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

Techniques d’enquête

Techniques d’enquête
44-1

Date de diffusion : le 21 juin 2018

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 STATCAN.infostats-infostats.STATCAN@canada.ca

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

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

Programme des services de dépôt

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

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 « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Signes conventionnels dans les tableaux

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

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

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2018

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête

Nº 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Juin 2018

•

Volume 44

•

Numéro 1

Statistique Canada Statistics Canada

Canada

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 C. Julien
Anciens présidents J. Kovar (2009-2013)
D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Membres G. Beaudoin
S. Fortier (Gestionnaire de la production)
W. Yung

COMITÉ DE RÉDACTION

Rédacteur en chef W. Yung, *Statistique Canada*

Ancien rédacteur en chef M.A. Hidiroglou (2010-2015)
J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Rédacteurs associés

J.-F. Beaumont, *Statistique Canada*
M. Brick, *Westat Inc.*
P. Brodie, *Office for National Statistics*
P.J. Cantwell, *U.S. Bureau of the Census*
J. Chipperfield, *Australian Bureau of Statistics*
J. Dever, *RTI International*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
D. Haziza, *Université de Montréal*
M.A. Hidiroglou, *Statistique Canada*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
D. Judkins, *Abt Associates*
J. Kim, *Iowa State University*
P. Kott, *RTI International*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistique Canada*
I. Molina, *Universidad Carlos III de Madrid*
J. Opsomer, *Colorado State University*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
F. Scheuren, *National Opinion Research Center*
P.L.N.D. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *University of Southampton*
D. Steel, *University of Wollongong*
M. Thompson, *University of Waterloo*
D. Toth, *U.S. Bureau of Labor Statistics*
J. van den Brakel, *Statistics Netherlands*
C. Wu, *University of Waterloo*
A. Zaslavsky, *Harvard University*
L.-C. Zhang, *University of Southampton*

Rédacteurs adjoints C. Bocci, K. Bosa, C. Boulet, H. Mantel, S. Matthews, C.O. Nambeu, Z. Patak et Y. You, *Statistique Canada*

POLITIQUE DE RÉDACTION

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

Présentation de textes pour la revue

Techniques d'enquête est publiée en version électronique deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (statcan.smj-rte.statcan@canada.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca/techniquesdenquete).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 44, numéro 1, juin 2018

Table des matières

Articles réguliers

Omer Ozturk et Konul Bayramoglu Kavlak
Inférence sous un modèle pour des échantillons fondés sur des ensembles ordonnés.....1

Guillaume Chauvet et Camelia Goga
Linéarisation contre Bootstrap pour estimer la variance de l'évolution de l'indice de Gini 19

Jacco Daalmans, Tommaso Di Fonzo, Nino Mushkudiani et Reinier Bikker
Étalonnage temporel avec préservation des taux de croissance (PTC) : Inconvénients et autres solutions 45

Phillip S. Kott, Dan Liao, Jeremy Aldworth, Sarra L. Hedden, Joseph C. Gfroerer, Jonaki Bose et Lisa Colpe
Étude de divers estimateurs de la prévalence de la maladie mentale grave fondés sur un échantillon à deux phases.....65

Katherine Jenny Thompson, Stephen Kaputa et Laura Bechtel
Stratégies de sous-échantillonnage des non-répondants pour les programmes économiques.....81

Malay Ghosh, Jiyoun Myung et Fernando A.S. Moura
Estimation bayésienne robuste sur petits domaines 109

Jack Kuang Tsung Chen, Richard L. Valliant et Michael R. Elliott
Calage assisté par un modèle pour des données de sondage non probabiliste en utilisant le LASSO adaptatif.....125

Autres revues 157

Techniques d'enquête, juin 2018
Vol. 44, nº 1, pp. 1-17
Statistique Canada, nº 12-001-X au catalogue

Inférence sous un modèle pour des échantillons fondés sur des ensembles ordonnés

Omer Ozturk et Konul Bayramoglu Kavlak[1]

Résumé

Le présent article traite de l'inférence statistique sous un modèle de superpopulation en population finie quand on utilise des échantillons obtenus par échantillonnage d'ensembles ordonnés (EEO). Les échantillons sont construits sans remise. Nous montrons que la moyenne d'échantillon sous EEO est sans biais par rapport au modèle et présente une plus petite erreur de prédiction quadratique moyenne (EPQM) que la moyenne sous échantillonnage aléatoire simple (EAS). Partant d'un estimateur sans biais de l'EPQM, nous construisons aussi un intervalle de confiance de prédiction pour la moyenne de population. Une étude en simulation à petite échelle montre que cet estimateur est aussi bon qu'un estimateur sur échantillon aléatoire simple (EAS) quand l'information pour le classement est de qualité médiocre. Par ailleurs, cet estimateur est plus efficace que l'estimateur EAS quand la qualité de l'information de classement est bonne et que le ratio des coûts de l'obtention d'une unité sous EEO et sous EAS n'est pas très élevé. L'étude en simulation indique aussi que les probabilités de couverture des intervalles de prédiction sont très proches des probabilités de couverture nominales. La procédure d'inférence proposée est appliquée à un ensemble de données réel.

Mots-clés : Échantillonnage d'ensembles ordonnés; population finie; erreur de prédiction quadratique moyenne; modèle de coût d'échantillonnage; classement cohérent; classement concomitant; classement visuel.

