). Si, au contraire, la fraction d'échantillonnage est grande, les deux termes de (3.10) sont requis.

3.3.1 Variance sous le plan de la fonction d'estimation

Afin d'obtenir une expression pour $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$, nous supposons que $J = 3$, comme auparavant. La méthodologie est celle de l'échantillonnage à deux phases (plus précisément, l'échantillonnage a plusieurs phases), comme il est discuté au chapitre 9 de Särndal et coll.

(1992). Après certains calculs (voir l'annexe), et en définissant $B_i = (\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N}$ $V_i^{-1}, \boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N), \boldsymbol{e}_{i(1\cdots3)} = \boldsymbol{e}_i, \boldsymbol{e}_{i(2\cdots3)} = (0, e_{i2}, e_{i3})'$, et $\boldsymbol{e}_{i(3\cdots3)} = (0, 0, e_{i3})'$, nous obtenons :

$$\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] = \sum_{j=1}^{3} D_{(j)} = \sum_{j=1}^{3}\sum_{k=j}^{3} D_{(j)k}, \tag{3.11}$$

où $D_{(j)} \overset{\text{def}}{=} N^{-2}\mathrm{Var}_p\left(\sum_{i\in s_{j(j)}} B_i W_i \boldsymbol{e}_i\right) = \sum_{k=j}^{3} D_{(j)k}$, pour $j = 1, 2, 3,$

$$N^2 D_{(j)j} \overset{\text{def}}{=} \mathrm{Var}\left[\sum_{i\in s_{j(j)}} w_{ij} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots3)}\right], \text{ pour } j = 1, 2, 3,$$

$$N^2 D_{(j-1)j} \overset{\text{def}}{=} E\left\{\mathrm{Var}\left[\sum_{i\in s_{j(j-1)}} w_{ij} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots3)} \mid s_{j-1(j-1)}\right]\right\}, \text{ pour } j = 2, 3,$$

$$N^2 D_{(1)3} \overset{\text{def}}{=} E\left\{E\left[\mathrm{Var}\left(\sum_{i\in s_{3(1)}} w_{i3} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots3)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\},$$

et, en annexe, nous montrons que :

$$N^2 D_{(j)k} = \mathrm{Var}\left[\sum_{i\in s_{k(j)}} w_{ik} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(k\cdots3)}\right] - \mathrm{Var}\left[\sum_{i\in s_{k-1(j)}} w_{i,k-1} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(k\cdots3)}\right],$$

pour $j = 1, 2, 3$ et $3 \geq k > j$. De manière générale, nous avons prouvé ce qui suit.

Propriété 3.1 La variance (sous le plan) de $\Psi_s(\boldsymbol{\beta}_N)$ peut être décomposée en :

$$\begin{aligned}
&\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] \\
&= \frac{1}{N^2}\sum_{j'=1}^{J}\sum_{j=j'}^{J}\left\{\mathrm{Var}_p\left[\sum_{i\in s_{j(j')}} w_{ij} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_{j-1(j')}} w_{i,j-1} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right]\right\} && (3.12)\\
&= \frac{1}{N^2}\sum_{j=1}^{J}\left\{\mathrm{Var}_p\left[\sum_{i\in s_j} w_{ij} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_{j-1}} w_{i,j-1} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right]\right\}, && (3.13)
\end{aligned}$$

où nous posons que $w_{i,j-1} = 0$ quand $j = j', w_{i0} = 0,$ et pour obtenir (3.13), nous avons changé les variable et utilisé la propriété d'indépendance entre les cohortes.

Dans (3.11), (3.12) et (3.13), nous avons supposé que les cohortes sont indépendantes sous le plan. Cependant, dans certains cas, cette hypothèse ne tient pas; un exemple est celui d'une base

de sondage multiple dont nous avons discuté dans la première partie de la section 3.2. Une autre situation dans laquelle il pourrait ne pas être approprié de supposer que les cohortes sont indépendantes est celle où les ajustements de pondération recoupent les cohortes, ce qui est le cas de la SDR; nous discutons de ce problème à la section 5. Les calculs pour le cas des trois cohortes, fournis en annexe, montrent que l'équation (3.13) est vérifiée pour les termes de variance, même sans indépendance. Nous précisons aussi dans l'annexe les conditions sous lesquelles il s'agit d'une bonne approximation pour les termes de covariance.

3.3.2 Estimation

L'estimation de V_{Tot} dans (3.10) peut être effectuée comme il suit. $H(\boldsymbol{\beta}_N), \hat{H}(\boldsymbol{\beta})$ et $H(\boldsymbol{\beta})$ peuvent être estimés par $\hat{H}(\hat{\boldsymbol{\beta}})$. $\hat{H}_{\Sigma V}(\boldsymbol{\beta})$ peut être estimé par $\hat{H}_{\Sigma V}(\hat{\boldsymbol{\beta}})$, où $\Sigma_i = \text{Cov}[Y_i \mid X_i]$ peut être estimé par $\hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i'$.

Nous utilisons (3.13) dans la propriété 3.1 pour estimer $\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$. À condition qu'il existe une méthode pour estimer la variance des estimateurs (transversaux) de Horvitz-Thompson (H-T), l'expression (3.13) peut être utilisée. Si nous définissons $Z_{ij} = B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}$, nous constatons que chaque terme intervenant dans le calcul de (3.13) tel que $\text{Var}_p\left[\sum_{i\in s_j} w_{ij} Z_{ij}\right]$ est simplement la variance d'un estimateur H-T de la vague j. De toute évidence, la méthode d'estimation de la variance doit prendre en considération à la fois le plan d'échantillonnage et toute correction pour tenir compte de la non-réponse et du calage, mais cela ne présente aucune difficulté de plus que celle posée par tout problème transversal, car tous les éléments sont appliqués transversalement. Dans le cas de la SDR, les variances des estimateurs transversaux sont estimées par rééchantillonnage, mais toute méthode d'estimation de la variance sous le plan peut être utilisée.

Nous utilisons les poids de rééchantillonnage transversaux fournis par le programme de la SDR, mais nous ne réestimons pas le paramètre d'intérêt pour chaque réplique. Premièrement, notons que nous ne devons effectuer le rééchantillonnage que pour l'estimation de la « partie

substantielle » $(\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)])$ de la variance sous le plan $(E_\xi V_p)$. Deuxièmement, bien que $\hat{\boldsymbol{\beta}}$ ne figure pas dans l'expression de l'estimateur H-T dont la variance doit être calculée (et recalculée à chaque réplique), les travaux de Roberts, Binder, Kovačević, Pantel et Phillips (2003), qui appliquent la méthode du « bootstrap de la fonction d'estimation » (Hu et Kalbfleisch 2000) à des données d'enquête, montrent que dans des conditions telles que les nôtres, il n'est pas nécessaire de recalculer l'estimateur à chaque réplique, mais que l'estimateur sur l'échantillon complet suffit. Cette simplification accélère le calcul des estimations répétées.

En guise d'illustration, disons que nous en sommes à la vague j, c'est-à-dire que nous estimons le j^{e} terme dans (3.13). La r^{e} réplique du premier terme est $\sum_{i\in s_j} w_{ij}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, où $w_{ij}^{(r)}$ est le r^{e} poids de rééchantillonnage pour le sujet i à la vague j, et la r^{e} réplique du deuxième terme est $\sum_{i\in s_{j-1}} w_{i,j-1}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, où $w_{i,j-1}^{(r)}$ est le r^{e} poids de rééchantillonnage pour le sujet i à la vague $j-1$.

4 Application à la SDR

Nous utilisons pour l'application le jeu de données restreint de la SDR, aux termes d'un accord de licence conclu avec la NSF. La SDR collecte des données sur, entre autres, la situation d'emploi, l'employeur principal, l'emploi principal, les emplois antérieurs, les études récentes, les caractéristiques démographiques et l'incapacité, qui varient d'une vague à l'autre. Nous n'utilisons que l'information recueillie durant toutes les vagues d'intérêt, à savoir 1995, 1997, 1999, 2001, 2003, 2006 et 2008.

Pour illustrer notre méthodologie, nous avons construit un modèle du salaire des individus au fil du temps. La réponse est le logarithme du salaire (dans l'emploi principal), avec une fonction de lien identité et plusieurs covariables; la modélisation du logarithme du salaire (par opposition au salaire) est une pratique courante. Le modèle contient des covariables indépendantes du temps (comme le sexe) ainsi que des covariables dépendant du temps (tel que le secteur d'emploi).

Nous avons quatre grandes catégories de covariables. Les *variables relatives au diplôme* sont le domaine du diplôme, le nombre d'années écoulées depuis l'obtention du diplôme et l'âge au moment de l'obtention du diplôme. Les *variables relatives à l'emploi* sont le domaine ou la catégorie d'emploi, le secteur, l'indicateur d'études postdoctorales, l'indicateur de membre auxiliaire du corps professoral, le nombre d'heures travaillées par semaine dans l'emploi principal, le nombre de semaines par année dans l'emploi principal, le lien entre l'emploi et le diplôme de doctorat, le travail à temps partiel pour différentes raisons, le nombre de mois depuis le début de l'emploi principal, le mois de début de l'emploi principal, le fait que l'employeur/le type d'emploi a changé ou non depuis la vague précédente et le fait que le changement d'employeur/de type d'emploi depuis la vague précédente était dû ou non à une mise à pied ou à une cessation d'emploi. Les *caractéristiques démographiques de la personne* sont le sexe, la citoyenneté, la race/ethnicité, la présence d'enfants dans la famille, l'état matrimonial, la situation de travail du (de la) conjoint(e). Enfin, les *variables d'« environnement »* sont le nombre d'années écoulées depuis 1995, l'État (d'emploi) et l'indice des prix à la consommation (de la région d'emploi). La liste complète des variables, des interactions et des catégories figure dans Carrillo et Karr (2011). Pour les variables catégoriques, la catégorie de référence est celle pour laquelle la fréquence est la plus élevée.

Le jeu de données pour notre modèle comprend 59 346 sujets et 190 693 observations, réparties comme il suit : $n_{95} = 30\,234$, $n_{97} = 30\,652$, $n_{99} = 26\,732$, $n_{01} = 26\,778$, $n_{03} = 24\,956$, $n_{06} = 25\,910$ et $n_{08} = 25\,431$. Ces données correspondent à des salaires non manquants compris entre 5 000 $ et 999 995 $, pour les personnes dont l'âge est cohérent d'une vague à l'autre et pour lesquelles ne manque pas la valeur de la variable indiquant si l'employeur (établissement d'enseignement postsecondaire) était un établissement public ou privé. Les poids de sondage (transversaux) moyens pour chacune de ces vagues sont : $\overline{w}_{95} = 15{,}37$, $\overline{w}_{97} = 16{,}28$, $\overline{w}_{99} = 19{,}96$, $\overline{w}_{01} = 20{,}74$, $\overline{w}_{03} = 22{,}71$, $\overline{w}_{06} = 22{,}93$ et $\overline{w}_{08} = 24{,}88$.

Les poids de sondage que nous utilisons pour chaque vague sont les poids corrigés finaux. Ces poids correspondent aux poids de sondage originaux corrigés pour la non-réponse et la poststratification. Cependant, la théorie que nous avons élaborée à la section 3 repose sur l'hypothèse que les poids sont les inverses des probabilités de sélection; autrement dit, les poids de sondage originaux. Il s'agit d'une disparité dont nous prévoyons étudier les effets dans l'avenir. Par ailleurs, les calculs de la dernière partie de l'annexe (qui ne reposent sur aucune hypothèse concernant les poids) donnent à penser que l'effet de cette disparité est faible.

Les covariables et les interactions que nous avons prises en considération ont été choisies parce qu'elles étaient suggérées par les analyses exploratoires ou par les spécialistes du domaine de la NSF. Carrillo et Karr (2011) présentent les coefficients β estimés du modèle $y_{ij} = \log(\text{SALAIRE}_{ij}) = X'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}$, où X_{ij} comprend l'ordonnée à l'origine ainsi que les autres covariables. Ce $\boldsymbol{\beta}$ correspond à celui du modèle ξ, dans la formule (3.1), dont les propriétés sont discutées à la section 3. La matrice de covariance de travail est estimée être $\hat{V}_i = \hat{\phi}\mathbf{R}(\hat{\alpha})$, avec $\hat{\phi} = \hat{\sigma}^2 = \left(\sum_{i\in s}\sum_{j=95}^{08} w_{ij}\hat{e}_{ij}^2\right)\Big/\left(\sum_{i\in s}\sum_{j=95}^{08} w_{ij} - p\right) = 0{,}196$, où $\hat{e}_{ij} = y_{ij} - X'_{ij}\hat{\boldsymbol{\beta}}$ et $p = 208$ est le nombre de covariables dans X_{ij}, w_{ij} est le poids transversal pour le sujet i à la vague j à condition que $i \in s_j$ et zéro autrement. L'estimation $\hat{\alpha}$ contient les $21 = (7\times 6)/2$ autocorrélations estimées $\hat{\alpha}_{jj'} = \hat{\alpha}_{j'j} = \left(\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}}\hat{e}_{ij}\hat{e}_{ij'}\right)\Big/\left(\hat{\phi}\left[\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}} - p\right]\right)$, pour $j \neq j' = 1995$, 1997, 1999, 2001, 2003, 2006, 2008, et $\hat{\alpha}_{jj} = 1$ pour tout j. Les valeurs estimées forment la matrice d'autocorrélation :

$$\mathbf{R}(\hat{\alpha}) = \begin{pmatrix} 1 & \hat{\alpha}_{95,97} & \hat{\alpha}_{95,99} & \hat{\alpha}_{95,01} & \hat{\alpha}_{95,03} & \hat{\alpha}_{95,06} & \hat{\alpha}_{95,08} \\ & 1 & \hat{\alpha}_{97,99} & \hat{\alpha}_{97,01} & \hat{\alpha}_{97,03} & \hat{\alpha}_{97,06} & \hat{\alpha}_{97,08} \\ & & 1 & \hat{\alpha}_{99,01} & \hat{\alpha}_{99,03} & \hat{\alpha}_{99,06} & \hat{\alpha}_{99,08} \\ & & & 1 & \hat{\alpha}_{01,03} & \hat{\alpha}_{01,06} & \hat{\alpha}_{01,08} \\ & & & & 1 & \hat{\alpha}_{03,06} & \hat{\alpha}_{03,08} \\ & \text{sym} & & & & 1 & \hat{\alpha}_{06,08} \\ & & & & & & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0{,}38 & 0{,}36 & 0{,}32 & 0{,}30 & 0{,}28 & 0{,}27 \\ & 1 & 0{,}42 & 0{,}36 & 0{,}33 & 0{,}32 & 0{,}31 \\ & & 1 & 0{,}46 & 0{,}38 & 0{,}36 & 0{,}34 \\ & & & 1 & 0{,}47 & 0{,}40 & 0{,}38 \\ & & & & 1 & 0{,}49 & 0{,}44 \\ & \text{sym} & & & & 1 & 0{,}55 \\ & & & & & & 1 \end{pmatrix}.$$

Nous présentons maintenant certaines conclusions concernant les salaires de la population active possédant un doctorat fondées sur les coefficients estimés, qui figurent dans Carrillo et Karr (2011). Avant tout, une estimation raisonnable du salaire moyen prend en considération l'ordonnée à l'origine, le nombre d'heures travaillées par semaine (dont la moyenne est 47) et le nombre d'années écoulées depuis l'obtention du diplôme (en moyenne 15), de sorte qu'une estimation de la moyenne globale est $\exp(9,4 + 47 \times 0,038 - 47^2 \times 0,0003 + 15 \times 0,03 - 15^2 \times 0,0006) = 52\ 067\ \$$, pour un sujet pour lequel toutes les autres covariables continues sont égales à zéro et dans la catégorie de référence pour toutes les variables catégoriques.

Toutes choses étant constantes par ailleurs, les salaires des femmes sont environ égaux à 93,4 % de ceux des hommes, tandis que la race ne semble pas avoir d'effet sur les salaires. L'effet du terme d'interaction du sexe et du nombre d'années écoulées depuis 1995 n'est pas significatif; par conséquent, cette différence de salaire ne varie pas au cours du temps. Soulignons qu'en n'utilisant que les données pour une seule année, nous ne serions pas capables d'évaluer l'effet du temps. Fait encore plus important, en utilisant uniquement les données d'une seule vague, disons 2008, nous ne serions pas capables de déterminer si l'effet d'être de sexe féminin varie au cours du temps.

Les titulaires d'un doctorat ayant un emploi de gestionnaire sont ceux dont les salaires sont les plus élevés, suivis par ceux appartenant aux professions du secteur de la santé; par ailleurs, ceux dont les salaires sont les plus faibles sont ceux employés dans les « autres » professions, suivis de ceux travaillant dans le domaine des sciences politiques.

Parmi les secteurs d'emploi, les salaires les plus élevés sont observés dans le secteur à but lucratif (20 % plus élevés que pour la catégorie de référence correspondant à professeur permanent dans les établissements publics offrant des programmes de quatre années), suivi, dans l'ordre, par l'administration fédérale, le secteur du travail autonome et le secteur sans but lucratif, qui offrent tous des salaires plus élevés que celui de la catégorie de référence. Les salaires les

plus faibles sont ceux offerts par les collèges avec programmes de deux ans et les établissements avec programmes de deux et de quatre ans pour lesquels la permanence ne s'applique pas.

L'effet négatif le plus important sur les salaires est également observé dans le secteur de l'enseignement. Les personnes ayant un poste de membre auxiliaire du corps professoral ont des salaires qui correspondent environ à 59 % des salaires de titulaires d'un doctorat comparable. Fait qui n'est pas étonnant, les salaires postdoctoraux ne correspondent qu'à environ 74 % des salaires de personnes comparables occupant d'autres types de poste.

Le secteur d'emploi est aussi un facteur qui contribue à la dépendance, difficile à interpréter, du salaire à l'égard du mois où a débuté l'occupation du poste courant : les salaires sont plus faibles pour les mois de début correspondant à août et à septembre. Des analyses supplémentaires montrent que l'effet du mois ne s'observe que dans le secteur de l'enseignement, où, comme nous l'avons vu, les salaires sont plus faibles que dans l'industrie ou l'administration publique, et où il est fréquent que les emplois débutent en août ou en septembre. Par conséquent, le secteur d'emploi donne partiellement, mais pas entièrement la réponse. Une ventilation plus fine du secteur de l'enseignement, en utilisant les classifications de Carnegie, réduit, mais ne supprime pas le caractère significatif de l'effet du mois. La SDR ne semble pas fournir suffisamment de données pour éliminer entièrement les effets du mois, de sorte que nous avons gardé la définition des secteurs d'emploi de la SDR.

Les personnes titulaires d'un diplôme en informatique et en sciences de l'information sont celles dont les salaires sont les plus élevés (environ 20 % plus élevés que pour les sciences biologiques), suivies par celles titulaires d'un diplôme en génie électrique et informatique et en économie (environ 16 % plus élevés). Les titulaires d'un doctorat en sciences agricoles et alimentaires, en sciences de l'environnement et de la vie, en sciences de la terre, de l'atmosphère et des océans, et en « autres » sciences sociales reçoivent les salaires les plus faibles. Les « autres » sciences sociales sont les sciences sociales à l'exclusion des sciences économiques et politiques.

Les personnes mariées sont celles dont les salaires sont les plus élevés, suivies par les personnes dans une relation de type mariage, les personnes veuves, les personnes séparées, les personnes divorcées et les personnes jamais mariées. Ces dernières ont des salaires de l'ordre de 89 % seulement de celui des personnes mariées; on pourrait soutenir qu'il existe une certaine association entre le fait de n'avoir jamais été marié et l'âge. La présence d'enfants de plus de deux ans est associée à des salaires plus élevés, mais la présence d'enfants plus jeunes ne l'est pas.

Les titulaires d'un doctorat occupant un emploi qui n'est relié que dans une certaine mesure au domaine de leur diplôme gagnent environ 93 % de ce que les personnes possédant un emploi étroitement associé à leur domaine d'études (la catégorie de référence) gagnent. Si l'emploi n'est pas relié au diplôme de doctorat à cause d'un changement de carrière ou d'intérêt professionnel, le salaire est de l'ordre de 82 % de ce gagnent les personnes occupant un emploi étroitement associé à leur domaine d'études. Par ailleurs, les personnes dont l'emploi n'est pas relié à leur domaine d'études pour d'autres raisons ne gagnent qu'environ 76 % du salaire de celles appartenant à la catégorie de référence.

On observe une augmentation de salaire d'environ 3 % pour chaque année supplémentaire depuis l'obtention du diplôme de doctorat, mais il existe un effet de diminution pour les nombres d'années élevés. Nous interprétons cela comme l'effet de l'expérience. Une petite pénalité est associée à l'obtention du doctorat plus tard dans la vie; pour chaque année d'âge supplémentaire au moment de l'obtention du diplôme, le salaire est réduit de 1 %.

Nous avons également constaté que l'indice des prix à la consommation (IPC) régional a un effet significatif. Le salaire est d'autant plus élevé que l'IPC est élevé. Nous n'avons pas pu utiliser l'IPC associé au marché du travail auquel appartient l'emploi parce que les données de la SDR n'indiquent pas l'emplacement géographique de manière plus précise que l'État. Nous avons inclus l'État dans le modèle comme mesure de substitution du coût de la vie; l'effet de l'État est fortement significatif et les coefficients de certains États comptent parmi les plus élevés

dans l'ensemble. Les salaires les plus élevés sont offerts en Californie, à Washington D.C. et ses faubourgs, ainsi que dans la ville de New York et ses faubourgs. Par ailleurs, les salaires les plus faibles sont observés à Puerto Rico, dans le Vermont, au Montana, dans le Maine, dans l'Idaho, dans le Dakota du Sud, dans le Dakota du Nord, et dans les territoires/à l'étranger.

Le fait d'avoir un emploi à temps partiel en raison de la retraite ou d'une semi-retraite a un effet significatif et figure dans plusieurs termes d'interaction significatifs. Étant donné ces résultats, nous ne pensons pas que les données à notre disposition à l'heure actuelle brossent le tableau complet de la retraite, par exemple, en ce qui concerne les personnes qui sont (semi) retraitées et continuent d'avoir un emploi à temps plein.

Enfin, nous avons analysé les résidus; les figures 4.1 et 4.2 montrent un graphique en boîtes à moustaches des résidus standardisés par année et un graphique spaghetti des résidus standardisés, respectivement.

La figure 4.1 montre que l'ajustement du modèle est raisonnablement bon pour toutes les années de référence, car la plupart des résidus standardisés sont compris entre -2 et 2. En outre, les distributions des résidus ne semblent pas différer fortement d'une année à l'autre.

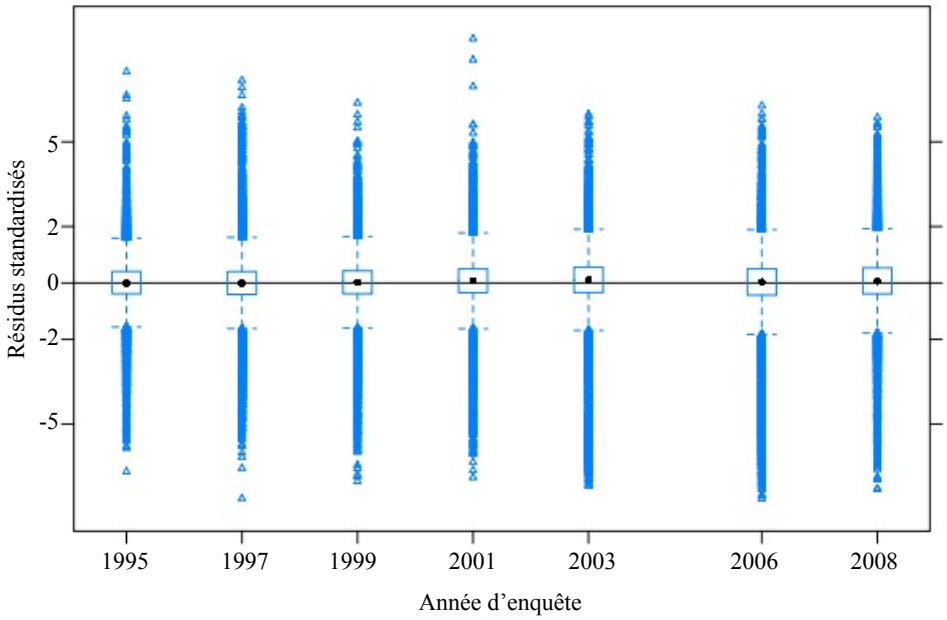


Figure 4.1 Boîtes à moustaches des résidus standardisés par année

La figure 4.2 nous permet également de conclure que l'ajustement du modèle est assez bon pour la plupart des personnes, car la plupart des lignes fluctuent entre -2 et 2. Néanmoins, pour quelques personnes, le modèle semble prédire de façon excessivement élevé fortement en 2003, ainsi qu'en 2006. Nous avons inclus plusieurs termes dans le modèle pour corriger ce problème, mais manifestement aucun ne semble le faire complètement.

