Nº 12-001-X au catalogue

Techniques d'enquête

Juin 2012

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 ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

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

Appels locaux ou internationaux :

Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

Programme des services de dépôt

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

Comment accéder à ce produit ou le commander

Le produit nº 12-000-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 ».

Statistique Canada
Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2012

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2012

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/copyright-droit-auteur-fra.htm).

Juin 2012

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

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

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-X)

Note de reconnaissance

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

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*

Rédacteur en chef délégué H. Mantel, *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, *Westat Inc*
D. Kasprzyk, *NORC at the University of Chicago*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistique Canada*
P. Lynn, *Univerity of Essex*
D.J. Malec, *National Center for Health Statistics*
G. Nathan, *Hebrew University*
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*
N. Schenker, *National Center for Health Statistics*
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, *Iowa State University*
C. Wu, *University of Waterloo*
W. Yung, *Statistique Canada*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints C. Bocci, K. Bosa, P. Dick, G. Dubreuil, S. Godbout, C. Leon, 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 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).

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 $ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 $ CA (6 $ × 2 exemplaires); autres pays, 20 $ CA (10 $ × 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 38, numéro 1, juin 2012

Table des matières

Articles réguliers

Sebastian Bredl, Peter Winker et Kerstin Kötschau
Une approche statistique pour déceler la falsification des données d'enquête par les intervieweurs 1

Steven Elliott
Application de la théorie des graphes à l'élaboration et à l'essai d'instruments d'enquête 13

G. Hussain Choudhry, J.N.K. Rao et Michael A. Hidiroglou
À propos de la répartition de l'échantillon pour une estimation sur domaine efficace 25

Ted Chang
Options de calage pour remplacer la poststratification de données croisées à deux dimensions 33

Paul Knottnerus et Arnout van Delden
À propos de la variance des variations estimées d'après des panels rotatifs et des strates dynamiques 45

Dan Liao et Richard Valliant
Facteurs d'inflation de la variance dans l'analyse des données d'enquêtes complexes 57

Hung-Mo Lin, Hae-Young Kim, John M. Williamson et Virginia M. Lesser
Estimation de coefficients d'accord d'après des données d'enquête 69

Jörg Drechsler et Jerome P. Reiter
Conjuguer des données synthétiques et le sous-échantillonnage pour créer des fichiers de microdonnées à grande diffusion pour les enquêtes à grande échelle 81

Balgobin Nandram et Myron Katzoff
Un modèle hiérarchique bayésien de non-réponse pour les données catégoriques d'un tableau à double entrée provenant de petits domaines avec incertitude au sujet de l'ignorabilité 89

Communications brèves

Phillip S. Kott
Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes 103

Autres revues 109

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.
∞

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.
∞

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 1-12
Statistique Canada, N° 12-001-X au catalogue

Une approche statistique pour déceler la falsification des données d'enquête par les intervieweurs

Sebastian Bredl, Peter Winker et Kerstin Kötschau [1]

Résumé

Les données d'enquête peuvent être falsifiées par les intervieweurs, la fabrication de données étant la forme de falsification la plus flagrante. Même un petit nombre d'interviews contrefaites peuvent fausser gravement les résultats d'analyses empiriques subséquentes. Outre l'exécution de réinterviews, certaines approches statistiques ont été proposées pour repérer ce genre de comportement frauduleux. À l'aide d'un petit ensemble de données, le présent article illustre comment la classification automatique, qui n'est ordinairement pas employée dans ce contexte, pourrait être utilisée pour repérer les intervieweurs qui falsifient les tâches qui leur sont assignées. Plusieurs indicateurs sont combinés pour classer les intervieweurs « à risque » en se fondant uniquement sur les données recueillies. Cette classification multivariée semble supérieure à l'utilisation d'un seul indicateur tel que la loi de Benford.

Mots clés : Fabrication de données ; falsificateur ; loi de Benford ; classification automatique.

1. Introduction

Lorsque la collecte des données se fait par interview, il convient de se préoccuper de la qualité des données. Cette dernière peut souffrir si le répondant fournit des réponses fausses ou imprécises ou que le questionnaire est mal conçu, ou encore si l'intervieweur s'écarte de la procédure d'interview établie. S'il le fait délibérément, on parle de « falsification des données par l'intervieweur » (Schreiner, Pennie et Newbrough 1988) ou de « tromperie » (Schräpler et Wagner 2003).

L'intervieweur peut falsifier les données de nombreuses façons (voir Guterbock 2008). Des formes assez subtiles de falsification consistent à interviewer le mauvais membre du ménage ou à mener l'enquête par téléphone quand l'interview devrait avoir lieu sur place. La forme de falsification la plus grave est la fabrication d'interviews complètes sans jamais prendre contact avec le ménage concerné. Dans notre analyse, nous traitons de ce dernier cas.

Les interviews contrefaites peuvent avoir de graves répercussions sur les statistiques produites d'après les données d'enquête. Schnell (1991), ainsi que Schräpler et Wagner (2003) donnent des preuves que l'effet sur les statistiques univariées pourrait être moins prononcé, à condition que la proportion de falsificateurs demeure suffisamment faible et que la « qualité » des données contrefaites soit élevée. En revanche, même une faible proportion d'interviews contrefaites peut suffire à introduire un biais important dans les statistiques multivariées. Schräpler et Wagner (2003) constatent que l'inclusion de données contrefaites provenant du panel socioéconomique de l'Allemagne (GSOEP) dans une régression multivariée réduit d'environ 80 % l'effet de la formation sur le logarithme de la rémunération brute, bien que la part d'interviews contrefaites soit inférieure à 2,5 %. Cet exemple montre qu'il est important de repérer ces interviews.

Le moyen le plus fréquent de déceler les intervieweurs qui falsifient les données consiste à procéder à une nouvelle interview (Biemer et Stokes 1989). Dans ce cas, un superviseur prend contact avec certains ménages qui auraient dû être interviewés afin de vérifier si l'intervieweur leur a effectivement rendu visite. Cependant, pour des raisons budgétaires, il est impossible de réinterviewer tous les ménages qui participent à une enquête (voir Forsman et Schreiner 1991). Par conséquent, il faut déterminer comment optimiser l'échantillon de nouvelles interviews de manière à déceler le mieux possible les falsificateurs. En général, il semble utile de sélectionner pour une nouvelle interview les ménages interrogés par un intervieweur qui, selon des caractéristiques associées aux réponses obtenues dans ces interviews, est plus susceptible que les autres de fabriquer des données. Dans ce contexte, Hood et Bushery (1997) utilisent le terme d'intervieweur « à risque ». Si la sélection des cas à réinterviewer se fait par échantillonnage en deux étapes, où les intervieweurs sont sélectionnés à la première étape et les enquêtés interrogés par ces intervieweurs, à la deuxième étape [comme le recommandent Forsman et Schreiner (1991)], les intervieweurs à risque pourraient être suréchantillonnés à la première étape.

Dans le présent article, nous montrons une approche purement statistique qui s'appuie sur les données contenues

1. Sebastian Bredl, Département de statistique et d'économétrie, Université Justus-Liebig, 35394 Gießen, Licher Straße 64, Allemagne. Courriel : sebastian.bredl@wirtschaft.uni-giessen.de ; Peter Winker, Département de statistique et d'économétrie, Université Justus-Liebig, 35394 Gießen, Licher Straße 64, Allemagne. Courriel : peter.winker@wirtschaft.uni-giessen.de ; Kerstin Kötschau, Hanse Parlament, 22587 Hamburg, Blankeneser Landstrasse 7, Allemagne. Courriel : kkoetschau@hanse-parlament.eu.

dans les questionnaires pour définir un groupe d'intervieweurs à risque. L'idée n'est pas nouvelle ; on trouve dans la littérature spécialisée plusieurs exemples de ce genre d'approche (Hood et Bushery 1997 ; Diekmann 2002 ; Turner, Gribbe, Al-Tayyip et Chromy 2002 ; Schräpler et Wagner 2003 ; Swanson, Cho et Eltinge 2003 ; Murphy, Baxter, Eyerman, Cunningham et Kennet 2004 ; Porras et English 2004 ; Schäfer, Schräpler, Müller et Wagner 2005 ; Li, Brick, Tran et Singer 2009). Cependant, à l'exception des travaux de Li et coll. (2009), les tests effectués dans ces études reposent sur l'examen d'indicateurs uniques dérivés des données des intervieweurs pour déceler les falsificateurs. Certaines études comprennent le calcul de plusieurs indicateurs, mais en les considérant tous séparément. Nous combinons plusieurs indicateurs dans des analyses par classification automatique, ce qui permet d'obtenir une meilleure classification des falsificateurs éventuels comparativement aux approches antérieures. Autant que nous sachions, cette procédure est une innovation dans le contexte du dépistage des intervieweurs qui falsifient les données, mais elle a déjà été employée dans d'autres domaines afin de déceler des comportements frauduleux. L'idée fondamentale est que les caractéristiques des « cas » frauduleux (ce qui définit un cas dépend du contexte) présentent, comparativement aux cas honnêtes, des schémas frappants qu'il est possible de révéler si ces caractéristiques sont considérées simultanément dans une classification automatique. Murad et Pinkas (1999) essayent de déceler la fraude dans l'industrie des télécommunications par classification des profils d'appels des clients. Un appel est caractérisé par plusieurs indicateurs, dont l'heure de l'appel ou la destination de celui-ci. Thiprungsri (2010) regroupe les demandes d'indemnisation en exécution d'un contrat d'assurance-vie collective soumises par les clients des compagnies d'assurance-vie en s'appuyant sur plusieurs caractéristiques des demandes. Celles qui forment de très petits groupes sont considérées comme étant suspectes. Donoho (2004) utilise la classification automatique, entre autres, pour dégager les tendances des marchés des options susceptibles de révéler des opérations d'initié.

Nous disposons d'un petit ensemble de données d'enquête (voir la sous-section 3.1 pour une description plus détaillée de cet ensemble) qui est constitué en partie de données falsifiées. Issu de 250 questionnaires administrés par 13 intervieweurs, l'ensemble de données est de taille assez limitée et la mesure dans laquelle nos résultats peuvent être généralisés à de plus grands ensembles de données n'est pas claire. Toutefois, cet ensemble de données nous permet d'illustrer notre approche. Le fait que nous sachions quelles données ont été recueillies honnêtement et lesquelles ont été contrefaites permet de procéder à une première évaluation de l'approche. Il convient toutefois de souligner que cette connaissance a priori n'est pas une condition préalable à l'emploi de la méthode.

La question de l'identification des intervieweurs à risque a été abordée durant les années 1980, mais la littérature traitant du sujet demeure peu abondante. En 1982, le U.S. Census Bureau a mis en œuvre l'Interviewer Falsification Study. En s'appuyant sur l'information recueillie dans le contexte de cette étude, Schreiner et coll. (1988) constatent que les intervieweurs ayant peu d'ancienneté sont plus susceptibles que les autres de fabriquer des données. Hood et Bushery (1997) utilisent plusieurs indicateurs pour découvrir les intervieweurs à risque dans la National Health Interview Survey (NHIS). Par exemple, ils calculent le taux par intervieweur de ménages désignés comme étant inadmissibles ou de ménages n'ayant pas de numéro de téléphone, et comparent ces taux aux données de recensement pour les régions pertinentes. En cas d'écart important, l'intervieweur est signalé et une nouvelle interview est effectuée. Les taux de détection parmi les intervieweurs signalés s'avèrent plus élevés que ceux observés pour des échantillons aléatoires de cas à réinterviewer. Turner et coll. (2002) constatent aussi, en examinant les données de la Baltimore STD and Behaviour Survey, que les intervieweurs qui fabriquent des données indiquent moins fréquemment des numéros de téléphone que les intervieweurs honnêtes. Dans le cas de l'interview assistée par ordinateur, Bushery, Reichert, Albright et Rossiter (1999), ainsi que Murphy et coll. (2004) proposent d'utiliser l'horodatage, c'est-à-dire l'enregistrement de l'heure et de la durée de l'interview par l'ordinateur, pour découvrir les intervieweurs suspects. Ceux qui ont besoin d'un temps anormalement long ou court pour administrer le questionnaire complet ou certains modules, ou ceux qui administrent un nombre étonnamment grand de questionnaires durant une période donnée pourraient alors être signalés comme des intervieweurs à risque. Schäfer et coll. (2005) supposent que les falsificateurs évitent les réponses extrêmes lorsqu'ils fabriquent les données. D'après des données du GSOEP, ces auteurs calculent la variance des réponses pour chaque question de tous les questionnaires administrés par un intervieweur et totalisent ces variances. Grâce à d'autres mécanismes de contrôle intégrés dans le GSOEP, les falsificateurs sont connus et il s'avère qu'on les retrouve parmi les intervieweurs ayant les variances globales les plus faibles. Porras et English (2004) adoptent une approche similaire et constatent aussi que les falsificateurs produisent des variances plus faibles que celles observées pour les questionnaires remplis honnêtement. Li et coll. (2009) combinent plusieurs indicateurs prédictifs dans un modèle de régression logistique dans lequel l'état de falsification connu d'un intervieweur sert de variable dépendante binaire. Les auteurs constatent que les échantillons de nouvelles

interviews dans lesquelles sont surpondérés les cas ayant une probabilité élevée d'être frauduleux selon le modèle de régression logistique donnent lieu à la détection d'un plus grand nombre de cas réels de fabrication de données que les échantillons sélectionnés purement au hasard. Cependant, il est évident qu'il faut disposer de données de réinterviews antérieures pour lesquelles l'état de falsification est connu pour exécuter la régression logistique.

Le nombre de combinaisons de réponses rares ou improbables dans les questionnaires soumis par un intervieweur (Murphy et coll. 2004 ; Porras et English 2004) et la comparaison de la composition ou des statistiques descriptives des ménages dans les questionnaires d'un intervieweur avec les données pour l'échantillon complet (Turner et coll. 2002 ; Murphy et coll. 2004) sont d'autres indicateurs discutés dans la littérature.

Un autre moyen de déceler les données contrefaites devenu populaire ces dernières années est l'utilisation de la loi de Benford (Schräpler et Wagner 2003 ; Swanson et coll. 2003 ; Porras et English 2004 ; Schäfer et coll. 2005), dont nous discuterons à la section 2, y compris le succès avec lequel elle a permis de détecter les interviews contrefaites dans le cadre d'études antérieures. La section 2 décrit aussi notre approche statistique en vue de dépister les falsificateurs. La section 3 présente les données sur lesquelles s'appuie notre analyse ainsi que nos résultats. Nous concluons l'article par une discussion de nos résultats.

2. Méthodes

2.1 Loi de Benford

Quand le physicien Frank Benford a remarqué que les pages des tables de logarithmes contenant les logarithmes des nombres faibles (1 et 2) étaient plus souvent utilisées que celles contenant les logarithmes de nombres plus élevés (8 et 9), il a commencé à étudier la distribution du premier chiffre d'une grande gamme de types de nombres, comme ceux figurant à la première page d'un journal, dans les adresses de voirie ou dans les poids moléculaires (Benford 1938). Benford a constaté que la distribution du premier chiffre non nul des nombres pouvait être décrite par la formule qui suit que l'on a appelée « loi de Benford » :

$$\text{Prob}(\text{premier chiffre} = d) = \log_{10}\left(1 + \frac{1}{d}\right). \qquad (1)$$

Toutefois, les séries de nombres étudiées par Benford (1938) ne semblaient pas se conformer toutes à cette loi. Par conséquent, la question qui se posait était de savoir quelle sorte de données on pouvait supposer produire des premiers chiffres de nombre obéissant à la loi. Des discussions de cette question peuvent être consultées dans Hill (1995), Nigrini (1996), Hill (1999), et Scott et Fasli (2001). La détection de la fraude financière est un domaine dans lequel l'application de la loi de Benford s'est beaucoup répandue durant la dernière décennie (Nigrini 1996 ; 1999 ; Saville 2006). Bien que les résultats de ces études ne soient pas pertinents dans notre contexte, il est intéressant de mentionner que le consensus dans la littérature semble être que l'on peut supposer que les valeurs monétaires suivent la loi de Benford. Swanson et coll. (2003) montrent que la distribution des premiers chiffres des nombres dans la Consumer Expenditure Survey des États-Unis est proche de la distribution de Benford.

La notion fondamentale qui sous-tend l'utilisation de la loi de Benford pour détecter les données contrefaites est que les falsificateurs ne connaissent vraisemblablement pas la loi ou qu'ils ne sont pas capables de fabriquer des données qui la suivent. Par conséquent, un écart important de la distribution des premiers chiffres par rapport à la distribution de Benford dans un ensemble de données indique que les données pourraient être contrefaites. Naturellement, il faut se demander si la nature des données permet de supposer qu'elles suivent la loi de Benford si elles sont authentiques. La loi de Benford ne peut pas être appliquée si les questionnaires ne contiennent que très peu de variables métriques, voire aucune.

Schräpler et Wagner (2003), ainsi que Schäfer et coll. (2005) utilisent la loi de Benford pour déceler la fabrication des données dans le GSOEP. Dans les deux études, tous les questionnaires administrés par chaque intervieweur sont combinés et vérifiés afin de déterminer si la distribution des premiers chiffres des nombres figurant dans les questionnaires s'écarte de manière significative de la loi de Benford. Cela peut se faire en calculant la statistique χ^2 :

$$\chi_i^2 = n_i \sum_{d=1}^{9} \frac{(h_{id} - h_{bd})^2}{h_{bd}} \qquad (2)$$

où n_i est le nombre de premiers chiffres dans l'ensemble des questionnaires provenant de l'intervieweur i, h_{id} est la proportion observée du premier chiffre d dans l'ensemble de premiers chiffres dans les questionnaires de l'intervieweur i et h_{bd} est la proportion du premier chiffre d dans l'ensemble des premiers chiffres sous la loi de Benford. Les valeurs élevées de χ^2 indiquent un écart par rapport à la distribution de Benford et signalent des intervieweurs à risque. Schräpler et Wagner (2003) utilisent différents types de variables continues dans leur analyse, tandis que Schäfer et coll. (2005) limitent la leur à des valeurs monétaires. Dans les deux études, les valeurs critiques de χ^2 sont supposées dépendre de la taille de l'échantillon n et sont par conséquent corrigées pour ce paramètre. Les résultats obtenus semblent prometteurs. L'ajustement de la distribution des premiers chiffres des nombres à la loi de Benford est en général nettement moins bon pour les questionnaires des

falsificateurs (qui étaient connus à l'avance) que pour ceux des intervieweurs honnêtes ; il semble donc approprié d'utiliser la loi de Benford comme moyen de dépistage des intervieweurs à risque.

Cependant, quand nous avons comparé les données des intervieweurs honnêtes figurant dans notre ensemble de données à la distribution de Benford, nous avons observé un écart important pour le chiffre 5. Cet écart pourrait être dû à l'arrondissement des nombres par les répondants. Le même problème est mentionné par Swanson et coll. (2003) et par Porras et English (2004), qui choisissent d'appliquer une autre approche « dans l'esprit de Benford » (Porras et English 2004, page 4224). Nous adoptons cette dernière approche qui consiste à comparer la distribution des premiers chiffres des nombres dans les questionnaires d'un intervieweur à la distribution des premiers chiffres des nombres dans l'ensemble des questionnaires sauf les siens. La valeur de χ^2 au niveau de l'intervieweur est calculée comme il est décrit plus haut, mais la proportion prévue d'un chiffre selon la loi de Benford h_{bd} est remplacée par la proportion du chiffre dans tous les autres questionnaires. Nous utilisons ensuite la valeur de χ^2 résultante comme indicateur dans la classification automatique.

Pour ce qui est de la sélection des variables dont nous examinons le premier chiffre des valeurs, nous suivons l'approche de Schäfer et coll. (2005) et n'incluons dans l'analyse que le premier chiffre des valeurs monétaires. L'enquête que nous utilisons pour les besoins de l'illustration contient des valeurs monétaires exprimées en devise locale pour les dépenses des ménages consacrées à divers articles, comme le loyer ou l'achat de terrains, de semences ou d'engrais, ou le paiement des impôts, et pour le revenu du ménage en provenance de diverses sources, comme le travail autonome agricole et non agricole et les transferts publics ou privés. Globalement, nous incluons le premier chiffre de 26 valeurs monétaires différentes par interview, en écartant les valeurs nulles déclarées. Ensuite, nous regroupons les premiers chiffres des nombres figurant dans tous les questionnaires livrés par un intervieweur et comparons leur distribution à celle obtenue pour toutes les autres interviews conformément à la méthode décrite plus haut. Le fait de se limiter aux valeurs monétaires constitue un critère catégorique durant le processus de sélection des données. En outre, comme nous l'avons mentionné plus haut, il est généralement reconnu que les données financières se prêtent à l'analyse fondée sur la loi de Benford. Il importe toutefois de mentionner que nous n'appuyons pas notre analyse sur la loi de Benford, mais sur une approche s'inspirant de cette loi.

2.2 Analyses multivariées

Notre idée consiste à combiner plusieurs indicateurs que nous dérivons directement des questionnaires remplis par chaque intervieweur et que nous supposons être différents pour les falsificateurs et les intervieweurs honnêtes. Nous recourons pour cela à la classification automatique et à l'analyse discriminante. Tous les indicateurs sont calculés au niveau de l'intervieweur. Cela signifie que nous regroupons pour l'analyse tous les questionnaires produits par un intervieweur, ce qui augmente la quantité de données sur laquelle est fondée la valeur de chaque indicateur unique. Les valeurs des indicateurs devraient donc être plus fiables et moins sensibles aux valeurs aberrantes. Par ailleurs, il est évident que le pouvoir discriminant des indicateurs au niveau des intervieweurs diminue lorsque ces derniers ne falsifient qu'une partie de leurs tâches. Par conséquent, l'examen des indicateurs au niveau du questionnaire semble préférable si la quantité de données par questionnaire est suffisamment grande.

La classification automatique constitue la méthode réelle de dépistage des intervieweurs à risque. Les intervieweurs sont répartis en deux groupes avec l'intention d'en obtenir un qui contient une part élevée de falsificateurs et l'autre, une part élevée d'intervieweurs honnêtes. La classification ne requiert pas d'information a priori sur les intervieweurs qui fabriquent des données et ceux qui ne le font pas. En fait, c'est ce que l'analyse est censée révéler. Puisque nous savons au départ à quel groupe appartient chaque intervieweur, nous pouvons découvrir si la classification automatique repère les « vrais falsificateurs » comme étant des intervieweurs à risque. Manifestement, l'hypothèse voulant que notre approche soit capable de faire parfaitement la distinction entre les deux groupes n'est pas réaliste. L'idée est plutôt d'obtenir un groupe d'intervieweurs à risque présentant une part élevée de falsificateurs comparativement à l'autre groupe. Si une nouvelle interview est possible, les efforts subséquents de réinterview seront axés sur les intervieweurs appartenant au groupe à risque.

Afin d'évaluer la performance de la classification automatique, nous considérons le nombre de falsificateurs non décelés, ainsi que le nombre de « fausses alarmes ». Les deux types d'« erreurs » ont un coût : les données des falsificateurs non décelés altèreront vraisemblablement les résultats des analyses statistiques subséquentes. Les fausses alarmes entraînent un coût en ce sens que l'on pourrait s'efforcer inutilement de réinterviewer les ménages concernés ou supprimer inutilement des données de l'échantillon. En outre, les intervieweurs honnêtes dont le travail fait l'objet de nouvelles interviews pourraient se démoraliser, particulièrement s'ils savent que c'est le travail des intervieweurs à risque qui est principalement visé. La façon de pondérer un falsificateur non détecté comparativement à une fausse alarme dans une fonction de perte est une question très subjective. En général, il paraît raisonnable d'accorder plus de poids au premier qu'à la seconde.

L'analyse discriminante requiert avant de l'exécuter que l'on sache pour chaque intervieweur s'il s'agit ou non d'un falsificateur. Par conséquent, ce n'est pas un instrument pour déceler les falsificateurs. Nous utilisons l'analyse discriminante pour vérifier nos hypothèses sur le comportement des falsificateurs, dont nous discuterons plus loin, et pour évaluer dans quelle mesure les indicateurs employés permettent de distinguer les deux groupes.

L'un de nos indicateurs est la valeur de χ^2, calculée en comparant la distribution des premiers chiffres des nombres figurant dans les questionnaires fournis par chaque intervieweur à la distribution correspondante dans tous les autres questionnaires, comme nous l'avons décrit à la sous-section précédente. En outre, nous calculons trois autres indicateurs en partant d'hypothèses concernant le comportement des falsificateurs de données. Schäfer et coll. (2005) supposent que les falsificateurs ont tendance à répondre à chaque question, donc à produire moins de valeurs manquantes. En outre, comme Porras et English (2004), ils s'attendent à ce que les falsificateurs choisissent des réponses moins extrêmes pour les questions ordinales. Hood et Bushery (1997) émettent l'hypothèse que les falsificateurs vont « essayer de ne pas compliquer les choses et de fabriquer un minimum de données falsifiées » (Hood et Bushery 1997, page 820).

Partant de ces hypothèses, nous calculons trois proportions, qui serviront de variables indicatrices dans les analyses multivariées, de même que la valeur de χ^2. Les trois variables indicatrices sont calculées comme il suit :

- Le « ratio de non-réponse partielle » est la proportion de questions qui demeurent sans réponse dans l'ensemble des questions. Nous nous attendons à ce que ce ratio soit plus faible pour les falsificateurs que pour les intervieweurs honnêtes.
- Le « ratio de réponses extrêmes » concerne les réponses mesurées sur une échelle ordinale. Le ratio indique la part de réponses extrêmes (la catégorie la plus faible ou la plus élevée sur l'échelle) dans toutes les réponses ordinales. Selon les hypothèses susmentionnées, ce ratio devrait être plus faible pour les falsificateurs.
- Le « ratio de réponses 'Autre' » concerne les questions qui, outre plusieurs réponses définies, offrent l'option « Autre » comme réponse possible. Le choix de cette option requiert la déclaration explicite d'une autre réponse. Si les falsificateurs ont tendance à ne pas compliquer les choses, nous pouvons nous attendre à ce qu'ils préfèrent les réponses définies à la déclaration d'une autre réponse. Donc, ce ratio également (calculé comme étant la proportion de réponses « Autre » dans l'ensemble des réponses pour lesquelles l'option « Autre » peut être choisie) devrait être plus faible pour les falsificateurs.

Naturellement, la liste des variables indicatrices susceptibles d'être incluses dans la classification automatique peut être allongée. En général, il est possible de dériver de nombreuses autres variables de ce genre des hypothèses sur le comportement des intervieweurs qui fabriquent des données ou d'utiliser celles qui ont déjà été proposées dans la littérature, quoique dans d'autres contextes que celui de la classification automatique. Par exemple, en se fondant sur l'hypothèse que les falsificateurs essayent de fabriquer un minimum de données falsifiées, Hood et Bushery (1997) s'attendent à ce qu'ils sélectionnent souvent de manière disproportionnée la réponse « Non » aux questions, qui selon la réponse, mènent à un ensemble de nouvelles questions ou les évitent (en supposant que « Non » est généralement la réponse qui évite les questions supplémentaires). On pourrait donc calculer le ratio de réponses « Non » à ce genre de questions et l'utiliser comme variable dans l'analyse par classification automatique. Nous ne nous servons pas de ce ratio, car deux versions légèrement différentes du questionnaire ont été utilisées dans notre exemple empirique. Il n'y a qu'un faible nombre de questions menant ou non à de nouvelles questions selon la réponse, le même dans les deux versions du questionnaire.

En outre, quand l'interview assistée par ordinateur permet d'utiliser les données d'horodatage comme il est discuté dans Bushery et coll. (1999), le temps moyen nécessaire pour mener une interview ou le nombre d'interviews réalisées en une journée peuvent servir d'indicateur. Les enquêtes par panel fournissent certains renseignements supplémentaires pour construire des indicateurs. Stokes et Jones (1989) proposent de comparer le taux réel de membres d'un ménage non appariés dans les questionnaires d'un intervieweur particulier au taux prévu de non-appariement calculé conditionnellement à plusieurs caractéristiques du ménage. Les auteurs emploient cette procédure dans l'enquête postcensitaire qui est menée à titre d'enquête de suivi du recensement des États-Unis. Si le taux réel de non-appariement dépasse fortement le taux prévu, les auteurs considèrent qu'il s'agit d'un indicateur de données contrefaites. Habituellement, cette approche peut être appliquée aussitôt que l'on dispose des données d'au moins deux vagues d'une enquête par panel.

Il devient évident que les premières étapes de notre approche consistent à examiner la structure du questionnaire et d'autres types de données, comme les données d'horodatage recueillies durant les opérations d'enquête. Ensuite, il convient de déterminer les indicateurs susceptibles d'être différents pour les falsificateurs et pour les intervieweurs honnêtes que l'on peut dériver de ces sources. Une autre

approche consiste à recourir à des techniques d'exploration de données (*data mining*) pour dégager les tendances qui sont fréquentes dans les données contrefaites ou celles qui ne sont pas les mêmes dans les données contrefaites que dans les données recueillies honnêtement (Murphy, Eyerman, McCue, Hottinger et Kennet 2005). S'il est possible de les déceler, de telles tendances pourraient servir d'indicateurs à la place de ceux dérivés d'après des hypothèses quant au comportement des falsificateurs. Toutefois, cette approche nécessite un énorme ensemble de données comprenant des cas connus de falsification afin de procéder à l'exploration des données. Ce genre d'ensemble des données n'est pas toujours disponible.

3. Résultats

3.1 Sources des données

Les données sur lesquelles porte la présente étude proviennent d'enquêtes-ménages menées en novembre 2007 et en février 2008 dans un pays de la Communauté des États indépendants (CEI) (ancienne Union soviétique). L'enquête a été réalisée dans le cadre d'un projet de recherche international sur les réformes agraires et la pauvreté rurale. Nous avions l'intention d'interviewer 200 ménages dans quatre villages en 2007. Après avoir déterminé que toutes les interviews avaient été contrefaites dans le premier village étudié, nous avons interrompu l'enquête et lancé un nouveau cycle avec de nouveaux intervieweurs dans d'autres villages en février 2008. Tous les villages ont été sélectionnés en s'appuyant sur des critères qualitatifs tels que la structure de la production agricole et la mise en œuvre de réformes agraires. Dans chaque village, les ménages ont été échantillonnés aléatoirement en se servant de listes de ménages fournies par les maires des villages. Non seulement cette procédure assurait que tous les ménages soient sélectionnés au hasard, mais elle fournissait aussi le fondement pour des réinterviews, puisque tous les ménages étaient définis exactement. Cependant, ces nouvelles interviews n'avaient pas été planifiées au tout début de l'enquête. Comme les ménages possédaient rarement un téléphone, les appels de vérification n'étaient pas possibles et les nouvelles interviews de ces ménages ont nécessité des déplacements dans le village pour procéder à la réinterview sur place, ce qui a entraîné d'importantes dépenses d'argent et de temps. Cinq intervieweurs ont été recrutés au moment de la première enquête en 2007. Deux d'entre eux étaient les partenaires locaux du projet de recherche. Ils avaient participé à l'élaboration du questionnaire et étaient responsables de la coordination des enquêtes dans leur pays. Les trois autres intervieweurs étaient des étudiants engagés par les partenaires. Le questionnaire comprenait diverses sections concernant les caractéristiques du ménage, la richesse en ressources, ainsi que le revenu et les dépenses. La plupart des questions étaient de type fermé. Quelques-unes seulement comprenaient une échelle. Des données sur des variables métriques ont été recueillies pour les dépenses des ménages, comme le loyer ou l'achat de terrains, de semences ou d'engrais, ou les impôts, ainsi que pour le revenu du ménage provenant de diverses sources, comme le travail autonome agricole et non agricole, et les transferts publics et privés.

Au moment où les interviews de l'enquête de 2007 ont été menées, aucun des chercheurs allemands n'était présent dans les villages. Les questionnaires ont été recueillis immédiatement après la réalisation de l'enquête dans le premier village. Un premier examen de ces questionnaires a suscité des soupçons, parce que le papier des questionnaires paraissait très propre et très blanc. Le papier ne présentait pas de salissures ni de coins cornés. En comparant les réponses figurant sur divers questionnaires soumis par un intervieweur, nous avons découvert deux questionnaires contenant des réponses identiques. Étant donné que nous avions demandé d'indiquer le montant du revenu en provenance de diverses sources en valeur métrique, il était fort peu probable que les réponses sur les deux questionnaires soient identiques. Comme nous n'obtenions aucune explication de la part des partenaires du projet, nous avons réinterviewé sur place un sous-échantillon de 10 % de l'échantillon original. Aucun des ménages réinterviewés n'a déclaré qu'il avait déjà été interviewé. Après que nous ayons détecté la fabrication des interviews, les partenaires ont reconnu que toutes les interviews avaient été contrefaites. Nous avons naturellement cessé de travailler avec tous les intervieweurs et partenaires, et créé un nouveau groupe de recherche local.

En février 2008, l'enquête a été répétée dans le même pays. Comme nous l'avons mentionné plus haut, nous avons sélectionné de nouveaux villages et ménages conformément aux critères susmentionnés. Nous avons recruté neuf étudiants pour les interviews et organisé une supervision sur place durant l'enquête. Dans la plupart des cas, les interviews ont eu lieu dans une école ou à l'hôtel de ville afin que nous puissions surveiller tous les intervieweurs. Quand les interviews ont eu lieu au domicile des familles participant à l'enquête, nous avons assisté à certaines d'entre elles. Étant donné cette procédure, nous présumons que les réponses au questionnaire de l'enquête de 2008 n'ont pas été contrefaites.

Le présent article porte sur un total de 250 interviews de ménages réalisées par 13 intervieweurs, dont quatre étaient des falsificateurs de l'enquête de 2007 (les interviews soumises par l'un d'eux ont été exclues, car il n'a remis que trois questionnaires) qui ont définitivement contrefait les

résultats et qui sont désignés par F1 à F4, et neuf intervieweurs censés être honnêtes, désignés par H1 à H9. Le tableau 1 donne un aperçu du nombre de questionnaires par intervieweur qui ont été inclus dans l'analyse.

Tableau 1
Nombre de questionnaires par intervieweur

Intervieweur	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Nombre de questionnaires	10	12	10	10	22	23	23	24	23	23	23	23	24

3.2 Classification automatique

À la présente sous-section, nous présentons les résultats de notre analyse par classification automatique. D'après ces résultats, nous évaluons dans quelle mesure notre procédure permet de détecter les intervieweurs qui fabriquent des données. Comme nous l'avons déjà mentionné, nous utilisons quatre variables indicatrices dans la classification automatique, à savoir le ratio de non-réponse partielle, la proportion de réponses extrêmes sur échelle ordinale parmi l'ensemble des réponses sur échelle ordinale que nous appelons ratio de réponses extrêmes, la proportion de réponses pour lesquelles l'option « Autre » nécessitant une autre réponse a été choisie parmi l'ensemble des réponses offrant cette option (appelée ratio de réponses « Autre ») et la valeur de χ^2 découlant de la comparaison de la distribution du premier chiffre dans les nombres figurant dans les questionnaires d'un intervieweur à la distribution correspondante pour l'ensemble des autres questionnaires.

Le tableau 2 donne les valeurs des quatre variables indicatrices incluses dans la classification automatique pour les 13 intervieweurs. Il montre que le ratio de non-réponse partielle et le ratio de réponses « Autre » sont manifestement plus faibles pour les quatre falsificateurs que pour les intervieweurs honnêtes. F1 et F4 n'ont pas choisi l'option « Autre » du tout. Pour le ratio de réponses extrêmes, les résultats paraissent moins clairs. Toutes les valeurs sont comprises entre 40 % et 70 %, sauf celles pour l'intervieweur F1, qui est clairement plus faible. Les valeurs de χ^2 sont assez élevées pour les falsificateurs F2 et F4. Les valeurs pour les deux autres diffèrent peu de celles observées pour les intervieweurs honnêtes.

L'idée générale de la classification automatique est de déterminer des sous-groupes d'éléments dans un espace d'éléments qui sont tous caractérisés par des mesures multivariées (voir Härdle et Simar (2007) pour une introduction à la classification automatique). À la première étape, il faut choisir une mesure pour évaluer la distance ou la ressemblance entre les éléments. À la deuxième étape, les éléments sont affectés à divers sous-groupes ou classes. Les éléments d'une classe particulière doivent se ressembler du point de vue de la mesure choisie, tandis que les éléments appartenant à des classes différentes doivent être distincts. Une grande variété de méthodes sont applicables pour affecter les éléments aux classes, le nombre de classes pouvant être fixe ou déterminé par la méthode de classification.

Tableau 2
Valeurs des variables incluses dans la classification automatique pour chaque intervieweur (toutes les valeurs sauf celles de χ^2 sont exprimées en pourcentage)

Intervieweur	Non-réponse partielle	Réponses « Autre »	Réponses extrêmes	Valeur de χ^2
F1	1,36	0,00	28,33	19,63
F2	0,71	0,65	40,85	29,70
F3	0,68	2,33	56,90	11,34
F4	0,51	0,00	58,62	27,33
H1	3,85	18,01	65,12	14,48
H2	1,99	2,40	59,42	6,91
H3	3,10	9,47	70,07	15,49
H4	4,52	13,04	56,43	16,61
H5	1,18	4,48	70,07	12,16
H6	3,46	1,37	50,75	15,42
H7	2,51	12,72	45,65	9,11
H8	1,77	10,95	69,85	3,63
H9	0,14	1,61	69,44	19,14

Nous avons choisi comme mesure de proximité le carré de la distance euclidienne et employé plusieurs méthodes de classification pour vérifier la robustesse des résultats. Dans tous les cas, les intervieweurs ont été classés en deux groupes avec l'intention d'obtenir un « groupe de falsificateurs » et un « groupe d'intervieweurs honnêtes ». L'avantage de cette approche est que l'on obtient une classification nette. Par contre, lorsque l'on examine séparément les variables indicatrices, il n'est pas évident de savoir où tirer la ligne qui sépare les falsificateurs des interviews honnêtes. Avant de procéder à la classification automatique, nous avons transformé toutes les variables en variables centrées réduites de moyenne nulle et de variance égale à l'unité. Les effets d'échelle sont ainsi éliminés, car les distances sont mesurées en écarts-types et non en différentes unités.

La première méthode de classification que nous avons utilisée est la classification hiérarchique. Il s'agit d'une procédure classique qui peut aussi être appliquée à de plus grands ensembles de données et qui est implémentée dans les progiciels statistiques classiques. La classification hiérarchique consiste à fusionner les classes pas à pas, en combinant les deux classes les plus proches. Au début, chaque élément est considéré comme constituant une classe distincte. Nous mesurons la distance entre deux classes comme étant le carré de la distance euclidienne entre toutes les paires d'éléments possibles, où le premier élément de la paire provient d'une classe et le deuxième, de l'autre classe. Nous avons utilisé le progiciel STATA avec l'option « average linkage » pour effectuer la classification hiérarchique.

Dans la classification hiérarchique, deux éléments demeurent dans la même classe une fois qu'ils ont été appariés. Donc, la procédure n'aboutit pas forcément à un optimum global par rapport à une mesure de distance

donnée. Dans notre cas, le nombre relativement faible d'intervieweurs nous permet de réaliser une autre analyse en examinant simplement toutes les compositions de classes possibles et de choisir la meilleure en fonction d'une certaine fonction cible. (L'analyse a été exécutée en MATLAB et le code du programme peut être obtenu sur demande.) Cette procédure est clairement supérieure à la classification hiérarchique, car elle fait en sorte que soit déterminée la composition globalement optimale des classes. Cependant, nous donnons aussi les résultats de la classification hiérarchique, car elle est assez pratique comparativement à l'approche consistant à essayer toutes les compositions possibles qui requiert d'importants calculs quand le nombre d'intervieweurs augmente. À la place, on pourrait recourir à des techniques d'optimisation heuristiques.

Pour examiner toutes les compositions de classes possibles, nous employons deux fonctions cibles. La première combine deux notions, à savoir qu'une grande distance entre deux centres de classe est admissible, de même qu'une petite distance entre les éléments d'une classe et le centre de la classe. Nous recherchons la composition des classes qui maximise l'expression suivante :

$$\frac{\sum_{i=1}^{4}(\bar{d}_{1i} - \bar{d}_{2i})^2}{\sum_{j=1}^{n_1}\sum_{i=1}^{4}(d_{ij} - \bar{d}_{1i})^2 + \sum_{j=n_1+1}^{13}\sum_{i=1}^{4}(d_{ij} - \bar{d}_{2i})^2}. \qquad (3)$$

L'indice i représente les quatre variables indicatrices différentes $\bar{d}_{ai}$ où $a = 1, 2$ est la moyenne de la variable i dans la classe a, j symbolise les différents éléments (intervieweurs) dans la classe 1 et dans la classe 2, d_{ij} est la valeur de la variable i pour l'élément j, et n_1 est le nombre d'éléments dans la classe 1. Donc, le numérateur mesure la distance entre les deux classes, et le dénominateur, la distance à l'intérieur des classes, et la distance est mesurée par le carré de la distance euclidienne.

Il serait également intéressant de voir quelle composition optimale des classes on obtiendrait si, au lieu de maximiser l'équation (3), on minimisait la moyenne du carré de la distance euclidienne entre toutes les paires possibles dans une classe. En fait, cette idée ressemble fort à la fonction cible pertinente dans les procédures de classification hiérarchique présentées plus haut. Notre deuxième mesure de distance, qui cette fois doit être minimisée, est calculée comme il suit :

$$\frac{\sum_{j=1}^{n_1-1}\sum_{k=j+1}^{n_1}\mathrm{CDE}_{jk} + \sum_{j=n_1+1}^{13-1}\sum_{k=j+1}^{13-1}\mathrm{CDE}_{jk}}{(n_1(n_1-1))/2 + ((13-n_1)(13-n_1-1))/2} \qquad (4)$$

où CDE_{jk} est le carré de la distance euclidienne entre les éléments j et k, calculée comme $\mathrm{CDE}_{jk} = \sum_{i=1}^{4}(d_{ij} - d_{ik})^2$. Le numérateur est égal à la somme des distances entre toutes les paires possibles d'éléments dans la même classe. En divisant cette somme par le nombre de paires possibles, on obtient la distance intra-classe moyenne.

Le tableau 3 donne les résultats des trois méthodes de classification. Dans l'analyse hiérarchique avec lien entre groupes, les trois falsificateurs F1, F2 et F4 forment la classe 1, et le falsificateur F3 ainsi que tous les intervieweurs honnêtes, la classe 2. Donc, nous arrivons à séparer les deux groupes d'intervieweurs, à l'exception d'un falsificateur. Cependant, si nous ne savions pas au départ quels intervieweurs ont fabriqué des données et lesquels sont honnêtes, nous devrions décider laquelle des deux classes contient les intervieweurs à risque. Cela peut se faire en comparant les moyennes des variables indicatrices dans chaque classe présentées au tableau 4. Pour la classification hiérarchique, les moyennes pour le ratio de non-réponse partielle et pour le ratio de réponses « Autre » sont manifestement plus faibles dans la classe 1. Il en est de même de la moyenne pour le ratio de réponses extrêmes, quoique la différence entre les deux classes soit moins frappante. Enfin, une moyenne plus élevée de la valeur de χ^2 peut être observée pour la classe 1. Étant donné ces résultats, en nous en tenant aux hypothèses susmentionnées concernant le comportement des falsificateurs, nous déterminerions correctement que la classe 1 est celle qui contient tous les intervieweurs à risque. Nous avons également essayé d'améliorer les résultats de la classification hiérarchique en utilisant les moyennes de classe présentées au tableau 4 comme point de départ pour la classification par la méthode des K-moyennes. Cependant, l'application de l'algorithme des K-moyennes n'a donné lieu à aucune modification de la composition des classes.

La composition des classes qui maximise l'équation (3) est identique à celle obtenue en utilisant la classification hiérarchique. Par conséquent, comme le montre le tableau 4, les moyennes des indicateurs dans les deux classes sont également identiques.

La composition des classes qui minimise l'équation (4) est légèrement différente. La classe 1 contient maintenant tous les falsificateurs et un intervieweur honnête. Les moyennes des variables indicatrices indiquent de nouveau clairement que la classe 1 est celle qui contient les intervieweurs à risque. Ce résultat est très satisfaisant. Tous les falsificateurs sont repérés et une seule fausse alarme est produite. Il ne faut toutefois pas perdre de vue que cela ne signifie pas que cette méthode de classification particulière est celle qui donne les meilleurs résultats lorsqu'elle est appliquée à un autre ensemble de données.

Afin de déterminer dans quelle mesure un plus grand nombre d'indicateurs produit de meilleurs résultats, nous avons répété notre approche de classification fondée sur les équations 3 et 4 en nous servant de toutes les combinaisons

possibles d'indicateurs, y compris les cas ne s'appuyant que sur un seul indicateur. Les résultats (voir le tableau 7 en annexe) indiquent généralement que l'augmentation du nombre d'indicateurs améliore les résultats. Toutefois, il existe aussi des combinaisons comptant un plus petit nombre d'indicateurs qui produisent des résultats semblables à ceux fondés sur les quatre indicateurs pris ensemble. Pour déterminer quelle combinaison d'indicateurs est la meilleure, il faudrait s'appuyer sur l'établissement hautement subjectif du coût relatif de la non-détection d'un falsificateur comparativement à celui d'une fausse alarme. Toutefois, on peut déterminer quelles combinaisons d'indicateurs sont non Pareto-dominées en ce sens qu'il n'existe aucune autre combinaison présentant moins de falsificateurs non décelés (fausses alarmes) et en même temps ne présentant pas plus de fausses alarmes (falsificateurs non décelés). La combinaison d'indicateurs constituée des quatre indicateurs est la seule qui est non Pareto-dominée, quelle que soit l'équation utilisée. Par contre, les combinaisons ne comprenant qu'un seul indicateur sont Pareto-dominées dans six cas sur huit.

Tableau 3
Résultats des trois méthodes de classification employées

Classification hiérarchique													
Intervieweur	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Classe	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance entre les classes divisée par la distance intra-classe													
Intervieweur	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Classe	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance entre les éléments d'une classe													
Intervieweur	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Classe	1	1	1	1	2	2	2	2	2	2	2	2	1

Tableau 4
Moyennes des variables indicatrices par classe pour les trois compositions des classes

	Non-réponse partielle		Réponses « Autre »		Réponses extrêmes		Valeur de χ^2	
Classification hiérarchique								
Classe	1	2	1	2	1	2	1	2
Moyenne	0,86	2,32	0,22	7,64	42,60	61,37	25,55	12,43
Distance entre les classes divisée par la distance intra-classe								
Classe	1	2	1	2	1	2	1	2
Moyenne	0,86	2,32	0,22	7,64	42,60	61,37	25,55	12,43
Distance entre les éléments d'une classe								
Classe	1	2	1	2	1	2	1	2
Moyenne	0,68	2,80	0,92	9,06	50,83	60,92	21,43	11,73

3.3 Analyse discriminante

Enfin, nous nous tournons vers l'analyse discriminante pour vérifier si les hypothèses concernant le comportement des falsificateurs sur lesquelles est fondée notre classification automatique sont valides. L'analyse discriminante peut être appliquée si les classes sont connues afin d'évaluer dans quelle mesure les indicateurs employés dans l'analyse permettent de bien séparer les divers groupes et si l'appartenance à un groupe peut être prédite correctement [voir Härdle et Simar (2007) pour une introduction à l'analyse discriminante]. Dans une analyse discriminante linéaire, les coefficients b_0 et b_i de la fonction discriminante $D = b_0 + \sum_{i=1}^{n} b_i x_i$ sont déterminés de façon telle qu'ils maximisent une fonction qui augmente avec l'écart entre les valeurs D moyennes des deux groupes distincts et simultanément diminue avec les écarts entre les valeurs D des éléments à l'intérieur des groupes. Dans notre cas, les x_i sont nos quatre variables indicatrices et nous obtenons deux groupes en séparant les falsificateurs et les intervieweurs honnêtes.

Nous utilisons les probabilités a priori correspondant à la taille de groupe relative (4/13 et 9/13) afin de prédire l'appartenance à un groupe. Le tableau 5 donne les résultats. Manifestement, les quatre variables permettent d'obtenir une bonne séparation des falsificateurs et des intervieweurs honnêtes, car l'appartenance à un groupe est prédite correctement dans tous les cas sauf un.

Comme le montre le tableau 5, les valeurs négatives de la fonction discriminante sont associées au groupe de falsificateurs. Par conséquent, le tableau 6 indique que les signes de trois des quatre coefficients sont en harmonie avec le comportement prévu des falsificateurs. Des ratios de non-réponse partielle et de réponses extrêmes plus élevés mènent à une plus forte probabilité d'observer un intervieweur honnête, de même qu'une valeur de χ^2 plus faible. Le coefficient estimé pour les autres ratios est négatif. Donc, une augmentation des autres ratios, toutes choses étant égales par ailleurs, augmente la probabilité qu'un intervieweur ait fabriqué des données. Cela pourrait sembler contredire nos hypothèses susmentionnées. Une explication possible serait que l'effet des autres ratios est déjà reflété par le ratio de non-réponse partielle. En fait, le coefficient de corrélation entre les deux variables est assez élevé, sa valeur étant de 0,71. Le lambda de Wilks de l'analyse discriminante est statistiquement significatif au seuil de signification de 5 %.

Tableau 5
Résultats de l'analyse discriminante par intervieweur

Intervieweur	Groupe prédit	Groupe réel	Fonction discriminante
F1	1	1	-2,878
F2	1	1	-3,376
F3	2	1	-0,541
F4	1	1	-1,955
H1	2	2	1,828
H2	2	2	1,060
H3	2	2	1,747
H4	2	2	1,616
H5	2	2	0,706
H6	2	2	0,777
H7	2	2	-0,041
H8	2	2	1,765
H9	2	2	-0,710

Tableau 6
Coefficients estimés standardisés et non standardisés (analyse discriminante)

Variable	Coefficient (non standardisé)	Coefficient (standardisé)
Non-réponse partielle	0,767	0,917
Réponses « Autre »	-0,025	-0,129
Réponses extrêmes	0,075	0,821
Valeur de χ^2	-0,092	-0,562
Constante	-4,250	–
Lambda de Wilks (Prob > F)	0.0254	

4. Conclusion

Les données d'enquête peuvent être affectées par les intervieweurs qui fabriquent des données. La fabrication de données est un problème qu'il ne faut pas négliger, car il peut causer des biais importants. Même une petite quantité de données contrefaites peut altérer gravement les résultats des analyses empiriques ultérieures. Nous étendons les approches antérieures en vue de repérer les intervieweurs à risque en combinant plusieurs indicateurs dérivés directement des données d'enquête par classification automatique. Afin de démontrer notre approche, nous l'appliquons à un petit ensemble de données qui a été fabriqué en partie par des falsificateurs. Le fait que nous sachions dès le départ quels sont les falsificateurs nous permet d'évaluer les résultats de la classification automatique et de procéder par après à une analyse discriminante pour révéler dans quelle mesure les deux groupes d'intervieweurs peuvent être bien séparés au moyen des variables indicatrices. Des classifications automatiques de divers types sont effectuées. Toutes donnent lieu à la détermination d'une classe d'intervieweurs à risque, le ratio de non-réponse partielle et le ratio de réponses « Autre » étant les deux indicateurs les plus clairs. Nous n'arrivons pas à identifier parfaitement les falsificateurs. Cependant, dans tous les cas, la classe des intervieweurs à risque contient une part nettement plus élevée de falsificateurs que la deuxième classe. Les avantages de la classification automatique tient au fait que l'on obtient une classification nette des intervieweurs qui sont à risque et des autres intervieweurs, ce qui n'est pas le cas quand des indicateurs tels que la valeur de χ^2 sont examinés individuellement. En outre, elle nous permet de combiner l'information provenant de plusieurs indicateurs. En étudiant la performance de tous les sous-ensembles possibles d'indicateurs, nous constatons qu'en général, un grand nombre d'indicateurs permet de mieux repérer les falsificateurs. Le fait que diverses méthodes de classification produisent des résultats différents ne devrait pas être nécessairement considéré comme un défaut de notre approche. Selon la pondération choisie du coût d'un falsificateur non détecté comparativement à celui d'une fausse alarme, on pourrait en dernière analyse n'affecter au groupe des falsificateurs éventuels que les intervieweurs qui se retrouvent systématiquement dans la classe des intervieweurs à risque, quelle que soit la méthode de classification appliquée (ce qui impliquerait un coût élevé des fausses alarmes), affecter au groupe des falsificateurs éventuels tous les intervieweurs qui se retrouvent dans la classe des intervieweurs à risque au moins une fois (ce qui impliquerait un coût élevé des falsificateurs non détectés) ou choisir une solution intermédiaire.

L'application à un petit ensemble de données démontre un autre mérite de notre approche : elle a été testée et a donné de bons résultats dans une situation dans laquelle le nombre de questionnaires par intervieweur était assez limité (trois des falsificateurs n'ont soumis que dix questionnaires). Si un petit nombre de questionnaires par intervieweur est suffisant pour effectuer l'analyse, on pourrait aussi imaginer de la mettre en œuvre durant la période principale de travail sur le terrain, quand les intervieweurs n'ont soumis qu'un certain pourcentage de leurs questionnaires. Les falsificateurs pourraient alors être remplacés par d'autres intervieweurs qui mèneront l'enquête auprès des unités qui auraient dû être interviewées par les falsificateurs.

Évidemment, en examinant nos résultats, nous ne devons pas perdre de vue que nous avons appliqué notre méthode à un ensemble de données dans lequel une forme très grave de fabrication des données a eu lieu : d'une part, nous avons des falsificateurs qui ont contrefait les données de tous leurs questionnaires (presque) complètement et d'autre part, nous avons des intervieweurs qui (on le présume) ont effectué tout leur travail honnêtement, ce qui facilite la discrimination entre les intervieweurs honnêtes et les intervieweurs malhonnêtes. En outre, la taille de notre échantillon, qui ne comprend que 13 intervieweurs, est assez limitée. Il serait intéressant d'explorer l'utilité de notre approche lorsqu'on l'applique à de plus grands ensembles de données, étant donné que la part d'interviews falsifiées dans les grandes enquêtes s'avère plus faible que dans notre cas. Qui plus est, les grands ensembles de données pourraient permettre de construire des indicateurs additionnels pour la classification automatique. Si l'enquête comprend un programme de réinterview, il serait possible d'évaluer l'utilité de notre approche en comparant le « succès » d'une réinterview aléatoire avec celui d'une réinterview axée sur les intervieweurs considérés comme étant à risque. Nous avons également l'intention de poursuivre l'analyse dans des conditions expérimentales. Des conditions appropriées permettent de s'assurer que l'on obtient un ensemble de données qui a été recueilli partiellement en menant des interviews réelles et partiellement fabriquées en disant à certains participants à l'expérience de remplir le questionnaire eux-mêmes.

Annexe

Tableau 7
Résultats des classifications automatiques fondées sur les équations 3 et 4 pour toutes les combinaisons possibles de classes

Indicateurs				Équation 3		Équation 4	
Non-réponse partielle	Réponses « Autre »	Réponses extrêmes	Valeur de χ^2	Falsificateurs non décelés	Fausses alarmes	Falsificateurs non décelés	Fausses alarmes
			X	2	0	1	1
		X		2	1	2	2
		X	X	2	0	1[1]	0
	X			0[1]	4	0	4
	X		X	2	0	0	2
	X	X		3	0	0	3
	X	X	X	1[1]	0	1	1
X				0[1]	4	0	4
X			X	2	1	0	2
X		X		3	0	-[2]	-
X		X	X	1[1]	0	1	1
X	X			0[1]	4	0	4
X	X		X	1	1	0	2
X	X	X		0[1]	4	0	4
X	X	X	X	1[1]	0	0[1]	1

[1] Combinaison d'indicateurs non Pareto-dominée.
[2] Les valeurs moyennes de classe n'ont pas permis de déterminer la classe « à risque ».

Remerciements

Nous tenons à souligner l'appui financier de la Deutsche Forschungsgemeinschaft par la voie du projet « PP 1292: *Survey Methodology* ». Nous remercions en outre John Bushery et quatre examinateurs anonymes de leurs commentaires constructifs concernant notre article.

Bibliographie

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(1), 551-572.

Biemer, P., et Stokes, S. (1989). The optimal design quality control sample to detect interviewer cheating. *Journal of Official Statistics*, 5(1), 23-29.

Bushery, J., Reichert, J., Albright, K. et Rossiter, J. (1999). Using date and time stamps to detect interviewer falsification. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 316-320.

Diekmann, A. (2002). Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung. Rapport technique manuscript 06/2002, Institut für Technikfolgenabschätzung (ITA), Wien.

Donoho, S. (2004). Early detection of insider trading in option markets. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 420-429.

Eyerman, J., Murphy, J., McCue, C., Hottinger, C. et Kennet, J. (2005). Dépistage de la falsification des données par l'intervieweur par l'exploration de données. Dans le *Recueil : Symposium 2005, Défis méthodologiques reliés aux besoins futurs d'information*. Statistique Canada.

Forsman, G., et Schreiner, I. (1991). The design and analysis of reinterview: An overview. Dans *Measurement Errors in Surveys*, (Éds., P.B. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz et S. Sudman), New York : John Wiley & Sons, Inc, 279-301.

Guterbock, T.M. (2008). Falsification. Dans *Encyclopedia of Survey Research Methods*, (Éd., P.J. Lavrakas), Sage Publications, Thousand Oaks, 1, 267-270.

Härdle, W., et Simar, L. (2007). *Applied Multivariate Statistical Analysis*, 2e Édition. Springer, Berlin.

Hill, T. (1995). A statistical derivation of the significant digit law. *Statistical Science*, 10(4), 354-363.

Hill, T. (1999). The difficulty of faking data. *Chance*, 26, 8-13.

Hood, C., et Bushery, M. (1997). Getting more bang from the reinterviewer buck: Identifying 'At risk' interviewers. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 820-824.

Li, J., Brick, J., Tran, B. et Singer, P. (2009). Using statistical models for sample design of a reinterview program. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 4681-4695.

Murad, U., et Pinkas, G. (1999). Unsupervised Profiling for Identifying Superimposed Fraud. Lecture Notes in Computer Science, 1704, 251-261.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D. et Kennet, J. (2004). A system for detecting interviewer falsification. Article présenté à l'American Association for Public Opinion Research 59th Annual Conference.

Nigrini, M. (1996). A taxpayers compliance application of Benford's law. *Journal of the American Taxation Association*, 18, 72-91.

Nigrini, M. (1999). I've got your Number. *Journal of Accountancy*, 187(5), 79-83.

Porras, J., et English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 4223-4228.

Saville, A. (2006). Using Benford's law to predict data error and fraud - An examination of companies listed on the JSE Securities Exchange. *South African Journal of Economic and Management Sciences*, 9(3), 341-354.

Schäfer, C., Schräpler, J., Müller, K. et Wagner, G. (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. *Schmollers Jahrbuch*, 125, 183-193.

Schnell, R. (1991). Der einfluss gefälschter Interviews auf survey ergebnisse. *Zeitschrift für Soziologie*, 20(1), 25-35.

Schräpler, J., et Wagner, G. (2003). Identification, Characteristics and Impact of Faked Interviews in Surveys - An analysis by means of genuine fakes in the raw data of SOEP. Document de discussion IZA séries, 969.

Schreiner, I., Pennie, K. et Newbrough, J. (1988). Interviewer falsification in census bureau surveys. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 491-496.

Scott, P., et Fasli, M. (2001). Benford's law: An empirical investigation and a novel explanation. Rapport technique de la CSM, Department of Computer Science, University Essex.

Stokes, L., et Jones, P. (1989). Evaluation of the interviewer quality control procedure for the post-enumeration survey. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 696-698.

Swanson, D., Cho, M. et Eltinge, J. (2003). Detecting possibly fraudulent data or error-prone survey data using Benford's law. Dans *Proceedings of the Survey Research Method Section*, American Statistical Association, 4172-4177.

Thiprungsri, S. (2010). Cluster Analysis for Anomaly Detection in Accounting Data. Collected Papers of the Nineteenth Annual Strategic and Emerging Technologies Research Workshop San Francisco, California.

Turner, C., Gribbe, J., Al-Tayyip, A. et Chromy, J. (2002). Falsification in Epidemiologic Surveys: Detection and Remediation (Ébauche de prépublication). Papier technique sur l'Health and Behavior Measurement. Washington DC : Research Triangle Institute. No. 53.

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 13-23
Statistique Canada, N° 12-001-X au catalogue

Application de la théorie des graphes à l'élaboration et à l'essai d'instruments d'enquête

Steven Elliott [1]

Résumé

La présente étude porte sur l'application de la théorie des graphes à l'élaboration et à l'essai d'instruments d'enquête. Une approche fondée sur la théorie des graphes présente plusieurs avantages par rapport aux approches conventionnelles au chapitre de la structure et des caractéristiques du système de spécifications des instruments de recherche, en particulier les instruments assistés par ordinateur et de vaste portée. La possibilité de vérifier la connectivité de tous les éléments constitue l'un de ces avantages, de même que la capacité de simuler un instrument. Une telle approche permet aussi de produire des mesures servant à décrire l'instrument, par exemple le nombre de chemins et de parcours. Le concept de la « base » est abordé dans le contexte des essais des logiciels. Ce terme désigne le plus petit ensemble de parcours, parmi tous ceux que comporte un instrument, qui couvre tous les appariements de liens et de nœuds. Ces parcours peuvent être utilisés à titre d'ensemble complet et économique de cas types pour faire l'essai des instruments.

Mots clés : Théorie des graphes ; interview assistée par ordinateur (IAO) ; élaboration de questionnaires ; essai de logiciel ; essai portant sur la base ; cas types.

1. Introduction

La théorie des graphes est une théorie mathématique relative aux ensembles de nœuds et de liens. Le terme « graphe » désigne la représentation visuelle d'un tel ensemble. Les graphes ont été utilisés dans de nombreux domaines d'étude pour modéliser des phénomènes de la réalité. Les premiers exemples ont trait à l'analyse de la logistique des transports (Berge 1976, page VII). Dans le cadre de telles analyses, une approche fondée sur la théorie des graphes sera utile pour déterminer entre autres les parcours offrant une efficience maximale par rapport à différents endroits. Dans le graphe, ces endroits sont représentés par les nœuds, tandis que les liens représentent les chemins allant d'un endroit à l'autre.

Il existe aussi des applications de la théorie des graphes dans le domaine de la méthodologie d'enquête. Si l'on représente les questions d'un questionnaire d'enquête sous forme de nœuds et les chemins engendrés par les enchaînements de questions sous forme de liens, il devient possible de recourir à un graphe pour modéliser un questionnaire. Dès lors, bon nombre des théorèmes et des mesures descriptives associés à la théorie des graphes peuvent s'appliquer au regard des questionnaires. De plus, une approche fondée sur la théorie des graphes sera utile dans l'optique des processus de documentation et d'essai d'instruments d'enquête. Par exemple, un système de documentation contenant un tableau pour les questions et un autre pour les options de réponse permet de vérifier la connectivité de tous les éléments des instruments ainsi que de procéder à des simulations d'instruments fonctionnels. Une procédure d'essai où l'ensemble de cas types couvre à tout le moins la « base » d'un graphe d'instrument garantit que toutes les combinaisons de liens et de nœuds font l'objet d'un essai à partir du plus petit nombre possible de cas.

Aux fins de l'élaboration, de la documentation et de l'essai de la plupart des instruments d'enquête, il n'est pas nécessaire de recourir à une représentation fondée sur la théorie des graphes. En effet, la plupart du temps, les instruments d'enquête comportent un nombre relativement peu élevé de questions et ne comptent pas beaucoup de points d'embranchement. Pensons notamment aux enquêtes sur la satisfaction de la clientèle et aux courtes enquêtes avec papier et crayon, comme le recensement américain. Pour ce genre d'instruments, les procédures conventionnelles de documentation et d'essai sont appropriées. Par contre, une approche reposant sur la théorie des graphes pourrait avoir sa place dans le cas des enquêtes complexes et de grande envergure, comme il y en a beaucoup à l'heure actuelle. Par exemple, en 2009, on a procédé à l'Enquête canadienne sur les capacités financières (ECCF) pour déterminer les connaissances et les comportements des Canadiens lorsqu'il est question de la prise de décisions financières ; les interviews téléphoniques assistées par ordinateur ont porté sur douze sections dont chacune comptait environ douze questions (Statistique Canada 2010). Mentionnons encore les interviews trimestrielles (IPAO) de la Consumer Expenditure Survey (2010), enquête menée par le Bureau of Labor Statistics du département américain du Travail. Cette enquête compte 22 sections, dont la plupart comptent au moins trois sous-sections ; le nombre de questions par sous-section va de six à 90 (US Bureau of Labor Statistics 2010). Dans l'un et l'autre cas, une approche fondée sur la théorie des graphes serait tout à fait envisageable aux fins de documentation et d'essai.

1. Steven Elliott, Westat Incorporated, 1500 Research Blvd., Rockville, MD 20850-3158, États-Unis. Courriel : sdelliott2@verizon.net.

Le présent document porte sur l'application de la théorie des graphes aux instruments de recherche et d'enquête. La section 2 propose la description d'un questionnaire sous forme de graphe et délimite les priorités particulières qui distinguent un graphe de questionnaire des autres types de graphes. La section 3 énonce les conséquences d'une représentation fondée sur la théorie des graphes de la structure des bases de données associées aux systèmes de documentation et de spécifications des enquêtes assistées par ordinateur. Les caractéristiques des structures de données selon la théorie des graphes sont commentées à la section 4. Les sections 5 et 6 portent sur les essais de logiciels et les conséquences associées à la théorie des graphes dans l'optique des essais. Un exposé est fait des raisons justifiant l'utilisation d'une « base » constituée d'un ensemble de cas types englobant toutes les paires de nœuds unis par des liens. Les parcours que comporte cette base constituent un ensemble complet de cas types aux fins de procéder à l'essai d'instruments.

2. Envisager le questionnaire comme un graphe

Il est possible de représenter un graphe de la façon suivante : $G = (V, E)$, où $V = \{v_1, v_2, v_3, ..., v_n\}$ est un ensemble de nœuds ou de sommets, et $E = \{(v_i, v_j), (v_1, v_k), ...\}$ est un ensemble de liens, ou de relations, entre paires de sommets. Dans le contexte de la théorie des graphes, on utilise le terme « arêtes » pour désigner les liens, qui sont représentés au moyen du symbole E (Chartrand 1985, page 27). Un graphe n'a pas à comporter d'autres caractéristiques particulières. Cela dit, un graphe auquel on incorpore des caractéristiques spéciales sera utile pour modéliser de nombreux phénomènes en sciences et en génie. À titre d'exemple, les graphes comportant des arêtes non orientées (c'est-à-dire où les deux nœuds liés entre eux peuvent l'un comme l'autre être le prédécesseur ou le successeur) peuvent servir à modéliser des circuits électriques à courant alternatif, tandis que les graphes dont les arêtes sont orientées pourront se prêter à la modélisation des problèmes entourant la conception des circuits de circulation. D'autres encore seront utilisés pour modéliser des réseaux dans des domaines comme l'informatique, les communications, la sociologie et la psychologie.

Dans le cas des questionnaires d'enquête, les nœuds du graphe représentent différents éléments de l'instrument d'enquête. Le plus souvent, il s'agira des questions de fond de l'enquête ou des points de décision qui servent à déterminer l'enchaînement des questions. Les arêtes représentent pour leur part les options de réponse ou les résultats associés à un nœud. Elles représentent également le cheminement d'un nœud à l'autre, chaque arête comportant un nœud prédécesseur et un nœud successeur uniques. Le graphe présenté à la figure 1 représente un instrument d'enquête simple comportant douze questions. Les cercles noirs (c'est-à-dire les nœuds) correspondent aux éléments de l'instrument, et les lignes qui relient ces cercles sont les arêtes unissant une question donnée à une autre question. Par exemple, le premier nœud peut représenter une question à laquelle deux réponses sont possibles, comme « oui » et « non ». Le deuxième nœud peut être une question avec cinq réponses possibles, les trois premières réponses établissant un lien avec le nœud 3 et les deux dernières, avec le nœud 4.

Le graphe qui est utilisé pour représenter un questionnaire se voit attribuer différentes propriétés particulières. Ces propriétés définissent la nature logique du questionnaire. Bethlehem et Hundepool (2004) ont exposé un certain nombre de ces propriétés. En premier lieu, un questionnaire comporte un nœud qui marque le début et un autre qui marque la fin. En deuxième lieu, tous les nœuds, autres que celui du début et celui de la fin, sont liés, c'est-à-dire que, pour chaque nœud du graphe, il y a au moins un chemin vers le nœud du début et au moins un vers le nœud de la fin. En troisième lieu, toutes les arêtes sont orientées. Cela veut dire que le chemin découlant de l'enchaînement d'un nœud à l'autre va toujours dans la même direction. En quatrième lieu, il peut exister plus d'une arête entre une paire donnée de nœuds. Dans de nombreux types de graphes, une paire de nœuds donnée sera reliée par une seule arête. Cette restriction ne s'applique toutefois pas dans le cas d'un graphe de questionnaire parce qu'un questionnaire comportera souvent plus d'une option de réponse déterminant l'enchaînement d'une question à une autre. Enfin, en cinquième lieu, il peut y avoir des structures en boucle, de sorte qu'un nœud pourra figurer plus d'une fois à l'intérieur d'un même chemin. Les structures en boucle sont fréquemment utilisées dans les questionnaires pour modifier les réponses dont on a déterminé qu'elles étaient incorrectes. Par exemple, les questions d'ordre financier ou portant sur l'utilisation du temps peuvent être vérifiées au moyen de contrôles donnant lieu à un retour en arrière si la somme des questions composantes ne correspond pas au bon total.

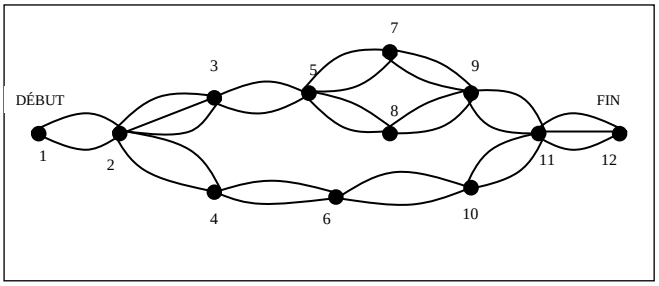


Figure 1 Représentation d'un instrument d'enquête sous forme de graphe

Les caractéristiques d'un graphe de questionnaire peuvent être résumées comme suit :

1. Présence d'un nœud marquant le début et d'un autre marquant la fin.
2. Connectivité (chaque nœud est connecté aux nœuds de début et de fin).
3. Toutes les arêtes sont orientées.
4. Les paires de nœuds peuvent être connectées par des arêtes multiples ou parallèles.
5. Les nœuds peuvent figurer plus d'une fois à l'intérieur d'un même chemin.

À partir d'un ensemble de propriétés définitionnelles, il est possible d'établir différents descripteurs, dont le nombre de chemins et une base. Il est également possible de modéliser un système de documentation d'après la structure du graphe, ainsi qu'on le verra à la section suivante.

3. Systèmes de documentation et de spécifications des questionnaires d'enquête

Les systèmes de documentation de questionnaires entrent généralement dans deux catégories selon qu'il s'agit d'un document-texte ou d'une base de données relationnelle. Dans le cas d'un système qui consiste en un document-texte, l'information relative à une question de fond ou à une autre catégorie d'éléments de l'instrument sera le plus souvent présentée sous forme de section comportant le texte de la question, les options de réponse, les enchaînements et des instructions à l'intention des programmeurs. Le système de documentation n'a en soi aucune fonctionnalité, si ce n'est les fonctions de recherche et d'impression que comporte le logiciel de traitement de texte ayant servi à rédiger la documentation. Pour leur part, les systèmes reposant sur une base de données relationnelle sont habituellement organisés sous forme de tableau dont les lignes représentent les questions de l'enquête et les colonnes, les attributs des questions. Chaque enregistrement du tableau correspond à un n-tuple des attributs des questions. Par exemple, les attributs d'une question peuvent comprendre l'intitulé, le numéro d'ordre, le texte de la question, les options de réponse, l'information sur les enchaînements et les notes techniques pertinentes. Le *Tool for the Analysis and Documentation of Electronic Questionnaires* (TADEQ) (Bethlehem et Hundepool 2004) est un système de spécifications correspondant à cette description. On peut aussi mentionner les systèmes conçus à Westat Inc. pour le *Medicare Current Beneficiary Survey* (MCBS), commandité par les US Centers for Medicare and Medicaid Services (*Medicare Current Beneficiary Survey*: *Overview* 2010), ou encore le *Medical Expenditure Panel Survey* (MEPS), commandité par le US Department of Health and Human Services (*MEPS*: *Survey Instruments and Associated Documentation* 2010). Ces systèmes de base de données ont en commun une structure comportant un tableau primaire où chaque enregistrement représente une question.