1 Introduction

Dans de nombreuses études par sondage, il est courant que la base de sondage contienne de l'information auxiliaire en plus de la caractéristique d'intérêt. Sous une hypothèse de modélisation assez forte, cette information auxiliaire améliore l'inférence statistique. Par exemple, les estimateurs par le ratio et par la régression utilisent l'information sur les covariables sous une hypothèse de linéarité pour estimer la moyenne ou le total de population. L'information auxiliaire peut aussi être utilisée sous une hypothèse plus faible dans le cas d'un échantillon fondé sur des ensembles ordonnés (EEO; en anglais, *ranked set sample* ou *RSS*) ou d'un échantillon poststratifié au jugé (PSJ). Ces échantillons sont sélectionnés en se servant de l'information auxiliaire pour accroître le contenu informationnel de chaque unité mesurée grâce à un processus de classement. Le processus de classement est appliqué à un petit ensemble de taille H formé en groupant l'unité mesurée avec $H-1$ unités non mesurées supplémentaires tirées de la population. Le processus de classement, qui est exécuté avant ou après la mesure, détermine la position relative de chaque unité mesurée. L'information pour le classement peut découler d'une inspection visuelle ou d'une autre méthode de classement. Un mécanisme de classement raisonnable requiert une relation monotone entre la variable de classement et la réponse, condition qui est nettement plus faible que la forte hypothèse de linéarité des estimateurs par la régression et par le ratio.

Un échantillon EEO équilibré pour un ensemble de taille H et un cycle de taille d peut être construit en tirant d'abord $n = Hd$ échantillons aléatoires simples de taille H de la population, puis en classant les unités de chaque échantillon de la plus petite à la plus grande sans effectuer de mesure. Dans ces n

1. Omer Ozturk, département des statistiques, The Ohio State University, 1958 Neil Avenue, Columbus, OH, 43210, États-Unis. Courriel : omer@stat.osu.edu; Konul Bayramoglu Kavlak, département des sciences de l'actuariat, Hacettepe University, Ankara, Turquie. Courriel : konul.bayramoglu@hacettepe.edu.tr.

ensembles (échantillons) ordonnés, on mesure alors les unités de rang 1 dans les d premiers ensembles, l'unité de rang 2 dans les d ensembles suivants, et ainsi de suite. On obtient ainsi des échantillons de H ensembles différents de statistiques d'ordre au jugé, qui contiennent chacun d statistiques d'ordre au jugé indépendantes et identiquement distribuées.

Un contraste prononcé existe entre une observation sous EAS et sous EEO, l'observation sur un échantillon EAS ne fournissant de l'information que sur l'unité mesurée, alors que l'observation sur un échantillon EEO fournit, outre l'information issue de l'unité mesurée, une information limitée au sujet des $(H-1)$ autres unités non mesurées dans l'ensemble qui découle de la position relative (rang) de l'unité mesurée. Puisque le processus d'attribution de rang (classement) ne nécessite pas de mesure formelle et qu'il est habituellement moins onéreux que cette mesure formelle, l'échantillon EEO permet de réduire considérablement le coût de l'échantillonnage.

Un échantillon PSJ diffère d'un échantillon EEO en ce sens que l'étape de classement a lieu après le tirage d'un échantillon EAS. La construction d'un échantillon PSJ de taille n requiert une taille d'ensemble H. Une fois que la taille d'ensemble H est déterminée, on tire d'abord un échantillon aléatoire simple de taille n et on effectue une mesure sur chacune des n unités. Pour chaque unité mesurée dans l'échantillon, on sélectionne alors $H-1$ unités supplémentaires pour former un ensemble de taille H. Les unités dans cet ensemble sont triées de la plus petite à la plus grande sans effectuer de mesure, et le rang de l'unité mesurée dans l'ensemble est enregistré. L'échantillon PSJ est alors constitué de n valeurs mesurées, ainsi que des rangs de ces valeurs.

Les échantillons EEO et PSJ induisent tous deux parmi les unités mesurées une structure stochastique telle que les observations contenues dans la classe au jugé h sont habituellement plus petites que les observations contenues dans la classe au jugé h', $h < h'$. Cette caractéristique de classement stochastique répartit les unités mesurées dans le support de la distribution et crée un échantillon plus représentatif qu'un échantillon aléatoire simple. La nature du classement stochastique dans un échantillon PSJ diffère considérablement de celle du classement stochastique dans un échantillon EEO. Un échantillon PSJ consiste en un échantillon aléatoire simple et un vecteur de rangs connexe. Ce vecteur de rangs est relié lâchement à l'échantillon et peut être ignoré si on le souhaite. Par contre, un échantillon EEO est mesuré sous forme de statistiques d'ordre établies au jugé, de sorte que les rangs au jugé ne peuvent pas être séparés des valeurs observées. Un échantillon EEO ne peut pas être traité comme un échantillon EAS.

Les plans d'échantillonnage PSJ et EEO en population finie ont suscité les uns et les autres beaucoup d'intérêt chez les chercheurs. Patil, Sinha et Taillie (1995) ont utilisé l'échantillonnage d'ensembles ordonnés pour estimer la moyenne de population pour une population de taille N quand l'échantillon est construit sans remise. Takahasi et Futatsuya (1998) ont montré que l'estimateur de la moyenne de population calculé pour un échantillon fondé sur des ensembles ordonnés est plus précis que l'estimateur calculé pour un échantillon aléatoire simple quand les échantillons sont tirés sans remise à partir d'une population finie. Deshpande, Frey et Ozturk (2006) ont décrit trois plans d'échantillonnage différents et construit des intervalles de confiance non paramétriques pour les quantiles de population. Al-Saleh et Samawi (2007), Ozdemir et Gokpinar (2007 et 2008), Gokpinar et Ozdemir (2010), Ozturk et Jozani (2013), Frey (2011) et Ozturk (2014, 2015, 2016a) ont calculé les probabilités d'inclusion et construit les estimateurs de type

Horwitz-Thompson pour la moyenne et le total de population à partir d'un échantillon fondé sur des ensembles ordonnés. Ces documents de recherche montrent qu'un plan EEO améliore considérablement l'efficacité par rapport au plan d'échantillonnage aléatoire simple habituel. Ozturk (2016b) a élaboré des estimateurs de la moyenne de population basés sur un échantillon PSJ et a montré que ces estimateurs nécessitaient un facteur de correction pour population finie similaire à celui utilisé dans le cas d'un échantillon aléatoire simple.