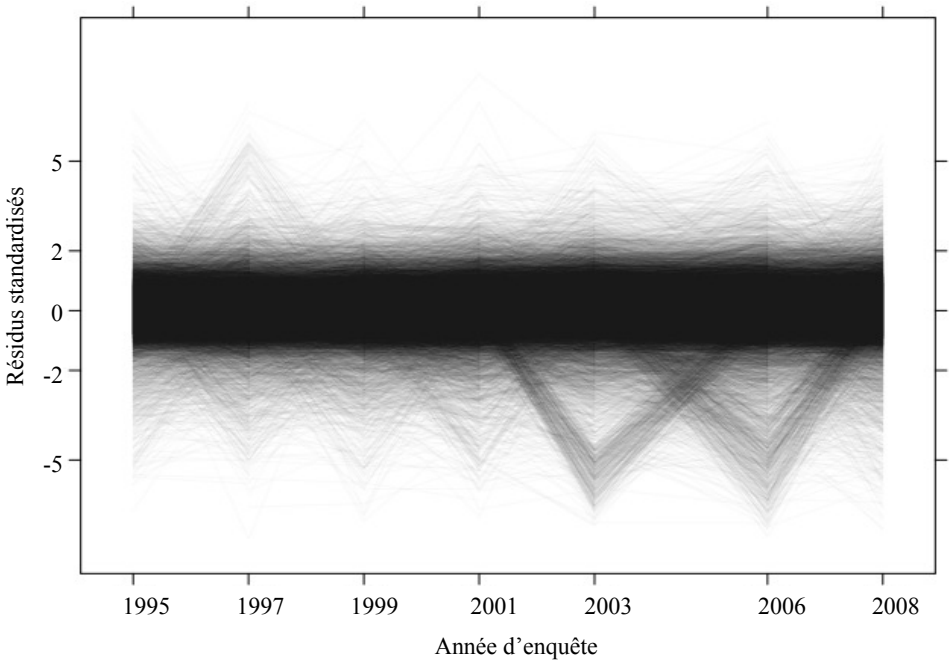


Figure 4.2 Graphique spaghetti des résidus standardisés

En dernier lieu, nous avons tenté de produire des arbres de classification exploratoires pour ces anomalies résiduelles. Nous avons constaté que, pour l'ensemble de données disponible, le seul élément relié à ces anomalies était le mode d'enquête. En 2003, les anomalies sont disproportionnellement importantes pour les réponses en ligne, et en 2006, elles sont disproportionnellement importantes pour les réponses par ITAO. Nous concluons qu'il existe un effet de mode de collecte pour ces deux années bien que, ces années-là, les répondants présentaient une caractéristique différente qui n'est pas incluse dans les variables existantes.

Enfin, la représentation graphique des valeurs prédites en fonction des valeurs observées (que l'on peut consulter dans Carrillo et Karr [2011]) montre la même chose. Pour la plupart des

observations, le modèle donne de bons résultats, sauf pour ces quelques cas en 2003 et en 2006 pour lesquels il produit une surestimation importante.

5 Conclusion et travaux de recherche à venir

Nous avons proposé une nouvelle approche pour combiner différentes cohortes d'une enquête longitudinale. La principale exigence de notre méthode est qu'il existe un poids de sondage transversal pour chaque vague, ou que l'on puisse en construire un d'après les données à notre disposition. Ce poids doit permettre de faire une inférence statistique pour la population d'intérêt durant la vague correspondante. Dans ces conditions, notre méthode devrait donner de meilleurs résultats que les procédures d'estimation habituelles (dans lesquelles l'autocorrélation n'est pas intégrée) dans de nombreuses situations pratiques, en particulier lorsque l'autocorrélation entre les réponses fournies par un même sujet est forte.

En général, les praticiens des sondages évitent autant que possible d'utiliser des poids de sondage multiples. Cependant, dans le cas des panels rotatifs, il s'agit d'une approche intéressante pour au moins deux raisons. D'une part, elle permet d'utiliser toutes les données d'une manière claire et cohérente dans une seule procédure d'analyse. D'autre part, nous avons montré comment des poids de sondage transversaux auxquels ont a facilement accès peuvent être utilisés directement pour l'analyse longitudinale, sans qu'il soit nécessaire d'élaborer, de sauvegarder et de diffuser un ou plusieurs poids longitudinaux supplémentaires.

Notre méthode est directement applicable à toute forme d'enquête longitudinale, à condition qu'il existe des poids de sondage transversaux à chaque vague (ou qu'ils puissent être créés), et que ces poids représentent la population d'intérêt durant la vague en question.

Pour la théorie que nous avons élaborée au sujet de la variance de l'estimateur proposé, nous avons utilisé les poids de sondage (transversaux) w_{ij}, qui sont les inverses des probabilités d'inclusion. Cependant, pour l'application à la SDR, nous avons utilisé les poids de sondage (transversaux) finaux dans notre modèle pour les salaires, qui ne sont pas les poids de sondage

originaux, mais les poids corrigés (de la manière habituelle). Cette disparité nécessite une étude plus approfondie.

De même, dans nos calculs de la variance, nous avons supposé que les cohortes étaient indépendantes. Cependant, la SDR ne satisfait pas entièrement cette hypothèse pour deux raisons. Premièrement, à n'importe quelle vague, la sélection de l'échantillon provenant des anciennes cohortes n'est pas effectuée indépendamment d'une cohorte à l'autre. Afin de réduire le nombre de strates, depuis 1991, la NSF les a regroupées en fonction de l'année d'obtention du diplôme pour les anciennes cohortes. En outre, les corrections pour la poststratification apportées aux poids de sondage ne sont pas conditionnées sur les cohortes non plus, et par conséquent, les poids sont partagés entre les cohortes. Ce schéma de sélection de l'échantillon et la procédure de correction de la pondération violent l'hypothèse d'indépendance entre les cohortes. Certains calculs supplémentaires (inclus en annexe) ont montré que l'indépendance entre les cohortes n'est pas une exigence vraiment cruciale pour que notre méthode d'estimation de la variance produise de bonnes approximations, comme il est expliqué à la section 3.3.1. Durant de futurs travaux de recherche, nous prévoyons évaluer plus en détail l'effet de ce problème.

Remerciements

La présente étude a été financée par la subvention SRS-1019244 accordée par la NSF au National Institute of Statistical Sciences (NISS). Les opinions, constations et conclusions ou recommandations exposées dans la présente publication sont celles des auteurs et ne reflètent pas nécessairement celles de la National Science Foundation. Les auteurs remercient Paul Biemer de RTI International, Stephen Cohen et Nirmala Kannankutty du National Center for Science and Engineering Statistics à la NSF, et Criselda Toto, anciennement du NISS, de leurs nombreuses discussions éclairantes durant l'étude. Nous remercions aussi le rédacteur associé et deux examinateurs de leurs suggestions constructives.

Annexe – Preuves

- Pour développer une expression pour C_ξ, nous commençons par simplifier $\Psi_s(\boldsymbol{\beta})\Psi'_U(\boldsymbol{\beta})$. Soit $\boldsymbol{F}_{i(k)} = B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(k\cdots 3)}$ pour $k = 1, 2, 3,$ alors nous avons :

$$\begin{aligned} N^2\Psi_s(\boldsymbol{\beta})\Psi'_U(\boldsymbol{\beta}) &= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\in U} \boldsymbol{e}'_i \mathrm{I}_i(\mathrm{U}) B'_i = \left[\sum_{i\in s} B_i W_i \boldsymbol{e}_i\right]\left[\sum_{i\in s} \boldsymbol{F}'_{i(1)} + \sum_{i\notin s} \boldsymbol{F}'_{i(1)}\right] \\ &= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\in s} \boldsymbol{F}'_{i(1)} + \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\notin s} \boldsymbol{F}'_{i(1)} \\ &= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_i B'_i + \sum_{i\in s}\sum_{\substack{k\in s\\ k\neq i}} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_k \mathrm{I}_k(\mathrm{U}) B'_k + \mathrm{A}, \end{aligned}$$

où $\mathrm{A} = \left(\sum_{i\in s} B_i W_i \boldsymbol{e}_i\right)\left(\sum_{i\notin s} \boldsymbol{F}'_{i(1)}\right)$, et soit $\mathrm{B} = \sum_{i\in s}\sum_{\substack{k\in s\\ k\neq i}} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_k \mathrm{I}_k(\mathrm{U}) B'_k$. Les deux sommes dans A sont indépendantes du modèle, $\boldsymbol{e}_i$ et $\boldsymbol{e}'_k$ (dans B) sont deux termes dépendants du modèle, et les expressions A et B ont toutes deux une espérance nulle sous le modèle; par conséquent, $E_\xi[\Psi_s(\boldsymbol{\beta})\Psi'_U(\boldsymbol{\beta})] = N^{-2}\sum_{i\in s} B_i W_i E_\xi[\boldsymbol{e}_i \boldsymbol{e}'_i] B'_i = N^{-2}\sum_{i\in s} B_i W_i \Sigma_i B'_i = N^{-1}\hat{H}_{\Sigma V}(\boldsymbol{\beta})$; l'équation (3.9) s'ensuit.

- Nous développons maintenant l'expression pour $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$, la variance sous le plan de la fonction d'estimation; nous redéfinissons $B_i = (\partial\boldsymbol{\mu}'_i / \partial\boldsymbol{\beta})\,|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N}\, V_i^{-1}$ et $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)$; alors

$$\begin{aligned} \mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] &= \mathrm{Var}_p\left(\frac{1}{N}\sum_{i\in s} B_i W_i \boldsymbol{e}_i\right) \\ &= \frac{1}{N^2}\mathrm{Var}_p\left(\sum_{i\in s_{1(1)}} B_i W_i \boldsymbol{e}_i + \sum_{i\in s_{2(2)}} B_i W_i \boldsymbol{e}_i + \sum_{i\in s_{3(3)}} B_i W_i \boldsymbol{e}_i\right) \\ &= \frac{1}{N^2}\mathrm{Var}_p\left(\sum_{i\in s_{1(1)}} B_i W_i \boldsymbol{e}_i\right) + \frac{1}{N^2}\mathrm{Var}_p\left(\sum_{i\in s_{2(2)}} B_i W_i \boldsymbol{e}_i\right) \\ &\quad + \frac{1}{N^2}\mathrm{Var}_p\left(\sum_{i\in s_{3(3)}} B_i W_i \boldsymbol{e}_i\right) = D_{(1)} + D_{(2)} + D_{(3)}, \end{aligned} \tag{A.1}$$

où, pour la ligne (A.1), nous supposons que les (trois) cohortes sont indépendantes sous le plan. Maintenant, $N^2 D_{(1)} = \mathrm{Var}_p\left[\sum_{i\in U_{1(1)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{I}_{i(1)}\}\,\boldsymbol{e}_i\right] = \mathrm{Var}_p\left[\sum_{i\in U_{1(1)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{e}_i\}\,\boldsymbol{I}_{i(1)}\right]$, où $\mathrm{Diag}\{\boldsymbol{e}\}$ est, pour un vecteur colonne $\boldsymbol{e}$, une matrice diagonale dont les entrées sur la diagonale sont les éléments de $\boldsymbol{e}$, et

$\boldsymbol{I}_{i(1)} = \left(I_i(s_{1(1)}),\ I_i(s_{2(1)})I_i(s_{1(1)}),\ I_i(s_{3(1)})I_i(s_{2(1)})\, I_i(s_{1(1)}) \right)'$. Similairement, nous pouvons obtenir $N^2 D_{(2)} = \mathrm{Var}_p \left[\sum_{i \in U_{2(2)}} B_i W_i \ \mathrm{Diag}\{\boldsymbol{e}_i\}\, \boldsymbol{I}_{i(2)} \right]$ et $N^2 D_{(3)} = \mathrm{Var}_p \left[\sum_{i \in U_{3(3)}} B_i W_i \ \mathrm{Diag}\{\boldsymbol{e}_i\}\, \boldsymbol{I}_{i(3)} \right]$, où $\boldsymbol{I}_{i(2)} = \left(0, I_i(s_{2(2)}), I_i(s_{3(2)})I_i(s_{2(2)}) \right)'$, et $\boldsymbol{I}_{i(3)} = \left(0, 0, I_i(s_{3(3)}) \right)'$. Maintenant, concentrons-nous sur $D_{(1)}$; en posant que $C_i = B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}$, nous avons :

$$
\begin{aligned}
N^2 D_{(1)} &= \mathrm{Var}_p \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \right) = \mathrm{Var} \left\{ E \left[\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{1(1)} \right] \right\} \\
&\quad + E \left\{ \mathrm{Var} \left[\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var} \left\{ E \left[E \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&\quad + E \left\{ \mathrm{Var} \left[E \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right. \\
&\qquad \left. + E \left[\mathrm{Var} \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= N^2 D_{(1)1} + N^2 D_{(1)2} + N^2 D_{(1)3}. \qquad \text{(A.2)}
\end{aligned}
$$

Occupons-nous de chacun des termes de (A.2) tour à tour; en commençant par $N^2 D_{1(1)}$, nous avons :

$$
E \left(\sum_{i \in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) = \sum_{i \in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}_{i(1)}^{(1)},
$$

où $\boldsymbol{I}_{i(1)}^{(1)} = (I_i(s_{1(1)}), I_i(s_{2(1)})I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} I_i(s_{2(1)})I_i(s_{1(1)}))'$, alors

$$
\begin{aligned}
&E \left[E \left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \\
&\qquad = \sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)}^{(2)} = \sum_{i \in U_{1(1)}} B_i \mathrm{I}_i^{(1)}(\mathrm{U}) \mathrm{Diag}\{\boldsymbol{I}_{i(1)}^{(2)}\} \boldsymbol{e}_i \\
&\qquad = \sum_{i \in U_{1(1)}} F_i \left[\frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{1(1)})}{\pi_{i1}} \right]' \\
&\qquad = \sum_{i \in U_{1(1)}} F_i \mathbf{1}_3 \frac{I_i(s_{1(1)})}{\pi_{i1}} = \sum_{i \in U_{1(1)}} w_{i1(1)} \boldsymbol{F}_{i(1)} I_i(s_{1(1)}),
\end{aligned}
$$

où $\boldsymbol{I}_{i(1)}^{(2)} = (I_i(s_{1(1)}),\ \pi_{i2|s_{1(1)}} I_i(s_{1(1)}),\ \pi_{i3|s_{2(1)}} \pi_{i2|s_{1(1)}} I_i(s_{1(1)}))'$, $\mathrm{I}_i^{(1)}(\mathrm{U}) = \mathrm{diag}[I_i(U_1)/\pi_{i1},$ $I_i(U_2)\ /\ (\pi_{i1}\pi_{i2|s_{1(1)}}), I_i(U_3)\ /\ (\pi_{i1}\pi_{i2|s_{1(1)}}\pi_{i3|s_{2(1)}})]$, $F_i = B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}$, et $\mathbf{1}_3 = (1,1,1)'$; cela implique que $N^2 D_{(1)1} = \mathrm{Var}\left[\sum_{i\in U_{1(1)}} w_{i1} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i I_i(s_{1(1)})\right] = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i1}\boldsymbol{F}_{i(1)}\right]$.

Pour $N^2 D_{(1)2}$, nous avons :

$$
\begin{aligned}
&E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \\
&\qquad = \sum_{i\in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}^{(1)} = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i^{(1)}(\mathrm{U})\mathrm{Diag}\{\boldsymbol{I}_{i(1)}^{(3)}\}\boldsymbol{e}_i \\
&\qquad = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\left[\frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{2(1)})I_i(s_{1(1)})}{\pi_{i1}\pi_{i2|s_{1(1)}}}, \frac{I_i(s_{2(1)})I_i(s_{1(1)})}{\pi_{i1}\pi_{i2|s_{1(1)}}}\right]' \\
&\qquad = \sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}[\pi_{i2|s_{1(1)}}, I_i(s_{2(1)}), I_i(s_{2(1)})]',
\end{aligned}
$$

où $\boldsymbol{I}_{i(1)}^{(3)} = (I_i(s_{1(1)}), I_i(s_{2(1)})I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} I_i(s_{2(1)})I_i(s_{1(1)}))'$; alors,

$$
\begin{aligned}
&\mathrm{Var}\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right] \\
&\qquad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}^{(4)} \mid s_{1(1)}\right] \\
&\qquad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}[0, I_i(s_{2(1)}), I_i(s_{2(1)})]' \mid s_{1(1)}\right] \\
&\qquad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\} I_i(s_{2(1)})\mathbf{1}_{02} \mid s_{1(1)}\right] \\
&\qquad = \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots 3)} \mid s_{1(1)}\right], \qquad \text{(A.3)}
\end{aligned}
$$

où $\boldsymbol{I}_{i(1)}^{(4)} = [\pi_{i2|s_{1(1)}}, I_i(s_{2(1)}), I_i(s_{2(1)})]', \mathbf{1}_{02} = (0,1,1)'$, et la ligne (A.3) est obtenue parce que, conditionnellement à $s_{1(1)}, \pi_{i2|s_{1(1)}}$ est constant et par conséquent la variance de cette composante est nulle. Cela signifie que :

$$
\begin{aligned}
N^2 D_{(1)2} &= E\left\{ \mathrm{Var}\left[E\left(\sum_{i \in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= E\left\{ \mathrm{Var}\left[\sum_{i \in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var}\left[\sum_{i \in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \right] - \mathrm{Var}\left\{ E\left[\sum_{i \in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var}\left[\sum_{i \in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \right] - \mathrm{Var}\left\{ E\left[\sum_{i \in s_{2(1)}} w_{i2|s_{1(1)}} w_{i1} \boldsymbol{F}_{i(2)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var}\left[\sum_{i \in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(2)} \right] - \mathrm{Var}\left\{ \sum_{i \in s_{1(1)}} w_{i1} \boldsymbol{F}_i \right\}.
\end{aligned}
$$

Nous pouvons, similairement, montrer que :

$$
\begin{aligned}
N^2 D_{(1)3} &= E\left\{ E\left[\mathrm{Var}\left(\sum_{i \in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= E\left\{ E\left[\mathrm{Var}\left(\sum_{i \in s_{3(1)}} w_{i3} I_i(s_{2(1)}) I_i(s_{1(1)}) \boldsymbol{F}_{i(3)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= E\left\{ \mathrm{Var}\left[\sum_{i \in s_{3(1)}} w_{i3} I_i(s_{2(1)}) \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right] \right. \\
&\qquad \left. -\mathrm{Var}\left[E\left(\sum_{i \in s_{3(1)}} w_{i3} I_i(s_{2(1)}) \boldsymbol{F}_{i(3)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\} \\
&= E\left\{ \mathrm{Var}\left[\sum_{i \in s_{3(1)}} w_{i3} \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right] - \mathrm{Var}\left[\sum_{i \in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right] \right\} \\
&= \mathrm{Var}\left[\sum_{i \in s_{3(1)}} w_{i3} \boldsymbol{F}_{i(3)} \right] - \mathrm{Var}\left[E\left(\sum_{i \in s_{3(1)}} w_{i3} \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right) \right] \\
&\quad -\mathrm{Var}\left[\sum_{i \in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(3)} \right] + \mathrm{Var}\left[E\left(\sum_{i \in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(3)} \mid s_{1(1)} \right) \right] \\
&= \mathrm{Var}\left[\sum_{i \in s_{3(1)}} w_{i3} \boldsymbol{F}_{i(3)} \right] - \mathrm{Var}\left[\sum_{i \in s_{1(1)}} w_{i1} \boldsymbol{F}_{i(3)} \right] \\
&\quad -\mathrm{Var}\left[\sum_{i \in s_{2(1)}} w_{i2} \boldsymbol{F}_{i(3)} \right] + \mathrm{Var}\left[\sum_{i \in s_{1(1)}} w_{i1} \boldsymbol{F}_{i(3)} \right].
\end{aligned}
$$

Au moyen de calculs similaires, nous obtenons les expressions correspondantes pour $N^2 D_{(2)}, N^2 D_{(2)2}, N^2 D_{(2)3}$ et $N^2 D_{(3)} = N^2 D_{(3)3}$.

- Enfin, nous esquissons le développement d'une expression pour $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$ sans supposer que les cohortes sont indépendantes. Premièrement, notons que $\Psi_s(\boldsymbol{\beta}_N)$ peut s'écrire :

$$\sum_{i\in s_1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} w_{i1} & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{bmatrix} + \sum_{i\in s_2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$-\sum_{i\in s_1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$+\sum_{i\in s_3} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & 0 & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix} - \sum_{i\in s_2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & 0 & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix}$$

$$= \sum_{i\in s_1} w_{i1} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_i - \sum_{i\in s_1} w_{i1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix} + \sum_{i\in s_2} w_{i2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix} - \sum_{i\in s_2} w_{i2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix}$$

$$+\sum_{i\in s_3} w_{i3} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix};$$

en posant que $\boldsymbol{z}_i = B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_i, \boldsymbol{z}_{i(2\cdots3)} = B_i \mathrm{I}_i(\mathrm{U})[0, e_{i2}, e_{i3}]'$ et $\boldsymbol{z}_{i(3\cdots3)} = B_i \mathrm{I}_i(\mathrm{U})[0, 0, e_{i3}]'$, $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$ peut être développée comme il suit :

$$\mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i\right] + \mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_{i(2\cdots3)}\right]$$

$$+ \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\boldsymbol{z}_{i(2\cdots3)}\right] + \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\boldsymbol{z}_{i(3\cdots3)}\right]$$

$$+ \mathrm{Var}_p\left[\sum_{i\in s_3} w_{i3}\boldsymbol{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i, \sum_{i\in s_1} w_{i1}\boldsymbol{z}_{i(2\cdots3)}\right]$$

$$+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i, \sum_{i\in s_2} w_{i2}\boldsymbol{z}_{i(2\cdots3)}\right]$$

$$- 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i, \sum_{i\in s_2} w_{i2}\boldsymbol{z}_{i(3\cdots3)}\right] + 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\boldsymbol{z}_i, \sum_{i\in s_3} w_{i3}\boldsymbol{z}_{i(3\cdots3)}\right]$$

$$
\begin{aligned}
&- 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}\right] + 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] \\
&- 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] \\
= {} & \mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_i\right] + \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}\right] \\
&+ \mathrm{Var}_p\left[\sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots1)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots1)}, \sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] \qquad \text{(A.4)} \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots3)}\right].
\end{aligned}
$$

Dans cette dernière expression, la première chose que nous constatons est que *tous* les éléments diagonaux dans *tous* les termes de covariance sont exactement égaux à zéro; cela signifie que l'expression (3.13) est exacte pour les termes de variance, que les cohortes soient indépendantes ou non les unes des autres.

Pour analyser l'importance des termes de covariance, nous nous concentrons sur le terme de la ligne (A.4); la conclusion pour les autres termes est la même; notons que ce terme peut s'écrire :

$$
2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} e_{i1} \\ e_{i2} \\ 0 \end{pmatrix}, \sum_{i\in s_3} w_{i3}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} 0 \\ 0 \\ e_{i3} \end{pmatrix} - \sum_{i\in s_2} w_{i2}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} 0 \\ 0 \\ e_{i3} \end{pmatrix}\right];
$$

La propriété 3.1 énonce que, si les *cohortes* sont indépendantes sous le plan, tous les termes de covariance sont exactement égaux à zéro. En outre, de cette dernière

expression, nous concluons, trivialement, que si les *vagues* sont indépendantes sous le plan, tous les termes de covariance sont égaux à zéro également. Cette formule pour le terme de la ligne (A.4) implique aussi que, si les poids individuels ne varient pas fortement d'une vague à la suivante, et que le chevauchement entre vague consécutive est important, les termes de covariance ne sont pas trop grands. Enfin, si le chevauchement est faible, il est raisonnable de supposer l'indépendance des vagues sous le plan, et les termes de covariance peuvent alors être approximés sans risque par zéro.

Bibliographie

Ardilly, P., et Lavallée, P. (2007). Pondération dans les échantillons rotatifs : le cas de l'Enquête SILC en France. *Techniques d'enquête*, 33, 2, 149-156.

Berger, Y.G. (2004a). Variance estimation for change: An evaluation based upon the 2000 finnish labour force survey. Proceedings. European Conference on Quality and Methodology in Official Statistics.

Berger, Y.G. (2004b). Variance estimation for measures of change in probability sampling. *La Revue Canadienne de Statistique*, 32, 4, 451-467.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Carrillo, I.A., Chen, J. et Wu, C. (2010). The pseudo-GEE approach to the analysis of longitudinal surveys. *La Revue Canadienne de Statistique*, 38, 4, 540-554.

Carrillo, I.A., Chen, J. et Wu, C. (2011). A pseudo-GEE approach to analyzing longitudinal surveys under imputation for missing responses. *Journal of Official Statistics*, 27, 2, 255-277.

Carrillo, I.A., et Karr, A.F. (2011). Combining cohorts in longitudinal surveys. Rapport technique 180, National Institute of Statistical Sciences, Research Triangle Park, NC. Adresse URL http://www.niss.org/sites/default/files/tr180.pdf.

Carrillo, I.A., et Karr, A.F. (2012). Estimating change with multi-cohort longitudinal surveys. En préparation.

Cox, B.G., Grigorian, K., Wang, R. et Harter, R. (2010). 2008 Survey of Doctorate Recipients Weighting Implementation Report, document préparé par la National Opinion Research Center (NORC) for the National Science Foundation (NSF).

Diggle, P., Heagerty, P., Liang, K.-Y. et Zeger, S. (2002). *Analysis of Longitudinal Data*, 2e Édition. Oxford University Press, New York.

Hedeker, D., et Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. New Jersey : John Wiley & Sons, Inc., Hoboken.

Hirano, K., Imbens, G.W., Ridder, G. et Rubin, D.B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69, 6, 1645-1659.

Hu, F., et Kalbfleisch, J.D. (2000). The estimating function bootstrap (Pkg: P449-495). *La Revue Canadienne de Statistique*, 28, 3, 449-481.

Larsen, M.D., Qing, S., Zhou, B. et Foulkes, M.A. (2011). Calibration estimation and longitudinal survey weights: Application to the NSF Survey of Doctorate Recipients. *Proceedings of the Survey Research Method Section*, American Statistical Association, 1360-1374.

Liang, K.-Y., et Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lohr, S. (2007). Recent developments in multiple frame surveys. *Joint Statistical Meeting of the American Statistical Association*, 3257-3264.