Malgré les avantages découlant de la nature simple des systèmes conventionnels, un système de spécifications modélisé sous forme de graphe présentera de plus grandes capacités. Avant de décrire ces capacités et la structure sous-jacente requise, il convient de souligner qu'il existe beaucoup de façons d'élaborer une structure de données selon la théorie des graphes (on se reportera à ce sujet à Gibbons (1985, page 73), qui a décrit et catégorisé certaines de ces structures). Le système proposé ici consiste en une structure relationnelle (liste) comportant deux tableaux principaux. Le premier tableau représente les nœuds du graphe et le second, les arêtes. Dans le tableau des nœuds, chaque enregistrement (ligne) représente un élément donné de l'instrument (question d'enquête, contrôle, point de décision déterminant l'enchaînement). Dans le second tableau, chaque enregistrement représente une arête donnée (option de réponse ou condition particulière applicable à un point de décision). Chaque enregistrement faisant partie de l'un des tableaux comporte des attributs qui lui sont propres. Les différents attributs sont énoncés dans les colonnes du tableau. Dans le tableau des nœuds, chaque colonne représente un attribut particulier, par exemple l'identificateur et la catégorie de l'élément. Dans le tableau des arêtes, chaque colonne représente un attribut, comme le texte d'une option de réponse. Il faut mentionner deux distinctions importantes entre un système de documentation reposant sur cette structure et un système de documentation plus conventionnel : 1) l'information relative aux arêtes n'est pas contenue dans le tableau des éléments de l'instrument ; 2) le tableau des arêtes (liens) contient des identificateurs concernant le prédécesseur et le successeur d'une arête. Ainsi que cela est exposé à la section suivante, ces distinctions font en sorte qu'un tel système de documentation se prête à des utilisations que ne permet pas un système conventionnel.

4. Caractéristiques d'un système de spécifications fondé sur la théorie des graphes

L'utilisation de tableaux distincts pour les nœuds et pour les liens à titre d'éléments de base d'un système de spécifications présente plusieurs avantages, en particulier la possibilité de simuler une interview. Le concepteur ou l'évaluateur peut sélectionner les options de réponse de l'instrument au fil de l'enchaînement des éléments de ce dernier comme s'il faisait remplir l'instrument par un répondant. La figure 2 constitue un exemple d'écran de lancement de la simulation d'un instrument. L'élément qui marque le début de la simulation est choisi à partir de cet

écran. La figure 3 représente l'écran de simulation proprement dit. L'élément pertinent est affiché au centre de l'écran, ainsi que le texte de la question ou la condition applicable. La partie inférieure gauche de l'écran présente l'ensemble des éléments à partir desquels on peut avoir fait la transition vers l'élément affiché (prédécesseurs). On parlera de « points d'origine » dans l'écran. La partie inférieure droite est réservée aux « points de destination », c'est-à-dire les éléments vers lesquels il est possible d'aller à partir de l'élément affiché (successeurs). Il est donc possible de cheminer dans un instrument à la fois dans l'une ou l'autre direction, un élément à la fois, en choisissant soit un point d'origine, soit un point de destination. Le questionnaire utilisé à titre d'exemple dans les figures 2 et 3 a trait aux connaissances générales sur le cancer, et la question exposée à la figure 4 compte un seul prédécesseur et un seul successeur. Ce sera le cas de la plupart des questions d'enquête ; si toutefois il existait de multiples prédécesseurs ou successeurs, l'écran en afficherait la liste.

Il est possible de simuler l'application d'un instrument d'enquête parce qu'un tableau distinct est utilisé pour les liens. Ce tableau peut se prêter à des recherches afin de trouver tous les prédécesseurs et successeurs de n'importe quel élément du questionnaire. Au cours de la phase de conception, cette caractéristique peut servir à s'assurer que chacune des sections et des questions est connectée comme il se doit et que tous les enchaînements sont exacts. Lors de la phase d'essai, la même caractéristique peut servir à effectuer une comparaison en parallèle d'un instrument et des spécifications sur lesquelles il repose. L'évaluateur peut recourir au système de spécifications pour simuler l'instrument sur un moniteur tout en exécutant l'instrument sur un second moniteur. Une telle comparaison peut servir à vérifier la formulation et la présentation des questions ainsi que les options de réponse ; elle permet aussi de vérifier que l'instrument assure la transition vers la question appropriée au moment approprié. Les rapports d'erreurs ou de problèmes peuvent dès lors être intégrés directement au système de spécifications à titre d'attributs d'un élément de l'instrument.

Une autre méthode d'évaluation de l'intégrité d'un questionnaire consiste à déterminer les éléments « orphelins » de l'instrument. En effet, il peut arriver lors de la création ou de la modification d'un questionnaire qu'un élément devienne inaccessible – ce que l'on désignera au moyen du terme « orphelin ». Étant donné qu'il existe un tableau des liens (options de réponse et conditions), il est possible de lancer des recherches à partir de ce tableau afin de déterminer si une question donnée constitue le successeur d'un lien quelconque. La question qui ne succède à aucun lien est un « orphelin ». On trouvera à la figure 4 l'écran affichant la liste des éléments classés selon la fréquence à laquelle ils constituent des successeurs (ce que l'on a appelé le « rapport sur les orphelins »). On peut voir que la première question de l'enquête n'a pas de point d'origine, ce qui est normal puisque le premier élément ne peut avoir de prédécesseur. Tout autre élément dépourvu de tout point d'origine sera un orphelin. Le rapport sur les orphelins sert également à décrire les éléments de l'instrument. Par exemple, une question ou un élément comportant un grand nombre de points d'origine peut être la première question d'une section prévue dans les cas d'interruption prématurée. Une telle section sera accessible à partir de n'importe quelle autre section que comporte l'interview, de sorte qu'elle aura un grand nombre de prédécesseurs.

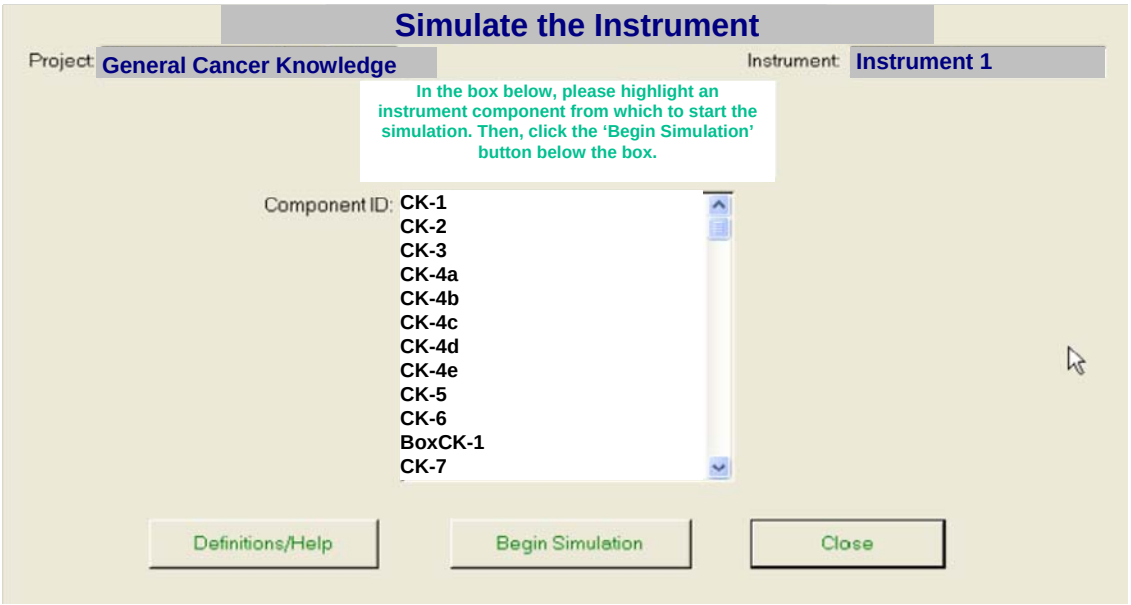


Figure 2 Écran de lancement de la simulation

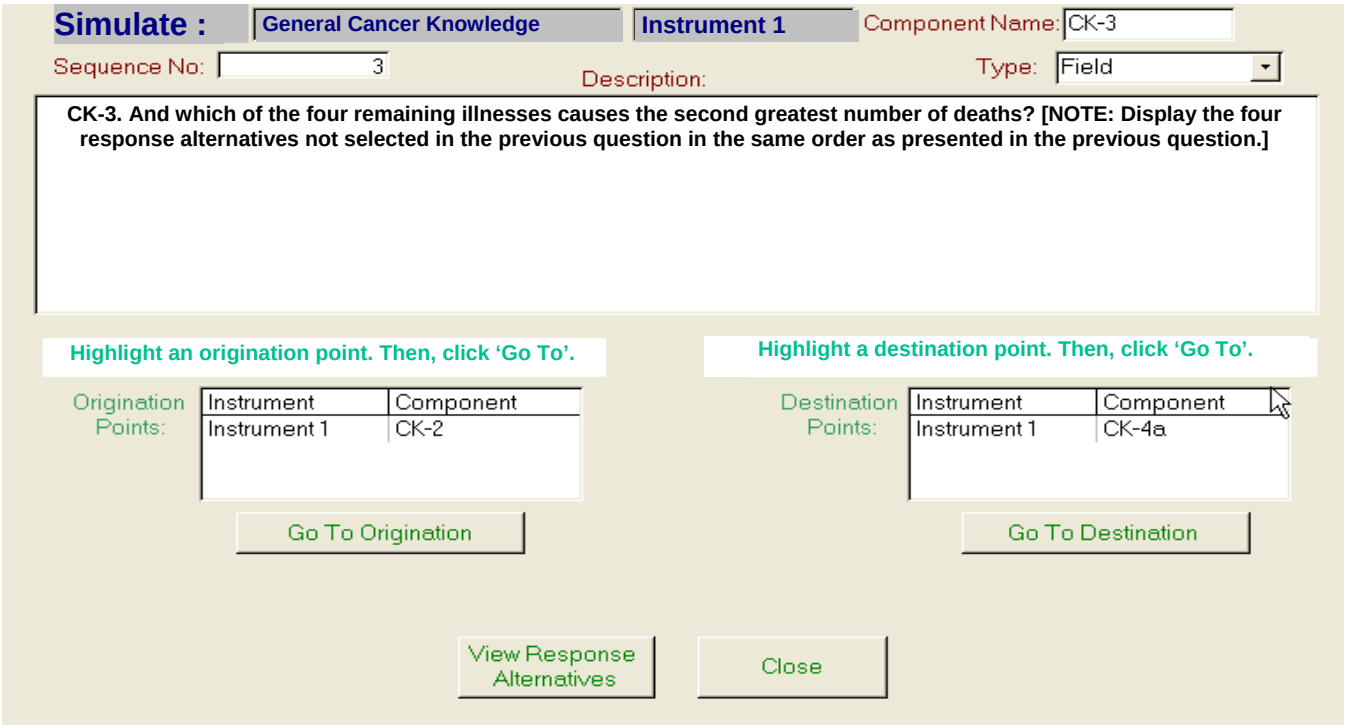


Figure 3 Écran de simulation

5. Essai

Le processus d'essai d'un instrument d'enquête assisté par ordinateur consiste à vérifier que l'exécution de l'instrument est conforme aux spécifications conceptuelles. Plusieurs approches ont été utilisées à cette fin. L'une d'entre elles consiste à tester d'abord les éléments de base du système, puis de passer à des ensembles de plus en plus grands et intégrés d'éléments (essai ascendant). On parle d'« essai d'unités » dans le cas des essais portant sur les éléments de base (Beizer 1995, page 5). Une fois que chaque élément de base a fait l'objet d'un essai exclusif, les éléments sont assemblés et les essais portent sur leurs interactions. C'est ce que l'on appelle l'« essai d'intégration » (Hetzel 1984, page 11), dont l'étape finale est l'essai du système dans son ensemble, en conformité avec l'utilisation qui en serait faite dans un environnement de production (Myers 1979, page 110).

D'autres approches et d'autres termes ont été employés en ce qui touche les procédures d'essai, par exemple les méthodes de la boîte noire et de la boîte blanche ainsi que les essais de régression. Dans la méthode de la boîte noire, un programme est considéré comme étant une boîte noire, c'est-à-dire que son fonctionnement interne ne peut être observé. Les intrants et les extrants sont les seuls aspects observables du fonctionnement du programme (Beizer 1995, page 8). Dans la méthode de la boîte blanche, on se fonde sur la connaissance du code du programme pour décider de la manière de procéder aux essais et des cas qui feront l'objet des essais (Patton 2006, page 55). Par exemple, un programmeur peut mener un ensemble d'essais selon la méthode de la boîte blanche, de manière à exécuter chaque ligne de code (« couverture du code ») ou chaque point d'embranchement (« couverture des points d'embranchement »). Les essais de régression ont pour objet de garantir l'intégrité du code lorsque des ajouts ou des changements sont apportés à un programme opérationnel (Beizer 1995, page 235). Les essais de régression portent sur un ensemble de cas types choisis de manière à assurer l'exécution de chacun des principaux embranchements du programme. D'autres types d'essais sont utilisés dans le cadre de l'élaboration de logiciels (versions alpha et bêta, convivialité), et de nombreuses sources fournissent une description plus complète des procédures d'essai (se reporter à Kaner, Falk et Nguyen 1999, page 277).

Dans toute procédure d'essai, l'une des principales préoccupations aura trait au biais d'essai. Cela survient lorsque certains éléments ou certaines fonctionnalités d'un instrument sont exclus des essais, ce qui peut arriver par exemple dans le cas de questions posées vers la fin d'une enquête ou faisant partie d'une section plus obscure. Ce biais est entièrement éliminé si l'on sélectionne un ensemble de cas types de manière à inclure l'ensemble des éléments de l'instrument, des liens entre ces éléments et des aspects fonctionnels. Cela dit, compte tenu de la longueur et de la complexité de certaines enquêtes, il n'est pas judicieux dans

la pratique de tenir des essais exhaustifs. Prenons l'exemple du questionnaire représenté à la figure 1. Ce questionnaire compte seulement douze questions et 28 options de réponse, ce qui donne tout de même 672 chemins possibles. Dans le cas d'enquêtes plus longues, il pourrait y avoir bien au-delà de 10 000 chemins. Dès lors, à partir du moment où une couverture exhaustive n'est pas une solution viable pour les enquêtes de grande envergure, le biais d'essai peut être évité au moyen d'un échantillonnage probabiliste de cas types possibles. Une approche fondée sur la théorie des graphes pourra être utile dans l'une et l'autre spécifications de l'univers des cas types ainsi qu'aux fins de déterminer une approche rationnelle d'échantillonnage de cas types.

6. Approche d'essai fondée sur la théorie des graphes

Il existe plusieurs moyens de définir un univers d'éléments d'essai. On peut par exemple utiliser les cas types, dont il a déjà été question, chaque cas type étant une interview fictive. Il pourrait aussi s'agir de questions d'enquête ou d'options de réponse, ou encore d'une combinaison quelconque des unes et des autres. Nos commentaires porteront uniquement sur les cas types, de sorte qu'il est utile de donner une définition précise des cas types et de deux termes étroitement liés, soit le « parcours » et le « chemin ».

Un parcours se compose d'un ensemble unique et ordonné de nœuds allant du début à la fin de l'instrument. Chaque nœud faisant partie d'un parcours donné, exception faite d'un nœud de début ou de fin, est lié à la fois à un prédécesseur et à un successeur (cette définition concorde avec les travaux de Bethlehem et Hundepool 2004). Un parcours unique existe chaque fois qu'un élément compte plus d'un successeur. À la figure 1, il existe de multiples successeurs dans le cas des éléments 2 et 4. Ces deux points d'embranchement engendrent trois parcours :

Parcours 1 - 1, 2, 3, 5, 7, 9, 11, 12
Parcours 2 - 1, 2, 3, 5, 8, 9, 11, 12
Parcours 3 - 1, 2, 4, 6, 10, 11, 12

Un chemin constitue pour sa part un ensemble unique de nœuds et de liens alternés à partir du nœud du début jusqu'au nœud de la fin. Tout comme un parcours, un chemin doit satisfaire aux conditions de connectivité et d'orientation. En théorie des graphes, ce terme est synonyme de cas type ou encore de scénario d'essai dans le domaine des essais de logiciels. Étant donné que le chemin englobe le lien unissant une paire de nœuds, le nombre de chemins à l'intérieur d'un graphe sera supérieur ou égal au nombre de parcours. Le nombre de chemins que contient un parcours donné est égal au produit du nombre de liens entre chaque paire de nœuds que compte le parcours. Ainsi, à la figure 1, le nombre de chemins pour chaque parcours est le suivant :

Parcours 1 - 2 x 3 x 2 x 2 x 2 x 2 x 3 = 288
Parcours 2 - 2 x 3 x 2 x 2 x 2 x 2 x 3 = 288
Parcours 3 - 2 x 2 x 2 x 2 x 2 x 3 = 96

Le nombre total de chemins correspond à la somme des chemins associés à tous les parcours (dans notre exemple, 288 + 288 + 96 = 672). On peut effectuer ce calcul au moyen de la formule suivante :

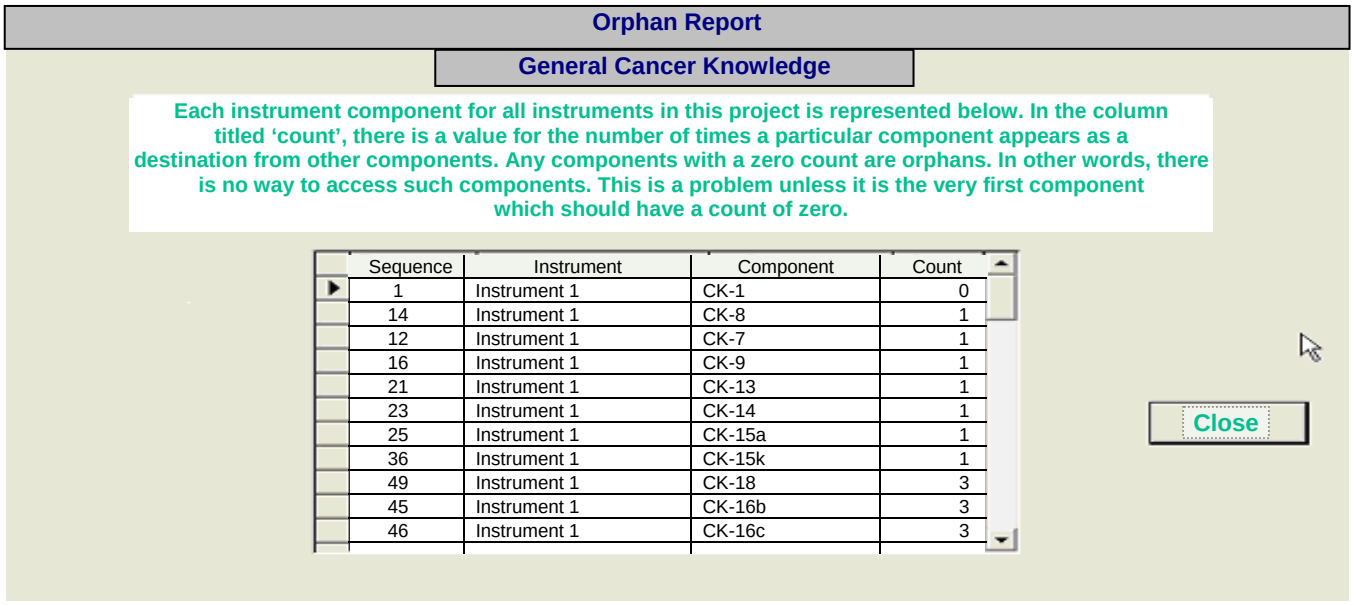
$$\text{Chemins} = \sum_{i}^{P} \prod_{j}^{NP_i} \text{liens}_{ij}$$

où $\boldsymbol{i}$ représente le i^{e} parcours, $\mathbf{P}$ est le nombre total de parcours, $\boldsymbol{j}$ est le j^{e} ensemble de liens à l'intérieur d'un parcours donné, $\mathbf{NP}_{\boldsymbol{i}}$ est le nombre de paires de nœuds liés entre eux à l'intérieur d'un parcours donné et *liens* est le nombre de liens unissant une paire de nœuds.

Si un protocole d'essai repose sur un échantillon de chemins, la base du graphe comportera un ensemble à la fois minimal et exhaustif, ou un univers, de cas types. Dans ce contexte, le terme « base » a un sens analogue à celui applicable en géométrie. En effet, la base d'un espace géométrique est composée d'un ensemble de vecteurs qui suffit à couvrir l'espace, ou si l'on veut à localiser n'importe quel point à l'intérieur de l'espace. De même, la base d'un graphe est un ensemble de parcours qui suffit à englober toutes les paires de nœuds prédécesseurs-successeurs. Cela signifie que la base comprend tous les nœuds ainsi qu'au moins un des liens reliant chaque paire de nœuds connectés entre eux. Une base constituera un sous-ensemble de l'ensemble des parcours possibles. Tout questionnaire comporte un ensemble de parcours (**P**) dont chaque membre satisfait à la définition des parcours énoncée précédemment (séquence unique de nœuds). À l'intérieur de cet ensemble, il y a un sous-ensemble dont la caractéristique est que chaque parcours qui en fait partie contient au moins une paire de nœuds connectés qui n'est contenue dans aucun autre parcours du sous-ensemble. Il s'agit du sous-ensemble de « parcours de base » (**PB**).

Pour mieux comprendre la différence entre les parcours faisant partie du sous-ensemble **PB** et ceux du sous-ensemble qui constitue son complément (soit **P** – **PB**), examinons le graphe présenté à la figure 5. L'ensemble des parcours (**P**) du graphe est :

Parcours 1 - 1, 2, 4, 5, 7
Parcours 2 - 1, 2, 4, 6, 7
Parcours 3 - 1, 3, 4, 5, 7
Parcours 4 - 1, 3, 4, 6, 7



Sequence	Instrument	Component	Count
1	Instrument 1	CK-1	0
14	Instrument 1	CK-8	1
12	Instrument 1	CK-7	1
16	Instrument 1	CK-9	1
21	Instrument 1	CK-13	1
23	Instrument 1	CK-14	1
25	Instrument 1	CK-15a	1
36	Instrument 1	CK-15k	1
49	Instrument 1	CK-18	3
45	Instrument 1	CK-16b	3
46	Instrument 1	CK-16c	3

Figure 4 Rapport sur les orphelins

Que l'on élimine n'importe quel de ces quatre parcours, les trois autres incluront chaque paire de nœuds connectés ; dès lors, n'importe quel trio de parcours constituera un ensemble de parcours de base (**PB**). Ainsi, si le parcours 1 est éliminé, chaque paire de nœud continuera d'être incluse dans les parcours 2, 3 et 4. Par contre, si l'on élimine les parcours 1 et 2, les paires de nœud 1 - 2 et 2 - 4 seraient exclus. L'ensemble de deux parcours ne suffirait donc pas à couvrir l'ensemble des séquences indépendantes de nœuds dans le graphique.

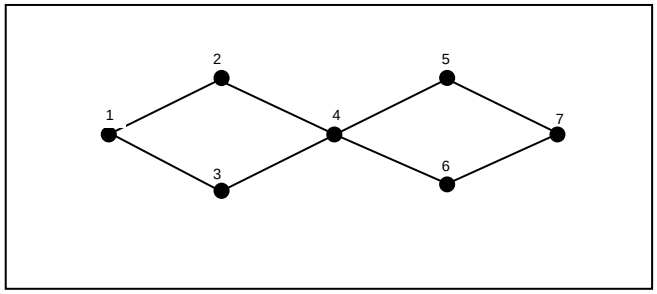


Figure 5 Représentation des parcours et de l'ensemble de parcours de base

Ainsi que cela ressort de la figure 1, de nombreux questionnaires existants comptent tant de chemins qu'il est peu judicieux dans la pratique de vouloir faire l'essai de la totalité d'entre eux. Il faut aussi considérer qu'un chemin typique à l'intérieur d'un instrument sera assorti d'un ou de plusieurs chemins similaires, c'est-à-dire des chemins composés du même ensemble de nœuds et pouvant ne différer que par un unique lien parallèle. De ce fait, non seulement faire l'essai de tous les chemins serait fastidieux, du fait du grand nombre de chemins, mais aussi redondant, étant donné l'existence de nombreux chemins similaires. Le concepteur de l'essai aura donc pour tâche de choisir un sous-ensemble de chemins afin de maximiser la couverture et de limiter le plus possible la redondance. Il est possible d'y arriver en utilisant au départ le sous-ensemble **PB** à titre de première étape de l'échantillonnage à partir de l'univers de chemins. Utiliser **PB** ainsi équivaut à amorcer le processus d'échantillonnage avec un échantillon par choix raisonné (Cochran 1977, page 10). On peut aussi concevoir cette première étape comme une redéfinition de l'univers des éléments dans le but d'éliminer la redondance. L'univers ainsi constitué a une couverture complète et contient le plus petit ensemble de cas requis pour inclure toutes les paires de nœuds connectés. L'étape suivante pourrait consister à choisir un ou plusieurs chemins à partir de chacun des parcours faisant partie du sous-ensemble **PB**, ce que l'on peut faire de plusieurs manières, notamment en envisageant chaque parcours comme une grappe de cas types puis en procédant à un échantillonnage probabiliste à partir de chaque grappe. Ou encore, il serait possible de choisir un chemin dans chaque grappe en sélectionnant au hasard un lien parallèle pour chaque nœud.

À partir du moment où l'on accepte la notion d'essai portant sur la base, il faut déterminer quelle portion de la base doit faire l'objet des essais. Si les essais portent sur tous les parcours du sous-ensemble **PB**, les seuls éléments d'un instrument qui ne font pas l'objet d'essais sont les liens redondants. Ces liens peuvent comporter des fautes d'orthographe ou des erreurs de formatage, mais il est peu probable

qu'ils contiennent des erreurs d'enchaînement. Cela tient à la nature des travaux de programmation entourant la création d'un instrument d'IAO. Les options de réponse sont habituellement « groupées », en ce sens que les options qui font passer à la même question suivante seront probablement toutes soit mal orientées, soit bien orientées. C'est pourquoi soumettre une base à des essais complets constitue une méthode efficace de limiter les erreurs les plus susceptibles de donner lieu à la perte de données.

Cela dit, il se peut que des essais non exhaustifs constituent la seule stratégie raisonnable s'il existe des contraintes au chapitre du temps ou des efforts requis et si la base compte un grand nombre de parcours. Malgré le fait que n'importe quelle partie d'un instrument qui n'a pas fait l'objet d'essais peut contenir une erreur, toute portion de parcours dans la base peut constituer un essai non biaisé. Par conséquent, le pourcentage de parcours à inclure dans les essais dépendra probablement de facteurs propres à la situation conceptuelle particulière. Par exemple, un instrument peut compter des modules ayant été utilisés antérieurement ou qui n'ont fait l'objet que de modifications mineures depuis leur utilisation précédente. Il ne sera pas nécessaire de soumettre ces modules à des essais aussi poussés que s'il s'agit de modules entièrement nouveaux. En règle générale, un échantillon de cas types de la taille minimale requise devrait englober chaque section de l'instrument faisant partie d'un ou de plusieurs parcours, et les parcours inclus devraient couvrir toutes les connexions entre sections.

7. Discussion et conclusion

Une approche de développement de logiciels fondée sur la théorie des graphes présente deux grands avantages par rapport aux approches conventionnelles. D'abord, on peut ainsi disposer d'un système de documentation capable de simuler le déroulement d'une interview assistée par ordinateur. Cette caractéristique sera utile pour vérifier les enchaînements et aidera les évaluateurs à comparer directement le déroulement de l'instrument et les spécifications conceptuelles. L'autre grand avantage tient à la sélection de cas aux fins des essais. L'utilisation de la base d'un questionnaire rend possible la spécification d'un univers de cas types couvrant toutes les paires de nœuds à partir du nombre minimal de parcours requis. L'échantillonnage probabiliste effectué à partir de cet univers prévient tout biais dans les procédures d'essai.

Dans la pratique, on peut obtenir le premier de ces deux avantages en structurant la base de données d'après un système de spécifications qui fera en sorte que l'on dispose d'un tableau des nœuds et d'un tableau des liens. Si le tableau des liens précise le nœud prédécesseur et le nœud successeur, les recherches dans les tableaux permettront de procéder à la vérification des enchaînements et à des simulations. L'autre avantage peut être obtenu au moyen d'un algorithme servant à déterminer une base. Ainsi que l'a mentionné Poole (1995), l'une des principales choses à faire lorsque l'on prévoit faire l'essai d'un logiciel consiste à déterminer quels cas types utiliser. Poole a présenté un algorithme qui accomplit cette tâche à partir du graphe de déroulement du programme. L'utilisation d'un tel graphe à cette fin est appropriée dans la mesure où le programme n'est pas de trop grande taille. Dans le cas de programmes compliqués et de grande envergure, les diagrammes de déroulement deviennent trop complexes. Il en va de même avec les questionnaires très étendus et compliqués (Bethlehem et Hundepool 2004). On trouvera en annexe le résultat obtenu au moyen d'un algorithme qui, à partir d'un graphe de questionnaire servant d'exemple, produit une base, fait le compte des chemins et détermine les parcours de base. (Il est possible d'obtenir cet algorithme en communiquant avec l'auteur (sdelliott2@verizon.net). L'algorithme en question ne traite pas les structures en boucle inhérentes aux contrôles ou aux fonctions de retour en arrière. Ces structures peuvent faire l'objet d'essais en dehors du graphe du questionnaire. Un algorithme traitant les boucles est en cours d'élaboration.)

Une approche reposant sur la théorie des graphes est en outre utile parce qu'elle permet de recourir à différentes mesures descriptives des questionnaires, comme le nombre de chemins, le nombre de parcours, la complexité cyclomatique (soit la mesure de la complexité du code d'un logiciel (se reporter à Hetzel 1984, McCabe 1976, Watson et McCabe 1996) ; elle est égale au nombre de parcours qui constituent la base du graphe. Dans le cas d'un graphe orienté où il ne peut y avoir de liens parallèles, la complexité cyclomatique ($\mathbf{CC}$) = $\mathbf{L} - \mathbf{N} + 2$, où $\mathbf{L}$ est le nombre de liens et $\mathbf{N}$, le nombre de nœuds) et plusieurs types de matrices descriptives (se reporter à l'annexe). Les améliorations qui seront apportées aux approches fondées sur la théorie des graphes dans l'avenir prendront probablement la forme, par exemple, 1) de taxonomies relatives aux éléments, aux liens et aux erreurs, 2) de tableaux secondaires dans la base de données des spécifications concernant les attributs propres aux différents types de nœuds et de liens, 3) de plans d'échantillonnage sophistiqués pour la sélection des cas types et 4) du recours à l'échantillonnage des chemins par choix raisonné.

Les taxonomies favoriseront la spécification de types particuliers d'éléments des instruments ainsi que l'incorporation de tableaux secondaires au système de documentation. Un type particulier d'éléments comprendra par exemple des éléments comportant une fonction de randomisation.

De tels éléments pourraient servir à des fins de sélection à plusieurs phases, de sorte qu'un répondant qui fait état, par exemple, d'une maladie donnée soit plus susceptible d'être acheminé vers une section de suivi portant sur la maladie en question. Dans un tel cas, la question initiale sur la maladie pourrait entrer dans une catégorie spéciale appelée « sélection des répondants ». Le tableau secondaire relatif aux questions de sélection des répondants dans le système de documentation pourra comporter des attributs ayant trait à un générateur de nombres aléatoires (générateur avec élément initial (« seed »), seuils de sélection, *etc.*).

Au chapitre de l'échantillonnage, les améliorations pourraient consister entre autres en un échantillonnage stratifié (Cochran 1977, page 89) ou en un échantillonnage avec probabilité proportionnelle à la taille. L'échantillonnage stratifié pourrait servir à garantir que toutes les sections que comporte un questionnaire sont prises en compte. Les parcours seraient stratifiés en fonction des sections qu'ils traversent. Dans le contexte de l'échantillonnage avec probabilité proportionnelle à la taille, la taille pourrait servir de mesure de la longueur du parcours, la probabilité de sélection d'un parcours donné étant fonction du nombre de nœuds que compte le parcours. Les parcours plus longs seraient donc sélectionnés plus fréquemment. On pourra utiliser l'échantillonnage des chemins par choix raisonné pour faire l'essai de caractéristiques d'un instrument autres que les erreurs de programmation. Par exemple, lors des dernières étapes de l'élaboration d'un questionnaire, on peut mettre l'accent sur des séquences de questions précises pour faire l'essai des caractéristiques cognitives de l'instrument.

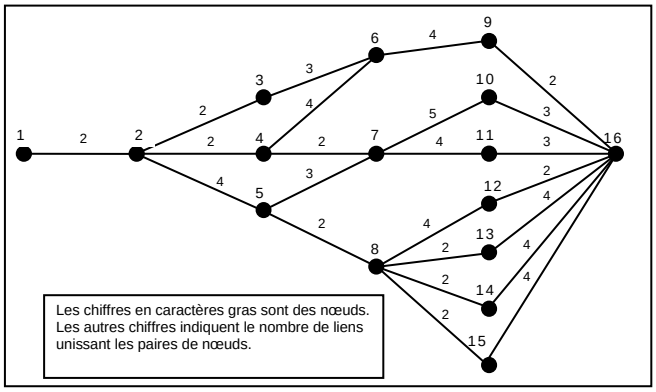
D'autres chercheurs du domaine mettront sans doute de l'avant d'autres améliorations touchant l'application de la théorie des graphes à l'élaboration de questionnaires. Il semble bien que l'on puisse tirer parti de cette théorie pour la description, l'élaboration et l'essai d'instruments d'IAO complexes. En matière d'IAO, la tendance actuelle semble aller dans le sens du recours à des instruments plus sophistiqués et de plus vaste portée, d'où l'utilité de disposer, dans le cadre des travaux d'élaboration, d'outils qui aident à documenter les éléments des instruments et à repérer les erreurs.

Remerciements

Il convient de souligner l'appui de Westat Incorporated, de Rockville (Maryland) à l'égard d'une partie importante des travaux ayant abouti à la rédaction de ce document. Mentionnons également la collaboration des réviseurs et des rédacteurs de *Techniques d'enquête* ainsi que les améliorations qu'ils ont apportées au document.

Annexe

Exemple de génération d'une base



Liens (en excluant les liens redondants) = 23
Nœuds = 16

Figure 6 Graphique de questionnaire

Tableau 1
Nombre d'embranchements pour chaque nœud

Nœud	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Nombre d'embranchements	1	3	1	2	2	1	2	4	1	1	1	1	1	1	1	0

Tableau 2
Matrice des liens

Nœud	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		2														
2			2	2	4											
3						3										
4						4	2									
5							3	2								
6									4							
7										5	4					
8												4	2	2	2	
9																2
10																3
11																3
12																2
13																4
14																4
15																4
16																

Chaque cellule contient une valeur relative au nombre de liens entre les nœuds figurant dans les colonnes et ceux figurant dans les lignes.

Tableau 3
Matrice des parcours

	1er nœud	2e nœud	3e nœud	4e nœud	5e nœud	6e nœud	7e nœud	8e nœud	9e nœud	10e nœud
Parcours 1	1	2	3	6	9	16				
Parcours 2	1	2	4	6	9	16				
Parcours 3	1	2	5	7	10	16				
Parcours 4	1	2	4	7	10	16				
Parcours 5	1	2	5	8	12	16				
Parcours 6	1	2	5	7	11	16				
Parcours 7	1	2	4	7	11	16				
Parcours 8	1	2	5	8	13	16				
Parcours 9	1	2	5	8	14	16				
Parcours 10	1	2	5	8	15	16				

Les valeurs dans les cellules représentent les nœuds. Chaque ligne correspond à un parcours.
[*Remarque* : *Dans cet exemple, les parcours comptent tous six nœuds. Toutefois, de manière générale, les parcours n'auront pas tous le même nombre de nœuds.*]

Tableau 4
Nombre de liens et nombre de chemins pour chaque parcours

	Appariement des nœuds					Chemins
	1er au 2e	2e au 3e	3e au 4e	4e au 5e	5e au 6e	
Parcours 1	2	2	3	4	2	96
Parcours 2	2	2	4	4	2	128
Parcours 3	2	4	3	5	3	360
Parcours 4	2	2	2	5	3	120
Parcours 5	2	4	2	4	2	128
Parcours 6	2	4	3	4	3	288
Parcours 7	2	2	2	4	3	96
Parcours 8	2	4	2	2	4	128
Parcours 9	2	4	2	2	4	128
Parcours 10	2	4	2	2	4	128

Parcours = 10 Nombre total de chemins = 1 600
Les cellules indiquent le nombre total de liens entre nœuds successifs à l'intérieur d'un parcours.

Tableau 5
Matrice des parcours de base

	1er nœud	2e nœud	3e nœud	4e nœud	5e nœud	6e nœud	7e nœud	8e nœud	9e nœud	10e nœud
Parcours de base 1	1	2	3	6	9	16				
Parcours de base 2	1	2	4	6	9	16				
Parcours de base 3	1	2	5	7	10	16				
Parcours de base 4	1	2	4	7	10	16				
Parcours de base 5	1	2	5	8	12	16				
Parcours de base 6	1	2	5	7	11	16				
Parcours de base 7	1	2	5	8	13	16				
Parcours de base 8	1	2	5	8	14	16				
Parcours de base 9	1	2	5	8	15	16				

Les valeurs dans les cellules représentent les nœuds. Chaque ligne correspond à un parcours de base.

Tableau 6
Nombre de liens et nombre de chemins pour chaque parcours de base

	Appariement des nœuds					Chemins
	1er au 2e	2e au 3e	3e au 4e	4e au 5e	5e au 6e	
Parcours de base 1	2	2	3	4	2	96
Parcours de base 2	2	2	4	4	2	128
Parcours de base 3	2	4	3	5	3	360
Parcours de base 4	2	2	2	5	3	120
Parcours de base 5	2	4	2	4	2	128
Parcours de base 6	2	4	3	4	3	288
Parcours de base 7	2	4	2	2	4	128
Parcours de base 8	2	4	2	2	4	128
Parcours de base 9	2	4	2	2	4	128

Parcours de base = 9 Nombre total de chemins = 1 504
Les cellules indiquent le nombre total de liens entre nœuds successifs d'après la matrice des parcours de base ci-avant.

Bibliographie

Balakrishnan, V.K. (1997). *Graph theory*. New York : McGraw Hill, Inc.

Beizer. B. (1995). *Black-box testing*. New York : John Wiley & Sons, Inc.

Berge, C. (1976). *Graphs and hypergraphs*. New York : North-Holland Publishing Company - Amsterdam, London and American Elsevier Publishing Company, Inc.

Bethlehem, J., et Hundepool, A. (2004). TADEQ: A tool for the documentation and analysis of electronic questionnaires. *Journal of Official Statistics*, 20, 233-264.

Centers for Medicare and Medicaid Services (2010). *Medicare Current Beneficiary Survey*: *Overview*. Août 2002. Adresse URL : https://www.cms.gov/LimitedDataSets/11_MCBS.asp.

Chartrand, G. (1985). *Introductory Graph Theory*. New York : Dover Publications, Inc., Mineola.

Cochran, W.G. (1977). *Sampling Techniques*, 3ième Édition. New York : John Wiley & Sons, Inc.

Cohen, J. (1997). *Design and methods of the Medical Expenditure Panel Survey Household Component*. Rockville (MD) : Agency for Health Care Policy and Research. MEPS Methodology Report No. 1. AHCPR Pub. No. 97-0026.

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L. et O'Reilly, J.M. (1988). *Computer Assisted Information Collection*. New York : John Wiley & Sons, Inc.

Gibbons, A. (1985). *Algorithmic Graph Theory*. Cambridge University Press.

Harary, F., et Palmer, E. (1973). *Graphical Enumeration*. New York : Academic Press.

Hetzel, W. (1983). *The Complete Guide to Software Testing, QED*. Massachusetts : Information Sciences, Inc., Wellesley.

Kaner, C., Falk, J. et Nguyen, H. (1999). *Testing Computer Software*, 2ième Édition. New York : John Wiley & Sons, Inc.

Medical Expenditure Panel Survey (MEPS) (2010). *Survey Instruments and Associated Documentation*. http://www.meps.ahrq.gov/mepsweb/.

Myers, G.J. (1979). *The Art of Software Testing*. New York : John Wiley & Sons, Inc.

Patton, R. (2006). *Software Testing*, 2ième Édition. Sams Publishing, Inc.

Poole, J. (1995). NISTIR 5737 – A Method to Determine a Basis Set of Paths to Perform Program Testing, National Institute of Standards and Technology, Gaithersburg, MD, Novembre, 1995.

Statistique Canada (2010). *Enquête Canadienne sur les capacités financières* (*ECCF*): *Questionnaire 2009*. http://www.statcan.gc.ca.

Statistics Netherlands (2002). Blaise Developer's Guide. Department of Statistical Informatics, Statistics Netherlands, Heerlen.

US Bureau of Labor Statistics (2010). *Consumer Expenditure Surveys Quarterly Interview CAPI Survey 2010*. United States Department of Labor, http://www.bls.gov/cex/capi/2010/cecapihome.htm.

Watson, A., et McCabe, T. (1996). NIST Special Publication 500-235 Structured Testing: A Testing Methodology Using the Cyclomatic Complexity Metric. National Institute of Standards and Technology, Gaithersburg, MD, Septembre, 1996.

ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 25-32
Statistique Canada, N° 12-001-X au catalogue

À propos de la répartition de l'échantillon pour une estimation sur domaine efficace

G. Hussain Choudhry, J.N.K. Rao et Michael A. Hidiroglou [1]

Résumé

Les questions concernant la répartition de l'échantillon sont étudiées dans le contexte de l'estimation des moyennes de sous-population (strate ou domaine), ainsi que de la moyenne de population agrégée sous échantillonnage aléatoire simple stratifié. Une méthode de programmation non linéaire est utilisée pour obtenir la répartition « optimale » de l'échantillon entre les strates qui minimise la taille totale d'échantillon sous la contrainte des tolérances spécifiées pour les coefficients de variation des estimateurs des moyennes de strate et de la moyenne de population. La taille totale d'échantillon résultante est alors utilisée pour déterminer les répartitions de l'échantillon par les méthodes de Costa, Satorra et Ventura (2004) s'appuyant sur une répartition intermédiaire ou de compromis et de Longford (2006) fondée sur des « priorités inférencielles » spécifiées. En outre, nous étudions la répartition de l'échantillon entre les strates quand sont également spécifiées des exigences de fiabilité pour des domaines qui recoupent les strates. Les propriétés des trois méthodes sont étudiées au moyen de données provenant de l'Enquête mensuelle sur le commerce de détail (EMCD) menée par Statistique Canada auprès d'établissements uniques.

Mots clés : Estimateurs composites ; répartition intermédiaire ; estimateurs directs ; moyennes de domaine ; programmation non linéaire.

1. Introduction

L'usage de l'échantillonnage simple stratifié est très répandu dans les enquêtes auprès des entreprises et d'autres enquêtes auprès des établissements employant des listes comme base de sondage. La moyenne de population $\bar{Y} = \sum_h W_h \bar{Y}_h$ est estimée par la moyenne d'échantillon pondérée $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, où $W_h = N_h / N$ est la taille relative de la strate h $(= 1, ..., L)$, et $\bar{Y}_h$ et $\bar{y}_h$ sont les moyennes de population et d'échantillon, respectivement, de la strate. La méthode bien connue de Neyman en vue de répartir l'échantillon entre les strates est optimale pour l'estimation de la moyenne de population si l'on considère la minimisation de la variance de $\bar{y}_{st}$ sous la contrainte que $\sum_h n_h = n$, où n est fixe, ou la minimisation de $\sum_h n_h$ sous la contrainte que la variance de $\bar{y}_{st}$ est fixe, où n_h est la taille d'échantillon de la strate. Toutefois, la répartition de Neyman peut donner lieu à de grands coefficients de variation (CV) des moyennes $\bar{y}_h$. Par ailleurs, l'équirépartition de l'échantillon, $n_h = n/L$, est efficace pour l'estimation des moyennes de strate, mais peut produire de beaucoup plus grands CV de l'estimateur $\bar{y}_{st}$ que la méthode de Neyman.

Bankier (1988) a proposé une « répartition exponentielle » (*power allocation*) comme compromis entre la répartition de Neyman et l'équirépartition. En posant que $C_h = S_h / \bar{Y}_h$ est le CV de la strate, la répartition exponentielle est donnée par

$$n_h^B = n \frac{C_h X_h^q}{\sum_h C_h X_h^q}, \; h = 1, \ldots, L \qquad (1.1)$$

où X_h est une mesure de taille ou d'importance de la strate h, et q est une constante de mise au point. La répartition exponentielle (1.1) s'obtient en minimisant $\sum_h \{X_h^q \, \mathrm{CV}(\bar{y}_h)\}^2$ sous la contrainte $\sum_h n_h = n$, où $\mathrm{CV}(\bar{y}_h)$ est le CV de la moyenne d'échantillon de la strate $\bar{y}_h$. Le choix $q = 1$ et $X_h = N_h \bar{Y}_h$ dans (1.1) mène à la répartition de Neyman

$$n_h^N = n \frac{N_h S_h}{\sum_h N_h S_h}, \; h = 1, \ldots, L \qquad (1.2)$$

et le choix $q = 0$ donne l'équirépartition si $C_h = C$ pour tout h, où S_h^2 est la variance dans la strate. Bankier (1988) a considéré les valeurs de q comprises entre 0 et 1 comme produisant des répartitions intermédiaires (*compromise allocations*). Il a donné un exemple numérique pour illustrer comment q peut être choisi en pratique. Le choix de $X_h = N_h$ et $q = 1/2$ dans (1.1) donne la « répartition proportionnelle à la racine carrée de la population » $n_h = n\sqrt{N_h} / \sum_h \sqrt{N_h}$ si $C_h = C$. La répartition exponentielle (1.1) et certains autres modes de répartition dépendent généralement de la variable d'intérêt y, de sorte qu'en pratique, on utilise à la place de cette dernière une variable indirecte dont les valeurs de population sont connues.

Costa et coll. (2004) ont proposé une répartition intermédiaire fondée sur une combinaison convexe d'une répartition proportionnelle, $n_h = nW_h$, et d'une équirépartition, $n_h = n/L$ (voir la section 2.1). Longford (2006) a étudié systématiquement la répartition sous échantillonnage aléatoire simple stratifié en introduisant des « priorités

1. G.H. Choudhry, Division de la recherche et de l'innovation en statistique, Statistique Canada. Courriel : ghchoudhry@gmail.com ; J.N.K. Rao, École de mathématiques et de statistique, Université Carleton. Courriel : jrao@math.carleton.ca ; M.A. Hidiroglou, Division de la recherche et de l'innovation en statistique, Statistique Canada. Courriel : mike.hidiroglou@statcan.gc.ca.

inférencielles » P_h pour les strates h et G pour la population. En particulier, il a supposé que $P_h = N_h^q$ pour une valeur spécifiée de q $(0 \le q \le 2)$ (voir la section 2.4). Il a également étudié le cas de petites tailles d'échantillon de strate n_h pour lesquelles les estimateurs composites des moyennes de strate $\overline{Y}_h$ peuvent être utilisés.

L'objectif principal du présent article est de proposer une méthode de répartition « optimale », fondée sur la programmation non linéaire (NLP, pour *non-linear programming*) (voir la section 2.3). La méthode consiste à minimiser la taille totale d'échantillon $\sum_h n_h$ sous la contrainte de tolérances spécifiées pour les CV des moyennes d'échantillon de strate $\overline{y}_h$ et de la moyenne de population estimée $\overline{y}_{st}$. Le cas des estimateurs indirects (composites) des moyennes de strate est étudié à la section 3. À la section 4, nous examinons la répartition optimale de l'échantillon entre les strates quand des exigences de fiabilité pour des domaines recoupant les strates sont également spécifiées. La méthode proposée s'étend facilement à plusieurs variables, mais pour simplifier, nous omettons les détails. À la section 5, en partant de la taille totale d'échantillon optimale obtenue par NLP, nous procédons à une étude numérique des propriétés des méthodes de Costa et coll. et de Longford en ce qui concerne le respect des exigences de fiabilité.

2. Répartition pour les estimateurs directs

À la présente section, nous considérons les estimateurs directs, $\overline{y}_h$, des moyennes de population de strate sous échantillonnage aléatoire simple stratifié. Le cas des estimateurs indirects des moyennes de strate est étudié à la section 3. Les estimateurs indirects sont utilisés lorsque les tailles d'échantillon de strate n_h sont faibles.

2.1 Répartition de Costa et coll.

La répartition de l'échantillon de Costa et coll. (2004) est donnée par

$$n_h^C = k(nW_h) + (1-k)(n/L) \tag{2.1}$$

pour une constante spécifiée $k(0 \le k \le 1)$. Cette répartition se réduit à l'équirépartition quand $k = 0$ et à la répartition proportionnelle quand $k = 1$. La formule (2.1) doit être modifiée quand $n / L > N_h$ pour une strate h dans un ensemble de strates A. La répartition modifiée est donnée par

$$\tilde{n}_h^C = k(nW_h) + (1-k)\, n_h^0, \tag{2.2}$$

où $n_h^0 = N_h$ si $h \in A$ et $n_h^0 = (n - \sum_{h\in A} N_h) / (L - m)$, autrement où m est le nombre de strates dans l'ensemble A. Notons que, si $k = 0$, (2.2) donne l'équirépartition modifiée. Nous examinons différents choix de la constante k dans l'étude numérique fondée sur des données provenant de l'Enquête mensuelle sur le commerce de détail (EMCD) de Statistique Canada présentée à la section 5.

2.2 Répartition de Longford

La méthode de Longford (2006) vise à contrôler simultanément la fiabilité des estimations des moyennes de strate $\overline{y}_h$ et de la moyenne de population $\overline{y}_{st}$ en minimisant la fonction d'objectif

$$\sum_{h=1}^{L} P_h V(\overline{y}_h) + (GP_+)V(\overline{y}_{st}) \tag{2.3}$$

par rapport aux tailles d'échantillon de strate n_h sous la contrainte $\sum_h n_h = n$, où $P_+ = \sum_h P_h$. Dans (2.3), le premier terme spécifie l'importance relative, P_h, de chaque strate h, tandis que le deuxième accorde une importance relative à $\overline{y}_{st}$ au moyen du poids G. Longford (2006) a supposé que $P_h = N_h^q$ pour une certaine constante $q(0 \le q \le 2)$. Le terme P_+ dans (2.3) neutralise l'effet des tailles P_h et du nombre de strates sur le poids G.

Sous échantillonnage aléatoire simple stratifié, la répartition de l'échantillon qui minimise (2.3) est

$$n_h^L = n\frac{S_h\sqrt{P_h'}}{\sum_h S_h\sqrt{P_h'}},\ h = 1, ..., L \tag{2.4}$$

où $P_h' = P_h + GP_+W_h^2$. Si $q = 2$, l'expression (2.4) ne dépend pas de la valeur de G et elle se réduit à la répartition de Neyman, n_h^N donnée par (1.2).

2.3 Répartition par programmation non linéaire (NLP)

Nous passons maintenant à la méthode NLP de détermination des tailles d'échantillon de strate n_h sous la contrainte des exigences de fiabilité spécifiées pour les moyennes d'échantillon de strate ainsi que pour la moyenne de population estimée. En posant que $\mathbf{f} = (f_1, ..., f_L)^T$ avec $f_h = n_h / N_h$, nous minimisons la taille totale d'échantillon

$$g(\mathbf{f}) = \sum_{h=1}^{L} f_h N_h \tag{2.5}$$

par rapport à f sous les contraintes

$$\mathrm{CV}(\overline{y}_h) \le \mathrm{CV}_{0h},\ h = 1, ..., L \tag{2.6}$$

$$\mathrm{CV}(\overline{y}_{st}) \le \mathrm{CV}_o \tag{2.7}$$

$$0 < f_h \le 1,\ h = 1, ..., L \tag{2.8}$$

où CV_{0h} et CV_0 sont les tolérances spécifiées pour les CV de la moyenne d'échantillon de strate $\overline{y}_h$ et de la moyenne de population estimée $\overline{y}_{st}$, respectivement. Des signes

d'inégalité sont utilisés dans (2.6) et (2.7) parce que les CV résultants pour certaines strates h et/ou pour l'agrégat peuvent être plus petits que les tolérances spécifiées (Cochran 1977, page 122).

En posant que $k_h = f_h^{-1}$, l'expression (2.5) devient une fonction convexe séparable des variables k_h,

$$\tilde{g}(\mathbf{k}) = \sum_{h=1}^{L} N_h k_h^{-1}. \qquad (2.9)$$

Nous spécifions à nouveau les contraintes (2.6) et (2.7) en fonction des variances relatives afin qu'elles soient linéaires en les variables k_h. La variance relative (VR) de $\bar{y}_h$ est le carré de son CV,

$$\mathrm{VR}(\bar{y}_h) = \frac{k_h - 1}{N_h} C_h^2. \qquad (2.10)$$

De même, la variance relative de $\bar{y}_{st}$ est le carré de son CV,

$$\mathrm{VR}(\bar{y}_{st}) = \bar{Y}^{-2} \sum_{h=1}^{L} W_h^2 \frac{k_h - 1}{N_h} S_h^2. \qquad (2.11)$$

Nous avons utilisé la procédure NLP de SAS avec l'option Newton-Raphson pour trouver la valeur optimale de k_h qui minimiserait (2.9) sous les contraintes

$$\mathrm{VR}(\bar{y}_h) \le \mathrm{VR}_{oh},\ h = 1, ..., L, \qquad (2.12)$$

$$\mathrm{VR}(\bar{y}_{st}) \le \mathrm{VR}_0, \qquad (2.13)$$

$$k_h \ge 1,\ h = 1, ..., L. \qquad (2.14)$$

$\mathrm{VR}(\bar{y}_h)$ et $\mathrm{VR}(\bar{y}_{st})$ sont données par (2.10) et (2.11) où $\mathrm{VR}_{0h} = \mathrm{VR}_{0h}^2$ et $\mathrm{VR}_0 = \mathrm{VR}_0^2$. En exprimant les contraintes sous forme linéaire et la fonction d'objectif sous forme d'une fonction convexe séparable, nous arrivons plus rapidement à la convergence du problème de NLP reformulé. En désignant la solution NLP comme étant $\mathbf{k}^0 = (k_1^0, ..., k_L^0)^T$, le vecteur correspondant de tailles optimales d'échantillon de strate est donné par $\mathbf{n}^0 = (n_1^0, ..., n_L^0)^T$, où $n_h^0 = N_h / k_h^0$. Nous pouvons modifier (2.14) pour nous assurer que $n_h^0 \ge 2$ pour tout h, ce qui permet d'obtenir une estimation sans biais des variables.

La méthode NLP peut être étendue facilement à plusieurs variables $y_1, ..., y_P$ en spécifiant des tolérances pour les CV des moyennes de strate et de la moyenne de population estimées pour chaque variable $(p = 1, ..., P)$. Si le nombre de variables P n'est pas faible, la taille d'échantillon totale optimale résultante $n^0 = \sum_h n_h^0$ peut devenir considérablement plus grande que n^0 pour une seule variable. Huddleston, Claypool et Hocking (1970), Bethel (1989) et d'autres chercheurs ont étudié la NLP pour obtenir une répartition optimale de l'échantillon dans le cas de l'estimation des moyennes de population de plusieurs variables sous échantillonnage aléatoire stratifié.

3. Répartition pour les estimateurs composites

Longford (2006) a étudié les estimateurs composites des moyennes de strate de la forme

$$\hat{\theta}_h = \alpha_h \bar{y}_h^S + (1 - \alpha_h) \bar{y}_h \qquad (3.1)$$

où $\bar{y}_h^S$ est un estimateur synthétique ; ici, nous prenons $\bar{y}_h^S = \bar{y}_{st}$. L'EQM de $\hat{\theta}_h$ est

$$\begin{aligned} \mathrm{EQM}(\hat{\theta}_h) &= V(\hat{\theta}_h) + [B(\hat{\theta}_h)]^2 \\ &= \alpha_h^2 \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} + (1 - \alpha_h)^2 W_h^2 \frac{S_h^2}{n_h} \\ &\quad + 2\alpha_h (1 - \alpha_h) W_h \frac{S_h^2}{n_h} + \alpha_h^2 (\bar{Y}_h - \bar{Y})^2 \\ &\quad + \text{termes ne dépendant pas des } n_h. \end{aligned} \qquad (3.2)$$

Longford (2006) a montré que, dans (3.1), le coefficient optimal α_h qui minimise (3.2) est approximativement égal à $\alpha_h^* = S_h^2 (S_h^2 + n_h \Delta_h^2)^{-1}$, où $\Delta_h = \bar{Y}_h - \bar{Y}$. Il a ensuite remplacé Δ_h^2 dans α_h^* par sa moyenne sur les strates, désignée par $\sigma_B^2 = L^{-1} \sum_h (\bar{Y}_h - \bar{Y})^2$, ce qui donne $\alpha_h^* \approx (1 + n_h \omega_h)^2$, où $\omega_h^2 = \sigma_B^2 / S_h^2$. L'EQM résultante de $\hat{\theta}_h$ est approximée par

$$\mathrm{EQM}(\hat{\theta}_h) \approx \frac{\sigma_B^2}{1 + n_h \omega_h}. \qquad (3.3)$$

La répartition de Longford s'obtient en minimisant la fonction d'objectif

$$\sum_{h=1}^{L} P_h \mathrm{EQM}(\hat{\theta}_h) + (\mathrm{GP}_+) V(\bar{y}_h) \qquad (3.4)$$

par rapport aux n_h. La solution résultante satisfait à

$$\frac{P_h \sigma_B^2 \omega_h}{(1 + n_h \omega_h)^2} + (\mathrm{GP}_+) W_h^2 \frac{S_h^2}{n_h} = \text{const.},\ h = 1, ..., L. \qquad (3.5)$$

Longford a utilisé une méthode itérative pour obtenir la solution de (3.5) puisqu'il n'y a pas de solution analytique.

Notre procédure NLP minimise $g(\mathbf{f})$ donnée par (2.5) sous les contraintes

$$\mathrm{REQM}(\hat{\theta}_h) \le \mathrm{REQM}_{0h},\ h = 1, ..., L;\ \mathrm{VR}(\bar{y}_{st}) \le \mathrm{VR}_o \qquad (3.6)$$

et (2.8), où $\text{REQM}(\hat{\theta}_h) = \text{EQM}(\hat{\theta}_h)/\bar{Y}_h^2$ et REQM_{0h} est une tolérance spécifiée. L'approximation (3) de $\text{EQM}(\hat{\theta}_h)$ est utilisée dans (3.6).

4. Répartition pour l'estimation sur domaine

Supposons que la population U est partitionnée en domaines ${}_dU$ $(d = 1, ..., D)$ qui recoupent les strates. En outre, supposons que les estimateurs des moyennes de domaine doivent satisfaire à des tolérances de variance relative spécifiées, ${}_d\text{VR}_0$, $d = 1, ..., D$. Nous utilisons la méthode NLP pour trouver les tailles additionnelles optimales d'échantillon de strate nécessaires pour satisfaire aux tolérances de domaine.

Un estimateur de la moyenne de domaine ${}_d\bar{Y} = {}_dN^{-1}\sum_{k\in U_d} y_k$ est l'estimateur par le ratio

$$ {}_d\hat{\bar{Y}} = \frac{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k\in s_h} {}_d\delta_k\, y_k}{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k\in s_h} {}_d\delta_k}, \qquad (4.1) $$

où ${}_d\delta_k = 1$ si $k \in {}_dU$ et ${}_d\delta_k = 0$ autrement, s_h est l'échantillon tiré de la strate h et ${}_dN$ est la taille du domaine d. La variance relative de l'estimateur par le ratio (4.1) est $\text{VR}({}_d\hat{\bar{Y}}) = V({}_d\hat{\bar{Y}})/{}_d\bar{Y}^2$, où la variance $V({}_d\hat{\bar{Y}})$ s'obtient au moyen de la formule de linéarisation habituelle pour un estimateur par le ratio.

Soit $\tilde{n}_h$ la taille totale révisée de l'échantillon tiré de la strate h, telle que l'accroissement de l'échantillon provenant de la strate h est $\tilde{n}_h - n_h^0$. Soit $\tilde{f}_h = \tilde{n}_h / N_h$ la fraction d'échantillonnage correspondante. Nous obtenons la taille optimale $\tilde{\mathbf{n}} = (\tilde{n}_1, ..., \tilde{n}_L)^T$ en minimisant l'accroissement de l'échantillon

$$ g(\tilde{\mathbf{f}}) - \sum_{h=1}^{L} n_h^0 N_h = \sum_{h=1}^{L} (\tilde{f}_h - f_h^0) N_h \qquad (4.2) $$

par rapport à $\tilde{\mathbf{f}} = (\tilde{f}_1, ..., \tilde{f}_L)^T$ sous les contraintes

$$ f_h^0 \le \tilde{f}_h \le 1, \quad h = 1, ..., L \qquad (4.3) $$

$$ \text{VR}({}_d\hat{\bar{Y}}) \le {}_d\text{VR}_0, \quad d = 1, ..., D. \qquad (4.4) $$

Comme précédemment, nous reformulons le problème en exprimant (4.2), (4.3) et (4.4) en fonction de $\tilde{\mathbf{k}} = (\tilde{k}_1, ..., \tilde{k}_L)^T$, où $\tilde{k}_h = \tilde{f}_h^{-1}$. Cela nous mène à la minimisation de la fonction convexe séparable

$$ g^*(\tilde{\mathbf{k}}) = \sum_{h=1}^{L} N_h \tilde{k}_h^{-1} \qquad (4.5) $$

par rapport à $\tilde{\mathbf{k}}$ sous les contraintes linéaires

$$ 1 \le \tilde{k}_h \le k_h^0, \quad h = 1, ..., L \qquad (4.6) $$

et

$$ \text{VR}({}_d\hat{\bar{Y}}) = {}_d\bar{Y}^{-2} \sum_{h=1}^{L} \left(\frac{N_h}{{}_dN}\right)^2 \frac{\tilde{k}_h - 1}{N_h} {}_dS_{eh}^2 \le {}_d\text{VR}_0, \quad d = 1, ..., D \qquad (4.7) $$

où ${}_d\text{VR}_0$ est la tolérance spécifiée, ${}_dS_{eh}^2$ désigne la variance de strate des résidus ${}_de_k = {}_d\delta_k(y_k - {}_d\bar{Y})$ pour $k \in U_h$ et U_h désigne la population de la strate. Désignons les valeurs optimales résultantes $\tilde{k}_h$ et $\tilde{n}_h$ par $\tilde{k}_h^0$ et $\tilde{n}_h^0$, respectivement, de sorte que l'accroissement optimal de l'échantillon dans la strate h soit $\tilde{n}_h^0 - n_h^0$.

On peut montrer que la minimisation de la taille totale d'échantillon sous toutes les contraintes $\text{VR}(\bar{y}_h) \le \text{VR}_{0h}$, $h = 1, ..., L$, $\text{VR}({}_d\hat{\bar{Y}}) \le {}_d\text{VR}_0$, $d = 1, ..., D$, $\text{VR}(\bar{y}_{st}) \le \text{VR}_0$, et $0 < f_h \le 1$, $h = 1, ..., L$ mènera à la même solution optimale, $\tilde{n}^0 = (\tilde{n}_1^0, ..., \tilde{n}_L^0)^T$. Cependant, les exigences de fiabilité au niveau du domaine peuvent souvent être spécifiées après avoir déterminé n^0.

5. Résultats empiriques

À la présente section, nous étudions les propriétés relatives des diverses méthodes de répartition de l'échantillon en nous servant de données tirées de l'EMCD. Aux sections 5.1 et 5.2, nous présentons nos résultats pour les estimateurs directs et les estimateurs composites des moyennes de strate, respectivement. À la section 5.3, nous donnons les résultats pour les moyennes de domaine.

5.1 Moyennes de strate : estimateurs directs

Pour l'étude empirique, nous avons utilisé un sous-ensemble des valeurs obtenues pour la population de l'EMCD limité aux établissements uniques. Les tailles de strate, N_h, les moyennes de population de strate, $\bar{Y}_h$, les écarts-types de strate, S_h, et les CV de strate, $C_h = S_h/\bar{Y}_h$, sont donnés au tableau 1 pour les dix provinces du Canada (traitées comme des strates). Pour la répartition NLP, les tolérances que nous avons choisies pour les CV étaient $\text{CV}_{0h} = 15\,\%$ pour les moyennes de strate $\bar{y}_h$ et $\text{CV}_0 = 6\,\%$ pour la moyenne d'échantillon pondérée $\bar{y}_{st}$ désignée Canada (CA).

La répartition NLP satisfaisant aux tolérances de CV spécifiées a donné une taille globale minimale d'échantillon $n^0 = 3\,446$. Le tableau 2 montre la répartition de l'échantillon n_h^0 et les coefficients de variation $\text{CV}(\bar{y}_h)$ et $\text{CV}(\bar{y}_{st})$ associés pour la répartition NLP. Son examen montre que la répartition NLP respecte la tolérance spécifiée $\text{CV}_0 = 6\,\%$, donne des CV plus petits que la tolérance

spécifiée $CV_{0h} = 15\,\%$ pour deux des grandes provinces (Qc : 11,4 % et Ont. : 11,0 %) et atteint un CV de 15 % pour les autres provinces.

En utilisant la taille globale optimale d'échantillon égale à 3 446, nous avons calculé les répartitions de l'échantillon n_h et les coefficients de variation $CV(\bar{y}_h)$ et $CV(\bar{y}_{st})$ associés pour l'équirépartition modifiée, la répartition proportionnelle et la répartition fondée sur la racine carrée, qui sont présentés au tableau 2. Ce dernier montre clairement que l'équirépartition modifiée ne convient pas pour ce qui est de satisfaire aux tolérances spécifiées pour les CV, car elle donne $CV(\bar{y}_{st}) = 9{,}1\,\%$, valeur significativement plus grande que la valeur spécifiée $CV_0 = 6\,\%$. En outre, sous l'équirépartition modifiée, $CV(\bar{y}_h)$ égale 19,3 %, 16,3 % et 16,2 % pour les grandes provinces que sont l'Ontario, l'Alberta et la Colombie-Britannique, respectivement. Notons que pour l'Île-du-Prince-Édouard, qui est la plus petite province, le tableau 2 donne $CV(\bar{y}_h) = 0\,\%$ pour l'équirépartition modifiée, car pour cette province, la méthode donne $n_h = N_h$.

Tableau 1
Valeurs de population pour l'EMCD

Provinces	N_h	$\bar{Y}_h$	S_h	C_h
Terre-Neuve (T.-N.)	909	963	1 943	2,02
Île-du-Prince-Édouard (Î.-P.-É.)	280	712	1 375	1,93
Nouveau-Brunswick (N.-B.)	1 333	1 368	3 200	2,34
Nouvelle-Écosse (N.-É.)	1 153	1 568	4 302	2,74
Québec (Qc)	11 135	2 006	4 729	2,36
Ontario (Ont.)	21 531	1 722	6 297	3,66
Manitoba (Man.)	1 700	1 295	2 973	2,30
Saskatchewan (Sask.)	1 743	1 212	3 019	2,49
Alberta (Alb.)	5 292	1 698	5 358	3,16
Colombie-Britannique (C.-B.)	7 803	1 291	4 013	3,11
Canada (CA)	52 879	1 654	-	-

Tableau 2
Équirépartition et répartitions proportionnelle, fondée sur la racine carrée et NLP, et CV (%) associés

Province	Équi-répartition		Proportionnelle		Racine carrée		NLP	
	n_h	CV_h	n_h	CV_h	n_h	CV_h	n_h	CV_h
T.-N.	352	8,4	59	25,4	169	14,0	151	15,0
Î.-P.-É.	280	0,0	18	44,1	94	16,2	104	15,0
N.-B.	352	10,7	87	24,2	205	15,0	206	15,0
N.-É.	352	12,2	75	30,6	191	18,1	259	15,0
Qc	352	12,4	726	8,5	593	9,4	410	11,4
Ont.	352	19,3	1 403	9,4	824	12,5	1 056	11,0
Man.	352	10,9	111	21,1	232	14,0	206	15,0
Sask.	352	11,9	114	22,6	234	15,2	238	15,0
Alb.	352	16,3	345	16,4	408	15,0	409	15,0
C.-B.	352	16,2	508	13,3	496	13,5	407	15,0
CA	3 446	9,1	3 446	5,2	3 446	6,3	3 446	6,0

Dans le cas de la répartition proportionnelle, le tableau 2 donne $CV(\bar{y}_{st}) = 5{,}2\,\%$, mais cette méthode donne des CV de strate considérablement plus grands que la valeur spécifiée de 15 % pour sept provinces, les chiffres variant de 16,4 % à 44,1 %. Par ailleurs, le tableau 2 montre que la répartition fondée sur la racine carrée offre un compromis raisonnable pour ce qui est du respect des tolérances spécifiées pour les CV. Nous avons $CV(\bar{y}_{st}) = 6{,}3\,\%$ et $CV(\bar{y}_h) \leq 15\,\%$ pour sept provinces, les trois provinces pour lesquelles le CV est supérieur à 15 % étant la Saskatchewan avec 15,2 %, Île-du-Prince-Édouard avec 16,2 % et la Nouvelle-Écosse avec 18,1 %.

Le tableau 3 donne les résultats pour la répartition de Costa et coll. (2.1) pour $k = 0{,}25$, 0,50 et 0,75, en utilisant la valeur $n = 3\,446$ obtenue par programmation non linéaire (NLP). Nous constatons d'après le tableau 2 que le choix $k = 0{,}25$, qui donne plus de poids à l'équirépartition, n'est pas satisfaisant pour l'estimation de la moyenne de population (Canada), le coefficient de variation étant $CV(\bar{y}_{st}) = 7{,}2\,\%$, mais donne de bons résultats pour les moyennes de strate, sauf dans le cas de l'Alberta pour laquelle $CV(\bar{y}_h) = 16{,}3\,\%$. Par ailleurs, le choix $k = 0{,}75$, qui accorde plus de poids à la répartition proportionnelle, donne des résultats médiocres pour l'estimation des moyennes provinciales, $CV(\bar{y}_h)$ variant de 16,2 % à 21,4 % pour sept provinces, quoi que $CV(\bar{y}_{st})$ soit inférieur à la tolérance souhaitée de 6 %. Le choix intermédiaire $k = 0{,}50$ donne d'assez bons résultats, les coefficients de variation étant $CV(\bar{y}_{st}) = 6{,}4\,\%$ et $CV(\bar{y}_h)$ environ égal à 15 % ou moins, sauf pour deux provinces (Nouvelle-Écosse et Alberta) pour lesquelles les CV sont de 17,0 % et 16,5 %, respectivement. Les propriétés de la méthode de Costa et coll. pour $k = 0{,}50$ et de la répartition fondée sur la racine carrée sont plus ou moins semblables, et ni l'une ni l'autre ne dépend de la variable d'intérêt, y, contrairement aux répartitions de Longford et NLP.