Tous les travaux de recherche publiés sur les plans d'échantillonnage PSJ et EEO en population finie s'appuient sur une approche fondée sur le plan de sondage. Autant que nous sachions, le modèle de superpopulation n'a pas été utilisé. Dans le présent article, nous élaborons une inférence statistique sous un modèle en utilisant un plan d'échantillonnage EEO pour la moyenne et le total de population en population finie. Des résultats similaires, avec une variation additionnelle due aux tailles aléatoires des échantillons des classes créées au jugé, peuvent également être établis pour un plan d'échantillonnage PSJ. En raison des tailles aléatoires d'échantillons des classes créées au jugé, les estimateurs fondés sur un échantillon PSJ sont moins efficaces que ceux fondés sur un échantillon EEO. Pour cette raison, nous n'examinerons pas davantage l'échantillon PSJ dans le présent article. La section 2 définit clairement le modèle et décrit les plans d'échantillonnage pour l'EEO sous un modèle de superpopulation. Nous montrons que les estimateurs de la moyenne et du total de population sont sans biais par rapport au modèle et que leurs erreurs de prédiction quadratique moyenne (EPQM) sont plus petites que l'EPQM des mêmes estimateurs pour un échantillon EAS. La section 3 décrit la construction d'estimateurs sans biais pour l'EPQM et fournit des intervalles de confiance approximatifs pour la moyenne et le total de population. La section 4 décrit les modèles de coût pour tenir compte de l'effet du coût supplémentaire (par rapport au coût de l'échantillon EAS) dans la construction de l'échantillon EEO. La section 5 fournit des données empiriques sur la performance des estimateurs. La section 6 décrit l'application des estimateurs proposés à un exemple en population finie. La section 7 donne quelques remarques finales.

2 Plans d'échantillonnage

Nous considérons des plans d'échantillonnage EEO à partir d'un modèle de superpopulation pour tirer des inférences statistiques dans un contexte de population finie. Soit Y la caractéristique d'intérêt. Les copies de Y, $Y_1, \ldots, Y_N$, sont considérées comme des variables aléatoires indépendantes et identiquement distribuées (iid) provenant d'une superpopulation. L'hypothèse de base pour ce modèle de superpopulation peut être énoncée sous la forme

$$\text{Modèle}: Y_1, \ldots, Y_N \text{ indépendantes et identiquement distribuées avec } E_M(Y_i) = \mu, V_M(Y_i) = \sigma^2. \quad (2.1)$$

L'indice M dans le modèle (2.1) est utilisé pour souligner que la moyenne et la variance sont calculées d'après un modèle de superpopulation plutôt que la distribution obtenue par randomisation comme dans Ozturk (2016b). Dans ce modèle de superpopulation, μ et σ^2 représentent les paramètres de population infinie inconnus.

Dans le modèle de superpopulation, une réalisation particulière, $y_1, \ldots, y_N$, des variables aléatoires $Y_1, \ldots, Y_N$ à partir du modèle (2.1) est considérée comme une population finie. Soit $P^N = \{y_1, \ldots, y_N\}$ cette population finie. Un échantillon est construit par échantillonnage d'ensembles ordonnés à partir de P^N. Sans perte de généralité, nous supposons que $y_{(1)} < y_{(2)} < \ldots < y_{(N)}$ représentent les valeurs ordonnées de $y_1, \ldots, y_N$, où $y_{(i)}$ est la i^{e} plus grande valeur de Y dans P^N. Dans tout le texte, H et d sont utilisés pour désigner les tailles d'ensemble et de cycle, respectivement.

Pour construire un échantillon à partir d'ensembles ordonnés (échantillon EEO), on sélectionne un ensemble de H unités expérimentales, $y_{s_1}, \ldots, y_{s_H}$, au hasard à partir de P^N et on classe ces unités par ordre croissant en fonction de leurs valeurs de Y sans effectuer de mesure. Le processus de classement peut être exécuté par inspection visuelle ou au moyen de certaines variables auxiliaires, et est donc sujet à une erreur de classement. L'unité qui correspond à la plus petite valeur de Y, $y_{[1]}$, est identifiée et mesurée, l'indice inférieur entre crochets, [1], désignant le rang de la plus petite unité (rang 1) dans l'ensemble $\{y_{[1]}, y^*_{[2]}, \ldots, y^*_{[H]}\}$. Les unités restantes non mesurées sont désignées par $\{y^*_{[2]}, \ldots, y^*_{[H]}\}$. Après avoir mesuré $y_{[1]}$, aucune des H unités contenues dans l'ensemble $\{y_{[1]}, y^*_{[2]}, \ldots, y^*_{[H]}\}$ n'est remise dans la population. On sélectionne alors un autre ensemble de H unités expérimentales au hasard parmi la population restante P^{N-H} et on les classe sans effectuer de mesure. Cette fois-ci, l'unité qui correspond à la deuxième plus petite valeur de Y, $y_{[2]}$, est identifiée et mesurée dans $\{y^*_{[1]}, y_{[2]}, y^*_{[3]}, \ldots, y^*_{[H]}\}$. Ce processus est poursuivi jusqu'à ce qu'un échantillon aléatoire simple de taille H soit tiré de la population réduite $P^{N-H(H-1)}$ et que la H^{e} plus petite unité soit identifiée et mesurée dans l'ensemble $\{y^*_{[1]}, y^*_{[2]}, \ldots, y^*_{[H-1]}, y_{[H]}\}$. Ce processus est appelé un cycle. Un cycle consiste à sélectionner H ensembles disjoints, chacun de taille H et à mesurer seulement H unités. Les $H(H-1)$ unités restantes sont utilisées uniquement pour les besoins du classement. Le cycle est répété d fois pour obtenir un échantillon EEO de taille $n = dH$ unités. Un échantillon EEO peut être représenté par

$$W_{h,i,H} = \{y^*_{[1]i}, \ldots, y^*_{[h-1]i}, y_{[h]i}, y^*_{[h+1]i}, \ldots, y^*_{[H]i}\}, \quad h = 1, \ldots, H, \quad i = 1, \ldots, d, \tag{2.2}$$