McLaren, C.H., et Steel, D.G. (2000). L'effet de divers plans de renouvellement sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance. *Techniques d'enquête*, 26, 2, 185-195.

National Science Foundation, National Center for Science and Engineering Statistics (2012). Survey of doctorate recipients. http://www.nsf.gov/statistics/srvydoctoratework/, consulté le 9 février 2012.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics*, 21, 1, 43-52.

Qualité, L., et Tillé, Y. (2008). Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'Enquête suisse sur la valeur ajoutée. *Techniques d'enquête*, 34, 2, 193-201.

Rao, J.N.K., et Wu, C. (2010). Pseudo-empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 492, 1494-1503.

Roberts, G., Binder, D., Kovačević, M., Pantel, M. et Phillips, O. (2003). Using an estimating function bootstrap approach for obtaining variance estimates when modelling complex health survey data. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, Halifax.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Smith, P., Lynn, P. et Elliot, D. (2009). Sample design for longitudinal surveys. Dans *Methodology of Longitudinal Surveys*, (Éd., P. Lynn). Wiley, Chichester, Chapitre 2, 21-33.

Song, P.X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer Series in Statistics. New York : Springer.

Steel, D., et McLaren, C. (2007). Design and analysis of repeated surveys. Discours principal, International Conference on Quality Management of Official Statistics, Corée.

Vieira, M.D.T. (2009). Analysis of Longitudinal Survey Data: Allowing for the Complex Survey Design in Covariance Structure Models. VDM Verlag.

Vieira, M.D.T., et Skinner, C.J. (2008). Estimating models for panel survey data under complex sampling. *Journal of Official Statistics*, 24, 3, 343-364.

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 207-241
Statistique Canada, N° 12-001-X au Catalogue

Le sondage indirect appliqué aux populations asymétriques

Pierre Lavallée et Sébastien Labelle-Blanchet[1]

Résumé

On recourt au sondage indirect quand la base de sondage ne coïncide pas avec la population cible, mais est reliée à celle-ci. L'estimation se fait alors par la Méthode généralisée du partage des poids (MGPP) qui est une procédure sans biais (voir Lavallée 2002, 2007). Dans le cas des enquêtes économiques, le sondage indirect s'applique comme suit : la base de sondage comprend les établissements, tandis que la population cible comprend les entreprises. Les entreprises sont sélectionnées au travers de leurs établissements. Cela permet de procéder à une stratification en fonction des caractéristiques des établissements au lieu de celles des entreprises. Comme les distributions des variables d'intérêt des établissements sont généralement très asymétriques (une faible proportion des établissements couvre la majeure partie de l'économie), la MGPP produit des estimations sans biais, mais dont la variance peut être importante. L'objectif du présent article est de proposer certaines corrections des poids en vue de réduire la variance des estimations dans le contexte des populations asymétriques, tout en maintenant l'absence de biais de la méthode. Après un court aperçu du sondage indirect et de la MGPP, nous décrivons les corrections qu'il faut apporter à cette dernière. Nous comparons les estimations obtenues en appliquant ces corrections à celles que donne la MGPP originale en nous servant d'un petit exemple numérique et de données réelles tirées du Registre des entreprises de Statistique Canada.

Mots clés : Méthode généralisée du partage des poids; liens pondérés; optimalité faible.

1 Introduction

Les enquêtes auprès des entreprises diffèrent des enquêtes sociales à plusieurs égards. L'une des différences tient au fait que la distribution des données associées à une base de sondage d'entreprises est fortement asymétrique, tandis que celle des données associées aux enquêtes sociales est beaucoup plus homogène. L'échantillonnage des entreprises s'effectue en transformant les structures opérationnelles en unités normalisées appelées unités statistiques. Celles-ci sont représentées sous forme d'une hiérarchie, c'est-à-dire une série de niveaux qui permettent l'intégration subséquente des divers éléments de données disponibles à différents niveaux de l'organisation. Le nombre de niveaux que compte la hiérarchie diffère selon

1. Pierre Lavallée, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada. Courriel : pierre.lavallee@statcan.gc.ca; Sébastien Labelle-Blanchet, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada. Courriel : sebastien.labelle-blanchet@statcan.gc.ca.

l'organisme statistique. Par exemple, le Registre des entreprises du Canada comprend quatre niveaux : l'entreprise, la société, l'établissement et l'emplacement. En général, l'entreprise statistique correspond à l'entité juridique. L'établissement statistique équivaut, dans la plupart des cas, à un centre de profit et fournit des données sur la valeur de la production et sur le coût des intrants et de la main-d'œuvre. Ces données suffisent pour calculer la valeur ajoutée (bénéfices, traitements et salaires). Dans le présent article, nous ne considérons que deux niveaux hiérarchiques, à savoir l'entreprise et l'établissement, comme ils sont définis dans le Registre des entreprises du Canada. Pour plus de renseignements, le lecteur pourra consulter Statistique Canada (2010).

Pour le tirage de l'échantillon, la stratification se fait souvent au niveau de l'établissement. Cela permet de contrôler la représentativité géographique (par exemple en stratifiant par province), industrielle (par exemple en stratifiant par activité industrielle) et de taille (par exemple en stratifiant par catégorie de revenus ou par nombre d'employés). Ces types de représentativité sont impossibles à contrôler si la stratification a lieu au niveau de l'entreprise. Cependant, en plus des statistiques au niveau de l'établissement, on doit souvent produire des statistiques au niveau de l'entreprise. Par conséquent, pour réaliser ces deux objectifs, nous sélectionnons un échantillon au niveau de l'établissement, puis nous l'étendons à l'ensemble des établissements appartenant aux entreprises propriétaires de ceux sélectionnés au départ. Soulignons que sélectionner des entreprises en passant par l'échantillonnage des établissements équivaut à sélectionner des grappes par la voie de leurs composantes. Cette procédure permet de produire des estimations aux niveaux de l'établissement et de l'entreprise, et de réduire aussi dans une certaine mesure les coûts de collecte en sélectionnant des groupes d'établissements.

Une façon d'envisager la production d'estimations au niveau de l'entreprise en se servant d'un échantillon d'établissements consiste à visualiser séparément la base de sondage et la population cible. La première est un ensemble d'établissements, tandis que la seconde est un ensemble d'entreprises correspondant à des groupes d'établissements. Lorsque la base de sondage ne

coïncide pas avec la population cible, mais est quand même reliée à cette dernière, nous nous trouvons dans une situation de *sondage indirect* (voir Lavallée 2002 [version française], 2007 [version anglaise]). Plus formellement, nous souhaitons produire une estimation pour une population cible U^B, en nous servant d'une base de sondage U^A, qui est reliée d'une certaine façon à U^B. Nous tirons un échantillon s^A de U^A pour produire une estimation pour U^B en utilisant les liens qui existent entre les deux populations. Afin de produire des estimations sans biais des quantités d'intérêt (par exemple totaux ou moyennes) pour la population cible U^B en utilisant s^A, nous obtenons des poids d'estimation par la Méthode généralisée du partage des poids (MGPP).

Même si la théorie de la MGPP est largement établie (voir Lavallée 2002, 2007), son application aux enquêtes auprès des entreprises pose certaines difficultés. Alors que les estimations produites sont sans biais, les variances ont tendances à être importantes. Ce manque de précision est dû à l'asymétrie de la population, un petit nombre d'établissements couvrant la majeure partie de l'économie.

L'objectif du présent article est de proposer certaines corrections des poids d'estimation en vue de réduire la variance des estimations dans le contexte des populations asymétriques, tout en préservant l'absence de biais de la méthode. Après un bref aperçu du sondage indirect et de la MGPP à la section 2, nous présentons le problème des populations à la section 3. Nous décrivons les corrections qu'il est proposé d'apporter à la MGPP à la section 4. À la section 5, nous comparons les estimations obtenues en appliquant ces corrections à celles produites par la MGPP originale en nous servant d'un petit exemple numérique et de données réelles provenant du Registre des entreprises de Statistique Canada. Une brève conclusion est présentée à la section 6.

2 Le sondage indirect et la MGPP

Pour commencer, nous donnons un aperçu du sondage indirect et de la MGPP. Le sondage indirect a été élaboré pour tout type de plan de sondage, mais nous nous concentrerons ici sur

l'échantillonnage aléatoire simple sans remise (EASSR) stratifié, puisque ce plan de sondage est celui utilisé le plus souvent dans les enquêtes auprès des entreprises.

Soit la population U^A de M^A établissements stratifiés en H strates, où la strate h contient M_h^A établissements. Dans chaque strate h, nous tirons un échantillon s_h^A de m_h^A établissements par EASSR. Soit $s^A = \bigcup_{h=1}^{H} s_h^A$ et $m^A = \sum_{h=1}^{H} m_h^A$. La population cible U^B contient N^B entreprises, où l'entreprise i contient les M_i^B établissements de U^A. Cette population peut aussi être considérée comme une population de M^B établissements, où chaque établissement k appartient à une entreprise i, avec $M^B = \sum_{i=1}^{N^B} M_i^B$.

Nous souhaitons produire une estimation pour la population cible U^B en utilisant la base de sondage U^A, ainsi que les liens qui existent entre les deux populations. Les liens entre la population U^A et la population U^B sont identifiés par la variable indicatrice $l_{j,i}$, où $l_{j,i} = 1$ s'il existe un lien entre l'établissement $j \in U^A$ et l'entreprise $i \in U^B$, et 0 autrement. Ici, $l_{j,i} = 1$ si l'établissement j de U^A appartient à l'entreprise i de U^B, et 0 autrement. Comme chaque établissement ne peut appartenir qu'à une seule entreprise, les liens entre U^A et U^B sont de type plusieurs à un ou un à un. Par conséquent, nous avons $L_j^A = \sum_{i=1}^{N^B} l_{j,i} = 1$, $L_i^B = \sum_{j=1}^{M^A} l_{j,i} = M_i^B$, pour tous les établissements $j \in U^A$ et pour toutes les entreprises $i \in U^B$.

Étapes du sondage indirect :

1. Pour chaque établissement j sélectionné dans s^A, nous identifions l'entreprise correspondante i de U^B.
2. Pour chaque entreprise i identifiée, nous supposons que nous pouvons dresser la liste U_i^B des M_i^B établissements de cette entreprise.
3. Pour chaque entreprise i identifiée, nous sondons *chacun* des M_i^B établissements de l'entreprise.
4. À la fin, nous obtenons un échantillon s^B de n^B entreprises, et cet échantillon contient $m^B = \sum_{i=1}^{n^B} M_i^B$ établissements.

Pour tous les établissements k liés aux entreprises $i \in s^B$, nous mesurons une variable d'intérêt y_{ik}. Nous voulons estimer le total $Y = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^{N^B} Y_i$ pour la population

cible U^B. Notons que le processus de collecte du sondage indirect aboutit à un nombre d'établissements sondés beaucoup plus grand que le nombre d'établissements compris dans l'échantillon initial s^A. Nous sélectionnons au départ m^A établissements dans l'échantillon s^A, et nous finissons par échantillonner $m^B = \sum_{i=1}^{n^B} M_i^B$ établissements, où $m^B \geq m^A$.

En pratique, il se peut que certaines entreprises ne fournissent des données qu'au niveau de l'entreprise. Cela signifie que nous obtenons les valeurs $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$ pour $i \in s^B$, mais non les valeurs y_{ik} mesurées au niveau de l'établissement. Comme nous le verrons, cela ne pose pas de problème pour les estimations globales, mais pourrait causer des difficultés pour certaines estimations détaillées. Le cas échéant, on procède à une désagrégation (également appelée allocation) des valeurs de l'entreprise au niveau de l'établissement en se basant principalement sur l'expérience des spécialistes du domaine (voir, par exemple, Delorme 2000).

Sous sondage indirect, une non-réponse peut exister dans l'échantillon s^A tiré de U^A, ou à l'intérieur des unités (entreprises ou établissements) désignées pour être sondées dans U^B. Puisque les unités de la population U^B sont en fait sondées par groupe (rappelons que les entreprises sont des groupes d'établissements), deux types de non-réponse sont issus de U^B : la non-réponse de groupe et la non-réponse d'unité. La non-réponse de groupe a lieu lorsque la variable d'intérêt y n'est mesurée pour aucun des établissements d'une entreprise sélectionnée pour participer à l'enquête. La non-réponse d'unité a lieu lorsque un ou plusieurs établissements de l'entreprise, mais pas tous, ne répondent pas. Dans le cas du sondage indirect, il existe aussi une autre forme de non-réponse découlant de la difficulté à déterminer certains liens. Ce type de non-réponse est associé à la situation où il est impossible de déterminer si un établissement k d'une entreprise i de U^B est relié ou non à un établissement j de U^A. On parle alors de problème de détermination de lien. Lavallée (2002, 2007) a proposé des solutions pour corriger ces types de non-réponse par ajustement des pondérations. Afin de limiter la portée du présent article, nous supposons qu'aucune non-réponse ne se produit à aucun niveau.

Conformément à la MGPP, pour estimer le total Y, nous utilisons l'estimateur

$$\hat{Y} = \sum_{i=1}^{n^B} w_i Y_i \tag{2.1}$$

où n^B est le nombre d'entreprises sondées. Les poids produits par la MGPP sont donnés par

$$w_i = \sum_{j=1}^{M^A} \frac{t_j^A}{\pi_j^A} \frac{l_{j,i}}{L_i^B} \tag{2.2}$$

où $t_j^A = 1$ si $j \in s^A$, et 0 autrement, et π_j^A est la probabilité de sélection de l'établissement j. Ici, nous avons $\pi_j^A = m_h^A / M_h^A$ pour $j \in h$. Il convient de souligner qu'en général, les poids (2.2) ne correspondent pas aux probabilités de sélection π_i^B des entreprises i. En utilisant (2.2), nous pouvons récrire l'estimateur (2.1) sous la forme

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \tag{2.3}$$

où

$$Z_j = \sum_{i=1}^{N^B} \frac{Y_i}{L_i^B} l_{j,i}. \tag{2.4}$$

En raison de la correspondance de type plusieurs à un entre U^A et U^B, nous avons

$$w_i = \frac{1}{M_i^B} \sum_{j=1}^{M_i^B} \frac{t_j^A}{\pi_j^A}. \tag{2.5}$$

En outre, la variable Z_j de (2.4) peut s'écrire $Z_j = Y_i / M_i^B = \bar{Y}_i$, pour $j \in i$, ce qui est la moyenne des M_i^B établissements appartenant à l'entreprise i. Nous obtenons donc

$$\hat{Y} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj} \tag{2.6}$$

où $Z_{hj} = Y_i / M_i^B = \bar{Y}_i$, pour $j \in i$.

On peut prouver que l'estimateur (2.1) – et par conséquent les estimateurs (2.3) et (2.6) – est sans biais pour Y (voir Lavallée 2002, 2007). Notons que l'estimateur $\hat{Y}$ est en fait simplement un estimateur de Horvitz-Thompson où la variable d'intérêt est la variable Z_{hj}. Sous EASSR stratifié, sa variance est donnée par

$$\mathrm{Var}(\hat{Y}) = \sum_{h=1}^{H} M_h^A \left(\frac{M_h^A - m_h^A}{m_h^A} \right) S_{Z,h}^2 \tag{2.7}$$

où $S_{Z,h}^2 = \sum_{j=1}^{M_h^A} \left(Z_{hj} - \bar{Z}_h \right)^2 \Big/ (M_h^A - 1)$ et $\bar{Z}_h = \sum_{j=1}^{M_h^A} Z_{hj} \Big/ M_h^A$. La variance $\mathrm{Var}(\hat{Y})$ peut être estimée en se servant de l'estimateur classique pour l'EASSR stratifié, ou au moyen d'autres estimateurs de variance proposés dans la littérature scientifique, tels les estimateurs par le jackknife et le bootstrap. Voir Wolter (2007) ou Särndal, Swensson et Wretman (1992).

La précision des estimations produites en appliquant la MGPP dépend uniquement de la variance, parce que l'estimateur (2.1) – et par conséquent les estimateurs (2.3) et (2.6) – est sans biais. L'examen de l'équation (2.7) montre que la précision dépend, comme dans le cas classique, des tailles d'échantillon et des fractions de sondage utilisées pour sélectionner s^A, mais aussi de la variabilité des variables dérivées Z. Puisque $Z_{hj} = Y_i / M_i^B = \bar{Y}_i$, pour $j \in i$, la valeur de Z_{hj} est la même pour tous les établissements j de l'entreprise i donnée. Autrement dit, le total Y_i de l'entreprise est partagé de manière égale entre ses établissements. Si tous les établissements d'une entreprise appartiennent à la même strate, la variabilité des variables Z dans une strate dépendra seulement de la différence entre les valeurs moyennes pour un nombre limité d'entreprises, ce qui pourrait rendre la variabilité relativement faible. Par ailleurs, si les établissements d'une entreprise appartiennent à différentes strates, la variabilité des variables Z dans une strate dépendra de la différence entre un nombre d'entreprises pouvant être aussi élevé que le nombre d'établissements, ce qui pourrait entraîner une assez forte variabilité. En raison de l'asymétrie de la population d'établissements et de la stratification appliquée à U^A, ce dernier cas est celui qui est le plus probable.

Il est intéressant de constater que la version actuelle du sondage indirect (conjugué à la MGPP) correspond mathématiquement du *sondage en grappes adaptatif* présenté par Thompson (1990, 1991, 1992, 2002) et par Thompson et Seber (1996). Sous sondage en grappes adaptatif, on sélectionnerait d'abord un échantillon d'établissements, puis on élaborerait une stratégie de collecte en vue de sonder tous les établissements des entreprises identifiés au moyen de l'échantillon initial d'établissements sélectionnés. Habituellement, la stratégie de collecte consisterait à étendre l'échantillon d'établissements en rendant visite l'un après l'autre aux établissements d'une même entreprise jusqu'à ce qu'ils soient tous couverts. Sous sondage indirect, la stratégie de collecte n'est pas spécifiée, mais on suppose qu'à la fin du processus de collecte, tous les établissements appartenant aux entreprises sélectionnées ont été sondés. On peut prouver que l'estimateur sous sondage en grappes adaptatif est le même que l'estimateur (2.1) obtenu en appliquant la MGPP (voir Lavallée 2002, 2007). Notons que les deux plans de sondage ne sont mathématiquement équivalents que dans certains cas particuliers. Il en est ainsi dans le présent article lorsque l'on utilise l'estimateur (2.1). Quand on utilise les liens pondérés (voir la section suivante), la MGPP produit un estimateur différent de celui associé au sondage en grappes adaptatif. En outre, quand les liens entre les populations U^A et U^B sont de type plusieurs à plusieurs, le sondage indirect et le sondage en grappes adaptatif ne sont plus équivalents.

2.1 Utilisation de liens pondérés

La variable indicatrice $l_{j,i}$ indique simplement s'il existe un lien entre l'établissement j et l'entreprise i provenant des populations U^A et U^B, respectivement. Il est toutefois possible de remplacer la variable indicatrice $l_{j,i}$ par une variable quantitative $\theta_{j,i}$ représentant l'importance que nous voulons donner au lien $l_{j,i}$. Autrement dit, la généralisation de la variable indicatrice l définie sur $\{0,1\}$ au moyen d'une variable quantitative θ définie sur $[0,+\infty[$, l'ensemble de nombres réels non négatifs, ne pose aucun problème. Ici, une valeur de $\theta_{j,i} = 0$ équivaut à un

lien $l_{j,i} = 0$. La théorie articulée sur la MGPP reste valide. Par exemple, l'estimateur résultant demeure sans biais. Comme nous le verrons plus loin, choisir les valeurs appropriées pour les liens pondérés $\theta_{j,i}$ sera le fondement des méthodes visant à réduire la variance des estimations obtenues en appliquant la MGPP.

Soit $\tilde{\theta}_{j,i} = \theta_{j,i} / \theta_i^B$, où $\theta_i^B = \sum_{j=1}^{M^A} \theta_{j,i}$. Partant de (2.2), nous définissons

$$w_i^\theta = \sum_{j=1}^{M_i^B} \frac{t_j^A}{\pi_j^A} \tilde{\theta}_{j,i}. \tag{2.8}$$

En utilisant (2.8), nous pouvons récrire l'estimateur (2.6) sous la forme

$$\hat{Y}_\theta = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^\theta \tag{2.9}$$

où

$$Z_{hj}^\theta = \sum_{i=1}^{N^B} \tilde{\theta}_{j,i} Y_i \tag{2.10}$$

pour $j \in h$. En raison de la correspondance de type plusieurs à un entre U^A et U^B, la variable Z_{hj}^θ dans (2.10) est une part pondérée du total Y_i pour les M_i^B établissements appartenant à l'entreprise i. La variance de (2.9) s'obtient en remplaçant Z_j par Z_j^θ dans (2.7) :

$$\mathrm{Var}(\hat{Y}_\theta) = \sum_{h=1}^{H} M_h^A \left(\frac{M_h^A - m_h^A}{m_h^A} \right) S_{\theta Zh}^2 \tag{2.11}$$

où $S_{\theta Zh}^2 = \sum_{j=1}^{M_h^A} \left(Z_{hj}^\theta - \bar{Z}_h^\theta \right)^2 \Big/ (M_h^A - 1)$ et $\bar{Z}_h^\theta = \sum_{j=1}^{M_h^A} Z_{hj}^\theta \Big/ M_h^A$.

2.2 Utilisation de liens pondérés optimaux

La MGPP offre une solution simple pour obtenir un poids d'estimation w_i pour chaque entreprise i sondée. Cependant, l'estimateur $\hat{Y}$ donné par (2.1) ou (2.3) résultant de l'application par défaut de la MGPP n'est pas toujours celui possédant la variance la plus faible. Il est possible

de l'améliorer en déterminant les poids optimaux pour les liens $\theta_{j,i}$. Ce problème a été résolu par Deville et Lavallée (2006).

Nous avons fait remarquer plus haut que la variance (2.7) dépend de la variabilité des variables dérivées Z. Sans liens pondérés, c'est-à-dire avec $Z_{hj} = Y_i / M_i^B = \bar{Y}_i$, pour $j \in i$, la valeur de Z_{hj} est la même pour tous les établissements j d'une entreprise i donnée. Comme il est probable que les établissements d'une entreprise appartiennent à différentes strates, la variabilité des variables Z dans une strate dépendra de la différence entre un nombre d'entreprises pouvant être aussi élevé que le nombre d'établissements. De plus, une entreprise i donnée fournira la même valeur de Z à tous ses établissements j, puisque $Z_{hj} = \bar{Y}_i$. Par conséquent, qu'un établissement fasse ou non partie d'une strate contenant de « grandes » ou de « petites » unités (selon une mesure de taille donnée), il recevra la valeur moyenne de son entreprise propriétaire. Cela fera augmenter la variabilité dans les strates et, donc, la variance (2.7). L'utilisation de liens pondérés repose sur l'idée de partager la valeur du total de l'entreprise Y_i inégalement entre les établissements de cette entreprise. Rechercher les liens pondérés optimaux équivaut à essayer de partager la valeur du total de l'entreprise Y_i de façon que la variance (2.11) soit minimale.

Deville et Lavallée (2006) ont obtenu un estimateur dont la variance est inférieure ou égale à celle de l'estimateur original $\hat{Y}$. Comme nous l'avons mentionné plus haut, l'estimateur $\hat{Y}_\theta$ donné par (2.9) produira encore des estimations sans biais. Or, la variance (2.11) de cet estimateur dépend des liens pondérés $\theta_{j,i}$. Le problème consiste alors à trouver au moins un jeu de valeurs $\theta_{j,i}$ tel que la variance de l'estimateur $\hat{Y}_\theta$ soit minimale. Autrement dit, pour les $\theta_{j,i}$ plus grands que 0, nous voulons déterminer les valeurs qui nous permettent d'obtenir l'estimateur le plus précis $\hat{Y}_\theta$. La solution s'obtient en minimisant la variance (2.11) par rapport aux liens pondérés $\theta_{j,i}$, problème relativement classique et facile à résoudre. Cependant, la solution n'est pas simple à écrire, et elle dépend souvent de la variable d'intérêt y.

Si les liens pondérés optimaux $\theta_{j,i}^{\text{opt}}$ dépendent de la variable d'intérêt y, alors les poids w_i^{θ} dépendront aussi de y. Cela signifie qu'un jeu différent de poids devra être calculé pour chaque variable d'intérêt. Pour contourner ce problème, Deville et Lavallée (2006) ont défini une *optimalité faible*, qui correspond à la minimisation de la variance (2.11) pour un choix très précis d'une variable d'intérêt : $Y_i = 1$ pour une entreprise i de U^B et $Y_{i'} = 0$ pour toutes les autres entreprises i' de $U^B (i' \neq i)$. Les liens pondérés faiblement optimaux résultants ne font pas intervenir, à proprement parler, la variable y et ils s'avèrent être relativement faciles à calculer, c'est-à-dire qu'ils peuvent être obtenus sous forme d'une solution explicite, sans que des calculs numériques soient nécessaires. En outre, si certaines conditions énoncées par Deville et Lavallée (2006) sont satisfaites, l'optimalité faible correspond à une *optimalité forte indépendante de* y. Autrement dit, les liens pondérés $\theta_{j,i}^{w-\text{opt}}$ obtenus par optimalité faible correspondent aux liens pondérés optimaux $\theta_{j,i}^{\text{opt}}$ obtenus en minimisant (2.11), et ils ne dépendent pas de la variable d'intérêt y. Malheureusement, ces conditions sont rarement satisfaites en pratique, même pour des plans de sondage simples tels que l'EASSR.