Tableau 3
Répartition de Costa et coll. et CV (%) associés pour $k = 0{,}25$, 0,50 et 0,75

Province	$k = 0{,}25$		$k = 0{,}50$		$k = 0{,}75$	
	n_d	CV_d	n_d	CV_d	n_d	CV_d
T.-N.	278	10,1	205	12,4	132	16,2
Î.-P.-É.	214	6,4	149	10,8	83	17,8
N.-B.	286	12,3	219	14,5	153	17,8
N.-É.	282	14,2	213	17,0	144	21,4
Qc	446	10,9	539	9,9	633	9,1
Ont.	615	14,5	878	12,1	1 140	10,5
Man.	292	12,2	231	14,0	171	16,6
Sask.	292	13,3	733	15,2	174	17,9
Alb.	350	16,3	349	16,3	347	16,4
C.-B.	391	15,3	430	14,6	469	13,9
CA	3 446	7,2	3 446	6,2	3 446	5,6

Nous examinons maintenant la répartition de Longford (2.4) qui dépend de q et de G. Le tableau 4 donne les résultats pour $q = 0,\ 0{,}5,\ 1{,}0,\ 1{,}5$ et $G = 0, 10, 100$, en utilisant $n = 3\,446$ obtenu par la méthode NLP. Pour $q =$ 2,0, la répartition de Longford ne dépend pas de G et, en fait, se réduit à la répartition de Neyman (1.2) qui minimise $\mathrm{CV}(\bar{y}_{st})$ pour une valeur n fixe, mais donne des valeurs très élevées de $\mathrm{CV}(\bar{y}_h)$, variant de 16 % à 85 % pour sept provinces. L'examen du tableau 4 montre que $\mathrm{CV}(\bar{y}_h)$, pour une valeur donnée de q, augmente rapidement avec G, tandis que $\mathrm{CV}(\bar{y}_{st})$ diminue lentement à mesure que G augmente et, en fait, est presque constant ($\approx 5{,}1$ %) pour $G > 100$ (valeurs non présentées ici). En outre, $\mathrm{CV}(\bar{y}_h)$, pour une valeur donnée de G, augmente rapidement à mesure que q augmente, tandis que $\mathrm{CV}(\bar{y}_{st})$ diminue. La répartition de Langford pour $q \geq 0{,}5$ et/ou $G \geq 10$, produit des valeurs de $\mathrm{CV}(\bar{y}_h)$ significativement plus élevées que la tolérance spécifiée $\mathrm{CV}_{0h} = 15$ % pour plusieurs provinces, même si $\mathrm{CV}(\bar{y}_{st})$ respecte la tolérance spécifiée de 6 %. Par ailleurs, pour $q = 0$ et $G = 0$, $\mathrm{CV}(\bar{y}_h)$ est inférieur à la tolérance spécifiée, sauf pour la Colombie-Britannique pour laquelle il est de 15,7 %, mais $\mathrm{CV}(\bar{y}_{st}) =$ 7,3 % et est significativement supérieur à la tolérance spécifiée. Pour $q = 1{,}0$ et $q = 1{,}5$, $\mathrm{CV}(\bar{y}_{st})$ demeure inférieur à 6 % quand $G = 0$, mais $\mathrm{CV}(\bar{y}_h)$ dépasse 15 % pour six provinces, variant de 17,7 % à 34,0 % pour $q = 1{,}0$ et de 22,0 % à 54,6 % pour $q = 1{,}5$. Dans l'ensemble, le tableau 4 donne à penser qu'il n'est possible de trouver aucune combinaison appropriée de q et G telle que toutes les exigences de fiabilité spécifiée soient remplies, même approximativement.

5.2 Moyennes de strate : estimateurs composites

Nous présentons maintenant certains résultats pour les estimateurs composites, $\hat{\theta}_h$, des moyennes de strate. Nous avons obtenu la taille totale optimale d'échantillon égale à $n = 3\,368$ en utilisant la procédure NLP et les exigences de fiabilité (3.6). Cette valeur est un peu plus faible que la valeur optimale $n^0 = 3\,446$ obtenue pour les estimateurs directs. Dans le cas de la répartition de Longford, nous avons utilisé $n = 3\,368$ et calculé la répartition de l'échantillon et les CV associés des estimateurs composites $\hat{\theta}_h$ et de la moyenne pondérée $\bar{y}_{st}$ pour des valeurs spécifiées de q et G, en contraignant n_h d'être au moins égal à deux. Pour les données de l'EMCD que nous avons utilisées, le premier terme de (3.5) est petit comparativement au deuxième. Par conséquent, la répartition de l'échantillon est uniforme pour les diverses valeurs de G pour une valeur donnée de q, ce qui signifie que, pour la répartition de Longford, les CV ne varient pas de manière significative en fonction de G.

Par conséquent, nous ne présentons au tableau 5 que les résultats pour $G = 0$ et $q = 0,\ 0{,}5,\ 1{,}0$ et 1,5. Nous constatons, en examinant ce tableau, que $\mathrm{CV}(\hat{\theta}_h)$ diminue en même temps que q pour les deux plus grandes provinces (Québec et Ontario) parce que l'échantillon se déplace des provinces plus petites vers ces deux provinces à mesure que q augmente. En outre, $\mathrm{CV}(\hat{\theta}_h)$ diminue au départ pour l'Alberta et la Colombie-Britannique, mais commence à augmenter quand q est grand, parce que l'échantillon commence à se déplacer de ces provinces vers le Québec et l'Ontario également. En outre, $\mathrm{CV}(\hat{\theta}_h)$ augmente avec q pour toutes les autres provinces, sauf la Nouvelle-Écosse pour laquelle il commence à diminuer quand la valeur de q est grande, en raison d'une composante synthétique plus importante et d'un biais très négligeable. En particulier, $\mathrm{CV}(\hat{\theta}_h)$ augmente rapidement pour Terre-Neuve et l'Île-du-Prince-Édouard à cause du biais très important.

Par ailleurs, $\mathrm{CV}(\bar{y}_{st})$ diminue au départ quand q augmente, mais commence à augmenter quand q est grand parce que la majeure partie de l'échantillon est affectée au Québec et à l'Ontario et qu'une très petite part est affectée aux provinces plus petites. Le tableau 5 suggère que la répartition de Longford ne donne d'assez bons résultats que pour $q = 0$ et $G = 0$, les valeurs de $\mathrm{CV}(\hat{\theta}_h)$ étant inférieures à 15 % pour toutes les provinces, au prix d'une valeur de $\mathrm{CV}(\bar{y}_{st})$ égale à 8,0 %.

Tableau 4
CV (%) pour la répartition de Longford pour q = 0, 0,5, 1,0 et 1,5

Province	q = 0			q = 0,5			q = 1,0			q = 1,5		
	G = 0	G = 10	G = 100	G = 0	G = 10	G = 100	G = 0	G = 10	G = 100	G = 0	G = 10	G = 100
T.-N.	13,5	19,3	29,7	17,2	23,0	33,4	22,7	29,0	38,3	30,4	36,2	40,6
Î.-P.-É.	12,7	20,4	34,6	21,4	29,6	48,5	34,0	45,4	67,3	54,6	67,3	85,6
N.-B.	12,0	17,1	25,0	14,5	19,4	26,8	18,3	23,1	29,0	23,5	27,6	30,3
N.-É.	11,1	16,7	25,5	14,2	19,5	27,9	18,7	24,1	30,9	24,9	29,4	32,8
Qc	11,0	9,8	9,1	9,9	9,4	9,0	9,2	9,0	8,9	8,9	8,9	8,8
Ont.	14,9	9,8	8,7	12,3	9,5	8,6	10,5	9,1	8,5	9,3	8,7	8,5
Man.	12,7	17,6	24,3	14,7	19,1	25,2	17,7	21,9	26,5	22,0	25,4	27,5
Sask.	13,6	18,9	25,9	15,7	20,5	26,9	19,0	23,5	28,3	23,5	27,0	29,4
Alb.	13,5	15,7	16,1	13,3	15,2	15,9	13,6	15,2	15,9	14,6	15,5	15,9
C.-B.	15,7	16,1	15,4	14,7	15,4	15,3	14,3	15,0	15,1	14,5	15,0	15,1
CA	7,3	5,5	5,1	6,2	5,3	5,1	5,5	5,2	5,1	5,2	5,1	5,1

Tableau 5
CV (%) pour les estimateurs composites en utilisant la répartition de Longford : $G = 0$ et $q = 0$, 0,5, 1,0 et 1,5

Province	$q = 0$	$q = 0{,}5$	$q = 1{,}0$	$q = 1{,}5$
T.-N.	12,7	17,0	24,2	37,3
Î.-P.-É.	12,4	23,8	46,0	112,2
N.-B.	10,4	12,8	16,1	20,4
N.-É.	9,4	11,9	14,5	11,7
Qc	10,3	9,0	8,3	8,0
Ont.	13,9	11,1	9,3	8,2
Man.	11,2	13,1	16,0	20,3
Sask.	12,4	14,6	17,9	23,2
Alb.	11,4	11,2	11,5	12,2
C.-B.	14,4	13,3	12,9	13,1
CA	8,0	6,3	5,4	5,6

5.3 Moyennes de domaine

Dans le Registre des entreprises du Canada, les établissements sont classés selon l'industrie en utilisant le Système de classification des industries de l'Amérique du Nord (SCIAN). Ce dernier est avant tout un système de classification des établissements destiné à la compilation de statistiques sur la production. L'industrie associée à chaque établissement inscrit dans le Registre des entreprises du Canada est désignée par un code à six chiffres conforme au SCIAN. En tout, 67 codes à six chiffres sont associés au secteur du commerce de détail. Ces codes sont regroupés en 19 groupes de commerce (GC) aux fins de la publication des données.

Nous avons considéré les groupes de commerce comme des domaines recoupant les provinces (strates). Le groupe de commerce comprenant le plus petit nombre d'établissements est le CG 110 (magasins de bière, de vin et de spiritueux) qui compte 307 établissements et le groupe de commerce comprenant le plus grand nombre d'établissements est le CG 100 (dépanneurs et magasins d'alimentation spécialisés) qui regroupe 7 752 établissements. Les établissements ont été codés en fonction des 19 groupes de commerce dans toutes les provinces sauf une, ceux de l'Île-du-Prince-Édouard ayant été codés en ne considérant que 16 groupes de commerce seulement.

Nous avons appliqué la procédure NLP fondée sur (4.5), (4.6) et (4.7), et obtenu l'augmentation de la taille totale optimale d'échantillon requise pour remplir les exigences de fiabilité spécifiée pour les estimateurs de domaine $_d\hat{\bar{Y}}$. Nous avons constaté qu'aucune augmentation de la taille totale de l'échantillon n'est nécessaire si la tolérance appliquée à $\mathrm{CV}(_d\hat{\bar{Y}})$ est inférieure ou égale à 30 % pour chaque domaine. Si la tolérance est réduite à 25 %, l'augmentation de la taille totale optimale de l'échantillon est de 622, ce qui donne une taille totale d'échantillon après l'augmentation de 4 068. Si la tolérance est abaissée à 20 %, l'augmentation de la taille totale optimale d'échantillon est de 2 100 et la taille totale de l'échantillon après l'augmentation est de 5 546, c'est-à-dire une valeur considérablement plus grande que la taille originale de 3 446. Il convient de souligner les CV des moyennes de strate $\bar{y}_h$ et de la moyenne d'échantillon pondérée $\bar{y}_{st}$ diminuent à mesure que la taille totale d'échantillon augmente.

6. Résumé et conclusion

Nous avons proposé une méthode par programmation non linéaire (NLP) de répartition de l'échantillon entre les strates sous échantillonnage aléatoire stratifié. Cette méthode minimise la taille totale de l'échantillon sous la contrainte des tolérances spécifiées pour les coefficients de variation des estimateurs des moyennes de strate et de la moyenne de population. Nous avons considéré des estimateurs directs ainsi que des estimateurs composites des moyennes de strate. Le cas de domaines recoupant les strates est également étudié. Les difficultés que posent d'autres méthodes en vue de satisfaire aux contraintes de fiabilité spécifiées sont illustrées au moyen de données provenant de l'Enquête mensuelle sur le commerce de détail auprès d'établissements uniques menée par Statistique Canada. Nous constatons aussi que la méthode NLP peut être étendue facilement en vue de traiter des contraintes de fiabilité pour des variables multiples. Les répartitions intermédiaires qui donnent d'assez bons résultats en ce qui concerne les contraintes de fiabilité sont également mentionnées.

Remerciements

Les auteurs remercient deux examinateurs et un rédacteur associé pour leurs commentaires constructifs et suggestions.

Bibliographie

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête*, 15, 49-60.

Bankier, M. (1988). Power allocation: Determining sample sizes for sub-national areas. *The American Statistician*, 42, 174-177.

Cochran, W.G. (1977). *Sampling Techniques*, 3e Édition. New York : John Wiley & Sons, Inc.

Costa, A., Satorra, A. et Ventura, E. (2004). Using composite estimators to improve both domain and total area estimation. *SORT*, 28, 69-86.

Huddleston, H.F., Claypool, P.L. et Hocking, R.R. (1970). Optimum allocation to strata using convex programming. *Applied Statistics*, 19, 273-278.

Longford, N.T. (2006). Calcul de la taille de l'échantillon pour l'estimation pour petits domaines. *Techniques d'enquête*, 32, 97-106.

North American Industry Classification System, Version 1.4 (2008). Catalogue 12F0074XCB. Statistics Canada.

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 33-44
Statistique Canada, N° 12-001-X au catalogue

Options de calage pour remplacer la poststratification de données croisées à deux dimensions

Ted Chang [1]

Résumé

Nous considérons diverses options de remplacement de la poststratification dans le cas de données croisées à deux dimensions pour lesquelles la fréquence d'au moins l'une des cellules du tableau à double entrée est trop faible pour permettre la poststratification en se basant sur cette classification double. Dans le jeu de données utilisé pour l'étude, la fréquence prévue dans la cellule la plus petite est de 0,36. Une approche consiste simplement à regrouper certaine cellules. Néanmoins, il est probable que cela détruise la structure de la classification double. Les approches de remplacement que nous proposons permettent de maintenir la classification à double entrée originale des données. Elles sont fondées sur l'étude du calage de Chang et Kott (2008). Nous choisissons des ajustements des pondérations qui dépendent des classifications marginales (mais non de la classification croisée complète) pour minimiser une fonction d'objectif correspondant aux différences entre les fréquences dans la population et dans les cellules du tableau à double entrée et leurs estimations sur échantillon. Selon la terminologie de Chang et Kott (2008), si les classifications par ligne et par colonne comprennent I et J cellules, respectivement, on obtient IJ variables *de référence* et $I + J - 1$ variables *du modèle*. Nous étudions les propriétés de ces estimateurs en créant des échantillons aléatoires simples pour la simulation d'après les données du Quarterly Census of Employment and Wages de 2005 tenues à jour par le Bureau of Labor Statistics. Nous procédons à la classification croisée de l'État et du groupe d'industries. Dans notre étude, comparativement à l'estimateur sans biais, les approches de calage introduisent un biais asymptotiquement négligeable, mais donnent une réduction de l'EQM pouvant aller jusqu'à 20 % pour un petit échantillon.

Mots clés : Calage ; poststratification ; modèle de prédiction.

1. Introduction

Supposons que nous ayons une population $\mathcal{U}$ doublement stratifiée par deux variables catégoriques désignées par les indices $(i, j), i = 1, ..., I, j = 1, ..., J$ et écrivons $\mathcal{U}_{ij}$ pour la strate (i et j). Si nous tirons un échantillon aléatoire simple $\mathcal{S}$ de taille n et que nous désignons par y la variable d'intérêt, un estimateur naturel du total $T_y = \sum_{k \in \mathcal{U}} y_k$ est l'estimateur poststratifié

$$\hat{t}_{y\text{PS}} = \sum_{i,j} N_{ij} \, \bar{y}_{ij} \qquad (1)$$

où N_{ij} est la taille de la population $\mathcal{U}_{ij}$ et $\bar{y}_{ij}$ est la moyenne d'échantillon de y sur $\mathcal{S} \cap \mathcal{U}_{ij}$. Cet estimateur est généralement utilisé à condition que toutes les tailles d'échantillon n_{ij} de $\mathcal{S} \cap \mathcal{U}_{ij}$ soient raisonnablement grandes.

Que faire si certaines tailles n_{ij} sont petites, ou même nulles ?

L'approche classique consiste à regrouper certaines cellules jusqu'à ce que toutes les tailles n_{ij} soient suffisamment grandes. Cependant, il pourrait être impossible d'effectuer ce groupement d'une façon qui maintienne le schéma de classification croisée : autrement dit, les indices j pourraient dépendre de i.

L'estimateur poststratifié $\hat{t}_{y\text{PS}}$ est un cas particulier d'un estimateur par calage. Définissons pour chaque $k \in \mathcal{U}$ le vecteur de variables $\mathbf{x}_k = (x_{11k}, \ldots, x_{IJk})^T$ de dimension $I \times J$ où $x_{ijk} = 1$ si $k \in \mathcal{U}_{ij}$ et $x_{ijk} = 0$ autrement. Le total de population $T_{\mathbf{x}}$ de $\mathbf{x}$ est $(N_{11}, \ldots, N_{IJ})^T$ et en supposant que $d_k = N / n$ est le poids d'échantillonnage et que $\beta = (N^{-1} n_{11}^{-1} N_{11} n, \ldots, N^{-1} n_{IJ}^{-1} N_{IJ} n)^T$

$$\hat{t}_{y\text{PS}} = \sum_{k \in \mathcal{S}} d_k (\mathbf{x}_k^T \beta) \, y_k$$

$$T_{\mathbf{x}} = \sum_{k \in \mathcal{S}} d_k (\mathbf{x}_k^T \beta) \mathbf{x}_k .$$

Ces équations établissent que si l'on utilise les variables $\mathbf{x}$ *de référence*, alors $\hat{t}_{y\text{PS}}$ est l'estimateur par calage résultant de T_y.

Chang et Kott (2008) ont déterminé les propriétés asymptotiques de l'estimation calée de la forme

$$\hat{t}_{y, \mathbf{z} f \mathbf{V}} = \sum_{k \in \mathcal{S}} d_k \, f(\mathbf{z}_k^T \hat{\beta}) \, y_k \qquad (2)$$

où $\hat{\beta}$ minimise une fonction d'objectif de la forme

$$Q(\beta) = \left(T_{\mathbf{x}} - \sum_{k \in \mathcal{S}} d_k \, f(\mathbf{z}_k^T \beta) \mathbf{x}_k \right)^T \mathbf{V}^{-1} \left(T_{\mathbf{x}} - \sum_{k \in \mathcal{S}} d_k \, f(\mathbf{z}_k^T \beta) \mathbf{x}_k \right). \qquad (3)$$

Dans les équations (2) et (3), $\mathbf{z}$ est un vecteur de variables du *modèle* dont la longueur Q est au plus égale à la

1. Ted Chang, University of Virginia, Department of Statistics, Charlottesville, VA, États-Unis. Courriel : tcc8v@virginia.edu.

longueur P du vecteur de variables *de référence* $\mathbf{x}$, f est une fonction à valeur réelle positive que Chang et Kott (2008) appellent *back link function* (fonction de *lien rétro-transformée* ou fonction de *lien inverse*), et $\mathbf{V}$ est une matrice définie positive symétrique de dimensions $P \times P$. $\mathbf{V}$ peut dépendre de β, comme cela se produirait si $\mathbf{V}(\beta)$ était une mesure de la variabilité de $\sum_{k \in \mathcal{S}} d_k \, f(\mathbf{z}_k^T \beta)\mathbf{x}_k$.

Dans Chang et Kott (2008), l'échantillon réalisé $\mathcal{S}$ est constitué des répondants d'un échantillon original auquel sont associés les poids d'échantillonnage d_k. L'échantillon de répondants $\mathcal{S}$ est supposé être un sous-échantillon de Poisson de l'échantillon original dont les probabilités de sélection poissonniennes sont $f(\mathbf{z}_k^T \beta_0)^{-1}$, pour un certain β_0. Les formules asymptotiques dérivées par Chang et Kott l'ont été sous un cadre asymptotique pour ce modèle de quasi-randomisation (fondé sur le plan). Nous utilisons le terme *quasi-randomisation* afin de nous rappeler que le mécanisme de réponse de Poisson supposé est, en fait, fondé sur un modèle.

Il convient de souligner que l'utilisation du calage pour corriger la non-réponse remonte à Fuller, Loughin et Baker (1994), du moins quand $\mathbf{z} = \mathbf{x}$ et $f(\eta) = 1 + \eta$.

Nous proposons d'utiliser la méthodologie de Chang et Kott (2008) en gardant $\mathbf{x}$ comme variables indicatrices pour la classification croisée $I \times J$ complète, mais en supposant que $\mathbf{z}$ est un vecteur de $I + J - 1$ variables indicatrices pour les classifications marginales. Autrement dit, nous proposons de rééquilibrer l'échantillon afin de nous rapprocher autant que possible, au sens de la minimisation de (3), des proportions de cellule correctes dans la classification complète, mais en exigeant que les poids de rééquilibrage ne dépendent que des classifications marginales.

Le cadre de Chang et Kott (2008) est applicable en présence de non-réponse (et/ou de non-couverture) si $f(\mathbf{z}_k^T \beta_0)^{-1}$ est la probabilité de réponse (ou de réponse et de couverture) combinée. Nous notons que la poststratification, qui est un cas particulier de calage, est souvent utilisée en vue de corriger la non-réponse ou la non-couverture. Dans l'exemple test que nous présentons plus bas, aucune non-réponse ni non-couverture ne doit être corrigée , et le cadre de Chang et Kott (2008) s'applique donc avec $\beta_0 = \mathbf{0}$ pour toute fonction f avec $f(0) = 1$. Autrement dit, si le calage est utilisé uniquement pour rééquilibrer l'échantillon, nous pouvons utiliser le cadre de Chang et Kott (2008) avec pratiquement n'importe quelle fonction f. Mais si nous essayons de corriger la non-réponse et/ou la non-couverture, des hypothèses plus fortes sont nécessaires.

Il convient de souligner que le ratissage (*raking*) est simplement l'estimation calée en utilisant les $I + J - 1$ variables indicatrices des classifications marginales comme variables de référence ainsi que variables du modèle, et en utilisant $f(\eta) = e^{\eta}$. Donc, nous examinerons également l'utilisation de cette fonction de lien rétrotransformée.

À la section 2, nous présentons les formules exactes pour les estimateurs que nous utiliserons dans la présente étude. Le cadre de Chang et Kott (2008) peut être appliqué pour établir les estimateurs de variance fondés sur l'échantillon et ces calculs sont exposés en annexe.

À la section 3, nous donnons les résultats d'une étude empirique réalisée en utilisant les données du Quarterly Census of Employment and Wages pour le premier trimestre de 2005 recueillies par le Bureau of Labor Statistics. Nous nous limiterons à cinq États que nous désignerons par A, B, C, D, E et à cinq groupes d'industries que nous désignerons par 1, 2, 3, 4, 5. Nous n'identifierons pas de manière plus précise les États ni les groupes d'industries afin d'empêcher l'identification de l'unité aberrante dans la discussion qui suit. Cette population comprend 283 725 entreprises. De cette population, nous tirerons par une méthode de Monte Carlo des échantillons aléatoires simples de taille $n = 200$, 1 000 et 5 000, et procéderons à la classification croisée de l'État par le groupe d'industries.

Il convient de souligner que 0,18 % de la population correspond à la classification croisée de l'État E et du groupe d'industries 5. Donc, quand $n = 200$, la taille prévue d'échantillon dans cette cellule est de 0,36, de sorte qu'il est hors de question de procéder à la poststratification en se basant sur la classification croisée.

Kott et Chang (2010) dérivent les propriétés de $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ en utilisant un cadre fondé sur un modèle. Les modèles qu'ils prennent en considération ne s'appliquent pas à notre choix des variables $\mathbf{x}$ et $\mathbf{z}$. Cependant, motivés par leur approche, nous examinons à la section 4 le comportement de l'estimateur $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ défini par l'équation (2), sous des hypothèses très simplifiées, y compris $f(\eta) = 1 + \eta$. Cela nous mène, à la section 5, au choix d'une nouvelle matrice de pondération $\mathbf{V}^{-1}$ à utiliser dans (3). Nous poursuivons ensuite notre exploration empirique en utilisant ce nouvel estimateur.

2. Formules mathématiques

À la présente section, nous donnons les formules utilisées dans l'étude. Toutes sont des cas particuliers des formules présentées dans Chang et Kott (2008). Nous supposons qu'un échantillon aléatoire simple de taille n est tiré d'une population de taille N et nous utilisons $\mathcal{S}$ et r pour désigner l'échantillon des répondants et sa taille. Nous supposons que la fonction des poids de calage possède un β_0 tel que $f(\mathbf{z}^T \beta_0)^{-1}$ est la probabilité de réponse pour un élément ayant les variables du modèle $\mathbf{z}$. En particulier, et sans perte de généralité, s'il n'existe aucun problème de non-réponse, nous supposons que $f(0) = 1$.

Les mêmes formules marchent pour la non-couverture, auquel cas $f(\mathbf{z}^T\beta_0)^{-1}$ est la probabilité combinée de réponse et de couverture.

Nous désignons par N_{ij}, $\mathcal{S}_{ij}$, et r_{ij} la taille de la population, l'échantillon de répondants et la taille de l'échantillon de répondants dans la cellule (i, j). Bien que nous supposions que N_{ij} est connu, notre méthodologie n'exige pas que les classifications par ligne et par colonne des non-répondants soient connues.

Nous définissons $N_{i\cdot} = \sum_j N_{ij}$ et, de manière analogue, nous définissons $N_{\cdot j}$.

Nous utiliserons les estimateurs d'un total T_y de la forme

$$\hat{t}_y = \frac{N}{n}\sum_i\sum_j\sum_{k\in\mathcal{S}_{ij}} w_{ij}\, y_k \qquad (4)$$

où les poids d'ajustement w_{ij} sont définis comme il est indiqué plus bas. Ces estimateurs sont tous des cas particuliers des équations (2) et (3) quand nous utilisons $\mathbf{V} = \mathbf{I}$.

Dans l'estimateur *à marges calées*, nous utilisons $f(\eta) = 1 + \eta$ et définissons $\mathbf{x} = \mathbf{z}$ comme étant $I + J - 1$ variables indicatrices indépendantes pour les catégories de marge. Dans ce cas, $T_{\mathbf{x}}$ est un vecteur de $N_{i\cdot}$ et $N_{\cdot j}$. Les poids d'ajustement $f(\mathbf{z}_k^T\beta)$ sont de la forme $w_{ij} = 1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}$ quand $\mathbf{z}$ est le vecteur des variables indicatrices de l'appartenance à la $(i, j)^{\text{e}}$ classification par ligne et colonne, respectivement. Puisque le nombre d'équations (la dimension de $\mathbf{x}$) est égal au nombre d'inconnues (la dimension de $\hat{\beta}$), nous devrions être capables de résoudre exactement les équations

$$T_{\mathbf{x}} = \sum_{k\in\mathcal{S}} d_k\, f(\mathbf{z}_k^T\beta)\mathbf{x}_k. \qquad (5)$$

Donc les $\hat{\beta}_{i\cdot}, \hat{\beta}_{\cdot j}$ résolvent les équations linéaires de rang $I + J - 1$

$$N_{i\cdot} = \frac{N}{n}\sum_j (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})\, r_{ij}$$
$$N_{\cdot j} = \frac{N}{n}\sum_i (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})\, r_{ij},$$

ce qui découle facilement de (5).

Dans l'estimateur *à fréquences de cellule calées*, nous utilisons $f(\eta) = 1 + \eta$ et définissons $\mathbf{x}$ comme étant les IJ variables indicatrices pour la classification croisée complète et $\mathbf{z}$ comme étant les $I + J - 1$ variables indicatrices indépendantes pour les catégories de marge. Dans ce cas, $T_{\mathbf{x}}$ est un vecteur de N_{ij} et, puisque $\mathbf{V} = \mathbf{I}$, les poids d'ajustement $w_{ij} = 1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}$ minimisent la fonction d'objectif

$$\sum_i\sum_j \left[N_{ij} - \frac{N}{n}\sum_i\sum_j (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})\, r_{ij} \right]^2.$$

Dans l'estimateur par *ratissage* (*raking*), nous utilisons $f(\eta) = e^{\eta}$ et définissons $\mathbf{x} = \mathbf{z}$ comme étant les $I + J - 1$ variables indicatrices indépendantes pour les catégories de marge. Les poids d'ajustement sont $w_{ij} = \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})$ où les $\hat{\beta}_{i\cdot}, \hat{\beta}_{\cdot j}$ résolvent les $I + J$ équations

$$N_{i\cdot} = \frac{N}{n}\sum_j \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})\, r_{ij}$$
$$N_{\cdot j} = \frac{N}{n}\sum_i \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})\, r_{ij}.$$

Puisque $\sum_i N_{i\cdot} = N = \sum_j N_{\cdot j}$, ces $I + J$ équations donnent seulement $I + J - 1$ contraintes. Il convient toutefois de souligner que, si une constante c est ajoutée à chaque $\hat{\beta}_{i\cdot}$ et est soustraite de chaque $\hat{\beta}_{\cdot j}$, les w_{ij} ne changent pas.

Dans l'estimateur *à fréquences de cellule calées exponentiel*, nous utilisons $f(\eta) = e^{\eta}$ et définissons $\mathbf{x}$ comme étant les IJ variables indicatrices pour la classification croisée complète et $\mathbf{z}$ comme étant les $I + J - 1$ variables indicatrices indépendantes pour les catégories de marge. Les poids d'ajustement $w_{ij} = \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})$ minimisent la fonction d'objectif

$$\sum_i\sum_j \left[N_{ij} - \frac{N}{n}\sum_i\sum_j \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})\, r_{ij} \right]^2.$$

Chang et Kott (2008) donnent les formules de l'estimation de la variance de $\hat{t}_y$, fondée sur l'échantillon. À l'annexe, nous appliquons ces formules aux quatre estimateurs susmentionnés.

3. Étude empirique

Nous utilisons comme population les données provenant du Quarterly Census of Employment et Wages (QCEW) recueillies pour le premier trimestre de 2005, en nous limitant à cinq États et cinq groupes d'industries. Le QCEW est compilé d'après des rapports qui doivent obligatoirement être soumis aux bureaux de l'emploi des États, de sorte qu'il s'agit pratiquement d'un recensement, et nous avons utilisé les données du QCEW complet pour les cinq États et les cinq groupes d'industries choisis. Cette population compte $N = 283\,725$ entreprises réparties comme il est indiqué au tableau 1.

Les variables réponse y sont l'emploi total et la rémunération (trimestrielle) totale. Pour ces variables $T_y = 2\,981\,364$ pour l'emploi total et $T_y = 2\,334\,400$ (en dizaines de milliers de dollars) pour la rémunération totale. Dans la présente étude, nous avons tiré 10 000 échantillons de taille $n = 200$, 1 000 et 5 000. Pour chacun des quatre estimateurs, nous présentons des estimations du biais, de l'erreur-type et de la racine carrée de l'erreur quadratique moyenne. Nous donnons aussi la racine carrée de la

moyenne des variances estimées en utilisant le premier terme de l'équation (15). À titre comparatif, nous présentons les valeurs théoriques et empiriques de l'estimateur non pondéré $N / n \sum_{k \in \mathcal{S}} y_k$. Ces résultats figurent au tableau 2 pour l'emploi total et au tableau 3 pour la rémunération totale.

Pour l'échantillon de taille $n = 5\,000$, la taille prévue d'échantillon dans la cellule la plus petite (État E et groupe d'industries 5) est de 9,07. Bien que cela soit un peu faible pour la poststratification, la probabilité que cette cellule ait une taille d'échantillon inférieure à 2, c'est-à-dire la taille minimale nécessaire pour l'estimation de la variance, est de 0,0011. Dans nos simulations, neuf exécutions ont produit une cellule ayant une taille inférieure à 2. Nous présentons aussi le comportement empirique de l'estimateur poststratifié en excluant les neuf cas posant problème, en utilisant l'estimation de variance (7.6.5) de Särndal, Swensson et Wretman (1992), ainsi que son comportement théorique en utilisant l'approximation de la variance donnée par (7.6.6) dans Särndal et coll. (1992).

Tableau 1
Entités commerciales selon l'État et le groupe d'industries

	Groupe d'industries					
	1	**2**	**3**	**4**	**5**	**somme**
A	5 986	5 548	7 712	3 969	1 299	24 514
	(2,11 %)	(1,96 %)	(2,72 %)	(1,40 %)	(0,46 %)	(8,64 %)
B	18 782	31 572	22 012	4 982	4 504	81 852
	(6,62 %)	(11,13 %)	(7,76 %)	(1,76 %)	(1,59 %)	(28,85 %)
C	13 518	13 099	17 837	5 610	3 001	53 065
	(4,76 %)	(4,62 %)	(6,29 %)	(1,98 %)	(1,06 %)	(18,70 %)
D	30 428	36 017	32 541	10 963	5 399	115 348
	(10,72 %)	(12,69 %)	(11,47 %)	(3,86 %)	(1,90 %)	(40,65 %)
E	2 225	2 020	3 110	1 076	515	8 946
	(0,78 %)	(0,71 %)	(1,10 %)	(0,38 %)	(0,18 %)	(3,15 %)
somme	70 939	88 256	83 212	26 600	14 718	283 725
	(25,00 %)	(31,11 %)	(29,33 %)	(9,38 %)	(5,19 %)	

Tableau 2
Comparaison empirique de quatre estimateurs de l'emploi total

Estimateur	**Biais**	**Erreur-type**	**REQM**	**R.C. var. est.**
	n = 200			
Non pondéré (théorique)	0	1 113 220		
Non pondéré (empirique)	-1 280	1 068 944	1 068 945	1 059 463
Marges calées	-1 394	1 105 201	1 105 201	1 048 873
Fréq. cell. calées	-218 751	1 008 436	1 031 889	975 140
Ratissage	-462	1 103 172	1 103 172	1 041 490
Fréq. cell. calées exponentiel	-227 578	1 000 154	1 025 719	962 153
	n = 1 000			
Non pondéré (théorique)	0	497 144		
Non pondéré (empirique)	-5 435	505 941	505 970	501 144
Marges calées	-6 212	506 239	506 277	498 946
Fréq. cell. calées	-56 118	493 611	496 790	488 222
Ratissage	-4 854	507 938	507 961	499 237
Fréq. cell. calées exponentiel	-58 891	492 939	496 445	487 281
	n = 5 000			
Non pondéré (théorique)	0	220 751		
Non pondéré (empirique)	1 516	224 088	224 093	222 034
Poststratifié (théorique)	0	220 315		
Poststratifié (empirique, 9 cas exclus)	1 234	223 225	223 228	221 094
Marges calées	1 649	223 091	223 098	220 833
Fréq. cell. calées	-8 606	222 170	222 337	220 347
Ratissage	3 632	236 355	236 383	220 606
Fréq. cell. calées exponentiel	-10 643	223 472	223 725	220 207

Tableau 3
Comparaison empirique de quatre estimateurs de la rémunération totale (dizaines de milliers de dollars)

Estimateur	Biais	Erreur-type	REQM	R.C. var. est.
	n = 200			
Non pondéré (théorique)	0	1 682 571		
Non pondéré (empirique)	-11 119	1 551 186	1 551 226	1 543 483
Marges calées	-11 474	1 582 383	1 582 425	1 510 413
Fréq. cell. calées	-214 323	1 451 931	1 467 664	1 413 411
Ratissage	-11 220	1 579 842	1 579 882	1 501 170
Fréq. cell. calées exponentiel	-221 435	1 438 810	1 455 750	1 393 246
	n = 1 000			
Non pondéré (théorique)	0	751 406		
Non pondéré (empirique)	-2 911	772 495	772 501	768 878
Marges calées	-4 372	776 955	776 968	768 869
Fréq. cell. calées	-51 649	756 201	757 963	751 384
Ratissage	-4 684	778 302	778 316	769 428
Fréq. cell. calées exponentiel	-54 305	754 963	756 913	749 832
	n = 5 000			
Non pondéré (théorique)	0	333 654		
Non pondéré (empirique)	2 678	336 057	336 068	337 239
Poststratifié (théorique)	0	333 765		
Poststratifié (empirique, 9 cas exclus)	1 802	335 271	335 276	336 192
Marges calées	2 510	334 910	334 920	336 064
Fréq. cell. calées	-7 149	333 560	333 637	335 006
Ratissage	-4 679	339 074	339 106	335 230
Fréq. cell. calées exponentiel	-9 251	334 365	334 493	334 755

Les distributions des variables réponse, c'est-à-dire l'emploi total et la rémunération totale, sont fortement asymétriques avec étalement vers la droite. Il existe une entreprise (dans l'État C et le groupe d'industries 4) dont l'emploi total est égal à plus du double de l'emploi total de l'entreprise suivante par ordre décroissant de taille et à plusieurs centaines de fois l'emploi moyen des entreprises restantes. Nous avons répété l'étude en utilisant une population dont cette entreprise a été supprimée. Les résultats sont présentés aux tableaux 4 et 5. En pratique, pour cette population, l'entreprise en question serait normalement échantillonnée avec certitude (*unité auto-représentative*) et les échantillons seraient construits en partant des entreprises restantes. Donc, les tableaux 4 et 5 reflètent peut-être davantage les propriétés relatives des estimateurs dans la pratique réelle.

Les échantillons utilisés pour les tableaux 4 et 5 sont identiques à ceux utilisés pour les tableaux 2 et 3, excepté que, dans les cas où l'unité aberrante était incluse dans l'échantillon, elle a été remplacée par une nouvelle observation provenant de la population. Ce remplacement a été effectué afin de rendre les résultats des tableaux 4 et 5 plus comparables à ceux des tableaux 2 et 3.

L'examen des tableaux 2 et 3 révèle que les méthodes pour lesquelles $P > Q$, c'est-à-dire celles consistant à caler les fréquences de cellule de la classification croisée en utilisant des poids de calage qui dépendent des classifications marginales, présentent manifestement un biais plus important que les autres. Cependant, relativement aux écarts-types, les biais de ces estimateurs diminuent quand la taille de l'échantillon augmente. Nous montrerons à la prochaine section que, sous un modèle très simplifié, le biais est d'ordre n^{-1} et l'écart-type est d'ordre $n^{-1/2}$. Considérons, par exemple, les résultats pour l'estimateur *à fréquences de cellule calées* du tableau 2. Dans ce cas, le biais divisé par l'erreur-type est égal à 0,217, 0,114 et 0,039 pour $n = 200$, 1 000 et 5 000, respectivement. Pour ces valeurs de n, les valeurs de $n^{-1/2}$ sont 0,0707, 0,0316 et 0,0141, et il semble que la première série de trois quotients est approximativement égale à trois fois la seconde série.

Il semble aussi que la fonction de lien rétrotransformée f exponentielle donne de légèrement de meilleurs résultats que le choix linéaire. Du point de vue de l'intensité des calculs, la première est beaucoup plus coûteuse que la seconde. Nous constatons aussi que, quand la taille de l'échantillon augmente, les propriétés des estimateurs semblent converger. Cela n'est pas inattendu, parce qu'en l'absence de non-réponse, quand $n \to \infty$, $\hat{\beta} \to 0$, de sorte que les poids d'ajustement $w = f(\mathbf{z}^T \hat{\beta}) \to 1$.

Si l'on compare l'estimateur à fréquences de cellule calées linéaire aux valeurs empirique de l'estimateur non pondéré, quand $n = 200$, le premier est, selon l'EQM, environ 7,3 % plus efficace pour l'emploi total et 11,7 % plus efficace pour la rémunération totale. (Cela signifie, par exemple, que l'EQM empirique de l'estimateur non pondéré est égale à 1,117 fois l'EQM empirique de l'estimateur à fréquences de cellule calées linéaire quand on estime la rémunération totale.) Pour l'estimateur à fréquences de

cellule calées exponentiel, l'accroissement de l'efficacité comparativement à l'EQM empirique de l'estimateur non pondéré est de 8,6 % pour l'emploi total et de 13,5 % pour la rémunération totale. La comparaison aux valeurs théoriques pour l'estimateur pondéré penche en faveur des estimateurs à fréquences de cellule calées, mais nous utilisons les résultats empiriques obtenus pour l'estimateur non pondéré, car les mêmes échantillons Monte Carlo ont été employés pour les divers estimateurs. L'estimateur à fréquences de cellule calées et l'estimateur à fréquences de cellule calées exponentiel continuent d'être meilleurs que l'estimateur sans biais en ce qui concerne l'EQM pour une taille d'échantillon de $n = 1\,000$.

Tableau 4
Comparaison empirique de quatre estimateurs de l'emploi total : population avec suppression de l'unité aberrante.

Estimateur	Biais	Erreur-type	REQM	R.C. var. est.
		$n = 200$		
Non pondéré (théorique)	0	950 688		
Non pondéré (empirique)	5 395	975 617	975 632	965 448
Marges calées	5 777	1 019 583	1 019 599	963 314
Fréq. cell. calées	-211 568	909 070	933 365	877 343
Ratissage	6 688	1 018 383	1 018 405	956 867
Fréq. cell. calées exponentiel	-217 810	902 756	928 660	868 797
		$n = 1\,000$		
Non pondéré (théorique)	0	424 552		
Non pondéré (empirique)	-8 393	422 116	422 199	414 019
Marges calées	-9 430	418 153	418 259	408 577
Fréq. cell. calées	-58 808	408 391	412 603	399 961
Ratissage	-8 135	419 938	420 016	408 611
Fréq. cell. calées exponentiel	-61 014	407 780	412 320	399 311
		$n = 5\,000$		
Non pondéré (théorique)	0	188 517		
Non pondéré (empirique)	702	191 631	191 632	188 089
Poststratifié (théorique)	0	187 691		
Poststratifié (empirique, 9 cas exclus)	563	190 854	190 855	187 180
Marges calées	820	190 662	190 664	186 664
Fréq. cell. calées	-9 376	189 884	190 115	186 202
Ratissage	2 933	205 924	205 944	186 618
Fréq. cell. calées exponentiel	-9 922	189 813	190 072	186 140

Tableau 5
Comparaison empirique de quatre estimateurs de la rémunération totale : population avec suppression de l'unité aberrante

Estimateur	Biais	Erreur-type	REQM	R.C. var. est.
		$n = 200$		
Non pondéré (théorique)	0	1 330 930		
Non pondéré (empirique)	711	1 341 900	1 341 901	1 334 556
Marges calées	1 256	1 387 484	1 387 485	1 318 285
Fréq. cell. calées	-201 575	1 225 852	1 242 314	1 194 071
Ratissage	1 473	1 386 978	1 386 979	1 311 353
Fréq. cell. calées exponentiel	-206 956	1 217 881	1 235 340	1 184 166
		$n = 1\,000$		
Non pondéré (théorique)	0	594 370		
Non pondéré (empirique)	-8 169	587 775	587 832	582 524
Marges calées	-10 093	583 606	583 693	576 251
Fréq. cell. calées	-56 429	569 158	571 948	563 022
Ratissage	-10 529	584 532	584 626	576 282
Fréq. cell. calées exponentiel	-58 435	568 277	571 273	562 061
		$n = 5\,000$		
Non pondéré (théorique)	0	263 923		
Non pondéré (empirique)	1 185	266 779	266 782	264 110
Poststratifié (théorique)	0	263 339		
Poststratifié (empirique, 9 cas exclus)	566	265 973	265 973	263 210
Marges calées	991	265 449	265 451	262 556
Fréq. cell. calées	-8 565	264 126	264 265	261 483
Ratissage	-6 008	271 535	271 602	262 021
Fréq. cell. calées exponentiel	-9 070	264 038	264 194	261 394

Quand la seule unité aberrante est supprimée, laissant 283 724 éléments de la population, les estimateurs à fréquences de cellule calées ont d'un peu meilleures propriétés que l'estimateur sans biais. Pour $n = 200$, l'estimateur à fréquences de cellules calées linéaire donne une amélioration de 9,3 % de l'efficacité pour l'emploi total et de 16,7 % pour la rémunération totale. Les ratios comparables pour l'estimateur à fréquences de cellule calées exponentiel sont de 10,4 % pour l'emploi total et de 18,0 % pour la rémunération totale.

Enfin, l'estimateur de variance donné par l'équation (15) présente un léger biais vers le bas.

4. Biais et variance fondés sur un modèle des estimateurs calés

Kott et Chang (2010) ont établi les priorités asymptotiques de $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ sous une structure probabiliste fondée sur un modèle différente. Dans leur article, $\mathcal{S}$ est un échantillon tiré avec les probabilités de sélection d_k^{-1} de sorte que la non-réponse n'est pas un problème. Si le nombre P de variables de référence $\mathbf{x}$ est égal au nombre Q de variables du modèle $\mathbf{z}$, Kott et Chang (2010) émettent plutôt l'hypothèse d'un modèle de *prédiction*

$$y_k = \mathbf{x}_k^T \theta + \varepsilon_k, k \in \mathcal{U}. \quad (6)$$

Ici, θ est un vecteur fixe inconnu, ε_k représente les erreurs indépendantes du modèle sous la contrainte

$$E(\varepsilon_k \mid \mathbf{z}_j, I_j, j \in \mathcal{U}) = 0, \quad (7)$$

et I_k est une variable aléatoire définie par $I_k = 1$ si $k \in \mathcal{S}$ et $I_k = 0$ autrement.

Quand $P > Q$, l'équation du modèle (6) doit être remplacée par

$$y_k = (\mathbf{A}_\infty \mathbf{x}_k)^T \theta + \varepsilon_k, k \in \mathcal{U} \quad (8)$$

pour une certaine matrice $\mathbf{A}_\infty$ de dimension $Q \times P$ [qui est définie dans un cadre asymptotique approprié, voir Kott et Chang (2010)].

Donc, quand $\mathbf{x}$ représente les variables indicatrices pour la classification croisée complète $I \times J$, nous avons que $\mathbf{x}_k^T\theta$, pour k dans la $(i, j)^{\text{e}}$ cellule, est la valeur moyenne de la variable de réponse sur la $(i, j)^{\text{e}}$ cellule. Donc, par définition, $E(\varepsilon_k \mid \mathbf{x}_j, j \in \mathcal{U}) = 0$ et, puisque $\mathbf{z}$ est une fonction de $\mathbf{x}$, le modèle donné par (6) et (7) est automatiquement vérifié quand l'échantillonnage (y compris la non-réponse) est non informatif.

Cependant, dans notre application de calage, $P = IJ > Q = I + J - 1$ et il n'existe a priori aucune raison que l'équation du modèle (8) soit vérifiée.

Motivés par Kott et Chang (2010), nous examinons le comportement des estimations calées sous le scénario suivant :

1. Les variables de référence $\mathbf{x}$ sont les variables indicatrices pour une partition de la population en classes $\mathcal{C}_r$. Le modèle (6) est automatiquement vérifié, où la r^{e} composante de θ est la moyenne de population de $\mathcal{C}_r$. Soit f_r la proportion de la population dans $\mathcal{C}_r$ et $V_r = \text{Var}(\varepsilon_k \mid k \in \mathcal{C}_r)$. Nous utiliserons aussi la notation $\text{Var}(\mathbf{x}_k)$ pour V_r quand $k \in \mathcal{C}_r$.
2. L'échantillon est un échantillon aléatoire simple de taille n sélectionné *avec remise*.
3. La fonction de lien rétrotransformée $f(\eta)$ dans l'estimateur $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ de l'équation (2) est $f(\eta) = 1 + \eta$.

Bien que ces hypothèses ne soient pas réalistes en pratique, l'objectif principal de la présente section est de justifier de manière heuristique un choix, décrit à la section suivante, pour la matrice $\mathbf{V}$. À ce stade, nous n'imposons aucune contrainte sur $\mathbf{z}$.

Nous notons que, dans cette situation, $E(\varepsilon_k \mid \mathbf{x}_j, I_j, j \in \mathcal{U}) = 0$. Notons que l'équation (7) sera vérifiée si les composantes des variables du modèle $\mathbf{z}$ sont des fonctions de $\mathbf{x}$, c'est-à-dire que chaque composante de $\mathbf{z}$ est constante dans chaque classe. Cependant, si $P > Q$, l'équation (8) ne sera généralement pas vérifiée. De toute façon, à la présente section, nous n'avons besoin ni de (7) ni de (8).

Nous posons que

$$\mu_{\mathbf{x}} = \frac{1}{N}\sum_{j\in\mathcal{U}} \mathbf{x}_j$$

$$\mu_{\mathbf{xz}} = \frac{1}{N}\sum_{j\in\mathcal{U}} \mathbf{x}_j \mathbf{z}_j^T$$

et la matrice $\mathbf{A}_\infty$ de l'équation (8) devient

$$\mathbf{A}_\infty = \mathbf{V}^{-1}\mu_{\mathbf{xz}}.$$

Soit $\hat{\mu}_{y,\mathbf{z}\mathbf{V}} = N^{-1}\hat{t}_{y,\mathbf{z}\mathbf{V}}$ où $\hat{t}_{y,\mathbf{z}\mathbf{V}}$ est défini comme dans (2). Nous avons supprimé f dans la notation $\hat{t}_{y,\mathbf{z}\mathbf{V}}$ parce qu'à la présente section, $f(\eta) = 1 + \eta$. En désignant par $\bar{y}_s$ et $\bar{\mathbf{x}}_s$ les moyennes d'échantillon indiquées et en utilisant Kott et Chang (2010), nous avons

$$\hat{\beta} = \left(\frac{1}{n}\sum_{j\in\mathcal{S}} \mathbf{z}_j \mathbf{x}_j^T \mathbf{A}_\infty\right)^{-1}\left(\frac{1}{n}\sum_{j\in\mathcal{S}} \mathbf{z}_j y_j\right) + O_p(n^{-1/2})$$

$$\begin{aligned}\hat{\mu}_{y,\mathbf{z}\mathbf{V}} &= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{A}_\infty (\mathbf{A}_\infty^T \mathbf{V} \mathbf{A}_\infty)^{-1}\left(\frac{1}{n}\sum_{j\in\mathcal{S}} \mathbf{z}_j y_j\right) + O_p(n^{-1}) \\ &= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{A}_\infty (\mathbf{A}_\infty^T \mathbf{V} \mathbf{A}_\infty)^{-1}\mu_{\mathbf{xz}}^T\theta + O_p(n^{-1}) \\ &= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{V}^{-1}\mu_{\mathbf{xz}}(\mu_{\mathbf{xz}}^T \mathbf{V}^{-1}\mu_{\mathbf{xz}})^{-1}\mu_{\mathbf{xz}}^T\theta + O_p(n^{-1}). \quad (9)\end{aligned}$$

Si $\hat{\mu}_{y,\mathbf{z}\mathbf{V}}$ est bornée, comme cela se produirait si l'on modifiait $f(\eta) = 1 + \eta$ quand η est grand afin d'éviter de grands ajustements des poids de calage, nous avons

$$\begin{aligned}
\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}}) &= \mathrm{E}(\bar{y}_s) + O(n^{-1}) = \mu_y + O(n^{-1}) \\
\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}}|\bar{\mathbf{x}}_s) &= \bar{\mathbf{x}}_s^T\theta + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T\mathbf{V}^{-1}\mu_{\mathbf{xz}}(\mu_{\mathbf{xz}}^T\mathbf{V}^{-1}\mu_{\mathbf{xz}})^{-1}\mu_{\mathbf{xz}}^T\theta \\
&\quad + O(n^{-1}) \\
\mathrm{Var}[\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}}|\bar{\mathbf{x}}_s)] &= \frac{1}{n}\theta^T(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})^T\theta \\
&\quad + o(n^{-1}),
\end{aligned}$$

où $\boldsymbol{\Sigma}_{\mathbf{x}}$ est la matrice de covariance de $\mathbf{x}$ et

$$\mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}} = \mu_{\mathbf{xz}}(\mu_{\mathbf{xz}}^T\mathbf{V}^{-1}\mu_{\mathbf{xz}})^{-1}\mu_{\mathbf{xz}}^T\mathbf{V}^{-1}. \qquad (10)$$

Maintenant,

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s) &= \mathrm{Var}(\bar{y}_s \mid \bar{\mathbf{x}}_s) + o(n^{-1}) \\
&= \frac{1}{n^2}\sum_{j\in\mathcal{S}} V(\mathbf{x}_j) + o(n^{-1}) \\
\mathrm{E}[\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s)] &= \mathrm{E}[\mathrm{Var}(\bar{y}_s \mid \bar{\mathbf{x}}_s)] + o(n^{-1}) \\
&= \frac{1}{n}\sum_r f_r V_r + o(n^{-1}).
\end{aligned}$$

Il est facile de voir que

$$\mathrm{Var}[\mathrm{E}(\bar{y}_s \mid \bar{\mathbf{x}}_s)] = \frac{1}{n}\,\theta^T\boldsymbol{\Sigma}_{\mathbf{x}}\theta.$$

Puisque
$\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}}) = \mathrm{Var}[\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}}|\bar{\mathbf{x}}_s)] + \mathrm{E}[\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}}|\bar{\mathbf{x}}_s)]$ et de même pour $\mathrm{Var}(\bar{y}_s)$, $\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}}) < \mathrm{Var}(\bar{y}_s)$ jusqu'aux termes d'ordre $o(n^{-1})$ quand

$$\theta^T(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})^T\theta < \theta^T\boldsymbol{\Sigma}_{\mathbf{x}}\theta. \qquad (11)$$

Les calculs établissent également que la contribution du biais quadratique à l'erreur quadratique moyenne de $\hat{\mu}_{y,\mathbf{zV}}$ est asymptotiquement négligeable.

5. Une nouvelle matrice de pondération $\mathbf{V}^{-1}$ proposée

À la présente section, nous retournons aux variables de référence $\mathbf{x}$ et du modèle $\mathbf{z}$ originales. Quand $\mathbf{V} = \mathbf{I}$, la matrice identité, nous voyons, en partant de (10), que $\mathbf{P}^T_{\mu_{\mathbf{xz}},\mathbf{I}}\theta = \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{I}}\theta$ est la projection de θ sur l'étendue des colonnes de $\mu_{\mathbf{xz}}$. Le premier membre de (11) sera nul si θ est dans cette étendue des colonnes.

Pour simplifier, nous écrivons $\mu_{\mathbf{xz}}$ sous forme d'une matrice singulière, de rang $I + J - 1$, contenant une ligne pour chaque cellule de classification croisée (i, j) possible et une colonne pour chaque classification par ligne i et chaque classification par colonne j. Donc, la $(i, j)^{\mathrm{e}}$ ligne de $\mu_{\mathbf{xz}}$ contient $f_{ij} = N_{ij}/N$ dans les colonnes correspondant à i et à j, et des zéros ailleurs. Donc, θ sera dans l'étendue des colonnes de $\mu_{\mathbf{xz}}$ si, et uniquement si, pour chaque i et j

$$\frac{\theta_{ij}}{f_{ij}} = \alpha_i + \beta_j \qquad (12)$$

pour un certain α_i et un certain β_j. Autrement dit, le ratio θ_{ij}/f_{ij} satisfait un modèle d'ANOVA à deux critères de classification, sans interaction, dans les classifications par colonne et ligne.

En nous souvenant que θ_{ij} représente la valeur moyenne de la variable d'intérêt y dans la $(i, j)^{\mathrm{e}}$ cellule, l'expression (12) ne semble pas être une approximation très prometteuse de la vérité. Une approximation plus probable serait le modèle d'ANOVA à deux critères de classification habituel

$$\theta_{ij} = \alpha_i + \beta_j. \qquad (13)$$

Supposons que nous remplacions les variables $\tilde{\mathbf{x}} = \mathbf{C}\mathbf{x}$ par certaines matrices diagonales $\mathbf{C}$. Notons que les lignes et les colonnes de $\mathbf{C}$ sont doublement indicées par (i, j) et désignons par c_{ij} l'entrée diagonale dans les $(i, j)^{\mathrm{e}}$ ligne et colonne. Soit $\tilde{\theta} = \tilde{\mathbf{C}}^{-1}\theta$, de sorte que le modèle (6) peut être réécrit sous la forme

$$y_k = \tilde{\mathbf{x}}_k^T\tilde{\theta} + \varepsilon_k.$$

Maintenant, la matrice $\mu_{\tilde{\mathbf{x}}\mathbf{z}}$ contient $c_{ij}f_{ij}$ dans la $(i, j)^{\mathrm{e}}$ ligne et dans les colonnes correspondant à i et à j. Ainsi, $\tilde{\theta}$ sera dans l'étendue des colonnes de $\mu_{\tilde{\mathbf{x}}\mathbf{z}}$ si et uniquement si

$$c_{ij}^{-1}\theta_{ij} = \tilde{\theta}_{ij} = c_{ij}f_{ij}(\alpha_i + \beta_j).$$

Donc, le modèle (13) est équivalent à $c_{ij} = f_{ij}^{-1/2}$. Il est facile de vérifier que

$$\begin{aligned}
&\tilde{\theta}^T(\mathbf{I} - \mathbf{P}_{\mu_{\tilde{\mathbf{x}}\mathbf{z}},\mathbf{I}})\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}}(\mathbf{I} - \mathbf{P}_{\mu_{\tilde{\mathbf{x}}\mathbf{z}},\mathbf{I}})\tilde{\theta} = \\
&\qquad \theta^T(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})^T\theta
\end{aligned}$$

quand $\mathbf{V} = \mathbf{C}^{-2}$. Nous proposons donc d'utiliser la matrice diagonale $\mathbf{V}_o$ dont les entrées diagonales sont f_{ij}.

Avec ce choix de $\mathbf{V}_o$, l'équation (9) suggère l'estimateur pour échantillonnage aléatoire simple

$$\hat{t}_{y,\mathbf{zV}_o} =$$

$$N\bar{y}_s + (T_{\mathbf{x}} - N\bar{x}_s)^T\mathbf{V}_{\mathbf{o}}^{-1}\hat{\mu}_{\mathbf{xz}}(\hat{\mu}_{\mathbf{xz}}^T\mathbf{V}_o^{-1}\hat{\mu}_{\mathbf{xz}})^{-1}\left(\frac{1}{n}\sum_{k\in\mathcal{S}}\mathbf{z}_k y_k\right) \qquad (14)$$

où $\hat{\mu}_{\mathbf{xz}} = n^{-1}\sum_{k\in\mathcal{S}}\mathbf{x}_k\mathbf{z}_k^T$. Dans notre cas, $\mu_{\mathbf{xz}}$ ainsi que $\mu_{\mathbf{x}}$ sont connues à partir de N_{ij}, mais dans l'esprit de l'estimation par le ratio, il est préférable d'utiliser $\hat{\mu}_{\mathbf{xz}}$ au lieu de $\mu_{\mathbf{xz}}$. Cette observation heuristique a été démontrée au moyen de simulations (non présentées) en se servant de la population du QCEW.

Nous donnons à l'estimateur $\hat{t}_{y,\mathbf{zV}_o}$ de l'équation (14) le nom d'estimateur *à fréquences de cellule calées pondéré*.

Des simulations exécutées en se servant de variables réponse artificielles y, bien que non présentées, montrent que, si le modèle (13) est vérifié, l'estimateur à fréquences de cellule calées pondéré $\hat{t}_{y,\mathbf{zV}_o}$ donne de nettement meilleurs résultats que les autres estimateurs pris en considération ici. Le tableau 6 donne les statistiques pour l'estimateur $\hat{t}_{y,\mathbf{zV}_o}$ pour les populations et les variables étudiées aux tableaux 2 à 5.

Comparativement aux tableaux 2 à 5, nous voyons que, dans tous les cas, $\hat{t}_{y,\mathbf{zV}_o}$ est, de tous les estimateurs considérés, celui qui possède le biais le plus élevé, mais l'EQM la plus faible. Pour n = 200 et la population complète, $\hat{t}_{y,\mathbf{zV}_o}$ produit un gain d'efficacité (mesuré par l'EQM) de 14,8 %, comparativement aux résultats empiriques obtenus pour l'estimateur sans biais pour l'estimation de l'emploi total et un gain d'efficacité de 21,1 % pour l'estimation de la rémunération totale. Pour n = 200 et la population dont a été supprimée une unité aberrante unique, les gains d'efficacité correspondants sont de 14,2 % et de 21,7 % pour l'emploi total et la rémunération totale, respectivement.

Le rédacteur associé a proposé que nous comparions nos estimateurs à un estimateur poststratifié en utilisant des cellules regroupées pour éviter le problème de cellules vides dans l'échantillon. Nous avons examiné cette question pour une taille d'échantillon n = 200 pour laquelle il est fort probable qu'il existe des cellules vides. Nous avons créé 14 poststrates. Neuf de ces poststrates correspondaient aux neuf plus grandes cellules dans les données originales. Les cinq autres poststrates étaient A1 et A2 ; A3, A5 et B4 ; A4, B5 et C4 ; C5 et D4 ; et toutes les cellules provenant de l'État E regroupées avec D5. Après ces combinaisons, les cinq poststrates combinées avaient une taille variant de 4,07 % à 5,06 % de la population, et les neuf cellules originales retenues avaient une taille variant de 4,62 % à 11,47 % de la population.

Malheureusement, l'auteur n'a plus accès à la base de données du QCEW. À part les fréquences de cellule du tableau 1, il ne dispose que des moyennes, des écarts-types et des valeurs maximales par cellule. Il a construit une pseudo population en utilisant les carrés de variables gamma générées aléatoirement. Les variables gamma au carré ont été créées de manière à ce qu'elles aient les mêmes moyennes et écarts-types de cellule que les moyennes et écarts-types de cellule des données originales. Ensuite, les variables gamma au carré ont été arrondies vers le haut à une valeur entière. Pour ces pseudo populations, T_y = 3 149 491 pour l'emploi et 2 305 273, en dizaines de milliers de dollars, pour la rémunération.

Une distribution gamma au carré a été utilisée parce que la distribution gamma n'est pas suffisamment asymétrique avec étalement vers la droite, même si, dans presque toutes les cellules, la plus grande valeur dans la population originale est supérieure à la plus grande valeur dans la pseudo population. Naturellement, sans les données originales, nous ne pouvons pas faire la distinction entre un étalement de la distribution vers la droite et une tendance à produire des valeurs aberrantes.

En tout, 10 000 échantillons Monte Carlo ont été tirés pour chaque taille d'échantillon. Les résultats sont présentés au tableau 7. Pour l'estimateur poststratifié, cinq des échantillons de taille 200 possédaient une poststrate vide et ces exécutions ont été exclues des résultats du tableau 7.

Tableau 6
Statistiques empiriques pour $\hat{t}_{y,\mathbf{zV}_o}$ de l'équation (14)

n	Biais	Erreur-type	REQM	R.C. var. est.
		Population complète – Emploi total		
200	-244 749	967 066	997 556	923 492
1 000	-64 839	490 758	495 023	483 550
5 000	-10 767	221 702	221 964	219 408
		Population complète – Rémunération totale		
200	-242 528	1 388 489	1 409 511	1 333 793
1 000	-62 091	752 603	755 160	744 315
5 000	-9 821	332 682	332 827	333 782
		Population avec unité aberrante supprimée – Emploi total		
200	-236 812	881 844	913 088	842 191
1 000	-67 468	405 215	410 793	396 105
5 000	-11 482	189 501	189 848	185 483
		Population avec unité aberrante supprimée – Rémunération totale		
200	-228 441	1 194 922	1 216 562	1 151 417
1 000	-66 765	565 008	568 939	557 676
5 000	-11 138	263 699	263 934	260 768

De toute évidence, la poststratification n'est pas utile. Même si aucune poststrate n'avait une fréquence prévue inférieure à huit, la taille des poststrates réelles était assez variable. En outre, les populations de cellule sont relativement asymétriques de sorte que les moyennes d'échantillon des poststrates sont assez variables.

Les autres conclusions concernant les pseudo populations reflètent celles tirées pour les populations réelles. En particulier, pour n = 200 et la pseudo population pour l'emploi, l'estimateur à fréquences de cellule calées pondéré $\hat{t}_{y,\mathbf{zV}_o}$ donne un gain d'efficacité de 11,8 %, comparativement à l'estimateur sans biais. Pour la pseudo population pour la rémunération et n = 200, le gain d'efficacité est de 21,1 %.

Tableau 7
Comparaison empirique de quatre estimateurs

Estimateur	Biais	Erreur-type Emploi total	REQM	Biais	Erreur-type Rémunération totale	REQM
			n = 200			
Non pondéré	644	1 006 956	1 006 956	-9 970	1 481 450	1 481 483
Poststratifié	-5 387	1 026 266	1 026 280	-2 149	1 548 833	1 548 834
Fréq. cell. calées	-224 198	942 164	968 472	-203 531	1 377 823	1 392 775
Fréq. cell. cal. pondéré	-248 937	919 419	952 523	-232 558	1 326 234	1 346 469
			n = 1 000			
Non pondéré	-3 317	445 676	445 687	1 544	679 148	679 150
Poststratifié	-2 967	448 218	448 228	1 672	685 370	685 372
Fréq. cell. calées	-54 311	436 821	440 185	-44 942	665 799	667 314
Fréq. cell. cal. pondéré	-63 327	432 396	437 008	-54 913	660 726	663 004
			n = 5 000			
Non pondéré	2 466	206 249	206 264	-2 539	304 852	304 863
Poststratifié	2 108	205 661	205 672	-2 705	304 751	304 763
Fréq. cell. calées	-8 265	204 693	204 859	-12 096	303 231	304 472
Fréq. cell. cal. pondéré	-10 551	204 080	204 352	-14 697	302 311	302 668

6. Conclusion

L'utilisation dans l'expression (3) de matrices de pondération $\mathbf{V}(\beta)^{-1}$ qui dépendent de β n'a pas été examinée dans le présent article. L'expérimentation avec ce genre de matrices n'a pas été encourageante. Le temps de calcul augmentait considérablement et, dans un nombre important de cas, la convergence numérique n'était pas atteinte, sans aucune amélioration de l'efficacité par rapport aux estimateurs avec matrice $\mathbf{V}$ fixe considérés ici. L'auteur n'a peut-être pas choisi la bonne matrice $\mathbf{V}(\beta)$. pour ses essais.

Outre la fonction de lien rétrotransformée exponentielle, l'auteur a essayé la fonction de lien rétrotransformée logistique $f(\eta) = (1 + e^{-\eta})^{-1}$. Ces exécutions n'ont pas convergé non plus. Tout bien réfléchi, la raison en est évidente : comme il n'y avait aucune non-réponse ni aucun problème de non-couverture dans les simulations, les ajustements des poids de calage $f(\mathbf{z}^T\beta) \to 1$ quand $n \to \infty$. Cependant, 1 n'est pas l'image de f. Il convient de souligner que, dans Chang et Kott (2008), une fonction de lien rétrotransformée logistique a été utilisée pour corriger la non-réponse.

Plusieurs questions évidentes se posent. Par exemple, comment les résultats de la présente étude évolueraient-ils si l'on utilisait un plan de sondage plus compliqué que l'échantillonnage aléatoire simple, ou s'il se produisait une non-réponse et/ou une non-couverture et que le calage était utilisé pour la corriger ? Falk (2010) examine ces questions tant théoriquement qu'en exécutant d'autres simulations au moyen de la population du QCEW. Falk (2010) considère également des fonctions de lien non linéaires.

Les extensions à des données croisées à trois dimensions (et plus) sont évidentes. Si I, J, K désignent le nombre de cellules dans chacune des trois classifications, il existe IJK cellules entièrement classées dont les totaux peuvent être utilisés comme variables $\mathbf{x}$ de référence. Il existe $IJ + IK + JK - I - J - K + 1$ variables marginales à une et à deux dimensions qui peuvent être utilisées pour les variables $\mathbf{z}$ du modèle. Il est clair que l'on pourrait ne pas vouloir utiliser la pléthore de variables disponibles.

Dans le contexte du calage linéaire en utilisant les mêmes variables $\mathbf{x}$ et $\mathbf{z}$, plusieurs études ont porté sur le choix des variables. Les études de Banker, Rathwell et Majkowski (1992), Silva et Skinner (1997), et de Clark et Chambers (2008) en sont des exemples. Les auteurs du dernier article font remarquer qu'un trop grand nombre de variables peut détériorer l'EQM de $\hat{T}_y$.

Les options de remplacement de la poststratification examinées ici peuvent être utilisées en présence de cellules dont l'effectif est faible, voire même de cellules vides. Par exemple, dans nos simulations, la fréquence prévue dans la cellule définie par l'État E et le groupe d'industries 5, est de 0,36 quand $n = 200$. On pourrait être tenté de regrouper certaines cellules et d'utiliser la poststratification. Toutefois, en général, il n'est pas possible de le faire et de maintenir la structure à deux dimensions des données. Comme la poststratification, notre approche comprend l'introduction de poids pour équilibrer l'échantillon, mais évite le regroupement de cellules. Ces approches augmentent généralement le biais, mais peuvent donner lieu à une réduction importante de l'EQM.

En outre, en présence de non-réponse ou de non-couverture, les inverses des ajustements des pondérations peuvent être considérés, sous un modèle de quasi-randomisation pour la réponse ou la couverture, comme probabilités estimées de réponse et/ou de couverture.

Dans nos approches de calage, nous supposons que ces probabilités sont une fonction des classifications par ligne et par colonne. Quand des cellules sont regroupées sans maintenir la classification double, ces probabilités sont difficiles à interpréter.

Remerciements

L'auteur remercie Phil Kott et John Eltinge de lui avoir fourni plusieurs éclaircissements très intéressants. Il remercie aussi Larry Huff et Ken Robertson du Bureau of Labor Statistics de l'avoir aidé à obtenir et à comprendre les données.

Annexe

Ici, nous établissons, en utilisant les équations (16) et (17) de Chang et Kott (2008), les estimateurs de variance fondés sur l'échantillon pour les quatre estimateurs étudiés à la section 2.

Soit

$$\hat{\mathbf{H}}_y = \frac{\partial \hat{t}_{y,\mathbf{z}f\mathbf{V}}}{\partial \beta}(\hat{\beta}).$$

Ici $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ est défini en (2). $\hat{\mathbf{H}}_y$ est un vecteur ligne contenant une entrée pour chaque variable $\mathbf{z}$. Dans le cas de la suppression d'une unité aberrante, $\hat{\mathbf{H}}_y$ contient $(I + J - 1)$ entrées, une pour chacune des $I + J - 1$ variables indicatrices linéairement indépendantes pour les classifications par ligne et par colonne.

Pour les estimateurs à marges calées et à fréquences de cellule calées, $f(\eta) = 1 + \eta$. Définissons les constantes s_{ij} et t_{ij} par

$$s_{ij} = \frac{N}{n} r_{ij}$$
$$t_{ij} = \frac{N}{n} \sum_{k \in \mathcal{S}_{ij}} y_k.$$

Alors, un simple calcul montre que, s'il existe une entrée dans $\hat{\mathbf{H}}_y$ pour la i^{e} classification par ligne, nous plaçons $\sum_j t_{ij}$. dans cette entrée. De même, si une entrée existe pour la j^{e} classification par colonne, nous plaçons $\sum_i t_{ij}$. dans cette entrée. Ici, nous utilisons la convention voulant que, si la i^{e} ligne ou la j^{e} colonne n'est pas l'une des $I + J - 1$ variables indicatrices linéairement indépendantes choisies, le $\beta_{i\cdot}$ ou $\beta_{\cdot j}$ correspondant est égal à 0.

Pour l'estimateur par ratissage (*raking)* et l'estimateur à fréquences de cellule calées exponentiel, $f(\eta) = e^{\eta}$ et nous pouvons calculer $\hat{\mathbf{H}}_y$ de manière similaire en utilisant

$$s_{ij} = \frac{N}{n} \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}$$
$$t_{ij} = \frac{N}{n} \sum_{k \in \mathcal{S}_{ij}} \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) y_k.$$

Ici, nous adoptons la convention voulant que, si la i^{e} ligne ou la j^{e} colonne n'est pas l'une des $I + J - 1$ variables indicatrices linéairement indépendantes choisies, le $\beta_{i\cdot}$ ou $\beta_{\cdot j}$ correspondant est égal à 1.