où seules les valeurs $y_{[h]i}$, $h = 1, \ldots, H$, $i = 1, \ldots, d$ sont mesurées. Les autres valeurs sont utilisées pour obtenir le rang des valeurs mesurées. Les unités dans les ensembles $W_{h,i,H}$ et $W_{h',i',H}$ sont toutes indépendantes si $h \neq h'$ ou $i \neq i'$, mais les unités dans $W_{h,i,H}$ sont toutes corrélées, puisqu'elles sont classées dans le même ensemble. Sous le modèle (2.1), les moyennes, les variances et les covariances des statistiques d'ordre obtenues au jugé sont données par

$$E_M(Y_{[h]i}) = \mu_{[h]}, \operatorname{Var}_M(Y_{[h]i}) = \sigma^2_{[h]},$$

$$\operatorname{Cov}_M(Y_{[h]i}, Y_{[h']i}) = \begin{cases} \sigma_{[h,h']} & \text{si } Y_{[h]i}, Y_{[h']i} \text{ proviennent du même ensemble} \\ 0 & \text{autrement.} \end{cases}$$

Il convient de souligner que, puisque tous les ensembles sont disjoints, aucune unité ne peut être utilisée plus d'une fois dans n'importe lequel des ensembles. Donc, toutes les unités échantillonnées sont distinctes. Puisque les ensembles sont classés indépendamment, les $Y_{[h]i}$ sont mutuellement indépendants. Les observations de même rang h, $Y_{[h]i}$, $i = 1, \ldots, d$ sont identiquement distribuées.

L'estimateur de la moyenne de population μ fondé sur les données d'échantillon EEO dans l'équation (2.2) peut être défini comme il suit.

$$\bar{Y}_R = \frac{1}{dH}\sum_{h=1}^{H}\sum_{i=1}^{d} Y_{[h]i}. \tag{2.3}$$

On peut constater immédiatement que l'estimateur $\bar{Y}_R$ est sans biais par rapport au modèle. Autrement dit, sous le modèle (2.1), $E_M\left(\bar{Y}_R - \bar{Y}_N\right) = 0,$ où $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$.

Nous considérons maintenant l'erreur de prédiction quadratique moyenne (EPQM) de l'estimateur $\bar{Y}_R$ sous le modèle (2.1)

$$\text{EPQM}_M\left(\bar{Y}_R\right) = E_M\left(\bar{Y}_R - \frac{1}{N}\sum_{i=1}^{N} Y_i\right)^2 = E_M\left(\bar{Y}_R - \bar{Y}_N\right)^2.$$

Puisque le prédicteur $\bar{Y}_R$ est sans biais par rapport au modèle pour $\bar{Y}_N$, $E_M\left(\bar{Y}_R - \bar{Y}_N\right) = 0,$ l'erreur de prédiction quadratique moyenne (EPQM) de $\bar{Y}_R$ est égal à $\text{Var}_M\left(\bar{Y}_R - \bar{Y}_N\right)$.

Théorème 1 : *Soit $Y_{[h]i}, h = 1,\ldots,H, i = 1,\ldots,d,$ un échantillon EEO obtenu à partir d'une population finie P^N. Sous un modèle de superpopulation dans l'équation (2.1), l'erreur de prédiction quadratique moyenne de l'estimateur $\bar{Y}_R$ est donnée par*

$$\sigma^2_{\text{EEO}} = \text{EPQM}_M\left(\bar{Y}_R\right) = \frac{N-n}{Nn}\sigma^2 - \frac{1}{nH}\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2. \tag{2.4}$$

Nous notons que l'expression de l'équation (2.4) est très similaire à celle de la variance d'échantillon d'un échantillon EEO en population infinie. La seule différence tient au coefficient $\frac{N-n}{Nn}$. En population infinie, la fraction $\frac{N-n}{Nn}$ de l'équation (2.4) devient $\frac{1}{n}$. Donc, $\left(1 - \frac{n}{N}\right)$ est le facteur de correction pour population finie (cpf) pour la variance de la moyenne d'échantillon EEO. Si la taille d'échantillon n'est pas petite comparativement à la taille de la population N, le facteur cpf, $\frac{N-n}{Nn}$, apporte une correction à la variance d'une moyenne d'échantillon EEO. Cette correction sera importante si n est relativement grand par rapport à N. Si n est petit, le facteur de correction pour population finie (cpf) est proche de 1 et son effet est minimal.

Corollaire 1 : *Supposons que n et N augmentent de telle façon que le ratio $\frac{n}{N}$ tend vers une limite a,* $\lim_{n\to\infty}\frac{n}{N} = a$.

i) Si $a > 0$, $\sigma^2_{\hat{E}EO}$ converge vers une forme simple

$$\lim_{n\to\infty} n\sigma^2_{\hat{E}EO} = (1-a)\sigma^2 - \frac{1}{H}\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2,$$

ii) si $a = 0$, $\lim_{n\to\infty} n\sigma^2_{\hat{E}EO} = \frac{1}{H}\sigma^2_{[h]}$, *ce qui est la même chose que la variance de la moyenne d'échantillon d'un échantillon EEO équilibré en population infinie,*

iii) si a est une valeur strictement positive, alors $\lim_{n\to\infty} n\sigma^2_{\hat{E}EO} < \frac{1}{H}\sigma^2_{[h]}$.

Le corollaire indique que, quand les tailles de l'échantillon et de la population augmentent à un certain taux, la variance de la moyenne d'échantillon d'un échantillon EEO $\left(\sigma^2_{\text{EEO}}\right)$ en population finie se réduit à une forme simple. Si a est strictement positive, la variance d'une moyenne d'échantillon EEO est plus petite que la variance d'une moyenne d'échantillon EEO en population infinie.