Sous EASSR *sans* stratification, on peut montrer que les liens pondérés faiblement optimaux sont donnés par $\tilde{\theta}_{j,i}^{w-\text{opt}} = \theta_{j,i}^{w-\text{opt}} / \sum_{j=1}^{M^A} \theta_{j,i}^{w-\text{opt}} = 1/M_i^B$ pour l'établissement $j \in U^A$ appartenant à l'entreprise $i \in U^B$, et 0 autrement. Cette solution concorde avec celle conjecturée par Kalton et Brick (1995). Ils ont obtenu ce résultat en se basant sur la situation simplifiée où $M^A = 2$ et où s^A est sélectionné par échantillonnage équiprobabiliste. Dans leurs conclusions, ils suggéraient d'utiliser les valeurs optimales $\theta_{j,i}^{\text{opt}} = 1$ lorsque $\theta_{j,i} > 0$, et $\theta_{j,i}^{\text{opt}} = 0$ lorsque $\theta_{j,i} = 0$. Lavallée (2002) et Lavallée et Caron (2001) ont obtenu des résultats allant dans le même sens en utilisant des simulations. Comme nous l'avons mentionné plus haut, malheureusement, les poids faiblement optimaux $\tilde{\theta}_{j,i}^{w-\text{opt}} = 1/M_i^B$ ne correspondent pas aux poids fortement optimaux indépendants de y.

3 Le problème des populations asymétriques

Comme il est mentionné dans l'introduction, l'application de la MGPP aux enquêtes auprès des entreprises peut produire des estimations dont la variance est importante. Ce manque de précision est dû à l'asymétrie de la population. Nous proposons d'illustrer le problème à l'aide d'un petit exemple donné à la figure 3.1.

Nous voulons étudier le revenu y de la population U^B de la figure 3.1 contenant $N^B = 3$ entreprises, où l'entreprise 1 contient $M_1^B = 4$ établissements, l'entreprise 2 contient $M_2^B = 4$ établissements, et l'entreprise 3 contient $M_3^B = 3$ établissements. Comme l'indique la figure 3.1, le revenu y des $M^B = 11$ établissements peut être considéré comme une population asymétrique.

Pour l'enquête, nous construisons une base de sondage U^A contenant les $M^A = 11$ établissements et nous décidons de répartir ces établissements en trois strates de taille, à savoir la strate $h = 1$ contenant les établissements pour lesquels $y \geq 750$, la strate $h = 2$ contenant ceux pour lesquels $100 \leq y < 750$, et la strate $h = 3$ contenant ceux pour lesquels $y < 100$ (en pratique, une telle stratification est impossible puisque la variable de stratification y est la même que la variable d'intérêt; nous pourrions alors utiliser une variable de taille x fortement corrélée à la variable d'intérêt y). Dans la strate $h = 1$, nous appliquons une fraction de sondage de 1 (c'est-à-dire $f_1 = m_1^A / M_1^A = 1$); pour $h = 2$, la taille d'échantillon est 1 (c'est-à-dire $f_2 = m_2^A / M_2^A = 1/3$); et pour $h = 3$, la taille d'échantillon est 2 (c'est-à-dire $f_3 = m_3^A / M_3^A = 2/6 = 1/3$).

En tout, $1 \times 3 \times 15 = 45$ échantillons s^A peuvent être tirés de U^A afin de produire une estimation du total réel $Y = 3\,800$. Pour chacun de ces 45 échantillons, nous calculons $\hat{Y}$ en utilisant (2.1). Les estimations sont présentées dans la boîte à moustaches de gauche de la figure 3.2.

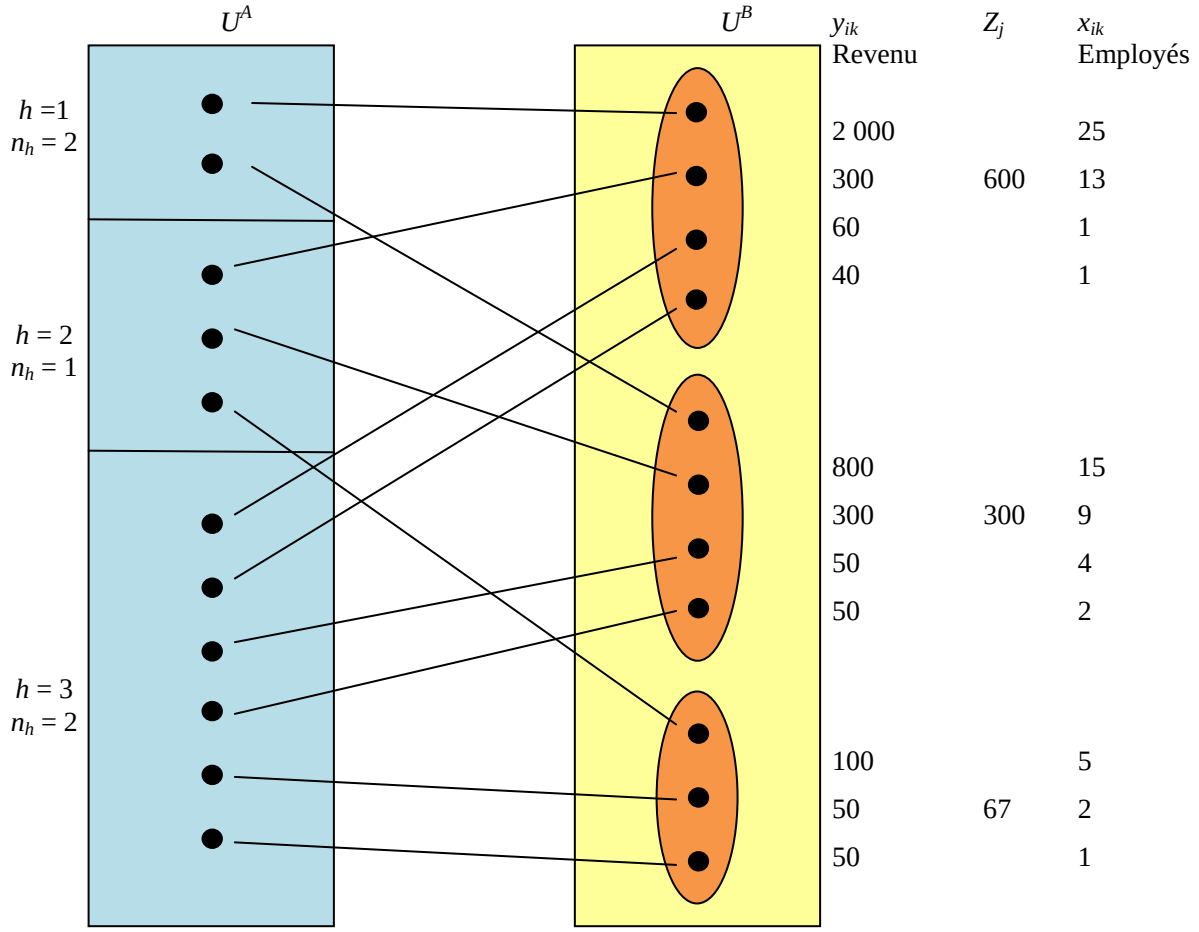


Figure 3.1 Petit exemple

Nous calculons aussi les estimations de Y en supposant que l'on utilise un EASSR stratifié *sans* sondage indirect. C'est-à-dire que, dans chaque strate h, nous sélectionnons un échantillon s_h^A de m_h^A établissements par EASSR et nous ne mesurons la variable d'intérêt y_{ik} que pour les établissements ik de U^B directement liés aux établissements échantillonnés j de U^A. En d'autres termes, nous mesurons la variable d'intérêt y_j pour les établissements échantillonnés j de U^A. Contrairement au sondage indirect, nous ne mesurons pas les variables d'intérêt pour les autres établissements des entreprises contenant les établissements échantillonnés au départ. Cela correspond à la théorie classique de l'échantillonnage. Donc, nous estimons Y en utilisant

$$\hat{Y}_{\text{classique}} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} y_{hj}. \tag{3.1}$$

On peut prouver que l'estimateur (3.1) est sans biais et que sa variance est donnée par

$$\text{Var}(\hat{Y}_{\text{classique}}) = \sum_{h=1}^{H} M_h^A \left(\frac{M_h^A - m_h^A}{m_h^A} \right) S_{y,h}^2 \tag{3.2}$$

où $S_{y,h}^2 = \sum_{j=1}^{M_h^A} \left(y_{hj} - \bar{Y}_h \right)^2 \Big/ (M_h^A - 1)$ et $\bar{Y}_h = \sum_{j=1}^{M_h^A} y_{hj} \Big/ M_h^A$. Les estimations sont présentées dans la boîte à moustaches de droite de la figure 3.2.

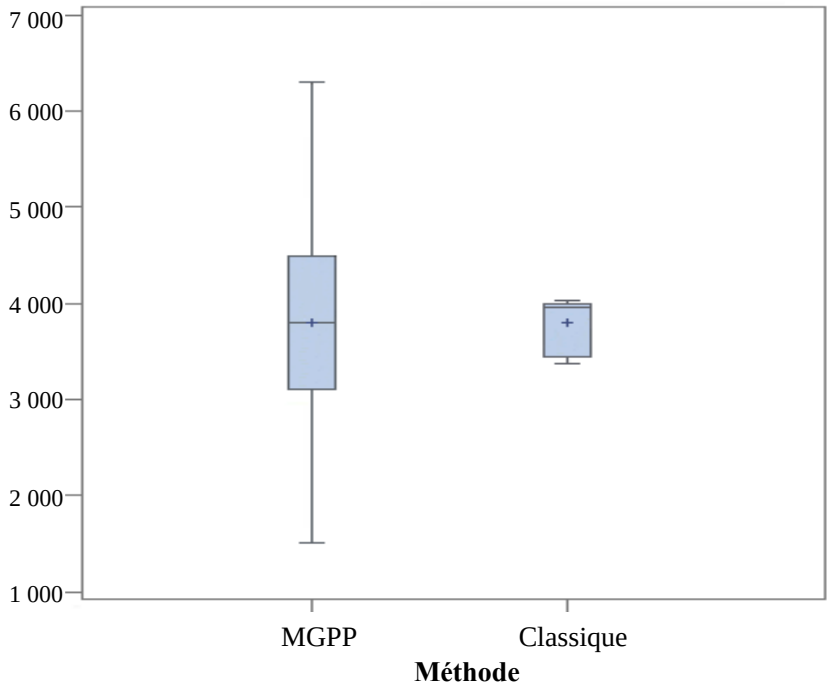


Figure 3.2 Sommaire des 45 estimations possibles

Comme le montre la figure 3.2, les estimations obtenues par sondage indirect (et la MGPP) varient assez bien d'un échantillon à l'autre. Si nous n'utilisons pas le sondage indirect (c'est-à-dire si nous suivons l'approche classique), la variabilité est beaucoup plus faible. Ce résultat peut être constaté directement en examinant les variances de $\hat{Y}$ et $\hat{Y}_{\text{classique}}$. En utilisant les formules (2.7) et (3.2), nous obtenons la variance $V(\hat{Y}_{\text{classique}}) = 80\ 480$, tandis que $V(\hat{Y}) = 1\ 115\ 111$!

À la section suivante, nous présentons les méthodes conçues en vue de réduire la variabilité des estimations produites par sondage indirect.

4 Méthodes proposées

Les méthodes proposées ici pour réduire la variance des estimations sont fondées principalement sur l'emploi de liens pondérés pour le calcul des estimations de Y sous sondage indirect. Nous utiliserons par conséquent l'estimateur (2.9) plutôt que l'estimateur (2.3). Le premier ensemble de méthodes est fondé sur l'emploi de liens pondérés $\theta_{j,i}$ qui sont proportionnels à une mesure de taille des établissements. Le deuxième correspond à l'utilisation des solutions optimales présentées à la section 2.2, sous diverses hypothèses. Enfin, le dernier consiste à se servir des probabilités de sélection exactes plutôt que des poids d'estimation obtenus en appliquant la MGPP, sous deux scénarios d'échantillonnage.

4.1 Méthodes fondées sur l'emploi de liens pondérés

Méthode 1 : $\theta_{j,i}$ proportionnel à π_j^A

Nous proposons d'abord de réduire la variance (2.11) en prenant $\theta_{j,i}$ proportionnel à π_j^A. Formellement, cela peut s'écrire $\theta_{j,i}^{\pi} = \pi_j^A l_{j,i}$. Dans les enquêtes auprès des entreprises, comme la stratification est habituellement effectuée selon la taille (en s'appuyant sur une mesure de taille), prendre $\theta_{j,i}$ proportionnel à π_j^A revient à attribuer des poids élevés aux liens des grands établissements et des poids faibles à ceux des petits établissements, ce qui est une approche naturelle.

Selon cette méthode, nous avons $\tilde{\theta}_{j,i}^{\pi} = \theta_{j,i}^{\pi} \big/ \theta_i^{\pi B} = \pi_j^A l_{j,i} \Big/ \sum_{j=1}^{M^A} \pi_j^A l_{j,i}$. Étant donné la correspondance de type plusieurs à un entre U^A et U^B, nous obtenons $\tilde{\theta}_{j,i}^{\pi} = \theta_{j,i}^{\pi} \big/ \theta_i^{\pi B} = \pi_j^A l_{j,i} \Big/ \sum_{j=1}^{M_i^B} \pi_j^A$. Par conséquent, partant de (2.8), nous avons

$$w_j^{\pi} = \frac{\sum_{j=1}^{M_i^B} t_j^A}{\sum_{j=1}^{M_i^B} \pi_j^A}. \tag{4.1}$$

En utilisant (4.1), nous pouvons récrire l'estimateur (2.9) sous la forme

$$\hat{Y}_{\pi} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{\pi} \tag{4.2}$$

où

$$Z_{hj}^{\pi} = \frac{\pi_j^A Y_i}{\sum_{j=1}^{M_i^B} \pi_j^A} \tag{4.3}$$

pour $j \in h$ et $j \in i$. Il convient de souligner que, si tous les établissements j d'une entreprise donnée appartiennent à la strate disons, h, nous avons $\tilde{\theta}_{j,i} = 1 / M_i^B$, et l'estimateur (4.2) est alors équivalent à l'estimateur (2.1) [et (2.3)].

Pour calculer la variance de $\hat{Y}_{\pi}$, nous utilisons la formule (2.11) avec les valeurs (4.3). Pour l'exemple de la section 3, nous obtenons $V(\hat{Y}_{\pi}) = 439\,111$, valeur qui représente une réduction importante comparativement à $V(\hat{Y}) = 1\,115\,111$, mais qui est encore relativement éloignée de $V(\hat{Y}_{\text{classique}}) = 80\,480$.

Méthode 2 : $\theta_{j,i}$ proportionnel à une mesure de taille x_j

Nous proposons de réduire la variance (2.11) en prenant $\theta_{j,i}$ proportionnel à une mesure de taille x corrélée à la variable d'intérêt y. Nous supposons que la variable x_j est disponible pour tous les établissements $j \in U^A$. Cette variable pourrait être utilisée, par exemple, pour stratifier la base de sondage U^A selon la taille. Comme dans la méthode 1, prendre $\theta_{j,i}^x = x_j l_{j,i}$ peut être considéré comme attribuer des poids élevés aux liens des grands établissements et des poids faibles à ceux des petits, ce qui de nouveau est une approche naturelle. Selon cette méthode, nous avons $\tilde{\theta}_{j,i}^x = \theta_{j,i}^x \big/ \theta_i^{xB} = x_j l_{j,i} \Big/ \sum_{j=1}^{M^A} x_j l_{j,i}$.

Nous avons $\tilde{\theta}_{j,i}^x = \theta_{j,i}^x \big/ \theta_i^{xB} = x_j l_{j,i} \Big/ \sum_{j=1}^{M_i^B} x_j = x_j l_{j,i} \big/ X_i$ en raison de la correspondance de type plusieurs à un entre U^A et U^B. Par conséquent, partant de (2.8), nous avons

$$w_i^x = \frac{1}{X_i} \sum_{j=1}^{M_i^B} \frac{t_j^A x_j}{\pi_j^A}. \tag{4.4}$$

En utilisant (4.4), nous pouvons récrire (2.9) sous la forme

$$\hat{Y}_x = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^x \tag{4.5}$$

où

$$Z_{hj}^x = \frac{Y_i}{X_i} x_j \tag{4.6}$$

pour $j \in h$ et $j \in i$.

Pour calculer la variance de $\hat{Y}_x$, nous utilisons la formule (2.11) avec les valeurs (4.6). Dans l'exemple de la section 3, la variable x correspond au nombre d'employés (voir la figure 3.1). La corrélation entre le revenu y et le nombre d'employés x est relativement forte (ρ = 92,8 %). Nous obtenons $V(\hat{Y}_x)$ = 686 540, ce qui de nouveau représente une réduction importante comparativement à $V(\hat{Y})$ = 1 115 111, mais est encore relativement éloigné de $V(\hat{Y}_{\text{classique}})$ = 80 480.

Méthode 3 : $\theta_{j,i}$ proportionnel à la variable d'intérêt y_j

La troisième méthode proposée consiste à réduire la variance (2.11) en prenant $\theta_{j,i}$ proportionnel à la variable d'intérêt y mesurée pour l'établissement j appartenant à l'entreprise i. Manifestement, prendre $\theta_{j,i}^y = y_j l_{j,i}$ revient à attribuer des poids élevés aux liens des grands établissements et des poids faibles à ceux des petits, ce qui de nouveau est une approche naturelle. Comme y_j est inconnue au début de l'enquête, cette méthode pourrait sembler impossible à appliquer, puisque $\theta_{j,i}^y$ dépend de y_j. Cependant, puisque la correspondance entre U^A et U^B est de type plusieurs à un, chaque quantité entrant dans $\theta_{j,i}^y$ est mesurée selon le processus de sondage indirect.

La méthode proposée est applicable en pratique et nous avons $\tilde{\theta}_{j,i}^y = \theta_{j,i}^y \big/ \theta_i^{yB} = y_j l_{j,i} \Big/ \sum_{j=1}^{M_i^B} y_j = y_j l_{j,i} / Y_i$. Les poids w_i^y sont donnés directement par (4.4) en remplaçant x par y. L'estimateur $\hat{Y}_y$ obtenu au moyen de (2.9) se réduit à

$$\hat{Y}_y = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} y_{hj}, \tag{4.7}$$

ce qui n'est autre que l'estimateur (3.1) découlant de la théorie classique de l'échantillonnage.

Notons qu'en général, cette méthode requiert un jeu de liens pondérés $\theta_{j,i}^{y}$ par variable d'intérêt y. Une solution consisterait à limiter la détermination des liens pondérés à quelques variables d'intérêt clés, chacune associée à un jeu plus grand de covariables corrélées. Cependant, ce genre de contrainte n'est pas nécessaire ici parce que l'estimateur (4.7) correspond simplement à l'estimateur (3.1). En effet, en fin de compte, nous obtenons des poids d'estimation qui correspondent simplement aux poids de sondage.

Pour calculer la variance de (4.7), nous employons simplement la formule (3.2). Pour l'exemple de la section 3, nous obtenons $V(\hat{Y}_y) = V(\hat{Y}_{\text{classique}}) = 80\ 480$, ce qui représente une réduction très importante comparativement à $V(\hat{Y}) = 1\ 115\ 111$.

4.2 Méthodes utilisant des liens pondérés faiblement optimaux

Méthode 4 : Utilisation de liens pondérés faiblement optimaux $\theta_{j,i}^{w-\text{opt},\text{EAS}}$ sous EASSR stratifié

Cette méthode consiste à utiliser les liens pondérés faiblement optimaux $\theta_{j,i}^{w-\text{opt},\text{EAS}}$ de Deville et Lavallée (2006) décrits à la section 2.2. Comme nous l'avons mentionné plus haut, ces liens sont obtenus en minimisant la variance (2.11) pour un choix très précis de la variable d'intérêt, à savoir $Y_i = 1$ pour une entreprise i de U^B et $Y_{i'} = 0$ pour toutes les autres entreprises i' de $U^B (i' \neq i)$. Les liens pondérés faiblement optimaux résultants ne font pas intervenir la variable y à proprement parler. Écrire les valeurs de $\theta_{j,i}^{w-\text{opt},\text{EAS}}$ comprend des expressions qui peuvent être exprimées adéquatement en notation matricielle. En utilisant des sommations, les expressions deviennent beaucoup plus compliquées, principalement parce qu'elles font intervenir un mélange de probabilités de sélection conjointes $\pi_{j,j'}^{A}$ des établissements j et j' qui peuvent appartenir ou non à la même strate.

Définissons la matrice carrée $\mathbf{\Delta}^A = \left[\Delta^A_{j,j'}\right]$ de taille M^A où $\Delta^A_{j,j'} = (\pi^A_{j,j'} - \pi^A_j \pi^A_{j'}) / \pi^A_j \pi^A_{j'}$. Soit $\mathbf{\Gamma}^A = \left[\gamma^A_{j,j'}\right]$ l'inverse de la matrice $\mathbf{\Delta}^A$, c'est-à-dire $(\mathbf{\Delta}^A)^{-1} = \mathbf{\Gamma}^A$. Soit $\mathbf{\Gamma}^A_i = \left[\gamma_{ij,ij'}\right]$ la sous-matrice carrée de $\mathbf{\Gamma}^A$ contenant tous les éléments (établissements) (j, j') appartenant à l'entreprise i. En suivant Deville et Lavallée (2006), nous obtenons $\tilde{\theta}^{w-\text{opt,EAS}}_{j,i} = l_{j,i} \sum_{j'=1}^{M_i^B} \gamma_{ij,ij'} \Big/ \sum_{j=1}^{M_i^B} \sum_{j'=1}^{M_i^B} \gamma_{ij,ij'}$. Malheureusement, ici, la correspondance de type plusieurs à un entre U^A et U^B n'aide pas à obtenir une forme plus simple pour $\tilde{\theta}^{w-\text{opt,EAS}}_{j,i}$.

Soulignons que, si une entreprise i contient un établissement j_0 dans la strate à tirage complet $h = 1$, nous avons $\Delta^A_{j_0,j'} = (\pi^A_{j_0,j'} - \pi^A_{j_0}\pi^A_{j'}) / \pi^A_{j_0}\pi^A_{j'} = 0$ et la matrice $\mathbf{\Delta}^A$ est singulière. Dans ce cas, la solution choisie consiste à prendre $\tilde{\theta}^{w-\text{opt,EAS}}_{j,i} = 1$ pour l'établissement à tirage complet j_0 de l'entreprise i, et $\tilde{\theta}^{w-\text{opt,EAS}}_{j',i} = 0$ pour les autres établissements $j' \neq j_0$ de l'entreprise i. Cela signifie que $Z^{w-\text{opt,EAS}}_{hj_0} = Y_i$ et $Z^{w-\text{opt,EAS}}_{hj'} = 0$ pour $j' \neq j_0$, et donc que la valeur complète Y_i sera attribuée à l'établissement j_0 dont la contribution à la variance est 0.

Nous avons

$$w_i^{w-\text{opt,EAS}} = \sum_{j=1}^{M_i^B} \frac{t_j^A}{\pi_j^A} \tilde{\theta}^{w-\text{opt,EAS}}_{j,i}. \tag{4.8}$$

En utilisant (4.8), nous pouvons récrire l'estimateur (2.9) sous la forme

$$\hat{Y}_{w-\text{opt,EAS}} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z^{w-\text{opt,EAS}}_{hj} \tag{4.9}$$

où

$$Z^{w-\text{opt,EAS}}_{hj} = \sum_{i=1}^{N^B} Y_i \tilde{\theta}^{w-\text{opt,EAS}}_{j,i} \tag{4.10}$$

pour $j \in h$ et $j \in i$.

Pour calculer la variance de $\hat{Y}_{w-\text{opt,EAS}}$, nous utilisons la formule (2.11) avec les valeurs (4.10). Pour l'exemple de la section 3, nous obtenons $V(\hat{Y}_{w-\text{opt,EAS}}) = 23\,111$, ce qui représente une réduction énorme de la variance comparativement à $V(\hat{Y}) = 1\,115\,111$ et $V(\hat{Y}_{\text{classique}}) = 80\,480$.

Méthode 5 : Utilisation de liens pondérés faiblement optimaux $\theta_{j,i}^{w-\mathrm{opt},\mathrm{EP}}$ sous échantillonnage de Poisson

Dans le contexte des enquêtes auprès des entreprises, l'échantillonnage de Poisson consiste à tirer l'échantillon s^A en passant par les M^A établissements de la population U^A et en sélectionnant l'établissement j si $u_j \leq \pi_j^A$, où $u_j \sim U(0,1)$. Les probabilités de sélection sont simplement données par $\pi_j^A = m_h^A / M_h^A$ pour $j \in h$ et la taille d'échantillon de strate réalisée résultante $\tilde{m}_h^A$ est aléatoire. Dans ce contexte, ce plan de sondage peut également être considéré comme un échantillonnage de Bernoulli stratifié (voir Särndal, Swensson et Wretman 1992).

L'échantillonnage de Poisson (ou l'échantillonnage de Bernoulli stratifié) est un plan de sondage très simple. Comme on peut le constater, chaque établissement de s^A est sélectionné indépendamment des autres. Cela signifie que la probabilité de sélection conjointe $\pi_{j,j'}^A$ de deux établissements j et j' de U^A est simplement donnée par $\pi_{j,j'}^A = \pi_j^A \pi_{j'}^A$. En conditionnant sur $\tilde{m}_h^A$, on peut montrer que l'échantillonnage de Bernouilli stratifié correspond à l'EASSR stratifié. L'estimateur qu'il faut utiliser sous échantillonnage de Bernoulli stratifié est l'estimateur par le ratio

$$\tilde{Y}_\theta = \sum_{h=1}^{H} \frac{M_h^A}{\tilde{m}_h^A} \sum_{j=1}^{\tilde{m}_h^A} Z_{hj}^\theta. \tag{4.11}$$

La variance de l'estimateur (4.11) est donnée approximativement par la formule (2.11) (voir Brewer et Hanif 1983). Étant donné la ressemblance relative de ces deux plans, supposer que l'on procède à un échantillonnage de Poisson peut être une approche raisonnable pour calculer les liens pondérés faiblement optimaux $\theta_{j,i}^{w-\mathrm{opt},\mathrm{EP}}$.