Par analogie avec (2), soit

$$\hat{t}_{\mathbf{x},\mathbf{z}f\mathbf{V}} = \sum_{k \in \mathcal{S}} d_k f(\mathbf{z}_k^T \hat{\beta}) x_k.$$

$\hat{t}_{\mathbf{x},\mathbf{z}f\mathbf{V}}$ est un vecteur colonne contenant une entrée pour chaque variable $\mathbf{x}$. Définissons la matrice $\hat{\mathbf{H}}$ comme étant

$$\hat{\mathbf{H}} = \frac{\partial \hat{t}_{\mathbf{x},\mathbf{z}f\mathbf{V}}}{\partial \beta}(\hat{\beta}).$$

$\hat{\mathbf{H}}$ est une matrice contenant une ligne pour chaque variable $\mathbf{x}$ et une colonne pour chaque variable $\mathbf{z}$.

Pour les estimateurs à fréquences de cellule calées et à fréquences de cellule calées exponentiel, la matrice $\hat{\mathbf{H}}$ est de dimensions $IJ \times (I + J - 1)$. Chacune des lignes de $\hat{\mathbf{H}}$ correspond à une paire (i, j) de classifications par ligne et colonne. Nous plaçons s_{ij} dans la ligne correspondant à (i, j) et dans les colonnes correspondant à la i^{e} classification par ligne et à la j^{e} classification par colonne (quand ces colonnes existent). Toutes les autres entrées de $\hat{\mathbf{H}}$ sont fixées à zéro.

Pour les estimateurs à marges calées et par ratissage, la matrice $\hat{\mathbf{H}}$ est de dimensions $(I + J - 1) \times (I + J - 1)$. Si une ligne (et donc une colonne) de $\hat{\mathbf{H}}$ existe pour la i^{e} classification par ligne, nous plaçons $\sum_j s_{ij}$ dans l'entrée diagonale correspondante de $\hat{\mathbf{H}}$. De même, si une ligne et une colonne existe pour la j^{e} classification par colonne, nous plaçons $\sum_i s_{ij}$ sur la diagonale de $\hat{\mathbf{H}}$. Nous plaçons s_{ij} dans l'entrée dont la ligne correspond à la i^{e} classification par ligne et dont la colonne correspond à la j^{e} classification par colonne (si celles-ci existent). Nous plaçons aussi s_{ij} dans l'entrée dont la colonne correspond à la i^{e} classification par ligne et dont la ligne correspond à la j^{e} classification par colonne (de nouveau si celles-ci existent). Les valeurs de toutes les autres entrées de $\hat{\mathbf{H}}$ sont fixées à zéro.

Soit $\mathbf{B} = \hat{\mathbf{H}}_{\mathbf{y}}^{\mathbf{T}}(\hat{\mathbf{H}}^{\mathbf{T}}\mathbf{V}^{-1}\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^{\mathbf{T}}\mathbf{V}^{-1}$ où, pour le moment, nous utilisons une matrice identité pour $\mathbf{V}$. $\mathbf{B}$ est de dimensions $1 \times (I + J - 1)$ pour les estimateurs à marges calées et par ratissage, et de dimensions $1 \times IJ$ pour les estimateurs à fréquences de cellule calées et à fréquences de cellule calées exponentiel. Dans les deux premiers cas, nous désignons les entrées de $\mathbf{B}$ par $b_{i\cdot}$ ou $b_{\cdot j}$, et pour le cas unique où un indice de colonne ou de ligne ne correspond pas à l'une des $I + J - 1$ variables indicatrices indépendantes,

nous fixons la valeur du b correspondant à zéro. Dans les deux derniers cas, nous désignons les entrées de **B** par b_{ij}. Pour $k \in \mathcal{S}_{ij}$, soit $u_k = w_{ij}(y_k - b_{i\cdot} - b_{\cdot j})$ pour les estimateurs à marges calées et par ratissage, et $u_k = w_{ij}(y_k - b_{ij})$ pour les estimateurs à fréquences de cellule calées et à fréquences de cellule calées exponentiel.

Essentiellement, Chang et Kott (2008) ont montré qu'asymptotiquement, l'estimateur calé a la même forme qu'un estimateur par la régression de la forme correspondant à l'équation (6.6.1) de Särndal et coll. (1992), où la matrice **B** susmentionnée joue le rôle du **B** dans (6.6.1) et les poids d'échantillonnage d_k sont remplacés par $d_k f(\mathbf{z}_k^T \hat{\beta})$. Pour l'échantillonnage sans remise, ils proposent d'estimer la variance de $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ en apportant les changements analogues à l'équation (6.6.3) de Särndal et coll. (1992).

Pour l'échantillonnage aléatoire simple et en l'absence de non-réponse et de non-couverture, l'estimateur de variance devient

$$\widehat{\mathbf{V}} = \frac{N^2}{n}(1 - n/N)s_u^2 \qquad (15)$$

où s_u^2 est la variance d'échantillon des u_k.

En présence de non-réponse, si l'on suppose que les répondants $\mathcal{S}$ représentent un échantillon de Poisson tiré de l'échantillon aléatoire simple original avec les probabilités poissonniennes $f(\mathbf{z}^T\beta_0)^{-1}$, l'estimateur de variance devient

$$\hat{\mathbf{V}} = \frac{N^2}{n}(1 - n/N)s_u^2 + \frac{N}{n}\sum_i\sum_j(1 - w_{ij})\sum_{k\in\mathcal{S}_i j}u_k^2 \qquad (16)$$

où s_u^2 est la variance d'échantillon des u_k. La même formule fonctionne pour la non-couverture où $f(\mathbf{z}^T\beta_0)^{-1}$ représente la probabilité de couverture et de réponse combinées dans un modèle à trois degrés dans lequel l'univers couvert est supposé être un échantillon de Poisson tiré de l'univers souhaité, l'échantillon est un échantillon aléatoire simple tiré de l'univers couvert, et les répondants représentent un échantillon de Poisson tiré de l'échantillon original.

Bibliographie

Banker, M.D., Rathwell, S. et Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 canadian census. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 764-769.

Chang, T., et Kott, S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 557-571.

Clark, R., et Chambers, R.L. (2008). Adaptive Calibration for Prediction of Finite Population Totals. University of Wollongong (document de travail en ligne).

Falk, G. (2010). *Calibration Adjustment for Nonresponse in Cross-Classified Data*, University of Virginia (dissertation).

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

Kott, S., et Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105, 1265-1275.

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

Silva, P.L.D.N., et Skinner, C.J. (1997). Sélection des variables pour l'estimation par régression dans le cas des populations finies. *Techniques d'enquête*, 23, 25-35.

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 45-56
Statistique Canada, N° 12-001-X au catalogue

À propos de la variance des variations estimées d'après des panels rotatifs et des strates dynamiques

Paul Knottnerus et Arnout van Delden [1]

Résumé

De nombreuses enquêtes-entreprises fournissent des estimations du chiffre d'affaires mensuel pour les principaux codes de la Classification type des industries. Cela inclut les estimations des variations du niveau du chiffre d'affaires mensuel comparativement à 12 mois plus tôt. Comme des échantillons chevauchant sont souvent utilisés dans les enquêtes-entreprises, les estimations du chiffre d'affaires durant des mois consécutifs sont corrélées, ce qui complique le calcul de la variance des variations. Le présent article décrit une procédure générale d'estimation de la variance qui comprend des corrections annuelles des strates quand des établissements passent dans d'autres strates en raison de leur taille réelle. La procédure tient également compte du renouvellement des échantillons, ainsi que des nouvelles unités et des unités disparues. L'article se termine par un exemple de calcul de la variance de l'estimation du taux de croissance annuel du chiffre d'affaires mensuel des supermarchés des Pays-Bas.

Mots clés : Covariances conditionnelles ; corrections des strates ; échantillons chevauchants ; enquêtes-entreprises ; nouvelles unités ; unités disparues.

1. Introduction

Dans de nombreuses enquêtes, une population en évolution est échantillonnée de manière répétée afin de pouvoir estimer le niveau et la variation de niveau d'une caractéristique entre deux éditions de l'enquête. Par exemple, beaucoup de pays réalisent une enquête-entreprise mensuelle pour estimer le niveau du chiffre d'affaires mensuel et la variation de ce niveau comparativement à un mois ou à une année plus tôt ; voir Konschnik, Monsour et Detlefsen (1985). Un autre exemple est celui de l'enquête sur la population active qui comprend l'échantillonnage mensuel de la population pour estimer le nombre de personnes occupées et le taux de chômage. L'estimation de la variance est nécessaire afin de savoir si les variations observées sont statistiquement significatives. L'estimation de la variance est également nécessaire à l'étape de la conception de l'enquête, afin de déterminer la taille et la répartition optimales de l'échantillon, ou l'estimateur optimal.

Dans les enquêtes répétées, on estime souvent les changements en s'appuyant sur une stratification de la population. Les entreprises sont très hétérogènes en ce qui concerne la taille et le type d'activité économique. Par conséquent, les enquêtes-entreprises sont habituellement réalisées selon un plan d'échantillonnage aléatoire simple stratifié (EASS) avec tirage sans remise ; voir Smith, Pont et Jones (2003). Dans les enquêtes auprès des ménages ou des particuliers, l'échantillon n'est en général pas stratifié, parce que les ménages sont moins hétérogènes. Certaines enquêtes sociales, telles que les enquêtes sur la population active, comprennent toutefois une poststratification afin de réduire la variance et le biais de l'estimateur.

En vue d'établir les formules de la variance d'une variation estimée dans une population comprenant des strates dynamiques, il convient de faire attention à trois facteurs de complication. Premièrement, la variation de niveau est le résultat de deux composantes. L'une est due à la variation de la moyenne de population des unités qui se trouvaient dans la même strate lors des deux éditions de l'enquête. L'autre est causée par le changement de composition de la strate entre deux éditions de l'enquête résultant de la création de nouvelles unités et de la disparition d'unités existantes au sein de la population, ainsi que de la migration des unités de population entre les strates ; voir Holt et Skinner (1989). Deuxièmement, en raison de la migration de certaines unités de population entre les strates, la moyenne estimée de la strate h à la période t peut être corrélée avec la moyenne de la strate ℓ à la période $t+1$. Troisièmement, une autre complication tient au fait que la population est échantillonnée répétitivement, ce qui donne des échantillons partiellement chevauchants entre deux périodes. Différents plans à panel rotatif peuvent être utilisés dans les enquêtes-entreprises.

Divers auteurs ont établi des formules d'estimateurs de la variance fondées sur le plan de sondage pour l'estimation des variations. En supposant une grande population sans création ni disparition d'unités, Kish (1965) a dérivé une expression pour la variance d'une variation estimée fondée sur des échantillons chevauchants. Tam (1984) a éliminé l'hypothèse d'une grande population. Partant des résultats de Tam, Qualité et Tillé (2008) comparent plusieurs estimateurs de variance d'une variation estimée. Wood (2008) généralise les résultats de Tam aux enquêtes avec probabilités de sélection inégales. Lowerre (1979) et Laniel (1987)

1. Paul Knottnerus et Arnout van Delden, Statistics Netherlands, CP 24500, 2490 HA La Haye, Pays-Bas. Courriel : pkts@cbs.nl et adln@cbs.nl.

traitent de l'estimation de la variance d'une variation dans des populations dynamiques, mais ils ne tiennent pas compte de la stratification. Hidiroglou, Särndal et Binder (1995) considèrent des populations dynamiques et la stratification, mais sans modification des strates. Nordberg (2000) et Berger (2004) établissent des formules pour la situation plus compliquée d'une population dynamique comprenant des unités qui migrent entre les strates. Pour le plan d'échantillonnage de la Suède, Nordberg (2000) établit des formules en utilisant des indicateurs d'inclusion qui requièrent certains calculs algébriques. En supposant que la taille du chevauchement des deux échantillons à deux périodes différentes est fixe, Berger (2004) établit ses formules en se basant sur un échantillonnage de Poisson conditionnellement à la taille d'échantillon par strate, ce qui requiert certaines opérations algébriques matricielles.

Dans le présent article, nous établissons les expressions pour l'échantillonnage aléatoire simple stratifié (EASS) d'une manière plus simple, sans supposer que les tailles des chevauchements sont fixes. En outre, contrairement au plan d'échantillonnage de la Suède, celui des Pays-Bas ne requiert pas de calculs très longs pour estimer l'une des composantes de la variance d'une variation. En outre, nous proposons une méthode d'estimation de rechange pour les plans d'échantillonnage comportant une telle composante non nulle. Afin de clarifier la procédure d'estimation de la variance, nous décrivons son application au taux de croissance annuel du chiffre d'affaires des supermarchés des Pays-Bas pour des périodes de quatre semaines.

La présentation de l'article est la suivante. À la section 2, nous décrivons brièvement l'enquête-entreprise sur le chiffre d'affaires mensuel menée aux Pays-Bas, y compris le plan d'échantillonnage. À la section 3, nous établissons les formules de la variance de l'estimateur d'une variation. À la section 4, nous illustrons la procédure d'estimation de la variance en comparant les variances de deux estimateurs différents du taux de croissance annuel du chiffre d'affaires mensuel des supermarchés des Pays-Bas au cours de la période 2003-2004. Enfin, à la section 5, nous résumons les principaux résultats et présentons nos conclusions.

2. Le plan d'échantillonnage des enquêtes-entreprises des Pays-Bas

Chaque mois, Statistics Netherlands estime le chiffre d'affaires mensuel pour certains des codes principaux de la CTI. Les données publiées comprennent les taux de croissance sur 12 mois du chiffre d'affaires mensuel, c'est-à-dire la variation relative du niveau mensuel du chiffre d'affaires comparativement à 12 mois plus tôt. Dans la suite de l'exposé, nous parlerons de ce taux de croissance comme étant le taux de croissance annuel.

Toutes les unités statistiques ou *établissements* figurent sur la liste du Registre général des entreprises (RGE) tenu à jour par Statistics Netherlands. Le registre est mis à jour chaque mois au moyen de sources administratives en ce qui concerne les créations et les disparitions d'établissements, et une fois par an, le 31 décembre, en ce qui concerne la catégorie de taille et le type d'activité économique (code de la CTI). Il convient de souligner que l'enregistrement dans le RGE peut avoir lieu avec un certain retard par rapport aux changements dans la population (création d'établissements, disparition d'établissements, changements de classe de taille, *etc.*). De surcroît, les unités disparues (inconnues) dans la base de sondage peuvent entraîner un biais dans les estimations du niveau du chiffre d'affaires. Afin d'éviter ce genre de biais, il est important de détecter rapidement les disparitions et de supprimer ces unités de la base de sondage. Les disparitions d'établissements décelées dans l'échantillon peuvent jouer un rôle ici. Cependant, une analyse plus approfondie et la correction de ces erreurs dépassent le cadre du présent article sur l'estimation de la variance des taux de croissance. Pour estimer ces variances, nous supposons que les unités de la population et leurs caractéristiques dans le registre sont correctes. De même, nous supposons que la non-réponse aux enquêtes est nulle.

Chaque premier jour du mois, un échantillon EASS est tiré du RGE pour estimer le chiffre d'affaires du mois courant. En fait, on utilise un échantillon rotatif. L'échantillon est stratifié selon la taille et selon le type d'activité économique. La probabilité réelle de sélection dépend de la taille et de l'activité économique. La probabilité de sélection augmente avec la taille de l'établissement, les plus grands étant inclus dans l'échantillon avec une probabilité de 1. Pour certains codes de la CTI, il existe non seulement des données d'enquête, mais aussi des données provenant de sources administratives. Les unités déjà présentes dans les fichiers administratifs sont considérées comme une strate distincte. Les estimations pour cette strate ont une variance nulle.

L'échantillon est mis à jour de deux façons. Chaque mois, il est corrigé pour tenir compte des nouvelles unités et des unités disparues dans la population. Une fois par an, en janvier, 10 % des unités de l'échantillon sont remplacées et des corrections sont apportées aux strates. Nous allons maintenant discuter plus en détail des mises à jour mensuelles et annuelles.

2.1 Mise à jour mensuelle

Chaque mois t ($t = 1, 2, ...$), une proportion fixe f_h des N_h^t unités de la strate U_h^t sont échantillonnées ($h = 1, ..., H$). Cela donne un échantillon s_h^t de taille $n_h^t = f_h N_h^t$. Donc, le nombre réel d'unités dans l'échantillon peut changer d'un mois à l'autre à cause de la création et de

la disparition d'unités dans la population. Notons qu'en dehors d'erreurs d'arrondissement mineures, la fraction d'échantillonnage réelle f_h ne dépend pas du mois t. En fait, la procédure de mise à jour de s_h^t au mois t est la suivante. Soit $U_{0h}^{t-1,t}$ l'ensemble de nouvelles unités dans la strate h au mois $t-1$ et désignons sa taille par $N_{0h}^{t-1,t}$. Le nombre d'unités échantillonnées à partir de $U_{0h}^{t-1,t}$ au mois t est $n_{0h}^{t-1,t} = f_h N_{0h}^{t-1,t}$. En outre, désignons la différence requise $n_h^t - n_{0h}^{t-1,t}$ par $n_{h,\text{REQ}}^{t-1,t}$ et définissons $s_{h,\text{PRE}}^t$ par $s_{h,\text{PRE}}^t = s_h^{t-1} \cap U_h^t$, c'est-à-dire l'ensemble des unités dans l'échantillon s_h^{t-1} à la période précédente qui sont encore présentes au mois t. Soit $n_{h,\text{PRE}}^t$ la taille de $s_{h,\text{PRE}}^t$. Quand $n_{h,\text{PRE}}^t \geq n_{h,\text{REQ}}^{t-1,t}$, on abandonne aléatoirement un nombre d'unités correspondant à la différence ; autrement, on sélectionne un nombre d'unités correspondant à la différence à partir de $U_h^t \setminus U_{0h}^{t-1,t} \setminus s_{h,\text{PRE}}^t$. Notons que les unités supprimées de l'échantillon au mois $t-1$ ou avant cela peut être à nouveau sélectionnées au mois t.

2.2 Mise à jour annuelle

Chaque année en janvier, l'échantillon est mis à jour pour tenir compte à la fois de la restratification des unités et du remplacement de 10 % de l'échantillon. Toutes les unités présentes dans l'échantillon de décembre qui existent encore en janvier sont stratifiées en fonction de leur taille réelle, c'est-à-dire le nombre d'employés et le code de la CTI en janvier. Les limites des classes de taille proprement dites ne changent pas. Par conséquent, l'échantillon provenant d'une strate conformément à la nouvelle stratification de janvier peut contenir des unités ayant des probabilités d'inclusion différentes, parce que les unités migrent entre les strates ayant des fractions d'échantillonnage différentes.

Afin d'apporter une correction pour tenir compte des probabilités d'inclusion éventuellement différentes dans la strate ℓ, désignons la sous-strate constituée des unités qui appartenaient à la strate h en décembre et à la strate ℓ en janvier par $U_{h\ell}^{\text{déc, janv}}$ et désignons sa taille par $N_{h\ell}^{\text{déc, janv}}$ ($h, \ell = 1, ..., H$). Par analogie avec la procédure de mise à jour mensuelle, définissons $s_{h\ell,\text{PRE}}^{\text{janv}}$ par $s_{h\ell,\text{PRE}}^{\text{janv}} = s_h^{\text{déc}} \cap U_{h\ell}^{\text{déc, janv}}$ et désignons par $n_{h\ell,\text{PRE}}^{\text{janv}}$ la taille de $s_{h\ell,\text{PRE}}^{\text{janv}}$. Puisque la taille requise de l'échantillon $s_{h\ell,\text{REQ}}^{\text{déc, janv}}$ provenant de $U_{h\ell}^{\text{déc, janv}}$ en janvier est $n_{h\ell,\text{REQ}}^{\text{déc, janv}} = f_\ell N_{h\ell}^{\text{déc, janv}}$, la mise à jour annuelle de l'échantillon $s_{h\ell,\text{PRE}}^{\text{janv}}$ s'effectue comme il suit.

Premièrement, quand $n_{h\ell,\text{PRE}}^{\text{janv}} \geq n_{h\ell,\text{REQ}}^{\text{déc, janv}}$, un nombre d'unités correspondant à la différence est supprimé aléatoirement de $s_{h\ell,\text{PRE}}^{\text{janv}}$. En outre, 10 % des $n_{h\ell,\text{REQ}}^{\text{déc, janv}}$ unités encore présentes dans $s_{h\ell,\text{PRE}}^{\text{janv}}$ sont remplacées par des unités provenant de $U_{h\ell}^{\text{déc, janv}} \setminus s_{h\ell,\text{PRE}}^{\text{janv}}$, à condition que ce dernier ensemble contienne suffisamment d'unités. Si le nombre d'unités disponibles n'est pas suffisant, le nombre d'unités remplacées est seulement $N_{h\ell}^{\text{déc, janv}} - n_{h\ell,\text{PRE}}^{\text{janv}}$. Deuxièmement, quand $n_{h\ell,\text{PRE}}^{\text{janv}} < n_{h\ell,\text{REQ}}^{\text{déc, janv}}$, un nombre d'unités correspondant à la différence sont sélectionnées dans $U_{h\ell}^{\text{déc, janv}} \setminus s_{h\ell,\text{PRE}}^{\text{janv}}$. Par la suite, un remplacement supplémentaire de $n_{h\ell,\text{PRE}}^{\text{janv}} - 0{,}9 n_{h\ell,\text{REQ}}^{\text{déc, janv}}$ unités est effectué dans $s_{h\ell,\text{REQ}}^{\text{déc, janv}}$ quand cette différence est positive et que suffisamment de nouvelles unités sont disponibles. Cette procédure est exécutée pour toutes les sous-strates $h\ell$, y compris $h = \ell$. Troisièmement, comme dans la procédure de mise à jour mensuelle, le nombre d'unités échantillonnées en janvier dans la sous-strate $U_{0\ell}^{\text{déc, janv}}$ d'unités nouvellement créées dans la strate ℓ est $n_{0\ell}^{\text{déc, janv}} = f_\ell N_{0\ell}^{\text{déc, janv}}$. En outre, notons que cette approche peut également être adoptée quand les limites des catégories de taille ou les fractions d'échantillonnage sont modifiées en janvier.

À part les corrections des strates effectuées en janvier, l'échantillon résultant au mois t peut être considéré plus ou moins comme un ensemble d'échantillons EAS tirés des strates U_h^t. Quand la population et les strates h sont stables au fil des ans, la procédure décrite jusqu'à présent équivaut à un plan d'échantillonnage aléatoire simple stratifié (EASS) classique pour le mois t. Par conséquent, Statistics Netherlands utilise les formules de variance bien connues pour l'EASS pour estimer la variance du niveau du chiffre d'affaires mensuel. À la section suivante, nous montrons comment la variance de l'estimation d'une variation du niveau peut être estimée sous l'hypothèse d'un échantillonnage EASS.

3. Variance du taux de croissance annuel du chiffre d'affaires mensuel

3.1 Variance du taux de croissance annuel

Soit O^t le chiffre d'affaires total de tous les établissements compris dans la population au mois t et $g^{t,s}$ la variation relative du niveau du chiffre d'affaires entre les mois t et s, c'est-à-dire

$$g^{t,\,s} = \frac{O^t}{O^s} - 1 \quad (t > s).$$

Pour les estimations correspondantes, il est vérifié par définition que

$$\hat{g}^{t,\,s} = \frac{\hat{O}^t}{\hat{O}^s} - 1, \qquad (1)$$

où un « chapeau » indique qu'il s'agit d'une estimation ; pour un estimateur, nous utilisons la même notation. En outre, définissons

$$G^{t,\,t-12} \equiv \frac{O^t}{O^{t-12}} = 1 + g^{t,\,t-12}.$$

Afin d'estimer la variance du taux de croissance annuel du chiffre d'affaires mensuel, nous utilisons le

développement en série de Taylor de premier ordre d'un ratio de deux estimateurs. C'est-à-dire,

$$\begin{aligned}
&\mathrm{var}(\hat{g}^{t,\,t-12}) \\
&= \mathrm{var}\left\{\frac{\hat{O}^{t}}{\hat{O}^{t-12}}\right\} \\
&\approx \frac{\mathrm{var}(\hat{O}^{t} - G^{t,\,t-12}\hat{O}^{t-12})}{(O^{t-12})^2} \\
&= \frac{\mathrm{var}(\hat{O}^{t}) + (G^{t,\,t-12})^2\mathrm{var}(\hat{O}^{t-12}) - 2G^{t,\,t-12}\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^{t})}{(O^{t-12})^2}. \quad (2)
\end{aligned}$$

Le problème principal est l'estimation de $\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^{t})$. Aux sections suivantes, nous examinons ce terme et son estimation.

3.2 Terme de covariance du taux de croissance annuel

En utilisant le plan d'échantillonnage stratifié, nous pouvons écrire le terme $\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^{t})$ provenant de (2) sous la forme

$$\begin{aligned}
\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^{t}) &= \mathrm{cov}\left(\sum_{h=1}^{H} N_h^{t-12}\,\overline{o}_h^{t-12}, \sum_{\ell=1}^{H} N_\ell^{t}\,\overline{o}_\ell^{t}\right) \\
&= \sum_{h=1}^{H}\sum_{\ell=1}^{H} N_h^{t-12} N_\ell^{t}\,\mathrm{cov}(\overline{o}_h^{t-12}, \overline{o}_\ell^{t}), \quad (3)
\end{aligned}$$

où $\overline{o}_h^{t-m}$ désigne la moyenne d'échantillon du chiffre d'affaires dans la strate h pendant le mois $t-m$ ($m = 0, 12$). Notons que la stratification des unités au mois $t-12$ peut différer de celle au mois t. Comme nous l'avons vu à la section 2.2, le renouvellement classique du panel a lieu en janvier. En outre, chaque établissement est affecté à la strate h correcte en fonction de son nombre réel d'employés en janvier ($h = 1, ..., H$). Afin de tenir compte de ces caractéristiques du plan d'échantillonnage, définissons

- $N_{h\ell}^{t-12,t}$: la taille de la sous-strate $U_{h\ell}^{t-12,t}$, c'est-à-dire l'ensemble d'unités qui, au mois $t-12$, appartenait à la strate h et, au mois t, à la strate ℓ ($h, \ell = 1, ..., H$) ;
- $O_{h\ell}^{t-m}$: le total de population de sous-strate du chiffre d'affaires dans $U_{h\ell}^{t-12,\,t}$ au mois $t-m$ ($m = 0, 12$) ;
- $\overline{O}_{h\ell}^{t-m}$: la moyenne de population de sous-strate du chiffre d'affaires dans $U_{h\ell}^{t-12,\,t}$ au mois $t-m$ [c'est-à-dire $\overline{O}_{h\ell}^{t-m} = O_{h\ell}^{t-m}/N_{h\ell}^{t-12,t}$ ($m = 0, 12$)] ;
- $n_{h\ell}^{t-m}$: la taille de l'échantillon $s_{h\ell}^{t-m}$, c'est-à-dire l'échantillon réel tiré de $U_{h\ell}^{t-12,t}$ au mois $t-m$ ($0 \le m \le 12$) ;
- $o_{h\ell}^{t-m}$: le total d'échantillon du chiffre d'affaires dans $s_{h\ell}^{t-m}$ ($m = 0, 12$) ;
- $\overline{o}_{h\ell}^{t-m}$: la moyenne d'échantillon du chiffre d'affaires dans $s_{h\ell}^{t-m}$ [c'est-à-dire $\overline{o}_{h\ell}^{t-m} = o_{h\ell}^{t-m}/n_{h\ell}^{t-m}$ ($m = 0, 12$)] ;
- $n_{h\ell}^{t-12,t}$: le nombre d'unités dans la partie chevauchante des échantillons $s_{h\ell}^{t-12,\,t} \equiv s_{h\ell}^{t-12} \cap s_{h\ell}^{t}$;
- $\overline{o}_{h\ell,\mathrm{CHEV}}^{t-m}$: la moyenne d'échantillon du chiffre d'affaires dans le chevauchement $s_{h\ell}^{t-12,\,t}$ au mois $t-m$. [c'est-à-dire $\overline{o}_{h\ell,\mathrm{CHEV}}^{t-m} = o_{h\ell,\mathrm{CHEV}}^{t-m}/n_{h\ell}^{t-12,t}$ ($m = 0, 12$)].

En plus de la notation de la section 2, définissons la strate auxiliaire 0 pour les *nouvelles unités* aux mois $t-12, ..., t-1$ et, de même, la strate $H+1$ pour les *unités disparues* durant cette période. Nous pouvons alors écrire $\overline{o}_h^{t-12}$ et $\overline{o}_\ell^{t}$ comme étant

$$\begin{aligned}
\overline{o}_h^{t-12} &= \sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}}\,\overline{o}_{hg}^{t-12} \\
\overline{o}_\ell^{t} &= \sum_{k=0}^{H} \frac{n_{k\ell}^{t}}{n_\ell^{t}}\,\overline{o}_{k\ell}^{t},
\end{aligned}$$

respectivement ($1 \le h, \ell \le H$). D'où, les covariances données par (3) peuvent être réécrites

$$\mathrm{cov}(\overline{o}_h^{t-12}, \overline{o}_\ell^{t}) = \mathrm{cov}\left(\sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}}\,\overline{o}_{hg}^{t-12}, \sum_{k=0}^{H} \frac{n_{k\ell}^{t}}{n_\ell^{t}}\,\overline{o}_{k\ell}^{t}\right) \quad (4a)$$

$$= \frac{1}{n_h^{t-12} n_\ell^{t}}\mathrm{cov}(n_{h\ell}^{t-12}\,\overline{o}_{h\ell}^{t-12}, n_{h\ell}^{t}\,\overline{o}_{h\ell}^{t}), \quad (4b)$$

où nous avons utilisé $\mathrm{cov}(n_{hg}^{t-12}\,\overline{o}_{hg}^{t-12}, n_{k\ell}^{t}\,\overline{o}_{k\ell}^{t}) = 0$ ($k \ne h$) et $\mathrm{cov}(n_{hg}^{t-12}\,\overline{o}_{hg}^{t-12}, n_{h\ell}^{t}\,\overline{o}_{h\ell}^{t}) = 0$ ($g \ne \ell$). Cette dernière covariance est nulle parce que

$$\begin{aligned}
&\mathrm{cov}(n_{hg}^{t-12}\,\overline{o}_{hg}^{t-12}, n_{h\ell}^{t}\,\overline{o}_{h\ell}^{t}) \\
&\quad = E\,\mathrm{cov}(n_{hg}^{t-12}\,\overline{o}_{hg}^{t-12}, n_{h\ell}^{t}\,\overline{o}_{h\ell}^{t} \mid n_{hg}^{t-12}, n_{h\ell}^{t}) \\
&\qquad + \mathrm{cov}\{E(n_{hg}^{t-12}\,\overline{o}_{hg}^{t-12} \mid n_{hg}^{t-12}, n_{h\ell}^{t}), E(n_{h\ell}^{t}\,\overline{o}_{h\ell}^{t} \mid n_{hg}^{t-12}, n_{h\ell}^{t})\} \\
&\quad = 0 + \overline{O}_{hg}^{t-12}\,\overline{O}_{h\ell}^{t}\,\mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0.
\end{aligned}$$

À la dernière ligne, nous avons également utilisé le fait que, pour $1 \le g \le H+1$, ,

$$\mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0. \quad (5)$$

Pour une justification et les hypothèses qui sous-tendent (5), voir l'annexe A. En outre, à l'annexe A, nous proposons une autre méthode d'estimation quand cette variance n'est pas négligeable. La covariance dans (4b) peut s'exprimer comme

$$\begin{aligned}\operatorname{cov}(n_{h\ell}^{t-12}\,\bar{o}_{h\ell}^{t-12},\, n_{h\ell}^{t}\,\bar{o}_{h\ell}^{t}) &= E\{\operatorname{cov}(n_{h\ell}^{t-12}\,\bar{o}_{h\ell}^{t-12},\, n_{h\ell}^{t}\,\bar{o}_{h\ell}^{t} \mid \nu_{h\ell})\} \\ &\quad + \operatorname{cov}\{E(n_{h\ell}^{t-12}\,\bar{o}_{h\ell}^{t-12} \mid \nu_{h\ell}), E(n_{h\ell}^{t}\,\bar{o}_{h\ell}^{t} \mid \nu_{h\ell})\} \quad (6)\end{aligned}$$

où $\nu_{h\ell} = (n_{h\ell}^{t-12}, n_{h\ell}^{t-12,t}, n_{h\ell}^{t})$. La première composante du deuxième membre est

$$\begin{aligned}E\{\operatorname{cov}(n_{h\ell}^{t-12}\,\bar{o}_{h\ell}^{t-12},\, n_{h\ell}^{t}\,\bar{o}_{h\ell}^{t} \mid \nu_{h\ell})\} &= E\{n_{h\ell}^{t-12} n_{h\ell}^{t} \operatorname{cov}(\bar{o}_{h\ell}^{t-12},\, \bar{o}_{h\ell}^{t} \mid \nu_{h\ell})\} \\ &= E\left\{ n_{h\ell}^{t-12} n_{h\ell}^{t} \left(\frac{n_{h\ell}^{t-12,t}/n_{h\ell}^{t-12}}{n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}} \right) S_{h\ell}^{t-12,t} \right\}. \quad (7)\end{aligned}$$

À la dernière ligne, nous avons utilisé l'expression (26) de l'annexe B. En outre,

$$S_{h\ell}^{t-12,t} = \frac{1}{N_{h\ell}^{t-12,t}} \sum_{i=1}^{N_{h\ell}^{t-12,t}} (O_{h\ell i}^{t-12} - \bar{O}_{h\ell i}^{t-12})(O_{h\ell i}^{t} - \bar{O}_{h\ell}^{t}). \quad (8)$$

La deuxième composante du deuxième membre de (6) est égale à $\bar{O}_{h\ell}^{t-12}\,\bar{O}_{h\ell}^{t} \operatorname{cov}(n_{h\ell}^{t-12}, n_{h\ell}^{t}) = 0$ compte tenu de (5). Il découle donc de (4) et de (6) que

$$\operatorname{cov}(\bar{o}_{h}^{t-12}, \bar{o}_{\ell}^{t}) = E\left\{ \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h}^{t-12} n_{\ell}^{t}} \left(\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}} \right) S_{h\ell}^{t-12,t} \right\}. \quad (9)$$

3.3 Estimation du terme de covariance du taux de croissance annuel

L'expression (9) peut être estimée d'après l'échantillon chevauchant $s_{h\ell}^{t-12,t}$ par

$$\hat{\operatorname{cov}}(\bar{o}_{h}^{t-12}, \bar{o}_{\ell}^{t}) = \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h}^{t-12} n_{\ell}^{t}} \left(\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}} \right) \hat{S}_{h\ell,\mathrm{CHEV}}^{t-12,t}, \quad (10)$$

où

$$\hat{S}_{h\ell,\mathrm{CHEV}}^{t-12,t} = \frac{1}{n_{h\ell}^{t-12,t} - 1} \sum_{i=1}^{n_{h\ell}^{t-12,t}} (o_{h\ell i}^{t-12} - \bar{o}_{h\ell,\mathrm{CHEV}}^{t-12})(o_{h\ell i}^{t} - \bar{o}_{h\ell,\mathrm{CHEV}}^{t}).$$

Notons que (10) est sans biais pour l'estimation de (9) car

$$E(\hat{S}_{h\ell,\mathrm{CHEV}}^{t-12,t} \mid \nu_{h\ell}) = S_{h\ell}^{t-12,t}.$$

Bien que (10) produise des estimations raisonnables si $n_{h\ell}^{t-12,t}$ est suffisamment grand, un inconvénient de l'estimateur de la covariance $\hat{S}_{h\ell,\mathrm{CHEV}}^{t-12,t}$ dans (10) est que, si $n_{h\ell}^{t-12,t}$ est petit, l'estimation de $\operatorname{var}(\hat{O}^{t} - G^{t,t-12}\hat{O}^{t-12})$ dans le numérateur de (2) peut être négative. Rappelons que cette variance est estimée par

$$\begin{aligned}\hat{\operatorname{var}}(\hat{O}^{t} - G^{t,t-12}\hat{O}^{t-12}) &= \hat{\operatorname{var}}(\hat{O}^{t}) + (\hat{G}^{t,t-12})^2\,\hat{\operatorname{var}}(\hat{O}^{t-12}) \\ &\quad - 2\hat{G}^{t,t-12}\,\hat{\operatorname{cov}}(\hat{O}^{t}, \hat{O}^{t-12}). \quad (11)\end{aligned}$$

Donc, nous proposons un estimateur pour remplacer $\hat{S}_{h\ell,\mathrm{CHEV}}^{t-12,t}$ donné dans (10). Définissons les écarts-types

$$\hat{S}_{h\ell}^{t-m} = \sqrt{\frac{1}{n_{h\ell}^{t-m} - 1} \sum_{i=1}^{n_{h\ell}^{t-m}} (o_{h\ell i}^{t-m} - \bar{o}_{h\ell}^{t-m})^2} \quad (m = 0, 12).$$

Nous proposons pour $S_{h\ell}^{t-12,t}$ l'estimateur modifié suivant

$$\hat{S}_{h\ell}^{t-12,t} = \hat{\rho}_{h\ell,\mathrm{CHEV}}^{t-12,t}\,\hat{S}_{h\ell}^{t-12}\hat{S}_{h\ell}^{t}, \quad (12)$$

où $\rho_{h\ell}^{t-12,t}$ est la corrélation entre les variables o^{t} et o^{t-12} dans $U_{h\ell}^{t-12,t}$, et $\hat{\rho}_{h\ell,\mathrm{CHEV}}^{t-12,t}$ est son estimation d'après $s_{h\ell}^{t-12,t}$. Selon (10) et (12), la covariance (3) peut être estimée par

$$\hat{\operatorname{cov}}(\hat{O}^{t-12}, \hat{O}^{t}) = \sum_{h=1}^{H} \sum_{\ell=1}^{H} \frac{N_{h}^{t-12} N_{\ell}^{t}}{n_{h}^{t-12} n_{\ell}^{t}}\, n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}} \right) \hat{S}_{h\ell}^{t-12,t}. \quad (13)$$

Pour l'estimation $\hat{\rho}_{h\ell,\mathrm{CHEV}}^{t-12,t}$, l'expression $\hat{\rho}_{h\ell,\mathrm{CHEV}}^{t-12,t} \le 1$ est toujours vérifiée, tandis que l'utilisation de (10) peut mener implicitement à une corrélation estimée plus grande que 1 et un résultat de (11) éventuellement négatif. Voir la section suivante pour un exemple. Dans toutes les applications rencontrées jusqu'à présent, les résultats négatifs de (11) pouvaient être expliqués par le fait que, contrairement à (12), l'utilisation de (10) mène implicitement à une corrélation estimée supérieure à 1. Cela concorde avec la constatation de Berger (2004, page 462) qu'une surestimation de la corrélation entre $\hat{O}^{t-12}$ et $\hat{O}^{t}$ peut donner lieu à une sous-estimation importante de la variance d'une variation. Néanmoins, dans certaines circonstances extraordinaires, l'utilisation de (12) pourrait produire un résultat négatif de (11) également. Des conditions suffisantes pour que l'utilisation de (12) produise un estimateur de variance non négatif avec une probabilité de 1 peuvent être obtenues sur demande auprès des auteurs. Pour une revue générale des méthodes d'estimation de la variance dans les enquêtes-entreprises, voir Brodie (2003).

L'application de (12) peut donner lieu à un problème particulier quand $n_{h\ell}^{t} = 1$ ou $n_{h\ell}^{t-12} = 1$. Pour évaluer les variances d'échantillon requises, on peut emprunter la variance d'échantillon d'une sous-strate apparentée ou de la même sous-strate durant un mois antérieur. On peut aussi imputer une variance quand il se dégage des données une relation de la forme $S_{h\ell}^{2} \approx \sigma^{2}\bar{O}_{h\ell}^{\beta}$; voir Särndal, Swensson et Wretman (1992, page 461). En outre, le terme de covariance correspondant pourrait être ignoré quand sa contribution (prévue) à la variance totale est faible. Cela est souvent le cas quand les fractions d'échantillonnage dans les strates h et ℓ sont faibles, c'est-à-dire dans les strates contenant des unités assez petites et, par conséquent, ayant de faibles variances comparativement aux strates contenant des unités plus grandes. Des remarques similaires s'appliquent aux $\rho_{h\ell}^{t-12,t}$ imputées quand $n_{h\ell}^{t-12,t} \le 2$ et $n_{h\ell}^{t-m} \ge 2$ ($m = 0, 12$). Puisque les $\rho_{h\ell}^{t-12,t}$ sont souvent assez élevées, cela semble être un moyen viable. Dans l'exemple donné à la section 4, les $\rho_{h\ell}^{t-12,t}$ ont une moyenne globale de 0,90 et une variance de 0,0074, de sorte que l'effet des $\rho_{h\ell}^{t-12,t}$ imputées sur les résultats finaux est vraisemblablement modéré.

En outre, notons que, si $n_{h\ell}^{t-m} = 0$ ($m = 0$ ou $m = 12$), le terme de covariance correspondant dans (13) peut être négligé sans affecter son absence de biais, à condition que les $S_{h\ell}^{t-12,t}$ restants soient estimés de façon non biaisée. Sous cette hypothèse, un terme de ce genre avec $n_{h\ell}^{t-m} = 0$ ($m = 0$ ou $m = 12$) peut être négligé parce que l'espérance de

$$n_{h\ell}^{t-12,t}\left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right)\hat{S}_{h\ell}^{t-12,t} \qquad (14)$$

provenant de (13) est égale à

$$E\left[E\left\{n_{h\ell}^{t-12,t}\left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right)\hat{S}_{h\ell}^{t-12,t} \middle| \nu_{h\ell}\right\}\right] = E\left\{n_{h\ell}^{t-12,t}\left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right)S_{h\ell}^{t-12,t}\right\},$$

et l'espérance dans le deuxième membre est le paramètre qui doit être estimé. En outre, quand $n_{h\ell}^{t-m} = 0$ ($m = 0$ ou $m = 12$) et conséquemment $n_{h\ell}^{t-12,t} = 0$, le résultat de (14) est nul et l'estimateur $\hat{S}_{h\ell}^{t-12,t}$ de $S_{h\ell}^{t-12,t}$ devient sans pertinence. Par conséquent, le fait d'ignorer ce terme quand $n_{h\ell}^{t-m} = 0$ ($m = 0$ ou $m = 12$) n'a pas d'incidence sur les espérances de (13) et (14).

3.4 Une comparaison avec les résultats de Nordberg

En utilisant la formulation classique des indicateurs d'inclusion δ_{hi}^{t} pour chaque strate, Nordberg (2000) dérive une expression différente pour la première composante de (6). Cependant, on peut montrer, après certaines manipulations algébriques, que notre expression (9) est équivalente à celle de Nordberg (3.4) ; une preuve peut être obtenue sur demande auprès des auteurs. En outre, Nordberg établit une expression non nulle pour la deuxième composante de (6), c'est-à-dire la covariance entre les deux espérances conditionnelles correspondantes. Notons que le plan d'échantillonnage de la Suède diffère légèrement du nôtre.

Selon Nordberg (2000, page 370), l'estimation de la deuxième composante pour le plan d'échantillonnage de la Suède nécessite une procédure gourmande en ressources informatiques qui comprend la simulation du procédé d'échantillonnage. Cependant, puisque les $n_{h\ell}^{t}$, $n_{h\ell}^{t-12}$ et $n_{h\ell}^{t-12,t}$ sont tous des statistiques auxiliaires, une autre option pourrait consister à conditionner sur ces statistiques afin de pouvoir ignorer la deuxième composante. Rappelons qu'une statistique est dite auxiliaire quand sa distribution marginale ne dépend pas des paramètres cibles à estimer ; voir Cox et Hinkley (1974, pages 31-35). Une approche de rechange de ce genre sans la deuxième composante est recommandée en particulier quand $\hat{g}_{\mathrm{EASS}}^{t-12,t} \approx \hat{g}_{\mathrm{PS,sous}}^{t-12,t}$ où $\hat{g}_{\mathrm{PS,sous}}^{t-12,t}$ est l'estimateur poststratifié basé sur les sous-strates $h\ell$. Cependant, quand la différence entre $\hat{g}_{\mathrm{EASS}}^{t-12,t}$ et $\hat{g}_{\mathrm{PS,sous}}^{t-12,t}$ est non négligeable, le calcul de la variance inconditionnelle paraît indispensable, y compris l'estimation de la deuxième composante selon Nordberg. Pour une approche différente du problème d'estimation de la deuxième composante, voir l'annexe A.

Pour une justification de l'utilisation d'une (co)variance conditionnelle, voir Holt et Smith (1979). Un avantage important de la (co)variance conditionnelle est que l'intervalle de confiance correspondant possède de meilleures propriétés de couverture que celui fondé sur la variance inconditionnelle. Désignons l'intervalle de confiance à 95 % conditionnel classique pour un paramètre arbitraire θ par $(\hat{\theta}_l, \hat{\theta}_u | \nu)$, où ν désigne le vecteur constitué de toutes les statistiques (auxiliaires) qui interviennent dans les (co)variances conditionnelles. Alors, sous l'hypothèse de normalité et certaines conditions faibles, il est vérifié que le niveau de confiance (*NC*) à 95 % réel est égal au niveau de confiance nominal parce que

$$\begin{aligned} \mathrm{NC} &= \sum_{\nu\in\Omega_\nu} P(\nu) P(\hat{\theta}_l < \theta < \hat{\theta}_u | \nu) \\ &= 0{,}95 \sum_{\nu\in\Omega_\nu} P(\nu) = 0{,}95, \end{aligned}$$

où Ω_ν représente l'ensemble de tous les résultats possibles du vecteur aléatoire ν. Lorsqu'on utilise des (co)variances inconditionnelles, les intervalles de confiance obtenus peuvent être assez inexacts pour une répartition donnée de l'échantillon. En outre, quand la moyenne est calculée sur l'ensemble des répartitions, *NC* peut différer de 0,95 ; par exemple, voir Knottnerus (2003, pages 133-135). Notons qu'à l'étape de la planification avant le tirage de l'échantillon, les variances inconditionnelles sont toujours utiles pour examiner l'effet de plan de Kish en vue de comparer divers plans d'échantillonnage. De surcroît, notons que, pour évaluer un intervalle de confiance conditionnel pour $g^{t-12,t}$, les variances sous-jacentes de $\hat{O}_{\mathrm{PS,sub}}^{t-m}$ devraient également être prises conditionnellement à $\nu_{h\ell}$ ($m = 0, 12$).

Enfin, l'estimateur sans biais proposé par Nordberg [2000, équation (3.9)] pour la première composante de (6) est assez différent de ceux décrits à la sous-section précédente. En fait, son estimateur est fondé sur la procédure d'estimation du terme de covariance $S_{h\ell}^{t-12,t}$ qui suit. Premièrement, estimer la quantité sous-jacente $\sum_{i=1}^{N_{h\ell}^{t-12,t}} O_{h\ell i}^{t-12} O_{h\ell i}^{t}$ à partir de l'échantillon chevauchant $s_{h\ell}^{t-12,t}$. Deuxièmement, estimer les moyennes du chiffre d'affaires correspondantes d'après $s_{h\ell}^{t-12}$ et $s_{h\ell}^{t}$, respectivement. Puisque les composantes estimées de cette façon proviennent d'échantillons différents, il n'est pas toujours possible d'éviter que (11) donne un résultat négatif. Pour un petit exemple portant sur des données réelles, voir la section qui suit. Dans la suite de l'exposé, l'estimateur sous-jacent de Nordberg pour $S_{h\ell}^{t-12,t}$ est désigné par $\hat{S}_{h\ell\mathrm{NBG}}^{t-12,t}$. La détermination de l'expression explicite pour $\hat{S}_{h\ell\mathrm{NBG}}^{t-12,t}$ peut être obtenue sur demande auprès des auteurs.

4. Une application à la variation du chiffre d'affaires des supermarchés des Pays-Bas

4.1 Deux estimateurs de la variation annuelle du chiffre d'affaires

Afin de déterminer l'effet sur les estimateurs de la variance, il est important de savoir qu'en janvier, le chiffre d'affaires est estimé deux fois. La première estimation, désignée par $\hat{O}^{\text{janvA}}$ (avec *A* pour *ancienne*), est produite avant la mise à jour annuelle de l'échantillon et est utilisée pour estimer la variation mensuelle du chiffre d'affaires en janvier comparativement à celle de décembre. La deuxième estimation, désignée par $\hat{O}^{\text{janvN}}$ (avec *N* pour *nouvelle*), est faite après la mise à jour annuelle de l'échantillon et est utilisée pour estimer la variation mensuelle du chiffre d'affaires en février comparativement à janvier. Cette procédure implique que les unités de l'ancien échantillon ainsi que celles du nouvel échantillon reçoivent un questionnaire en janvier.

Contrairement à l'estimateur (1), l'estimateur réel utilisé par Statistics Netherlands pour la variation annuelle du chiffre d'affaires mensuel est fondé sur une chaîne de 12 variations mensuelles du chiffre d'affaires.

$$\begin{aligned}\hat{G}_{\text{réel}}^{t,t-12} &= 1 + \hat{g}_{\text{réel}}^{t,t-12} = \prod_{j=0}^{11}(1+\hat{g}^{t-j,t-j-1}) \\ &= \frac{\hat{O}^{t}}{\hat{O}^{t-1}} \times \frac{\hat{O}^{t-1}}{\hat{O}^{t-2}} \times \ \ldots \ \times \ \frac{\hat{O}^{\text{fév}}}{\hat{O}^{\text{janvN}}} \\ &\quad \times \frac{\hat{O}^{\text{janvA}}}{\hat{O}^{\text{déc}}} \times \ \ldots \ \times \ \frac{\hat{O}^{t-11}}{\hat{O}^{t-12}} \\ &= \frac{\hat{O}^{t}}{\hat{O}^{t-12}} \times \frac{\hat{O}^{\text{janvA}}}{\hat{O}^{\text{janvN}}} \quad (t \neq \text{janv}). \qquad (15)\end{aligned}$$

À la présente section, nous allons comparer les variances des estimateurs (1) et (15). De la même façon que (2), les formules de variance pour $\hat{g}_{\text{réel}}^{t,t-12}$ peuvent être établies par un développement en série de Taylor de premier ordre.

4.2 Description des données

Dans le présent exemple, les calculs des variances et des intervalles de confiance sont fondés sur les données sur le chiffre d'affaires des supermarchés des Pays-Bas pour des périodes de quatre semaines en 2003 et en 2004 (c'est-à-dire $t = 1, ..., 26$). Donc, il existe 13 observations pour une année et, conséquemment, nous utilisons des symboles légèrement adaptés, tels que $g^{t,t-13}$ dans la suite de la présente section.

La population est constituée d'environ 3 500 établissements. Les données sur le chiffre d'affaires proviennent d'un échantillon stratifié et de fichiers administratifs. Un échantillon EASS brut d'environ 900 unités stratifiées selon la taille est tiré de la liste complète des unités de population du RGE qui englobe également les unités des fichiers administratifs. Les établissements comptant 50 *employés* ou plus sont inclus avec une probabilité de 1. Les autres sont échantillonnés avec des probabilités d'inclusion décroissantes allant de 1:2 (20 à 49 employés par établissement) à 1:40 pour la taille la plus petite (1 employé par établissement). Les fichiers administratifs contiennent environ 950 unités, présentes dans toutes les classes de taille. Environ 500 des 900 unités de l'échantillon brut figuraient déjà dans les fichiers administratifs, mais ne recevaient pas de questionnaire. Donc, l'échantillon net contient environ 400 unités. En fait, dans cet exemple particulier, la taille d'échantillon de chaque strate est aléatoire. Cependant, comme nous l'avons expliqué à la sous-section 3.4, nous estimons toutes les (co)variances conditionnellement à n_h dans un tel cas. Les données des unités provenant des fichiers administratifs sont placées dans une strate distincte dont la fraction d'échantillonnage est égale à l'unité.

4.3 Résultats

Le tableau 1 donne les taux de croissance annuels et leur marge d'erreur à 95 % pour $t = 16, ..., 24$. Nous constatons que les marges d'erreur à 95 % des taux de croissance estimés $\hat{g}_{\text{réel}}^{t,t-13}$, utilisées à l'heure actuelle par Statistics Netherlands, varient entre 0,8 et 1,0 (point de pourcentage). Par exemple, à la première période ($t = 16$), l'intervalle de confiance à 95 % du taux de croissance annuel va de -1,3 à 0,7 %. Comme prévu, les marges d'erreur à 95 % pour l'estimateur plus compliqué $\hat{g}_{\text{réel}}^{t,t-13}$ sont proches de celles pour l'estimateur plus simple $\hat{g}^{t,t-13}$ provenant de (1). Les marges d'erreur à 95 % de $\hat{g}^{t,t-13}$ varient entre 0,7 et 1,0 (point de pourcentage). L'estimateur du taux de croissance qu'il convient de privilégier est $\hat{g}_{\text{réel}}^{t,t-13}$ car il comprend une correction pour la mise à jour annuelle de l'échantillon en janvier. L'estimation de sa variance peut toutefois être simplifiée en utilisant l'estimateur de la variance décrit à la section 3, plutôt que l'expression plus laborieuse pour $\hat{\text{var}}(\hat{g}_{\text{réel}}^{t,t-13})$.

Tableau 1
Taux de croissance estimés avec les marges d'erreur à 95 %

t	$\hat{g}_{\text{réel}}^{t,t-13} \times 100\%$	$\hat{g}^{t,t-13} \times 100\%$
16	-0,3 (± 1,0) [1]	-0,4 (± 1,0)
17	-3,7 (± 1,0)	-3,8 (± 0,9)
18	1,6 (± 1,0)	1,5 (± 0,9)
19	-2,2 (± 0,9)	-2,3 (± 0,9)
20	0,5 (± 0,8)	0,4 (± 0,7)
21	-1,7 (± 0,8)	-1,8 (± 0,7)
22	-2,2 (± 0,8)	-2,3 (± 0,7)
23	0,0 (± 0,8)	-0,1 (± 0,7)
24	-2,3 (± 0,9)	-2,4 (± 0,9)

[1] Les marges d'erreur à 95 % sont données entre parenthèses.

Comme il est décrit à la section 3, nous avons utilisé la corrélation estimée $\hat{\rho}_{h\ell,\mathrm{CHEV}}^{t-13,t}$ provenant de la partie chevauchante de l'échantillon $s_{h\ell}^{t-13,t}$ pour estimer la covariance $S_{h\ell}^{t-13,t}$ afin d'éviter que (11) donne des résultats négatifs. Knottnerus et Van Delden (2006) ont évalué le biais de $\hat{\rho}_{h\ell,\mathrm{CHEV}}^{t-13,t}$ pour les données sur les supermarchés des Pays-Bas et découvert une faible sous-estimation de $\hat{\rho}_{h\ell,\mathrm{CHEV}}^{t-13,t}$ résultant d'une légère surestimation, inférieure à 5 %, de $\mathrm{var}(\hat{g}^{\,t,t-13})$.

L'utilisation de l'estimateur $\hat{S}_{h\ell,\mathrm{CHEV}}^{t-13,t}$ dans (10) peut donner un résultat négatif pour (11) et une corrélation estimée $\hat{\rho}_{h\ell}^{t-13,t}$ plus grande que 1. Par exemple, considérons la population spécifique pour laquelle $N = 50$ et $H = 1$ composée des unités de la sous-strate $h\ell = 65$. Au moyen des données de panel pour cette population, présentées au tableau 2 pour $t = 3$ et $t = 16$, nous obtenons après certains calculs $\hat{S}^{\,t-13} = 410,7$, $\hat{S}^{\,t} = 394,3$ et $\hat{G}^{\,t,t-13} = 1,028$. Notons que, dans la suite de la présente section, l'indice $h\ell = 11$ est omis dans les symboles parce qu'il n'existe qu'une seule strate. Le tableau 3 donne, pour trois approches différentes, certaines estimations supplémentaires pour les données de panel du tableau 2. Par exemple, l'utilisation de $\hat{S}_{\mathrm{CHEV}}^{t-13,t}$ dans (10) produit une corrélation estimée $\hat{\rho}^{\,t-13,t} =$ 1,39. Celle-ci donne alors, en utilisant (11), une estimation négative de la variance égale à moins 2,2 millions. Similairement, pour les mêmes données, l'estimateur de rechange $\hat{S}_{\mathrm{NBG}}^{t-13,t}$ de $S^{\,t-13,t}$ fondé sur Nordberg (2000) produit pour (11) le résultat moins 36,9 millions, parce que l'estimation correspondante $\hat{\rho}_{\mathrm{NBG}}^{t-13,t}$ devient 1,64. En revanche, l'utilisation de la corrélation estimée d'après l'échantillon chevauchant $s^{\,t-13,t}$ conformément à (12) donne $\hat{\rho}_{\mathrm{CHEV}}^{t-13,t} =$ 0,9997 et l'estimation de la variance au moyen de (11) devient la valeur positive de 52,1 millions. En outre, pour les données de panel du tableau 2, le résultat de l'estimateur de Nordberg (3.9) pour la covariance entre $\hat{O}^{\,t-13}$ et $\hat{O}^{\,t}$ est égal à 111,1 millions, tandis que l'estimateur de covariance (13) proposé donne 67,8 millions.

Tableau 2
Données de panel[1] provenant d'une population avec $N = 50$ et $H = 1$

Période	Chiffre d'affaires par unité (en milliers d'euros)				
	1	2	3	4	5
$t = 3$		493,9	264,3	1 179,1	380,0
$t = 16$	475,3	472,0	267,0	1 169,0	

[1] En réalité, les données de panel appartiennent à la sous-strate $h\ell = 65$.

5. Conclusion

Les formules de variance obtenues dans le présent article sont utiles pour calculer la variance du taux de croissance annuel estimé du chiffre d'affaires mensuel. L'utilisation de (13) comme estimateur de $\mathrm{cov}(\hat{O}^{\,t-12}, \hat{O}^{\,t})$ produit, en particulier, des estimations raisonnables de la covariance des variations. La procédure d'estimation de la variance tient compte de l'utilisation de panels rotatifs, des nouvelles unités, des unités disparues et des unités qui migrent entre les strates.

En outre, nous recommandons d'estimer une covariance de population au moyen de l'estimateur (12) fondé sur la corrélation correspondante estimée d'après l'échantillon chevauchant, ainsi que sur les variances correspondantes estimées d'après de grands échantillons distincts. Cela peut aider à éviter une sous-estimation importante ou la production d'un résultat négatif par l'estimateur de la variance du taux de croissance annuel. Les covariances estimées résultantes ne présentent qu'un léger biais.

Tableau 3
Estimations selon trois approches différentes

Approche		Paramètres à estimer		
		$S^{\,t-13,t}$	$\rho^{\,t-13,t}$	$\mathrm{var}(\hat{O}^{\,t} - G^{\,t,t-13}\hat{O}^{\,t-13})$
Nordberg (2000)	estimateur	$\hat{S}_{\mathrm{NBG}}^{t-13,t}$	$\dfrac{\hat{S}_{\mathrm{NBG}}^{t-13,t}}{\hat{S}^{\,t-13}\hat{S}^{\,t}}$	Éq. (11)
	résultat	$265,2 \times 10^3$	1,64	$-36,9 \times 10^6$
Éq. (10)	estimateur	$\hat{S}_{\mathrm{CHEV}}^{t-13,t}$	$\dfrac{\hat{S}_{\mathrm{CHEV}}^{t-13,t}}{\hat{S}^{\,t-13}\hat{S}^{\,t}}$	Éq. (11)
	résultat	$225,0 \times 10^3$	1,39	$-2,2 \times 10^6$
Éq. (12)	estimateur	$\hat{\rho}_{\mathrm{CHEV}}^{t-13,t}\,\hat{S}^{\,t-13}\hat{S}^{\,t}$	$\hat{\rho}_{\mathrm{CHEV}}^{t-13,t}$	Éq. (11)
	résultat	$161,9 \times 10^3$	1,00[1]	$52,1 \times 10^6$

[1] En fait, 0,9997.

Pour le plan d'échantillonnage des supermarchés des Pays-Bas, le deuxième terme de covariance dans (6) est négligeable parce que $n_{h\ell,\,\mathrm{REQ}}^{\text{déc, janv}}$ est fixé. Par contre, pour le plan d'échantillonnage SAMU utilisé en Suède, ce terme est non négligeable et son estimation prend beaucoup de temps ; le terme SAMU (SAMordnade Urval) est un acronyme suédois signifiant échantillons coordonnés. À l'annexe A, nous proposons une méthode de rechange pour estimer cette covariance. Cependant, sous la condition que $\hat{g}^{\,t,t-12} \approx \hat{g}^{\,t,t-12}_{\mathrm{PS,sous}}$, il suffit à notre avis d'utiliser uniquement la première covariance. Cela simplifie considérablement la procédure d'estimation. En outre, sous l'hypothèse de normalité, l'intervalle de confiance conditionnel possède de meilleures propriétés de couverture que l'intervalle inconditionnel.

L'exemple des supermarchés des Pays-Bas illustre l'une des applications pratiques des formules de variance, à savoir déterminer quel estimateur possède la variance la plus faible. Les résultats confirment que la variance de l'estimateur simple $\hat{g}^{\,t,t-13}$ est proche de celle de l'estimateur $\hat{g}^{\,t,t-13}_{\text{réel}}$ de la section 4 qui comprend une correction pour tenir compte du renouvellement de l'échantillon en janvier. Donc, pour les supermarchés des Pays-Bas, $\hat{\mathrm{var}}(\hat{g}^{\,t,t-13})$ pourrait être utilisé pour estimer $\mathrm{var}(\hat{g}^{\,t,t-13}_{\text{réel}})$. Pour des branches d'activité ayant un autre code de la CTI, il convient de vérifier si $\mathrm{var}(\hat{g}^{\,t,t-13}_{\text{réel}}) \approx \mathrm{var}(\hat{g}^{\,t,t-13})$, puisque l'effet du renouvellement de l'échantillon en janvier ne doit pas nécessairement être négligeable.

Remerciements

Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas forcément les politiques de Statistics Netherlands. Les auteurs remercient le rédacteur associé et deux examinateurs anonymes de leurs commentaires et suggestions constructifs qui leur ont permis d'améliorer considérablement l'article.

Annexe A

Justification de (5)

Premièrement, considérons le cas des strates sans nouvelles unités et sans unités disparues. À part la mise à jour annuelle de janvier, il n'y a dans ces conditions aucune mise à jour mensuelle. D'où, $n_{h\ell}^{t} = n_{h\ell,\,\mathrm{REQ}}^{\text{déc, janv}}$ est fixe et il en découle (5). Ce cas s'applique aux supermarchés des Pays-Bas, parce que cette population est demeurée assez stable au fil des ans. Deuxièmement, dans le cas des nouvelles unités et des unités disparues dans les strates, nous pouvons écrire $n_{h\ell}^{t}$ sous la forme

$$n_{h\ell}^{t} = n_{\ell}^{t} - n_{0\ell}^{t-12,\,t} - \sum_{k\neq h}^{H} n_{k\ell}^{t}, \qquad (16)$$

où $n_{0\ell}^{t-12,\,t}$ ou, plus brièvement, $n_{0\ell}^{t}$ est le nombre de nouvelles unités créées durant les mois $t-12, \ldots, t-1$ parmi s_{ℓ}^{t}. Comme la procédure d'échantillonnage des nouvelles unités après le mois $t-12$ est indépendante de n_{hg}^{t-12}, les variables aléatoires $n_{0\ell}^{t-12,\,t}$ et n_{hg}^{t-12} ont une covariance nulle. En outre, en utilisant $\mathrm{cov}(n_{hg}^{t-12}, n_{k\ell}^{t}) = 0$ pour $k \neq h$, il découle de (16) que $\mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0$ $(h = 1, \ldots, H)$.

En fait, nous avons supposé jusqu'à présent que la distribution de $n_{k\ell}^{t}$ $(k \neq h)$ peut être décrite par une loi hypergéométrique ayant les paramètres $(N_{\ell}^{t}, N_{k\ell}^{t-12,\,t}, n_{\ell}^{t})$, quelles que soient les valeurs de n_{hg}^{t-12}. Une remarque similaire s'applique à $n_{0\ell}^{t-12,\,t}$. Cependant, nous pouvons soutenir qu'en pratique, ces hypothèses donnent lieu à une erreur mineure, de deuxième ordre, dans les formules de variance. Afin de dépister cette erreur, nous supposons pour simplifier et sans perte de généralité que *i*) les nouvelles unités et les unités disparues ne migrent pas entre les strates, *ii*) il n'y a pas de disparitions parmi les nouvelles unités, *iii*) $n_{0h}^{t} = f_h N_{0h}^{t}$ est fixe, *iv*) après leur premier mois dans la population, les nouvelles unités sont sans pertinence pour les mises à jour mensuelles durant le reste de la période étudiée et *v*) les unités disparues ne sont pas sélectionnées dans l'échantillon ni supprimées de celui-ci lors des mises à jour mensuelles ; donc, une erreur de troisième ordre demeure ignorée. Sous ces hypothèses, nous examinons maintenant de plus près la seconde composante de la covariance pour $\ell = h$, disons $C_{hh,\,\mathrm{sec}}$, provenant de (4a). Par analogie avec (6), $C_{hh,\,\mathrm{sec}}$ peut s'écrire

$$\begin{aligned} C_{hh,\,\mathrm{sec}} &\equiv \frac{1}{n_h^{t-12} n_h^{t}} \\ &\quad \mathrm{cov}\left\{ E\left(\sum_{g=1}^{H+1} n_{hg}^{t-12}\, \bar{o}_{hg}^{\,t-12} \,|\, \nu_h \right), E\left(\sum_{k=0}^{H} n_{kh}^{t}\, \bar{o}_{kh}^{\,t} \,|\, \nu_h \right) \right\} \\ &= \frac{1}{n_h^{t-12} n_h^{t}} \sum_{g=1}^{H+1} \sum_{k=1}^{H} \bar{O}_{hg}^{\,t-12}\, \bar{O}_{kh}^{\,t}\, \mathrm{cov}(n_{hg}^{t-12}, n_{kh}^{t}), \end{aligned} \qquad (17)$$

où $\nu_h = (n_{h1}^{t-12}, \ldots, n_{h,H+1}^{t-12}, n_{1h}^{t}, \ldots, n_{Hh}^{t})$. Notons que, sous les hypothèses susmentionnées, $C_{h\ell,\mathrm{sec}} = 0$ pour $\ell \neq h$.

Afin d'estimer les covariances dans (17), considérons la formule de l'espérance conditionnelle de y sachant $x = x_0$, quand y et x suivent une loi normale bivariée. Autrement dit, en notation classique,

$$E(y \,|\, x_0) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2}(x_0 - \mu_x).$$

De plus, pour une variation donnée Δx_0 de x, l'espérance conditionnelle de la variation de y est égale à $E(\Delta y \,|\, \Delta x_0) = \sigma_{yx} \Delta x_0 / \sigma_x^2$ ou, de manière équivalente, à

$$\sigma_{yx} = \frac{E(\Delta y \mid \Delta x_0)}{\Delta x_0}\sigma_x^2. \qquad (18)$$

Donc, pour estimer, par exemple, $\mathrm{cov}(n_{h,H+1}^{t-12}, n_{kh}^{t})$ dans (17) sous l'hypothèse de normalité, il suffit d'évaluer l'effet prévu sur $y = n_{kh}^{t}$ causé par une variation des futures unités disparues $x = n_{h,H+1}^{t-12}$ dans s_h^{t-12}.

Soit $\Delta n_{h,H+1}^{t-12}$ une variation supplémentaire (positive) de ces unités disparues dans s_h^{t-12}. Définissons $p_{h,H+1}^{\mathrm{janv},t}$ par $p_{h,H+1}^{\mathrm{janv},t} = N_{h,H+1}^{\mathrm{janv},t} / N_{h,H+1}^{t-12}$ où $N_{h,H+1}^{\mathrm{janv},t}$ est le nombre d'unités disparues dans la strate h entre janvier et le mois t. En outre, $p_{hg}^{t-12} = N_{hg}^{t-12,t} / N_h^{t-12}$ $(g = 1, ..., H+1)$. En utilisant l'hypothèse (v), le nombre prévu d'unités disparues supplémentaires dans l'échantillon de janvier avant le renouvellement peut être estimé par $p_{h,H+1}^{\mathrm{janv},t}\Delta n_{h,H+1}^{t-12}$. D'où, le nombre prévu d'unités disparues supplémentaires dans l'échantillon après le renouvellement peut être estimé par

$$\gamma_{\text{réd}}^{\mathrm{janv}}\ p_{h,H+1}^{\mathrm{janv},t}\,\Delta n_{h,H+1}^{t-12}\ ;$$
$$\gamma_{\text{réd}}^{\mathrm{janv}} = (0{,}9 - f_h)/(1 - f_h), \qquad (19)$$

où $\gamma_{\text{réd}}^{\mathrm{janv}}$ est le facteur de réduction dû au renouvellement de l'échantillon en janvier. Pour l'établissement de l'expression (19), voir la fin de cette annexe. Les mises à jour mensuelles correspondantes entre janvier et le mois t dues à ces unités disparues supplémentaires dans l'échantillon provenant de la strate h mènent à l'estimation suivante de l'augmentation prévue du nombre d'unités *entrantes* n_{kh}^{t} provenant de la strate k $(k \neq h)$ dans l'échantillon du mois t

$$E(\Delta n_{kh}^{t} \mid \Delta n_{h,H+1}^{t-12}) = \gamma_{\text{red}}^{\mathrm{janv}}\ p_{h,H+1}^{\mathrm{janv},t}\,\Delta n_{h,H+1}^{t-12}\,p_{kh}^{t}, \qquad (20)$$

où $p_{kh}^{t} = N_{kh}^{t-12,t} / (N_h^{t} - N_{0h}^{t})$. Rappelons, d'après la sous-section 2.1, qu'une mise à jour au mois s a lieu uniquement quand $d_h^{s-1} \neq f_h D_h^{s-1}$, où D_h^{s} (d_h^{s}) représente le nombre d'unités disparues dans U_h^{s} (s_h^{s}), et que $n_{kh}^{t} = f_h N_{kh}^{t-12,t}$ est fixé quand $N_{h,H+1}^{\mathrm{janv},t} = 0$ $(k \neq h)$ est remplacé. De surcroît, notons que les nouvelles unités sont exclues de la définition de p_{kh}^{t} dans (20) à cause de l'hypothèse (iv).