3 Estimateurs sans biais

À la présente section, nous construisons un estimateur sans biais pour σ^2_{EEO}. En réécrivant l'estimateur de σ^2_{EEO} sous une forme légèrement différente, nous obtenons

$$\begin{aligned}\sigma^2_{\text{EEO}} &= \left(\frac{N-n}{Nn}\right)\sigma^2 - \frac{1}{nH}\sum_{h=1}^{H}(\mu_{[h]}-\mu)^2 \\ &= \left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2 - \frac{1}{nH}\left(H\sigma^2 - \sum_{h=1}^{H}\sigma^2_{[h]}\right) \\ &= \left(\frac{-1}{N}\right)\sigma^2 + \frac{1}{nH}\sum_{h=1}^{H}\sigma^2_{[h]}.\end{aligned}$$

Soit

$$\begin{aligned}T_1^* &= \frac{1}{2d^2H^2}\sum_{h=1}^{H}\sum_{h\neq h'}^{H}\sum_{i=1}^{d}\sum_{j=1}^{d}(Y_{[h]i}-Y_{[h']j})^2 \\ T_2^* &= \frac{1}{2d(d-1)H^2}\sum_{h=1}^{H}\sum_{i=1}^{d}\sum_{j\neq i}^{d}(Y_{[h]i}-Y_{[h]j})^2.\end{aligned}$$

En utilisant ces définitions, nous pouvons établir facilement le résultat qui suit.

Théorème 2 : *Soit* $Y_{[h]i}$, $i = 1,\ldots,n, h = 1,\ldots,H$ *un échantillon EEO de taille d'ensemble* H *provenant d'une population finie. Un estimateur sans biais de* σ^2_{EEO} *est donné par*

$$\hat{\sigma}^2_{EEO} = T_2^*\left(\frac{H}{n}\right) - (T_1^* + T_2^*)\frac{1}{N}. \tag{3.1}$$

Le théorème 2 indique que l'estimateur de variance est sans biais pour toutes tailles d'échantillon et d'ensemble indépendamment de la qualité de l'information de classement. Un estimateur sans biais de la variance de $\bar{Y}_R$ nous permet de construire un intervalle de confiance pour la moyenne et le total de population. En utilisant une approximation normale, un intervalle de confiance à $(1-\alpha)100$ % pour la moyenne de population est donné par

$$\bar{Y}_R \pm t_{n-H,\alpha/2}\hat{\sigma}^2_{\text{EEO}},$$

où $t_{df,a}$ est le a^{e} quantile supérieur de la distribution t avec degrés de liberté df. Le nombre $n-H$ de degrés de liberté est proposé pour tenir compte de l'hétérogénéité entre les H classes obtenues au jugé. Le choix de $df = n-H$ est également proposé dans Ahn, Lim et Wang (2014) en population infinie.

4 Modèle de coût

L'amélioration de l'efficacité qu'offre l'estimateur EEO résulte de l'information sur la position relative (rang) de l'observation mesurée parmi les $H-1$ unités non mesurées dans un ensemble. Cette information supplémentaire est obtenue au prix de l'échantillonnage d'un ensemble de taille H et du classement subséquent. Le classement peut être effectué à l'aide d'une variable (auxiliaire) concomitante ou par

inspection visuelle des unités physiques dans chaque ensemble. Donc, ces deux approches, modèles de classement visuel ou concomitant, peuvent mener à des structures de coût différentes. Dans l'un ou l'autre cas, le processus de classement doit avoir une certaine cohérence afin de pouvoir élaborer une fonction de coût ayant un sens. Patil, Sinha et Taillie (1997) ont défini un processus de classement cohérent dans lequel le classement d'un ensemble est cohérent pour tous les sous-ensembles et superensembles. Sous un scénario de classement cohérent, le classement par rang d'un ensemble de H unités demeure identique lors du classement de n'importe quel sous-ensemble ou superensemble contenant ces unités. Pour des renseignements plus détaillés sur le classement cohérent, le lecteur est invité à consulter Patil et coll. (1997) ou Nahhas, Wolfe et Chen (2002).

Le classement concomitant s'appuie sur une variable auxiliaire pour classer H unités dans un ensemble. La qualité du classement dépend de la relation monotone (qui ne doit pas nécessairement être linéaire) entre la variable d'intérêt et la variable auxiliaire. Par ailleurs, une inspection visuelle peut être effectuée de différentes façons. L'une des stratégies fait appel à la comparaison par paires. Sous un classement cohérent, les $\binom{H}{2}$ comparaisons par paires ne sont pas toutes nécessaires pour un classement visuel. Par exemple, dans un ensemble de taille $H = 3,$ si l'unité 1 est jugée plus petite que l'unité 2 et que l'unité 2 est jugée plus petite que l'unité 3, nous supposons raisonnablement que l'unité 1 est plus petite que l'unité 3 sans effectuer de comparaison. Afin de distinguer les effets des structures de coût des scénarios de classement concomitant et visuel, nous dénotons l'estimateur de l'équation (2.3) par $\overline{Y}_{\mathrm{CC}}$ pour classement concomitant et $\overline{Y}_{\mathrm{CV}}$ pour classement visuel.

Pour le classement visuel, nous utilisons le modèle d'inspection visuelle de Nahhas et coll. (2002). Ce modèle compare toujours l'unité sélectionnée avec l'élément le plus grand du classement précédent. Une unité est choisie au hasard et comparée à l'unité jugée précédemment comme étant la plus grande. Si l'unité examinée est jugée plus grande, elle devient la plus grande parmi toutes les unités jugées. Autrement, elle est comparée à la deuxième plus grande unité jugée précédemment et ainsi de suite jusqu'à ce qu'un rang lui soit attribué. Le nombre de comparaisons par paires requises sous cette stratégie de classement selon un scénario de classement cohérent est une variable aléatoire à valeur entière ayant le support $H-1, H, H+1, \ldots, \binom{H}{2}$. Le nombre prévu de comparaisons par paires pour ce scénario de classement est approximativement égal à $f(H) = (H+2)(H-1)/4.$ Le lecteur consultera Nahhas et coll. (2002) pour un exposé plus approfondi sur le nombre prévu de comparaisons par paires.