Les liens pondérés faiblement optimaux $\theta_{j,i}^{w-\mathrm{opt},\mathrm{EP}}$ s'obtiennent en calculant $\tilde{\theta}_{j,i}^{w-\mathrm{opt},\mathrm{EAS}}$ comme dans la méthode 4, mais en supposant que cette sélection d'échantillon est faite sous échantillonnage de Poisson. Cette hypothèse simplifie considérablement les calculs, parce que la matrice $\mathbf{\Delta}^A$ devient alors une matrice diagonale, facile à inverser. En raison de la correspondance de type plusieurs à un entre U^A et U^B, après le processus de minimisation, nous obtenons

$$\tilde{\theta}_{j,i}^{w-\text{opt,EP}} = \frac{\pi_j^A l_{j,i}}{(1-\pi_j^A)\tau_i} \tag{4.12}$$

où $\tau_i = \sum_{j=1}^{M_i^B} \pi_j^A \big/ (1-\pi_j^A)$. Par conséquent, partant de (2.8), nous avons

$$w_i^{w-\text{opt,EP}} = \frac{1}{\tau_i} \sum_{j=1}^{M_i^B} \frac{t_j^A}{(1-\pi_j^A)}. \tag{4.13}$$

En utilisant (4.13), nous pouvons récrire (2.9) sous la forme

$$\hat{Y}_{w-\text{opt,EP}} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{w-\text{opt,EP}} \tag{4.14}$$

où

$$Z_{hj}^{w-\text{opt,EP}} = \frac{\pi_j^A}{(1-\pi_j^A)} \frac{Y_i}{\tau_i} \tag{4.15}$$

pour $j \in h$ et $j \in i$. Notons que les résultats précédents reposent sur l'hypothèse que $0 < \pi_j^A < 1$ pour tous les établissements j de U^A. Dans le cas où $\pi_j^A = 1$ pour un établissement donné j_0 d'une entreprise i, nous prenons $\tilde{\theta}_{j_0,i}^{w-\text{opt,EP}} = 1$, et $\tilde{\theta}_{j',i}^{w-\text{opt,EP}} = 0$ pour $j' \neq j_0$. Pour calculer la variance de $\hat{Y}_{w-\text{opt,EP}}$, nous utilisons la formule (2.11) avec les valeurs Z données par (4.15).

Pour l'exemple de la section 3, nous obtenons $V(\hat{Y}_{w-\text{opt,EP}}) = 22\,857$. De nouveau, il s'agit d'une réduction très importante de la variance comparativement à $V(\hat{Y}) = 1\,115\,111$ ainsi que $V(\hat{Y}_{\text{classique}}) = 80\,480$.

Méthode 6 : Utilisation de liens pondérés faiblement optimaux $\theta_{j,i}^{w-\text{opt,grp}}$ sous échantillonnage poissonnien de groupes d'établissements

Cette méthode consiste de nouveau à utiliser les liens pondérés faiblement optimaux de Deville et Lavallée (2006) décrits à la section 2.2, mais pour des groupes d'établissements. En premier lieu, nous construisons des groupes d'établissements dans la population U^A où un

groupe d'établissements j^* comprend tous les établissements qui font partie de la même strate h et appartiennent aux mêmes entreprises i. Cela crée une nouvelle population U^{A^*} contenant M^{A^*} groupes d'établissements. L'échantillon s^{A^*} de m^{A^*} groupes d'établissements contient tous les groupes d'établissements formés au moyen des établissements de l'échantillon s^A. La probabilité de sélection du groupe d'établissements j^* est donnée par

$$\pi_{j^*}^{A^*} = 1 - \frac{\binom{M_h^A - M_{j^*}}{m_h^A}}{\binom{M_h^A}{m_h^A}} = 1 - \frac{(M_h^A - M_{j^*})(M_h^A - M_{j^*} - 1)\ldots(M_h^A - M_{j^*} - m_h^A + 1)}{M_h^A(M_h^A - 1)\ldots(M_h^A - m_h^A + 1)} \quad (4.16)$$

pour $j^* \in h$, où M_{j^*} est le nombre d'établissements dans le groupe d'établissements j^*.

L'idée derrière l'utilisation de groupes d'établissements est de n'avoir qu'une seule unité appartenant à une entreprise donnée par strate. Comme, par construction, les établissements du groupe j^* d'une entreprise i appartiennent à différentes strates, leur sélection est effectuée indépendamment d'un groupe d'établissements à l'autre. Cela implique que la solution pour une optimalité faible est alors similaire à celle obtenue à la section 4.5 pour l'échantillonnage de Poisson, mais avec des groupes d'établissements. Par conséquent, nous avons

$$\tilde{\theta}_{j^*,i}^{w-\mathrm{opt,grp}} = \frac{\pi_{j^*}^A l_{j^*,i}}{(1 - \pi_{j^*}^A)\tau_i^*} \quad (4.17)$$

où $\tau_i^* = \sum_{j^*=1}^{M_i^{B^*}} \pi_{j^*}^{A^*} \Big/ (1 - \pi_{j^*}^{A^*})$ et $M_i^{B^*}$ est le nombre de groupes d'établissements contenus dans l'entreprise i.

L'utilisation de groupes d'établissements peut être considérée comme une étape intermédiaire dans le processus de sondage indirect allant de la population U^A à la population U^B. Autrement dit, le processus de sondage indirect va de la population U^A à la population U^{A^*}, puis de la population U^{A^*} à la population U^B. Ici, nous avons $j \in j^* \in i$ pour tous les établissements. En suivant les règles de transitivité définies par Deville et Lavallée (2006), nous pouvons montrer

que les liens pondérés faiblement optimaux $\tilde{\theta}_{j,i}^{w-\text{opt,grp}}$ pour $j \in j^*$ et $j^* \in i$ (et donc, $j \in i$) sont donnés par

$$\tilde{\theta}_{j,i}^{w-\text{opt,grp}} = \frac{\pi_{j^*}^{A} l_{j,i}}{(1-\pi_{j^*}^{A})\tau_i^* M_{j^*}}. \tag{4.18}$$

Donc, partant de (2.8), nous avons

$$w_i^{w-\text{opt,grp}} = \frac{1}{\tau_i^*} \sum_{j^*=1}^{M_i^{B^*}} \frac{\pi_{j^*}^{A^*}}{(1-\pi_{j^*}^{A^*})M_{j^*}} \sum_{j=1}^{M_{j^*}} \frac{t_j^A}{\pi_j^A}. \tag{4.19}$$

En utilisant (4.19), nous pouvons récrire (2.9) sous la forme

$$\hat{Y}_{w-\text{opt,grp}} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{w-\text{opt,grp}} \tag{4.20}$$

où

$$Z_{hj}^{w-\text{opt,grp}} = \frac{\pi_{j^*}^{A^*}}{(1-\pi_{j^*}^{A^*})M_{j^*}} \frac{Y_i}{\tau_i^*} \tag{4.21}$$

pour $j \in h$ et $j^* \in h$. Soulignons que les résultats qui précèdent reposent sur l'hypothèque que $0 < \pi_{j^*}^{A^*} < 1$ pour tous les groupes d'établissements j^* de U^{A^*}. Pour le cas où $\pi_{j^*}^{A^*} = 1$ pour un groupe d'établissements donné j_0^* d'une entreprise i, nous prenons $\tilde{\theta}_{j,i}^{w-\text{opt,grp}} = l_{j,i} / M_{j^*}$ pour tous les établissements j de ce groupe d'établissements j_0^*, et $\tilde{\theta}_{j,i}^{w-\text{opt,grp}} = 0$ pour tous les autres établissements ne faisant pas partie du groupe d'établissements j_0^*. Nous avons $\pi_{j^*}^{A^*} = 1$ quand au moins un établissement j appartenant au groupe j^* a une probabilité de sélection $\pi_j^A = 1$. Pour calculer la variance de $\hat{Y}_{w-\text{opt,grp}}$, nous utilisons la formule (2.11) avec les valeurs (4.21).

Pour l'exemple de la section 3, nous obtenons $V(\hat{Y}_{w-\text{opt,grp}}) = 23\,000$. De nouveau, il s'agit d'une réduction très importante de la variance comparativement à $V(\hat{Y}) = 1\,115\,111$ ainsi qu'à $V(\hat{Y}_{\text{classique}}) = 80\,480$.

4.3 Autres méthodes

Méthode 7 : Utilisation d'un établissement désigné

Comme nous l'avons mentionné plus haut, la raison de l'utilisation de groupes d'établissements dans la méthode 6 est de n'avoir qu'une seule unité appartenant à une entreprise donnée par strate. Dans un même ordre d'idée, on peut décider qu'un seul établissement représentera l'entreprise complète. C'est-à-dire que, pour chaque entreprise appartenant à U^B, nous identifions un établissement de U^A qui sera utilisé pour sélectionner l'entreprise qui en est propriétaire. Un choix naturel pour l'*établissement désigné* dans l'entreprise est celui ayant la plus grande valeur pour une variable donnée x. Par exemple, x peut être le revenu de l'établissement.

Le choix d'un établissement désigné mène à une nouvelle base de sondage U^{A_+} qui contient le même nombre d'unités que la population cible U^B, c'est-à-dire $M^{A_+} = N^B$. Puisqu'il existe une correspondance de type un à un entre l'établissement désigné et l'entreprise qui en est propriétaire, l'établissement désigné de l'entreprise i peut également être étiqueté en utilisant i. La nouvelle base de sondage U^{A_+} peut garder la même définition de stratification que la base de sondage originale U^A. Cela signifie que, si la stratification de U^A a été faite par province et par catégorie d'industrie en se fondant sur les valeurs pour les établissements, la stratification de U^{A_+} est faite en se servant des mêmes catégories en se basant sur les valeurs pour les établissements désignés.

Dans la base de sondage U^{A_+}, nous sélectionnons un échantillon s^{A_+} de m^{A_+} établissements désignés par EASSR en utilisant des fractions de sondage égales aux fractions originales, c'est-à-dire $\pi_i^{A_+} = m_h^{A_+} / M_h^{A_+} = m_h^A / M_h^A$, pour $i \in h$. L'estimation du total Y s'obtient en utilisant l'estimateur suivant

$$\hat{Y}_+ = \sum_{h=1}^{H} \frac{M_h^{A_+}}{m_h^{A_+}} \sum_{i=1}^{m_h^{A_+}} Y_i . \tag{4.22}$$

On peut montrer que l'estimateur (4.22) est sans biais et que sa variance est donnée par

$$\mathrm{Var}(\hat{Y}_+) = \sum_{h=1}^{H} M_h^{A_+} \left(\frac{M_h^{A_+} - m_h^{A_+}}{m_h^{A_+}} \right) S_{+Yh}^2 \tag{4.23}$$

où $S_{+Yh}^2 = \sum_{i=1}^{M_h^{A_+}} \left(Y_{hi} - \bar{Y}_h \right)^2 \Big/ (M_h^{A_+} - 1)$ et $\bar{Y}_h = \sum_{i=1}^{M_h^{A_+}} Y_{hi} \Big/ M_h^{A_+}$.

Mentionnons que, bien que nous ne gardions qu'un seul établissement désigné par entreprise, nous pouvons encore produire des estimations par strate pour tout domaine d'intérêt (par exemple différentes activités industrielles). En guise d'illustration, considérons le petit exemple de la section 3. Pour la première entreprise de U^B (c'est-à-dire celle dont le revenu total est égal à 2 400), l'établissement désigné serait le premier établissement de la strate à tirage complet de U^A (c'est-à-dire l'établissement comptant 25 employés). Aucun des trois autres établissements de cette entreprise ne serait disponible pour l'échantillonnage. Toutefois, si nous voulions produire une estimation pour la deuxième strate, nous limiterions simplement le calcul des valeurs Y_i dans (4.22) aux établissements appartenant à cette deuxième strate. Ici, au lieu d'utiliser $Y_i = 2\,400$ dans (4.22), nous utiliserions $Y_i = 300$. Cela correspond à une estimation par domaine (Särndal et coll. 1992).

Pour l'exemple de la section 3, nous obtenons $V(\hat{Y}_+) = 1\,820\,000$! Par cette méthode, puisqu'un établissement hérite de tous les revenus de l'entreprise, l'utilisation d'un établissement désigné est avantageuse quand il se trouve dans la strate à tirage complet. Cependant, l'établissement désigné peut être contenu dans une strate à tirage partiel, ce qui donne une strate encore plus asymétrique. Le revenu total de l'entreprise, multiplié par le poids de sondage, est attribué à cette strate à tirage partiel, ce qui augmente considérablement la variance.

Méthode 8 : Utilisation des probabilités de sélection des entreprises

Comme il est mentionné dans Lavallée (2002, 2007), en faisant appel au théorème de Rao-Blackwell, des statistiques exhaustives peuvent améliorer un estimateur existant en produisant un nouvel estimateur dont l'erreur quadratique moyenne est inférieure ou égale à celle de

l'estimateur de départ (voir Cassel, Särndal et Wretman 1977). Notons que cette forme d'amélioration a été utilisée, par exemple, par Thompson (1990) dans le contexte du sondage en grappes adaptatif.

En partant de l'estimateur (2.1) [ou (2.3)], l'estimateur $\hat{Y}_{\mathrm{RB}}$ obtenu en appliquant le théorème de Rao-Blackwell est donné par

$$\hat{Y}_{\mathrm{RB}} = \sum_{i=1}^{n^B} \frac{Y_i}{L_i^B} \sum_{j=1}^{M^A} \frac{P(t_j^A = 1 \mid s^B)}{\pi_j^A} l_{j,i} \tag{4.24}$$

où $P(t_j^A = 1 \mid s^B)$ est la probabilité d'avoir sélectionné l'établissement j dans U^A, sachant que les n^B entreprises de s^B ont été tirées de U^B.

En s'appuyant sur la correspondance de type plusieurs à un entre U^A et U^B, on peut obtenir une approximation pour la probabilité $P(t_j^A = 1 \mid s^B)$ comme il suit. Pour $j \in i$, nous avons

$$\begin{aligned} P(t_j^A = 1 \mid s^B) &\approx P(t_j^A = 1 \mid i \in s^B) \\ &= P(t_j^A = 1, i \in s^B) \,/\, P(i \in s^B) \\ &= P(t_j^A = 1) \,/\, P(i \in s^B) \\ &= \pi_j^A \,/\, \pi_i^B \end{aligned} \tag{4.25}$$

où π_i^B est la probabilité de sélection de l'entreprise $i \in U^B$, qui correspond à la probabilité de sélectionner n'importe lequel de ses M_i^B établissements. Notons que le résultat (4.25) devient exact dans le cadre de l'échantillonnage de Poisson. En se basant sur (4.25), l'estimateur (4.24) est alors approximativement équivalent à l'estimateur de Horvitz-Thompson suivant

$$\hat{Y}_{\mathrm{HT}} = \sum_{i=1}^{n^B} \frac{Y_i}{\pi_i^B}. \tag{4.26}$$

Puisque l'estimateur (4.26) n'est rien d'autre que l'estimateur de Horvitz-Thompson basé sur la sélection des entreprises, sa variance est donnée par

$$\hat{Y}_{\mathrm{HT}} = \sum_{i=1}^{N^B} \sum_{i'=1}^{N^B} \frac{(\pi_{i,i'}^B - \pi_i^B \pi_{i'}^B)}{\pi_i^B \pi_{i'}^B} Y_i Y_{i'}. \tag{4.27}$$

Pour calculer la probabilité de sélection π_i^B, il faut connaître les probabilités de sélection π_j^A pour chacun des M_i^B établissements de l'entreprise i. Souvent, il peut être difficile, voire impossible, d'obtenir ces probabilités (voir Lavallée 2002, 2007). Cela peut constituer un obstacle important à l'utilisation de l'estimateur (4.26) en pratique, et en fait, il s'agit de l'une des raisons menant à l'utilisation de la MGPP. Cependant, ici, l'obtention des probabilités s'avère possible parce qu'on dispose de la base de sondage U^A complète pour sélectionner les établissements. L'utilisation de l'EASSR stratifié simplifie également la tâche. Les probabilités de sélection π_i^B peuvent alors être calculées en adaptant la formule (4.16). Il est également possible de calculer les probabilités de sélection conjointe $\pi_{i,i'}^B$, mais il s'agit d'un exercice plus difficile.

Pour l'exemple de la section 3, nous obtenons $V(\hat{Y}_{\text{HT}}) = 14\,545$, valeur qui correspond à la variance la plus faible donnée par les méthodes proposées.

5 Simulations en utilisant des données réelles

Les simulations reflètent une enquête auprès des entreprises type de Statistique Canada. Nous avons choisi trois populations qui sont fréquemment la cible d'enquêtes menées par Statistique Canada. Les populations d'établissements des secteurs de la fabrication, du commerce de détail et de la restauration ont été extraites du Registre des entreprises (RE). On sait que ces trois populations ont une distribution asymétrique pour des variables économiques telles que le revenu, surtout les deux premières. Les simulations ont été effectuées sous EASSR stratifié selon le secteur d'activité, la région et la catégorie de revenu. L'algorithme de Lavallée-Hidiroglou (1988) a été utilisé pour créer les catégories de revenu, déterminer la taille d'échantillon et effectuer la répartition. Les établissements étaient répartis en trois strates en fonction de leur taille, à savoir une strate à tirage complet et deux strates à tirage partiel. Un coefficient de variation de 5 % était visé dans chaque strate par secteur d'activité et par région. Le tableau qui suit contient certaines statistiques sur la population.

Tableau 5.1
Populations simulées, tailles d'échantillon et statistiques

Secteur d'activité	N^B	M^A	m^A	Revenu moyen	Variance	Asymétrie
Fabrication	96 955	100 109	2 223	4 364 808	$1{,}08 \times 10^{16}$	164
Commerce de détail	142 020	159 247	3 627	2 034 111	$3{,}29 \times 10^{14}$	133
Restauration	107 358	113 425	2 439	561 764	$4{,}43 \times 10^{12}$	106
Total	346 333	372 781	8 289		---	

La variable de revenu disponible dans le RE a été utilisée comme variable d'intérêt y. Puisque les valeurs de cette variable sont connues pour toutes les unités, aucun échantillon n'a dû être sélectionné pour la plupart des méthodes. Il convient de souligner que, pour la méthode 2 ($\theta_{j,i}$ proportionnel à une mesure de taille x_j), on s'est servi du nombre d'employés.

Sauf pour les méthodes 7 et 8, nous avons calculé les variances réelles d'après les données en utilisant la formule (2.11). Pour la méthode 7, nous avons utilisé la formule (4.23). Pour la méthode 8, nous avons dû calculer les probabilités de sélection π_i^B réelles de toutes les entreprises. Nous aurions pu calculer ces probabilités en nous servant de formules exactes, mais nous avons choisi de le faire en exécutant une simulation Monte Carlo en deux étapes. L'une des raisons est que la formule de variance (4.27) qui utilise les probabilités de sélection conjointe $\pi_{i,i'}^B$ fait intervenir un trop grand nombre de paires (j, j'). Pour la première étape de la simulation Monte Carlo, nous avons sélectionné 20 000 échantillons selon le plan d'EASSR stratifié décrit plus haut. Pour chaque échantillon, nous avons déterminé quelle entreprise finissait par être sélectionnée. Sur ces 20 000 échantillons, nous avons pu estimer la probabilité de sélection π_i^B de chaque entreprise i sous ce plan de sondage à probabilités inégales. Une fois que ces probabilités ont été calculées, nous avons exécuté une autre simulation Monte Carlo pour calculer la variance. Nous avons sélectionné $R = 20\,000$ échantillons s^A d'établissements en utilisant de nouveau le plan d'EASSR stratifié décrit plus haut. Nous avons alors obtenu les échantillons correspondants s^B d'entreprises. Pour chaque réplique $r, r = 1, \ldots, R,$ nous avons

produit une estimation $\hat{Y}_{\mathrm{HT},r}$ de Y en utilisant l'estimateur (4.26). Nous avons calculé la variance en utilisant la formule

$$V_{\mathrm{MC}}(\hat{Y}_{\mathrm{HT}}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_{\mathrm{HT},r} - \hat{\bar{Y}}_{\mathrm{HT}}^{(R)})^2 \tag{5.1}$$

où $\hat{\bar{Y}}_{\mathrm{HT}}^{(R)} = \sum_{r=1}^{R}\hat{Y}_{\mathrm{HT},r} / R$. Notons que, comme l'estimateur (4.26) est sans biais, nous avons $\hat{\bar{Y}}_{\mathrm{HT}}^{(R)} \approx Y$.

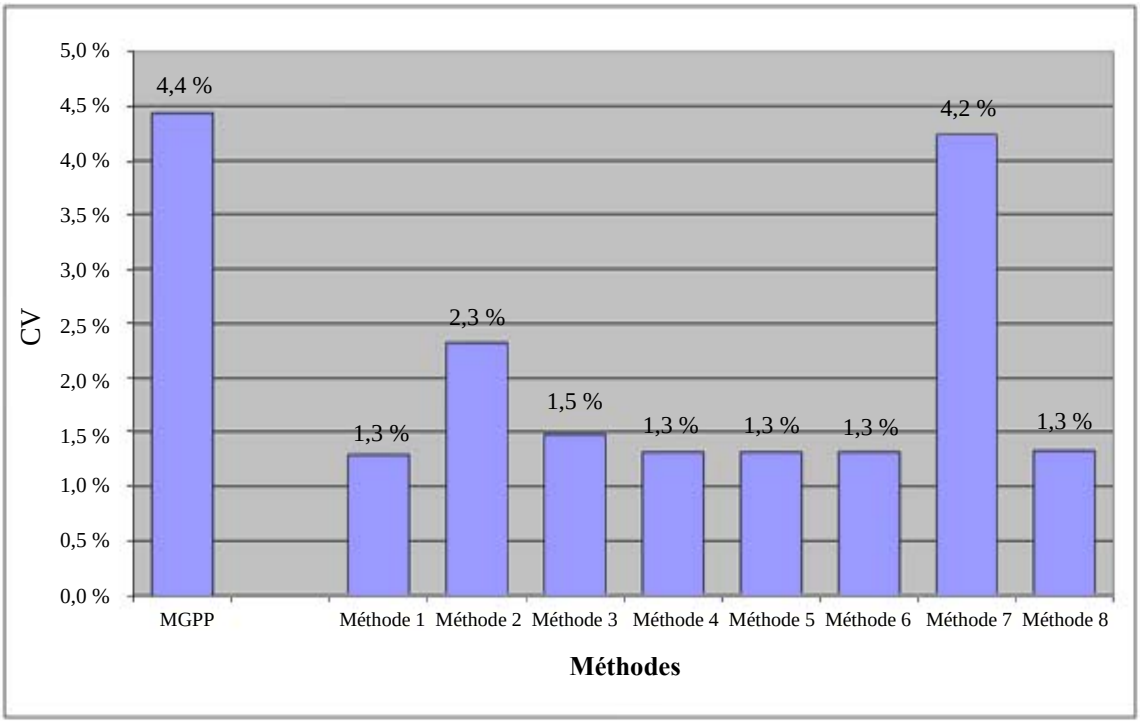
Pour chaque estimateur $\hat{Y}$, nous avons calculé le coefficient de variation en utilisant

$$\mathrm{CV}(\hat{Y}) = \frac{\sqrt{V(\hat{Y})}}{Y}. \tag{5.2}$$

5.1 Résultats de la simulation

Pour la MGPP classique et pour toutes les méthodes présentées, nous avons calculé les estimations, les variances et les coefficients de variation. Le graphique qui suit donne les CV obtenus au niveau national.