Ensuite, définissons pour $m = 0, 12$

$$\bar{O}_h^{t-m} = \frac{1}{N_h^{t-m}}\sum_{i=1}^{N_h^{t-m}} O_{hi}^{t-m}\ ;$$

$$(S_h^{t-m})^2 = \frac{1}{N_h^{t-m} - 1}\sum_{i=1}^{N_h^{t-m}} (O_{hi}^{t-m} - \bar{O}_h^{t-m})^2\ ;$$

$$p_{h,\le H}^{t-12} = 1 - p_{h,H+1}^{t-12}\ ;$$

$$p_{in,h}^{t} = 1 - p_{hh}^{t}\ ;$$

$$\bar{O}_{h,\le H}^{t-12} = \sum_{g=1}^{H}\frac{p_{hg}^{t-12}}{p_{h,\le H}^{t-12}}\,\bar{O}_{hg}^{t-12}\ ;$$

$$\bar{O}_{in,h}^{t} = \sum_{\substack{k=1\\k\neq h}}^{H}\frac{p_{kh}^{t}}{p_{in,h}^{t}}\,\bar{O}_{kh}^{t}.$$

Maintenant, en nous servant de (18) et (20), nous obtenons pour $k \neq h$ l'approximation de la covariance suivante

$$\begin{aligned}
&\mathrm{acov}(n_{h,H+1}^{t-12}, n_{kh}^{t}) \\
&\quad = \frac{E(\Delta n_{kh}^{t} \mid \Delta n_{h,H+1}^{t-12})}{\Delta n_{h,H+1}^{t-12}}\,\mathrm{var}(n_{h,H+1}^{t-12}) \\
&\quad \approx \gamma_{\text{réd}}^{\mathrm{janv}}\ p_{h,H+1}^{\mathrm{janv},t}\,p_{kh}^{t}\,n_h^{t-12}\,p_{h,H+1}^{t-12}(1 - p_{h,H+1}^{t-12})(1 - f_h) \\
&\quad = n_h^{t-12}\,p_{kh}^{t}A_h / p_{in,h}^{t}\ ; \qquad (21)
\end{aligned}$$

$$A_h = \gamma_{\text{réd}}^{\mathrm{janv}}\ p_{in,h}^{t}\,p_{h,H+1}^{\mathrm{janv},t}\,p_{h,H+1}^{t-12}(1 - p_{h,H+1}^{t-12})(1 - f_h),$$

où, pour simplifier, nous avons omis le terme $N_h^{t-12}/(N_h^{t-12} - 1)$ à la deuxième ligne. Comme n_h^{t-12} est fixé, l'expression $\mathrm{cov}(n_{h,H+1}^{t-12}, n_{kh}^{t}) = -\mathrm{cov}(n_{h1}^{t-12} + ... + n_{hH}^{t-12}, n_{kh}^{t})$ est vérifiée. D'où, par analogie avec la loi multihypergéométrique, nous pouvons utiliser pour $1 \le g \le H$ et $k \neq h$ la relation suivante comme approximation de $\mathrm{cov}(n_{hg}^{t-12}, n_{kh}^{t})$

$$\mathrm{acov}(n_{hg}^{t-12}, n_{kh}^{t}) = -\frac{p_{hg}^{t-12}}{p_{h,\le H}^{t-12}}\,\mathrm{acov}(n_{h,H+1}^{t-12}, n_{kh}^{t}) \qquad (22a)$$

$$= -n_h^{t-12}\,\frac{p_{hg}^{t-12}}{p_{h,\le H}^{t-12}}\,\frac{p_{kh}^{t}}{p_{in,h}^{t}}A_h, \qquad (22b)$$

où (21) est utilisée également. Ou bien, notons que

$$\begin{aligned}
\mathrm{cov}(n_{h,H+1}^{t-12}, n_{kh}^{t}) &= -\mathrm{cov}(n_{h,\le H}^{t-12}, n_{kh}^{t}) \\
&= -\sum_{g\le H}\ \sum_{i\in U_{hg}^{t-12,t}} \mathrm{cov}(\delta_{hgi}^{t-12}, n_{kh}^{t}),
\end{aligned}$$

où

$$\delta_{hgi}^{t-12} = \begin{cases} 1 & \text{si la } i^{\text{e}} \text{ unité dans } U_{hg}^{t-12,t} \\ & \text{est incluse dans l'échantillon } s_h^{t-12} \\ 0 & \text{autrement.} \end{cases}$$

Donc, par symétrie, $\mathrm{cov}(\delta_{hgi}^{t-12}, n_{kh}^{t}) = -\mathrm{cov}(n_{h,H+1}^{t-12}, n_{kh}^{t})/N_{h,\le H}^{t-12,t}$ d'où découle l'expression (22a) $(1 \le g \le H)$. De même, pour $k = h$, nous obtenons d'après (21) et (22b)

$$\begin{aligned}
\mathrm{acov}(n_{h,H+1}^{t-12}, n_{hh}^{t}) &= -n_h^{t-12}A_h\ ; \\
\mathrm{acov}(n_{hg}^{t-12}, n_{hh}^{t}) &= n_h^{t-12}\,p_{hg}^{t-12}A_h / p_{h,\le H}^{t},
\end{aligned} \qquad (23)$$

respectivement $(1 \le g \le H)$. Maintenant, en substituant (21) à (23) dans (17), nous obtenons l'approximation

$$C_{hh,\,\mathrm{sec}} = A_h(\bar{O}_{h,H+1}^{t-12} - \bar{O}_{h,\le H}^{t-12})(\bar{O}_{in,h}^{t} - \bar{O}_{hh}^{t})/n_h^{t}. \qquad (24)$$

En supposant que les deux termes entre parenthèses dans (24) sont absolument plus petits que S_h^{t}, il découle de (24) que

$$\left|C_{hh,\,\mathrm{sec}}\right| \le \frac{\gamma_{\text{réd}}^{\mathrm{janv}}\ p_{h,H+1}^{\mathrm{janv},t}\,p_{in,h}^{t}\,p_{h,H+1}^{t-12}(1 - p_{h,H+1}^{t-12})(1 - f_h)}{n_h^{t}}(S_h^{t})^2.$$

Donc, quand $p_{in,h}^{t}, p_{h,H+1}^{t-12} \le 0{,}1$, nous pouvons conclure que, sous les hypothèses susmentionnées, la contribution de la seconde composante de la covariance est inférieure à 1 %

de $\mathrm{var}(\bar{o}_h^t)$, de sorte que (5) peut être utilisée sans affecter gravement les résultats. Quand $C_{hh,\mathrm{sec}}$ est non négligeable, elle peut être estimée d'après l'échantillon conformément à (24) par

$$\hat{C}_{hh,\mathrm{sec}} = A_h\{(\bar{o}_{h,H+1}^{t-12} - \bar{o}_{h,\le H}^{t-12})(\bar{o}_{in,h}^{t} - \bar{o}_{hh}^{t}) - \hat{\mathrm{cov}}(\bar{o}_{h,\le H}^{t-12}, \bar{o}_{hh}^{t})\}/n_h^t, \quad (25)$$

où, par analogie avec (10) et (12), $\hat{\mathrm{cov}}(\bar{o}_{h,\le H}^{t-12}, \bar{o}_{hh}^{t})$ est définie par

$$\hat{\mathrm{cov}}(\bar{o}_{h,\le H}^{t-12}, \bar{o}_{hh}^{t}) = \frac{n_{hh}^{t-12}}{n_{h,\le H}^{t-12}}\left(\frac{n_{hh}^{t-12,t}}{n_{hh}^{t-12} n_{hh}^{t}} - \frac{1}{N_{hh}^{t-12,t}}\right)\hat{\rho}_{hh,\mathrm{CHEV}}^{t-12,t}\hat{S}_{hh}^{t-12}\hat{S}_{hh}^{t}.$$

Nous avons utilisé dans (25) le fait que *i*) pour deux estimateurs arbitraires (sans biais) $\hat{a}$ et $\hat{b}$, $E(\hat{a}\hat{b}) = ab + \mathrm{cov}(\hat{a}, \hat{b})$ et *ii*) $\mathrm{cov}(\bar{o}_{hg}^{t-12}, \bar{o}_{kh}^{t}) = 0$ ($g \neq h$ ou $k \neq h$).

Nous concluons cette annexe par la détermination de (19). Le nombre prévu d'unités disparues supplémentaires restant dans l'échantillon de janvier durant le renouvellement est $0{,}9\, p_{h,H+1}^{\mathrm{janv},t}\, \Delta n_{h,H+1}^{t-12}$. Le nombre d'unités disparues en dehors de l'échantillon juste avant le renouvellement peut être estimé par $N_{h,H+1}^{\mathrm{janv},t} - n_{h,H+1}^{\mathrm{janv}} - p_{h,H+1}^{\mathrm{janv},t}\Delta n_{h,H+1}^{t-12}$. Donc, le nombre de *nouvelles* unités disparues dans l'échantillon dues au renouvellement dans toutes les sous-strates $U_{hg}^{t-12,t}$ $(1 \le g \le H)$ en janvier peut être est estimé par

$$0{,}1(n_h^{\mathrm{janv}} - n_{0h}^{\mathrm{janv}})\frac{N_{h,H+1}^{\mathrm{janv},t} - n_{h,H+1}^{\mathrm{janv}} - p_{h,H+1}^{\mathrm{janv},t}\Delta n_{h,H+1}^{t-12}}{N_h^{\mathrm{janv}} - N_{0h}^{t-12,\,\mathrm{janv}} - (n_h^{\mathrm{janv}} - n_{0h}^{\mathrm{janv}})}.$$

Maintenant, en utilisant $n_{0h}^{\mathrm{janv}} = f_h\, N_{0h}^{t-12,\,\mathrm{janv}}$ conformément aux hypothèses susmentionnées, nous voyons qu'après le renouvellement, le nombre final d'unités disparues supplémentaires dans l'échantillon dues à $\Delta n_{h,H+1}^{t-12}$ peut être estimé par

$$\left\{0{,}9 - \frac{0{,}1(n_h^{\mathrm{janv}} - n_{0h}^{\mathrm{janv}})}{N_h^{\mathrm{janv}} - N_{0h}^{t-12,\,\mathrm{janv}} - (n_h^{\mathrm{janv}} - n_{0h}^{\mathrm{janv}})}\right\} p_{h,H+1}^{\mathrm{janv},t}\Delta n_{h,H+1}^{t-12} = \frac{0{,}9 - f_h}{1 - f_h}\, p_{h,H+1}^{\mathrm{janv},t}\Delta n_{h,H+1}^{t-12} = \gamma_{\mathrm{red}}^{\mathrm{janv}}\, p_{h,H+1}^{\mathrm{janv},t}\Delta n_{h,H+1}^{t-12}.$$

Annexe B

Quelques formules de covariance utiles pour les échantillons chevauchants

Soit s_{123} un échantillon mère constitué de trois sous-échantillons EAS mutuellement disjoints s_1, s_2 et s_3. Soit la variable x observée dans s_{12} et la variable y, dans s_{23}. Les moyennes d'échantillon correspondantes sont désignées par $\bar{x}_{12}$ et $\bar{y}_{23}$, respectivement. Désignons la taille de s_k par n_k ($k = 1, 2, 3, 12, 23$). Définissons $\lambda = n_2/n_{12}$, $\mu = n_2/n_{23}$ et $f_k = n_k/N$. En outre, définissons S_{xy} par

$$S_{xy} = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \bar{X}_p)(Y_j - \bar{Y}_p).$$

Alors, la covariance entre $\bar{x}_{12}$ et $\bar{y}_{23}$ est égale à

$$\mathrm{cov}(\bar{x}_{12}, \bar{y}_{23}) = \left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right)S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N}\right)S_{xy} = \left(\frac{\lambda}{n_{23}} - \frac{1}{N}\right)S_{xy}. \quad (26)$$

Cela peut être démontré comme il suit :

$$\begin{aligned}
&\mathrm{cov}(\bar{x}_{12}, \bar{y}_{23}) \\
&\quad = \mathrm{cov}\{(1-\lambda)\bar{x}_1 + \lambda\bar{x}_2, \mu\bar{y}_2 + (1-\mu)\bar{y}_3\} \\
&\quad = (1-\lambda)\mathrm{cov}(\bar{x}_1, \bar{y}_{23}) + \lambda\mu\,\mathrm{cov}(\bar{x}_2, \bar{y}_2) \\
&\qquad + \lambda(1-\mu)\mathrm{cov}(\bar{x}_2, \bar{y}_3) \\
&\quad = -(1-\lambda)\frac{S_{xy}}{N} + \lambda\mu\left(\frac{1}{n_2} - \frac{1}{N}\right)S_{xy} - \lambda(1-\mu)\frac{S_{xy}}{N} \\
&\quad = \left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right)S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N}\right)S_{xy} = \left(\frac{\lambda}{n_{23}} - \frac{1}{N}\right)S_{xy}.
\end{aligned}$$

À la troisième ligne, nous avons utilisé le fait que $\mathrm{cov}(\bar{x}_1, \bar{y}_{23}) = \mathrm{cov}(\bar{x}_2, \bar{y}_3) = -S_{xy}/N$. Cela découle de la formule de la variance conditionnelle

$$\begin{aligned}
\mathrm{cov}(\bar{x}_2, \bar{y}_3) &= E\{\mathrm{cov}(\bar{x}_2, \bar{y}_3|s_2)\} + \mathrm{cov}\{E(\bar{x}_2|s_2), E(\bar{y}_3|s_2)\} \\
&= 0 + \mathrm{cov}\left\{\bar{x}_2, \frac{\bar{Y}_p - f_2\bar{y}_2}{1 - f_2}\right\} \\
&= -\frac{f_2}{1 - f_2}\,\mathrm{cov}(\bar{x}_2, \bar{y}_2) = -\frac{S_{xy}}{N}.
\end{aligned}$$

Pour une autre preuve fondée sur le coefficient d'auto-corrélation d'échantillonnage, voir Knottnerus (2003, page 375).

Bibliographie

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics*, 32, 451-467.

Brodie, P. (2003). Review of recent work on variance estimation methods in business surveys. Rapport non-publié, Office for National Statistics, Londres.

Cox, D.R., et Hinkley, D.V. (1974). *Theoretical Statistics.* Londres : Chapman and Hall.

Hidiroglou, M.A., Särndal, C.-E. et Binder, D.A. (1995). Weighting and estimation in business surveys. Dans *Business Survey Methods*, (Éds., B.G. Cox et coll.). New York : John Wiley & Sons, Inc.

Holt, D., et Skinner, C.J. (1989). Components of change in repeated surveys. *Revue Internationale de Statistique*, 57, 1-18.

Holt, D., et Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, A, 142, 33-46.

Kish, L. (1965). *Survey sampling.* New York : John Wiley & Sons, Inc.

Knottnerus, P. (2003). *Sample Survey Theory*: *Some Pythagorean Perspectives*. New York : Springer-Verlag.

Knottnerus, P., et Van Delden, A. (2006). Estimation of changes in repeated surveys and their significance, http://www.iser.essex.ac.uk/ulsc/mols2006/programme/data/paper/Knottnerus.doc.

Konschnik, C.A., Monsour, N.J. et Detlefsen, R.E. (1985). Constructing and maintaining frames and samples for business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 113-122.

Laniel, N. (1987). Variances for a rotating sample from a changing population. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 496-500.

Lowerre, J.M. (1979). Sampling for change. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 343-347.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

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, 193-201.

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

Smith, P., Pont, M. et Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994-2000. *The Statistician*, 52, 257-295.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, 288-289.

Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, 24, 53-78.

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 57-67
Statistique Canada, N° 12-001-X au catalogue

Facteurs d'inflation de la variance dans l'analyse des données d'enquêtes complexes

Dan Liao et Richard Valliant [1]

Résumé

Les données d'enquêtes servent souvent à ajuster des modèles de régression linéaire. Les valeurs des covariables utilisées dans la modélisation n'étant toutefois pas contrôlées comme elles pourraient l'être dans une expérience, la colinéarité entre les covariables est un problème inévitable dans l'analyse des données d'enquêtes. Même si de nombreux livres et articles ont décrit le problème de la colinéarité et proposé des stratégies en vue de comprendre, d'évaluer et de traiter sa présence, la littérature sur les méthodes d'enquête n'a livré aucun outil diagnostique approprié pour évaluer son incidence sur l'estimation par la régression quand il est tenu compte de la complexité de l'enquête. Nous avons élaboré des facteurs d'inflation de la variance qui mesurent l'augmentation (« l'inflation ») de la variance des estimateurs des paramètres attribuable au fait que les variables explicatives ne sont pas orthogonales. Les facteurs d'inflation de la variance conviennent pour les estimateurs par la régression pondérée par les poids de sondage et tiennent compte des caractéristiques du plan de sondage complexe, par exemple, les pondérations, les grappes et les strates. Ces méthodes sont illustrées en utilisant un échantillon probabiliste provenant d'une enquête-ménage sur la santé et la nutrition.

Mots clés : Échantillon en grappes ; diagnostics de colinéarité ; estimateur de la variance par linéarisation ; moindres carrés pondérés par les poids de sondage ; échantillon stratifié.

1. Introduction

Dans une régression linéaire, la colinéarité des variables explicatives s'entend d'une situation où ces variables sont corrélées les unes aux autres. Les termes multicolinéarité et mauvais conditionnement sont également utilisés pour désigner ce genre de situation. La colinéarité est préoccupante pour des raisons tant numériques que statistiques. Les estimations des coefficients de pente peuvent être numériquement instables dans certains jeux de données en ce sens que de petites variations dans les $\mathbf{X}$ ou les $\mathbf{Y}$ peuvent produire de grandes variations dans les valeurs de ces estimations. Statistiquement, la corrélation entre les variables explicatives peut mener à des estimations de pente dont la variance est grande. En outre, quand les $\mathbf{X}$ sont fortement corrélés, le R^2 d'une régression peut être grand alors que les estimations de pente individuelles ne sont pas statistiquement significatives. Même si elles le sont, elles peuvent être de signe opposé à celui attendu (Neter, Kutner, Wasserman et Nachtsheim 1996). La colinéarité influe parfois aussi sur les prévisions (Smith 1974 ; Belsley 1984).

Dans les plans expérimentaux, il peut être possible de créer des situations où les variables explicatives sont orthogonales les unes par rapport aux autres. Par contre, dans de nombreuses enquêtes, des données sur des variables fortement corrélées sont recueillies pour l'analyse. Ainsi, le revenu total et ses composantes (par exemple, traitements et salaires, gains en capital, intérêts et dividendes) sont recueillis dans le cadre de la Panel Survey of Income Dynamics (http://psidonline.isr.umich.edu/) afin de suivre le bien-être économique au fil du temps. Lorsqu'une variable explicative est une combinaison linéaire des autres, on parle de colinéarité (ou de multicolinéarité) parfaite et celle-ci est facile à déceler. Les cas présentant un intérêt en pratique sont ceux où la colinéarité est imparfaite, mais a néanmoins une incidence sur la précision des estimations (Kmenta 1986, section 10.3).

Alors que la littérature sur les diagnostics de régression est abondante pour les données ne provenant pas d'enquêtes, elle l'est nettement moins pour les données d'enquêtes. Au cours de la dernière décennie, quelques articles ont présenté des techniques d'évaluation de la qualité de la régression appliquée à des données d'enquêtes complexes, principalement pour repérer les points et les groupes influents présentant des valeurs de données ou de poids de sondage anormales. Ainsi, Elliot (2007) a élaboré des méthodes bayésiennes de troncature des poids des estimateurs par la régression linéaire et par la régression linéaire généralisée pour les plans d'échantillonnage avec probabilités d'inclusion inégales. Li (2007a, b), de même que Li et Valliant (2009, 2011) ont adapté et étendu une série de techniques diagnostiques classiques à la régression sur des données d'enquêtes complexes, principalement pour la détection des observations influentes et des groupes d'observations influents. Les travaux de recherche de Li portent sur les résidus et les leviers, DFBETA, DFBETAS, DFFIT, DFFITS, la distance de Cook et l'approche *forward search* (recherche avant). Alors que de nombreuses publications de statistique appliquée offrent des suggestions et des lignes directrices précieuses pour aider les analystes des données à

1. Dan Liao, RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005. Courriel : dliao@rti.org ; Richard Valliant, University of Michigan et University of Maryland, Joint Program in Survey Methodology, 1218 Lefrak Hall, College Park, MD, 20742.

diagnostiquer la présence de colinéarité (par exemple, Farrar et Glauber 1967 ; Theil 1971 ; Belsley, Kuh et Welsch 1980 ; Fox 1984 ; Belsley 1991), aucun de ces travaux de recherche ne traite des diagnostics de colinéarité lorsque des modèles sont ajustés au moyen de données d'enquêtes.

Le facteur d'inflation de la variance (VIF, de l'anglais *variance inflation factors*) décrit à la section 2, qui représente l'une des techniques classiques de diagnostic de la colinéarité les plus répandues, s'applique principalement à des régressions par les moindres carrés ordinaires ou pondérés. Le VIF mesure l'augmentation (« inflation ») de la variance de l'estimation d'une pente causée par la non-orthogonalité des variables explicatives en sus de ce que la variance serait dans des conditions d'orthogonalité. À la section 3, nous examinons le cas d'un analyste qui estime les paramètres d'un modèle en appliquant la méthode des moindres carrés pondérés par les poids de sondage (MCPPS) et calculons les VIF appropriés pour cette méthode. Les composantes du VIF peuvent être estimées en utilisant les éléments d'un estimateur de variance d'usage fréquent dans les progiciels pour l'analyse des données d'enquêtes. Dans le cas de la régression linéaire, un estimateur de variance de type sandwich permettra d'estimer à la fois la variance sous le modèle et la variance sous le plan de sondage de l'estimateur MCPPS de la pente. Comme nous le montrerons à la section 3, la variance sous le modèle ou sous le plan de sondage de $\hat{\beta}_k$, un estimateur de la pente associée à la variable explicative $\mathbf{x}_k$, présente une certaine inflation quand diverses variables explicatives sont corrélées les unes aux autres, comparativement à ce qu'elle serait si $\mathbf{x}_k$ était orthogonale aux autres variables explicatives. La mesure de l'inflation, le VIF, est composée de termes qui doivent être estimés d'après l'échantillon. Notre approche consiste à substituer des estimateurs qui ont une interprétation à la fois sous le modèle et sous le plan, comme il est décrit à la section 3.5.

À la quatrième section, nous présentons une étude empirique portant sur des données de la National Health and Nutrition Examination Survey réalisée aux États-Unis. Nous montrons l'application de notre nouvelle approche et comparons les valeurs nouvellement dérivées du VIF pour l'estimateur MCPPS à celles obtenues pour les estimateurs par les moindres carrés ordinaires et par les moindres carrés pondérés, qui sont produites par les progiciels statistiques standard. Les comparaisons montrent que les valeurs du VIF diffèrent selon la méthode de régression et qu'un VIF spécial pour échantillon complexe devrait être utilisé pour évaluer l'effet nuisible de la colinéarité dans l'analyse des données d'enquêtes.

2. Diagnostics de colinéarité dans l'estimation par les moindres carrés ordinaires

Supposons que l'échantillon s contient n unités sur chacune desquelles sont observées p variables explicatives $\mathbf{x}$ et une variable d'analyse Y. Le modèle linéaire classique, dans un contexte autre qu'une enquête, est $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, où $\mathbf{Y}$ est un vecteur de dimension $n \times 1$ d'observations sur une variable réponse, ou variable dépendante ; $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ est une matrice de plan de dimensions $n \times p$ de constantes fixes avec $\mathbf{x}_k$, le vecteur de dimension $n \times 1$ des valeurs de la variable explicative k pour les n unités échantillonnées ; $\boldsymbol{\beta}$ est un vecteur de dimension $p \times 1$ des paramètres à estimer ; et $\boldsymbol{\epsilon}$ est un vecteur de dimension $n \times 1$ de termes d'erreur statistiquement indépendants de moyenne nulle et de variance constante σ^2. Nous supposons, pour simplifier, que $\mathbf{X}$ est de plein rang-colonne. L'estimation par les moindres carrés ordinaires (MCO) de $\boldsymbol{\beta}$ est $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, pour lequel la variance sous le modèle est $\mathrm{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Ici, nous utilisons l'indice M pour désigner l'espérance sous le modèle.

Les colinéarités des variables explicatives accroissent la variance des coefficients de régression du modèle comparativement à la situation où les $\mathbf{X}$ sont orthogonales. On peut voir cet effet dans la formule de la variance de l'estimation d'un coefficient de régression particulier autre que l'ordonnée à l'origine $\hat{\beta}_k$ (Theil 1971),

$$\mathrm{Var}_M(\hat{\beta}_k) = \frac{\sigma^2}{\sum_{i \in s} x_{ik}^2} \frac{1}{1 - R_k^2} \qquad (1)$$

où R_k^2 est le carré de la corrélation multiple provenant de la régression de la k^{e} colonne de $\mathbf{X}$ sur les autres colonnes. Ce R-carré, défini comme étant $R_k^2 = \hat{\beta}_{(k)}\mathbf{X}_{(k)}^T\mathbf{X}_{(k)}\hat{\beta}_{(k)} / \mathbf{x}_k^T\mathbf{x}_k$, où $\hat{\beta}_{(k)}$ est l'estimation par les MCO de la pente quand on fait la régression de $\mathbf{x}_k$ sur les autres $\mathbf{x}$ et que $\mathbf{X}_{(k)}$ est la matrice $\mathbf{X}$ dont la k^{e} colonne a été supprimée. Le terme $\sigma^2 / \sum x_{ik}^2$ est la variance de $\hat{\beta}_k$ sous le modèle si la k^{e} variable explicative est orthogonale à toutes les autres variables explicatives. La valeur de R_k^2 peut être non nulle parce que la k^{e} variable explicative est corrélée à une autre variable explicative ou à cause d'une relation plus complexe de dépendance entre $\mathbf{x}_k$ et plusieurs autres variables explicatives. Par conséquent, la colinéarité entre $\mathbf{x}_k$ et certaines autres variables explicatives peut entraîner l'inflation de la variance de $\hat{\beta}_k$ au-delà de la valeur qui serait obtenue si les $\mathbf{X}$ étaient orthogonales. Le deuxième terme de (1), $(1 - R_k^2)^{-1}$, est appelé facteur d'inflation de la variance, symbolisé par VIF, pour *variance inflation factor* (Theil 1971).

Une référence fondamentale sur la colinéarité et d'autres diagnostics de la régression par les MCO est le document de Belsley et coll. (1980). Les diagnostics de colinéarité sont abordés dans de nombreux autres traités, y compris Fox (1984) et Neter et coll. (1996). Dans certains cas, il est souhaitable de pondérer les cas différemment dans une analyse de régression afin d'intégrer une variance résiduelle non constante. Cette forme de pondération, qui est fondée

sur un modèle, est la méthode des moindres carrés pondérés (MCP). La plupart des progiciels statistiques courants (par exemple, SAS, Stata, S-Plus et R) utilisent $(1 - R^2_{k(\mathrm{MCP})})^{-1}$ comme VIF pour les MCP, où $R^2_{k(\mathrm{MCP})}$ est le carré de la corrélation multiple provenant de la régression par les MCP de la k^{e} colonne de $\mathbf{X}$ sur les autres colonnes. Fox et Monette (1992) ont également généralisé ce concept d'inflation de la variance en tant que mesure de colinéarité à un sous-ensemble de paramètres dans $\mathbf{b}$ et dérivé un facteur d'inflation de la variance généralisé (GVIF pour *generalized variance-inflation factor*). De surcroît, certains travaux intéressants ont abouti à l'élaboration de mesures de type VIF, telles que les *indices de colinéarité* de Steward (1987) qui sont simplement les racines carrées des VIF et la *tolérance* définie comme étant l'inverse du VIF dans Simon et Lesage (1988).

3. Le VIF dans la régression par les moindres carrés pondérés par les poids de sondage

3.1 Estimateurs par les moindres carrés pondérés par les poids de sondage

Supposons que le modèle structurel sous-jacent à la superpopulation est $\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{e}$, où les termes d'erreur du modèle ont une structure de variance générale $\mathbf{e} \sim (0, \sigma^2\mathbf{V})$ avec $\mathbf{V}$ et σ^2 connues. Soit $\mathbf{W}$ la matrice diagonale des poids de sondage. Nous supposons tout au long de l'exposé que les poids de sondage sont construits de telle façon qu'ils peuvent être utilisés pour estimer les totaux de population finie. L'estimateur par les moindres carrés pondérés par les poids de sondage (MCPPS) est $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}} = (\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{WY}$, en supposant que $\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}$ est inversible. Fuller (2002) décrit les propriétés de cet estimateur.

L'estimateur $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$ est un estimateur sans biais par rapport au modèle de $\boldsymbol{\beta}$ sous le modèle $\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{e}$, que $\mathrm{Var}_M(\mathbf{e}) = \sigma^2\mathbf{V}$ soit spécifiée correctement ou non, et est approximativement sans biais sous le plan pour le paramètre de population (recensement) $\mathbf{B}_U = (\mathbf{X}_U^T\mathbf{X}_U)^{-1}\mathbf{X}_U^T\mathbf{Y}_U$, dans la population finie de N unités. L'indice U désigne la population finie, $\mathbf{Y}_U = (Y_1, ..., Y_N)^T$, et $\mathbf{X}_U = (\mathbf{x}_1, ..., \mathbf{x}_p)$ avec $\mathbf{x}_k$ le vecteur de dimension $N \times 1$ des valeurs de la covariable *k*.

3.2 Variance des estimations des coefficients sous le modèle

La variance sous le modèle de l'estimateur du paramètre $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$, en supposant que $\mathrm{Var}_M(\mathbf{e}) = \sigma^2\mathbf{V}$, peut s'exprimer

$$\begin{aligned}\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}) &= \sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1} \\ &= \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\sigma^2 = \mathbf{G}\sigma^2, \qquad (2)\end{aligned}$$

où $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$, $\tilde{\mathbf{V}} = \mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}$, $\mathbf{A} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$, $\mathbf{B} = \tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}$, et $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$.

Si les colonnes de $\mathbf{X}$ sont orthogonales, $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathrm{diag}(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)$ et $\mathbf{A}^{-1} = \mathrm{diag}(1/\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)$, où $\tilde{\mathbf{x}}_k = \mathbf{w}_k^{1/2}\mathbf{x}_k$. Le ij^{e} élément de $\mathbf{G}$ devient alors $\tilde{\mathbf{x}}_i^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_j/(\tilde{\mathbf{x}}_i^T\tilde{\mathbf{x}}_i)^2$. Donc, quand les $\mathbf{X}$ sont orthogonales, la variance de $\hat{\beta}_{\mathrm{PPS}_k}$ sous le modèle est

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{PPS}_k}) = \sigma^2\,\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{V}}\,\tilde{\mathbf{x}}_k/(\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{x}}_k)^2, \qquad (3)$$

un fait que nous utiliserons plus tard. De manière plus générale, la variance sous le modèle de $\hat{\beta}_{\mathrm{PPS}_k}$, l'estimation du coefficient de la k^{e} variable explicative, est

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{PPS}_k}) = \mathbf{i}_k'\,\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}})\mathbf{i}_k = \sigma^2\,\mathbf{i}_k'\,\mathbf{G}\mathbf{i}_k = \sigma^2 g^{kk} \quad (4)$$

où $\mathbf{i}_k$ est un vecteur de dimension $p \times 1$ avec 1 en position k et des 0 ailleurs, et g^{kk} est le k^{e} élément diagonal de la matrice $\mathbf{G}$.

3.3 VIF fondé sur le modèle

Comme nous le montrons à l'annexe A, la variance sous le modèle de $\hat{\beta}_{\mathrm{PPS}_k}$ dans (4) peut s'écrire :

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{PPS}_k}) = g^{kk}\sigma^2 = \frac{\zeta_k\rho_k}{1 - R^2_{\mathrm{PPS}(k)}}\frac{\sigma^2\,\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{V}}\,\tilde{\mathbf{x}}_k}{(\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{x}}_k)^2}, \quad (5)$$

où

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{V}}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{e}}_{xk}} = \frac{\mathbf{e}_{xk}^T\mathbf{WVWe}_{xk}}{\mathbf{e}_{xk}^T\mathbf{We}_{xk}},$$

avec $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{\mathrm{PPS}(k)}$ représentant le résidu de la régression par les MCPPS de $\mathbf{x}_k$ sur $\mathbf{X}_{(k)}$, et $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\mathrm{PPS}(k)} = \mathbf{W}^{1/2}\mathbf{e}_{xk}$,

$$\rho_k = \frac{\tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k} = \frac{\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{WVW}\mathbf{x}_k}$$

et $R^2_{\mathrm{PPS}(k)}$, défini à l'annexe A, est le carré de la corrélation multiple provenant de la régression pondérée de la k^{e} colonne de $\mathbf{X}$ sur les autres colonnes. Donc, ζ_k et ρ_k dépendent de $\mathbf{W}$ et de $\mathbf{V}$. La variance sous orthogonalité donnée par (3) est augmentée

$$\mathrm{VIF}_k = \frac{\zeta_k\rho_k}{1 - R^2_{\mathrm{PPS}(k)}} \qquad (6)$$

fois lorsque l'on introduit les $p - 1$ autres variables explicatives dans les MCPPS. Dans ces derniers, le VIF fondé sur le modèle inclut non seulement le coefficient de corrélation multiple $R^2_{\mathrm{PPS}(k)}$, mais aussi deux coefficients d'ajustement, ζ_k et ρ_k, qui ne sont pas présents dans le cas des MCO et des MCP.

En procédant à la décomposition de $\tilde{\mathbf{V}}$ en valeurs singulières, nous pouvons borner le facteur $\zeta_k\rho_k$, qui est l'ajustement du VIF dans les MCP. Sur la base des extrema

du ratio des formes quadratiques (Lin 1984), les bornes du terme ζ_k sont celles de l'intervalle $\mu_{\min}(\tilde{\mathbf{V}}) \le \zeta_k \le \mu_{\max}(\tilde{\mathbf{V}})$, et les bornes de ρ_k, celles de l'intervalle

$$\frac{1}{\mu_{\max}(\tilde{\mathbf{V}})} \le \rho_k \le \frac{1}{\mu_{\min}(\tilde{\mathbf{V}})},$$

où $\mu_{\min}(\tilde{\mathbf{V}})$ et $\mu_{\max}(\tilde{\mathbf{V}})$ sont les valeurs singulières minimale et maximale de la matrice $\tilde{\mathbf{V}}$. En combinant ces résultats, le coefficient conjoint $\zeta_k \rho_k$ a pour bornes l'intervalle :

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \le \zeta_k \rho_k \le \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

Notons que, quand $\tilde{\mathbf{V}} = \mathbf{I}$, $\zeta_k = \rho_k = 1$ et (6) se réduit à

$$\frac{1}{1 - R^2_{\text{PPS}(k)}} \frac{\sigma^2}{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k},$$

qui est la variance sous le modèle de l'estimation par les MCP quand $\mathbf{V}$ est diagonale et que $\mathbf{W}$ est correctement spécifiée comme étant $\mathbf{W} = \mathbf{V}^{-1}$. Dans ce cas inhabituel, le VIF calculé les progiciels courants sera approprié pour les MCPPS. Cependant, l'hypothèse que $\mathbf{W} = \mathbf{V}^{-1}$ est rarement raisonnable dans l'estimation d'après des données d'enquête. Si $\tilde{\mathbf{V}} \ne \mathbf{I}$, alors ζ_k et ρ_k ne sont pas égaux à 1 et un calcul spécial du VIF demeure nécessaire. Quand $\mathbf{V} = \mathbf{I}$, soit l'application habituellement considérée par les analystes,

$$\tilde{\mathbf{V}} = \mathbf{W}, \zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T \mathbf{W} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}}, \rho_k = \frac{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T \mathbf{W} \tilde{\mathbf{x}}_k}$$

et

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}},$$

où $w_{\min}$ est la valeur minimale des poids de sondage et $w_{\max}$ est leur valeur maximale. Dans ce cas, l'intervalle de $\zeta_k \rho_k$ est borné par

$$\left[\frac{w_{\min}}{w_{\max}}, \frac{w_{\max}}{w_{\min}} \right].$$

Quand tous les poids de sondage sont constants, $\zeta_k \rho_k = 1$ et les VIF produits par les logiciels standard, $(1 - R^2_{\text{PPS}})^{-1}$, n'ont pas besoin d'être ajustés sous les MCPPS ; cependant, si l'intervalle des poids de sondage est grand, $\zeta_k \rho_k$ peut être très petit ou très grand, et peut être supérieur ou inférieur à 1. Dans ce cas, le VIF produit par les logiciels standard ne convient pas et un calcul spécial est nécessaire. Ces faits seront illustrés dans nos études expérimentales.

Le VIF donné par (6) est approprié, que le modèle comprenne ou non une ordonnée à l'origine. Une autre version peut être écrite en supposant que le modèle contient une ordonnée à l'origine quand on effectue la régression de $\tilde{\mathbf{x}}_k$ sur les autres $\mathbf{x}$. L'établissement de cette forme figure dans Liao (2010). Nous résumons le résultat ci-dessous.

La variance de $\hat{\beta}_{\text{PPS}_k}$ dans un modèle M2 qui contient une ordonnée à l'origine et dans lequel $\mathbf{x}_k$ est orthogonale aux autres $\mathbf{x}$ est :

$$\text{Var}_{M2}(\hat{\beta}_{\text{PPS}_k}) = \frac{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)^T \tilde{\mathbf{V}} (\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)}{\text{SST}^2_{\text{PPS}_{m(k)}}} \qquad (7)$$

où $\tilde{\mathbf{1}} = (w_1^{1/2}, ..., w_n^{1/2})$, $\bar{\tilde{x}}_k = \sum_{i \in s} w_i x_{ki} / \hat{N}$, $\hat{N} = \sum_{i \in s} w_i$ et $\text{SST}_{\text{PPS}_{m(k)}} = \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N} \bar{\tilde{x}}_k^2$. La variance de $\hat{\beta}_{\text{PPS}_k}$ peut alors s'écrire

$$\text{Var}_M(\hat{\beta}_{\text{PPS}_k}) = \frac{\zeta_k \rho_{mk}}{1 - R^2_{\text{PPS}m(k)}} \text{Var}_{M2}(\hat{\beta}_{\text{PPS}_k}) \qquad (8)$$

où $R^2_{\text{PPS}m(k)}$ est le R-carré de la régression par les MCPPS de $\tilde{\mathbf{x}}_k$ sur les $\mathbf{x}$ dans la partie restante de $\tilde{\mathbf{X}}$ (en excluant une colonne pour l'ordonnée à l'origine). Le terme ζ_k a été défini en suivant (5) et

$$\rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N} \bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)^T \tilde{\mathbf{V}} (\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)}.$$

La plupart des progiciels donnent systématiquement $(1 - R^2_{\text{PPS}m(k)})^{-1}$ comme VIF dans la sortie de la régression par les MCP. Il convient de souligner que ce paramètre est différent du VIF égal à $(1 - R^2_{\text{PPS}(k)})^{-1}$ présenté à la section 3.3, qui ne suppose pas qu'une ordonnée à l'origine est gardée dans le modèle. Habituellement, les progiciels ne fournissent pas $(1 - R^2_{\text{PPS}(k)})^{-1}$.

Au moyen d'arguments similaires à ceux de la section précédente, nous pouvons borner $\zeta_k \rho_{mk}$ par

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \le \zeta_k \rho_{mk} \le \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

La variance sous le modèle de $\hat{\beta}_{\text{PPS}_k}$ est augmentée d'une valeur

$$\text{VIF}_{mk} = \frac{\zeta_k \rho_{mk}}{1 - R^2_{\text{PPS}m(k)}}$$

comparativement à sa variance sous le modèle (M2) contenant uniquement la variable explicative $\tilde{\mathbf{x}}_k$ et l'ordonnée à l'origine. Le nouveau VIF_{mk} ajusté pour tenir compte de l'ordonnée à l'origine garde certaines propriétés du VIF_k donné en (6). Quand $\tilde{\mathbf{V}} = \mathbf{I}$, nous avons $\zeta_k = 1$, $\rho_{mk} = 1$ et le VIF corrigé pour l'ordonnée à l'origine dans (8) pour les MCPPS est égal au VIF corrigé pour l'ordonnée à l'origine classique : $(1 - R^2_{m(k)})^{-1}$. Quand $\mathbf{V} = \mathbf{I}$, nous avons $\tilde{\mathbf{V}} = \mathbf{W}$,

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk} \mathbf{W} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}}, \rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N} \bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)^T \mathbf{W} (\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)}$$

et

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}}.$$

L'intervalle de $\zeta_k \rho_{mk}$ dépend aussi de l'intervalle des poids de sondage, comme celui de $\zeta_k \rho_k$.

3.4 Estimation du VIF pour un modèle avec échantillonnage en grappes stratifié quand $\mathbf{V}$ est inconnue

Aux sections qui précèdent, nous avons utilisé des arguments fondés sur le modèle pour calculer les VIF. Ces derniers contiennent des termes, $\tilde{\mathbf{V}}$ en particulier, qui sont inconnus et doivent être estimés. À la présente section, nous construisons des estimateurs des composantes du VIF, de nouveau en nous servant d'arguments fondés sur le modèle. Cependant, un estimateur par linéarisation classique de la variance sous le plan permet aussi d'estimer la variance sous le modèle, comme nous le montrons plus bas, et fournit les composantes nécessaires pour estimer le VIF. Dans la suite de la présente section, nous présentons des estimateurs qui conviennent pour un modèle possédant une structure de covariance en grappes stratifiée.

Supposons que, dans un plan d'échantillonnage à plusieurs degrés stratifié, il existe $h = 1, ..., H$ strates dans la population, $i = 1, ..., N_h$ grappes dans la strate correspondante h et $t = 1, ..., M_{hi}$ unités dans la grappe hi. Nous sélectionnons $i = 1, ..., n_h$ grappes dans la strate h et $t = 1, ..., m_{hi}$ unités dans la grappe hi. Désignons l'ensemble de grappes échantillonnées dans la strate h par s_h et l'échantillon d'unités dans la grappe hi par s_{hi}. Le nombre total d'unités échantillonnées dans la strate h est $m_h = \sum_{i \in s_h} m_{hi}$, et le total dans l'échantillon est $m = \sum_{h=1}^{H} m_h$. Nous supposons que les grappes sont sélectionnées avec remise dans les strates et indépendamment entre les strates. Considérons le modèle qui suit :

$$\begin{aligned} &E_M(Y_{hit}) = \mathbf{x}_{hit}^T \boldsymbol{\beta} \\ &\qquad h = 1, \ldots, H, \quad i = 1, \ldots, N_h, \quad t = 1, \ldots, M_{hi} \\ &\mathrm{Cov}_M(Y_{hit}, Y_{h'i't'}) = 0 \\ &\qquad h \neq h', \text{ où, } h = h' \text{ et } i \neq i'. \end{aligned} \tag{9}$$

Nous supposons que les unités sont corrélées dans chaque grappe, mais il n'est pas nécessaire de spécifier la corrélation particulière des covariances pour la présente analyse. L'estimateur du paramètre de régression est :

$$\hat{\boldsymbol{\beta}}_{\mathrm{PPS}} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{Y}_{hi} \tag{10}$$

où $\mathbf{X}_{hi}$ est la matrice de dimensions $m_{hi} \times p$ des covariables pour les unités échantillonnées dans la grappe hi, $\mathbf{W}_{hi} = \mathrm{diag}(w_t)$, $t \in s_{hi}$ est la matrice diagonale des poids de sondage pour la grappe hi et $\mathbf{Y}_{hi}$ est le vecteur de dimensions $m_{hi} \times 1$ de variables réponse dans la grappe hi. La variance sous le modèle de $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$ est :

$$\begin{aligned} \mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}) &= \mathbf{A}^{-1} \left[\sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{V}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi} \right] \mathbf{A}^{-1} \\ &= \mathbf{A}^{-1} \left[\sum_{h=1}^{H} \mathbf{X}_h^T \mathbf{W}_h \mathbf{V}_h \mathbf{W}_h \mathbf{X}_h \right] \mathbf{A}^{-1}, \end{aligned} \tag{11}$$

où $\mathbf{V}_{hi} = \mathrm{Var}_M(\mathbf{Y}_{hi})$ et $\mathbf{V}_h = \mathrm{Blkdiag}(\mathbf{V}_{hi})$, $i \in s_h$. L'expression (11) est un cas particulier de (2) avec $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T, ..., \mathbf{X}_H^T)$, $\mathbf{X}_h$ est la matrice de dimensions $m_h \times p$ des covariables pour les unités échantillonnées dans la strate h, $\mathbf{W} = \mathrm{diag}(\mathbf{W}_{hi})$, pour $h = 1, ..., H$ et $i \in s_h$, et $\mathbf{V} = \mathrm{Blkdiag}(\mathbf{V}_h)$.

Désignons les résidus au niveau de la grappe comme un vecteur, $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$. Un estimateur par linéarisation sous le plan de sondage est donné par :

$$\begin{aligned} \mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}) &= \mathbf{A}^{-1} \left[\sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i \in s_h} (\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)(\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)^T \right] \mathbf{A}^{-1} \\ &= \mathbf{A}^{-1} \left[\sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left(\sum_{i \in s_h} \mathbf{z}_{hi} \mathbf{z}_{hi}^T - n_h \bar{\mathbf{z}}_h \bar{\mathbf{z}}_h^T \right) \right] \mathbf{A}^{-1}, \end{aligned} \tag{12}$$

où

$$\bar{\mathbf{z}}_h = \frac{1}{n_h} \sum\nolimits_{i \in s_h} \mathbf{z}_{hi}$$

et $\mathbf{z}_{hi} = \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{e}_{hi}$ avec $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$. Cette expression peut être réduite à la formule pour un plan de sondage stratifié à seul degré quand les tailles des échantillons de grappes sont toutes égales à 1, $m_{hi} = 1$. L'expression (12) est utilisée par les progiciels Stata et SUDAAN, entre autres. L'estimateur $\mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}})$ est convergent et approximativement sans biais par rapport au plan sous un plan dans lequel les grappes sont sélectionnées avec remise (Fuller 2002). Li (2007a, b) a montré que (12) est également un estimateur approximativement sans biais par rapport au modèle sous le modèle (11).

Le terme entre crochets dans (12) sert d'estimateur de la matrice $\mathbf{B}$ dans l'expression (2). Les composantes de $\mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}})$ peuvent être utilisées pour construire les estimateurs de ζ_k et ρ_k dans (5) et de ρ_{mk} dans (8). En particulier,

$$\hat{\zeta}_k = \frac{\mathbf{e}_{xk}^T \mathbf{W} \hat{\mathbf{V}} \mathbf{W} \mathbf{e}_{xk}}{\mathbf{e}_{xk}^T \mathbf{W} \mathbf{e}_{xk}}, \tag{13}$$

où

$$\hat{\mathbf{V}} = \mathrm{Blkdiag}\left[\frac{n_h}{n_h - 1} \left(\hat{\mathbf{V}}_h - \frac{1}{n_h} \mathbf{e}_h \mathbf{e}_h^T \right) \right], \; h = 1, 2, ..., H$$

avec $\hat{\mathbf{V}}_h = \mathrm{Blkdiag}(\mathbf{e}_{hi} \mathbf{e}_{hi}^T)$ et

$$\hat{\rho}_k = \frac{\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{W} \hat{\mathbf{V}} \mathbf{W} \mathbf{x}_k},$$

avec $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)} \hat{\boldsymbol{\beta}}_{\mathrm{PPS}(k)}$. L'estimation de $\hat{\rho}_{mk}$, défini selon (8), est

$$\hat{\rho}_{mk} = \frac{(\mathbf{x}_k^T \mathbf{W} \mathbf{x}_k - \hat{N} \bar{\bar{x}}_k^2)}{(\mathbf{x}_k - \mathbf{1}\bar{\bar{x}}_k)^T \mathbf{W} \hat{\mathbf{V}} \mathbf{W} (\mathbf{x}_k - \mathbf{1}\bar{\bar{x}}_k)}. \qquad (14)$$

Étant donné ces estimateurs des composantes, VIF_k est estimé par

$$\widehat{\mathrm{VIF}}_k = \frac{\hat{\zeta}_k \hat{\rho}_k}{1 - R^2_{\mathrm{PPS}(k)}}$$

et VIF_{mk} est estimé par

$$\widehat{\mathrm{VIF}}_{mk} = \frac{\hat{\zeta}_k \hat{\rho}_{mk}}{1 - R^2_{\mathrm{PPS}m(k)}}.$$

4. Étude expérimentale

Nous allons maintenant illustrer les diagnostics de colinéarité modifiés proposés et étudier leur comportement en utilisant des données sur l'apport alimentaire provenant de la National Health and Nutrition Examination Survey (NHANES) réalisée aux États-Unis en 2007-2008 (http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/datadoc_changes_0708.htm). Les données sur l'apport alimentaire sont utilisées pour estimer les types et les quantités d'aliments et de boissons consommés durant la période de 24 heures qui précède l'entrevue (de minuit à minuit) et pour estimer les apports d'énergie, de nutriments et d'autres composants alimentaires provenant de ces aliments et boissons. La NHANES est réalisée selon un plan d'échantillonnage probabiliste à plusieurs degrés complexe. Certains sous-groupes de population sont suréchantillonnés afin d'accroître la fiabilité et la précision des estimations des indicateurs de l'état de santé pour ces groupes. Parmi les participants qui répondent à l'interview sur place au centre d'examen mobile (CEM), environ 94 % fournissent des renseignements complets sur les apports alimentaires. Les poids de sondage pour ces données ont été construits en prenant les poids de sondage ajustés pour le CEM et en les rajustant en outre pour tenir compte de la non-réponse supplémentaire et de la différence de répartition selon le jour de la semaine pour la collecte des données sur les apports alimentaires. Ces derniers poids sont plus variables que les poids produits pour le CEM. Le jeu de données utilisé dans notre étude est un sous-ensemble des données de 2007-2008 composé de femmes de 26 à 40 ans ayant répondu à l'enquête. Les observations comportant des valeurs manquantes pour les variables choisies ont été exclues de l'échantillon qui, en dernière analyse, contient 672 réponses complètes. Les poids finaux dans notre échantillon varient de 6 028 à 330 067, avec un ratio de 55 pour 1. Le National Center for Health Statistics des États-Unis recommande que le plan de sélection de l'échantillon s'approche de la sélection stratifiée avec remise de 32 UPE dans 16 strates, avec 2 UPE dans chaque strate.

Pour notre étude empirique, nous avons ajusté un modèle de régression linéaire du poids corporel (en kg) par la méthode des moindres carrés pondérés par les poids de sondage. Les variables explicatives considérées comprennent l'âge, la race noire et les variables d'apport nutritionnel total quotidien, qui sont les calories (100 kcal), les protéines (100 g), les glucides (100 g), les sucres (100 g), les lipides totaux (100 g), les acides gras saturés totaux (100 g), les acides gras mono-insaturés totaux (100 g), les acides gras poly-insaturés totaux (100 g) et l'alcool (100 g). Toutes les variables d'apport nutritionnel total quotidien sont corrélées les unes aux autres à divers degrés comme l'illustre la figure 1.

Trois méthodes de régression ont été comparées dans l'étude. La première s'appuie sur la méthode des *moindres carrés ordinaires* (MCO) et ne tient pas compte des aspects complexes de l'échantillonnage, y compris la pondération. La deuxième est la méthode des *moindres carrés pondérés* (MCP), dans laquelle sont intégrés les poids de sondage en supposant que $\mathbf{V} = \mathbf{W}^{-1}$, mais qui ne tient pas compte de tous les aspects complexes de l'échantillonnage. La troisième est la méthode des *moindres carrés pondérés par les poids de sondage* (MCPPS), qui s'appuie sur le plan d'échantillonnage complexe réel comme il est décrit à la section 3.4. Les matrices des poids, les estimateurs des variances des coefficients et les diagnostics de colinéarité de ces trois méthodes sont présentés au tableau 1.

Les résultats de l'ajustement du modèle en utilisant les trois méthodes de régression différentes sont présentés au tableau 2. Le modèle contenant toutes les variables explicatives figure dans la partie supérieure du tableau. Dans le tiers inférieur du tableau, un modèle réduit, dans lequel le problème de quasi-dépendance est moindre, est ajusté au moyen de trois variables explicatives seulement : l'âge, la race noire et les calories. Dans le modèle réduit, la valeur du coefficient pour les calories est positive et significative quand les MCP ou les MCPPS sont utilisés, ce qui paraît logique et reflète la relation positive attendue entre le poids corporel d'une répondante et son apport calorique total quotidien. Cependant, lorsque l'on inclut les autres variables d'apport nutritionnel total dans le modèle, la valeur du coefficient des calories est négative et non significative à cause de l'inflation de sa variance. Il s'agit d'un exemple type dans lequel il y a inflation de la variance d'un coefficient et où son signe est illogique à cause de la colinéarité.

Le tableau 3 donne les valeurs du VIF pour les trois méthodes de régression utilisées. Les formules du VIF pour ces méthodes sont présentées au tableau 1. Quand toutes les

variables explicatives sont incluses dans le modèle, la variable des calories obtient la valeur du VIF la plus élevée dans toutes les régressions en raison de sa forte quasi-dépendance avec toutes les autres variables d'apport nutritionnel total. Comme le montre le tableau 1, le VIF pour les MCPPS peut être obtenu en multipliant le VIF pour les MCP par le coefficient d'ajustement $\zeta_k \rho_k$. Au tableau 3, les coefficients d'ajustement $\zeta_k \rho_k$ pour toutes les variables d'apport nutritionnel sauf les lipides sont inférieurs à 1, surtout celui des glucides, qui est de 0,46. Cela indique que les valeurs du VIF pour ces variables dans les MCPPS sont nettement plus faibles que celles dans les MCP, et que la colinéarité des variables explicatives dans le modèle a moins d'effet sur l'estimation des coefficients lorsque l'on utilise les MCPPS que quand on applique les MCP. Par contre, pour les variables d'apport nutritionnel reliées aux lipides, les $\zeta_k \rho_k$ sont tous plus grand que 1. Donc, la colinéarité entre ces variables est plus nuisible à l'estimation des coefficients sous les MCPPS que sous les MCP. Afin d'examiner ce problème de plus près, nous avons également ajusté un modèle ne contenant que deux variables d'apport nutritionnel : les lipides totaux et les acides gras mono-insaturés totaux. Les valeurs du VIF sous les MCPPS sont trois fois plus grandes que sous les MCO ou les MCP pour ces deux variables nutritionnelles. Si un analyste applique la méthode des MCPPS aux données d'enquête examinées ici et qu'il utilise les valeurs non ajustées du VIF fournies par les progiciels statistiques standard pour les MCO (présentées dans la première colonne) ou pour les MCP (présentées dans la deuxième colonne), son appréciation de la gravité de la colinéarité dans le modèle sera quelque peu erronée. Brièvement, même si les pentes estimées et les prédictions de la régression par les MCP et par les MCPPS sont les mêmes, les VIF risquent d'être sous-estimés ou surestimés si l'on ne tient pas compte des aspects complexes de l'enquête.

5. Conclusion

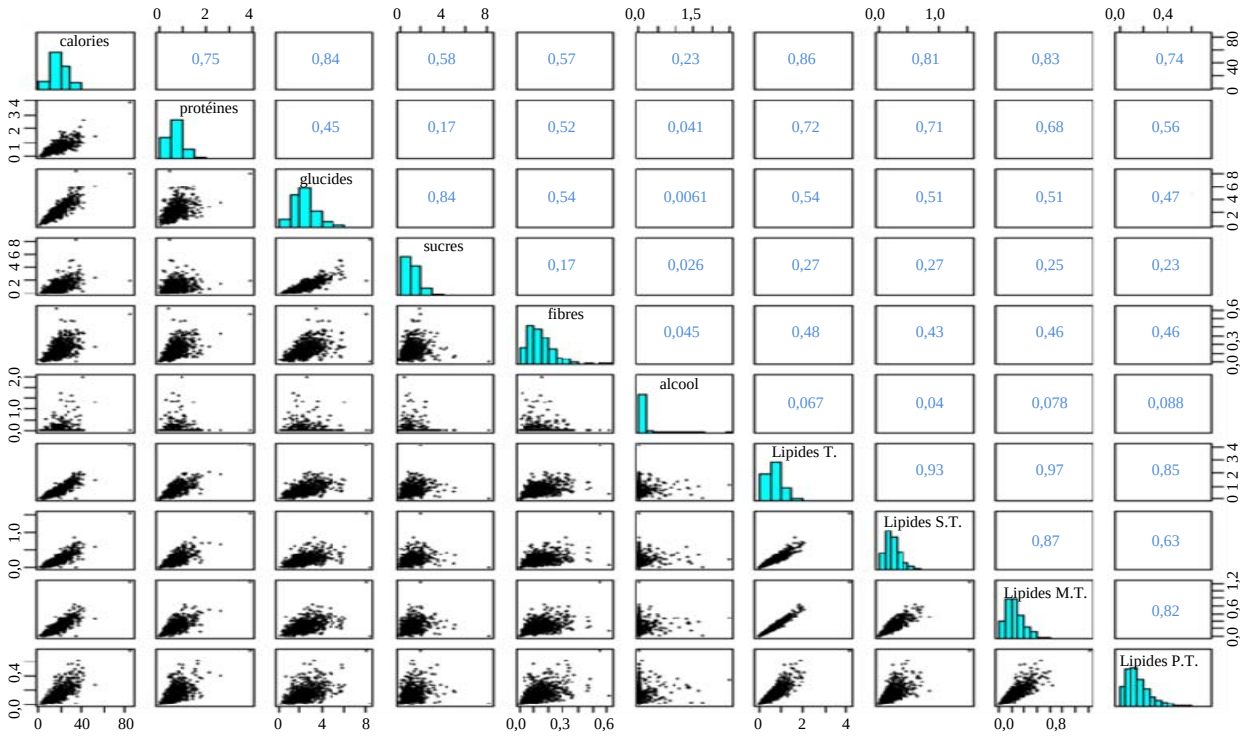
Afin qu'ils soient applicables à des modèles estimés d'après des données d'enquête, les diagnostics de régression doivent être adaptés de manière à tenir compte de la pondération et des caractéristiques du plan d'échantillonnage, telles que la stratification et la mise en grappes. Dans le présent article, nous avons élaboré une nouvelle formule pour calculer un facteur d'inflation de la variance (VIF) approprié pour les modèles linéaires. Un VIF est une mesure de l'augmentation, ou inflation, de la variance de l'estimateur d'un paramètre attribuable au fait que les variables explicatives sont corrélées au lieu d'être orthogonales. Bien que les progiciels standard permettent de produire des estimations de la pente d'une droite de régression pondérée par les poids de sondage en suivant des procédures qui reposent sur les moindres carrés pondérés, les VIF calculés par les routines développées pour des données ne provenant pas d'enquêtes sont incorrects. Le VIF pour un échantillon complexe est égal au VIF issu des moindres carrés pondérés multiplié par un facteur d'ajustement. Ce dernier est positif, mais peut être plus grand ou plus petit que 1, selon la nature des données analysées.

Dans une étude empirique, nous avons illustré l'application de notre nouvelle approche en utilisant des données provenant de la National Health and Nutrition Examination Survey de 2007-2008. Nous avons donné un exemple simple de la façon dont la colinéarité entre les variables explicatives influence l'estimation des coefficients dans la régression linéaire et démontré que, même si les coefficients estimés (et les valeurs prédites) sont les mêmes quand on utilise les moindres carrés pondérés ou les moindres carrés pondérés par les poids de sondage, leurs variances estimées et leurs valeurs du VIF (qui reflètent l'effet de la colinéarité sur l'estimation du coefficient) peuvent être différentes.

Tableau 1
Méthodes de régression et leurs statistiques de diagnostic de la colinéarité utilisées dans l'étude expérimentale

Type de régression	Matrice de pondération $\mathbf{W}^a$	Estimation de la variance de $\hat{\boldsymbol{\beta}}$	Formule du VIF
MCO	$\mathbf{I}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{1}{1-R^2_{m(k)}}$ [b]
MCP	$\mathbf{W}^c$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{1}{1-R^2_{\text{PPS}m(k)}}$
MCPPS	$\mathbf{W}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{\hat{\zeta}_k\hat{\rho}_{mk}}{1-R^2_{\text{PPS}m(k)}}$
		avec	avec $\hat{\zeta}_k = \dfrac{\mathbf{e}_{xk}^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{e}_{xk}}$,
		$\hat{\mathbf{V}} = \sum_{h=1}^{H}\dfrac{n_h}{n_h-1}\left[\text{Blkdiag}(\mathbf{e}_{hi}\,\mathbf{e}_{hi}^T) - \dfrac{1}{n_h}\mathbf{e}_h\,\mathbf{e}_h^T\right]$	$\hat{\rho}_{mk} = \dfrac{(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\hat{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}$

[a] Dans tous les modèles de régression, les paramètres sont estimés par : $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$.
[b] $R^2_{m(k)}$ est le R-carré de la régression par les MCO de $\mathbf{x}_k$ sur les $\mathbf{x}$ dans la partie restante de $\mathbf{X}$ (en excluant une colonne pour l'ordonnée à l'origine).
[c] $\mathbf{W}$ est la matrice diagonale qui contient les poids de sondage w_i sur la diagonale principale.



[a] Lipides T. : lipides totaux
Lipides S.T. : acides gras saturés totaux
Lipides M.T. : acides gras mono-insaturés totaux
Lipides P.T. : acides gras poly-insaturés totaux

Figure 1 Diagrammes de dispersion par paire et coefficients de corrélation des variables nutritionnelles[a]

Tableau 2
Estimations des paramètres et leurs erreurs-types connexes pour trois méthodes de régression différentes

	Modèle complet					
	MCO		MCP		MCPPS	
Variable	**Bêta**	**E.-T.**	**Bêta**	**E.-T.**	**Bêta**	**E.-T.**
Ordonnée à l'origine	63,90***[a]	6,95	67,47***	6,36	67,47***	8,76
Âge	0,26	0,19	0,08	0,18	0,08	0,25
Race noire	10,39***	2,07	10,59***	2,38	10,59***	2,20
Calories	-6,41	5,76	-8,19	5,56	-8,19	5,75
Protéines	25,72	24,76	40,98	23,60	40,98	25,38
Glucides	26,67	23,93	32,31	22,96	32,31	22,65
Sucres	-1,90	3,06	-0,30	2,82	-0,30	4,06
Fibres	-41,17	20,23	-34,20	17,98	-34,20	19,05
Alcool	38,84	39,45	49,37	38,28	49,37	40,10
Lipides totaux	150,25*	69,53	161,78*	72,12	161,78	94,76
Acides gras saturés totaux	-113,20*	49,81	-101,40	56,26	-101,40	82,71
Acides gras mono-insaturés totaux	-72,05	48,03	-92,44	51,52	-92,44	83,51
Acides gras poly-insaturés totaux	-92,60*	46,13	-75,55	51,16	-75,55	78,76
	Modèle réduit					
	MCO		MCP		MCPPS	
Variable	**Bêta**	**E.-T.**	**Bêta**	**E.-T.**	**Bêta**	**E.-T.**
Ordonnée à l'origine	62,26***	6,88	67,52***	6,29	67,52***	8,48
Âge	0,27	0,19	0,07	0,18	0,07	0,25
Race noire	12,54***	1,98	11,74***	2,32	11,74***	2,05
Calories	0,15	0,10	0,23*	0,09	0,23*	0,10

[a] valeurs p de signification : * $p = 0{,}05$; ** $p = 0{,}01$; *** $p = 0{,}005$.

Tableau 3
Valeurs du VIF pour trois méthodes de régression différentes

			Modèle complet	
	MCO	MCP	MCPPS	
Variable	VIF	VIF	VIF	$\zeta_k \rho_k$
Âge	1,02	1,03	0,96	0,94
Race noire	1,10	1,07	1,12	1,05
Calories	3 411,61	3 562,70	2 740,83	0,77
Protéines	123,12	127,35	103,50	0,81
Glucides	1 074,87	1 007,40	462,08	0,46
Sucres	8,37	7,03	4,87	0,69
Fibres	4,59	3,94	2,37	0,60
Alcool	120,56	115,67	89,92	0,78
Lipides totaux	1 190,24	1 475,27	2 513,69	1,70
Acides gras saturés totaux	76,80	112,61	202,91	1,80
Acides gras mono-insaturés totaux	82,37	107,34	286,24	2,67
Acides gras poly-insaturés totaux	34,73	49,45	118,21	2,39
			Modèle réduit	
	MCO	MCP	MCPPS	
Variable	VIF	VIF	VIF	$\zeta_k \rho_k$
Âge	1,00	1,00	0,98	0,98
Race noire	1,02	1,01	0,97	0,96
Lipides totaux	20,10	20,22	63,15	3,12
Acides gras mono-insaturés totaux	20,16	20,26	61,57	3,04
			Modèle réduit	
	MCO	MCP	MCPPS	
Variable	VIF	VIF	VIF	$\zeta_k \rho_k$
Âge	1,00	1,00	0,98	0,97
Race noire	1,00	1,03	1,00	1,00
Calories	1,00	1,01	0,96	0,95

Les objectifs de l'analyse doivent être pris en considération pour décider de la façon d'utiliser les VIF. Si l'objectif principal est la prédiction, le fait d'inclure des variables colinéaires ou de choisir des variables incorrectes n'est pas très préoccupant. Si des conclusions plus fondamentales sont souhaitées, l'analyste devrait examiner quelles variables il serait logique d'inclure comme variables explicatives au lieu de se fier à un algorithme automatique pour sélectionner les variables. Les VIF sont un outil utile pour repérer les variables explicatives dont les coefficients estimés ont une variance inutilement grande. Quoiqu'ils puissent être considérés comme un outil pour la sélection automatique des variables, les simulations présentées dans Liao (2010), non reproduites ici, montrent que leur utilisation n'est pas un moyen fiable d'identifier un vrai modèle sous-jacent.

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs dont les commentaires ont apporté des améliorations importantes à l'article.

Annexe A

Détermination de g^{kk}

De façon analogue à la détermination du VIF pour les MCO classiques dans Theil (1971), la somme des carrés des écarts et la matrice des produits croisés $\mathbf{A} = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$, qui peut être partitionné comme

$$\mathbf{A}_{p\times p} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T \tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}_{(k)}^T \tilde{\mathbf{X}}_{(k)} \end{pmatrix} \qquad (15)$$

où les colonnes de $\tilde{\mathbf{X}}$ sont réordonnées de manière que $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_k \; \tilde{\mathbf{X}}_{(k)})$ avec $\tilde{\mathbf{X}}_{(k)}$ désignant la matrice de dimensions $n \times (p-1)$ contenant toutes les colonnes sauf la k^{e} colonne de $\tilde{\mathbf{X}}$.

En utilisant la formule de l'inverse d'une matrice partitionnée, l'élément supérieur gauche de $\mathbf{A}^{-1}$ peut être exprimé par :

$$a^{kk} = \mathbf{i}_k^T \mathbf{A}^{-1} \mathbf{i}_k = \mathbf{i}_k^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{i}_k = \frac{1}{(1 - R^2_{\text{PPS}_{(k)}})\, \text{SST}_{\text{PPS}_{(k)}}} = \frac{1}{(1 - R^2_{\text{PPS}_{(k)}})\, \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \quad (16)$$

où

$$R^2_{\text{PPS}_{(k)}} = \frac{\hat{\beta}^T_{\text{PPS}_{(k)}} \tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{X}}_{(k)} \hat{\beta}_{\text{PPS}_{(k)}}}{\text{SST}_{\text{PPS}_{(k)}}},$$

avec $\hat{\beta}_{\text{PPS}_{(k)}} = (\tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{X}}_{(k)})^{-1} \tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{x}}_k$ est le coefficient de détermination correspondant à la régression de $\mathbf{x}_k$ sur les $p - 1$ autres variables explicatives. Le terme $\text{SST}_{\text{PPS}_{(k)}} = \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k$, est la somme totale des carrés des écarts dans cette régression.

Le terme $(1 - R^2_{\text{PPS}_{(k)}})^{-1}$ dans (16) est le VIF qui sera produit par les progiciels standard quand est exécutée une régression par les moindres carrés pondérés. Sous le modèle $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ avec $\boldsymbol{\epsilon} \sim (0, \sigma^2 \mathbf{W}^{-1})$, l'expression (16) est égale à $\text{Var}_M(\hat{\beta}_{\text{PPS}_k})/\sigma^2$. Cependant, elle n'est pas appropriée pour les régressions par les moindres carrés pondérés par les poids de sondage, parce que la variance de $\hat{\boldsymbol{\beta}}_{\text{PPS}}$ possède la forme plus complexe donnée par (2).

La matrice $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ peut être exprimée sous la forme :

$$\mathbf{G} = \begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix} \begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix} \begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix} \quad (17)$$

où la matrice inverse $\mathbf{A}^{-1} = [a^{hk}]$, $h, k = 1, ..., p$, $\mathbf{a}^{k(k)}$ est définie comme étant la k^{e} ligne de $\mathbf{A}^{-1}$ en excluant a^{kk}, $(a^{k1}, ..., a^{k(k-1)}, a^{k(k+1)}, ..., a^{kp})$, $\mathbf{a}^{(k)k} = [\mathbf{a}^{k(k)}]^T$ et $\mathbf{A}^{(k)(k)}$ est définie comme étant la partie $(k-1) \times (k-1)$ de la matrice $\mathbf{A}^{-1}$ en excluant les k^{e} ligne et colonne. La version partitionnée de $\mathbf{B}$ est

$$\mathbf{B} = \begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \end{pmatrix}. \quad (18)$$

En vertu de la symétrie de $\mathbf{A}$ et de $\mathbf{B}$, le k^{e} élément diagonal de $\mathbf{G}$ est

$$g^{kk} = a^{kk}(a^{kk} b_{kk} + 2\mathbf{b}_{k(k)} \mathbf{a}^{(k)k}) + \mathbf{a}^{(k)kT} \mathbf{B}_{(k)(k)} \mathbf{a}^{(k)k}. \quad (19)$$

En utilisant l'inverse partitionnée de la matrice $\mathbf{A}$, qui représente $(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}$, on peut montrer que

$$\mathbf{a}^{(k)k} = -a^{kk} (\tilde{\mathbf{X}}^T_{(k)} \tilde{\mathbf{X}}_{(k)})^{-1} \tilde{\mathbf{X}}_{(k)} \tilde{\mathbf{x}}_k = -a^{kk} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)}. \quad (20)$$

En substituant $a^{(k)k}$ dans (19), g^{kk} peut être exprimé de manière compacte en fonction de a^{kk}, $\hat{\boldsymbol{\beta}}_{\text{PPS}(k)}$ et la composante inférieure droite de la matrice $\mathbf{B}$:

$$\begin{aligned} g^{kk} &= (a^{kk})^2 \left(b_{kk} - 2\mathbf{b}_{k(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} + \hat{\boldsymbol{\beta}}^T_{\text{PPS}(k)} \mathbf{B}_{(k)(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right) \\ &= a^{kk} \times \frac{1}{1 - R^2_{\text{PPS}(k)}} \frac{1}{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \\ &\quad \times \left(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{x}}_k - 2\tilde{\mathbf{x}}_k^T \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} + \hat{\boldsymbol{\beta}}^T_{\text{PPS}(k)} \tilde{\mathbf{X}}_{(k)} \tilde{\mathbf{V}} \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right) \\ &= a^{kk} \times \frac{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)^T \tilde{\mathbf{V}} \left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)}{\left(1 - R^2_{\text{PPS}(k)} \right) \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \\ &= a^{kk} \times \frac{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)^T \tilde{\mathbf{V}} \left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)}{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)^T \left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)} \right)} \\ &= \frac{1}{1 - R^2_{\text{PPS}(k)}} \frac{1}{\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k} \frac{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{V}} \tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T \tilde{\mathbf{e}}_{xk}}, \end{aligned} \quad (21)$$

où $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)} \hat{\boldsymbol{\beta}}_{\text{PPS}(k)}$ est le résidu de la régression de $\tilde{\mathbf{x}}_k$ sur $\tilde{\mathbf{X}}_{(k)}$.

Bibliographie

Belsley, D.A. (1984). Collinearity and forecasting. *Journal of Forecasting*, 38, 73-93.

Belsley, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York : John Wiley & Sons, Inc.

Belsley, D.A., Kuh, E. et Welsch, R.E. (1980). *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York : Wiley Interscience.

Elliot, M. (2007). Réduction bayésienne des poids pour les modèles de régression linéaire généralisée. *Techniques d'enquête*, 33, 27-40.

Farrar, D.E., et Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.

Fox, J. (1984). *Linear Statistical Models and Related Methods, With Applications to Social Research*. New York : John Wiley & Sons, Inc.

Fox, J., et Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417), 178-183.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28(1), 5-25.

Kmenta, J. (1986). *Elements of Econometrics*. New York : Macmillan, 2e Éd.

Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3341-3348.

Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland. Available at http://drum.lib.umd.edu/bitstream/1903/7598/1/umi-umd-4863.pdf.

Li, J., et Valliant, R. (2009). Matrice chapeau et effets de levier pondérés par les poids de sondage. *Techniques d'enquête*, 35(1), 17-27.

Li, J., et Valliant, R. (2011). Linear regression influence diagnostics for unclustered survey data. *Journal of Official Statistics*, 20, 99-119.

Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Thèse de doctorat non-publiée, University of Maryland. Disponible au http://drum.lib.umd.edu/bitstream/1903/10881/1/Liao_umd_0117E_11537.pdf.

Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.

Neter, J., Kutner, M., Wasserman, W. et Nachtsheim, C. (1996). *Applied Linear Statistical Models*. New York : McGraw-Hill/Irwin, 4e Éd.

Simon, S.D., et Lesage, J.P. (1988). The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management*, 1, 137-152.

Smith, G. (1974). Multicollinearity and forecasting. Yale Univeristy Cowles Foundation Disscussion Papier No.383. Disponible au http://cowles.econ.yale.edu/P/cd/d03b/d0383.pdf.

Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.

Theil, H. (1971). *Principles of Econometrics*. New York : John Wiley & Sons, Inc.

ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 69-79
Statistique Canada, N° 12-001-X au catalogue

Estimation de coefficients d'accord d'après des données d'enquête

Hung-Mo Lin, Hae-Young Kim, John M. Williamson et Virginia M. Lesser [1]

Résumé

Nous présentons une approche fondée sur des équations d'estimation généralisées pour estimer le coefficient de corrélation de concordance et le coefficient kappa d'après des données d'enquête. Les estimations ainsi que leurs erreurs-types doivent tenir compte correctement du plan d'échantillonnage. Nous présentons des mesures pondérées du coefficient de corrélation de concordance et du coefficient kappa, ainsi que la variance de ces mesures tenant compte du plan d'échantillonnage. Nous utilisons la méthode de linéarisation par série de Taylor et la procédure du jackknife pour estimer les erreurs-types des estimations résultantes des paramètres. Des mesures anthropométriques et des données sur la santé buccodentaire provenant de la Third National Health and Nutrition Examination Survey sont utilisées pour illustrer cette méthodologie.

Mots clés : Mise en grappes ; coefficient de corrélation de concordance ; équations d'estimation généralisées ; estimateur jackknife ; coefficient kappa ; pondération de l'échantillon ; stratification ; linéarisation par série de Taylor.

1. Introduction

Les enquêtes comprennent souvent la collecte de mesures multiples d'états latents, tels que la qualité de vie ou le désir de poursuivre des études collégiales, ainsi que de mesures multiples d'états difficiles à classer, tels que le syndrome de fatigue chronique. Quand plusieurs mesures sont recueillies, il est naturel de vouloir évaluer leur accord et obtenir des intervalles de confiance pour ces mesures de l'accord. Il pourrait aussi être intéressant de comparer le degré d'accord chez divers sous-groupes de la population ou pour diverses paires de mesures. Dans ce contexte, on pourrait souhaiter tester l'égalité des mesures d'accord. Le présent article porte sur deux mesures de l'accord entre des mesures multiples de ce genre, le coefficient de corrélation de concordance (CCC, ρ_c) et le coefficient kappa (κ). Le premier est utile pour des mesures continues sur des échelles naturelles. Si la mesure d'un concept latent ne possède pas d'échelle naturelle, elle peut être recalculée arbitrairement de manière que sa moyenne soit nulle et sa variance égale, à l'unité. Quand cela est possible, il est dénué de sens de parler d'écarts entre les moments marginaux. Cependant, s'il existe une échelle naturelle, le recalcul n'est pas souhaitable et une bonne mesure de l'accord tiendra compte à la fois de la corrélation et de l'accord des moments marginaux. Le coefficient kappa est des plus utiles pour les classifications binaires.

Le CCC, comme on l'a montré, convient mieux que le coefficient de corrélation de Pearson (ρ) pour mesurer l'accord ou la reproductibilité (Lin 1989 ; Lin 1992). Il évalue l'exactitude entre deux lectures en mesurant la variation de la relation linéaire ajustée par rapport à la droite à 45° passant par l'origine (la droite de concordance), ainsi que la précision en mesurant de quelle distance chaque observation s'écarte de la droite ajustée. Soit Y_{i1} et Y_{i2} une paire de variables aléatoires continues mesurées sur le même sujet i en utilisant deux méthodes. Le CCC pour mesurer l'accord de Y_{i1} et Y_{i2} est défini comme suit :

$$\rho_c = 1 - \frac{E[(Y_{i1} - Y_{i2})^2]}{E_{\text{indep}}[(Y_{i1} - Y_{i2})^2]} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \quad (1)$$

où $\sigma_1^2 = \text{var}(Y_{i1})$, $\sigma_2^2 = \text{var}(Y_{i2})$ et $\sigma_{12} = \text{cov}(Y_{i1}, Y_{i2})$ (Lin 1989). Comme l'a souligné Lin (1989), $\rho_c = 0$ si et uniquement si $\rho = 0$. On peut également montrer algébriquement que ρ_c est proportionnel à ρ et que $-1 \leq -|\rho| \leq \rho_c \leq |\rho| \leq 1$ (Lin 1989). D'où l'imprécision peut être reflétée par une valeur plus faible de ρ et un biais systématique peut être reflété par un ratio de ρ_c à ρ plus faible. Ensemble, les renseignements sur ρ et ρ_c constituent un jeu d'outils pour déterminer quelle correction, en vue d'améliorer l'exactitude et/ou la précision, est la plus avantageuse (Lin et Chinchilli 1997).

Le coefficient de corrélation intraclasse (CCI) est aussi une mesure de l'accord très répandue pour les variables mesurées sur une échelle continue (Fleiss 1986). Supposons que Y_{i1} et Y_{i2} peuvent être décrites dans un modèle linéaire comme suit : $y_{ij} = \mu_j + \theta_i + e_{ij}$ où μ_j est la moyenne de la mesure fournie par la j^{e} méthode, $\theta_i \sim (0, \sigma_\theta^2)$ est la variable latente pour le i^{e} sujet, et les $e_{ij} \sim (0, \sigma_e^2)$ sont des termes d'erreur indépendants. Carrasco et Jover (2003, page 850) ont utilisé un modèle comprenant des composantes de

1. Hung-Mo Lin, Department of Anesthesiology, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1010, New York, NY 10029, États-Unis. Courriel : hung-mo.lin@mountsinai.org; Hae-Young Kim, Center for Statistical Analysis and Research, New England Research Institutes, 9 Galen Street, Watertown, MA 02472, États-Unis; John M. Williamson, Center for Global Health Research, Centers for Disease Control and Prevention/Kenya Medical Research Institute, 1578 Kisumu-Busia Road, Kisian Kisumu, Kenya; Virginia M. Lesser, Department of Statistics and Survey Research Center, Oregon State University, 44 Kidder Hall, Corvallis, OR 97731-4606, États-Unis.

variance pour démontrer que le CCC coïncide avec le coefficient de corrélation intraclasse (CCI) si l'on tient compte de l'écart entre les moyennes des méthodes :

$$\rho_{\text{ICC}} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_e^2 + \sigma_\mu^2} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}.$$

Par conséquent, on peut estimer le CCC en utilisant les composantes de variance d'un modèle à effets mixtes ou la méthode habituelle des moments. En raison de sa supériorité par rapport au coefficient de corrélation de Pearson et de son lien avec le CCI, l'application du CCC a gagné en popularité ces dernières années (Chinchilli, Martel, Kumanyika et Lloyd 1996 ; Zar 1996). En 2009 et en 2010, le CCC a été utilisé comme mesure d'accord dans plus de 60 publications médicales dans des domaines tels que la maladie respiratoire (Dixon, Sugar, Zinreich, Slavin, Corren, Naclerio, Ishii, Cohen, Brown, Wise et Irvin 2009 ; Kocks, Kerstjens, Snijders, de Vos, Biermann, van Hengel, Strijbos, Bosveld et van der Molen 2010), le sommeil (Khawaja, Olson, van der Walt, Bukartyk, Somers, Dierkhising et Morgenthaler 2010), la pédiatrie (Liottol, Radaelli, Orsi1, Taricco, Roggerol, Giann, Consonni, Mosca1 et Cetin 2010), la neurologie (MacDougall, Weber, McGarvie, Halmagyi et Curthoys 2009) et la radiologie (Mazaheri, Hricak, Fine, Akin, Shukla-Dave, Ishill, Moskowitz, Grater, Reuter, Zakian, Touijer et Koutcher 2009).

Le coefficient kappa (κ) (Cohen 1960) et le coefficient kappa pondéré (Cohen 1968) sont les indices les plus utilisés pour mesurer l'accord entre des résultats de type discret ou ordinal, respectivement (Fleiss 1981). Soient Y_{i1} et Y_{i2} deux variables aléatoires binaires prenant les valeurs 0 et 1 avec les probabilités $\pi_1 = \Pr(Y_{i1} = 1)$ et $\pi_2 = \Pr(Y_{i2} = 1)$. Kappa corrige le pourcentage d'accord entre les évaluateurs en tenant compte de la proportion de l'accord due au hasard (calculée sous hypothèse d'indépendance) et est défini comme suit :

$$\kappa = \frac{P_o - P_e}{1{,}0 - P_e}, \qquad (2)$$

où P_e est la probabilité que les réponses binaires formant la paire soient égales sous hypothèse d'indépendance $(\pi_1\pi_2 + (1-\pi_1)(1-\pi_2))$ et P_o est la probabilité que les réponses binaires formant la paire soient égales (Cohen 1960). La différence $P_o - P_e$ représente l'excès d'accord par rapport à l'accord dû au hasard. Une valeur de 0 pour κ indique qu'il n'y a pas d'accord au-delà de celui dû au hasard et une valeur de 1 indique un accord parfait (Fleiss 1981). Les inconvénients du kappa tiennent au fait qu'il s'agit d'une fonction de la distribution marginale des évaluateurs (Fleiss, Nee et Landis 1979 ; Tanner et Young 1985) et que son étendue dépend du nombre d'évaluations par sujet (Fleiss et coll. 1979). Robieson (1999) a remarqué que le CCC calculé d'après des données mesurées sur une échelle ordinale est équivalent au kappa pondéré lorsque l'on utilise des scores entiers. Le kappa a été employé pour mesurer la validité et la reproductibilité de la similarité entre jumeaux (Klar, Lipsitz et Ibrahim 2000), de divers outils épidémiologiques (Maclure et Willett 1987) et de l'accord témoin-informant dans les études cas-témoins (Korten, Jorm, Henderson, McCusker et Creasey 1992).

La valeur des enquêtes par sondage est généralement reconnue et les publications traitant de l'estimation d'après des données d'enquête est abondante (Hansen, Hurwitz et Madow 1953 ; Cochran 1963 ; Kish 1965). Ainsi, un certain nombre d'études fédérales réalisées aux États-Unis en vue d'obtenir des estimations de la santé de la population sont fondées sur des enquêtes nationales, telles que la National Health Interview Survey (NHIS), le Behavioral Risk Factor Surveillance System (BRFSS), et la National Health and Nutrition Examination Surveys (NHANES). Chacune de ces études s'appuie sur un plan de sondage complexe, comportant le suréchantillonnage des sous-populations, la stratification et la mise en grappes. Ces plans sont souvent utilisés pour améliorer la précision, fournir des estimations pour les sous-populations ou réduire les coûts associés à l'établissement des bases de sondage. Afin de faire une inférence fondée sur le plan de sondage concernant la population cible dans le cas de plans de sondage complexes, les estimateurs et leurs variances comprennent les poids de sondage et tiennent compte de la structure du plan en vue d'obtenir des estimations sans biais. En outre, grâce à l'intégration des poids de sondage et de la structure du plan de sondage dans les analyses, toute corrélation possible provenant des grappes dans un plan à plusieurs degrés est prise en compte, de sorte que les erreurs-types des estimateurs ne sont pas sous-estimées.

En général, les chercheurs n'essayent pas de vérifier si leur estimation de l'accord entre les mesures obtenue au moyen du CCC ou du coefficient kappa diffère de manière significative de zéro. Ils cherchent à publier des intervalles de confiance avec leurs estimations (par exemple Dixon et coll. 2009 ; Mazaheri et coll. 2009). Comme dans le cas du coefficient de corrélation de Pearson, il n'existe aucune valeur cible permettant de juger de la force de l'accord. Par conséquent, il est essentiel de porter le jugement quant à l'accord entre toute méthode d'essai et toute méthode de référence avec un degré de certitude établi. Dans certaines situations, les études qui sont réalisées requièrent la vérification d'hypothèses ou la comparaison d'indices d'accord pour plus d'une nouvelle méthode par rapport à une méthode de référence. Ainsi, Khawaja et coll. (2010) ont vérifié l'égalité de deux CCC calculés pour comparer

l'indice d'apnées-hypopnées (IAH) pour les deux et trois premières heures de sommeil en considérant l'IAH provenant de la polysomnographie (PSG) de nuit complète comme la norme de référence. Dans le domaine de la recherche en radiologie, les associations entre les mesures du volume d'une tumeur de la prostate d'après les résultats d'imagerie et d'après l'examen pathologique ont été évaluées en comparant les CCC. On a testé l'égalité de l'accord de chacune des deux méthodes d'imagerie avec les résultats de l'examen pathologique (Mazaheri et coll. 2009). Des tests de l'égalité du coefficient kappa ont été utilisés pour comparer l'évaluation visuelle et la planimétrie informatisée dans l'évaluation de l'ectopie cervicale (Gilmour, Ellerbrock, Koulos, Chiasson, Williamson, Kuhn et Wright 1997 ; Williamson, Manatunga et Lipsitz 2000), et pour comparer les taux de cholestérol chez les jumeaux monozygotes et dizygotes (Feinleib, Garrison, Fabsitz, Christian, Hrubec, Borhani, Kannel, Roseman, Schwartz et Wagner 1977).

Comme l'illustrent les deux exemples tirés de la NHANES III présentés à la section 3, d'importants écarts peuvent exister entre les estimations pondérées et non pondérées des erreurs-types des estimations des paramètres dans les études par sondage. Omettre d'inclure les poids de sondage et de tenir compte de la structure du plan de sondage dans les analyses donne lieu à une sous-estimation des erreurs-types et à une inférence incorrecte. Cet aspect est particulièrement important dans le cas d'enquêtes répétées tous les deux ou trois ans, et les chercheurs s'intéressent souvent tout spécialement à la comparaison des variations entre domaines ou sous-populations. Ainsi, dans la première application de la NHANES III, nous comparons l'accord entre le poids corporel autodéclaré et le poids corporel mesuré durant l'examen physique chez les adolescents. Il est nécessaire de calculer des erreurs-types (intervalles de confiance) exactes si l'on veut comparer le CCC pour divers domaines, tels que les sous-groupes de poids normal et d'obésité.

Nous donnons des mesures pondérées du CCC et du coefficient kappa, ainsi que les estimateurs de variance de ces mesures en tenant compte du plan de sondage. À la section 2, nous présentons une approche fondée sur des équations d'estimation généralisées (GEE pour *generalized estimating equations*) pour estimer ces deux coefficients d'accord d'après des données d'enquête. À la section 3, nous illustrons notre méthode au moyen de données recueillies dans le cadre de la NHANES III. Nous utilisons des mesures corporelles pour estimer ρ_c en vue d'évaluer l'accord entre le poids autodéclaré et le poids réel. Nous nous servons aussi de données sur la santé buccodentaire pour estimer κ en vue d'évaluer l'accord entre deux définitions de la maladie périodontique. Nous tenons compte de la stratification et de la mise en grappes, et intégrons les poids de sondage dans les deux exemples. Nous concluons par une brève discussion.

2. Méthodes

Nous proposons une approche générale d'estimation du CCC et du coefficient kappa d'après des données d'enquête en suivant deux méthodes GEE. Pour le CCC, trois ensembles d'équations d'estimation sont nécessaires. Un premier ensemble modélise la distribution des réponses continues. Comme l'ont fait Barnhart et Williamson (2001), nous utilisons un deuxième ensemble d'équations d'estimation pour estimer les variances des réponses continues. Enfin, le troisième ensemble sert à estimer le CCC en modélisant la covariance entre les réponses continues appariées et les estimations des moyennes et des variances d'après les deux premiers ensembles d'équations d'estimation. Pour κ, deux ensembles seulement d'équations d'estimation sont nécessaires. Un premier ensemble modélise la distribution marginale des réponses binaires. Comme dans Lipsitz, Laird et Brennan (1994), un deuxième ensemble d'équations d'estimation est introduit pour estimer κ en modélisant une variable aléatoire binaire qui dépeint l'accord entre deux réponses sur un sujet.

Afin de tenir compte des probabilités de sélection variables, des matrices de pondération sont incorporées dans chaque ensemble d'équations d'estimation. L'estimation de l'erreur-type des estimations $\hat{\rho}_c$ et $\hat{\kappa}$ proposées d'après les données d'enquête est effectuée par la méthode de linéarisation de Taylor. Nous montrons aussi comment l'estimation des erreurs-types des estimateurs proposés peut se faire par l'approche du jackknife.

Supposons qu'une enquête est réalisée selon un plan de sondage avec stratification, mise en grappes et probabilités de sélection inégales. Soit Y_{hij} la variable réponse pour le j^{e} membre ($j = 1, ..., m_{hi}$) de la i^{e} grappe ($i = 1, ..., n_h$) de la h^{e} strate ($h = 1, ..., H$). En calculant la moyenne sur tous les échantillons possibles, la valeur espérée correspondante est $E[Y_{hij}] = \mu_{hij}$ si Y_{hij} est une réponse continue, et la probabilité correspondante est $E[Y_{hij}] = \Pr[Y_{hij} = 1] = \pi_{hij}$ si Y_{hij} est une réponse binaire. Le poids de sondage w_{hij} est l'inverse de la probabilité de sélection du j^{e} membre de la i^{e} grappe de la h^{e} strate.

2.1 Le coefficient de corrélation de concordance

Liang et Zeger (1986) ont élaboré des méthodes fondées sur les moments pour analyser les observations corrélées provenant d'une même grappe (par exemple mesures répétées au cours du temps sur le même individu ou observations sur plusieurs membres de la même famille). L'approche GEE donne une estimation cohérente des paramètres

marginaux, même en cas de spécification incorrecte de la structure de corrélation en utilisant un estimateur « sandwich » robuste de la variance. Nous utilisons l'approche GEE pour analyser les données d'enquête en intégrant additionnellement une matrice de poids de sondage comme suit :

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathbf{D}'_{hi}\,\mathbf{W}_{hi}\,\mathbf{V}_{hi}^{-1}\,(\mathbf{Y}_{hi}-\boldsymbol{\mu}_{hi}(\hat{\boldsymbol{\mu}}))=\mathbf{0},$$

où $\mathbf{D}'_{hi}$ est la matrice des dérivées partielles $d[\boldsymbol{\mu}_{hi}]'/d\boldsymbol{\mu}$ de dimensions $(q\times m_{hi})$, $\mathbf{W}_{hi}$ est une matrice diagonale principale de dimensions $(m_{hi}\times m_{hi})$ constituée des poids de sondage au niveau de la personne w_{hij}, $\mathbf{V}_{hi}$ est une matrice de variance-covariance de travail de dimensions $(m_{hi}\times m_{hi})$ pour les réponses intragrappe, $\mathbf{Y}_{hi}$ est un vecteur de réponses de dimension $(m_{hi}\times 1)$ comprenant les réponses Y_{hij}, et $\boldsymbol{\mu}_{hi}=E[\mathbf{Y}_{hi}]$ est éventuellement une fonction du vecteur de paramètres $\boldsymbol{\beta}$ de dimension $(q\times 1)$. Les équations d'estimation généralisées (GEE) peuvent alors être résolues non itérativement, ce qui donne l'estimation habituelle

$$\hat{\mu}=\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}\,Y_{hij}\right)\Bigg/\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}\right)$$

si nous estimons une moyenne commune $\mu=\boldsymbol{\beta}$ $(q=1)$ et que nous utilisons une matrice de covariance de travail sous indépendance.

Supposons qu'une paire de réponses continues, Y_{hij1} et Y_{hij2}, est observée pour le j^{e} membre de la i^{e} grappe de la h^{e} strate et que leurs valeurs prévues sont μ_{hij1} et μ_{hij2}. De nouveau, supposons que nous estimons les moyennes communes μ_1 et μ_2 sans covariables pour la paire de réponses continues intrasujet, qui peuvent être estimées en utilisant l'équation d'estimation généralisée susmentionnée.

Barnhart et Williamson (2001) ont montré comment trois ensembles d'équations d'estimation généralisées peuvent être utilisés pour modéliser le CCC défini en (1) en utilisant des données corrélées. Nous étendons le deuxième ensemble de GEE de Barnhart et Williamson (2001) afin d'estimer les variances des réponses continues en intégrant de nouveau une matrice de pondération comme suit :

$$\nu_2(\hat{\sigma}_1^2,\hat{\sigma}_1^2,\hat{\mu}_1,\hat{\mu}_2)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathbf{F}'_{hi}\,\mathbf{W}_{hi}\,\mathbf{H}_{hi}^{-1}(\mathbf{Y}_{hi}^2-\boldsymbol{\delta}_{hi}^2(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\mu}_1,\hat{\mu}_2))=\mathbf{0},$$

où $\mathbf{F}'_{hi}$ est la matrice des dérivées partielles $d[\boldsymbol{\delta}_{hi}^2]'/d\boldsymbol{\sigma}^2$ de dimensions $(2\times 2m_{hi})$ avec $\boldsymbol{\sigma}^2=[\sigma_1^2,\sigma_2^2]'$, $\mathbf{W}_{hi}$ est une matrice diagonale principale des poids d'échantillonnage au niveau de la personne w_{hij} de dimensions $(2m_{hi}\times 2m_{hi})$, $\mathbf{H}_{hi}$ est une matrice de variance-covariance de travail de dimensions $(2m_{hi}\times 2m_{hi})$ pour les carrés des réponses intragrappe, $\mathbf{Y}_{hi}^2=[Y_{hi11}^2,Y_{hi12}^2,Y_{hi21}^2,Y_{hi22}^2,\ldots,Y_{him_{hi}1}^2,Y_{him_{hi}2}^2]'$ est un vecteur réponse des variables continues de dimensions $(2m_{hi}\times 1)$, et $\boldsymbol{\delta}_{hi}^2=E[\mathbf{Y}_{hi}^2]$. Bien que $\boldsymbol{\delta}_{hi}^2$ soit une fonction des termes de variance σ_1^2 et σ_2^2 ainsi que des moyennes μ_1 et μ_2, nous supposons que les moyennes sont fixes dans $\boldsymbol{\delta}_{hi}^2$ et nous ne calculons les dérivées de $\boldsymbol{\delta}_{hi}^2$ que par rapport aux variances. De nouveau, nous choisissons la matrice $\mathbf{H}_{hi}$ de dimensions $(2m_{hi}\times 2m_{hi})$ comme matrice de variance-covariance de travail sous « indépendance », ainsi que le vecteur colonne $\boldsymbol{\delta}_{hi}^2=[\sigma_1^2+\mu_1^2,\sigma_2^2+\mu_2^2,\ldots,\sigma_1^2+\mu_1^2,\sigma_2^2+\mu_2^2]'$ de dimension $(2m_{hi}\times 1)$ parce que nous supposons que les variances et les moyennes sont communes dans toutes les strates et toutes les grappes. Les équations d'estimation généralisées susmentionnées peuvent alors être résolues non itérativement :

$$\hat{\sigma}_p^2=\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}}W_{hijp}\,Y_{hijp}^2\right)\Bigg/\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}}W_{hijp}\right)-\hat{\mu}_p^2,$$

pour la p^{e} mesure dans la paire, $p=1,2$.

Le CCC peut être estimé au moyen d'un troisième ensemble d'équations d'estimation en utilisant les produits par paire des réponses pour modéliser σ_{12}, une fois que les moyennes et les variances sont estimées. Soit $\mathbf{U}_{hi}=[Y_{hi11}Y_{hi12},Y_{hi21}Y_{hi22},\ldots,Y_{him_{hi}1}Y_{him_{hi}2}]'$ un vecteur de dimension $(m_{hi}\times 1)$ des produits par paire des réponses et écrivons $\boldsymbol{\theta}_{hi}=E[\mathbf{U}_{hi}]$, qui est une fonction des moyennes, des variances et du CCC. Pour trouver la valeur de $\hat{\rho}_c$, nous résolvons un troisième ensemble d'équations d'estimation :

$$\nu_3(\hat{\rho}_c,\hat{\sigma}_1^2,\hat{\sigma}_1^2,\hat{\mu}_1,\hat{\mu}_2)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathbf{C}'_{hi}\,\mathbf{W}_{hi}\,\mathbf{K}_{hi}^{-1}(\mathbf{U}_{hi}-\boldsymbol{\theta}_{hi}(\hat{\rho}_c,\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\mu}_1,\hat{\mu}_2))=0,$$

où $\mathbf{C}'_{hi}$ est un vecteur de dérivées partielles $\partial\boldsymbol{\theta}_{hi}/\partial\rho_c$ de dimension $(1\times m_{hi})$, $\mathbf{W}_{hi}$ est une matrice diagonale principale de dimensions $(m_{hi}\times m_{hi})$ comprenant les poids d'échantillonnage au niveau de la personne w_{hij}, et $\mathbf{K}_{hi}$ est une matrice de covariance de travail de dimensions $(m_{hi}\times m_{hi})$ que nous choisissons comme matrice de covariance sous « indépendance ». Les équations d'estimation généralisées susmentionnées peuvent être résolues non itérativement :

$$\hat{\rho}_c=\frac{2\hat{\sigma}_{12}}{\hat{\sigma}_1^2+\hat{\sigma}_2^2+(\hat{\mu}_1-\hat{\mu}_2)^2},$$

où

$$\hat{\sigma}_{12} = \frac{\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} W_{hij12}\, Y_{hij1}\, Y_{hij2}\right)}{\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} W_{hij12}\right)} - \hat{\mu}_1\hat{\mu}_2.$$

2.2 Estimateur de la variance par linéarisation

Les estimateurs robustes habituels de la variance des moyennes et du CCC selon l'approche GEE ne sont pas valides ici puisqu'ils ne tiennent pas compte de la structure d'échantillonnage, ne prenant seulement en considération que la corrélation des observations faites sur le même individu. Nous proposons d'estimer l'erreur-type par la méthode de linéarisation de Taylor (Binder 1983 ; Binder 1996). Les dérivées premières de ρ_c (équation 1) par rapport à μ_1, μ_2, σ_1^2, σ_2^2, et σ_{12} sont :

$$\begin{aligned}
\frac{\partial \rho_c}{\partial \mu_1} &= \frac{-4\sigma_{12}(\mu_1-\mu_2)}{D^2},\\
\frac{\partial \rho_c}{\partial \mu_2} &= \frac{-4\sigma_{12}(\mu_2-\mu_1)}{D^2},\\
\frac{\partial \rho_c}{\partial \sigma_1^2} &= \frac{-2\sigma_{12}}{D^2},\\
\frac{\partial \rho_c}{\partial \sigma_2^2} &= \frac{-2\sigma_{12}}{D^2},\\
\frac{\partial \rho_c}{\partial \sigma_{12}} &= \frac{2}{D},
\end{aligned}$$

où $D = \sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2$. Donc,

$$\begin{aligned}
\hat{\rho}_c - \rho_c \approx{}& \left(\frac{\partial \rho_c}{\partial \mu_1}\right)(\hat{\mu}_1 - \mu_1) + \left(\frac{\partial \rho_c}{\partial \mu_2}\right)(\hat{\mu}_2 - \mu_2)\\
&+ \left(\frac{\partial \rho_c}{\partial \sigma_1^2}\right)(\hat{\sigma}_1^2 - \sigma_1^2)\\
&+ \left(\frac{\partial \rho_c}{\partial \sigma_2^2}\right)(\hat{\sigma}_2^2 - \sigma_2^2) + \left(\frac{\partial \rho_c}{\partial \sigma_{12}}\right)(\hat{\sigma}_{12} - \sigma_{12})\\
={}& \frac{-4\sigma_{12}(\mu_1-\mu_2)}{D^2}(\hat{\mu}_1 - \mu_1)\\
&+ \frac{-4\sigma_{12}(\mu_2-\mu_1)}{D^2}(\hat{\mu}_2 - \mu_2)\\
&+ \frac{-2\sigma_{12}}{D^2}(\hat{\sigma}_1^2 - \sigma_1^2) + \frac{-2\sigma_{12}}{D^2}(\hat{\sigma}_2^2 - \sigma_2^2)\\
&+ \frac{2}{D}(\hat{\sigma}_{12} - \sigma_{12}).
\end{aligned}$$

L'équation ci-dessus peut être réécrite en deux parties, l'une comprenant les estimations des paramètres $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, et $\hat{\sigma}_{12}$, et l'autre, uniquement les paramètres qui ne contribuent pas à l'estimation de la variance de $\hat{\rho}_c$. Donc, la première partie devient

$$\begin{aligned}
&- \frac{4\sigma_{12}(\mu_1-\mu_2)}{D^2}\hat{\mu}_1 - \frac{4\sigma_{12}(\mu_2-\mu_1)}{D^2}\hat{\mu}_2\\
&\quad - \frac{2\sigma_{12}}{D^2}\hat{\sigma}_1^2 - \frac{2\sigma_{12}}{D^2}\hat{\sigma}_2^2 + \frac{2}{D}\hat{\sigma}_{12}\\
&= - \frac{2\sigma_{12}}{D^2}(2(\mu_1-\mu_2)(\hat{\mu}_1 - \hat{\mu}_2) + \hat{\sigma}_1^2 + \hat{\sigma}_2^2) + \frac{2}{D}\hat{\sigma}_{12}\\
&= - \frac{2\sigma_{12}}{D^2}\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} 2(\mu_1-\mu_2)(w_{hij}^* Y_{hij1} - w_{hij}^* Y_{hij2})\right.\\
&\qquad \left. + w_{hij}^*(Y_{hij1}-\mu_1)^2 + w_{hij}^*(Y_{hij2}-\mu_2)^2\right)\\
&\quad + \frac{2}{D}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}^*(Y_{hij1}-\mu_1)(Y_{hij2}-\mu_2) \qquad (3)
\end{aligned}$$

où $w_{hij}^* = w_{hij} / (\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij})$. Après mise en évidence du terme de sommation, l'équation (3) devient une fonction linéaire des données qui peut alors être exprimée sous la forme $\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}^*\, z_{hij}$, où

$$\begin{aligned}
z_{hij} = {}& - \frac{2\sigma_{12}}{D^2}(2(\mu_1-\mu_2)(Y_{hij1} - Y_{hij2})\\
&+ (Y_{hij1}-\mu_1)^2 + (Y_{hij2}-\mu_2)^2)\\
&+ \frac{2}{D}(Y_{hij1}-\mu_1)(Y_{hij2}-\mu_2). \qquad (4)
\end{aligned}$$

On crée alors à partir de l'équation (4) une variable aléatoire $\hat{z}_{hij}$ qui remplace les paramètres par leurs estimations respectives. La variance de ce nouvel estimateur $\hat{z}_{hij}$ est une approximation de la variance de $\hat{\rho}_c$, qui peut être estimée en utilisant un logiciel standard d'analyse de données d'enquête (voir l'annexe).

2.3 Estimateur jackknife de la variance

Nous utilisons aussi la technique du jackknife pour estimer l'erreur-type des paramètres selon la méthode de Rust et Rao (1996, section 2.1) en vue de comparer les résultats aux estimations par linéarisation. La technique du jackknife est mise en œuvre en calculant un ensemble d'estimations répétées et en estimant la variance en se servant de ces estimations. Un jeu de données répétées est créé pour chaque grappe en supprimant de l'échantillon toutes les observations provenant de la grappe en question. Les poids de toutes les autres observations qui se trouvent dans la strate contenant la grappe sont augmentés d'un facteur $n_h / (n_h - 1)$. Les poids dans les autres strates restent inchangés. Donc, les nouveaux poids pour le jeu de données répétées créé en supprimant la grappe i de la strate h sont :

$\omega_{klj}^{(hi)} = w_{klj}$ si $k \neq h$ (strates différentes)

$\omega_{hlj}^{(hi)} = w_{hlj} n_h /(n_h - 1)$ si $l \neq i$ (même strate mais grappes différentes)

$\omega_{hij}^{(hi)} = 0$ (pour la grappe supprimée).

L'estimateur de variance jackknife résultant pour $\hat{\rho}_c$ est

$$v_J(\hat{\rho}_c) = \sum_{h=1}^{H} \left(\frac{n_h - 1}{n_h} \right) \sum_{i=1}^{n_h} (\hat{\rho}_{c(hi)} - \hat{\rho}_c)^2$$

où $\hat{\rho}_{c(hi)}$ est estimé de la même façon que $\hat{\rho}_c$, mais en utilisant les poids recalculés $\omega^{(hi)}$ au lieu des poids ω. Les estimateurs jackknife des moyennes sont calculés de la même façon.

2.4 Le coefficient kappa

Supposons qu'une paire de réponses binaires, Y_{hij1} et Y_{hij2}, est observée pour le j^{e} membre de la i^{e} grappe de la h^{e} strate et que leurs valeurs prévues sont les probabilités π_{hij1} et π_{hij2}. De nouveau, supposons que nous estimons les probabilités communes π_1 et π_2 sans covariables pour la paire de réponses binaires intrasujet. Lipsitz et coll. (1994) ont montré comment se servir de deux ensembles d'équations d'estimation généralisées pour élaborer de simples estimations non itératives du coefficient κ qui peuvent être utilisées pour des données non équilibrées, parce que les estimations antérieures de kappa et de sa variance n'étaient proposées que pour des données équilibrées. Ils ont défini la variable aléatoire binaire $U_{hij} = Y_{hij1}Y_{hij2} + (1 - Y_{hij1})(1 - Y_{hij2}) = 1$ si les deux réponses formant la paire concordent et 0 autrement. Par conséquent, $E[U_{hij}] = P_o$, qui désigne la probabilité d'observer un accord et que l'on suppose ici être constante sur toutes les strates, grappes et paires d'observations. Posons maintenant que $E[Y_{hij1}Y_{hij2}] = \Pr[Y_{hij1} = Y_{hij2} = 1] = \omega$. La probabilité d'observer un accord peut être exprimée par $P_o = 1 - \pi_1 - \pi_2 + 2\omega$. La probabilité d'un accord dû au hasard est définie comme étant $P_e = \pi_1\pi_2 + (1 - \pi_1)(1 - \pi_2)$ et est estimée par $\hat{P}_e = \hat{\pi}_1\hat{\pi}_2 + (1 - \hat{\pi}_1)(1 - \hat{\pi}_2)$, où $\hat{\pi}_1$ et $\hat{\pi}_2$ sont calculées dans le premier ensemble d'équations d'estimation.

Nous pouvons calculer des estimations de κ d'après des données d'enquête en suivant l'approche décrite pour le CCC à la section 2.1. Nous pouvons intégrer les matrices de poids de sondage dans les deux ensembles d'équations d'estimation généralisées de Lipsitz et coll. (1994) pour estimer kappa. Puis, en choisissant les matrices de covariance de travail sous « indépendance » pour les deux ensembles d'équations comme dans l'approche de Lipsitz et coll. (1994), nous arrivons à l'estimation non itérative suivante de kappa pour des données d'enquête :

$$\hat{\kappa} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} U_{hij} - \hat{P}_e \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} - \hat{P}_e \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}}. \quad (5)$$

Cet estimateur est identique à celui de Lumley (2010), qui peut être calculé en utilisant le progiciel R pour données d'enquête et la fonction svykappa.

On peut estimer l'erreur-type de $\hat{\kappa}$ de la même façon que celle de $\hat{\rho}_c$ en utilisant la méthode de linéarisation de Taylor. Les dérivées premières de kappa par rapport à $P_o, \pi_1,$ et π_2 sont :

$$\frac{\partial\kappa}{\partial P_o} = \frac{1}{1 - P_e},$$
$$\frac{\partial\kappa}{\partial \pi_1} = \frac{(1 - P_o)(1 - 2\pi_2)}{(1 - P_e)^2},$$
$$\frac{\partial\kappa}{\partial \pi_2} = \frac{(1 - P_o)(1 - 2\pi_1)}{(1 - P_e)^2}.$$

Donc

$$\begin{aligned}\hat{\kappa} - \kappa &\approx \left(\frac{\partial\kappa}{\partial P_o}\right)(\hat{P}_o - P_o) \\ &\quad + \left(\frac{\partial\kappa}{\partial \pi_1}\right)(\hat{\pi}_1 - \pi_1) + \left(\frac{\partial\kappa}{\partial \pi_2}\right)(\hat{\pi}_2 - \pi_2) \\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w^*_{hij} z_{hij},\end{aligned}$$

où $w^*_{hij} = w_{hij} / (\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij})$ et

$$\begin{aligned}z_{hij} &= \left(\frac{\partial\kappa}{\partial P_o}\right)U_{hij} + \left(\frac{\partial\kappa}{\partial \pi_1}\right)Y_{hij1} + \left(\frac{\partial\kappa}{\partial \pi_2}\right)Y_{hij2} \\ &= \frac{U_{hij}}{1 - P_e} + \frac{(1 - P_o)[Y_{hij1}(1 - 2\pi_2) + Y_{hij2}(1 - 2\pi_1)]}{(1 - P_e)^2}. \quad (6)\end{aligned}$$

En remplaçant dans (6) les paramètres par leurs estimations respectives, on traite ensuite $\hat{z}_{hij}$ comme une variable aléatoire et on estime sa variance en utilisant un logiciel standard d'analyse de données d'enquête qui tient compte du plan de sondage. La variance de ce nouvel estimateur $\hat{z}_{hij}$ est une approximation de la variance de $\hat{\kappa}$. La méthode du jackknife peut également être utilisée pour estimer la variance de $\hat{\kappa}$.

3. Enquête NHANES III

Nous avons utilisé des données provenant de la troisième National Health and Nutrition Examination Survey pour illustrer notre méthode. La NHANES III, qui a été réalisée

par le National Center for Health Statistics des Centers for Disease Control and Prevention, a été conçue comme une enquête de six années divisée en deux phases (1988 à 1991 et 1991 à 1994). Les données ont été recueillies suivant un plan de sondage probabiliste complexe, à plusieurs degrés, en vue de sélectionner des participants représentatifs de la population américaine civile, ne vivant pas en établissement. Des renseignements détaillés sur le plan de sondage ainsi que les lignes directrices concernant l'analyse et la diffusion des données ont été publiés dans les manuels de référence et les rapports de la NHANES III (National Center for Health Statistics 1996).

3.1 Étude du poids chez les adolescents

L'obésité est un problème de santé publique dont l'ampleur augmente rapidement et dont la surveillance est le plus souvent fondée sur des valeurs autodéclarées de la taille et du poids. Une série d'études récentes et d'examens systémiques ont eu pour but d'essayer d'évaluer l'accord entre le poids autodéclaré et le poids mesuré, en particulier chez la population d'adolescents. Selon les résultats généraux, le poids autodéclaré serait légèrement inférieur au poids mesuré, et un nombre considérable d'adolescents de poids normal se percevraient à tort comme ayant un excès de poids et adopteraient des comportements de contrôle du poids néfastes pour la santé (Field, Aneja et Rosner 2007 ; Gorber, Tremblay, Moher et Gorber 2007 ; Sherry, Jefferds et Grummer-Strawn 2007). Donc, les chercheurs ont laissé entendre que les programmes de prévention de l'obésité devraient aborder la question des perceptions fausses du poids et les effets indésirables des méthodes de contrôle du poids nuisibles pour la santé, même chez les adolescents ayant un poids normal (Talamayan, Springer, Kelder, Gorospe et Joye 2006). Une étude canadienne comparable fondée sur des données de l'Enquête sur la santé dans les collectivités canadiennes de 2005 portant sur la population adulte a également révélé que les associations entre l'obésité et les problèmes de santé peuvent être surestimées si l'on se sert du poids autodéclaré (Shield, Gorber et Tremblay 2008). Nous utilisons des données provenant de la composante des mesures corporelles (anthropométrie) de la NHANES III pour estimer le CCC qui mesure l'accord entre le poids autodéclaré et le poids mesuré (en livres) chez les adolescents (de 12 à 16 ans).

L'autodéclaration du poids a été obtenue juste avant sa mesure réelle. Nous utilisons les données couvrant la période complète de six ans de l'enquête (de 1988 à 1991 et de 1991 à 1994). Pour simplifier, nous avons exclu une strate qui ne contenait qu'une seule UPE. Le nombre de strates était donc de 48 et chaque strate contenait 2 UPE. Nous avons utilisé dans les analyses le poids de sondage étiqueté wtpfex6 qui tient compte des probabilités de sélection inégales. Des données complètes pour les deux mesures du poids étaient disponibles pour 1 651 sujets. Les estimations du poids autodéclaré et du poids réel (en livres), calculées en utilisant PROC SURVEYMEANS en SAS, étaient égales à 135,5 (e.-t. = 1,8) et 136,3 (e.-t. = 1,8), respectivement. Les estimations des erreurs-types obtenues par l'approche du jackknife sont les mêmes que celles susmentionnées.

Le CCC est un choix naturel pour évaluer l'accord entre deux mesures du poids, parce que celles-ci sont faites sur la même échelle et que leurs étendues sont comparables (poids autodéclaré : 78 lb ~ 350 lb et poids réel : 73 lb ~ 372 lb) (Lin et Chinchilli 1997). L'estimation du CCC pour évaluer l'accord entre les deux définitions du poids en utilisant la méthode proposée est égale à 0,93. L'erreur-type de l'estimation est égale à 0,021 en utilisant la méthode de linéarisation de Taylor. L'erreur-type par le jackknife de 0,021 concorde étroitement avec l'erreur-type par linéarisation. Ces statistiques sont résumées au tableau 1 ainsi que leur valeur calculée quand il n'est pas tenu compte de la structure d'échantillonnage. Les erreurs-types des estimations tenant compte du plan de sondage sont beaucoup plus grandes que les estimations non pondérées.

Tableau 1
Moyenne non pondérée et pondérée, CCC, et erreurs-types respectives pour le poids autodéclaré et le poids réel des adolescents, en livres

	Autodéclaré	Réel	CCC
Estimation non pondérée	135,31	136,96	0,890
E.-T.	0,76	0,80	0,0005
Estimation pondérée	135,47	136,30	0,926
E.-T.	1,75	1,82	0,0205

Comme le CCC, le coefficient de corrélation de Pearson usuel entre les mesures autodéclarées et le poids réel est également égal à 0,93. Dans ce cas, l'écart moyen entre les deux mesures du poids est à peine inférieur à une livre. Lorsque l'on examine les sous-populations, on constate des différences entre le CCC et le coefficient de corrélation de Pearson. Considérons la sous-population de personnes dont le poids mesuré était supérieur à 200 lb au moment de l'examen physique. Si l'on résume les données pour cette sous-population, le poids autodéclaré est, en moyenne, inférieur de 8 livres au poids mesuré (223,2 lb c. 231,4 lb). Il existe un léger écart du CCC (0,72) par rapport au coefficient de corrélation de Pearson (0,76). La différence entre les deux mesures augmente chez le sous-groupe des personnes plus obèses. Dans la sous-population pour laquelle le poids mesuré est supérieur à 220 lb, les moyennes des poids autodéclarés et mesurés sont 231,9 lb et 248,8 lb, respectivement. Le CCC est égal à 0,67, tandis que le

coefficient de corrélation de Pearson est égal à 0,85. Dans cette situation, le CCC reflète la reproductibilité ainsi que les différences entre les moyennes des poids autodéclarés et mesurés. Par conséquent, le CCC est informatif et avantageux si l'on s'intéresse à ces comparaisons, particulièrement dans l'analyse par domaine dans le cas d'une enquête complexe.

3.2 Étude de la santé buccodentaire

Slade et Beck (1999) ont utilisé la mesure de la profondeur de la poche parodontale et de la perte d'attachement comme indice de maladies parodontale. La prévalence de la maladie parodontale en utilisant les seuils publiés antérieurement de profondeur de la poche parodontale ≥ 4 mm et de perte d'attache ≥ 3 mm a été estimée par Slade et Beck (1999, tableau 1). La profondeur de la poche parodontale peut être le reflet d'une inflammation plutôt que d'une maladie parodontale chronique, de sorte que le niveau d'attachement de la gencive pourrait être une mesure plus significative de la destruction parodontale. Néanmoins, la profondeur de la poche parodontale demeure la mesure recommandée en pratique clinique (Winn, Johnson et Kingman 1999). Par conséquent, nous comparons l'accord de ces deux définitions de la maladie parodontale en utilisant le coefficient kappa.

Nous nous servons de l'échantillon qui a été analysé par Slade et Beck (1999). Les données portent sur 14 415 personnes de 13 ans et plus pour lesquelles une évaluation de la profondeur des poches parodontales et de la perte d'attachement avait été effectuée par six dentistes désignés. De nouveau, nous utilisons les données recueillies pour la période complète de six ans de l'enquête (1988 à 1991 ainsi que 1991 à 1994). Le nombre total de strates était de 49 et chaque strate contenait deux UPE. Nous avons utilisé dans nos analyses la variable étiquetée poids d'échantillonnage, wtpfex6, qui tient compte des probabilités de sélection différentes.

La première définition de la maladie parodontale correspond à une profondeur de la poche parodontale $\geq$ 4 mm et la deuxième, à une perte d'attachement maximale ≥ 3 mm. Pour les deux variables, nous utilisons les valeurs maximales pour l'ensemble des dents dans la bouche d'une personne. Les estimations de la probabilité pour les variables de perte d'attachement et de profondeur de la poche parodontale sont égales à 0,358 (e.-t. jackknife = 0,0088) et 0,212 (e.-t. jackknife = 0,016), respectivement, en utilisant la méthode proposée. Les erreurs-types asymptotiques fondées sur le développement en série de Taylor usuel (Woodruff 1971, produites par PROC SURVEYFREQ en SAS, version 9.1) sont égales à 0,0088 et 0,015, respectivement.

Le coefficient kappa est un choix naturel pour évaluer l'accord entre deux évaluations binaires car il comporte une correction pour l'accord dû au hasard (Fleiss 1981). L'estimation de kappa pour mesurer l'accord entre les deux définitions de la maladie parodontale (profondeur de la poche parodontale ≥ 4 mm et perte d'attachement ≥ 3 mm) en utilisant la méthode proposée est égale à 0,307. L'erreur-type de 0,0158 a été obtenue par la méthode de linéarisation de Taylor ainsi que par la méthode du jackknife. Au tableau 2, ces résultats sont comparés aux mesures obtenues quand on ne tient pas compte du plan de sondage complexe. L'erreur-type du coefficient kappa est plus grande lorsque l'on prend en considération le plan de sondage.

Tableau 2
Moyenne non pondérée et pondérée, kappa et erreurs-types respectives pour la perte d'attachement et la profondeur de la poche parodontale

	Perte d'attachement	Profondeur de la poche	Kappa
Estimation non pondérée	0,393	0,283	0,334
E.-T.	0,004	0,004	0,008
Estimation pondérée	0,358	0,212	0,307
E.-T.	0,009	0,016	0,0158

4. Discussion

Le CCC et le coefficient kappa évaluent l'accord entre deux mesures pour des réponses continues et catégorielles, respectivement. Dans le présent article, nous avons proposé une approche fondée sur des équations d'estimation généralisées pour estimer le CCC pour une paire de variables continues et le coefficient kappa pour une paire de variables binaires, d'après des données d'enquête, quand les données ont été recueillies selon un plan de sondage complexe comportant des caractéristiques telles que la stratification ou la mise en grappes. L'estimateur de type sandwich habituel de la variance ne tient compte que des mesures répétées faites sur la même personne et ne prend pas en considération le cadre d'échantillonnage (par exemple grappes, stratification et pondération). Dans l'approche GEE, les erreurs-types des estimateurs sont estimées par les méthodes de linéarisation de Taylor et du jackknife. Si les données ne sont pas recueillies selon un plan de sondage complexe, les estimateurs proposés sont identiques aux estimateurs habituels. Comme le montre les deux exemples tirés de l'étude NHANES III, il est nécessaire d'intégrer dans l'analyse les poids de sondage et les caractéristiques du plan de sondage afin que les erreurs-types ne soient pas sous-estimées quand les données sont recueillies selon un plan de sondage complexe. Les tableaux 1 et 2 révèlent des écarts importants

entre les erreurs-types des estimations pondérées et non pondérées des erreurs-types, pour le CCC ainsi que le coefficient kappa. Les intervalles de confiance qui tiennent compte des poids de sondage et des caractéristiques du plan permettent d'obtenir des inférences correctes.

À l'annexe, nous montrons les étapes du calcul des mesures pondérées du CCC et du coefficient kappa ainsi que de leurs erreurs-types en utilisant un logiciel standard d'analyse de données d'enquête qui prend en compte les poids de sondage, la mise en grappes et la stratification. L'approche GEE est avantageuse parce qu'il s'agit d'un cadre commode pour élaborer des estimateurs des coefficients d'accord et qu'il est facile de l'étendre à des évaluateurs multiples, des méthodes multiples, un ajustement des covariables et des tailles de grappes non équilibrées. Cette approche fondée sur le plan de sondage donne une estimation correcte de l'erreur-type sans que l'on doive émettre des hypothèses concernant un modèle sous-jacent et tenir compte du plan d'échantillonnage. Si l'on souhaite estimer l'accord entre deux variables ordinales à l'aide du coefficient kappa, l'approche des équations d'estimation généralisées de Williamson et coll. (2000) peut être étendue de la même façon à la méthode proposée.

Remerciements

Nous remercions les examinateurs et l'éditeur pour leurs commentaires utiles. Entre autres, l'éditeur a fourni des suggestions très judicieuses pour la section d'introduction.

Annexe

Étapes du calcul du CCC et de son erreur-type en utilisant un logiciel standard d'analyse de données d'enquête

Étape 1 : Calculer les moyennes des variables continues Y_{hij1} et Y_{hij2} en utilisant un logiciel d'analyse de données d'enquête qui intègre la stratification, la mise en grappes et la pondération de l'échantillon (par exemple PROC SURVEYMEANS en SAS).

Étape 2 : Élever au carré les valeurs de Y_{hij1} et Y_{hij2} centrées autour de leur moyenne respective.

Étape 3 : Calculer les moyennes des carrés des valeurs centrées de Y_{hij1} et Y_{hij2} en utilisant un logiciel standard pour l'analyse des données d'enquête. Ces moyennes sont les estimations de la variance de Y_{hij1} et Y_{hij2}. Calculer la moyenne du produit des valeurs centrées de Y_{hij1} et Y_{hij2} en utilisant un logiciel standard pour l'analyse de données d'enquête. Cette moyenne est la covariance estimée de Y_{hij1} et Y_{hij2}.

Étape 4 : Calculer le CCC en substituant les moyennes et les variances estimées dans l'équation (1). Créer la nouvelle variable Z_{hij} fondée sur l'équation (4).

Étape 5 : Calculer l'erreur-type de Z_{hij} en utilisant un logiciel standard d'analyse des données d'enquête. L'erreur-type de Z_{hij} estime l'erreur-type de $\hat{\rho}_c$.

CODE SAS :

Soient $y1$ et $y2$ les variables de la paire de réponses continues, et s, c et w, les variables pour les strates, les grappes et les poids :

```
PROC SURVEYMEANS DATA=dataset MEAN;
                                        /*Étape 1 susmentionnée*/ ;
      STRATA s;
      CLUSTER c;
      WEIGHT w;
      VAR y1 y2;
      ODS OUTPUT STATISTICS=stat;
data _null_;
      set stat (where=(varname='y1' ));
      call symputx('muy1', mean);
data _null_;
      set stat (where=(varname='y2' ));
      call symputx('muy2', mean);
data dataset; set dataset;              /*Étape 2 susmentionnée*/ ;
       cy1 = y1 – &muy1;
       cy2 = y2 – &muy2;
       vary1 = cy1 * *2;
       vary2 = cy2 * *2;
       covy12 = cy1 * cy2;
PROC SURVEYMEANS MEAN;                  /*Étape 3 susmentionnée*/ ;
      STRATA s;
      CLUSTER c;
      WEIGHT w;
      VAR vary1 vary2 covy12;
      ODS OUTPUT STATISTICS=stat;
run;
data _null_;
      set stat (where=(varname='vary1' ));
      call symputx('vary1', mean);
data _null_;
      set stat (where=(varname='vary2' ));
      call symputx('vary2', mean);
data _null_;
      set stat (where=(varname='covy12' ));
      call symputx('covy12', mean);
data dataset; set dataset;              /*Étape 4 susmentionnée*/ ;
       d = &vary1 + &vary2 + (&muy1 – &muy2) ** 2;
       CCC = 2 * &covy12 / d;
       z = (2 / d) * (cy1 * cy2) – (2 * &covy12 / d / d) * ((cy1 * *2) +
       (cy2 * *2) + 2 * (&muy1 – &muy2) * (y1 – y2));
PROC SURVEYMEANS MEAN;                  /*Étape 5 susmentionnée*/ ;
      STRATA s;
      CLUSTER c;
      WEIGHT w;
      VAR CCC z;
run;
```

Étapes du calcul de kappa et de son erreur-type en utilisant un logiciel standard d'analyse de données d'enquête

Étape 1 : Estimer les probabilités des variables binaires Y_{hij1} et Y_{hij2} en utilisant un logiciel d'analyse de données d'enquête qui intègre la stratification, la mise en grappes et la pondération de l'échantillon (par exemple PROC SURVEYFREQ en SAS).

Étape 2 : Estimer $P_e(= \hat{\pi}_1\hat{\pi}_2 + (1-\hat{\pi}_1)(1-\hat{\pi}_2))$.

Étape 3 : Créer la nouvelle variable d'accord $U_{hij}(=Y_{hij1}\ Y_{hij2} + (1-Y_{hij1})(1-Y_{hij2}))$.

Étape 4 : Calculer la somme des poids de sondage et la somme pondérée des U_{hij} (par exemple, en utilisant PROC SURVEYMEANS en SAS). Estimer kappa en utilisant l'équation (2).

Étape 5 : Créer une nouvelle variable z_{hij} en utilisant l'équation (6).

Étape 6 : Calculer l'erreur-type de z_{hij} en utilisant un logiciel standard d'analyse des données d'enquête. L'erreur-type de z_{hij} estime l'erreur-type de $\hat{\kappa}$.

Bibliographie

Barnhart, H.X., et Williamson, J.M. (2001). Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics*, 57, 931-940.

Behavioral Risk Factor Surveillance System (BRFSS). http://www.cdc.gov/BRFSS.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 20, 37-46.

Binder, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases : une approche de type « recette ». *Techniques d'enquête*, 22, 17-22.

Carrasco, J.L., et Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, 59, 849-858.

Chinchilli, V.M., Martel, J.K., Kumanyika, S. et Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, 52, 341-353.

Cochran, W.G. (1963). *Sampling Techniques*, 2e Éd. New York : John Wiley & Sons, Inc.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

Dixon, A.E., Sugar, E.A., Zinreich, S.J., Slavin, R.G., Corren, J., Naclerio, R.M., Ishii, M., Cohen, R.I., Brown, E.D., Wise, R.A. et et Irvin, C.G. (2009). Criteria to screen for chronic sinonasal disease. *Chest*, 136 (5), 1324-1332.

Feinleib, M., Garrison, R.J., Fabsitz, R.R., Christian, J.C., Hrubec, Z., Borhani, N.O., Kannel, W.B., Roseman, R., Schwartz, J.T. et Wagner, J.O. (1977). The NHLBI Twin Study of cardiovascular disease risk factors: Methodology and summary of results. *American Journal of Epidemiology*, 106, 284-295.

Field, A.E., Aneja, P. et Rosner, B. (2007). The validity of self-reported weight change among adolescents and young adults. *Obesity*, 15, 2357-2364.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2e Édition. New York : John Wiley & Sons, Inc.

Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York : John Wiley & Sons, Inc.

Fleiss, J.L., Nee, J.C.M. et Landis, J.R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86, 974-977.

Gilmour, E., Ellerbrock, T.V., Koulos, J.P., Chiasson, M.A., Williamson, J.M., Kuhn, L. et Wright, T.C. (1997). Measuring cervical ectopy: Direct visual assessment versus computerized planimetry. *American Journal of Obstetrics and Gynecology*, 176, 108-111.

Gorber, S.C., Tremblay, M., Moher, D. et Gorber, B. (2007). A comparison of direct *vs.* self-report measures for assessing height, weight and body mass index: A systematic review. *Obesity Review*, 8, 373-374.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York : John Wiley & Sons, Inc. Vols I and II.

Khawaja, I.S., Olson, E.J., van der Walt, C., Bukartyk, J., Somers, V., Dierkhising, R. et Morgenthaler, T.I. (2010). Diagnostic accuracy of split-night polysomnograms. *Journal of Clinical Sleep Medicine*, 6 (4), 357-362.

Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Klar, N., Lipsitz, S.R. et Ibrahim, J.G. (2000). An estimating equations approach for modeling kappa. *Biometrical Journal*, 42, 45-58.

Kocks, J.W., Kerstjens, H.A., Snijders, S.L., de Vos, B., Biermann, J.J., van Hengel, P., Strijbos, J.H., Bosveld, H.E. et van der Molen, T. (2010). Health status in routine clinical practice: validity of the clinical COPD questionnaire at the individual patient level. Health and Quality of Life Outcomes. 8, 135-141.

Korten, A.E., Jorm, A.F., Henderson, A.S., McCusker, E. et Creasey, H. (1992). Control-informant agreement on exposure history in case-control studies of Alzheimer's disease. *International Journal of Epidemiology*, 21, 1121-1131.

Liang, K.Y., et Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255-268.

Lin, L. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, 48, 599-604.

Lin, L., et Chinchilli, V. (1997). Rejoinder to the letter to the editor from Atkinson and Nevill. *Biometrics*, 53, 777-778.

Liottol, N., Radaelli, T., Orsi1, A., Taricco, E., Roggerol, P., Giann, M.L., Consonni, D., Mosca1, F. et Cetin, I. (2010). Relationship between in utero sonographic evaluation and subcutaneous plicometry after birth in infants with intrauterine growth restriction: An exploratory study. *Italian Journal of Pediatrics*, 36, 70-77.

Lipsitz, S.R., Laird, N.M. et Brennan, T.A. (1994). Simple moment estimates of the κ -coefficient and its variance. *Applied Statistics*, 43, 309-323.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R. Hoboken*. New Jersey : John Wiley & Sons, Inc.

MacDougall, H.G., Weber, K.P., McGarvie, L.A., Halmagyi, G.M. et Curthoys, I.S. (2009). The video head impulse test. Diagnostic accuracy in peripheral vestibulopathy. *Neurology*, 73, 1134-1141.

Maclure, M., et Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126, 161-169.

Mazaheri, Y., Hricak, H., Fine, S.W., Akin, O., Shukla-Dave, A., Ishill, N.M., Moskowitz, C.S., Grater, J.E., Reuter, V.E., Zakian, K.L., Touijer, K.A. et Koutcher, J.A. (2009). Prostate tumor volume measurement with combined T2-weighted imaging and diffusion-weighted MR: Correlation with pathologic tumor volume. *Radiology*, 252 (2), 449-457.

National Center for Health Statistics (2011). Third National Health and Nutrition Examination Survey, 1988-1994, NHANES III Examination data file (CD-ROM). http://www.cdc.gov/ nchs/nhanes.htm.

National Health Interview Survey (NHIS) (2011). http://www. cdc.gov/nchs/ nhis.htm.

Robieson, W. (1999). On weighted kappa and concordance correlation coefficient. Thèse de doctorat, University of Illinois in Chicago/Graduate College/Mathematics.

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.

Sherry, B., Jefferds, M.E. et Grummer-Strawn, L.M. (2007). Accuracy of adolescent self-report of height and weight in assessing overweight status: A literature review. *Archive of Pediatrics Adolescent Medicine*, 161, 1154-1161.

Shield, M., Gorber, S.C. et Tremblay, M.S. (2008). Effects of measurement on obesity and morbidity. *Health Reports*, 19, 77-84.

Slade, G.D., et Beck, J.D. (1999). Plausibility of periodontal disease estimates from NHANES III. *Journal of Public Health Dentistry*, 59, 67-72.

Talamayan, K.S., Springer, A.E., Kelder, S.H., Gorospe, E.C. et Joye, K.A. (2006). Prevalence of overweight misperception and weight control behaviors among normal weight adolescents in the United States. *The Scientific World Journal*, 6, 365-373.

Tanner, M.A., et Young, M.A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, 80, 175-180.

Williamson, J.M., Manatunga, A.K. et Lipsitz, S.R. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics*, 1, 191-202.

Winn, D.M., Johnson, C.L. et Kingman, A. (1999). Periodontal disease estimates in NHANES III: Clinical measurement and complex sample design issues. *Journal of Public Health Dentistry*, 59, 73-78.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

Zar, J.H. (1996). *Biostatistical Analysis. Upper Saddle River*. New Jersey : Prentice Hall International.

ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 81-88
Statistique Canada, N° 12-001-X au catalogue

Conjuguer des données synthétiques et le sous-échantillonnage pour créer des fichiers de microdonnées à grande diffusion pour les enquêtes à grande échelle

Jörg Drechsler et Jerome P. Reiter [1]

Résumé

Afin de créer des fichiers de données à grande diffusion à partir d'enquêtes à grande échelle, les organismes statistiques diffusent parfois des sous-échantillons aléatoires des enregistrements originaux. Le sous-échantillonnage aléatoire amenuise la taille des fichiers transmis aux analystes secondaires des données et réduit les risques de divulgation accidentelle de renseignements confidentiels sur les participants aux enquêtes. Cependant, le sous-échantillonnage n'élimine pas entièrement le risque, de sorte qu'il faut altérer les données avant leur diffusion. Nous proposons de créer des sous-échantillons protégés contre la divulgation provenant d'enquêtes à grande échelle en recourant à l'imputation multiple. L'idée consiste à remplacer dans l'échantillon original les valeurs identificatoires ou sensibles par des valeurs tirées de modèles statistiques et de diffuser des sous-échantillons de ces données protégées contre la divulgation. Nous présentons des méthodes permettant de faire des inférences fondées sur les multiples sous-échantillons synthétiques.

Mots clés : Confidentialité ; divulgation ; imputation multiple.

1. Introduction

Les instituts nationaux de la statistique (INS) tels que le U.S. Census Bureau et Statistique Canada réalisent des enquêtes à grande échelle, comme l'American Community Survey (ACS) et l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ), qui sont jugées très précieuses par les analystes secondaires des données. Alors que ceux-ci souhaitent avoir accès à autant de données que possible, l'INS doit veiller à ce que l'identité des participants à l'enquête et les attributs sensibles demeurent confidentiels. Une stratégie souvent adoptée pour réduire les risques de divulgation dans les études à grande échelle consiste à diffuser des sous-échantillons des données d'enquête originales. Ainsi, le Census Bureau diffuse un sous-échantillon des données de l'ACS englobant 1 % de l'ensemble des ménages américains (les données recueillies durant l'ACS couvrent 2,5 % de ces ménages), et Statistique Canada diffuse les données sur un échantillon de 20 % des personnes qui participent à l'ELNEJ. Pour une discussion de la protection de la confidentialité découlant de l'échantillonnage, consulter Willenborg et de Waal (2001), ainsi que Reiter (2005). Toutefois, le sous-échantillonnage n'élimine habituellement pas à lui seul les risques de divulgation, surtout si certaines unités du sous-échantillon présentent des combinaisons inhabituelles de caractéristiques. Par conséquent, les INS altèrent les données avant leur diffusion. Par exemple, dans le cas de l'ACS, le Census Bureau procède à la permutation de certaines données, au regroupement des valeurs extrêmes supérieures de certaines variables, à l'agrégation géographique et à la perturbation des données sur l'âge ; dans le cas de l'ELNEJ, Statistique Canada recourt à la permutation et à la suppression de données.

Quand elles sont appliquées intensivement, comme cela est parfois nécessaire pour assurer le respect de la confidentialité dans des enquêtes suscitant beaucoup d'attention, les stratégies classiques de contrôle de la divulgation peuvent fausser considérablement les inférences (Winkler 2007 ; Elliott et Purdam 2007 ; Drechsler et Reiter 2010). En outre, dans le cas de nombreuses techniques classiques, il est difficile pour les analystes des données, surtout ceux ne possédant pas une formation statistique poussée, de tenir compte comme il convient des effets du contrôle de la divulgation sur l'estimation. Motivés par ces limitations, nous proposons une nouvelle approche pour produire des échantillons de microdonnées à grande diffusion à partir d'enquêtes à grande échelle à laquelle nous donnons le nom de sous-échantillonnage avec synthèse. L'idée fondamentale consiste à remplacer dans l'échantillon original les valeurs identificatoires ou sensibles par des valeurs obtenues par tirages multiples à partir de modèles statistiques estimés en se servant du fichier de données originales, puis de diffuser des sous-échantillons de ces données protégées contre la divulgation. Les sous-échantillons peuvent être composés d'un ensemble commun d'enregistrements ou être tirés de manière indépendante.

Cette approche est une variante de celle des données partiellement synthétiques (Little 1993 ; Reiter 2003), qui a été utilisée aux États-Unis pour créer plusieurs produits de données à grande diffusion, y compris la Survey of Income

1. Jörg Drechsler, Institute for Employment Research, Department for Statistical Methods, Regensburger Straße 104, 90478 Nürnberg, Allemagne. Courriel : joerg.drechsler@iab.de ; Jerome P. Reiter, Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251. Courriel : jerry@stat.duke.edu.

and Program Participation, la Longitudinal Business Database, la Survey of Consumer Finances, les données sur les logements de groupe (*group quarters*) de l'American Community Survey et l'application Web *OnTheMap*. L'approche proposée ici diffère de la synthèse partielle en raison du sous-échantillonnage, qui nécessite que l'on ajuste les méthodes inférentielles de Reiter (2003) ; ces ajustements sont présentés ici. L'approche diffère aussi des méthodes de création d'échantillons de microdonnées à grande diffusion synthétiques d'après des données de recensement élaborées récemment par Drechsler et Reiter (2010). Dans le sous-échantillonnage avec synthèse, les données initiales proviennent d'une enquête et non d'un recensement ; donc, les inférences doivent tenir compte de l'incertitude additionnelle qui résulte de l'échantillonnage initial.

2. Approche générale

Nous décrivons maintenant la génération des données et les procédures d'inférence pour les deux approches de sous-échantillonnage avec synthèse, à savoir la diffusion de sous-échantillons différents (indépendants) ou la diffusion d'un même ensemble d'enregistrements dans chaque sous-échantillon. Les méthodes de génération des données, ainsi que celles permettant de faire des inférences valides d'après plusieurs ensembles de données dépendent de l'approche de sous-échantillonnage. Dans les deux approches, nous désignons par D les données d'échantillon originales recueillies auprès de n_1 unités échantillonnées dans une population constituée de N unités. Nous supposons pour commencer que le plan de sondage original est un échantillonnage aléatoire simple ; par après, nous étendons l'exposé à l'échantillonnage stratifié. Nous supposons que toutes les unités échantillonnées répondent complètement dans D. Contrairement à l'approche de synthèse partielle classique (Reiter 2004), les méthodes présentées ici n'ont pas été élaborées en vue de traiter simultanément les données manquantes et la synthèse avec sous-échantillonnage. Nous nous concentrons ici sur les descriptions générales des approches et sur la présentation des méthodes inférentielles. Nous ne discutons pas des stratégies de construction de modèles de synthèse ; voir Drechsler et Reiter (2009) et les références incluses pour des directives à ce sujet.

2.1 Diffusion de différents sous-échantillons aléatoires

2.1.1 Résumé de l'approche

Pour commencer, l'INS crée m ensembles de données partiellement synthétiques, $D_{syn} = \{D_i : i = 1, \ldots, m\}$, pour l'enquête originale en suivant l'approche de Reiter (2003). En particulier, l'INS remplace les valeurs identificatoires ou sensibles présentes dans D par des imputations multiples. Les modèles de synthèse sont estimés en ne se servant que des enregistrements dont les valeurs seront synthétisées. La synthèse est effectuée indépendamment m fois, ce qui donne D_{syn}. L'INS tire alors un sous-échantillon aléatoire simple de $n_2 < n_1$ enregistrements de chaque D_i. Ces m sous-échantillons, $d_{syn} = \{d_i : i = 1, \ldots, m\}$, sont diffusés aux membres du public.

L'analyste de d_{syn} cherche à faire des inférences au sujet d'un paramètre Q, par exemple une moyenne de population ou un coefficient de régression. Dans chaque sous-échantillon d_i, il estime Q au moyen d'un estimateur ponctuel q, et la variance de q, au moyen d'un estimateur u, en spécifiant q et u comme si d_i comprenait les données collectées durant l'enquête. Ici, u est spécifié en faisant abstraction de tout facteur de correction pour population finie ; par exemple, quand q est la moyenne d'échantillon, $u = s^2/n_2$, où s^2 représente la variance d'échantillon. Pour $i = 1, \ldots, m$, soit q_i et u_i les valeurs de q et u dans d_i. Les quantités qui suivent sont nécessaires pour les inférences.

$$\bar{q}_m = \sum_{i=1}^{m} q_i / m \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2 / (m-1) \tag{2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u_i / m. \tag{3}$$

L'analyste peut alors utiliser $\bar{q}_m$ pour estimer Q et

$$T_d = (n_2/n_1 - n_2/N)\,\bar{u}_m + b_m/m \tag{4}$$

pour estimer la variance de $\bar{q}_m$. Les dérivations de ces estimations sont présentées à la section 2.1.2. Il convient de noter que, sans le sous-échantillonnage, c'est-à-dire si $n_2 = n_1$, (4) est égal à l'estimation de la variance pour la synthèse partielle classique (Reiter 2003). Quand n_2 est grand, les inférences sont fondées sur une loi t, $(\bar{q}_m - Q) \sim t_{\nu_d}(0, T_d)$, dont le nombre de degrés de liberté est égal à $\nu_d = (m-1)(1 + (n_2/n_1 - n_2/N)\, m\bar{u}_m / b_m)^2$.

Les méthodes d'inférence peuvent être étendues à l'échantillonnage stratifié dans les conditions où l'INS utilise les mêmes strates pour le sous-échantillon et l'échantillon original. Soit N_h la taille de la population dans la strate h, où $h = 1, \ldots, H$. Pour chaque h, soit $\bar{q}_{mh}$ et T_{dh} les valeurs de (1) et (4) calculées en se servant uniquement des enregistrements compris dans d_{syn} dans la strate h. Ces estimations sont utilisées dans les inférences des quantités de population dans la strate h. Pour les inférences au sujet

de la moyenne de la population totale, l'estimation ponctuelle de Q est $\bar{q}_m = \sum_h (N_h / N)\, \bar{q}_{mh}$, et sa variance estimée est $T_d = \sum_h (N_h / N)^2 T_{dh}$. Les estimations ponctuelles et les estimations de variance pour des fonctions non linéaires des moyennes peuvent être obtenues en utilisant des développements en série de Taylor. Il convient de mentionner que les INS devraient diffuser les valeurs de n_{2h} / n_{1h} pour toutes les strates pour permettre l'estimation de la variance.

2.1.2 Dérivation des inférences pour l'approche des sous-échantillons aléatoires différents

L'analyste s'intéresse à $f(Q \mid d_{syn})$, qui peut s'écrire sous la forme

$$f(Q \mid d_{syn}) = \int f(Q \mid D_{syn}, d_{syn})\, f(D_{syn} \mid d_{syn})\, dD_{syn}. \quad (5)$$

Dans toutes les dérivations mathématiques de la section 2.1.2, nous supposons que l'analyste utilise les mêmes distributions que celles employées par l'INS pour créer D_{syn}. Nous supposons aussi que les tailles d'échantillons sont suffisantes pour permettre des approximations normales de ces distributions. Donc, pour chaque distribution, nous n'avons besoin que des deux premiers moments, que nous dérivons en nous servant d'arguments bayésiens en grand échantillon classiques. Nous émettons l'hypothèse de priors diffus pour tous les paramètres.

Soit Q_i et U_i l'estimation ponctuelle de Q et sa variance que l'analyste calculerait en se servant de D_i (dont il ne dispose pas). Soit les quantités $\bar{Q}_m$, $\bar{U}_m$, et B_m définis dans (1) à (3), mais en utilisant Q_i et U_i. Partant des résultats de synthèse partielle classique (Reiter 2003), nous avons $(Q \mid D_{syn}) \sim N(\bar{Q}_m, \bar{U}_m + B_m / m)$. Nous supposons que $(q_i \mid D_i) \sim N(Q_i, (1 - n_2 / n_1) u_i)$ et, comme il est habituel dans les contextes d'imputation multiple, que $u_i \approx \bar{u}_m$. Donc, en utilisant la théorie bayésienne classique, nous avons $(\bar{Q}_m \mid d_{syn}) \sim N(\bar{q}_m, (1 - n_2 / n_1)\, \bar{u}_m / m)$ et $((m-1) b_m / (B_m + (1 - n_2 / n_1)\, \bar{u}_m) \mid d_{syn}) \sim \chi^2_{m-1}$. D'où, nous avons $f(Q \mid d_{syn}, B_m, \bar{U}_m) = N(\bar{q}_m, \bar{U}_m + B_m / m + (1 - n_2 / n_1)\, \bar{u}_m / m)$.