Nous présentons maintenant les définitions de coût pour trois modèles, à savoir les modèles de classement concomitant, de classement visuel et d'échantillonnage aléatoire simple : C_C = coût total pour le classement concomitant, C_V = coût total pour le classement visuel, C_S = coût total pour l'échantillonnage aléatoire simple, c_i = coût de l'échantillonnage d'une seule unité, c_{qy} = coût de la quantification de la variable d'intérêt (Y) pour une unité, c_{qx} = coût de la quantification de la variable concomitante (auxiliaire X) pour une unité, c_r = coût d'une comparaison par paires. Nous supposons que le coût indirect de l'échantillonnage aléatoire simple est nul, mais que le coût indirect (en sus du coût indirect de l'EAS) du modèle de classement concomitant (visuel) de l'EEO est absorbé dans $c_{qx}(c_r)$. Les coûts totaux pour ces trois modèles sont donnés par

$$C_S = n_s(c_i + c_{qy}),\ C_C = n_c(Hc_i + Hc_{qx} + c_{qy}),\ C_V = n_V(Hc_i + f(H) + c_{qy}),$$

où n_S, n_C et n_V sont le nombre total d'observations (mesurées) dans les modèles d'EAS, d'EEO avec classement concomitant et d'EEO avec classement visuel. Le lecteur consultera Nahhas et coll. (2002) pour des renseignements plus détaillés sur ces fonctions de coût.

Nous fixons maintenant le coût total sur ces trois modèles $C_S = C_C = C_V = C$. Sous ce coût fixé, nous examinons l'efficacité relative de $\overline{Y}_{\mathrm{CC}}$ et $\overline{Y}_{\mathrm{CV}}$ comparativement à la moyenne d'échantillon EAS $\overline{Y}_{\mathrm{EAS}}$. Soit

$$\mathrm{PR} = \frac{1}{1-D}, \qquad D = 1 - \frac{1}{H\sigma^2}\sum_{h=1}^{H}(\mu_{[h]} - \mu)^2,$$

où PR est la précision relative de la moyenne d'échantillon EEO par rapport à la moyenne d'échantillon EAS en population infinie. Sous le modèle de superpopulation, nous pouvons établir le résultat d'efficacité qui suit.

Théorème 3 : *Soit* $Y_{[h]i}$, $h = 1, \ldots, H$, $i = 1, \ldots, d$, *un échantillon EEO provenant d'une population finie* P^N. *Pour un coût fixé, sous un modèle de superpopulation et un scénario de classement cohérent, les résultats d'efficacité qui suivent sont établis.*

$$\mathrm{ER}\left(\overline{Y}_{\mathrm{CC}}, \overline{Y}_{\mathrm{EAS}}\right) = \frac{\mathrm{Var}\left(\overline{Y}_{\mathrm{EAS}}\right)}{\mathrm{Var}\left(\overline{Y}_{\mathrm{CC}}\right)} \geq 1, \qquad \text{si PR} \geq \frac{Hc_i + Hc_{qx} + c_{qy}}{c_i + c_{qy}}$$

$$\mathrm{ER}\left(\overline{Y}_{\mathrm{CV}}, \overline{Y}_{\mathrm{EAS}}\right) = \frac{\mathrm{Var}\left(\overline{Y}_{\mathrm{EAS}}\right)}{\mathrm{Var}\left(\overline{Y}_{\mathrm{CV}}\right)} \geq 1, \qquad \text{si PR} \geq \frac{Hc_i + f(H)\,c_r + c_{qy}}{c_i + c_{qy}}.$$

Les fractions dans le deuxième membre des inégalités du théorème susmentionné représentent le ratio du coût de la sélection et de la mesure d'une seule unité sous EEO et sous EAS, respectivement. Si le coût d'échantillonnage d'une unité et le coût de classement d'un ensemble sont négligeables (gratuits), le ratio des coûts devient 1. L'une des hypothèses de base, dans les conditions où l'EEO est utilisé, est que le coût du classement des unités est relativement faible comparativement au coût de la mesure. Donc, il n'est pas déraisonnable de supposer que le ratio des coûts sera très proche de 1 pour les conditions où l'utilisation de l'EEO est appropriée. Il est établi dans la littérature que la précision relative (PR) est toujours supérieure ou égale à 1 (voir Dell et Clutter (1972), Patil et coll. (1997), Nahhas et coll. (2002)). Elle est égale à 1 uniquement sous classement aléatoire. Les valeurs de PR dans le cas d'une population normale pour différentes valeurs de ρ (coefficient de corrélation entre la réponse Y et la variable auxiliaire X) et de tailles d'ensemble sont données au tableau 4.1. Il est maintenant raisonnable de dire que l'estimateur EEO sous un modèle de superpopulation est plus efficace si le coût de l'échantillonnage et du classement d'une unité est relativement bas comparativement au coût de la mesure.

Tableau 4.1
Précision relative (PR) de la moyenne d'échantillon EEO par rapport à la moyenne d'échantillon EAS en population infinie pour une distribution normale $N(0,1)$. ρ est le coefficient de corrélation entre la réponse et la variable auxiliaire, et H est la taille de l'ensemble

ρ	$H = 2$	$H = 3$	$H = 4$	$H = 5$	$H = 6$
1,00	1,467	1,914	2,347	2,770	3,186
0,90	1,347	1,631	1,869	2,073	2,251
0,75	1,218	1,367	1,477	1,561	1,628
0,50	1,086	1,136	1,168	1,190	1,207