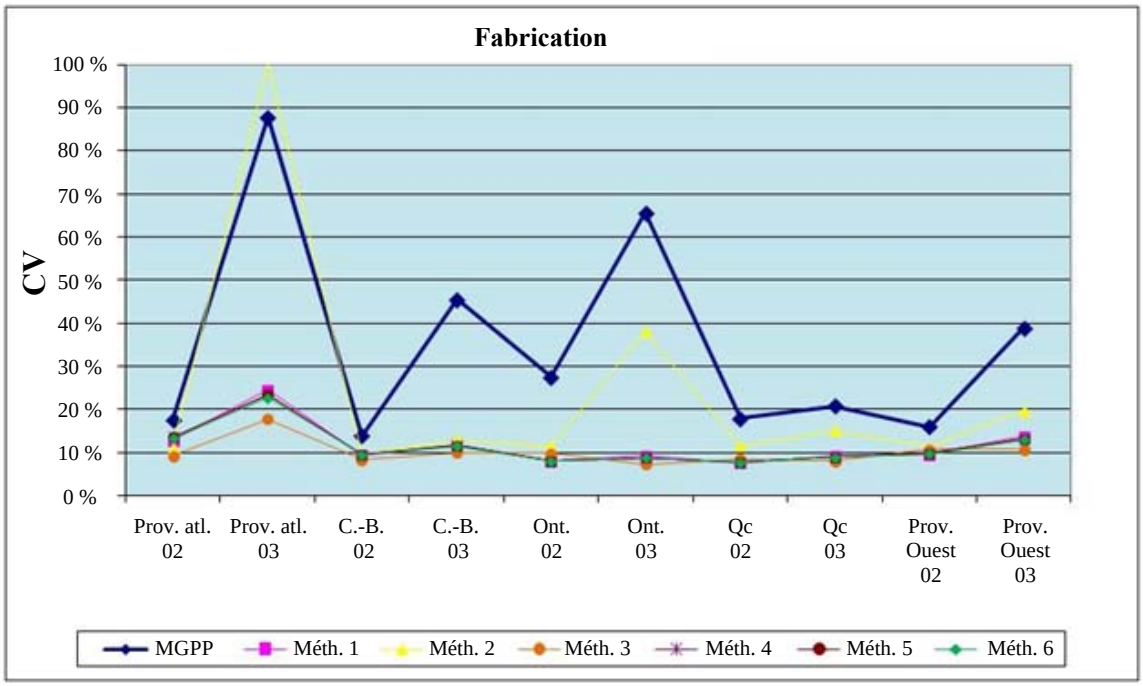
Graphique 5.1 Coefficients de variation par méthode



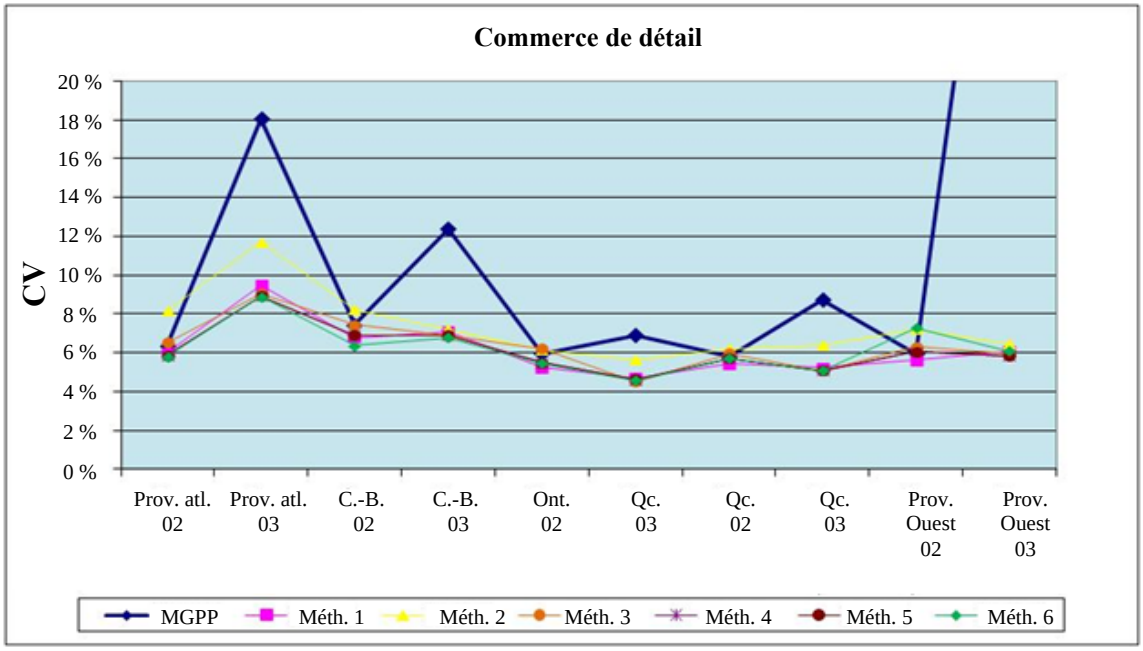
À part la méthode 7, toutes les méthodes donnent une diminution de la variance et la réduction est souvent considérable. Comme il est décrit plus haut, la méthode 7 (utilisant les établissements désignés) consiste à sélectionner un seul établissement au sein d'une entreprise en se basant sur la variable auxiliaire et à affecter l'entreprise complète à cet établissement. Autrement dit, un établissement donné hérite de tous les revenus de l'entreprise. Cette approche est avantageuse quand l'établissement désigné se trouve dans une strate à tirage complet. Cependant, si l'établissement désigné se trouve dans une strate à tirage partiel, la distribution à l'intérieur de cette strate devient encore plus asymétrique. On attribue à cette strate la totalité des revenus de l'entreprise multipliés par le poids de sondage de cette entreprise, ce qui fait augmenter considérablement la variance. Toutes les autres méthodes donnent des résultats raisonnables que nous analysons en détail au moyen des graphiques qui suivent.

Les graphiques qui suivent donnent le CV pour chaque strate à tirage partiel par industrie.

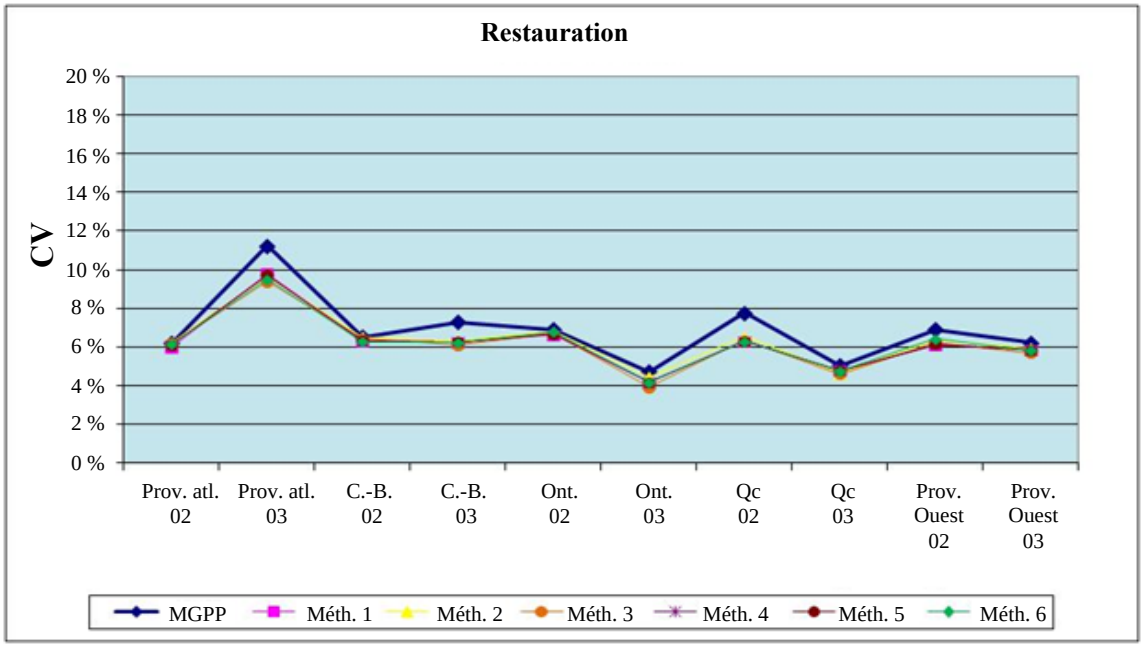
Graphique 5.2 CV pour la fabrication par strate



Graphique 5.3 CV pour le commerce de détail par strate



Graphique 5.4 CV pour la restauration par strate



Note 1 : L'échelle du CV n'est pas la même pour la fabrication que pour les deux autres industries.

Note 2 : Les CV par strate des méthodes 7 et 8 ne sont pas présentés ici parce qu'ils ne sont pas pertinents pour la comparaison qui est faite. En effet, la notion de strate n'est pas la même pour ces deux méthodes que pour les autres. La stratification définie par le plan de sondage original est effectuée au niveau de l'établissement. Pour les méthodes 7 et 8, l'échantillonnage est effectué au niveau de l'entreprise et, par conséquent, une strate type pour les méthodes 1 à 6 devient un domaine pour les méthodes 7 et 8. Évidemment, les variances associées aux méthodes 7 et 8 sont beaucoup plus grandes, ce qui rend non pertinente toute comparaison avec les autres méthodes.

Les CV sont particulièrement élevés pour la MGPP classique dans certaines strates, surtout pour le secteur de la fabrication. Nous nous y attendions, parce que ce dernier est, des trois secteurs examinés, celui pour lequel l'asymétrie de la distribution de la variable d'intérêt était la plus prononcée. En outre, dans ce secteur, nous avons des établissements dont les revenus peuvent varier fortement au sein d'une même entreprise, et ces établissements peuvent être répartis entre plusieurs strates. Ce sont les raisons pour lesquelles la variance de la MGPP classique devient très élevée.

Tous les graphiques montrent que l'utilisation de n'importe laquelle des méthodes proposées donne lieu à une réduction du CV, donc de la variance. Les CV sont généralement plus faibles pour les autres méthodes que pour la MGPP classique (représentée par la ligne bleu foncé avec les losanges).

5.2 Comparaison des méthodes proposées

La méthode 1 donne des résultats très prometteurs, étant donné sa simplicité. Elle produit certains des CV les plus faibles observés pour l'ensemble des méthodes. Elle cible réellement la source du problème de la MGPP, à savoir la nécessité d'une répartition inégale des poids, proportionnelle dans une certaine mesure à la taille de la variable d'intérêt. Puisque la plupart des enquêtes auprès des entreprises s'appuient sur une stratification par taille, cette méthode donne de bons résultats. Elle a également l'avantage de ne pas dépendre directement de la variable d'intérêt.

La méthode 2 utilise une variable auxiliaire (ici, le nombre d'employés) pour répartir les poids. Cette variable n'est pas bien corrélée avec la variable d'intérêt, ce qui explique pourquoi cette méthode produit la plus petite réduction de la variance. En fait, il s'agit d'une version plus faible de la méthode 3.

La méthode 3 répartit le poids proportionnellement à la variable d'intérêt y, c'est-à-dire le revenu de l'établissement dans l'entreprise. La méthode donne de très bons résultats au niveau

national ainsi que provincial. Le résultat est généralement plus élevé que pour les méthodes 1, 4, 5 et 6, à cause de l'asymétrie prononcée de la distribution du revenu.

Les méthodes 4, 5 et 6 donnent des résultats fort semblables, produisant un CV compris entre 6 % et 10 % pour le commerce de détail et la restauration, et entre 10 % et 25 % pour la fabrication. La similarité des résultats obtenus par les trois méthodes est raisonnable parce que toutes visent à produire la variance la plus faible possible. Chaque fois qu'un établissement d'une entreprise se trouve dans une strate à tirage complet, ces méthodes concentrent toutes les valeurs sur cet établissement et attribuent des liens nuls à tous les autres établissements de l'entreprise. Il s'agit d'un choix naturel pour minimiser la variance, puisque la contribution de cette entreprise à la variance devient nulle. Comme un établissement d'une grande entreprise peut se trouver dans une strate à tirage complet, la variance devient plus faible que celle produite par n'importe quelle autre méthode. Ce sont les raisons pour lesquelles ces trois méthodes représentent le meilleur moyen de partager les poids. Cependant, il n'a pas été possible de déterminer laquelle des trois est la meilleure.

La méthode 7 ne donne pas de bons résultats. Rappelons que pour cette méthode, nous avons permis qu'un seul établissement désigné représente l'ensemble de l'entreprise. Les valeurs de la variable d'intérêt de tous les établissements sont totalisées au niveau de l'entreprise, puis attribuées à cet établissement désigné. La distribution de la variable d'intérêt par strate devient encore plus asymétrique, ce qui produit une plus grande variance. Du point de vue de l'échantillonnage, une entreprise se retrouve dans une strate unique (parce qu'elle est représentée par un seul établissement) qui pourrait ne pas être une strate à tirage complet. En outre, la méthode n'est pas efficace quand il faut produire des estimations au niveau provincial ou du secteur d'activité. Les estimations ne peuvent pas bénéficier de la stratification des établissements, alors que cela est possible dans le cas des autres méthodes. Ce fait contribue également à une plus grande variance.

La méthode 8 utilise les probabilités de sélection des entreprises obtenues par des simulations. Elle donne de bons résultats, le CV étant de 1,3 % au niveau national. Sa performance égale celle des méthodes 4, 5 et 6. Cependant, elle peut être difficile à appliquer en pratique. Nous devons soit calculer explicitement les probabilités de sélection de deuxième ordre $\pi_{i,i'}^{B}$, ce qui peut être très compliqué, ou les estimer par simulation, ce qui prend beaucoup de temps.

6 Conclusion

La MGPP est applicable dans le contexte des enquêtes auprès des entreprises, mais elle peut donner lieu à des variances importantes à cause de la distribution asymétrique des données associées à ce genre de population. Nous avons proposé huit méthodes de rechange pour partager les poids afin de réduire la variance. Nous avons ensuite exécuté une simulation pour comparer ces méthodes en utilisant des données réelles. Les simulations reproduisaient une enquête auprès des entreprises type réalisée selon un plan EASSR stratifié. Toutes les méthodes proposées ont réduit la variance. Les simulations ont montré que les meilleures méthodes s'obtenaient en partageant les poids proportionnellement à π_j^A (méthode 1) ou en utilisant des liens pondérés faiblement optimaux (méthodes 4, 5 et 6). L'utilisation de liens pondérés sous échantillonnage de Poisson (méthode 5) est la méthode que nous privilégions. En théorie, cette méthode est proche de la méthode optimale et elle est également la plus simple à mettre en œuvre.

Bibliographie

Brewer, K.R.W., et Hanif, M. (1983). *Sampling With Unequal Probabilities.* New York : Springer-Verlag, 164 pages.

Cassel, C.-M., Särndal, C.-E. et Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling.* New York : John Wiley & Sons, Inc.

Delorme, C. (2000). Data allocation: A new survey process for business surveys. ICES-II, Proceedings of the Second International Conference on Establishment Surveys, Survey Methods for Businesses, Farms, and Institutions, Buffalo, New York, 17 au 21 juin 2000, 1285-1290.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, Vol. 32, 2, 185-196.

Kalton, G., et Brick, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, Vol. 21, 1, 37-49.

Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode Généralisée du Partage des Poids*. Éditions de l'Université de Bruxelles, Bruxelles.

Lavallée, P. (2007). *Indirect Sampling*. New York : Springer.

Lavallée, P., et Caron, P. (2001). Estimation par la méthode généralisée du partage des poids : le cas du couplage d'enregistrements. *Techniques d'enquête*, Vol. 27, 2, 171-187.

Lavallée, P., et Hidiroglou, M.A. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, Vol. 14, 1, 35-45.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Statistique Canada (2010). *Guide sommaire sur le Registre des entreprises*. Division du Registre des Entreprises, Statistique Canada, Ottawa, 9 pages, juillet 2010.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, Vol. 85, 412, 1050-1059.

Thompson, S.K. (1991). Stratified adaptive cluster sampling. *Biometrika*, Vol. 78, 2, 389-397.

Thompson, S.K. (1992). *Sampling*. New York : John Wiley & Sons, Inc.

Thompson, S.K. (2002). *Sampling*, 2ième Édition. New York : John Wiley & Sons, Inc.

Thompson, S.K., et Seber, G.A. (1996). *Adaptive Sampling*. New York : John Wiley & Sons, Inc.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2ième Édition. New York : Springer.

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 243-255
Statistique Canada, N° 12-001-X au Catalogue

De la performance des estimateurs sur petits domaines autocalés sous le modèle au niveau du domaine de Fay-Herriot

Yong You, J.N.K. Rao et Mike Hidiroglou[1]

Résumé

Nous considérons deux méthodes distinctes d'autocalage pour l'estimation des moyennes de petit domaine fondée sur le modèle au niveau du domaine de Fay-Herriot (FH), à savoir la méthode de You et Rao (2002) appliquée au modèle FH et la méthode de Wang, Fuller et Qu (2008) basée sur des modèles augmentés. Nous établissons un estimateur de l'erreur quadratique moyenne de prédiction (EQMP) de l'estimateur de You-Rao (YR) d'une moyenne de petit domaine qui, sous le modèle vrai, est correct jusqu'aux termes de deuxième ordre. Nous présentons les résultats d'une étude en simulation du biais relatif de l'estimateur de l'EQMP de l'estimateur YR et de l'estimateur de l'EQMP de l'estimateur de Wang, Fuller et Qu (WFQ) obtenu sous un modèle augmenté. Nous étudions aussi l'EQMP et les estimateurs de l'EQMP des estimateurs YR et WFQ obtenus sous un modèle mal spécifié.

Mots clés : Modèle augmenté; meilleure prédiction linéaire sans biais empirique; erreur quadratique moyenne de prédiction; erreur de spécification du modèle.

1 Introduction

On parle de petit domaine quand la taille de l'échantillon d'une sous-population ou d'un domaine est trop petite pour produire des estimations directes fiables des paramètres du domaine. Il faut alors utiliser des estimateurs indirects fondés sur un modèle qui s'appuient sur des données d'échantillon empruntées à des domaines apparentés grâce à des modèles de lien afin de réaliser un important gain d'efficacité par rapport aux estimateurs directs. Dans le présent article, nous nous concentrerons sur un modèle de base au niveau du domaine, appelé modèle de Fay-Herriot (FH) (Fay et Herriot 1979), et sur les meilleurs prédicteurs linéaires sans biais empiriques (EBLUP, pour *Empirical Best Linear Unbiased Predictor*) associés des moyennes de petit domaine.

1. Yong You, Division de la recherche et de l'innovation en statistique, Statistique Canada, Ottawa, Canada. Courriel : yong.you@statcan.gc.ca; J.N.K. Rao, École de mathématiques et de statistique, Université Carleton, Ottawa, Canada; Mike Hidiroglou, Division de la recherche et de l'innovation en statistique, Statistique Canada, Ottawa, Canada.

Ces EBLUP ne concordent pas forcément avec les estimateurs directs que l'on privilégie en pratique pour les agrégats parce qu'ils sont fondés sur des échantillons de suffisamment grande taille pour satisfaire les exigences de fiabilité et qu'ils ne dépendent pas de modèles. Par conséquent, on procède souvent à un calage pour contraindre les estimations à concorder. Le présent article porte sur deux méthodes d'autocalage fondées sur des modifications des prédicteurs EBLUP. La première méthode (WFQ), que l'on doit à Wang, Fuller et Qu (Wang et coll. 2008), s'appuie sur un modèle FH augmenté et sur les EBLUP associés, appelés estimateurs WFQ. La deuxième méthode (YR), également proposée par Wang, Fuller et Qu, est fondée sur une approche utilisée par You et Rao (2002) dans le contexte de modèles au niveau de l'unité. Les estimateurs YR s'obtiennent en modifiant les estimateurs optimaux des paramètres de régression utilisés dans les EBLUP afin d'imposer la concordance. Étant donné les modifications des EBLUP, les erreurs quadratiques moyennes de prédiction (EQMP) des estimateurs WFQ et YR autocalés seront plus importantes.

Le présent article a deux grands objectifs. Le premier consiste à comparer les EQMP et leurs estimateurs pour les estimateurs YR et WFQ des moyennes de petit domaine. À la section 2, nous examinons les expressions des EQMP et des estimateurs associés pour l'EBLUP et pour l'estimateur WFQ. À la section 3, nous développons les expressions de l'EQMP et de son estimateur pour l'estimateur YR. À la section 4, nous comparons les EQMP et leurs estimateurs au moyen d'une étude en simulation.

Le deuxième objectif est d'examiner les propriétés des estimateurs WFQ et YR et des estimateurs de leur EQMP lorsque le modèle de lien est mal spécifié dans le modèle FH en raison de l'omission d'une variable. Wang, Fuller et Qu ont aussi étudié l'effet de l'erreur de spécification du modèle de lien dans un cas particulier, pour lequel ils ont montré que l'estimateur YR produit un grand biais, mais non l'estimateur WFQ. Cependant, ce résultat tenait au fait que, dans leur étude en simulation, la variable d'augmentation était fortement corrélée avec une covariable omise (coefficient de corrélation $\rho = 0{,}983$). Par conséquent, leur modèle

augmenté était, en réalité, proche du modèle vrai, d'où la performance supérieure de l'estimateur WFQ. À la section 4, nous considérons une variable omise différente, qui est faiblement corrélée à la variable d'augmentation (ρ =-0,116), de sorte que le modèle augmenté est également mal spécifié. Nous comparons les biais, ainsi que les EQMP et leurs estimateurs pour les EBLUP et les estimateurs YR et WFQ obtenus sous le modèle mal spécifié.

2 Estimateurs EBLUP et WFQ

Supposons que nous ayons m petits domaines et des estimateurs directs sans biais sous le plan, y_i, des moyennes de domaine $\theta_i, i = 1, \ldots, m.$ Le modèle FH se rapporte à (y_i, θ_i) et aux variables auxiliaires au niveau du domaine associées $x_i = (x_{i1}, \ldots, x_{ip})'$ avec $x_{i1} = 1.$ Sous l'hypothèse d'échantillonnage indépendant dans les divers domaines, le modèle FH peut s'écrire sous forme d'un modèle linéaire mixte donné par

$$y_i = \theta_i + e_i = x_i'\beta + v_i + e_i, i = 1, \ldots, m, \tag{2.1}$$

où $\theta_i = x_i'\beta + v_i$ est le modèle de lien, et e_i est l'erreur d'échantillonnage de moyenne 0 et de variance connue σ_i^2, qui est indépendante de l'effet aléatoire propre au domaine v_i. L'échantillonnage est indépendant dans les divers domaines et les v_i sont supposés indépendants et identiquement distribués de moyenne 0 et de variance σ_v^2.

Le meilleur prédicteur linéaire sans biais (BLUP, pour *Best Linear Unbiased Predictor*) de θ_i sous le modèle « vrai » (2.1) est donné par

$$\tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i) x_i'\tilde{\beta}, \tag{2.2}$$

où $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$ et $\tilde{\beta}$ est l'estimateur par les moindres carrés pondérés (MCP) optimal de β donné par

$$\tilde{\beta} = \left(\sum_{i=1}^{m} \gamma_i x_i x_i' \right)^{-1} \left(\sum_{i=1}^{m} \gamma_i x_i y_i \right), \tag{2.3}$$

voir Rao (2003, page 116). L'estimateur $\tilde{\theta}_i$ dépend de la variance sous le modèle σ_v^2 inconnue, et en remplaçant σ_v^2 dans (2.2) par un estimateur approprié $\hat{\sigma}_v^2$, nous obtenons l'EBLUP :

$$\hat{\theta}_i^{\text{EBLUP}} = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i) x_i' \hat{\beta}, \tag{2.4}$$

où $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \sigma_i^2)$ et $\hat{\beta}$ s'obtient à partir de (2.3) en remplaçant γ_i par $\hat{\gamma}_i$. Dans le présent article, nous utilisons l'estimateur du maximum de vraisemblance restreint (REML pour *Restricted Maximum Likelihood*) de σ_v^2, en supposant que v_i et e_i suivent une loi normale. La somme pondérée $\sum_{i=1}^m w_i \hat{\theta}_i^{\text{EBLUP}}$ des EBLUP (2.4) ne concorde pas nécessairement avec l'estimateur direct pondéré correspondant de l'agrégat $\sum_{i=1}^m w_i y_i$, où les w_i sont des poids spécifiés au préalable de façon que $\sum_{i=1}^m w_i y_i$ soit un estimateur de l'agrégat (total ou moyenne) convergent sous le plan. Un écart important entre $\sum_{i=1}^m w_i \hat{\theta}_i^{\text{EBLUP}}$ et $\sum_{i=1}^m w_i y_i$ peut être le signe d'un échec du modèle dont il faudra s'occuper avant de procéder au calage, comme l'a souligné le rédacteur associé.

Un estimateur de l'erreur quadratique moyenne de prédiction $\text{EQMP}(\hat{\theta}_i^{\text{EBLUP}}) = \text{E}(\hat{\theta}_i^{\text{EBLUP}} - \theta_i)^2$ correct jusqu'aux termes de deuxième ordre, sous l'estimation par la méthode du REML, est donné par

$$\text{eqmp}(\hat{\theta}_i^{\text{EBLUP}}) = g_{1i} + g_{2i} + 2g_{3i}, \tag{2.5}$$

où $g_{1i} = \hat{\gamma}_i \sigma_i^2$ est le terme principal d'ordre $O(1)$, et g_{2i} et g_{3i} sont des termes d'ordre plus faible $O(m^{-1})$ qui tiennent compte de la variabilité de $\hat{\beta}$ et $\hat{\sigma}_v^2$ respectivement (Rao 2003, page 128). Nous avons

$$g_{2i} = (1 - \hat{\gamma}_i)^2 x_i' \hat{V}(\tilde{\beta}) x_i = \hat{\sigma}_v^2 (1 - \hat{\gamma}_i)^2 x_i' \left(\sum_{i=1}^m \hat{\gamma}_i x_i x_i' \right)^{-1} x_i$$

et

$$g_{3i} = 2(\sigma_i^2)^2 (\hat{\sigma}_v^2 + \sigma_i^2)^{-3} \left\{ \sum_{i=1}^m (\hat{\sigma}_v^2 + \sigma_i^2)^{-2} \right\}^{-1},$$

où $\hat{V}(\tilde{\beta})$ est l'estimateur de $\mathrm{V}(\tilde{\beta}) = \sigma_v^2 \left(\sum_{i=1}^m \gamma_i x_i x_i' \right)^{-1}$. L'estimateur de l'EQMP (2.5) est presque sans biais en ce sens que

$$\mathrm{E}\{\mathrm{eqmp}(\hat{\theta}_i^{\mathrm{EBLUP}})\} = \mathrm{EQMP}(\hat{\theta}_i^{\mathrm{EBLUP}}) + O(m^{-2}).$$

Wang, Fuller et Qu ont obtenu l'estimateur EBLUP, $\hat{\eta}_i^{\mathrm{EBLUP}}$, sous le modèle FH augmenté qui suit :

$$y_i = x_i'\delta + w_{ei}\lambda + u_i + e_i \equiv \eta_i + e_i, \tag{2.6}$$

où les effets aléatoires u_i sont indépendants avec $\mathrm{E}(u_i) = 0$ et $\mathrm{Var}(u_i) = \sigma_u^2$, et sont indépendants de e_i. La variable auxiliaire d'augmentation utilisée est $w_{ei} = w_i\sigma_i^2$. Ils ont montré que l'estimateur EBLUP de η_i sous le modèle augmenté (2.6), $\hat{\eta}_i^{\mathrm{EBLUP}} \equiv \hat{\theta}_i^{\mathrm{WFQ}}$, possède la propriété d'autocalage au sens où il satisfait

$$\sum_{i=1}^m w_i\hat{\theta}_i^{\mathrm{WFQ}} = \sum_{i=1}^m w_i y_i.$$

L'EBLUP $\hat{\theta}_i^{\mathrm{WFQ}}$ sous le modèle augmenté (2.6) s'obtient à partir de (2.4) en remplaçant x_i' par (x_i', w_{ei}), $\hat{\sigma}_v^2$ par $\hat{\sigma}_u^2$ et $\hat{\beta}'$ par $(\hat{\delta}', \hat{\lambda})$. L'estimateur de $\mathrm{EQMP}(\hat{\theta}_i^{\mathrm{WFQ}})$ s'obtient de manière similaire à partir de (2.5). Sous le modèle vrai (2.1), l'EQMP de $\hat{\theta}_i^{\mathrm{WFQ}}$ sera plus grande que celle de $\hat{\theta}_i^{\mathrm{EBLUP}}$, mais son estimateur, $\mathrm{eqmp}(\hat{\theta}_i^{\mathrm{WFQ}})$, restera presque sans biais sous le modèle vrai, comme l'ont souligné le rédacteur associé et un examinateur, parce que le modèle vrai est un cas particulier du modèle augmenté avec $\lambda = 0$.