Pour obtenir $f(Q \mid d_{syn})$, nous devons éliminer par intégration B_m et $\bar{U}_m$ de cette distribution. Pour cela, nous substituons à B_m et à $\bar{U}_m$ leurs valeurs prévues approximatives. Pour approximer $E(B_m \mid d_{syn})$, nous utilisons $b_m - (1 - n_2 / n_1)\, \bar{u}_m$. Pour approximer $E(\bar{U}_m \mid d_{syn})$, nous notons que

$$\begin{aligned} \mathrm{Var}(Q \mid d_i) &= E[\mathrm{var}(Q \mid D_i) \mid d_i] + \mathrm{var}[E(Q \mid D_i) \mid d_i] \\ &= E(U_i \mid d_i) + \mathrm{var}(Q_i \mid d_i). \end{aligned} \quad (6)$$

Ici, $\mathrm{var}(Q \mid d_i) = (1 - n_2 / N) u_i$. En résolvant (6), nous obtenons $E(\bar{U}_m \mid d_{syn}) \approx (n_2 / n_1 - n_2 / N)\, \bar{u}_m$. Après substitution de ces valeurs prévues, nous avons $\mathrm{var}(Q \mid d_{syn}) = T_d$.

Puisque nous utilisons une variance estimée pour Q, nous approximons $f(Q \mid d_{syn})$ par une loi t de moyenne $\bar{q}_m$ et de variance T_d. Le nombre de degrés de liberté, ν_d, est dérivé en faisant concorder les deux premiers moments de $(\nu_d T_d) / \{(n_2 / n_1 - n_2 / N)\, \bar{u}_m + B_m / m + (1 - n_2 / n_1)\, \bar{u}_m / m\}$ avec ceux d'une loi $\chi^2_{\nu_d}$.

2.2 Diffusion du même sous-échantillon aléatoire

À première vue, diffuser un même ensemble d'enregistrements dans chaque sous-échantillon ressemble à la synthèse partielle classique. Cependant, l'estimateur de variance de Reiter (2003) peut présenter un biais positif dans ce contexte. Pour l'illustrer, supposons que D ne contient qu'une seule variable dont la moyenne d'échantillon est $\bar{x}_1$. Supposons aussi que nous créons D_{syn} en remplaçant toutes les valeurs de x, et que nous sélectionnons aléatoirement un ensemble commun de n_2 enregistrements pour le sous-échantillon. Soit $m = \infty$, et soit Q la moyenne de population de x. Si les remplacements sont simulés d'après le modèle correct, qui est estimé en se servant de D, alors $\bar{q}_\infty = \bar{x}_1$. D'où, $\mathrm{var}(\bar{q}_\infty)$ est identique à la variance de $\bar{x}_1$, qui est $(1 - n_1 / N)\, s_1^2 / n_1$. Toutefois, l'estimation de variance de Reiter (2003) comprend $\bar{u}_m$ basé sur $(1 - n_2 / N)\, s_2^2 / n_2$, où $E(s_2^2) = s_1^2$. Donc, en général, la variance de Reiter (2003) aura un biais positif pour les sous-échantillons avec synthèse.

Au lieu de la synthèse partielle classique, nous adoptons l'approche suivie par Reiter (2008) sous imputation multiple pour remplacer les données manquantes quand les enregistrements qui servent pour l'imputation ne sont ni utilisés ni diffusés pour l'analyse. Ces conditions ressemblent au tirage des mêmes enregistrements dans chaque sous-échantillon d_i parce que les modèles utilisés pour la synthèse sont estimés en se servant de D, mais l'analyste ne dispose que de d_{syn} pour l'analyse ; autrement dit, les enregistrements utilisés pour l'imputation ne sont pas tous diffusés pour l'analyse.

Par souci de commodité, nous résumons ici la méthodologie de Reiter (2008) sans inclure les dérivations mathématiques. Premièrement, comme dans la synthèse partielle classique, l'INS estime les modèles de synthèse en utilisant uniquement les enregistrements dont les valeurs seront synthétisées. Soit θ les paramètres qui régissent la distribution des modèles de données synthétiques. Deuxièmement, l'INS tire un échantillon de m valeurs de θ de la distribution *a posteriori* de ce dernier. Troisièmement, pour chaque tirage $\theta^{(l)}$ où $l = 1, \ldots, m$, l'INS tire un ensemble de données de remplacement $D^{(l,\, p)}$ des modèles de synthèse fondés sur $\theta^{(l)}$. L'INS répète ce processus r fois pour chaque $\theta^{(l)}$. Enfin, l'INS diffuse la série de $M = mr$ sous-échantillons provenant de ces ensembles de données,

$d^* = \{d^{(l,\,p)} : l = 1, \ldots, m;\ p = 1, \ldots, r\}$. Chaque $d^{(l,\,p)}$ comprend un indice l désignant son emboîtement (*nest*).

Pour $l = 1, \ldots, m$ et $p = 1, \ldots, r$, soit $q^{(l,\,p)}$ et $u^{(l,\,p)}$ l'estimation de Q et l'estimation de sa variance calculée en se servant de $d^{(l,\,p)}$. Ici, $u^{(l,\,p)}$ comprend le facteur de correction pour population finie. Les quantités qui suivent sont utilisées pour les inférences :

$$\bar{q}_M = \sum_{l=1}^{m}\sum_{p=1}^{r} q^{(l,\,p)}/(mr) = \sum_{l=1}^{m} \bar{q}_r^{(l)}/m, \qquad (7)$$

$$\bar{w}_M = \sum_{l=1}^{m}\sum_{p=1}^{r}(q^{(l,\,p)} - \bar{q}_r^{(l)})^2 / \{m(r-1)\} = \sum_{l=1}^{m} w_r^{(l)}/m, \quad (8)$$

$$b_M = \sum_{l=1}^{m}(\bar{q}_r^{(l)} - \bar{q}_M)^2 / (m-1), \qquad (9)$$

$$\bar{u}_M = \sum_{l=1}^{m}\sum_{p=1}^{r} u^{(l,\,p)}/(mr). \qquad (10)$$

L'analyste peut se servir de $\bar{q}_M$ pour estimer Q et de $T_s = \bar{u}_M - \bar{w}_M + (1+1/m)\, b_M - \bar{w}_M / r$ pour estimer la variance de $\bar{q}_M$. Quand r est grand, les inférences sont fondées sur une loi t, $(\bar{q}_M - Q) \sim t_{\nu_s}(0, T_s)$, dont le nombre de degrés de liberté est

$$\nu_s = \left(\frac{\{(1+1/m)b_M\}^2}{(m-1)T_s^2} + \frac{\{(1+1/r)\bar{w}_M\}^2}{\{m(r-1)\}T_s^2} \right)^{-1}. \quad (11)$$

Il se peut que $T_s < 0$, particulièrement si m et r sont petits. L'analyste peut utiliser à la place l'estimateur de variance toujours positif, mais prudent, $T_s^* = \lambda T_s + (1-\lambda)(1+1/m)b_M$, où $\lambda = 1$ quand $T_s > 0$ et $\lambda = 0$, autrement. Les facteurs qui motivent l'emploi de cet estimateur sont donnés dans Reiter (2008). En général, il est possible d'éviter les valeurs négatives de T_s en donnant une valeur élevée à m et à r. Quand $T_s < 0$, les inférences sont fondées sur une loi t à $(m-1)$ degrés de liberté qui découle de l'utilisation du premier terme seulement et de T_s^* dans (11).

Pour les plans stratifiés, l'estimation ponctuelle des quantités pour l'ensemble de la population est $\bar{q}_M = \sum_h (N_h/N)\, \bar{q}_{Mh}$, et sa variance estimée est $T_s = \sum_h (N_h/N)^2 T_{sh}$, où $\bar{q}_{Mh}$ et T_{sh} sont l'estimation ponctuelle et sa variance dans la strate h. Le nombre de degrés de liberté de la loi t pour l'échantillonnage stratifié est égal à

$$\nu_{st} = \left\{ \frac{\left\{\sum_h ((N_h/N)^2 (1+1/m) b_{Mh})\right\}^2}{(m-1)\sum_h (N_h/N)^2 T_{sh}^2} + \frac{\left\{\sum_h ((N_h/N)^2 (1+1/k)\bar{w}_{Mh})\right\}^2}{\{m(r-1)\}\sum_h (N_h/N)^2 T_{sh}^2} \right\}^{-1}. \quad (12)$$

Il est dérivé par appariement des moments à ceux d'une variable aléatoire χ^2.

3. Simulations illustratives sous un plan d'échantillonnage stratifié

À la présente section, nous examinons les propriétés analytiques des procédures d'inférence proposées pour le sous-échantillonnage avec synthèse dans le cas de l'échantillonnage aléatoire simple stratifié. Nous générons une population de $N = 1\,000\,000$ enregistrements comprenant cinq variables, $Y_1, \ldots, Y_5$, dans $H = 4$ strates. Y_1 est une variable catégorique comprenant dix catégories générées conformément à la distribution du tableau 1. Les distributions pour $(Y_2, \ldots, Y_5)$ sont présentées au tableau 2, de même que les tailles de strates.

Tableau 1
Distribution empirique de Y_1 dans la population générale

	1	2	3	4	5	6	7	8	9	10
Pourcentage	24,77	32,63	16,38	15,06	7,13	2,53	0,95	0,33	0,15	0,09

Pour créer D, nous tirons aléatoirement $n_{1h} = 7\,500$ enregistrements dans chaque strate. Chaque sous-échantillon comprend $n_{2h} = 5\,000$ enregistrements pour chaque strate. En pratique, l'INS pourrait recourir à la répartition proportionnelle pour fixer chaque n_{1h} et choisir des taux d'échantillonnage plus faibles pour fixer n_{2h}. Nous utilisons une taille d'échantillon commune et de grandes fractions d'échantillonnage pour illustrer le fait que les formules de variance pour le sous-échantillonnage avec synthèse traite correctement les facteurs de correction pour population finie non négligeables, par exemple 50 % des enregistrements sont échantillonnés dans la strate 4.

Nous considérons que Y_4 et Y_5 sont des variables confidentielles et illustrons deux scénarios de synthèse. Dans le premier, nous synthétisons les valeurs de Y_4 et Y_5 de tous les enregistrements. Pour cela, dans chaque strate, nous simulons Y_{4h} en utilisant une régression de Y_{4h} sur (Y_{1h}, Y_{2h}, Y_{3h}) estimée en nous servant de D, et nous simulons Y_{5h} en utilisant une régression de Y_{5h} sur $(Y_{1h}, Y_{2h}, Y_{3h}, Y_{4h})$ estimée en nous servant de D. Les prédictions de Y_{5h} sont fondées sur les valeurs synthétisées de Y_{4h}. Dans la deuxième approche, dans chaque strate, nous ne remplaçons Y_{4h} et Y_{5h} que dans les enregistrements pour lesquels $Y_{3h} > p_h$, où p_h est le 90e centile de Y_3 dans la population dans la strate h. Nous générons des valeurs de remplacement par échantillonnage à partir de modèles de régression ; cependant, dans chaque strate, nous estimons les modèles en nous servant uniquement des enregistrements qui satisfont $Y_{3h} > p_h$.

Dans le cas de l'approche fondée sur des sous-échantillons différents, nous générons $m = 5$ enquêtes synthétiques comme il est décrit à la section 2.1. Pour l'approche fondée sur le même sous-échantillon, nous commençons par tirer $m = 5$ valeurs de θ, les coefficients de régression et les variances. Pour chaque $\theta^{(l)}$, nous générons $r = 5$ ensembles de données synthétiques pour chaque emboîtement (*nest*) de premier degré.

Pour tous les scénarios, nous répétons le processus consistant à i) créer D par échantillonnage à partir de la population et ii) générer des sous-échantillons avec synthèse 5 000 fois en tout. Pour chacune de ces 5 000 exécutions, nous obtenons des inférences pour 50 paramètres, y compris les moyennes de population et les moyennes de strates de Y_4 et Y_5, les coefficients d'une régression de Y_3 sur toutes les autres variables, et les coefficients d'une régression de Y_5 sur toutes les autres variables. Les régressions sont estimées séparément dans chaque strate.

La figure 1 donne les principaux résultats des simulations. Le graphique de gauche présente les ratios de la moyenne simulée de T_d (et de T_s) à la var($\overline{q}_m$) simulée correspondante pour les 50 paramètres estimés. Les ratios médians sont proches de 1 dans tous les scénarios, et les moyennes de T_d (et de T_s) ne diffèrent jamais de plus de 10 % des variances réelles correspondantes. Donc, T_s et T_d semblent être tous deux des estimateurs de variance approximativement valides.

Le graphique du milieu de la figure 1 résume les pourcentages des 5 000 intervalles de confiance à 95 % synthétiques basés sur T_d (et sur T_s) qui couvrent leur grandeur Q correspondante. Les taux de couverture sont proches de 0,95, excepté pour les coefficients de régression sous l'approche fondée sur le même sous-échantillon avec synthèse totale. Pour ces coefficients, $T_s < 0$ dans une proportion allant jusqu'à 38 % des exécutions de la simulation, de sorte que les intervalles de confiance sont fondés sur l'estimateur T_s^* prudent. La fraction plus élevée de variances négatives survient dans la strate la plus petite dont le taux d'échantillonnage est de 50 %. Toutes les estimations de variance sont positives quand 10 % seulement des enregistrements sont synthétisés.

Tableau 2
Paramètres pour le tirage de $(Y_2, \ldots, Y_5)$ pour la population

	Taille de strate	Modèle	Distribution du terme d'erreur
Strate 1	750 000	$Y_2 = Y_1 + e$	$e \sim N(0, 5)$
		$Y_3 = Y_1 + Y_2 + e$	
		$Y_4 = Y_1 + Y_2 + Y_3 + e$	
		$Y_5 = Y_1 + Y_2 + Y_3 + Y_4 + e$	
Strate 2	200 000	$Y_2 = 2Y_1 + e$	$e \sim N(0, 10)$
		$Y_3 = 2Y_1 + 0,5Y_2 + e$	
		$Y_4 = 2Y_1 + 0,5Y_2 + Y_3 + e$	
		$Y_5 = 2Y_1 + 0,5Y_2 + 0,5Y_3 - 0,25Y_4 + e$	
Strate 3	40 000	$Y_2 = -3Y_1 + e$	$e \sim N(0, 30)$
		$Y_3 = -3Y_1 - 1,5Y_2 + e$	
		$Y_4 = -3Y_1 + Y_2 - 1 / 3Y_3 + e$	
		$Y_5 = -3Y_1 + Y_2 - 1 / 3Y_3 + 1 / 9Y_4 + e$	
Strate 4	10 000	$Y_2 = -2Y_1 + e$	$e \sim N(0, 20)$
		$Y_3 = -Y_1 - 1,5Y_2 + e$	
		$Y_4 = -2Y_1 + Y_2 + 1 / 4Y_3 + e$	
		$Y_5 = 2Y_1 - Y_2 - 1 / 4Y_3 + 1 / 16Y_4 + e$	

Le graphique de droite de la figure 1 donne les ratios de la racine carrée de l'erreur quadratique moyenne (REQM) simulée de $\overline{q}_m$ sur les REQM simulées pour les sous-échantillons sans aucune synthèse. Pour l'approche fondée sur le même sous-échantillon, les REQM des sous-échantillons synthétiques ont tendance à être plus faibles que celles calculées pour les sous-échantillons sans aucune synthèse, particulièrement dans le cas de la synthèse totale. Les REQM plus faibles sont dues au fait que les modèles synthétiques sont déterminés en se servant de D, c'est-à-dire les données d'enquête avant de tirer le sous-échantillon, de sorte qu'ils contiennent des renseignements supplémentaires qui ne figurent pas dans les sous-échantillons sans synthèse. Pour les divers sous-échantillons synthétiques, les ratios des REQM sont habituellement supérieurs à 1. Ici, l'accroissement de la synthèse entraîne une plus grande perte d'efficacité. Il convient de souligner qu'à la figure 1, les REQM pour l'approche fondée sur différents échantillons et celle fondée sur le même échantillon ne sont pas directement comparables, parce qu'elles sont basées sur des dénominateurs différents.

Afin de permettre de comparer les méthodes, et pour illustrer les pertes d'efficacité dues au sous-échantillonnage, nous répétons le plan de simulation en utilisant $m = 25$ pour l'approche des sous-échantillons indépendants et $mr = 25$ pour l'approche des mêmes sous-échantillons. Le graphique de gauche de la figure 2 montre les ratios des REQM simulées pour les 50 paramètres estimés dans les divers scénarios, où les dénominateurs sont les REQM moyennes fondées sur les données originales avant toute protection de la confidentialité. Le graphique de droite de la figure 2 donne les ratios des longueurs moyennes simulées des intervalles de confiance à 95 %, où les dénominateurs sont les longueurs moyennes basées sur les données originales avant toute protection de la confidentialité. Si l'on s'en tient au graphique de gauche, pour un nombre total donné d'ensembles de données diffusés et un pourcentage de synthèse donné, l'approche des échantillons indépendants produit des estimations plus efficaces que celle fondée sur le même échantillon. Le graphique de droite mène à des conclusions semblables, bien que cela soit plus difficile à voir à cause de la mise à l'échelle. Ici, l'approche fondée sur le même échantillon avec une synthèse totale produit des fractions importantes d'estimations de variance négatives, de sorte que la variance corrigée T_s^* est souvent utilisée, ce qui agrandit les longueurs des intervalles. La figure 2 comprend aussi les résultats pour la synthèse sans aucun sous-échantillonnage, qui fournit généralement des estimations plus efficaces que l'une et l'autre approche de sous-échantillonnage.

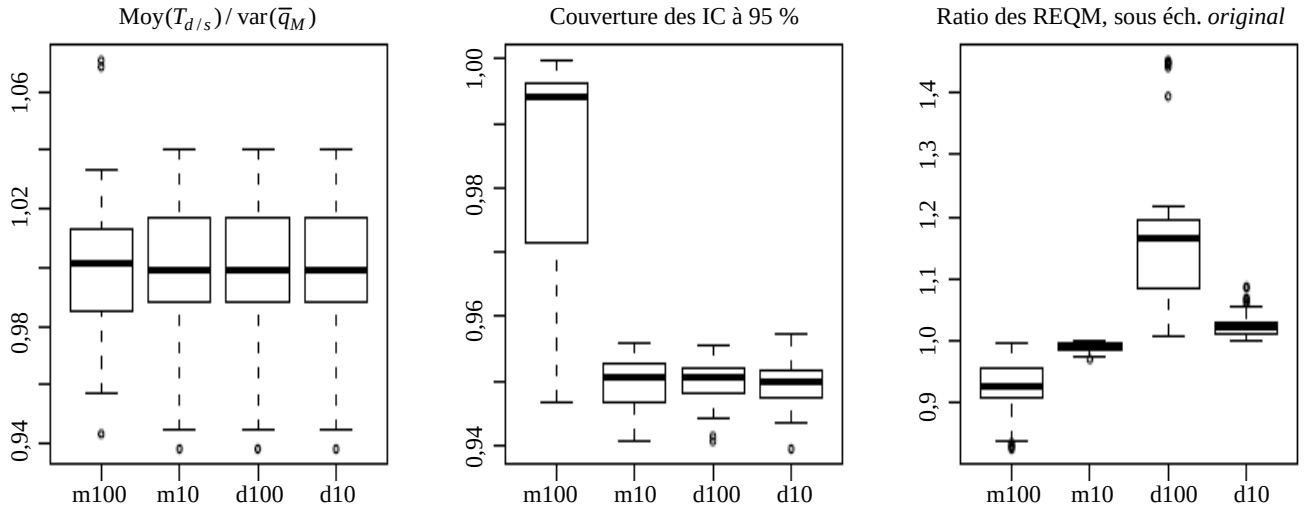


Figure 1 Résultats des simulations pour le plan d'échantillonnage stratifié. Dans les étiquettes, m et d indiquent l'approche fondée sur le même sous-échantillon et celle fondée sur des sous-échantillons différents. Les chiffres donnent le pourcentage d'enregistrements qui sont synthétisés. Les dénominateurs de la REQM sont fondés sur les estimations ponctuelles d'après les sous-échantillons sans synthèse. Pour l'approche des sous-échantillons différents, la REQM est calculée en se servant de la moyenne des *m* estimations ponctuelles. Chaque boîte à moustaches comprend 50 paramètres estimés

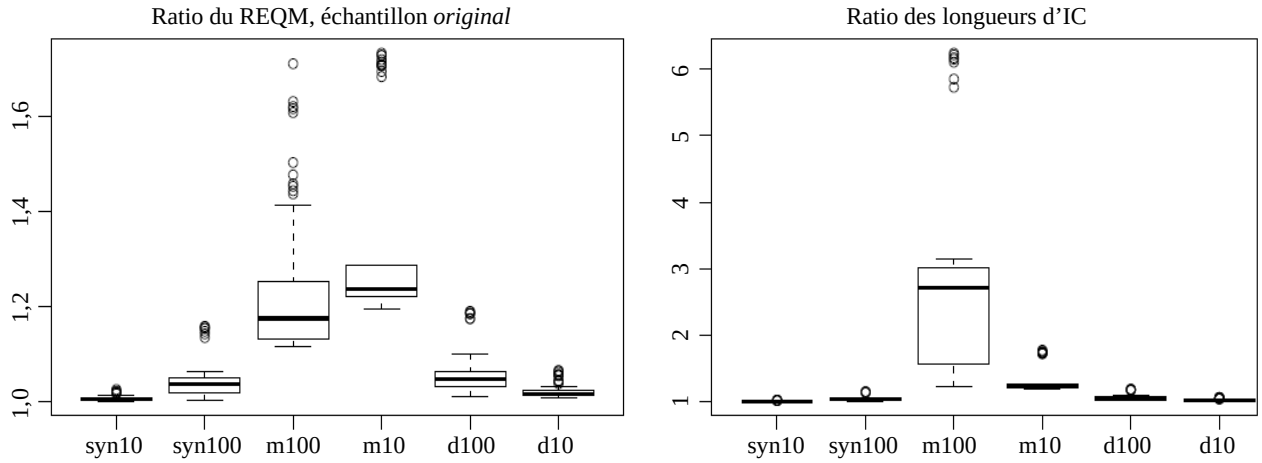


Figure 2 Comparaisons d'efficacité pour le plan d'échantillonnage stratifié. Dans les étiquettes, *original* et *syn* font référence à l'échantillon original et à l'échantillon synthétique avant le sous-échantillonnage, et m, d et les chiffres ont la même signification qu'à la figure 1. Les dénominateurs de la REQM sont basés sur les estimations ponctuelles d'après l'échantillon original sans synthèse. Chaque boîte à moustaches englobe 50 paramètres estimés

4. Conclusion

L'approche des sous-échantillons différents et celle des mêmes sous-échantillons ont des avantages concurrents. Pour un nombre fixe d'ensembles de données diffusés M, l'approche des sous-échantillons différents permet une estimation plus efficace que l'approche des mêmes sous-échantillons, comme en témoigne la figure 2, puisque les sous-échantillons diffusés sont indépendants plutôt que corrélés. L'approche des sous-échantillons différents garantit aussi des estimations de variance positives, contrairement à l'approche fondée sur le même échantillon. Cependant, si M est grand, l'approche des sous-échantillons différents affaiblit la protection de la confidentialité découlant du sous-échantillonnage, puisque les ensembles de données combinés contiennent vraisemblablement la plupart des enregistrements issus de l'enquête originale. Donc, à moins que le taux de sous-échantillonnage soit faible (par exemple 1 % ou 2 %), l'INS peut être obligé de choisir une valeur de m modeste (par exemple, $m = 5$) pour utiliser l'approche des sous-échantillons différents. Par conséquent, cette approche n'est pas viable quand la taille de l'échantillon original est modeste.

Au lieu de procéder au sous-échantillonnage avec synthèse, les organismes statistiques pourraient diffuser des données partiellement synthétiques qui comprennent tous les enregistrements provenant de l'échantillon original, pour autant qu'ils soient prêts à diffuser des fichiers de cette taille. La synthèse partielle sur les données originales produit généralement des estimations dont la variance est plus faible que le sous-échantillonnage avec synthèse, comme le montre la figure 2, puisque qu'un plus grand nombre d'enregistrements sont diffusés. Cependant, la synthèse partielle sur les données originales engendre généralement de plus grands risques de divulgation que le sous-échantillonnage avec synthèse, puisqu'un plus grand nombre d'enregistrements à risque figurent parmi les données diffusées et que la protection supplémentaire due au sous-échantillonnage fait défaut. Les organismes statistiques peuvent comparer les deux options en fonction des risques de divulgation en se servant des méthodes de Drechsler et Reiter (2008), qui tiennent compte de la protection résultant de l'échantillonnage, ainsi qu'en fonction de l'utilité des données en comparant les inférences pour les analyses représentatives.

Il se peut également que le processus de sous-échantillonnage engendre une protection supplémentaire suffisante pour permettre une quantité plus faible de synthèse qu'il ne serait nécessaire dans le cas d'une synthèse partielle de l'ensemble de données originales complètes. L'évaluation de l'utilité des données pour le sous-échantillonnage avec synthèse par opposition à la synthèse seulement pour des risques de divulgation donnés dépasse le cadre de ce bref exposé, mais il s'agit d'un domaine intéressant dans lequel poursuivre de futures études.

Nous n'avons pas élaboré d'approches de sous-échantillonnage avec synthèse pour d'autres plans de sondage que l'échantillonnage aléatoire simple (stratifié). Pour l'approche des sous-échantillons différents, les méthodes d'inférence appropriées nécessitent une estimation approximativement sans biais de la variance provenant de la première phase d'échantillonnage qui peut être calculée d'après le sous-échantillon seulement. Sous des plans de sondage

compliqués, cette estimation est difficile à obtenir. Dans le cas de l'approche fondée sur le même sous-échantillon, nous conjecturons que les analystes peuvent utiliser les méthodes d'inférence présentées à la section 2.2, pourvu que $\overline{u}_M$ tienne compte comme il convient des deux phases d'échantillonnage. Nous notons que les formules pour $\overline{w}_M$ et b_M restent les mêmes pour d'autres plans de sondage. L'évaluation de cette conjecture sera le sujet de futures recherches.

Remerciements

Ces travaux ont été financés par une bourse du U.S. National Science Foundation (SES-0751671).

Bibliographie

Drechsler, J., et Reiter, J.P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. Dans *Privacy in Statistical Databases*, (Éds., J. Domingo-Ferrer et Y. Saygin), New York : Springer, 227-238.

Drechsler, J., et Reiter, J.P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *Journal of Official Statistics*, 25, 589-603.

Drechsler, J., et Reiter, J.P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105, 1347-1357.

Elliott, M., et Purdam, K. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records. *Environment and Planning A*, 39, 1101-1118.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Reiter, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 203-211.

Reiter, J.P. (2004). Utilisation simultanée de l'imputation multiple pour les données manquantes et le contrôle de la divulgation. *Techniques d'enquête*, 30, 263-271.

Reiter, J.P. (2005). Estimating identification risks in microdata. *Journal of the American Statistical Association*, 100, 1103-1113.

Reiter, J.P. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*, 95, 933-946.

Willenborg, L., et de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York : Springer-Verlag.

Winkler, W.E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Rapport technique, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 89-102
Statistique Canada, N° 12-001-X au catalogue

Un modèle hiérarchique bayésien de non-réponse pour les données catégoriques d'un tableau à double entrée provenant de petits domaines avec incertitude au sujet de l'ignorabilité

Balgobin Nandram et Myron Katzoff [1]

Résumé

Nous étudions le problème de la non-réponse non ignorable dans un tableau de contingence bidimensionnel qui peut être créé individuellement pour plusieurs petits domaines en présence de non-réponse partielle ainsi que totale. En général, le fait de prendre en considération les deux types de non-réponse dans les données sur les petits domaines accroît considérablement la complexité de l'estimation des paramètres du modèle. Dans le présent article, nous conceptualisons le tableau complet des données pour chaque domaine comme étant constitué d'un tableau contenant les données complètes et de trois tableaux supplémentaires pour les données de ligne manquantes, les données de colonne manquantes et les données de ligne et de colonne manquantes, respectivement. Dans des conditions de non-réponse non ignorable, les probabilités totales de cellule peuvent varier en fonction du domaine, de la cellule et de ces trois types de « données manquantes ». Les probabilités de cellule sous-jacentes (c'est-à-dire celles qui s'appliqueraient s'il était toujours possible d'obtenir une classification complète) sont produites pour chaque domaine à partir d'une loi commune et leur similarité entre les domaines est quantifiée paramétriquement. Notre approche est une extension de l'approche de sélection sous non-réponse non ignorable étudiée par Nandram et Choi (2002a, b) pour les données binaires ; cette extension crée une complexité supplémentaire qui découle de la nature multivariée des données et de la structure des petits domaines. Comme dans les travaux antérieurs, nous utilisons un modèle d'extension centré sur un modèle de non-réponse ignorable de sorte que la probabilité totale de cellule dépend de la catégorie qui représente la réponse. Notre étude s'appuie sur des modèles hiérarchiques bayésiens et des méthodes Monte Carlo par chaîne de Markov pour l'inférence a posteriori. Nous nous servons de données provenant de la troisième édition de la National Health and Nutrition Examination Survey pour illustrer les modèles et les méthodes.

Mots-clés : Échantillonneur de Metropolis-Hastings ; algorithme SIR ; modèle de non-réponse non ignorable ; modèle d'extension.

1. Introduction

Généralement, les données des enquêtes par sondage sont résumées dans des tableaux de contingence à double entrée. Nous considérons le problème de la non-réponse non ignorable pour un grand nombre de tableaux de contingence de dimensions $r \times c$, pour chacun des domaines spécifiques. Dans nombre de ces enquêtes, des données manquent, si bien que la classification des individus échantillonnés n'est que partielle. Chaque tableau à double entrée présente donc à la fois des cas de non-réponse partielle (données manquantes pour l'une des deux catégories) et des cas de non-réponse totale (données manquantes pour les deux catégories). Comme on ne sait pas nécessairement de quelle façon les données manquent, il peut être souhaitable de privilégier un modèle dans lequel existe une certaine différence entre les données observées et les données manquantes (c'est-à-dire que les données manquantes ne sont pas ignorables). Pour un tableau de contingence $r \times c$ général, nous abordons la question de l'estimation des probabilités de cellule des tableaux à double entrée lorsque la non-réponse est peut-être non ignorable, mais que l'on ne dispose vraiment d'aucune information au sujet de l'ignorabilité. Dans de telles conditions, nous aimerions exprimer le degré d'incertitude au sujet de l'ignorabilité. Nandram et Choi (2002a, b) ont décrit un modèle d'extension approprié pour les données binaires lorsqu'il existe des données provenant de nombreux petits domaines. Nous étendrons ces travaux aux tableaux de contingence $r \times c$.

En désignant par x les covariables et par y la variable réponse, Little et Rubin (2002) décrivent trois types de mécanismes de création de données manquantes. Ils diffèrent selon que la probabilité de réponse a) est indépendante de x et de y ; b) dépend de x, mais non de y, ou c) dépend de y et éventuellement de x. Les données manquent entièrement au hasard (MCAR, *missing completely at random*) sous (a), manquent au hasard (MAR, *missing at random*) sous (b) et ne manquent pas au hasard (MNAR, *missing not at random*) sous (c). Les modèles pour les mécanismes de création de données manquantes MCAR et MAR sont dits ignorables si les paramètres de la variable dépendante du modèle et ceux de la variable réponse sont distincts (Rubin 1976). Les modèles pour les mécanismes de création des données manquantes de type MNAR sont dits

1. Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institue, 100 Institute Road, Worcester, MA 01609. Courriel : balnan@wpi.edu ; Myron Katzoff, Office of research and Methodology, National Center for Health Statisitcs, CDC, 3311 Toledo Road, Hyattsville, MD 20782. Courriel : mjk5@cdc.gov.

non ignorables. La difficulté générale que pose un modèle de non-réponse non ignorable tient au fait que les paramètres ne sont pas identifiables [par exemple, voir Nandram et Choi (2004, 2005, 2008, 2010), et Nandram, Han et Choi (2002)].

Pour un tableau de contingence $r \times c$, soit $I_{ijkl} = 1$ si le l^{e} individu dans le i^{e} domaine se trouve dans la j^{e} ligne et la k^{e} colonne, et 0 autrement. En outre, soit $J_{il} = 1$ si le l^{e} individu dans le i^{e} domaine a fourni des renseignements complets et 0 autrement. Enfin, soit $P(J_{il} = 1 \mid I_{ijkl} = 1, I_{ij'k'l} = 0, j' \neq j, k' \neq k) = \pi_{ijk}$. Pour la non-réponse totale (ou non-réponse d'une unité), si $\pi_{ijk} = \pi_i$, le modèle est ignorable ; pour la non-réponse partielle (non-réponse à certaines questions), si les valeurs de colonne manquent, mais que les valeurs de ligne sont observées et que $\pi_{ijk} = \pi_{ij}$ (ou $\pi_{ijk} = \pi_i$), le modèle est ignorable ; si les valeurs des lignes manquent, mais que les valeurs de colonne sont observées et que $\pi_{ijk} = \pi_{ik}$ (ou $\pi_{ijk} = \pi_i$), le modèle est ignorable. Tous les autres modèles sont non ignorables ; voir Rubin (1976) pour une explication plus détaillée.

Nandram et Choi (2002a, b) se servent d'un modèle d'extension pour étudier les données binaires en présence de non-réponse non ignorable. Le modèle d'extension, qui est un modèle de non-réponse non ignorable, dégénère en un modèle de non-réponse ignorable (dans l'esprit de Draper 1995) quand la valeur d'un paramètre de centrage est fixée à l'unité. Cela permet d'exprimer la certitude au sujet de l'ignorabilité ; voir également Forster et Smith (1998).

Nous discutons du modèle proposé par Nandram et Choi (2002a, b) pour des données binaires provenant de petits domaines. De sorte que J_{il} désigne les indicateurs de réponse et I_{il}, la réponse binaire. Spécifiquement, en introduisant les paramètres de centrage γ_i pour le domaine *i* afin d'intégrer l'incertitude au sujet de l'ignorabilité, le modèle de Nandram et Choi (2002a, b) est

$$
\begin{aligned}
I_{il} \mid p_i &\overset{\text{iid}}{\sim} \text{Bernoulli}\,(p_i), \\
J_{il} \mid \{\pi_i, J_{il} = 0\} &\overset{\text{iid}}{\sim} \text{Bernoulli}\,(\pi_i), l = 1, \ldots, n_i, i = 1, \ldots, l, \\
J_{il} \mid \{\pi_i, \gamma_i, y_{il} = 1\} &\overset{\text{iid}}{\sim} \text{Bernoulli}\,(\gamma_i \pi_i),\ 0 < \gamma_i \pi_i < 1.
\end{aligned}
$$

Quand $\gamma_i = 1$, le modèle de non-réponse non ignorable dégénère en un modèle de non-réponse ignorable. Ici, γ_i est le ratio des chances de succès parmi les répondants aux chances de succès parmi l'ensemble des individus pour le i^{e} domaine. Le paramètre γ_i décrit la portée de la non-ignorabilité du mécanisme de réponse pour le domaine *i*, et c'est donc grâce à ce paramètre γ_i qu'est intégrée l'incertitude au sujet de l'ignorabilité. Nandram et Choi (2002a, b) définissent $\delta_i = \pi_i \{\gamma_i p_i + (1 - p_i)\}$ comme étant la probabilité qu'un individu du domaine *i* réponde dans l'ensemble de la population et, en croyant que tous les domaines sont semblables, ils considèrent que les $(p_i,\ \delta_i, \gamma_i)$ suivent une loi commune. A priori, ils choisissent des lois bêta pour p_i et π_i, respectivement.

Ici, les paramètres ne sont pas identifiables. Cependant, quand $\gamma_i = 1$, ils le sont tous. Autrement dit, le caractère identifiable des paramètres dépend de γ_i. Notons que si $\gamma_i = 1$, nous obtenons un modèle ignorable pour un mécanisme MAR. Comme les paramètres de ce modèle sont identifiables, il est relativement logique de l'utiliser (ou des modèles similaires) comme modèle de référence. Il faut toutefois souligner que ce modèle n'est toujours pas justifié, parce qu'il repose sur l'hypothèse que les données manquantes ressemblent aux données observées. Donc, pour rendre ce modèle de non-réponse ignorable plus souple, nous utilisons le paramètre γ_i.

Soit γ_{iuv} le nombre d'individus pour lesquels $I_{il} = u$, $J_{il} = v (u, v = 0, 1)$ dans le i^{e} domaine. Alors, sous le modèle,

$$
\begin{aligned}
(y_{i00}, y_{i01}, y_{i10}, y_{i11}) \mid \pi_i, p_i, \gamma_i &\overset{\text{ind}}{\sim} \text{Multinomiale}\{n_i, \\
&(1 - p_i)(1 - \pi_i), (1 - p_i)\pi_i, (1 - \gamma_i \pi_i) p_i, \gamma_i \pi_i p_i\}
\end{aligned}
$$

avec indépendance sur les domaines. Ici, y_{i01} et y_{i11} seulement sont observés et, par conséquent, tous les paramètres sont non identifiables si les γ_i sont inconnus. Nous obtenons la fonction de vraisemblance de la même manière pour le tableau de contingence $r \times c$ plus complet avec données manquantes.

Nous partons d'une loi gamma et, pour permettre le centrage sur le modèle de non-réponse ignorable, nous devons sélectionner chaque γ_i de manière que sa moyenne soit égale à 1. Cependant, nous devons utiliser une loi gamma tronquée, parce que $0 < \pi_i < 1$ et $0 < \gamma_i \leq 1 / \pi_i$. Une idée intéressante de Nandram et Choi (2002a, b) consiste à modéliser le centrage sous forme d'une loi gamma tronquée

$$
\gamma_i \mid \nu \overset{\text{iid}}{\sim} \text{Gamma}(\nu, \nu), 0 < \gamma_i < 1 / \pi_i, 0 < \pi_i < 1.
$$

Le modèle est complet et possède des densités de probabilité a priori non informatives sur tous les hyperparamètres. D'autres distributions peuvent être choisies (par exemple une densité lognormale tronquée) pour les γ_i, mais il ne s'agit pas d'un problème essentiel et cela n'aurait pas beaucoup d'importance.

On peut se servir d'un modèle au niveau du domaine avec effets aléatoires dans lequel, conditionnellement aux données observées, la non-réponse dépend des effets aléatoires au niveau du domaine. Ce modèle peut être formulé en utilisant une fonction de lien logit, mais nous

n'avons pas suivi cette direction pour élaborer nos modèles, en partie parce que nous n'utilisons pas de covariables ici ; voir Nandram et Choi (2010) pour l'utilisation de covariables et d'effets aléatoires.

L'approche décrite dans Nandram et Choi (2002a, b) est intéressante, mais elle ne s'applique pas directement au problème des tableaux de contingence $r \times c$ qui nous occupe. Plus précisément, dans Nandram et Choi (2002a, b), un seul paramètre de centrage est nécessaire par domaine. Dans notre formulation, nous avons besoin de *rc* paramètres de centrage par domaine ; chacun de ces paramètres doit suivre une loi centrée sur l'unité pour permettre la dégénérescence en modèle de non-réponse ignorable. Des contraintes d'inégalité doivent également être incluses dans le modèle de non-réponse non ignorable. En outre, on ne peut écarter la possibilité que ces paramètres soient corrélés. La méthodologie nécessaire pour appliquer les travaux de Nandram et Choi (2002a, b) au tableau de contingence $r \times c$ n'est pas simple. Constatant ces difficultés, Nandram, Liu, Choi et Cox (2005) (avec un seul tableau supplémentaire) et Nandram, Cox et Choi (2005) (avec les trois tableaux supplémentaires) appliquent une idée plus simple, mais moins élégante que dans Nandram et Choi (2002a, b) pour procéder au centrage ; voir aussi Nandram et Choi (2005).

Essentiellement, Nandram, Cox et Choi (2005), ainsi que Nandram, Liu, Cox et Choi (2005) émettent l'hypothèse d'un modèle ignorable, obtiennent des échantillons des probabilités de réponse et se servent de ces probabilités de réponse échantillonnées pour ajuster les probabilités de réponse d'un modèle de non-réponse non ignorable en « contrôlant » ce paramètre. Naturellement, une alternative est possible lorsqu'il existe de l'information sur le degré de non-ignorabilité. Toutefois, l'intégration d'information a priori au sujet d'un écart systématique par rapport à l'ignorabilité est plus complexe dans le cas de notre problème et elle nécessiterait du travail sur le terrain supplémentaire coûteux afin d'obtenir cette information.

Nous discutons maintenant de notre conception du problème de non-réponse non ignorable, fondamentalement un problème de distorsion. En fait, ce problème est extrêmement difficile et nous pensons qu'il n'a vraiment aucune solution, mais que nous devons essayer d'en trouver une. Sans aucune information, il n'est pas possible de dire quelles sont les différences entre les répondants et les non-répondants. Un modèle de non-réponse ignorable est restreint, parce qu'il suppose que les répondants et les non-répondants sont semblables, alors qu'ils peuvent être différents. Les statisticiens doivent non seulement faire face à l'imprécision (erreur d'échantillonnage), mais aussi être suffisamment audacieux pour étudier la subjectivité (l'ignorance découlant de l'information manquante). Malheureusement, il est bien connu que les modèles de non-réponse non ignorable contiennent des paramètres non identifiables. Nous discutons de la façon dont sont identifiés les paramètres clés de non-ignorabilité. Nous savons que si les répondants et les non-répondants sont semblables, les γ_i sont égaux à l'unité, et nous obtenons le modèle de non-réponse ignorable à l'aide de tous les paramètres identifiés. Nous pouvons maintenant étendre le modèle de non-réponse ignorable à un modèle de non-réponse non ignorable en donnant à ces paramètres γ_i une loi centrée à 1, tout en maintenant l'identifiabilité. Un modèle de non-réponse non ignorable peut être formulé pour ajouter de la souplesse au modèle de non-réponse ignorable, comme nous l'avons fait dans nos travaux ; la souplesse est une forme d'analyse de sensibilité, cohérente dans le cas qui nous occupe, et il s'agit en effet d'une évaluation bayésienne de l'incertitude (du risque) (par exemple Greenland 2009). C'est ce que nous avons fait ou essayé de faire dans nos travaux.

Dans le présent article, nous tentons de résoudre le problème difficile de Nandram et Choi (2002a, b) sous sa forme originale pour les tableaux $r \times c$ pour de nombreux domaines. Le plan de l'exposé est le suivant. À la section 2, nous décrivons le modèle hiérarchique bayésien. Précisément, nous décrivons le mécanisme de non-réponse non ignorable et nous construisons une loi a priori propre. À la section 3, nous montrons comment ajuster le modèle en utilisant l'algorithme d'échantillonnage avec rééchantillonnage par importance (SIR, pour *sampling importance resampling*) pour effectuer un sous-échantillonnage à partir d'une densité de probabilité a posteriori approximative après une agrégation innovatrice de la densité a posteriori conjointe complète. À la section 4, nous illustrons notre méthodologie en nous servant de données à grande diffusion recueillies dans 13 États dans le cadre de la troisième édition de la National Health and Nutrition Examination Survey (NHANES III). À la section 5, nous présentons nos conclusions.

2. Le modèle de non-réponse non ignorable

Dans le contexte du problème de non-réponse dans un tableau à double entrée, on peut avoir affaire aussi bien à la non-réponse partielle qu'à la non-réponse totale. Donc, on peut considérer la totalité du tableau de données comme étant constitué de quatre tableaux, à savoir un tableau pour les données complètes et trois tableaux supplémentaires, pour l'information de ligne manquante, pour l'information de colonne manquante et pour l'information de ligne ainsi que de colonne manquante, respectivement. Dans tous l'exposé, nous donnons aux lignes l'indice $j = 1, \ldots, r$, aux colonnes l'indice $k = 1, \ldots, c$, et aux quatre tableaux l'indice $s = 1, 2, 3, 4$. Nous donnons aux domaines l'indice

$i = 1, 2, \ldots, A$ et aux individus dans les domaines, l'indice $l = 1, 2, \ldots, n_i$. Nous allons maintenant décrire le modèle de non-réponse non ignorable (c'est-à-dire le modèle avec extension).

2.1 Processus d'échantillonnage

Nous adaptons la terminologie et les définitions utilisées dans Nandram, Cox et Choi (2005) à la situation considérée ici. Pour l'individu *l* échantillonné dans le domaine *i*, soit

$$I_{ijkl} = \begin{cases} 1, & \text{si la catégorie de résultat est } (j, k) \\ 0, & \text{autrement,} \end{cases}$$

et soit $\boldsymbol{J}_{il}$ l'un des quadruplets $(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)$. Nous supposons que

$$\boldsymbol{I}_{il} \overset{\text{def}}{=} \text{vec}(\{I_{ijkl} \mid j = 1, \ldots, r;\ k = 1, \ldots, c\}) \mid \boldsymbol{p}_i \overset{\text{iid}}{\sim} \text{Mult}\{1, \boldsymbol{p}_i\} \qquad (1)$$

et

$$\boldsymbol{J}_{il} \mid \{I_{ijkl} = 1,\ I_{ij'k'l} = 0 \text{ pour tout } j' \neq j \text{ et } k' \neq k \mid \boldsymbol{\pi}_{ijk}\} \overset{\text{iid}}{\sim} \text{Mult}\{1, \boldsymbol{\pi}_{ijk}\}, \qquad (2)$$

où $\boldsymbol{p}_i \overset{\text{def}}{=} \text{vec}(\{p_{ijk} \mid j = 1, 2, \ldots, r; k = 1, 2, \ldots, c\})$ est un vecteur de probabilités pour le tableau de *rc* catégories pour la variable observée dont la somme doit être égale à l'unité, et pour la cellule (j, k) dans ce tableau bidimensionnel,

$$\boldsymbol{\pi}_{ijk} \overset{\text{def}}{=} \text{vec}(\{\pi_{isjk}\} \text{ pour } s = 1,2,3,4)$$

est un vecteur de probabilités dont la somme doit être égale à l'unité.

Ensuite, définissons les fréquences de cellule y_{isjk}, pour chaque tableau $s = 1, \ldots, 4$ pour le domaine *i*, telles que, pour la cellule (j, k),

$$(y_{i1\,jk}, y_{i2\,jk}, y_{i3\,jk}, y_{i4\,jk}) = \sum_{l=1}^{n_i} I_{ijkl}\, \boldsymbol{J}_{il},$$

où les $y_{i1\,jk}$ sont observées et les y_{isjk}, pour $s = 2, 3, 4$, sont des variables latentes qui satisfont les contraintes observées $\sum_k y_{i2\,jk} = u_{ij}, \sum_j y_{i3\,jk} = v_{ik}$ et $\sum_{j,k} y_{i4\,jk} = w_i$. Toutes les inférences seront faites conditionnellement aux quantités observées, u_{ij}, v_{ik} et w_i. Voir Nandram (2009) pour l'analyse d'un seul tableau $r \times c$ sous non-réponse quand les marges sont également aléatoires. Nous désignerons le vecteur des $y_{i1\,jk}$ par $\boldsymbol{y}_1$, le vecteur des y_{isjk}, $s = 2, 3, 4$, par $\boldsymbol{y}_{(1)}$, et le vecteur complet par $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_{(1)})'$.

Les paramètres π_{isjk} ne sont pas identifiables. Si les distributions de ces paramètres sont entièrement connues, la non-identifiabilité disparaît. Donc, le problème essentiel est la façon d'identifier ces paramètres. Nous savons que, si les répondants et les non-répondants sont semblables (c'est-à-dire en ce qui concerne les quatre tableaux, complet et partiellement complets) nous pouvons prendre $\pi_{isjk} = \pi_{is}$, ce qui est le modèle de non-réponse ignorable. Les π_{is} peuvent être estimées par les proportions de cas qui, pour chaque domaine, se retrouvent dans chacun des quatre tableaux. Il s'agit d'un point de départ naturel. Afin d'étendre le modèle de non-réponse ignorable à un modèle de non-réponse non ignorable tout en maintenant l'identifiabilité, nous devons d'abord procéder à une simplification. Nous prenons $\pi_{ijks} = \psi_{ijk}\ \pi_{is}$, qui donne un modèle de non-réponse non ignorable dans lequel les paramètres ψ_{ijk} ne sont pas identifiables.

Pour centrer le modèle non ignorable sur le modèle ignorable, nous prenons

$$\pi_{isjk} = \begin{cases} \tilde{\psi}_{ijk}\pi_{is}, & \text{pour } s = 1, \\ \psi_{ijk}\pi_{is}, & \text{pour } s = 2,3,4, \end{cases} \qquad (3)$$

et exigeons que $\sum_{s=1}^{4} \pi_{is} = 1$. Quelques opérations algébriques donnent alors la relation

$$\begin{aligned} \tilde{\psi}_{ijk}\pi_{i1} &= \left[1 + (1 - \psi_{ijk})\left(\frac{1 - \pi_{i1}}{\pi_{i1}}\right)\right]\pi_{i1} \\ &= a_{ijk}(\pi_{i1}, \psi_{ijk})\,\psi_{ijk}\pi_{i1}, \end{aligned} \qquad (4)$$

où $a_{ijk}(\pi_{i1}, \psi_{ijk}) = \{\psi_{ijk}^{-1} + (\psi_{ijk}^{-1} - 1)(\pi_{i1}^{-1} - 1)\}$, dont il découle clairement que $\tilde{\psi}_{ijk} = 1$ si, et seulement si, $\psi_{ijk} = 1$. Soulignons que, puisque $0 \le \pi_{isjk} \le 1$ et $(1 - \pi_{i1})^{-1} \le \min\{\pi_{is}^{-1}; s = 2,3,4\}$, il s'ensuit que $0 < \psi_{ijk} \le (1 - \pi_{i1})^{-1}$.

En combinant (1) et (2), et en notant la définition de π_{isjk} dans (3), comme dans le cas binaire, nous obtenons une distribution multinomiale pour $\boldsymbol{y}$ sachant $\boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{p}$, et nous voyons maintenant que la fonction de vraisemblance pour l'échantillon est

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{p}) &= \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \left\{ \left[\prod_{j,k} (\tilde{\psi}_{ijk}\pi_{i1}p_{ijk})^{y_{i1\,jk}} \right] \right. \\ &\qquad \left. \prod_{s=2}^{4} \prod_{j,k} (\psi_{ijk}\pi_{is}p_{ijk})^{y_{isjk}} \right\} \\ &= \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \left\{ \prod_{s=1}^{4} \prod_{j,k} (\psi_{ijk}\pi_{is}p_{ijk})^{y_{isjk}} \right. \\ &\qquad \left. \prod_{j,k} \left[a_{ijk}(\pi_{i1}, \psi_{ijk}) \right]^{y_{i1\,jk}} \right\}, \end{aligned} \qquad (5)$$

où

$$\boldsymbol{y}_{is}^{rc\times 1} = \text{vec}(\{y_{isjk} \mid j=1,\ldots,r; k=1,\ldots,c\}),$$
$$\boldsymbol{y} = (\boldsymbol{y}'_{11}, \boldsymbol{y}'_{12}, \boldsymbol{y}'_{13}, \boldsymbol{y}'_{14}, \boldsymbol{y}'_{21}, \ldots, \boldsymbol{y}'_{24}, \ldots, \boldsymbol{y}'_{A1}, \boldsymbol{y}'_{A2}, \boldsymbol{y}'_{A3}, \boldsymbol{y}'_{A4})',$$
$$\boldsymbol{\pi}^{A4\times 1} = (\pi_{11}, \ldots, \pi_{14}, \pi_{21}, \ldots, \pi_{24}, \ldots, \pi_{A1}, \ldots, \pi_{A4})',$$
$$\boldsymbol{\psi}^{Arc\times 1} = (\psi_{111}, \ldots, \psi_{1rc}, \psi_{211}, \ldots \psi_{2rc}, \ldots, \psi_{A11}, \ldots, \psi_{Arc}),$$

et

$$\boldsymbol{p}^{Arc\times 1} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_A)'.$$

En obtenant des facteurs qui sont des puissances de π_{is}, la fonction de vraisemblance peut également être exprimée sous la forme

$$f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{p}) = \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \left\{ \prod_{s=1}^{4} \pi_{is}^{y_{is..}} \times \prod_{j,k} \{p_{ijk}\ \psi_{ijk}\}^{y_{i.jk}} \left[a_{ijk}(\pi_{i1}, \psi_{ijk}) \right]^{y_{i1jk}} \right\}, \quad (6)$$

où $0 \le \pi_{is} \le 1, \sum_s \pi_{is} = 1$ et $0 \le \psi_{ijk} \le (1-\pi_{i1})^{-1}$. Ici, nous notons que $y_{is..}$ et y_{i1jk} sont des variables observées, mais que les $y_{i.jk}$ sont des variables latentes.

2.2 Construction des lois a priori

Les hypothèses qui suivent décrivent les lois a priori pour le modèle de non-réponse non ignorable :

1. Pour le vecteur de probabilités de cellule $\boldsymbol{p}_i$, nous supposons que

$$\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1 \tau_1),$$

où $\mu_1 = (\mu_{111}, \mu_{112}, \ldots, \mu_{11k}, \mu_{121}, \ldots, \mu_{1rc})'$; $\mu_{1jk} \ge 0$ et $\sum_{j=1}^{r}\sum_{k=1}^{c} \mu_{1jk} = 1$. Le paramètre τ_1 nous informe de la similarité entre les $\boldsymbol{p}_i$: plus la valeur de τ_1 est grande, plus les $\boldsymbol{p}_i$ se ressemblent. Il en est ainsi parce qu'une grande valeur de τ_1 signifie que les variances des $\boldsymbol{p}_i$ sont faibles, et comme elles ont la même moyenne, cela signifie qu'elles sont plus semblables quand τ_1 est grand.

Donc, pour $\boldsymbol{p}$, la densité de probabilité est

$$g_1(\boldsymbol{p} \mid \boldsymbol{\mu}_1, \tau_1) = \prod_{i=1}^{A} g_{1i}(\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1) = \prod_{i=1}^{A} \left\{ \frac{\prod_{j,k} p_{ijk}^{\mu_{1jk}\tau_1 - 1}}{\text{D}(\boldsymbol{\mu}_1 \tau_1)} \right\}, \quad (7)$$

où, pour un k-uplet $\boldsymbol{c}$ et un scalaire t

$$\text{D}(\boldsymbol{ct}) = \frac{\prod_{j=1}^{k} \Gamma(\boldsymbol{c}_j t)}{\Gamma(t)}$$

pour $c_j > 0$ et $\sum_{j=1}^{k} c_j = 1$.

2. Indépendamment des $\boldsymbol{p}_i$, les $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})'$ suivent la spécification

$$\boldsymbol{\pi}_i \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2 \tau_2),$$

avec $\pi_{is} \ge 0$ et $\sum_s \pi_{is} = 1$, où $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22}, \mu_{23}, \mu_{24})', \mu_{2s} \ge 0, \sum_{s=1}^{4} \mu_{2s} = 1$ et τ_2 est une mesure de la similarité entre les $\boldsymbol{\pi}_i$. Donc, pour $\boldsymbol{\pi}_i$, , la densité de probabilité est

$$g_{2i}(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2) = \frac{\prod_{s=1}^{4} \pi_{is}^{\mu_{2s}\tau_2 - 1}}{\text{D}(\boldsymbol{\mu}_2 \tau_2)}. \quad (8)$$

3. Pour chaque i, soit $\boldsymbol{\psi}_i = (\psi_{i11}, \ldots, \psi_{i1c}, \psi_{i21}, \ldots, \psi_{i2c}, \ldots, \psi_{irc})'$ de sorte que $\boldsymbol{\psi} = (\boldsymbol{\psi}'_1, \ldots, \boldsymbol{\psi}'_A)'$. Nous supposons que, pour chaque i, les ψ_{ijk} sont indépendants et identiquement distribués suivant une distribution dérivée de la loi Gamma(β, β), où le support est limité à l'intervalle ouvert $(0, (1-\pi_{i1})^{-1})$; autrement dit, la loi gamma ordinaire est tronquée comme il suit

$$\psi_{ijk} \mid \beta, \boldsymbol{\pi}_i \overset{\text{ind}}{\sim} \text{Gamma}(\beta, \beta)$$

de sorte que $0 < \psi_{ijk} < (1-\pi_{i1})^{-1}$.

Il convient de souligner que ces ψ_{ijk} sont identiquement distribués sur j et sur k. De nouveau, nous pourrions utiliser d'autres distributions, telles que la densité lognormale tronquée, mais cela ne change pas grand-chose. Dans cette formulation, il existe une certaine information au sujet de β parce que nous supposons que les petits domaines partagent un effet commun.

Donc, pour le domaine i, la densité de probabilité pour $\boldsymbol{\psi}_i$ est

$$g_{3i}(\boldsymbol{\psi}_i \mid \beta, \boldsymbol{\pi}_i) = \prod_{j=1}^{r}\prod_{k=1}^{c} \left\{ \frac{\beta^\beta \psi_{ijk}^{\beta-1} e^{-\beta\psi_{ijk}}}{\Gamma(\beta)} \Big/ \int_0^{(1-\pi_{i1})^{-1}} \frac{\beta^\beta \psi_{ijk}^{\beta-1} e^{-\beta\psi_{ijk}}}{\Gamma(\beta)} \text{d}\psi_{ijk} \right\},$$

pour $0 < \psi_{ijk} < (1-\pi_{i1})^{-1}$. En procédant à la transformation $t_{ijk} = \beta\psi_{ijk}$, nous voyons que la constante de normalisation dans le dénominateur de chacun des facteurs qui figurent dans $g_{3i}(\boldsymbol{\psi}_i \mid \beta, \boldsymbol{\pi}_i)$ est $G_\beta[\beta(1-\pi_{i1})^{-1}]$, où $G_\beta(\cdot)$ est la fonction gamma avec paramètre d'échelle β. Pour que l'intervalle d'intégration ne dépende pas de π_{i1}, posons que $\phi_{ijk} = (1-\pi_{i1})\psi_{ijk}$ et que $\boldsymbol{\phi}_i = (\phi_{i11}, \ldots, \phi_{i1c}, \phi_{i21}, \ldots, \phi_{i2c}, \ldots, \phi_{irc})'$. Alors

$$g_{3i}(\boldsymbol{\phi}_i \mid \beta, \boldsymbol{\pi}_i) = \prod_{j=1}^{r}\prod_{k=1}^{c}\left\{ \frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)\, G_{\beta}[\beta\,(1-\pi_{i1})^{-1}]} \right\}, \quad (9)$$

pour $0 < \phi_{ijk} < 1$. La loi priori conjointe de $\boldsymbol{\pi}_i$ et $\boldsymbol{\phi}_i$ est simplement le produit de $g_{3i}(\boldsymbol{\phi}_i \mid \beta, \boldsymbol{\pi}_i)$ et de $g_{2i}(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2)$. Donc, la loi a priori conjointe de $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_A)'$ et de $\boldsymbol{\pi}$ est

$$g^*(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{\mu}_2, \tau_2, \beta) \overset{\text{def}}{=} \prod_{i=1}^{A} \left\{ g_{3i}(\boldsymbol{\phi}_i \mid \boldsymbol{\beta}, \boldsymbol{\pi}_i) \cdot g_{2i}(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2) \right\}.$$

Autrement dit

$$g^*(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{\mu}_2, \tau_2, \beta) = \prod_{i=1}^{A} \left\{ \frac{\prod_{s=1}^{4} \pi_{is}^{\mu_{2s}\tau_2 - 1}}{\mathrm{D}(\boldsymbol{\mu}_2 \tau_2)} \prod_{j=1}^{r}\prod_{k=1}^{c} \frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)\, G_{\beta}[\beta(1-\pi_{i1})^{-1}]} \right\}. \quad (10)$$

Pour achever la description du modèle, nous spécifions les hypothèses concernant les hyperparamètres. Comme il n'existe pas de lois a priori conjuguées, nous utilisons des lois a priori de rétrécissement pour τ_1, τ_2 et β, parce qu'elles sont propres et non informatives. Les lois a priori de la forme $p(\tau_1) \propto 1/\tau_1$, en particulier les lois a priori gamma diffuses propres, sont déconseillées ; voir, par exemple, Gelman (2006). Des demi-densités de la loi de Cauchy et des densités de probabilité de la loi gamma sont d'autres options (pour lesquelles il faudrait spécifier les hyperparamètres). Donc, nous prenons

1. τ_1, τ_2 et β ayant des lois a priori de rétrécissement indépendantes de la forme

   $$f(x) = \frac{a_0}{(a_0 + x)^2}, \text{ pour } x \geq 0,$$

   où a_0 est spécifié ; il est courant en pratique de prendre $a_0 = 1$.

2. Nous supposons aussi que $\boldsymbol{\mu}_1 \sim \text{Dirichlet}(1, 1, \ldots, 1)$ et $\boldsymbol{\mu}_2 \sim \text{Dirichlet}(1, 1, 1, 1)$.

Soit $\Omega = (\beta, \boldsymbol{\mu}_1, \tau_1, \boldsymbol{\mu}_2, \tau_2)$. La densité de probabilité pour Ω est alors

$$p(\Omega) = \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2} (rc - 1)!\, 3!$$

pour τ_1, τ_2 et $\beta \geq 0$, $\sum_{j,k} \mu_{1jk} = 1$ et $\sum_{s=1}^{4} \mu_{2s} = 1$.

En vertu du théorème de Bayes, la densité de probabilité a posteriori conjointe est

$$\begin{aligned} h(\Omega, \boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \propto\ & f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{p})\, g_1(\boldsymbol{p} \mid \boldsymbol{\mu}_1, \tau_1)\, g^*(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{\mu}_2, \tau_2, \beta)\, p(\Omega) \\ =\ & \prod_{i=1}^{A} \left[\binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} (1-\pi_{i1})^{-y_{i..}} \right. \\ & \left\{ \prod_{s=1}^{4} \pi_{is}^{y_{is.}} \times \prod_{j,k} \{p_{ijk}\phi_{ijk}\}^{y_{i.jk}} \left[a_{ijk}(\pi_{i1}, \phi_{ijk}) \right]^{y_{i1jk}} \right\} \\ & \times \left\{ \frac{\prod_{j,k} p_{ijk}^{\mu_{1jk}\tau_1 - 1}}{\mathrm{D}(\boldsymbol{\mu}_1 \tau_1)} \right\} \left\{ \frac{\prod_{s=1}^{4} \pi_{is}^{\mu_{2s}\tau_2 - 1}}{\mathrm{D}(\boldsymbol{\mu}_2 \tau_2)} \right\} \\ & \left. \prod_{j=1}^{r}\prod_{k=1}^{c} \frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)\, G_{\beta}[\beta(1-\pi_{i1})^{-1}]} \right] \\ & \times \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2}, \end{aligned} \quad (11)$$

où, en substituant $(1-\pi_{i1})^{-1}\phi_{ijk}$ pour ψ_{ijk},

$$a_{ijk}(\pi_{i1}, \phi_{ijk}) = \left(\frac{1-\pi_{i1}}{\phi_{ijk}} \right) \left[1 + \frac{1}{\pi_{i1}} \left\{ 1 - \pi_{i1} - \phi_{ijk} \right\} \right]. \quad (12)$$

Pour faire des inférences au sujet des p_{ijk}, nous tirerons des échantillons de $h(\Omega, \boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ selon des méthodes Monte Carlo par chaîne de Markov. Cette procédure est décrite à la section 3.

3. Calculs

Nous utilisons l'algorithme SIR pour sous-échantillonner un échantillon aléatoire tiré d'une densité de probabilité a posteriori approximative. L'exécution de cette tâche se fait en trois étapes. Nous agrégeons sur les $\boldsymbol{p}_i, \boldsymbol{\pi}_i$ et $\boldsymbol{\phi}_i$, approximons la densité de probabilité agrégée par une densité plus simple et effectuons l'échantillonnage à partir de cette densité, puis nous sous-échantillonnons ces échantillons pour obtenir des échantillons tirés de la densité de probabilité originale. Nous montrons à la présente section comment exécuter ces trois étapes.

Pour obtenir l'approximation et pour simplifier les calculs, à l'annexe A, nous procédons à l'agrégation sur les $\boldsymbol{p}_i, \boldsymbol{\pi}_i$ et $\boldsymbol{\phi}_i$ pour obtenir

$$h(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \pi_a(\Omega\, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \cdot \prod_{i=1}^{A} I_i,$$

où

$$I_i = \iiint_0^{\frac{\beta b_i}{1-\pi_{i1}}} \frac{G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)}{\left[G_\beta\left(\frac{\beta}{1-\pi_{i1}}\right)\right]^{rc}}$$
$$\prod_{j,k}\left\{\left(\frac{W_i}{\beta}\right)\sum_{s=2}^{4} y_{isjk}\left[\frac{1}{\phi^*_{ijk}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{W_i\phi^*_{ijk}}{\beta}\right\}\right)\right]^{y_{i1jk}}\right\}$$
$$\frac{W_i^{rc\beta-1}\,e^{-W_i}}{\Gamma(rc\beta)G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)}\,\mathrm{d}W_i\left\{\frac{\prod_{j,k}\phi^{*\ y_{i\cdot jk}+\beta-1}_{ijk}}{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\boldsymbol{j})}\right\}$$
$$\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot}+\mu_2 s\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}\right\}\mathrm{d}\boldsymbol{\phi}^*_i\,\mathrm{d}\boldsymbol{\pi}_i, \qquad (13)$$

avec $b_i = \min\{\{1/\phi^*_{ijk}\},\ j=1,\ldots,r;\ k=1,\ldots,c\}$ et

$$\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$$
$$= \frac{a_0}{(a_0+\tau_1)^2}\cdot\frac{b_0}{(b_0+\tau_2)^2}\cdot\frac{c_0}{(c_0+\beta)^2}$$
$$\prod_{i=1}^{A}\frac{\Gamma(rc\beta)}{[\Gamma(\beta)]^{rc}}\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\boldsymbol{j})$$
$$\times\prod_{i=1}^{A}\binom{n_i}{\boldsymbol{y}'_{i1},\boldsymbol{y}'_{i2},\boldsymbol{y}'_{i3},\boldsymbol{y}'_{i4}}\frac{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\boldsymbol{\mu}_1\tau_1)}{\mathrm{D}(\boldsymbol{\mu}_1\tau_1)}\frac{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}{\mathrm{D}(\boldsymbol{\mu}_2\tau_2)}. \qquad (14)$$

Pour évaluer I_i pour chaque $i=1,\ldots,A$, nous procédons comme il suit sachant $(\Omega, \boldsymbol{y}_{i(1)})$:

1. Tirer des échantillons indépendants des vecteurs $\boldsymbol{\pi}_i$ et $\boldsymbol{\phi}^*_i$ des distributions Dirichlet$(\boldsymbol{y}_i^{(2)} + \boldsymbol{\mu}_2\tau_2)$ et Dirichlet$(\boldsymbol{y}_i^{(1)} + \beta\,\boldsymbol{j})$, respectivement. Pour chaque $\boldsymbol{\pi}_i$ et $\boldsymbol{\phi}^*_i$, tirer un échantillon des valeurs de W_i de la distribution gamma tronquée sur l'intervalle $(0, \{\beta b_i/1-\pi_{i1}\})$ avec le paramètre $rc\beta$.
2. Pour chaque $\boldsymbol{\pi}_i, \boldsymbol{\phi}^*_i$ et W_i sélectionné à l'étape (1), calculer R_1R_2, où

$$R_1 = G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)\Bigg/\left[G_\beta\left(\frac{\beta}{1-\pi_{i1}}\right)\right]^{rc} \qquad (15)$$

et

$$R_2 = \prod_{j,k}\left\{\left(\frac{W_i}{\beta}\right)\sum_{s=2}^{4} y_{isjk}\left[\frac{1}{\phi^*_{ijk}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{W_i\phi^*_{ijk}}{\beta}\right\}\right)\right]^{y_{i1jk}}\right\}. \qquad (16)$$

3. Répéter les étapes (1) et (2) 1 000 fois. Puis, calculer la moyenne de R_1R_2 sur ces 1 000 valeurs.

La suite des calculs comporte deux parties. Premièrement, nous utilisons l'échantillonneur de Metropolis-Hastings « griddy » (technique d'approximation par des grilles) pour tirer des échantillons de $\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$. Nous échantillonnons μ_1, μ_2, τ_1 et τ_2 à partir de leur densité de probabilité a posteriori conditionnelle en utilisant des grilles ; cela comporte la transformation de τ_1 et τ_2 pour les faire varier dans l'intervalle unitaire $(0, 1)$. Pour chaque distribution, nous utilisons 100 grilles ; voir Nandram, Cox et Choi (2005) pour une procédure similaire. Ici, le tirage de $\boldsymbol{y}_{(1)}$ est effectué par échantillonnage de la fonction de masse de probabilité conditionnelle par composante. Des échantillons sont tirés de la densité a posteriori conditionnelle de β en utilisant un pas de Metropolis de manière similaire à Nandram et Choi (2002a, b). Nous avons exécuté cet algorithme 11 000 fois en permettant un « rodage » de 1 000 itérations. Nous avons constaté que les autocorrélations entre les itérations étaient faibles, ce qui indiquait que l'échantillonneur produisait un bon mélange. Nous avons également utilisé la méthode des moyennes de lot pour étendre l'évaluation des calculs. Nous avons utilisé des lots de 25 pour calculer les erreurs-types numériques.

Deuxièmement, nous nous servons de l'algorithme SIR pour sous-échantillonner l'échantillon de 10 000 itérations tirées de $\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$. Pour chacune des 10 000 itérations, nous calculons les poids

$$w_m = \frac{h(\Omega^{(m)}, \boldsymbol{y}_{(1)}^{(m)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}{\pi_a(\Omega^{(m)}, \boldsymbol{y}_{(1)}^{(m)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})},$$
$$m = 1, \ldots, M = 10\,000, \qquad (17)$$

et nous rééchantillonnons $\{\Omega^{(m)}, \boldsymbol{y}_{(1)}^{(m)}\}$ avec probabilités proportionnelles aux poids w_m pour $m=1,\ldots,M$, sans remise. Nous utilisons un échantillonnage à 10 % et nous sous-échantillonnons les 10 000 itérations pour en obtenir 1 000 ; l'échantillonnage sans remise est une bonne idée, parce qu'il permet d'éviter les valeurs répétées qui existent déjà parce que l'échantillonneur de Metropolis-Hastings n'est pas vraiment un échantillonneur de type acceptation-rejet et qu'il donne des valeurs répétées. Comme d'habitude sous échantillonnage sans remise, les poids sont calculés chaque fois qu'une valeur est sélectionnée.

Enfin, nous pouvons maintenant faire une inférence exacte (dans les limites des méthodes Monte Carlo par chaîne de Markov) au sujet de $\boldsymbol{p}_i$ a posteriori. Soit $y_{i\cdot jk} = \sum_{s=1}^{4} y_{isjk}$ et $\boldsymbol{y}_i^*$ le vecteur de $y_{i\cdot jk}$. Alors,

$$\boldsymbol{p}_i \mid \boldsymbol{y}_i^*, \boldsymbol{\mu}_1, \tau_1 \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{y}_i^* + \boldsymbol{\mu}_1\tau_1),\ i=1,\ldots,A.$$

Donc, pour chaque valeur de $\boldsymbol{y}_i^*$, $\boldsymbol{\mu}_1$ et τ_i que nous obtenons au moyen de l'algorithme SIR, nous tirons une valeur de $\boldsymbol{p}_i, i=1,\ldots,A$. D'où nous obtenons une densité Rao-blackwellisée pour chaque $\boldsymbol{p}_i$, et l'inférence se poursuit de la manière habituelle.