5 Résultats empiriques

À la présente section, nous réalisons une étude en simulation pour vérifier les propriétés en échantillon fini de l'estimateur pour différentes valeurs des paramètres de simulation. Les ensembles de données sont générés à partir de superpopulations normales $(\mu = 10, \sigma = 4)$ et lognormales $(\mu = 0, \sigma = 1)$. Nous considérons deux populations finies de tailles différentes, $N = 150$ et $N = 1\,000$, pour examiner l'effet des tailles de population sur les estimateurs. Les combinaisons de tailles d'échantillon et d'ensemble (n, H) choisies sont (10, 2), (15, 3), (20, 4), (25, 5). La qualité de l'information de classement est modélisée à l'aide d'un modèle d'erreur perceptuelle donné dans Dell et Clutter (1972). Le modèle de Dell et Clutter contient deux variables, la variable d'intérêt Y et une variable de classement corrélée X. La variable de classement est modélisée au moyen d'un modèle additif $X = Y + \epsilon$, où ϵ est un bruit aléatoire généré indépendamment de Y. Pour appliquer le modèle d'erreur perceptuelle, nous générons un ensemble (de taille H) d'échantillons aléatoires simples, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_H)$, de moyenne μ et de variance σ^2 à partir de la vraie population d'intérêt. Un autre ensemble (de taille H) de nombres aléatoires est généré à partir d'une distribution normale de moyenne nulle et de variance σ_ϵ^2, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_H)$. Le modèle d'erreur perceptuelle est alors défini par $X_i = Y_i + \epsilon_i \quad i = 1, 2, \ldots, H$. Les nombres aléatoires (X_i, Y_i) sont classés par ordre en fonction des premières composantes $(X_{(i)})$ et les secondes composantes sont considérées comme étant les statistiques d'ordre classées au jugé $(Y_{[i]})$. La qualité de l'information de classement est contrôlée par le coefficient de corrélation entre Y et X, $\rho = \text{corr}(Y, X) = \left(\frac{\sigma^2}{\sigma^2 + \sigma_\epsilon^2}\right)^{1/2}$. Puisque les unités sont classées en se basant sur la variable concomitante X, le modèle de classement équivaut au classement concomitant de la section 3. Dans l'étude en simulation, nous avons utilisé $\rho = 1$ pour un classement parfait et $\rho = 0{,}75; 0{,}50$ pour un classement imparfait.

Dans chaque réplique de la simulation, une population finie de taille P^N est générée à partir d'une superpopulation normale de moyenne μ et d'écart-type σ spécifiés, $P^N = \{y_1, \ldots, y_N\}$. Un échantillon EEO est alors construit à partir de cette population finie, une réalisation issue de la superpopulation normale, avec des tailles d'ensemble et de cycle spécifiées. La qualité de l'information de classement dans chaque échantillon EEO est contrôlée par génération d'un vecteur de bruits aléatoires $\boldsymbol{\epsilon}$ avec ρ (ou, de façon équivalente σ_ϵ) spécifié dans le modèle d'erreur perceptuelle. La taille de la simulation est de 50 000.

Les résultats de la simulation sont présentés aux tableaux 5.1, 5.2, 5.3 et 5.4. Plusieurs aspects de ces tableaux doivent être discutés. Pour différentes combinaisons de valeurs de ρ et de tailles d'échantillon (n, H), les efficacités relatives de l'estimateur EEO comparativement à l'estimateur EAS sont données par

$$\text{ER}_{\text{CC}} = \frac{V\left(\bar{Y}_{\text{EAS}}\right)}{V\left(\bar{Y}_{\text{CC}}\right)} \tag{5.1}$$

où $V\left(\bar{Y}_{\text{EAS}}\right)$ et $V\left(\bar{Y}_{\text{CC}}\right)$ sont les EPQM des moyennes d'échantillon EAS et d'échantillon EEO provenant de l'étude en simulation sous le modèle de superpopulation de l'équation (2.1), respectivement. Dans l'équation (5.1), les valeurs de l'efficacité relative (ER_{CC}) plus grandes que 1 indiquent que l'estimateur EEO est plus efficace que l'estimateur EAS. Dans tous ces tableaux, l'estimateur de la moyenne d'échantillon EEO donne de meilleurs résultats que l'estimateur EAS. Son efficacité est une fonction croissante de la taille d'ensemble H et du coefficient de corrélation ρ, comme prévu. Sous le modèle de

coût concomitant, si le ratio des coûts de l'obtention d'une unité sous EEO et d'une unité sous EAS est inférieur aux valeurs de PR dans le tableau 4.1, l'efficacité de la moyenne d'échantillon EEO est supérieure à celle de la moyenne d'échantillon EAS.

L'effet de la taille de population finie N peut être observé en comparant les résultats d'efficacité dans les tableaux 5.1 et 5.2 pour la superpopulation normale, et dans les tableaux 5.3 et 5.4 pour la superpopulation lognormale. Quand $\rho >$ 0,50, les efficacités relatives $(\mathrm{ER}_{\mathrm{CC}})$ sont plus élevées dans le tableau 5.1 $(N = 150)$ que dans le tableau 5.2 $(N = 1\,000)$. Dans le tableau 5.1, le facteur de correction pour population finie est plus petit que ce facteur dans le tableau 5.2. Donc, la réduction de l'EPQM est plus faible dans l'estimateur EEO. Un effet similaire est également observé dans les tableaux 5.3 et 5.4.

L'étude en simulation visait aussi à étudier les propriétés de l'estimateur de l'EPQM de l'estimateur de la moyenne d'échantillon EEO. La valeur théorique de l'estimateur de l'EPQM est donnée sous l'en-tête σ^2_{EEO} quand $\rho =$ 1,0. L'estimation simulée (sans biais) de l'EPQM est donnée dans la cinquième (sixième) colonne dans les tableaux 5.1 à 5.4. Il est très clair que les estimations simulées et sans biais de l'EPQM sont presque identiques quand $\rho \neq 1$, comme prévu. Sous un classement parfait $(\rho = 1)$, les valeurs théoriques de l'EPQM, ainsi que les estimations simulées et sans biais de l'EPQM sont toutes proches les unes des autres dans les variations de la simulation.