3 Estimateur YR

Wang, Fuller et Qu ont appliqué la méthode de You et Rao (2002) au modèle (2.1) et obtenu un estimateur de θ_i donné par

$$\hat{\theta}_i^{\mathrm{YR}} = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)x_i'\hat{\beta}^{\mathrm{YR}}, \tag{3.1}$$

où $\hat{\beta}^{\mathrm{YR}}$ s'obtient à partir de

$$\tilde{\beta}^{\mathrm{YR}} = \left\{ \sum_{i=1}^{m} w_i (1-\gamma_i) x_i x_i' \right\}^{-1} \left\{ \sum_{i=1}^{m} w_i (1-\gamma_i) x_i y_i \right\} \tag{3.2}$$

en remplaçant γ_i par $\hat{\gamma}_i$. Notons que l'estimateur YR (3.1) a la même forme que l'EBLUP $\hat{\theta}_i^{\mathrm{EBLUP}}$, mais qu'il utilise un estimateur non optimal pour β. Les estimateurs YR $\hat{\theta}_i^{\mathrm{YR}}$ ont la propriété d'autocalage, c'est-à-dire $\sum_{i=1}^{m} w_i \hat{\theta}_i^{\mathrm{YR}} = \sum_{i=1}^{m} w_i y_i$, puisque, en vertu de (3.2),

$$\sum_{i=1}^{m} w_i (1-\gamma_i)(y_i - x_i' \tilde{\beta}^{\mathrm{YR}}) = 0.$$

Cependant, l'EQMP de $\hat{\theta}_i^{\mathrm{YR}}$ sera légèrement plus grande que celle de $\hat{\theta}_i^{\mathrm{EBLUP}}$ fondé sur $\hat{\beta}$, parce que $\hat{\beta}$ est asymptotiquement plus efficace que $\hat{\beta}^{\mathrm{YR}}$.

Dans le cas de $\hat{\theta}_i^{\mathrm{EBLUP}}$, l'estimateur de $\mathrm{EQMP}(\hat{\theta}_i^{\mathrm{YR}})$ contient les termes g_{1i}, g_{2i} et g_{3i}. Nous devons estimer la variance de $\tilde{\beta}^{\mathrm{YR}}$ afin d'obtenir le terme g_{2i} dans l'estimateur de $\mathrm{EQMP}(\hat{\theta}_i^{\mathrm{YR}})$; les autres termes g_{1i} et g_{3i} ne sont pas affectés. Il découle de (3.2) que

$$\mathrm{V}(\tilde{\beta}^{\mathrm{YR}}) = \sigma_v^2 \left\{ \sum_{i=1}^{m} w_i (1-\gamma_i) x_i x_i' \right\}^{-1} \left\{ \sum_{i=1}^{m} w_i^2 (1-\gamma_i)^2 \gamma_i^{-1} x_i x_i' \right\} \left\{ \sum_{i=1}^{m} w_i (1-\gamma_i) x_i x_i' \right\}^{-1}. \tag{3.3}$$

L'estimateur $\hat{V}(\tilde{\beta}^{\mathrm{YR}})$ s'obtient en substituant $\hat{\sigma}_v^2$ et $\hat{\gamma}_i$ à σ_v^2 et γ_i dans (3.3).

L'estimateur de $\mathrm{EQMP}(\hat{\theta}_i^{\mathrm{YR}})$ est donné par

$$\mathrm{eqmp}(\hat{\theta}_i^{\mathrm{YR}}) = g_{1i} + g_{2i}^{\mathrm{YR}} + 2 g_{3i}, \tag{3.4}$$

où

$$g_{2i}^{\mathrm{YR}} = (1-\hat{\gamma}_i)^2 x_i' \hat{V}(\tilde{\beta}^{\mathrm{YR}}) x_i.$$

L'estimateur de l'EQMP (3.4) est presque sans biais sous le modèle vrai (2.1), comme l'estimateur de l'EQMP (2.5) de $\hat{\theta}_i^{\mathrm{EBLUP}}$.

Remarque : Tout estimateur $\hat{y}_i$ of θ_i peut être ajusté comme il suit :

$$\hat{y}_i^a = \hat{y}_i + a_i \left(\sum_{i=1}^{m} w_i y_i - \sum_{i=1}^{m} w_i \hat{y}_i \right)$$

pour a_i spécifié de manière à satisfaire la contrainte de calage $\sum_{i=1}^{m} w_i \hat{y}_i^a = \sum_{i=1}^{m} w_i y_i$, où $\sum_{i=1}^{m} w_i a_i = 1$. En particulier, nous pouvons utiliser $\hat{y}_i = \hat{\theta}_i^{\text{EBLUP}}$ pour obtenir l'estimateur EBLUP ajusté. Comme l'ont mentionné WFQ, $\hat{\theta}_i^{\text{YR}}$ ainsi que $\hat{\theta}_i^{\text{WFQ}}$ sont des estimateurs de la forme $\hat{y}_i^a$ parce que $\sum_{i=1}^{m} w_i y_i - \sum_{i=1}^{m} w_i \hat{y}_i$ est égal à zéro si l'on fixe $\hat{y}_i$ égal à $\hat{\theta}_i^{\text{YR}}$ ou à $\hat{\theta}_i^{\text{WFQ}}$. Tout ensemble d'estimateurs $\{\hat{y}_i\}$ qui satisfait $\sum_{i=1}^{m} w_i y_i = \sum_{i=1}^{m} w_i \hat{y}_i$ possède la propriété d'autocalage.

4 Étude en simulation

Nous avons réalisé une petite étude en simulation du biais et de l'EQMP des estimateurs YR et WFQ, ainsi que du biais relatif et du coefficient de variation (CV) des estimateurs YR et WFQ de l'EQMP, en utilisant les conditions de simulation de WFQ. Comme ces auteurs, nous avons considéré $m = 50$ petits domaines (États), avec les tailles d'échantillon n_i variant de 7 à 58 et les variances d'échantillonnage connues $\sigma_i^2 = 16 / n_i$. Les poids w_i utilisés dans la contrainte de calage sont donnés par $w_i = \left(\sum_{i=1}^{50} P_i\right)^{-1} P_i$, où P_i est la population de l'État i indiquée dans le tableau 4.1 de WFQ.

Dans WFQ, le modèle vrai est donné par

$$y_i = \beta_0 + \beta_1 z_i + v_i + e_i, \tag{4.1}$$

avec $\beta_0 = 6$, $\beta_1 = 3$, $v_i \sim \text{N}(0, \sigma_v^2 = 1)$ et $e_i \sim \text{N}(0, 16 / n_i)$. WFQ ont utilisé $z_i = P_i^{0,2} - \left(\sum_{i=1}^{50} P_i^{0,2}\right) \Big/ 50$ comme valeurs des covariables dans le modèle vrai, et les z_i choisies sont fortement corrélées avec la variable augmentée pondérée associée $w_{ei} = w_i \sigma_i^2$; coefficient de corrélation $\rho = 0{,}983$.

Dans WFQ, le modèle mal spécifié choisi est le modèle de la moyenne

$$y_i = \mu + \tilde{v}_i + e_i \tag{4.2}$$

avec $\tilde{v}_i \sim \text{N}(0, \tilde{\sigma}_v^2)$ et $e_i \sim \text{N}(0, 16 / n_i)$, et en raison de ce choix, le modèle augmenté associé

$$y_i = \delta_0 + \lambda w_i \sigma_i^2 + u_i + e_i \tag{4.3}$$

est presque identique au modèle vrai (4.1).

Afin de refléter l'erreur de spécification au moyen du modèle augmenté, nous avons produit 50 valeurs $(z_1, \ldots, z_{50})$ à partir de $\mathrm{N}(1, 1)$ et les avons maintenues fixes dans les diverses simulations. Le coefficient de corrélation résultant entre les valeurs produites z_i et $w_i\sigma_i^2$ est égal à -0,116. Nous avons ensuite pris le modèle correspondant

$$y_i = \beta_0 + \beta_1 z_i + v_i + e_i \tag{4.4}$$

avec $\beta_0 = 6$, $\beta_1 = 3$, $v_i \sim \mathrm{N}(0, 1)$ et $e_i \sim \mathrm{N}(0, 16 / n_i)$ comme étant le modèle vrai et le modèle de la moyenne (4.2) comme étant le modèle mal spécifié. Nous avons considéré deux scénarios : i) le modèle vrai est donné par (4.4) et il est modélisé et spécifié correctement; ii) le modèle vrai est donné par (4.4) et il est modélisé et spécifié incorrectement comme étant le modèle de la moyenne (4.2).

Nous avons créé $R = 10\,000$ jeux de données $\{(y_i^{(r)}, z_i), i = 1, \ldots, 50; r = 1, ..., R\}$ sous le modèle vrai (4.4), où $y_i^{(r)} = \beta_0 + \beta_1 z_i + v_i^{(r)} + e_i^{(r)}$ avec $v_i^{(r)}$ tiré de la loi $\mathrm{N}(0, 1)$ et $e_i^{(r)}$ de la loi $\mathrm{N}(0, 16 / n_i)$, respectivement. Soit $\theta_i^{(r)} = \beta_0 + \beta_1 z_i + v_i^{(r)}$ la valeur de la vraie moyenne θ_i et $\hat{\theta}_i^{(r)}$ la valeur associée d'un estimateur $\hat{\theta}_i$ (EBLUP, YR ou WFQ) pour la r^{e} exécution de la simulation sous le scénario (i) ou (ii). Alors, le biais absolu et l'EQMP simulés de $\hat{\theta}_i$ sont donnés par

$$\left| \mathrm{B}(\hat{\theta}_i) \right| = \left| R^{-1} \sum_{r=1}^{R} (\hat{\theta}_i^{(r)} - \theta_i^{(r)}) \right|$$

et

$$\mathrm{EQMP}(\hat{\theta}_i) = R^{-1} \sum_{r=1}^{R} (\hat{\theta}_i^{(r)} - \theta_i^{(r)})^2.$$

Le biais relatif absolu (BRA) d'un estimateur de l'EQMP, $\mathrm{eqmp}(\hat{\theta}_i)$, est calculé comme étant $\left| \mathrm{BR}\{\mathrm{eqmp}(\hat{\theta}_i)\} \right|$, où

$$\mathrm{BR}\{\mathrm{eqmp}(\hat{\theta}_i)\} = \frac{1}{R} \sum_{r=1}^{R} \{\mathrm{eqmp}(\hat{\theta}_i^{(r)}) - \mathrm{EQMP}(\hat{\theta}_i)\} \Big/ \mathrm{EQMP}(\hat{\theta}_i),$$

où $\text{eqmp}(\hat{\theta}_i^{(r)})$ est la valeur de $\text{eqmp}(\hat{\theta}_i)$ pour la r^{e} exécution de la simulation. De même, le coefficient de variation (CV) de $\text{eqmp}(\hat{\theta}_i)$ est calculé comme

$$\text{CV}\{\text{eqmp}(\hat{\theta}_i)\} = \left[\frac{1}{R} \sum_{r=1}^{R} \{\text{eqmp}(\hat{\theta}_i^{(r)}) - \text{EQMP}(\hat{\theta}_i)\}^2 \right]^{\frac{1}{2}} \Big/ \text{EQMP}(\hat{\theta}_i) .$$

Nous résumons les résultats de la simulation en présentant le premier quartile, le troisième quartile, la médiane et la moyenne des valeurs de $\left|\text{B}(\hat{\theta}_i)\right|$, $\text{EQMP}(\hat{\theta}_i)$, $\left|\text{BR}\{\text{eqmp}(\hat{\theta}_i)\}\right|$ %, et $\text{CV}\{\text{eqmp}(\hat{\theta}_i)\}$ %. Nous avons également regroupé les domaines en fonction des variances d'échantillonnage, $\sigma_i^2 = 16 / n_i$, en trois groupes, puis nous avons calculé les mesures susmentionnées séparément pour chaque groupe. Les résultats étaient très semblables à ceux obtenus sans groupement. Par conséquent, pour simplifier, nous ne présentons les résultats que pour le cas non groupé.

Le tableau 4.1 donne les mesures sommaires pour le biais absolu des estimateurs EBLUP, YR et WFQ, sous les scénarios (i) et (ii). Comme prévu, les valeurs des mesures du biais absolu sont toutes négligeables pour les trois estimateurs sous le scénario (i), confirmant l'absence de biais des estimateurs sous ce scénario. Par ailleurs, le biais absolu est grand relativement à l'erreur-type sous le scénario (ii) pour les trois estimateurs, ce qui donne une beaucoup plus grande EQMP que sous le scénario (i). Il est intéressant de noter, lorsque l'on examine les valeurs pour le scénario (ii) au tableau 4.1, que le biais absolu est presque le même pour les divers estimateurs, ce qui donne à penser que l'autocalage ne réduit peut-être pas le biais absolu en cas d'erreur de spécification du modèle. Une explication plausible est que l'estimateur YR donne le même poids, $\hat{\gamma}_i$, à l'estimateur direct; de même, l'estimateur WFQ attribue un poids proche de $\hat{\gamma}_i$. Nos résultats ne contredisent pas la conclusion de WFQ selon laquelle leur estimateur réduit le biais absolu, car leur modèle augmenté a quasiment éliminé l'erreur de spécification du modèle.

Tableau 4.1
Mesures sommaires du biais absolu des estimateurs EBLUP, YR et WFQ

Mesure	EBLUP	YR	WFQ
	Scénario (i)		
1er quartile	0,002	0,002	0,002
Médiane	0,005	0,005	0,005
Moyenne	0,007	0,007	0,007
3e quartile	0,009	0,010	0,009
	Scénario (ii)		
1er quartile	0,092	0,085	0,083
Médiane	0,144	0,153	0,152
Moyenne	0,254	0,255	0,251
3e quartile	0,316	0,311	0,314

Il convient aussi de mentionner que l'estimateur EBLUP ajusté $\hat{y}_i^a$ défini dans la remarque en utilisant $\hat{y}_i = \hat{\theta}_i^{\text{EBLUP}}$ pourrait donner de meilleurs résultats que les estimateurs YR et WFQ en ce qui concerne le biais absolu sous le scénario (ii). Par exemple, si l'estimateur EBLUP donne une sous-estimation dans la plupart des domaines, le terme de correction $a_i \left(\sum_{i=1}^m w_i y_i - \sum_{i=1}^m w_i \theta_i^{\text{EBLUP}} \right)$ sera positif et compensera donc la sous-estimation. Par ailleurs, le terme de correction est nul pour les estimateurs avec autocalage YR et WFQ.

Le tableau 4.2 donne les mesures sommaires de l'EQMP des estimateurs EBLUP, YR et WFQ, sous les scénarios (i) et (ii). Il montre que, sous le scénario (i), les EQMP des trois estimateurs EBLUP, YR et WFQ sont pratiquement les mêmes, excepté que celle de l'EBLUP est légèrement plus faible que prévu. Sous le scénario (ii), nous observons de nouveau très peu de différence entre les EQMP des trois estimateurs. Toutefois, l'erreur de spécification du modèle les accroît chacune considérablement. Par exemple, pour le troisième quartile (correspondant aux faibles valeurs de n_i), la valeur passe de 0,621 à 1,100 pour l'estimateur EBLUP, de 0,626 à 1,094 pour l'estimateur YR et de 0,634 à 1,117 pour l'estimateur WFQ. Cette augmentation de l'EQMP est due au biais induit par l'erreur de spécification du modèle.

Tableau 4.2
Mesures sommaires pour l'EQMP des estimateurs EBLUP, YR et WFQ

Mesure	EBLUP	YR	WFQ
	Scénario (i)		
1er quartile	0,403	0,406	0,403
Médiane	0,471	0,476	0,472
Moyenne	0,499	0,505	0,507
3e quartile	0,621	0,626	0,634
	Scénario (ii)		
1er quartile	0,562	0,567	0,567
Médiane	0,751	0,748	0,752
Moyenne	0,885	0,886	0,883
3e quartile	1,100	1,094	1,117

Le tableau 4.3 donne les mesures sommaires du biais relatif absolu en pourcentage et du CV en pourcentage des estimateurs de l'EQMP associés aux estimateurs EBLUP, YR et WFQ, sous le scénario (i) correspondant à l'absence d'erreur de spécification du modèle vrai (4.4). Le tableau 4.4 donne les résultats pour les estimateurs de l'EQMP sous le scénario (ii), qui correspond à l'erreur de spécification du modèle vrai (4.4) sous forme du modèle de la moyenne (4.2).

Tableau 4.3
Mesures sommaires pour le biais relatif absolu (%) et le CV (%) des estimateurs de l'EQMP associés aux estimateurs EBLUP, YR et WFQ sous le scénario (i) : Modèle (4.4) spécifié correctement

Mesure	EBLUP	YR	WFQ
	Biais relatif absolu (BRA) %		
1er quartile	0,678	0,650	0,627
Médiane	1,186	1,216	1,173
Moyenne	1,354	1,349	1,372
3e quartile	1,674	1,716	1,863
	Coefficient de variation (CV) %		
1er quartile	12,93	12,44	13,19
Médiane	15,64	15,40	15,78
Moyenne	16,02	15,69	15,78
3e quartile	19,39	19,07	19,05

Tableau 4.4
Mesures sommaires pour le biais relatif absolu (%) et le CV (%) des estimateurs de l'EQMP associés aux estimateurs EBLUP, YR et WFQ sous le scénario (ii) : Modèle (4.4) mal spécifié

Mesure	EBLUP	YR	WFQ
	Biais relatif absolu (BRA) %		
1er quartile	4,848	4,850	4,857
Médiane	7,653	7,776	7,466
Moyenne	10,412	10,469	10,221
3e quartile	13,423	13,611	13,776
	Coefficient de variation (CV) %		
1er quartile	5,510	5,579	4,700
Médiane	7,894	7,317	7,610
Moyenne	10,52	10,55	10,17
3e quartile	13,42	13,03	13,11

Pour ce qui est des mesures du biais relatif absolu (BRA) et du CV des estimateurs de l'EQMP, le tableau 4.3 montre que, sous le scénario (i), le BRA en pourcentage est très faible (< 2 %) pour les trois estimateurs de l'EQMP, ce qui confirme que ces estimateurs sont presque sans biais. En outre, le CV en pourcentage est pratiquement le même pour les trois estimateurs de l'EQMP. Sous le scénario (ii), le tableau 4.4 montre que le BRA en pourcentage augmente considérablement, mais est pratiquement le même pour les trois estimateurs de l'EQMP. Par exemple, la valeur pour le troisième quartile passe de 1,674 à 13,423 dans le cas de l'EBLUP, de 1,716 à 13,611 dans le cas de l'estimateur YR et de 1,863 à 13,776 dans le cas de l'estimateur WFQ. De plus, le CV en pourcentage est presque le même pour les trois estimateurs de l'EQMP. La comparaison des valeurs du CV du tableau 4.4 (scénario (ii)) aux valeurs correspondantes du tableau 4.3 (scénario (i)) permet de constater que les CV sont effectivement beaucoup plus faibles sous le scénario (ii) que sous le scénario (i). Cela tient au fait que l'EQMP (utilisée dans le dénominateur du calcul du CV) est nettement plus grande sous le scénario (ii) que sous le scénario (i).

5 Conclusion

Nous avons étudié les propriétés de l'EBLUP et de deux estimateurs avec autocalage, YR et WFQ, sous le modèle au niveau du domaine de Fay-Herriot. Nous avons présenté un estimateur presque sans biais de l'EQMP de l'estimateur YR à la section 3. Notre étude en simulation indique que les trois méthodes donnent des résultats très semblables en ce qui concerne l'EQMP des estimateurs, ainsi que le biais relatif et le CV des estimateurs de l'EQMP. Toutefois, sous une erreur grossière de spécification du modèle due à l'omission d'une variable, l'EQMP augmente considérablement pour les trois estimateurs, sauf si le modèle augmenté de WFQ élimine presque complètement l'erreur de spécification. En pratique, il est difficile de tenir compte de l'erreur de spécification du modèle due à des variables omises inconnues. Il est également intéressant de souligner que l'autocalage pourrait ne pas réduire le biais sous une erreur de spécification du modèle que le modèle augmenté de WFQ ne prend pas en compte.

Remerciements

Nous remercions un examinateur et le rédacteur associé de leurs nombreux commentaires et suggestions constructifs.

Bibliographie

Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Rao, J.N.K. (2003). *Small Area Estimation*, New York : John Wiley & Sons, Inc.

Wang, J., Fuller, W.A. et Qu, Y. (2008). Estimation pour petits domaines sous une contrainte. *Techniques d'enquête*, 34, 1, 33-40.

You, Y., et Rao, J.N.K. (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 257-267
Statistique Canada, N° 12-001-X au Catalogue

Estimation prudente de la variance pour des plans d'échantillonnage avec probabilités d'inclusion par paire nulles

Peter M. Aronow et Cyrus Samii[1]

Résumé

Nous considérons une estimation prudente de la variance pour l'estimateur de Horvitz-Thompson d'un total de population sous des plans d'échantillonnage avec probabilités d'inclusion par paire nulles, appelés plans « non mesurables ». Nous décomposons l'estimateur de variance de Horvitz-Thompson classique sous ce genre de plan et caractérisons le biais de manière précise. Nous élaborons une correction du biais qui est garantie d'être faiblement prudente (non biaisée négativement) quelle que soit la nature de la non-mesurabilité. L'analyse jette de la lumière sur les conditions sous lesquelles l'estimateur de variance de Horvitz-Thompson classique donne de bons résultats malgré la non-mesurabilité et où la correction du biais prudente peut être meilleure que les approximations utilisées habituellement.

Mots clés : Estimation de Horvitz-Thompson; plans non mesurables; estimation de la variance.

1 Introduction

Il arrive que les plans d'échantillonnage donnent des paires d'unités ayant une probabilité nulle d'être incluses conjointement dans l'échantillon. L'énoncé des propriétés du total de population finie de Horvitz et Thompson (1952) indique clairement qu'une estimation générale sans biais de la variance pour les estimateurs des totaux de population est impossible pour ce genre de plan *non mesurable* (Särndal, Swensson et Wretman 1992, page 33). L'obtention de méthodes optimales pour l'estimation de la variance dans ces cas demeure une question ouverte. Le présent article analyse la nature du biais que la non-mesurabilité introduit pour l'estimateur de Horvitz-Thompson classique et étudie une approche en vue de corriger ce biais de manière prudente. Bien que nos résultats ne puissent offrir une solution au problème de non-mesurabilité pour toutes les applications pratiques, nous clarifions les conditions sous lesquelles l'estimateur

1. Peter M. Aronow, Department of Political Science, Yale University, 77 Prospect St., New Haven, CT 06520. Courriel : peter.aronow@yale.edu; Cyrus Samii, Department of Politics, New York University, 19 West 4th St., New York, NY 10012. Courriel : cds2083@nyu.edu.

classique donne de bons résultats et où la correction du biais prudente est meilleure que les approximations utilisées habituellement.

Malgré leurs inconvénients théoriques, les plans d'échantillonnage avec probabilités d'inclusion par paire nulles sont assez fréquents. Un plan non mesurable commun est celui comportant le tirage d'unités simples ou de grappes à partir d'un ensemble de strates. Cela peut se produire si la population ou une sous-population d'intérêt est accidentellement peu nombreuse dans les cellules de stratification. Un autre plan non mesurable est celui d'un échantillon systématique dans lequel les indices d'unité sont échantillonnés par multiples à partir d'une liste en partant d'une valeur initiale aléatoire. Dans ces plans, les unités dont les indices sont des multiples de différentes valeurs initiales ont une probabilité d'inclusion conjointe nulle.

Des méthodes approximatives ont été proposées pour des cas particuliers, comme il en est question dans Hansen, Hurwitz et Madow (1953, section 9.15), Särndal et coll. (1992, chapitre 3) et Wolter (2007, chapitre 2 et 8). Dans le cas d'une seule unité par strate, une approche courante est de regrouper les strates et à supposer que les unités ont été tirées par échantillonnage aléatoire simple à partir de la strate plus grande, fusionnée. Pour les échantillons systématiques, l'approche classique consiste à utiliser une approximation fondée sur une hypothèse d'échantillonnage aléatoire simple avec remise. Ces méthodes approximatives sont généralement biaisées dans une mesure qu'il est impossible de déterminer d'après les données. Dans certains cas, on peut montrer que le biais a tendance à être positif, mais il n'en est généralement pas ainsi, surtout quand les probabilités d'inclusion par paire nulles surviennent de manière désordonnée.