4. Un exemple

En guise d'illustration, nous choisissons un exemple dans le domaine de la statistique de la santé. À la section 4.1, nous décrivons brièvement les données provenant de la troisième édition de la National Health and Nutrition Examination Survey (NHANES III) que nous utilisons. En particulier, nous étudions la relation entre la densité minérale osseuse et le revenu familial ; voir Nandram, Cox et Choi (2005) pour une discussion de ce problème. À la section 4.2, après une brève discussion de nos calculs, nous présentons l'inférence a posteriori sur les probabilités de cellule. À la section 4.3, en nous servant du facteur de Bayes, nous discutons de la relation entre la densité minérale osseuse et le revenu familial.

4.1 Données de la NHANES III

Le plan de sondage est un plan probabiliste stratifié à plusieurs degrés qui est représentatif de l'ensemble de la population civile ne vivant pas en établissement, âgée de deux mois et plus, des États-Unis. Des renseignements plus détaillés sur le plan de sondage de la NHANES III sont disponibles ailleurs (National Center for Health Statistics 1992, 1994). La collecte des données de la NHANES III comporte deux volets : le premier comprend la sélection de l'échantillon et l'interview des membres des ménages échantillonnés en vue de recueillir des renseignements personnels et le second comprend l'examen physique des personnes interviewées dans un centre d'examen mobile (CEM). L'évaluation de la santé s'appuie sur un examen physique, des tests et des mesures faites par des techniciens, ainsi que des prélèvements pour l'analyse. L'échantillon a été sélectionné auprès des ménages de 81 unités primaires d'échantillonnage à travers les États-Unis continentaux d'octobre 1988 à septembre 1994. Les données finales retenues pour l'étude proviennent des 35 plus grandes unités primaires d'échantillonnage dont la population est égale ou supérieure à 500 000 habitants, et nous considérons 13 régions infranationales.

La non-réponse peut avoir lieu dans les volets interview et examen physique de l'enquête. La non-réponse à l'interview se produit lorsque les personnes échantillonnées ne participent pas à l'interview. Certaines personnes interviewées et incluses dans le sous-échantillon pour l'évaluation de la santé ont manqué l'examen physique à la maison ou au centre d'examen mobile et n'ont donc pas subi la totalité ou une partie des examens.

Les médecins pensent que les personnes obèses ou ayant un excès de poids ne se présentent généralement pas au CEM. Cohen et Duffy (2002) remarquent que les enquêtes sur la santé sont un bon exemple de situation où il paraît plausible qu'il existe un lien entre la propension à répondre et l'état de santé. La NHANES III est en effet un bon exemple.

Les personnes échantillonnées pour participer à la NHANES III peuvent être classées en fonction d'un grand nombre d'attributs, et les chercheurs analysent ces tableaux de contingence afin de déterminer la qualité de l'ajustement des modèles ou l'indépendance. Ici, nous étudions la densité minérale osseuse (DMO) et le revenu familial (RF). Mentionnons ici que, même si le RF est une variable discrète, nous avons classé la DMO en trois catégories (normale, ostéopénie et ostéoporose) et le RF en trois catégories (faible, moyen et élevé). Cependant, nous ne disposons que d'une classification partielle des individus, parce que certains sont classés en fonction d'un seul attribut, tandis que d'autres ne sont pas classés du tout. Parmi les ménages qui ont participé au volet de l'examen physique, environ 62 % ont fourni des données sur le RF et la DMO, 8 % ont fourni des données sur la DMO seulement, 29 % ont fourni des données sur le revenu seulement et 1 % n'ont fourni de données ni sur le revenu ni sur la DMO. Notre problème consiste à estimer les probabilités de cellule et de tester l'association entre la DMO et le RF pour chacune des 13 régions infranationales en utilisant notre modèle d'extension qui regroupe les données de manière adaptative.

Dans le tableau 1, nous présentons les tableaux 3×3 de la DMO et du RF pour les 13 régions susmentionnées. Notons que les données pour les régions 6 et 48 sont suffisantes pour traiter ces régions individuellement. Par contre, les autres régions sont très petites. Les fréquences dans le tableau contenant des totaux de ligne sont généralement faibles, sauf pour la région 17, et les fréquences dans le tableau contenant juste le total sont faibles. Même pour le tableau contenant les données complètes, les fréquences de cellule sont généralement faibles, ce qui nous oblige à utiliser des techniques d'estimation sur petits domaines pour emprunter de l'information.

4.2 Inférence a posteriori des probabilités de cellule

Nous discutons de la performance de nos calculs pour le modèle d'extension, puis de l'inférence a posteriori au sujet des probabilités de cellule. Nous utilisons la moyenne a posteriori (MP), l'écart-type a posteriori (ETP) et l'intervalle de crédibilité à 95 % pour chaque paramètre d'intérêt. Nous présentons aussi les erreurs-types numériques (ETN) pour évaluer la répétabilité de nos calculs.

Au tableau 2, nous présentons des résumés des distributions a posteriori de $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \tau_1, \tau_2$ et β, avant et après l'application de l'algorithme SIR. Ces résumés sont fort semblables, ce qui indique que l'approximation $\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ donnée par l'algorithme SIR n'est pas déraisonnable. Par exemple, les intervalles de crédibilité à 95 % obtenus pour β avant et après l'application de l'algorithme SIR sont (1,081 ; 1,940) et (1,086 ; 1,947),

respectivement, ce qui représente une bonne concordance. Les estimations de τ_1 et τ_2 devraient présenter les écarts les plus importants, mais elles sont également raisonnablement proches [par exemple pour τ_1, l'intervalle de crédibilité à 95 % est (28,282 ; 64,204) avec l'approximation et (27,962 ; 64,425) avec l'algorithme SIR]. Dans les deux cas, les ETN sont faibles, ce qui signifie que les calculs sont répétables.

Au tableau 3, nous comparons notre modèle d'extension (modèle 3) à deux autres. Le modèle 1 (un modèle de non-réponse ignorable) et le modèle 2 (un modèle de non-réponse non ignorable) (pas de centrage) sont décrits à l'annexe B. Pour l'illustration, nous avons choisi trois régions, une grande, une moyenne et une petite. Des différences se dégagent entre les trois modèles. En général, les grandes estimations ont tendance à être plus petites pour le modèle 2 et encore plus petites pour le modèle 1 que pour le modèle 3 (c'est-à-dire que les estimations provenant du modèle 3 sont naturellement les plus proches du modèle 1 et non du modèle 2). Le modèle 2 produit la variabilité la plus importante ; comme prévu, le modèle 3 donne une variabilité légèrement plus grande que le modèle 1. Faute d'espace, nous ne présentons pas les ETM, mais nous constatons qu'elles sont toutes inférieures à 0,005.

Tableau 1
Fréquences des tableaux 3×3 de la DMO et du RF correspondant aux 13 régions infranationales dans la NHANES III

État	Tableau complet									Total de colonne			Total de ligne			Total
4	21	14	9	8	7	3	2	2	0	11	5	6	4	0	1	1
6	257	127	106	92	51	32	32	5	7	178	54	82	65	28	4	20
12	33	18	21	22	4	4	15	5	0	18	11	16	5	6	2	1
17	25	15	13	8	5	3	0	0	1	18	10	16	17	2	2	4
25	9	7	12	6	5	9	2	1	0	9	6	12	1	4	5	1
26	18	11	18	6	5	9	2	1	1	10	5	11	4	3	0	1
29	9	4	10	3	2	4	3	1	2	9	2	9	0	2	4	1
36	42	17	27	32	13	18	9	6	1	43	21	42	9	7	6	1
39	8	6	14	2	5	4	3	0	1	9	7	5	2	3	0	0
42	14	8	11	12	8	4	8	1	2	35	15	24	3	1	0	0
44	12	9	6	8	5	0	5	1	0	19	4	12	7	1	0	1
48	159	44	22	51	11	13	9	6	2	88	12	23	16	8	2	14
53	14	10	15	10	10	14	3	1	1	9	4	8	2	4	1	0

Nota : Dans le tableau 3×3 complet, le premier (deuxième, troisième) ensemble de trois nombres correspond à la première (deuxième, troisième) ligne ; le total de colonne (ligne) renvoie au tableau 3×3 ne contenant que les totaux de colonne (ligne) ; le total renvoie au tableau 3×3 contenant les totaux uniquement.

Tableau 2
Données de la NHANES pour 13 régions : Comparaison de la densité de probabilité a posteriori approximative et de la densité de probabilité a posteriori correcte en utilisant les moyennes a posteriori (MP), les écarts-types a posteriori (ETP), les erreurs-types numériques (ETN) et les intervalles de crédibilité à 95 % des hyperparamètres

	Approximation				Corrigée			
	MP	ETP	ETN	Int. à 95 %	MP	ETP	ETN	Int. à 95 %
μ_{21}	0,528	0,031	0,001	(0,463 ; 0,582)	0,525	0,031	0,008	(0,456 ; 0,578)
μ_{22}	0,131	0,021	0,001	(0,096 ; 0,181)	0,133	0,021	0,002	(0,094 ; 0,179)
μ_{23}	0,328	0,028	0,001	(0,274 ; 0,383)	0,328	0,028	0,005	(0,269 ; 0,383)
μ_{24}	0,013	0,006	0,000	(0,004 ; 0,027)	0,014	0,006	0,000	(0,004 ; 0,029)
τ_2	21,638	9,559	0,255	(8,347 ; 46,587)	20,078	8,632	0,303	(8,538 ; 38,625)
μ_{111}	0,280	0,023	0,001	(0,234 ; 0,324)	0,277	0,023	0,004	(0,228 ; 0,319)
μ_{112}	0,133	0,016	0,000	(0,102 ; 0,165)	0,134	0,017	0,002	(0,101 ; 0,165)
μ_{113}	0,200	0,019	0,000	(0,163 ; 0,238)	0,199	0,019	0,003	(0,162 ; 0,236)
μ_{121}	0,105	0,015	0,000	(0,078 ; 0,135)	0,107	0,015	0,002	(0,079 ; 0,135)
μ_{122}	0,065	0,011	0,000	(0,044 ; 0,088)	0,065	0,011	0,001	(0,044 ; 0,087)
μ_{123}	0,072	0,012	0,000	(0,050 ; 0,096)	0,073	0,012	0,001	(0,049 ; 0,097)
μ_{131}	0,061	0,011	0,000	(0,041 ; 0,083)	0,061	0,011	0,001	(0,040 ; 0,083)
μ_{132}	0,037	0,008	0,000	(0,023 ; 0,054)	0,036	0,008	0,001	(0,022 ; 0,054)
μ_{133}	0,048	0,009	0,000	(0,031 ; 0,068)	0,048	0,009	0,001	(0,031 ; 0,068)
τ_1	45,960	10,094	0,153	(28,282 ; 64,204)	45,177	10,562	0,679	(27,962 ; 64,423)
β	1,472	0,218	0,004	(1,081 ; 1,940)	1,449	0,208	0,022	(1,086 ; 1,947)

Nota : Les hyperparamètres sont $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \tau_1, \tau_2$ et β.

Tableau 3
Moyennes a posteriori des probabilités de cellule et des intervalles de crédibilité (IC) à 95 % pour trois régions (grande, moyenne et petite) selon les trois modèles

	Modèle 1			Modèle 2			Modèle 3		
Cellule	**MP**	**ETP**	**IC à 95 %**	**MP**	**ETP**	**IC à 95 %**	**MP**	**ETP**	**IC à 95 %**
a. Grande									
(1,1)	0,239	0,044	(0,157 ; 0,326)	0,196	0,046	(0,117 ; 0,295)	0,259	0,038	(0,189 ; 0,335)
(1,2)	0,140	0,035	(0,078 ; 0,213)	0,127	0,035	(0,068 ; 0,200)	0,132	0,029	(0,082 ; 0,197)
(1,3)	0,240	0,044	(0,159 ; 0,332)	0,198	0,047	(0,118 ; 0,301)	0,248	0,037	(0,175 ; 0,322)
(2,1)	0,092	0,032	(0,039 ; 0,162)	0,098	0,040	(0,037 ; 0,188)	0,077	0,022	(0,039 ; 0,126)
(2,2)	0,074	0,028	(0,029 ; 0,136)	0,077	0,030	(0,030 ; 0,144)	0,056	0,020	(0,024 ; 0,099)
(2,3)	0,133	0,036	(0,070 ; 0,210)	0,121	0,042	(0,056 ; 0,219)	0,110	0,028	(0,058 ; 0,168)
(3,1)	0,036	0,020	(0,008 ; 0,083)	0,069	0,039	(0,013 ; 0,153)	0,047	0,018	(0,018 ; 0,086)
(3,2)	0,023	0,015	(0,003 ; 0,061)	0,043	0,025	(0,007 ; 0,100)	0,032	0,014	(0,009 ; 0,063)
(3,3)	0,025	0,017	(0,003 ; 0,066)	0,071	0,040	(0,010 ; 0,154)	0,042	0,016	(0,016 ; 0,079)
b. Moyenne									
(1,1)	0,233	0,034	(0,169 ; 0,302)	0,213	0,043	(0.141 ; 0.305)	0,254	0,032	(0,194 ; 0,318)
(1,2)	0,143	0,028	(0,093 ; 0,200)	0,127	0,032	(0.072 ; 0.196)	0,146	0,024	(0,102 ; 0,197)
(1,3)	0,190	0,031	(0,132 ; 0,254)	0,140	0,034	(0.084 ; 0.218)	0,208	0,027	(0,156 ; 0,259)
(2,1)	0,174	0,031	(0,118 ; 0,237)	0,160	0,042	(0.092 ; 0.249)	0,154	0,027	(0,106 ; 0,211)
(2,2)	0,043	0,018	(0,015 ; 0,083)	0,060	0,028	(0.017 ; 0.124)	0,032	0,012	(0,012 ; 0,059)
(2,3)	0,049	0,020	(0,017 ; 0,095)	0,065	0,031	(0.018 ; 0.136)	0,042	0,014	(0,020 ; 0,072)
(3,1)	0,112	0,025	(0,068 ; 0,167)	0,120	0,041	(0.059 ; 0.209)	0,092	0,020	(0,056 ; 0,134)
(3,2)	0,047	0,018	(0,018 ; 0,088)	0,059	0,026	(0.019 ; 0.118)	0,040	0,014	(0,018 ; 0,071)
(3,3)	0,010	0,009	(0,000 ; 0,033)	0,056	0,032	(0.006 ; 0.122)	0,032	0,012	(0,013 ; 0,059)
c. Petite									
(1,1)	0,196	0,052	(0,103 ; 0,305)	0,164	0,055	(0.077 ; 0.288)	0,253	0,043	(0,175 ; 0,334)
(1,2)	0,081	0,034	(0,028 ; 0,158)	0,081	0,032	(0.030 ; 0.155)	0,091	0,028	(0,043 ; 0,152)
(1,3)	0,213	0,052	(0,118 ; 0,323)	0,175	0,055	(0.087 ; 0.300)	0,220	0,043	(0,137 ; 0,306)
(2,1)	0,093	0,041	(0,028 ; 0,186)	0,111	0,055	(0.029 ; 0.234)	0,073	0,028	(0,030 ; 0,139)
(2,2)	0,056	0,029	(0,012 ; 0,126)	0,066	0,031	(0.018 ; 0.136)	0,045	0,020	(0,014 ; 0,094)
(2,3)	0,115	0,045	(0,042 ; 0,215)	0,118	0,053	(0.038 ; 0.240)	0,092	0,030	(0,041 ; 0,158)
(3,1)	0,115	0,048	(0,036 ; 0,222)	0,113	0,056	(0.031 ; 0.239)	0,081	0,030	(0,033 ; 0,148)
(3,2)	0,044	0,028	(0,006 ; 0,113)	0,065	0,035	(0.013 ; 0.144)	0,043	0,020	(0,012 ; 0,086)
(3,3)	0,087	0,042	(0,022 ; 0,184)	0,107	0,055	(0.023 ; 0.227)	0,103	0,034	(0,047 ; 0,181)

Nota : Voir l'annexe B pour une description des modèles 1 et 2.

4.3 Facteur de Bayes pour la preuve d'association

Nous avons également considéré l'association entre la densité minérale osseuse et le revenu familial. Bien que l'existence d'une telle association paraisse peu probable, il est intéressant d'examiner cette question ; voir Nandram, Cox et Choi (2005) pour une discussion de ce problème. Nous nous servons du facteur de Bayes (Kass et Raftery 1995) pour mesurer la force de la preuve d'une association comparativement à l'absence d'association dans le tableau de contingence $r \times c$. Nous le faisons pour chacune des 13 régions et pour toutes les régions confondues.

Nous utilisons deux procédures, l'une sans modélisation étendue et l'autre s'appuyant sur notre modèle (étendu) de non-réponse non ignorable. La méthode simple consiste à produire les fréquences de cellule selon une technique de ratissage ordinaire, et nous supposons qu'il n'y a pas d'erreur à le faire. Il s'agit d'une procédure dictée par le bon sens que les praticiens des enquêtes utilisent régulièrement. Au moyen de la deuxième procédure fondée sur notre modèle de non-réponse non ignorable, nous avons obtenu 1 000 tableaux combinés pour chaque région, comme il est décrit à la section 3 sur les calculs. Pour chaque région, nous avons obtenu les fréquences de cellule pour les quatre tableaux, et nous les avons totalisées pour obtenir un seul tableau de toutes les fréquences.

Voici la description de la procédure de ratissage pour obtenir les fréquences de cellule. Soit n_{jk} les fréquences de cellule pour les quatre tableaux combinés. Soit $n_{jk}^{(1)}$ les fréquences de cellule pour le tableau contenant les données complètes, $n_{j,c+1}^{(2)}$ celles pour le tableau des totaux de ligne, $n_{r+1,k}^{(3)}$ celles pour le tableau des totaux de colonne et $n_{r+1,c+1}^{(4)}$ celles pour le tableau des totaux. Les fréquences de cellule pour les quatre tableaux sont estimées par

$$n_{jk} = n_{jk}^{(1)} + \left(\frac{n_{jk}^{(1)}}{n_{j\cdot}^{(1)}}\right) n_{j,c+1}^{(2)} + \left(\frac{n_{jk}^{(1)}}{n_{\cdot k}^{(1)}}\right) n_{r+1,k}^{(3)} + \left(\frac{n_{jk}^{(1)}}{n_{\cdot\cdot}^{(1)}}\right) n_{r+1,c+1}^{(4)},$$

$j = 1, \ldots, r, k = 1, \ldots, c.$

Dans chaque cas, nous désignons la somme des fréquences de cellule pour chaque région par n_{jk}. Pour la procédure de ratissage, nous n'avons qu'un seul tableau pour chaque région, tandis que pour le modèle de non-réponse

non ignorable, nous avons un échantillon de 1 000 tableaux pour chaque région. Nous avons aussi un seul tableau combiné pour toutes les régions sous la procédure de ratissage et 1 000 tableaux pour toutes les régions combinées. Nous obtenons le facteur de Bayes pour chaque tableau sous un modèle multinomial-Dirichlet. Il convient de souligner que notre méthode s'appuie sur le modèle d'extension de sorte que les fréquences de cellule sont produites en empruntant de l'information à d'autres régions contrairement à la procédure de ratissage.

Puis, pour chaque tableau, nous prenons

$$\boldsymbol{n}\,|\,\boldsymbol{\pi} \sim \text{Multinomiale}(n, \boldsymbol{\pi}) \text{ et } \boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{1}).$$

Autrement dit, nous prenons une loi a priori uniforme pour $\boldsymbol{\pi}$ avec $\pi_{jk} > 0$ et $\sum_{j=1}^{r}\sum_{k=1}^{c}\pi_{jk} = 1$. Sous l'hypothèse de l'absence d'association, nous avons $\pi_{jk} = \pi_j \pi_k$, où $\pi_j > 0, \sum_{j=1}^{r}\pi_j = 1$ et $\pi_k > 0, \sum_{k=1}^{c}\pi_k = 1$. Donc, l'hypothèse d'une association est que les π_{jk} ne sont soumises à aucune contrainte (à part le fait qu'elles sont non négatives et que leur somme est égale à l'unité), tandis que, pour l'hypothèse d'absence d'association, $\pi_{jk} = \pi_j \pi_k$.

Le facteur de Bayes est le ratio de la vraisemblance marginale sous association par rapport à la vraisemblance marginale en l'absence d'association. Ce ratio mesure la force de la preuve d'association par opposition à l'absence d'association ; voir Kass et Raftery (1995). Soit $p_a(\boldsymbol{n})$ la vraisemblance marginale sous association et $p_0(\boldsymbol{n})$ la vraisemblance marginale en l'absence d'association. Alors, en posant que $n_{j\cdot} = \sum_{k=1}^{c} n_{jk}$ et $n_{\cdot k} = \sum_{j=1}^{r} n_{jk}$, il est facile de montrer que

$$p_a(\boldsymbol{n}) = p_0(\boldsymbol{n})\left\{\prod_{u=0}^{n-1}\frac{u+rc}{(u+r)(u+c)}\right\}\frac{\prod_{j=1}^{r} n_{j\cdot}!\prod_{k=1}^{c} n_{\cdot k}!}{\prod_{j=1}^{r}\prod_{k=1}^{c} n_{jk}!},$$

où $p_0(\boldsymbol{n}) = n!\prod_{u=0}^{n-1}(u+rc)^{-1}$. Observons que $p_0(\boldsymbol{n})$ n'est pas une fonction de $\{n_{jk}\}$. Donc, en tant que mesure d'association, c'est l'écart de $\prod_{j=1}^{r} n_{j\cdot}!\prod_{k=1}^{c} n_{\cdot k}!$ par rapport à $\prod_{j=1}^{r}\prod_{k=1}^{c} n_{jk}!$ qui importe. Par contre, nous notons que, pour la statistique d'indépendance de Pearson classique, ce sont les écarts de n_{jk} par rapport à $n_{j\cdot}n_{\cdot k}$ qui importent. Cependant, soulignons que ce test ne peut pas être appliqué, parce que bon nombre de fréquences de cellule prévues sont inférieures à 5 sous l'hypothèse de l'absence d'association et l'échantillonnage multinomial

Nous présentons nos résultats au tableau 4 et à la figure 1 qui correspond aux données du tableau 1 pour la classification croisée de la densité minérale osseuse et du revenu familial. Nous présentons les logarithmes des vraisemblances marginales (base *e*) et des facteurs de Bayes ; ceux-ci doivent être interprétés en appliquant la règle empirique de Kass et Raftery (1995).

À la figure 1, nous voyons que les boîtes à moustaches se trouvent toutes au-dessus de zéro, sauf celle pour la troisième région qui ne donne aucune preuve d'association ; il n'existe peut-être pas non plus de preuve d'association pour la région 42 (10 dans la figure). Un résumé de ces résultats est présenté au tableau 4. Les facteurs de Bayes indiquent une association dans toutes les régions, sauf la région 12, et leur valeur réelle est nettement plus grande sous le modèle de non-réponse non ignorable. Les valeurs pour la région 6 et pour toutes les régions confondues sont élevées (336,3 c. 5,8 et 3 798,2 c. 0,183).

Tableau 4
Données de la NHANES pour 13 régions. Comparaison des vraisemblances marginales négatives et des facteurs de Bayes ou de l'association de la DMO et du RF d'après la procédure de ratissage et le modèle d'extension, selon la région

	Ratissage			Extension	
Région	$-\ln\{p_0(n)\}$	$-\ln\{p_a(n)\}$	***FB***	$-\ln\{p_a(n)\}$	***FB***
4	26,19	23,07	22,855	$23{,}5_{0{,}014}$	$14{,}78_{0{,}169}$
6	45,73	43,98	5,766	$40{,}5_{0{,}038}$	$336{,}27_{11{,}465}$
12	31,14	38,01	0,001	$33{,}4_{0{,}054}$	$0{,}37_{0{,}027}$
17	29,13	27,03	8,134	$27{,}0_{0{,}026}$	$10{,}27_{0{,}191}$
25	25,44	26,02	0,558	$23{,}8_{0{,}029}$	$9{,}55_{0{,}202}$
26	26,89	23,18	40,562	$23{,}9_{0{,}018}$	$24{,}71_{0{,}370}$
29	23,21	20,87	10,301	$21{,}3_{0{,}018}$	$8{,}40_{0{,}115}$
36	34,99	36,09	0,330	$33{,}1_{0{,}064}$	$21{,}13_{0{,}928}$
39	23,77	24,89	0,325	$23{,}6_{0{,}044}$	$2{,}24_{0{,}68}$
42	29,51	30,21	0,497	$30{,}3_{0{,}099}$	$4{,}33_{0{,}255}$
44	25,61	30,48	0,008	$24{,}4_{0{,}027}$	$5{,}19_{0{,}137}$
48	38,83	35,34	32,650	$39{,}1_{0{,}060}$	$2{,}15_{0{,}081}$
53	27,11	24,82	9,865	$24{,}2_{0{,}017}$	$19{,}40_{0{,}282}$
All	53,43	55,13	0,183	$46{,}1_{0{,}049}$	$3{,}798{,}24_{151{,}82}$

Nota : Dans la colonne des régions, « Toutes » désigne toutes les régions confondues ; la notation a_b signifie que la moyenne est *a* et que l'erreur-type est *b* sur les 1 000 itérations. Le terme $\ln\{p_0(n)\}$ est le même pour les deux procédures.

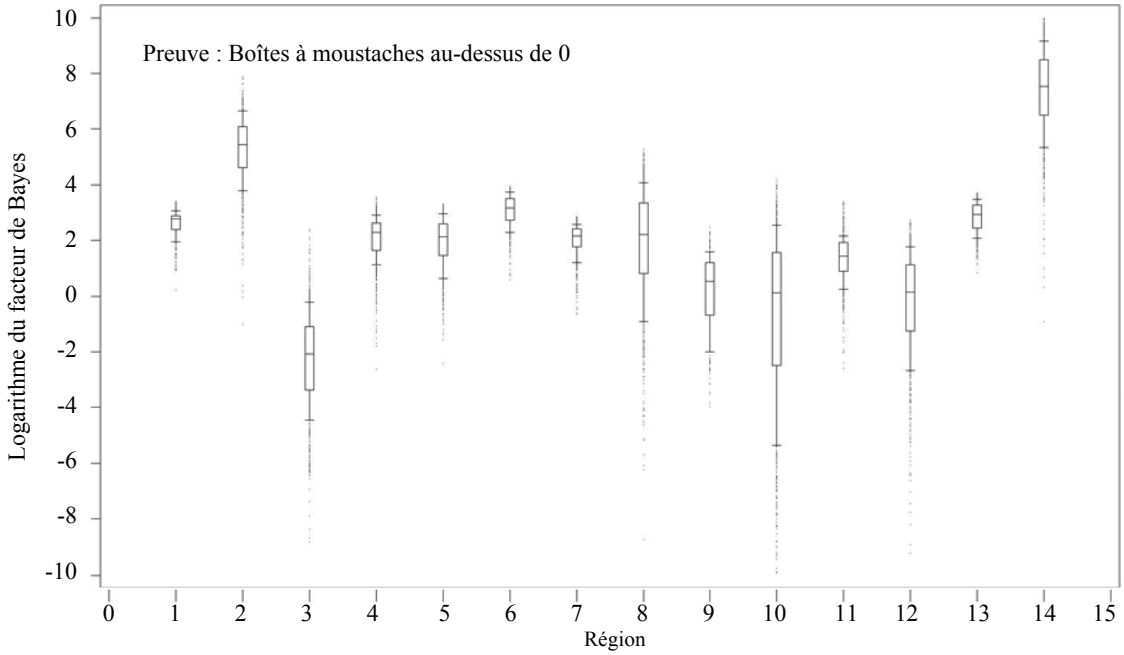


Figure 1 Boîtes à moustaches des logarithmes du facteur de Bayes selon la région pour évaluer la preuve de l'association entre la DMO et le RF

5. Conclusion

L'objectif du présent article était d'élaborer une méthodologie pour analyser les données provenant de tableaux incomplets de contingence à deux entrées, chaque tableau correspondant à un domaine ou région. Nous avons pour cela étendu la méthodologie bayésienne de Nandram et Choi (2002a, b) pour données binaires à des tableaux de contingences $r \times c$ pour petits domaines. Nous avons construit un nouveau modèle de non-réponse non ignorable bayésien (c'est-à-dire le modèle d'extension) qui est centré sur le modèle de non-réponse ignorable. Nous avons utilisé des méthodes Monte Carlo par chaîne de Markov (spécifiquement l'échantillonneur Metropolis-Hastings « griddy ») pour ajuster le modèle. Nous avons comparé notre modèle à un modèle de non-réponse ignorable et un modèle de non-réponse non ignorable. Enfin, nous avons illustré notre méthode en estimant les probabilités de cellule pour le tableau de contingence 3×3 de la densité minérale osseuse et du revenu sur 13 régions infranationales.

Nous avons montré qu'il existe des différences entre les trois modèles. En utilisant les données sur la densité minérale osseuse et le revenu familial, nous avons montré que notre modèle d'extension est un compromis entre le modèle de non-réponse ignorable et le modèle de non-réponse non ignorable. À l'aide du facteur de Bayes, nous avons montré qu'il existe des différences entre les tests de l'association de la densité minérale osseuse et du revenu familial quand les fréquences de cellule sont estimées au moyen de notre modèle et en utilisant une procédure de ratissage. En fait grâce à l'emprunt d'information, nous constatons que la preuve de l'association est nettement plus forte sous notre modèle que sous la procédure de ratissage.

Trois pistes supplémentaires pourraient être explorées. Premièrement, nous pouvons construire un modèle en vue d'intégrer l'écart systématique par rapport à l'ignorabilité. Cette tâche suscitera du travail sur le terrain coûteux supplémentaire pour obtenir l'information nécessaire. Deuxièmement, il serait également intéressant de relâcher l'hypothèse voulant que les marges du tableau de contingence soient fixes ; voir, par exemple, Nandram (2009) qui a procédé à cet examen pour une seule grande région. Troisièmement, le calage pourrait être encore amélioré (c'est-à-dire en intégrant de l'information sur les marges).

Remerciements

Nous exprimons notre reconnaissance à l'égard de deux examinateurs et d'un rédacteur associé pour l'aide apportée dans la présentation du document.

Annexe A

Densité a posteriori conjointe du modèle d'extension

Premièrement, en intégrant la fonction de densité a posteriori conjointe sur $\boldsymbol{p}$, nous obtenons

$$h(\Omega,\boldsymbol{\pi},\boldsymbol{\phi},\boldsymbol{y}_{(1)}\,|\,\boldsymbol{y}_1,\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) \propto \prod_{i=1}^{4}\frac{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\boldsymbol{\mu}_1\tau_1)}{\mathrm{D}(\boldsymbol{\mu}_1\tau_1)}$$
$$\left[\binom{n_i}{\boldsymbol{y}'_{i1},\,\boldsymbol{y}'_{i2},\,\boldsymbol{y}'_{i3},\boldsymbol{y}'_{i4}}(1-\pi_{i1})^{-y_{i\ldots}}\right.$$
$$\times\left\{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot\cdot}}\prod_{j,k}\phi_{ijk}^{y_{i\cdot jk}}[a_{ijk}(\pi_{i1},\phi_{ijk})]^{y_{i1jk}}\right\}$$
$$\times\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{\mu_{2s}\tau_2-1}}{\mathrm{D}(\boldsymbol{\mu}_2\tau_2)}\prod_{j=1}^{r}\prod_{k=1}^{c}\frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}}\right.$$
$$\left.\left.\frac{\phi_{ijk}^{\beta-1}e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)G_{\beta}[\beta(1-\pi_{i1})^{-1}]}\right\}\right]$$
$$\times\frac{a_0}{(a_0+\tau_1)^2}\cdot\frac{b_0}{(b_0+\tau_2)^2}\cdot\frac{c_0}{(c_0+\beta)^2}, \qquad \text{(A.1)}$$

où le vecteur de dimensions $rc\times 1$ est

$$\boldsymbol{y}_i^{(1)} \overset{\text{def}}{=} (y_{i.11}, y_{i.12}, \ldots, y_{i.1c}, y_{i.21}, \ldots, y_{i.2c}, y_{i.r1}, \ldots, y_{i.rc})'.$$

Maintenant, désignons par $\boldsymbol{j}$ un vecteur de dimensions $rc\times 1$ de valeurs 1 et soit

$$\boldsymbol{y}_i^{(2)} \overset{\text{def}}{=} (y_{i1..}, y_{i2..}, y_{i3..}, y_{i4..})'.$$

Alors, en agrégeant sur $\boldsymbol{\pi}$ et $\boldsymbol{\phi}$, nous obtenons

$$h(\Omega,\boldsymbol{y}_{(1)}\,|\,\boldsymbol{y}_1,\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) \propto \frac{a_0}{(a_0+\tau_1)^2}\cdot\frac{b_0}{(b_0+\tau_2)^2}\cdot\frac{c_0}{(c_0+\beta)^2}$$
$$\times\prod_{i=1}^{A}\binom{n_i}{\boldsymbol{y}'_{i1},\,\boldsymbol{y}'_{i2},\,\boldsymbol{y}'_{i3},\boldsymbol{y}'_{i4}}\frac{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\boldsymbol{\mu}_1\tau_1)}{\mathrm{D}(\boldsymbol{\mu}_1\tau_1)}\,\frac{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}{\mathrm{D}(\boldsymbol{\mu}_2\tau_2)}$$
$$\times\prod_{i=1}^{A}\left[\frac{\beta^{\beta}}{\Gamma(\beta)}\right]^{rc}\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\,\boldsymbol{j})\,I_i, \qquad \text{(A.2)}$$

où

$$I_i=\int\!\!\int\!\!\int\prod_{j,k}\left\{\frac{\left[\left(\frac{1-\pi_{i1}}{\phi_{ijk}}\right)\left(1+\frac{1}{\pi_{i1}}\{1-\pi_{i1}-\phi_{ijk}\}\right)\right]^{y_{i1jk}}}{(1-\pi_{i1})^{y_{i\cdot jk}+\beta}\,G_{\beta}\left(\frac{\beta}{1-\pi_{i1}}\right)}\right\}$$
$$\times\left\{\prod_{j,k}e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}\right\}\left\{\frac{\prod_{j,k}\phi_{ijk}^{y_{i\cdot jk}+\beta-1}}{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\,\boldsymbol{j})}\right\}$$
$$\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot\cdot}+\mu_{2s}\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}\right\}\mathrm{d}\boldsymbol{\phi}_i\,\mathrm{d}\,\boldsymbol{\pi}_i. \qquad \text{(A.3)}$$

Notons que $0\le\pi_{is}\le 1, \sum_{s=1}^{4}\pi_{is}=1$ et $0\le\phi_{ijk}\le 1$. Nous simplifions le calcul pour I_i dans (A.3) en deux étapes.

Premièrement, dans (A.3), nous effectuons la transformation

$$\phi_{ijk}=T_i\,\phi_{ijk}^{*} \qquad \sum_{j=1}^{r}\sum_{k=1}^{c}\phi_{ijk}=T_i.$$

Les nouvelles variables ϕ_{ijk}^{*} satisfont les relations $0\le\phi_{ijk}^{*}\le 1, \sum_{j=1}^{r}\sum_{k=1}^{c}\phi_{ijk}^{*}=1$ et les T_i sont soumis aux contraintes $0\le T_i\le 1/\phi_{ijk}^{*}$, pour $j=1,\ldots,r, k=1,\ldots,c$ et $0\le T_i\le rc$. Suite à cette transformation, nous avons

$$I_i=\int\!\!\int\!\!\int_0^{b_i}\prod_{j,k}\left\{\left[\frac{T_i}{1-\pi_{i1}}\right]^{\sum_{s=2}^{4}y_{isjk}}\right.$$
$$\left.\left[\frac{1}{\phi_{ijk}^{*}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{T_i}{1-\pi_{i1}}\phi_{ijk}^{*}\right\}\right)\right]^{y_{i1jk}}\right\}$$
$$\times\left\{\frac{\left(\frac{T_i}{1-\pi_{i1}}\right)^{rc\beta-1}e^{-\frac{\beta T_i}{1-\pi_{i1}}}}{(1-\pi_{i1})\left[G_{\beta}\left(\frac{\beta}{1-\pi_{i1}}\right)\right]^{rc}}\right\}\left\{\frac{\prod_{j,k}\phi_{ijk}^{*\,y_{i\cdot jk}+\beta-1}}{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\,\boldsymbol{j})}\right\}$$
$$\times\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot\cdot}+\mu_{2s}\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}\right\}\mathrm{d}T_i\,\mathrm{d}\boldsymbol{\phi}_i^{*}\mathrm{d}\boldsymbol{\pi}_i,$$

où $b_i=\min\{\{1/\phi_{ijk}^{*}\}\,j=1,\ldots,r\,;\,k=1,\ldots,c\}$.

Deuxièmement, en posant que $W_i=\{\beta T_i/1-\pi_{i1}\}$ et en absorbant le facteur $\beta^{rc\beta}/\Gamma(rc\beta)$ dans I_i, avec certaines opérations algébriques supplémentaires, nous obtenons

$$I_i=\int\!\!\int\!\!\int_0^{\frac{\beta b_i}{1-\pi_{i1}}}\frac{G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)}{\left[G_{\beta}\left(\frac{\beta}{1-\pi_{i1}}\right)\right]^{rc}}$$
$$\prod_{j,k}\left\{\left(\frac{W_i}{\beta}\right)^{\sum_{s=2}^{4}y_{isjk}}\left[\frac{1}{\phi_{ijk}^{*}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{W_i\phi_{ijk}^{*}}{\beta}\right\}\right)\right]^{y_{i1jk}}\right\}$$
$$\times\frac{W_i^{rc\beta-1}e^{-W_i}}{\Gamma(rc\beta)G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)}$$
$$\mathrm{d}W_i\left\{\frac{\prod_{j,k}\phi_{ijk}^{*\,y_{i\cdot jk}+\beta-1}}{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\,\boldsymbol{j})}\right\}\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot\cdot}+\mu_2 s\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}\right\}\mathrm{d}\boldsymbol{\phi}_i^{*}\,\mathrm{d}\boldsymbol{\pi}_i. \qquad \text{(A.4)}$$

Annexe B

Modèles de non-réponse ignorable et non ignorable

Posons que $\psi_{ijk} \equiv 1$ dans le modèle d'extension pour former le modèle de non-réponse ignorable. Pour $i = 1, \ldots, A$, nous prenons alors

$$\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2 \tau_2)$$

et indépendamment

$$\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1 \tau_1).$$

En outre, $p(\tau_2) = \{1/(1+\tau_1)^2\}, \tau_1 \geq 0, \mu_1 \sim \text{Dirichlet}(\mathbf{1})$, $p(\tau_2) = \{1/(1+\tau_2)^2\}, \tau_1 \geq 0$ et $\mu_2 \sim \text{Dirichlet}(\mathbf{1})$. Ici, nous avons l'indépendance à tous les niveaux et les vecteurs **1** sont de dimension appropriée, chaque coordonnée étant égale à l'unité. Notons que tous les paramètres du modèle ignorable doivent être identifiés et estimés.

Soit $\pi_{isjk} = \pi_{is}\psi_{ijk}$ dans le modèle d'extension pour former le modèle de non-réponse non ignorable. Dans ce cas, pour $i = 1, \ldots, A$,

$$\boldsymbol{\pi}_{ijk} \mid \boldsymbol{\mu}_2, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2 \tau_2)$$

et indépendamment

$$\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1 \tau_1).$$

Dans ce modèle, les paramètres $\boldsymbol{\pi}_{ijk}$ ne sont pas identifiables et nous prenons $\tau_2 \sim \text{Gamma}(\alpha_0, \beta_0)$, où α_0 et β_0 doivent être spécifiés. La spécification du modèle est alors achevée en attribuant à $\tau_1, \boldsymbol{\mu}_1$ et à $\boldsymbol{\mu}_2$ les mêmes propriétés distributionnelles qu'au paragraphe précédent.

Comme dans Nandram, Cox et Choi (2005), α_0 et β_0 sont spécifiés comme il suit. Le modèle de non-réponse ignorable et ajusté de manière à obtenir un échantillon de la densité de probabilité a posteriori de τ_2. Puis, α_0 et β_0 sont obtenus en utilisant la méthode des moments. Nandram, Cox et Choi (2005) ont constaté que l'inférence au sujet de $\boldsymbol{p}_i$ n'est pas très sensible au choix de ces paramètres.

Bibliographie

Cohen, G., et Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics*, 18, 13-23.

Draper, D. (1995). Assessment and propagation of model uncertainty (avec discussion). *Journal of the Royal Statistical Society*, Séries B, 57, 45-97.

Forster, J.J., et Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society*, Séries B, 60, 57-70.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Greenland, S. (2009). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Sciences*, 24, 195-210.

Kass, R.E., et Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, deuxième édition. New York : John Wiley & Sons, Inc.

Nandram, B. (2009). Bayesian inference of the cell probabilities of a two-way categorical table under non-ignorability. *Communications in Statistics - Theory and Methods*, 38, 3015-3030.

Nandram, B., et Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., et Choi, J.W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., et Choi, J.W. (2004). A nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse. *Journal of Nonparametric Statistics*, 16, 821-839.

Nandram, B., et Choi, J.W. (2005). Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines : une application aux données de la NHANES. *Techniques d'enquête*, 31, 79-92.

Nandram, B., et Choi, J.W. (2008). Une répartition bayésienne des électeurs indécis. *Techniques d'enquête*, 34, 41-54.

Nandram, B., et Choi, J.W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, 105, 120-135.

Nandram, B., Cox, L.H. et Choi, J.W. (2005). Analyse bayésienne des données catégoriques manquantes non ignorables : une application à la densité minérale osseuse et au revenu familial. *Techniques d'enquête*, 31, 233-247.

Nandram, B., Han, G. et Choi, J.W. (2002). Un modèle bayesien hiérarchique de non-réponse non-ignorable pour les données multinomiales des petites régions. *Techniques d'enquête*, 28, 157-170.

Nandram, B., Liu, N., Choi, J.W. et Cox, L. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.

National Center for Health Statistics (1992). Third national health and nutrition examination survey. *Vital and Health Statistics*, Series 2, 113.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Smith, A.F.M., et Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46, 84-88.

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 103-107
Statistique Canada, N° 12-001-X au catalogue

Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes

Phillip S. Kott [1]

Résumé

En cas de non-réponse totale d'une unité dans un échantillon tiré suivant les principes de l'échantillonnage probabiliste, une pratique courante consiste à diviser l'échantillon en groupes mutuellement exclusifs de manière qu'il soit raisonnable de supposer que toutes les unités échantillonnées dans un groupe ont la même probabilité de ne pas répondre. De cette façon, la réponse d'une unité peut être traitée comme une phase supplémentaire de l'échantillonnage probabiliste en se servant de l'inverse de la probabilité de réponse estimée d'une unité dans un groupe comme facteur de correction pour calculer les poids finaux pour les répondants du groupe. Si l'objectif est d'estimer la moyenne de population d'une variable d'enquête qui se comporte plus ou moins comme une variable aléatoire dont la moyenne est constante dans chaque groupe indépendamment des poids de sondage originaux, il est habituellement plus efficace d'intégrer les poids de sondage dans les facteurs de correction que de ne pas le faire. En fait, si la variable d'enquête se comportait exactement comme une telle variable aléatoire, l'estimation de la moyenne de population calculée en se servant des facteurs de correction pondérés selon le plan de sondage serait presque sans biais dans un certain sens (c'est-à-dire sous la combinaison du mécanisme d'échantillonnage probabiliste original et d'un modèle de prédiction), même si les unités échantillonnées dans un groupe n'ont pas toutes la même probabilité de répondre.

Mots clés : Double protection ; modèle de prédiction ; échantillonnage probabiliste ; modèle de réponse ; phase d'échantillonnage ; échantillonnage bernoullien stratifié.

1. Introduction

En l'absence de non-réponse, il est possible d'estimer la moyenne d'une population finie d'après un échantillon sans avoir à recourir à un modèle statistique qui, aussi raisonnable qu'il soit, pourrait ne pas être vérifié. Pour cela, on attribue à chaque unité de la population une probabilité positive de sélection dans l'échantillon et l'on crée des estimateurs en s'appuyant sur ce mécanisme de sélection aléatoire. Malheureusement, dans des conditions réelles, les enquêtes souffrent souvent de non-réponse.

Deux types distincts de modèle peuvent être utilisés pour faire face à la non-réponse totale d'une unité. L'un est un modèle de prédiction, ou de résultat, dans lequel on suppose que la variable d'enquête se comporte comme une variable aléatoire dont on connaît les caractéristiques, mais non les paramètres. L'autre est un modèle de réponse, ou de sélection, dans lequel le simple fait qu'une unité réponde à une enquête est traité comme une phase supplémentaire de la sélection aléatoire de l'échantillon.

Habituellement, les statisticiens d'enquête préfèrent les modèles de réponse pour deux raisons. Outre le fait que la modélisation de la réponse est commode parce qu'elle permet de traiter la réponse d'une unité comme une phase supplémentaire de l'échantillonnage aléatoire, une enquête est habituellement conçue afin de recueillir des renseignements sur plusieurs variables auprès des unités échantillonnées. La modélisation de la prédiction requiert que l'on formule un modèle hypothétique différent pour chaque variable d'enquête, chacun de ces modèles pouvant ne pas être vérifié. Par contre, la modélisation de la réponse ne nécessite l'hypothèse que d'un seul modèle. Il n'en est toutefois plus ainsi en cas de non-réponse partielle (pour une variable particulière de l'enquête). Par conséquent, les modèles de prédiction sont souvent préférés pour traiter la non-réponse partielle par imputation. Cela étant dit, la non-réponse partielle dépasse le cadre du présent article.

Sous un modèle de réponse hypothétique, les probabilités de réponse des unités sont traitées comme étant inconnues, ce qui signifie qu'elles doivent être estimées d'après l'échantillon. Habituellement, on suppose que le mécanisme de réponse des diverses unités est indépendant et qu'il ne dépend pas de la sélection de l'unité dans l'échantillon (chaque unité possède une probabilité a priori de réponse qui devient opérationnelle si elle est sélectionnée dans l'échantillon). Le modèle de réponse le plus simple et le plus fréquemment utilisé consiste à diviser l'échantillon et, implicitement la population complète, en groupes mutuellement exclusifs, appelés « groupes de réponse homogènes » par Särndal, Swensson et Wretman (1992) (le terme « classes de pondération » est plus courant ; voir, par exemple, Lohr [2009, pages 340-341]), et à supposer que chaque unité d'un groupe possède la même probabilité de ne pas répondre, quelle que soit sa probabilité de sélection dans l'échantillon original, π_k. Donc, le mécanisme de réponse produit un sous-échantillon bernoullien stratifié dans lequel les groupes constituent les strates.

1. Phillip S. Kott, RTI International, Suite 902, 6100 Executive Blvd., Rockville, MD 20852, États-Unis. Courriel : pkott@rti.org.

Conditionnellement aux tailles des échantillons de répondants dans les groupes, un sous-échantillon bernoullien stratifié dont les probabilités de sélection (réponse) sont inconnues est converti en un sous-échantillon aléatoire simple stratifié dont les probabilités de sélection sont connues : r_g / n_g pour les unités d'un groupe g quand ce groupe contient n_g unités échantillonnées, dont r_g fournissent une réponse.

Bien que la probabilité conditionnelle de réponse dans le groupe g sous le modèle de réponse bernoullien stratifié soit r_g / n_g, nous verrons qu'il est souvent préférable de multiplier le poids de sondage, $d_k = 1/\pi_k$, d'une unité répondante dans le groupe non pas par n_g / r_g, mais par

$$f_g = \frac{\sum_{k \in S_g} d_k}{\sum_{k \in R_g} d_k}, \qquad (1)$$

où S_g est l'échantillon original et R_g, le sous-échantillon de répondants dans le groupe g. Ce *facteur de correction* peut différer de n_g / r_g quand les d_k dans le groupe g varient.

Little et Vartivarian (2003) affirment que f_g est habituellement utilisé en pratique. Toutefois, ils soutiennent qu'intégrer de cette façon les poids de sondage dans le facteur de correction peut « accroître la variance ».

À la section 2, nous établissons la notation pour estimer la moyenne de population d'une variable d'enquête. L'utilisation du rapport n_g / r_g produit un estimateur à facteur d'extension double tandis que l'utilisation de f_g produit un estimateur à facteur d'expansion repondéré. Nous pouvons exprimer les deux estimateurs en nous servant d'une formulation donnée dans Kim, Navarro et Fuller (2006). Cette expression permet de voir que, si la variable d'enquête se comporte approximativement comme une variable aléatoire de moyenne constante à l'intérieur de chaque groupe, quels que soient les poids de sondage, l'utilisation de f_g est souvent plus efficace que l'utilisation de n_g / r_g. En fait, si la variable d'enquête se comporte exactement comme une telle variable aléatoire, l'estimation de la moyenne de population calculée en se servant de f_g sera presque sans biais sous la combinaison du plan de sondage original et de ce modèle de prédiction, même si le modèle de réponse ne tient pas.

À la section 3, nous montrons que les résultats empiriques présentés dans Little et Vartivarian (2003) concordent avec ces arguments et nous offrons certaines conclusions.

2. Les deux estimateurs

Supposons que nous voulions estimer la moyenne de population d'une variable d'enquête y_k :

$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{\sum_{g=1}^{G} \sum_{k \in U_g} y_k}{\sum_{g=1}^{G} N_g} = \frac{\sum_{g=1}^{G} N_g \bar{y}_{U_g}}{\sum_{g=1}^{G} N_g},$$

où la population U est divisée en G groupes, $U_1, ..., U_G$, chaque U_g contenant N_g unités, et $N = N_1 + ... + N_G$. En l'absence de non-réponse, chaque N_g est estimé sans biais sous la théorie de l'échantillonnage probabiliste par $\hat{N}_g = \sum_{k \in S_g} d_k$, et chaque $\bar{y}_{U_g}$ est estimé de façon presque sans biais (c'est-à-dire asymptotiquement)

$$\bar{y}_{S_g} = \frac{\sum_{k \in S_g} d_k y_k}{\sum_{k \in S_g} d_k}, \qquad (2)$$

sous des contraintes faibles quand n_g est suffisamment grand. Nous émettons ces deux hypothèses ici.

Pour un énoncé formel des conditions dans lesquelles chaque $\bar{y}_{S_g}$ est convergeant sous la théorie de l'échantillonnage probabiliste et par conséquent presque sans biais, voir Fuller (2009, page 115). Le lecteur que cela intéresse est invité à consulter Fuller chaque fois qu'un résultat du présent exposé dépend d'hypothèses au sujet du plan de sondage et de la population à mesure que la taille d'échantillon devient arbitrairement grande. Un traitement plus rigoureux de la plupart de la matière discutée ici sous le modèle de réponse peut être consulté dans Kim, Navarro et Fuller (2006).

Désignons l'estimateur de $\bar{y}_U$ sous échantillon complet dont nous avons discuté par $\bar{y}_S = \sum^G \hat{N}_g \bar{y}_{S_g}$. Il existe des moyens plus directs de rendre $\bar{y}_S$, mais la version susmentionnée servira mieux nos objectifs.

Si nous faisons une correction pour tenir compte de la non-réponse en utilisant le facteur f_g dans l'équation (1), nous obtenons l'estimateur à facteur d'extension repondéré :

$$\hat{\bar{y}}_{rw} = \frac{\sum_{g=1}^{G} \left(f_g \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^{G} \left(f_g \sum_{k \in R_g} d_k \right)}$$

$$= \frac{\sum_{g=1}^{G} \left(\frac{\sum_{k \in S_g} d_k}{\sum_{k \in R_g} d_k} \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^{G} \sum_{k \in S_g} d_k} = \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k y_k}{\sum_{k \in R_g} d_k} \right)}{\sum_{g=1}^{G} \hat{N}_g}.$$

Techniquement, $\hat{\bar{y}}_{rw}$ est le ratio de deux estimateurs à facteur d'extension repondéré, mais nous utilisons la terminologie plus simple ici.

L'emploi de n_g / r_g donne l'estimateur à facteur d'extension double :

$$\hat{\bar{y}}_{de} = \frac{\sum_{g=1}^{G}\left(\frac{n_g}{r_g}\sum_{k\in R_g} d_k y_k\right)}{\sum_{g=1}^{G}\left(\frac{n_g}{r_g}\sum_{k\in R_g} d_k\right)}.$$

Pour les besoins de notre étude, cet estimateur peut également être exprimé sous la forme

$$\hat{\bar{y}}_{de} = \frac{\sum_{g=1}^{G}\left(\frac{\sum_{k\in S_g} d_k p_k}{\sum_{k\in R_g} d_k p_k}\sum_{k\in R_g} d_k y_k\right)}{\sum_{g=1}^{G}\left(\frac{\sum_{k\in S_g} d_k p_k}{\sum_{k\in R_g} d_k p_k}\sum_{k\in R_g} d_k\right)} = \frac{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k y_k}{\sum_{k\in R_g} d_k p_k}\right)}{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k}{\sum_{k\in R_g} d_k p_k}\right)},$$

où

$$p_k = \frac{1}{d_k}\frac{\sum_{j\in S_g} d_j}{n_g} \quad \text{pour } k \in S_g \tag{3}$$

(de sorte que $\sum_{S_g} d_k p_k = \sum_{S_g} d_k = \hat{N}_g$).

Les estimateurs $\hat{\bar{y}}_{rw}$ et $\hat{\bar{y}}_{de}$ peuvent s'écrire tous deux sous la forme :

$$\hat{\bar{y}}_{S,\mathbf{q}} = \frac{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k y_k}{\sum_{k\in R_g} d_k q_k}\right)}{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k}{\sum_{k\in R_g} d_k q_k}\right)}. \tag{4}$$

Pour l'estimateur à facteur d'extension repondéré, tous les $q_k = 1$, tandis que pour l'estimateur à facteur d'extension double, $q_k = p_k$ tel qu'il est défini par l'équation (3).

Nous nous servirons bientôt de l'expression qui suit pour nos deux estimateurs :

$$\hat{\bar{y}}_{S,\mathbf{q}} - \bar{y}_S = \frac{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k e_k}{\sum_{k\in R_g} d_k q_k}\right)}{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k}{\sum_{k\in R_g} d_k q_k}\right)} \approx \frac{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k e_k}{\sum_{k\in R_g} d_k q_k}\right)}{\sum_{g=1}^{G}\hat{N}_g}, \tag{5}$$

où $e_k = y_k - \bar{y}_S$. L'équation (5) est vérifiée exactement quand tout $q_k = 1$. Si $q_k = p_k$, la presque égalité dépendra du fait que r_g est suffisamment grand, ainsi que d'autres contraintes faibles.

Supposons maintenant que le modèle de réponse qui suit est vérifié : Chaque unité k d'un groupe a une probabilité positive de réponse qui ne varie pas en fonction de π_k ni de y_k. Autrement dit, l'indicateur de réponse ρ_k, qui vaut 1 quand l'unité k répond si elle est échantillonnée et 0 autrement, est une variable aléatoire de Bernoulli dont la moyenne est commune dans U_g indépendamment des valeurs de π_k et y_k.

En traitant de cette façon la réponse d'une unité comme une deuxième phase de l'échantillonnage probabiliste, nous pouvons exprimer la variance/l'erreur quadratique moyenne supplémentaire due à la non-réponse, sachant l'échantillon original et les r_g pour les deux estimateurs, sous la forme

$$\begin{aligned} A_{\mathbf{q}} &= E_\rho[(\hat{\bar{y}}_{S,\mathbf{q}} - \bar{y}_s)^2 \mid S, \{r_g\}] \\ &\approx \frac{\sum_{g=1}^{G} \hat{N}_g{}^2 \operatorname{Var}_\rho(\hat{\bar{e}}_{S_g,\mathbf{q}} \mid S_g, r_g)}{\left(\sum_{g=1}^{G}\hat{N}_g\right)^2}, \end{aligned} \tag{6}$$

où $\hat{\bar{e}}_{S_g,\mathbf{q}} = \hat{\bar{y}}_{S_g,\mathbf{q}} - \bar{y}_S$, $\bar{e}_{S_g} = \bar{y}_{S_g} - \bar{y}_S$, et

$$\begin{aligned} &\operatorname{Var}_\rho(\hat{\bar{e}}_{S_g,\mathbf{q}} \mid S_g, r_g) \\ &\quad\approx \left(\frac{n_g}{r_g} - 1\right)\frac{\sum_{k\in S_g} d_k{}^2 (e_k - q_k \bar{e}_{S_g})^2}{\left(\sum_{k\in S_g} d_k q_k\right)^2} \\ &\quad= \left(\frac{n_g}{r_g} - 1\right)\frac{\sum_{k\in S_g} d_k{}^2 ([y_k - \bar{y}_S] - q_k[\bar{y}_{S_g} - \bar{y}_S])^2}{\left(\sum_{k\in S_g} d_k q_k\right)^2}, \end{aligned} \tag{7}$$

en appliquant à la population et au plan de sondage original des contraintes faibles que nous supposons être vérifiées, y compris (de nouveau) le fait que les tailles r_g sont suffisamment grandes. Ces conditions rendent les deux estimateurs presque sans biais sous la théorie de l'échantillonnage quasi probabiliste (théorie probabiliste augmentée d'un modèle de réponse) et rend discutable la distinction entre la variance et l'erreur quadratique moyenne en grand échantillon. La théorie de l'échantillonnage quasi probabiliste est également appelée théorie de l'échantillonnage « quasi fondée sur un plan » ou « quasi aléatoire ».

Si nous examinons les équations (6) et (7), nous constatons qu'à l'un des extrêmes, $\hat{\bar{y}}_{rw}$ possède une variance supplémentaire due à la non-réponse (approximativement) nulle quand tous les y_k échantillonnés originalement dans un groupe sont égaux, tandis qu'à l'autre extrême, $\hat{\bar{y}}_{de}$ présente une variance supplémentaire nulle quand tous les

$d_k e_k$ (ou, exprimés d'une autre façon, les $d_k[y_k - \bar{y}_S]$) échantillonnés au départ dans un groupe sont égaux.

Sur le plan heuristique, l'estimateur à facteur d'extension repondéré est plus efficace que l'estimateur à facteur d'extension double quand $\bar{e}_{S_g}$ est un meilleur prédicteur de e_k que $p_k \bar{e}_{S_g}$ pour $k \in S_g$. Donc, quand les groupes sont construits comme l'ont recommandé Little et Vartivarian (2003), et auparavant Little (1986), de façon à ce que la réponse y_k dans un groupe soit homogène (par opposition au fait que $d_k[y_k - \bar{y}_S]$ soit homogène), l'estimateur à facteur d'extension repondéré calculé en se servant de f_g sera généralement plus efficace que l'estimateur à facteur d'extension double calculé en se servant de n_g / r_g.

L'observation heuristique peut être exprimée formellement en utilisant une autre justification de l'utilisation de l'estimateur à facteur d'extension repondéré. Supposons que le modèle de prédiction qui suit est vérifié : Chaque y_k dans U_g est une variable aléatoire de moyenne commune, μ_g, indépendamment de π_k et ρ_k. Alors, $\hat{\bar{y}}_{rw}$ est presque sans biais sous des contraintes faibles en ce qui concerne la combinaison du mécanisme d'échantillonnage original (qui traite les d_k comme étant aléatoires, où $d_k = 0$ pour $k \neq S$) et du modèle de prédiction (qui traite les y_k comme étant aléatoires). C'est-à-dire que $\mathrm{E}_d[\mathrm{E}_y(\hat{\bar{y}}_{rw} - \bar{y}_U \mid S)] \approx 0$, puisque l'espérance double de $\hat{\bar{y}}_{rw}$ ainsi que $\bar{y}_U$ est presque $\sum^G N_g \mu_g / \sum^G N_g$. Cette absence combinée de biais est exacte quand le plan est tel que $\sum_S d_k \equiv N$. L'échantillonnage aléatoire simple stratifié est un exemple de ce type de plan. L'échantillonnage non stratifié avec probabilités inégales et de nombreux plans à plusieurs degrés ne le sont pas.

Il n'est pas difficile de voir que $\hat{\bar{y}}_{rw}$ est également sans biais par rapport à cette espérance double (c'est-à-dire $\mathrm{E}_d[\mathrm{E}_y(\hat{\bar{y}}_{rw} - \bar{y}_U \mid S)] = 0$) quand tous les μ_g sont égaux. En fait, l'espérance de $\hat{\bar{y}}_{rw}$ ainsi que de $\hat{\bar{y}}_{de}$ sous le modèle de prédiction est égale à cette moyenne commune, de même que l'espérance sous le modèle de prédiction d'un estimateur sans aucune correction pour tenir compte de la non-réponse totale, c'est-à-dire avec le remplacement de f_g dans $\hat{\bar{y}}_{rw}$ par 1. L'avantage de $\hat{\bar{y}}_{rw}$ par rapport à $\hat{\bar{y}}_{de}$ sous le modèle de prédiction s'obtient uniquement quand les μ_g varient, c'est-à-dire quand la moyenne de prédiction de la variable d'enquête varie d'un groupe à l'autre.

Il convient de souligner que si le modèle de réponse *ou* le modèle de prédiction tient, l'estimateur à facteur d'extension repondéré est presque sans biais dans un certain sens (c'est-à-dire sous la combinaison du plan de sondage original et du modèle de réponse ou sous le plan de sondage original et le modèle de prédiction). Cette propriété a été appelée « double protection » contre le biais de non-réponse. Voir, par exemple, Bang et Robins (2005).

3. Conclusion

Le présent exposé porte sur deux types de modèles distincts. Nous avons décrit un modèle de réponse dans lequel les indicateurs de réponse, ρ_k, sont traités comme une variable aléatoire de Bernoulli dans chaque groupe, mais dont les paramètres sont inconnus. Nous avons également décrit un modèle de prédiction dans lequel les valeurs observées, y_k, sont traitées comme des variables aléatoires de moyenne inconnue pouvant varier entre les groupes mais non à l'intérieur de ceux-ci.

Dans le modèle de réponse, nous supposons qu'à l'intérieur d'un groupe, les ρ_k ne dépendent pas des y_k. Par analogie, dans le modèle de prédiction, nous supposons que, dans un groupe, les y_k ne dépendent pas des ρ_k. Quand ρ_k ainsi que y_k sont traités comme des variables aléatoires, la première hypothèse, à savoir que les non-répondants *manquent au hasard*, équivaut à la seconde hypothèse, à savoir que le mécanisme de réponse est *ignorable* (voir, par exemple, Little et Rubin 1987). Il convient toutefois de bien saisir que les y_k n'ont pas à être traités comme des variables aléatoires sous le modèle de réponse et que les ρ_k n'ont pas à être traités comme des variables aléatoires sous le modèle de prédiction. Les deux concepts (réponse manquant au hasard et non-réponse ignorable) sont peut-être équivalents dans un certain sens, mais ils ne sont pas identiques.

L'exposé de Little et Vartivarian (2003) a pour élément central une série de simulations comportant une variable d'enquête binaire, deux groupes de réponses possibles et deux probabilités de sélection originales. La variable d'enquête ainsi que les indicateurs de réponse sont générés sous cinq modèles. La valeur prévue de chacune de ces variables 1) dépend du groupe de réponse seulement, 2) dépend de la probabilité de sélection seulement, 3) ne dépend ni de l'un ni de l'autre, ou 4) et 5) dépend de l'une de deux combinaisons égales de groupe de réponse et de probabilité de sélection. Cela produit 25 scénarios dont dix nous intéressent tout spécialement. Ces dix scénarios sont ceux dans lesquels la variable d'enquête est une fonction du groupe de réponse seulement ou n'est une fonction ni du groupe de réponse ni de la probabilité de sélection.

Comme le prédit notre théorie quand la variable d'enquête n'est fonction ni du groupe de réponse ni de la probabilité de sélection, l'estimateur à facteur d'extension repondéré et l'estimateur à facteur d'extension double ont tous deux un biais empirique presque nul (tableau 5 dans Little et Vartivarian) parce qu'ils sont tous deux presque sans biais sous la combinaison du plan d'échantillonnage original et d'un modèle de prédiction valide : toutes les unités de population ont la même moyenne. Quand la variable d'enquête est une fonction du groupe de réponse *et* que l'indicateur de réponse est entièrement ou partiellement une fonction de la probabilité de sélection, seul l'estimateur à

facteur d'extension repondéré est presque sans biais empiriquement, puisqu'il est le seul à être sans biais sous la combinaison du plan d'échantillonnage original et d'un modèle de prédiction valide. Par conséquent, $\hat{\bar{y}}_{rw}$ donne aussi une racine carrée de l'erreur quadratique moyenne empirique plus faible et une erreur absolue moyenne significativement plus faible qu'un estimateur de $\bar{y}_S$ (tableaux 4 et 6 dans Little et Vartivarian, respectivement ; le test de signification traite la valeur moyenne sur l'ensemble des simulations de $|\hat{\bar{y}}_{rw} - \bar{y}_S| - |\hat{\bar{y}}_{de} - \bar{y}_S|$ comme étant asymptotiquement normale).

Quand la variable d'enquête ainsi que les indicateurs de réponse sont des fonctions du groupe de réponse seulement, l'estimateur à facteur d'extension repondéré produit une racine carrée de l'erreur quadratique moyenne empirique et une erreur absolue moyenne légèrement plus faibles que l'estimateur à facteur d'extension double, mais la différence n'est pas significative pour le second paramètre.

Il n'est guère surprenant que la réduction de la racine carrée de l'erreur quadratique moyenne empirique soit modeste. La contribution de la non-réponse à la variance sous le modèle de réponse exprimé par les équations (6) et (7) est conditionnelle à l'échantillon original (techniquement, la contribution de la non-réponse à la variance totale sous échantillonnage quasi probabiliste de $\hat{\bar{y}}_{S,\mathbf{q}}$ est l'espérance de $A_{\mathbf{q}}$ dans l'équation (6) sous le mécanisme d'échantillonnage original). Dans les applications où les taux de réponse sont assez élevés (dans les simulations ils valent en moyenne 0,5), cette contribution peut être dominée par la variance/l'erreur quadratique moyenne sous échantillonnage probabiliste de l'estimateur sous échantillon complet, $\hat{\bar{y}}_U$.

Deux mises en garde sont de rigueur. La taille de l'échantillon de répondants dans chaque groupe doit être suffisamment grande pour que l'estimateur à facteur d'extension répondéré soit presque sans biais sous la théorie de l'échantillonnage quasi probabiliste. Pour l'estimateur à facteur d'extension double, il suffit que chaque taille r_g soit positive. En outre, le fait que l'estimateur à facteur d'extension repondéré est doublement protégé contre le biais de non-réponse n'est utile que si le modèle de réponse ou le modèle de prédiction hypothétique est correct. Si les probabilités de réponse *ainsi que* les valeurs des variables observées varient avec les poids de sondage, l'estimateur à facteur d'extension repondéré peut présenter un biais significatif. Malgré la perspective adoptée dans le présent exposé, il s'agit du message que Little et Vartivarian (2003) souhaitaient communiquer et il ne peut pas être contesté.

Remerciements

Je remercie le rédacteur associé et deux examinateurs de leur lecture attentive de versions antérieures du manuscrit qui m'a permis d'améliorer considérablement la qualité et l'exactitude des travaux résultants. Toute erreur qui subsiste dans le texte est entièrement imputable à l'auteur.

Bibliographie

Bang, H., et Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-972.

Fuller, W. (2009). *Sampling Statistics*, Hoboken, New Jersey : Wiley.

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

Little, R. (1986). Survey nonresponse adjustments. *Revue Internationale de Statistique*, 54, 139-157.

Little, R., et Rubin, D. (1987). *Statistical Analysis with Missing Data*, New York : John Wiley & Sons, Inc.

Little, R., et Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.

Lohr, S (2009). *Sampling: Design and Analysis*, *Second Edition*, Boston : Brooks/Cole.

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

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 27, No. 4, 2011

An Index Number Formula Problem: The Aggregation of Broadly Comparable Items
Mick Silver 553

Identifying Sources of Error in Cross-national Questionnaires: Application of an Error Source Typology to Cognitive Interview Data
Rory Fitzgerald, Sally Widdop, Michelle Gray, Debbie Collins 569

Using a Geographic Segmentation to Understand, Predict, and Plan for Census and Survey Mail Nonresponse
Nancy Bates, Mary H. Mulry 601

Are They Really Too Busy for Survey Participation? The Evolution of Busyness and Busyness Claims in Flanders
Anina Vercruyssen, Bart van de Putte, Ineke A.L. Stoop 619

Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys
Avinash C. Singh, Fulvia Mecatti 633

On Variance Estimation for the French Master Sample
Guillaume Chauvet 651

Secure Multiple Linear Regression Based on Homomorphic Encryption
Rob Hall, Stephen E. Fienberg, Yuval Nardi 669

Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey
Nicole Watson, Rosslyn Starick 693

Book Reviews 717
Editorial Collaborators 723
Index to Volume 27, 2011 727

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 28, No. 1, 2012

Capture–Recapture Sampling and Indirect Sampling
Pierre Lavallée, Louis-Paul Rivest 1

An Examination of Within-Person Variation in Response Propensity over the Data Collection Field Period
Kristen Olson, Robert M. Groves 29

Concording U.S. Harmonized System Codes over Time
Justin R. Pierce, Peter K. Schott 53

Uncertainty Analysis in Statistical Matching
Pier Luigi Conti, Daniela Marella, Mauro Scanu 69

Constrained Small Area Estimators Based on M-quantile Methods
Enrico Fabrizi, Nicola Salvati, Monica Pratesi 89

A Multiplicative Masking Method for Preserving the Skewness of the Original Micro-records
Nicolas Ruiz 107

Cell Bounds in *k*-way Tables Given Conditional Frequencies
Byran J. Smucker, Aleksandra Slavković, Xiaotian Zhu 121

Book Review 141
In Other Journals 151

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 39, No. 4, December/décembre 2011

Zhiguo Li and Bin Nan
Relative risk regression for current status data in case-cohort studies 557

Liuquan Sun, Xingqiu Zhao and Jie Zhou
A class of mixed models for recurrent event data 578

Chong Gu and Ping Ma
Nonparametric regression with cross-classified responses 591

David R. Bickel
A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons 610

Xiaoming Wang and Keumhee C. Carriere
Assessing additivity in nonparametric models - A kernel-based method 632

Chenlei Leng and Cheng Yong Tang
Improving variance function estimation in semiparametric longitudinal data analysis 656

Jihnhee Yu, Albert Vexler, Seong-Eun Kim and Alan D. Hutson
Two-sample empirical likelihood ratio tests for medians in application to biomarker evaluations 671

Marwan Zidan, Jung-Chao Wang and Magdalena Niewiadomska-bugaj
Comparison of *k* independent, zero-heavy lognormal distributions 690

Ivan Kojadinovic, Johan Segers and Jun Yan
Large-sample tests of extreme-value dependence for multivariate copulas 703

Jeffrey S. Rosenthal
Was the conservative majority predictable? 721

Errata

Hideatsu Tsukahara
Erratum to “Semiparametric estimation in copula models” 734

CONTENTS TABLE DES MATIÈRES

Volume 40, No. 1, March/mars 2012

D.M. Hajducek and J.F. Lawless
Duration analysis in longitudinal studies with intermittent observation times and losses to followup 1

Yongseok Park, John D. Kalbfleisch and Jeremy M.G. Taylor
Constrained nonparametric maximum likelihood estimation of stochastically ordered survivor functions..... 22

Binbing Yu and Ram C. Tiwari
A Bayesian approach to mixture cure models with spatial frailties for population-based cancer relative survival data 40

Yasaman Hosseinkashi, Shojaeddin Chenouri, Christopher G. Small and Rob Deardon
A stochastic graph process for epidemic modelling..... 55

E.C. Brechmann, C. Czado and K. Aas
Truncated regular vines in high dimensions with application to financial data 68

Alexander Bauer, Claudia Czado and Thomas Klein
Pair-copula constructions for non-Gaussian DAG models..... 86

Liang Peng
Approximate jackknife empirical likelihood method for estimating equations 110

Guillaume Chauvet and David Haziza
Fully efficient estimation of coefficients of correlation in the presence of imputed survey data..... 124

Zilin Wang and Mary E. Thompson
A resampling approach to estimate variance components of multilevel models..... 150

Huybrechts F. Bindele and Asheber Abebe
Bounded influence nonlinear signed-rank regression 172

Mary C. Meyer
Constrained penalized splines 190