Le présent article débute à la section 2 par la décomposition du biais de l'estimateur de variance de Horvitz-Thompson sous non-mesurabilité. Cette décomposition montre précisément comment les conditions appliquées aux données sous-jacentes donnent lieu à un biais plus ou moins important. Nous montrons aussi, à la section 3, comment une simple application de l'inégalité de Young donne une correction du biais et une classe d'estimateurs dont on est certain

qu'ils présenteront un biais faiblement positif et seront sans biais dans des conditions particulières. Nous discutons des implications pour les travaux appliqués à la section 4.

2 Estimation de la variance pour l'estimateur de Horvitz-Thompson

Considérons une population U indicée par $1, \ldots, k, \ldots N$ et un plan d'échantillonnage tel que la probabilité d'inclusion de l'unité k dans l'échantillon est π_k, et la probabilité d'inclusion conjointe des unités k et l est π_{kl}. Sous un plan mesurable, il y a deux conditions : 1) $\pi_k > 0$ et π_k est connu pour tout $k \in U$ et 2) $\pi_{kl} > 0$ et π_{kl} est connu pour tous $k, l \in U$. Les plans non mesurables comprennent ceux pour lesquels l'une ou l'autre de ces deux conditions n'est pas vérifiée. Ne pas satisfaire à la première condition empêche l'obtention d'estimation sans biais des totaux.

L'estimateur de Horvitz-Thompson d'un total de population est $\hat{t} = \sum_{k \in s} y_k / \pi_k = \sum_{k \in U} I_k \, y_k / \pi_k$, où $I_k \in \{0,1\}$ est l'indicateur d'inclusion de l'unité k, la seule composante stochastique de l'expression, avec $\mathrm{E}(I_k) = \pi_k$, la probabilité d'inclusion, et s et U désignent l'échantillon et la population, respectivement. Définissons $\mathrm{E}(I_k I_l) = \pi_{kl}$, qui est la probabilité que les unités k et l provenant de U soient toutes deux incluses dans l'échantillon. Puisque $I_k I_k = I_k, \mathrm{E}(I_k I_k) = \pi_{kk} = \pi_k$ par construction. Lorsque la condition 1 est vérifiée, comme nous le supposons tout au long du présent exposé, l'estimateur de Horvitz-Thompson est sans biais.

2.1 Propriétés de l'estimateur de variance de Horvitz-Thompson sous mesurabilité

Selon Horvitz et Thompson (1952), la variance de l'estimateur de Horvitz-Thompson pour le total est donnée par

$$\mathrm{Var}(\hat{t}) = \sum_{k \in U} \sum_{l \in U} \mathrm{Cov}(I_k, I_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_{k \in U} \mathrm{Var}(I_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in U \setminus k} \mathrm{Cov}(I_k, I_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Nous désignons un échantillon provenant d'un plan mesurable par s^M, et un estimateur sans biais de $\mathrm{Var}(\hat{t})$ sur s^M est

$$\widehat{\mathrm{Var}}(\hat{t}) = \sum_{k \in s^M} \sum_{l \in s^M} \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_{k \in U} \sum_{l \in U} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$

où la seule partie stochastique de l'expression est $I_k I_l$, et l'absence de biais est relative à $\mathrm{E}(I_k I_l) = \pi_{kl}$.

2.2 Propriétés de l'estimateur de variance de Horvitz-Thompson sous non-mesurabilité

Nous examinons maintenant le cas où la condition 2 n'est pas vérifiée : $\pi_{kl} = 0$ pour certaines unités $k, l \in U$. Comme I_k est une variable aléatoire de Bernoulli avec probabilité $\pi_k, \mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$ pour $k \neq l$, et $\mathrm{Cov}(I_k, I_k) = \mathrm{Var}(I_k) = \pi_k(1 - \pi_k)$. Alors, nous pouvons réexprimer la variance susmentionnée sous la forme

$$\begin{aligned}
\mathrm{Var}(\hat{t}) &= \sum_{k \in U} \pi_k (1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in U \setminus k} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \\
&= \sum_{k \in U} \pi_k (1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} > 0\}} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} - \underbrace{\sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} y_k y_l}_{A}.
\end{aligned}$$

Pour k et l tels que $\pi_{kl} = 0$, le plan d'échantillonnage ne permettra jamais une estimation sans biais de la composante de la variance étiquetée A ci-dessus, puisque nous n'observerons jamais y_k et y_l ensemble. Nous désignons un échantillon provenant d'un plan dans lequel la condition 2 n'est pas respectée par s^0. Quand $\widehat{\mathrm{Var}}(\hat{t})$ est appliqué à s^0, le résultat est sans biais pour $\mathrm{Var}(\hat{t}) + A$. Nous énonçons cela formellement comme il suit :

Proposition 1. Quand s^0 désigne un échantillon provenant d'un plan avec une probabilité $\pi_{kl} = 0$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}(\hat{t})\right] = \mathrm{Var}(\hat{t}) + \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} y_k y_l = \mathrm{Var}(\hat{t}) + A.$$

Preuve. Le résultat découle de

$$\begin{aligned}
\mathrm{E}\left[\sum_{k\in s^0}\sum_{l\in s^0}\frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\right] &= \mathrm{E}\left[\sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\right] \\
&= \sum_{k\in U}\mathrm{Var}(I_k)\left(\frac{y_k}{\pi_k}\right)^2 + \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}>0\}}\mathrm{Cov}(I_k, I_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \\
&= \mathrm{Var}(\hat{t}) + \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} y_k y_l = \mathrm{Var}(\hat{t}) + A.
\end{aligned}$$

L'estimateur de variance de Horvitz-Thompson classique, s'il est appliqué à des plans avec probabilités d'inclusion par paire nulles, peut par conséquent présenter un biais positif ou négatif. Si les valeurs de y_k, y_l sont systématiquement non négatives (ou systématiquement non positives), alors le biais est systématiquement non négatif. Quand les valeurs peuvent être positives ou négatives, A est égal à la somme de produits croisés de résultats qui n'apparaissent jamais ensemble sous le plan d'échantillonnage. Si les résultats conjointement exclusifs sont centrés sur zéro, l'absence de corrélation entre ces résultats aurait alors tendance à produire un biais faible, l'existence d'une corrélation positive, un biais positif et celle d'une corrélation négative, un biais négatif.

3 Correction du biais prudente pour l'estimateur de variance de Horvitz-Thompson sous non-mesurabilité

Le cas où A peut être inférieur à zéro pour un plan non mesurable laisse entendre qu'il est nécessaire d'appliquer une certaine correction garantissant que le biais soit faiblement borné vers le bas par zéro. Nous commençons par élaborer une correction générale du biais qui est garantie d'être prudente, et nous fournissons plus tard un cas particulier simplifié d'usage pratique.

3.1 Formulation générale

Considérons l'estimateur de variance suivant :

$$\widehat{\mathrm{Var}}_C(\hat{t}) = \sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k} + I_l \frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l} \right),$$

où a_{kl}, b_{kl} sont des nombres réels positifs tels que $1/a_{kl} + 1/b_{kl} = 1$ pour toutes les paires k, l avec $\pi_{kl} = 0$. Il est garanti que l'estimateur produira une valeur espérée égale ou supérieure à la variance réelle pour tous les plans et il est par conséquent prudent. Nous énonçons cette propriété formellement :

Proposition 2. La valeur espérée de $\widehat{\mathrm{Var}}_C(\hat{t})$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right] \geq \mathrm{Var}(\hat{t}).$$

Preuve. En vertu de l'inégalité de Young,

$$\frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} \geq |y_k||y_l|,$$

si $1/a_{kl} + 1/b_{kl} = 1$. Définissons A^* tel que

$$A^* = \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} \frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} \geq \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} |y_k||y_l| \geq \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} y_k y_l = A$$

et

$$A^* \geq \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} |y_k||y_l| \geq \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} -y_k y_l = -A.$$

Par conséquent

$$\mathrm{Var}(\hat{t}) + A + A^* \geq \mathrm{Var}(\hat{t}).$$

L'estimateur de Horvitz-Thompson associé de A^* sera

$$\hat{A}^* = \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k} + I_l \frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l} \right),$$

qui n'est pas biaisé par $E(I_k) = \pi_k$ et $E(I_l) = \pi_l$.

Puisque $E(\hat{A}^*) = A^*$, en vertu de la proposition 1,

$$\mathrm{E}\left[\sum_{k\in s^0}\sum_{l\in s^0}\frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \hat{A}^*\right] = \mathrm{Var}(\hat{t}) + A + A^*$$

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right] \geq \mathrm{Var}(\hat{t}).$$

Par substitution des termes,

$$\mathrm{E}\left[\sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}\left(I_k\frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k} + I_l\frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l}\right)\right] \geq \mathrm{Var}(\hat{t}).$$

$\widehat{\mathrm{Var}}_C(\hat{t})$ est justifié en tant qu'estimateur prudent pour le cas où l'on ne sait pas si A est positif. Cet estimateur est sans biais sous une condition spéciale :

Corollaire 1. Si, pour toutes les paires k, l telles que $\pi_{kl} = 0$, i) $|y_k|^{a_{kl}} = |y_l|^{b_{kl}}$ et ii) $-y_k y_l = |y_k||y_l|$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right] = \mathrm{Var}(\hat{t}).$$

Preuve. En vertu de i), ii) et l'inégalité de Young,

$$\frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} = |y_k||y_l| = -y_k y_l.$$

Donc,

$$A^* = \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}\frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} = \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}|y_k||y_l| = \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} -y_k y_l = -A.$$

Il s'ensuit que $\mathrm{Var}(\hat{t}) + A + A^* = \mathrm{Var}(\hat{t})$ et $\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right] = \mathrm{Var}(\hat{t})$.

Si certaines unités k, l se trouvent dans des grappes (c'est-à-dire $\Pr(I_k \neq I_l) = 0$), ces unités doivent être totalisées dans une unité plus grande avant l'estimation. Le regroupement des unités aura tendance à réduire le biais de l'estimateur de variance, parce que seules les paires de totaux au niveau de la grappe seront incluses dans A^*, plutôt que toutes les paires constituantes.

3.2 Cas particulier simplifié

En général, il serait difficile d'attribuer des valeurs optimales de a_{kl} et b_{kl} à toutes les paires k, l de sorte que $\pi_{kl} = 0$. Nous examinons donc plutôt un cas intuitif, en attribuant les valeurs $a_{kl} = b_{kl} = 2$ à toutes les paires :

$$\widehat{\mathrm{Var}}_{C2}(\hat{t}) = \sum_{k \in U} \sum_{l \in \{U:\pi_{kl}>0\}} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \sum_{k \in U} \sum_{l \in \{U \setminus k:\pi_{kl}=0\}} \left(I_k \frac{y_k^2}{2\pi_k} + I_l \frac{y_l^2}{2\pi_l} \right).$$

Un cas particulier de $\widehat{\mathrm{Var}}_C(\hat{t}), \widehat{\mathrm{Var}}_{C2}(\hat{t})$ est également prudent :

Corollaire 2. La valeur espérée de $\widehat{\mathrm{Var}}_{C2}(\hat{t})$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_{C2}(\hat{t}) \right] \geq \mathrm{Var}(\hat{t}).$$

Preuve. Pour toutes les paires k, l telles que $\pi_{kl} = 0, 1 / a_{kl} + 1 / b_{kl} = 1$. La proposition 1 est par conséquent vérifiée.

La décision de fixer toutes les valeurs à $a_{kl} = b_{kl} = 2$ est justifiée par le fait que cela donnera la valeur la plus faible de l'estimateur $\widehat{\mathrm{Var}}_C(\hat{t})$ sous la contrainte que a_{kl} et b_{kl} soit fixes en tant que constantes a et b sur tous les k, l.

Corollaire 3. Parmi la classe d'estimateurs $\widehat{\mathrm{Var}}_{\mathrm{Cab}}(\hat{t})$, définie comme l'ensemble d'estimateurs $\widehat{\mathrm{Var}}_C(\hat{t})$ tels que tous les $a_{kl} = a$ et tous les $b_{kl} = b, \widehat{\mathrm{Var}}_{C2}(\hat{t}) = \min_{a,b} \left[\widehat{\mathrm{Var}}_{\mathrm{Cab}}(\hat{t}) \right]$.

Preuve. Par de simples calculs algébriques,

$$\begin{aligned}
\widehat{\mathrm{Var}}_{\mathrm{Cab}}(\hat{t}) &= \sum_{k \in U} \sum_{l \in \{U:\pi_{kl}>0\}} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \sum_{k \in U} \sum_{l \in \{U \setminus k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^a}{a\pi_k} + I_l \frac{|y_l|^b}{b\pi_l} \right) \\
&= \widehat{\mathrm{Var}}(\hat{t}) + \sum_{k \in U} \sum_{l \in \{U \setminus k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^a}{a\pi_k} + I_l \frac{|y_l|^b}{b\pi_l} \right) \\
&= \widehat{\mathrm{Var}}(\hat{t}) + \frac{1}{2} \sum_{k \in U} \sum_{l \in \{U \setminus k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^a}{a\pi_k} + I_l \frac{|y_l|^b}{b\pi_l} + I_k \frac{|y_k|^b}{b\pi_k} + I_l \frac{|y_l|^a}{a\pi_l} \right) \\
&= \widehat{\mathrm{Var}}(\hat{t}) + \frac{1}{2} \sum_{k \in U} \sum_{l \in \{U \setminus k:\pi_{kl}=0\}} \left(\frac{I_k}{\pi_k} \left(\frac{|y_k|^a}{a} + \frac{|y_k|^b}{b} \right) + \frac{I_l}{\pi_l} \left(\frac{|y_l|^a}{a} + \frac{|y_l|^b}{b} \right) \right).
\end{aligned}$$

En vertu de l'inégalité de Young, étant donné $1/a + 1/b = 1, \left|y_k\right|^a / a + \left|y_k\right|^b / b \geq y_k^2$, mais l'égalité doit être vérifiée si $a = b = 2$. De même, $\left|y_l\right|^a / a + \left|y_l\right|^b / b \geq y_l^2$, mais l'égalité doit être vérifiée si $a = b = 2$. Puisque $I_k / \pi_k \geq 0$ et $I_l / \pi_l \geq 0$, tout choix $a \neq b$ ne peut donner que $\widehat{\mathrm{Var}}_{\mathrm{Cab}}(\hat{t}) \geq \widehat{\mathrm{Var}}_{C2}(\hat{t})$.

Sachant toutes les valeurs de y_k et y_l, il est possible de dériver un vecteur optimal des valeurs de a_{kl} et b_{kl} qui varient sur k, l, mais un tel exercice pourrait ne pas avoir de valeur pratique.

4 Applications

La proposition 1 montre que le biais de l'estimateur de variance de Horvitz-Thompson sous non-mesurabilité est

$$A = \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} y_k y_l .$$

Cette expression, conjuguée au fait que $A^* \geq A$ indique clairement que le degré de biais dans $\widehat{\mathrm{Var}}(\hat{t})$ et $\widehat{\mathrm{Var}}_C(\hat{t})$ dépend beaucoup du nombre de paires pour lesquelles les probabilités d'inclusion sont nulles. Pour les plans pour lesquels ce nombre est faible, $\widehat{\mathrm{Var}}(\hat{t})$ peut fournir un estimateur raisonnable et prudent quand y_k est de même signe pour tout k, et $\widehat{\mathrm{Var}}_C(\hat{t})$ peut fournir un estimateur raisonnable et prudent quand y_k peut prendre des signes différents pour certaines valeurs de k. Un exemple observé fréquemment est celui de l'échantillonnage stratifié où, pour une proportion relativement faible de cas, nous avons de petites strates desquelles nous ne tirons qu'une seule unité.

Pour les plans qui aboutissent à de nombreuses paires dont les probabilités d'inclusion sont nulles, $\widehat{\mathrm{Var}}(\hat{t})$ et $\widehat{\mathrm{Var}}_C(\hat{t})$ pourraient être exagérément prudents et d'autres estimateurs pourraient être préférés sur la base de critères tels que l'erreur quadratique moyenne. Un exemple frappant est celui de l'échantillonnage systématique. En effet, Särndal et coll. (1992, page 76) suggèrent que, sous échantillonnage systématique, l'estimateur de variance de Horvitz-Thompson, $\widehat{\mathrm{Var}}(\hat{t})$, peut donner un « résultat absurde ». L'expression pour A indique

clairement pourquoi il en serait ainsi. Wolter (2007, chapitre 8) montre que des estimateurs biaisés plus simples, tels que l'estimateur de variance avec remise (Hansen-Hurwitz), peuvent être fiables, si ce n'est légèrement prudents, pour une grande gamme de scénarios de données sous échantillonnage systématique avec probabilités égales et avec probabilités proportionnelles à la taille (PPT). Néanmoins, l'estimateur avec remise ne tient pas compte adéquatement de la variance d'échantillonnage quand la variance du résultat à l'intérieur des grappes de l'échantillon systématique est plus faible que la variance entre grappes. Dans de tels cas, $\widehat{\mathrm{Var}}(\hat{t})$ bornerait cette variance en espérance quand les résultats sont tous de même signe, et $\widehat{\mathrm{Var}}_C(\hat{t})$ bornerait systématiquement cette variance en espérance. Naturellement, il se pourrait encore que le biais soit trop grand pour que l'estimateur soit de beaucoup d'utilité, et nous ne suggérons donc pas que $\widehat{\mathrm{Var}}(\hat{t})$ et $\widehat{\mathrm{Var}}_C(\hat{t})$ fourniraient une solution complète au problème d'estimation de la variance pour l'échantillonnage systématique sous forte corrélation intra-grappe.

Les résultats des études en simulation peuvent être consultés dans un supplément (à l'adresse https://files.nyu.edu/cds2083/public/docs/smj_suppl.pdf). Ils illustrent les propriétés de $\widehat{\mathrm{Var}}(\hat{t})$ et $\widehat{\mathrm{Var}}_C(\hat{t})$ comparativement à d'autres options utilisées fréquemment dans des scénarios appliqués. Les simulations illustrent des situations dans lesquelles ces estimateurs sont préférables aux autres options. Pour l'échantillonnage d'une unité par strate, nous montrons que ces estimateurs sont moins biaisés que l'estimateur sur « strate fusionnée » dans une gamme de scénarios. Pour l'échantillonnage systématique PPT, ces estimateurs donnent de bons résultats quand la population présente une périodicité importante, situation dans laquelle l'estimateur avec remise fréquemment utilisé peut présenter un biais négatif important.

5 Conclusion

Nous avons caractérisé de manière précise le biais de l'estimateur de variance de Horvitz-Thompson sous des conditions de non-mesurabilité et utilisé cette caractérisation pour

élaborer une correction prudente du biais. Ces estimateurs reflètent l'incertitude fondamentale inhérente aux plans non mesurables. Comparativement aux méthodes approximatives disponibles, ces estimateurs donnent parfois de meilleurs résultats et parfois de moins bons résultats dans une perspective pratique. Mais les méthodes approximatives peuvent présenter des biais dont on ne peut pas toujours évaluer la grandeur ni le signe. Les estimateurs élaborés dans le présent article fournissent par conséquent une mesure informative de la variabilité d'échantillonnage avec laquelle les analystes peuvent être d'accord, sans formuler d'hypothèses supplémentaires ni ressortir à des méthodes posant le risque d'introduire un biais négatif. Le terme de biais, A, possède une forme simple qui laisse penser qu'il serait possible de perfectionner les estimateurs élaborés ici, ce qui pourrait faire l'objet de futurs travaux de recherche.

Remerciements

Les auteurs remercient Joel Middleton, Allison Carnegie, un examinateur anonyme et le rédacteur associé de leurs commentaires constructifs.

Bibliographie

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory: Volume 1 Methods and Theory*. New York : John Wiley & Sons, Inc.

Horvitz, D., et Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 260, 663-685.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer.

Wolter, K. (2007). Introduction to Variance Estimation, Deuxième Édition. New York : Springer.

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

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 28, No. 4, 2012

Letter to the Editor
D. Salgado, C. Pérez-Arriero, M. Herrador, I. Arbués 471

Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection
James Wagner, Brady T. West, Nicole Kirgis, James M. Lepkowski,
William G. Axinn, Shonda Kruger Ndiaye 477

Editorial Note
Editors-in-Chief 501

Evaluating Survey Questions: A Comparison of Methods
Ting Yan, Frauke Kreuter, Roger Tourangeau 503

Discussion
Jennifer H. Madans, Paul C. Beatty 531

Discussion Evaluation Procedures for Survey Questions
Willem E. Saris 537

Rejoinder
Ting Yan, Frauke Kreuter, Roger Tourangeau 553

Disfluencies and Gaze Aversion in Unreliable Responses to Survey Questions
Michael F. Schober, Frederick G. Conrad, Wil Dijkstra, Yfke P. Ongena 555

Inferentially Valid, Partially Synthetic Data: Generating from Posterior Predictive Distributions not Necessary
Jerome P. Reiter, Satkartar K. Kinney 583

Confidentialising Exploratory Data Analysis Output in Remote Analysis
Christine M. O'Keefe 591

Editorial Collaborators 615
Index to Volume 28, 2012 621

All inquires about submissions and subscriptions should be directed to jos@scb.se

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 29, No. 1, 2013

Prelude to the Special Issue on Systems and Architectures for High-Quality Statistics Production 1

Consolidation and Standardization of Survey Operations at a Decentralized Federal Statistical Agency
Jack Nealon, Elvera Gleaton..... 5

The General Survey System Initiative at RTI International: An Integrated System for the Collection and Management of Survey Data
Lisa Thalji, Craig A. Hill, Susan Mitchell, R. Suresh, Howard Speizer, Daniel Pratt..... 29

Redesign of Statistics Production within an Architectural Framework: The Dutch Experience
Peter Struijs, Astrea Camstra, Robbert Renssen, Barteld Braaksma..... 49

Innovative Production Systems at Statistics New Zealand: Overcoming the Design and Build Bottleneck
Allyson Seyb, Ron McKenzie, Andrew Skerrett..... 73

Addressing Disclosure Concerns and Analysis Demands in a Real-Time Online Analytic System
Tom Krenzke, Jane F. Gentleman, Jianzhu Li, Chris Moriarity..... 99

A Potential Framework for Integration of Architecture and Methodology to Improve Statistical Production Systems
John L. Eltinge, Paul P. Biemer, Anders Holmberg 125

Discussion
Daniel Defays, Jean-Marc Museux 147
Alan F. Karr 157
Frauke Kreuter 165
Thomas Lumley 171
Svein Nordbotten 177
Stephen Penneck 187
Claude Poirier, Claude Julien, Robert McLellan 193
Siu-Ming Tam, Bill Gross 203
Dennis Trewin..... 213

In Other Journals 221

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 40, No. 4, December/décembre 2012

Changbao Wu
A special issue of CJS in honour of Mary Thompson 619

David Bellhouse
Mary Thompson.....623

Special Issue Articles

Erica E.M. Moodie, Bibhas Chakraborty and Michael S. Kramer
Q-learning for estimating optimal dynamic treatment rules from observational data.....629

Mingyuan Zhang and Dylan S. Small
Effect of vitamin A deficiency on respiratory infection: Causal inference for a discretely observed continuous time non-stationary Markov process 646

Hugh Chipman, Pritam Ranjan and Weiwei Wang
Sequential design for computer experiments with a flexible Bayesian additive model.....663

Sharon L. Lohr and J. Michael Brick
Blending domain estimates from two victimization surveys with possible bias.....679

Chris Skinner and Ben Mason
Weighting in the regression analysis of survey data with a cross-national application 697

Bruce G. Lindsay and Weixin Yao
Fisher information matrix: A tool for dimension reduction, projection pursuit, independent component analysis, and more 712

N. Reid
Likelihood inference in complex settings 731

Zilin Li, Sijian Wang and Xihong Lin
Variable selection and estimation in generalized linear models with the seamless L_0 penalty 745

Erratum

Parthasarathy Bagchi and Joseph B. Kadane
Erratum: Note of correction: “Laplace approximations to posterior moments and marginal distributions on circles, spheres, and cylinders” 770

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 41, No. 1, March/mars 2013

Jinhong You, Yong Zhou and Gemai Chen
Statistical inference for multivariate partially linear regression models 1

Lee Dicker and Xihong Lin
Parallelism, uniqueness, and large-sample asymptotics for the Dantzig selector 23

Irène Gijbels and Dominik Sznajder
Positive quadrant dependence testing and constrained copula estimation 36

Jean-François Quessy, Mériem Saïd and Anne-Catherine Favre
Multivariate Kendall's tau for change-point detection in copulas 65

Xiaoyi Min and Dongchu Sun
A matching prior based on the modified profile likelihood in a generalized Weibull stress-strength model 83

Piet Groeneboom, Geurt Jongbloed and Stefanie Michael
Consistency of maximum likelihood estimators in a large class of deconvolution models 98

Mariela Sued and Victor J. Yohai
Robust location estimation with missing data 111

Wenyu Jiang and Bingshu E. Chen
Estimating prediction error in microarray classification: Modifications of the 0.632+ bootstrap when $n < p$ 133

Zhiqiang Tan
A cluster-sample approach for Monte Carlo integration using multiple samplers 151

Hui Zhao, Yang Li and Jianguo Sun
Analyzing panel count data with a dependent observation process and a terminal event 174

Zhiming Li, Tianfang Zhang and Runchu Zhang
Three-level regular designs with general minimum lower-order confounding 192

Acknowledgement of referees' services/Remerciements aux membres des jurys 211