Les probabilités de couverture des intervalles de confiance sont données sous l'en-tête $C(\overline{Y}_{\mathrm{CC}})$ dans la septième colonne des tableaux 5.1 à 5.4. Dans les tableaux 5.1 et 5.2, les probabilités de couverture des intervalles de confiance fondés sur l'approximation t sont raisonnablement proches de la probabilité de couverture nominale de 0,950. Par contre, dans les tableaux 5.3 et 5.4, les probabilités de couverture sont plus petites que la probabilité de couverture nominale de 0,95 pour la superpopulation lognormale. Les probabilités de couverture se rapprochent des valeurs nominales quand la taille d'échantillon augmente, ce qui indique que, pour des populations asymétriques, les tailles d'échantillon devraient être suffisamment grandes pour obtenir une probabilité de couverture raisonnable pour les intervalles de confiance.

Tableau 5.1
Estimation de l'EPQM et efficacité relative de l'estimateur sur échantillon EEO, et probabilité de couverture d'un intervalle de confiance à 95 % d'une moyenne de population. Les ensembles de données sont générés à partir d'une superpopulation normale avec $\mu = 10$, $\sigma = 4$ et une taille de population $N = 150$

		Est. par équations		Est. par simul.	Est. sans biais	Prob. de couverture	Eff. relative
H	ρ	σ^2_{EEO}	σ^2_{EAS}	$V(\overline{Y}_{\mathrm{CC}})$	$\hat{\sigma}^2_{\mathrm{EEO}}$	$C(\overline{Y}_{\mathrm{CC}})$	$\mathrm{ER}_{\mathrm{CC}}$
2,0	0,50	-	1,493	1,355	1,365	0,949	1,102
3,0	0,50	-	0,960	0,840	0,833	0,947	1,143
4,0	0,50	-	0,693	0,572	0,578	0,948	1,213
5,0	0,50	-	0,533	0,435	0,432	0,948	1,226
2,0	0,75	-	1,493	1,195	1,205	0,949	1,250
3,0	0,75	-	0,960	0,675	0,674	0,947	1,423
4,0	0,75	-	0,693	0,433	0,436	0,946	1,600
5,0	0,75	-	0,533	0,302	0,304	0,945	1,768
2,0	1,00	0,984	1,493	0,974	0,984	0,948	1,534
3,0	1,00	0,451	0,960	0,455	0,451	0,940	2,111
4,0	1,00	0,234	0,693	0,233	0,235	0,936	2,971
5,0	1,00	0,124	0,533	0,125	0,126	0,922	4,273

Tableau 5.2
Estimation de l'EPQM et efficacité relative de l'estimateur sur échantillon EEO, et probabilité de couverture d'un intervalle de confiance à 95 % d'une moyenne de population. Les ensembles de données sont générés à partir d'une superpopulation normale avec $\mu = 10$, $\sigma = 4$ et une taille de population $N = 1\,000$

		Est. par équations		Est. par simul.	Est. sans biais	Prob. de couverture	Eff. relative
H	ρ	σ^2_{EEO}	σ^2_{EAS}	$V\left(\bar{Y}_{\text{CC}}\right)$	$\hat{\sigma}^2_{\text{EEO}}$	$C\left(\bar{Y}_{\text{CC}}\right)$	ER_{CC}
2,0	0,50	-	1,584	1,461	1,455	0,950	1,084
3,0	0,50	-	1,051	0,931	0,924	0,949	1,129
4,0	0,50	-	0,784	0,665	0,670	0,950	1,180
5,0	0,50	-	0,624	0,524	0,522	0,950	1,191
2,0	0,75	-	1,584	1,304	1,295	0,949	1,215
3,0	0,75	-	1,051	0,770	0,765	0,948	1,365
4,0	0,75	-	0,784	0,525	0,526	0,951	1,494
5,0	0,75	-	0,624	0,392	0,395	0,951	1,590
2,0	1,00	1,075	1,584	1,075	1,076	0,950	1,473
3,0	1,00	0,541	1,051	0,538	0,541	0,951	1,954
4,0	1,00	0,325	0,784	0,327	0,325	0,949	2,398
5,0	1,00	0,215	0,624	0,217	0,215	0,948	2,877

Tableau 5.3
Estimation de l'EPQM et efficacité relative de l'estimateur sur échantillon EEO, et probabilité de couverture d'un intervalle de confiance à 95 % d'une moyenne de population. Les ensembles de données sont générés à partir d'une superpopulation lognormale avec $\mu = 0$, $\sigma = 1$ et une taille de population $N = 150$

		Est. par équations		Est. par simul.	Est. sans biais	Prob. de couverture	Eff. relative
H	ρ	σ^2_{EEO}	σ^2_{EAS}	$V\left(\bar{Y}_{\text{CC}}\right)$	$\hat{\sigma}^2_{\text{EEO}}$	$C\left(\bar{Y}_{\text{CC}}\right)$	ER_{CC}
2,0	0,50	-	0,436	0,400	0,400	0,852	1,089
3,0	0,50	-	0,280	0,243	0,242	0,869	1,153
4,0	0,50	-	0,202	0,160	0,162	0,883	1,262
5,0	0,50	-	0,156	0,117	0,116	0,886	1,336
2,0	0,75	-	0,436	0,371	0,372	0,855	1,176
3,0	0,75	-	0,280	0,216	0,217	0,867	1,300
4,0	0,75	-	0,202	0,146	0,146	0,874	1,388
5,0	0,75	-	0,156	0,103	0,103	0,878	1,514
2,0	1,00	0,362	0,436	0,361	0,364	0,839	1,207
3,0	1,00	0,201	0,280	0,197	0,198	0,849	1,423
4,0	1,00	0,128	0,202	0,128	0,127	0,847	1,586
5,0	1,00	0,086	0,156	0,085	0,085	0,845	1,833

Tableau 5.4
Estimation de l'EPQM et efficacité relative de l'estimateur sur échantillon EEO, et probabilité de couverture d'un intervalle de confiance à 95 % d'une moyenne de population. Les ensembles de données sont générés à partir d'une superpopulation lognormale avec $\mu = 0$, $\sigma = 1$ et une taille de population $N = 1\,000$

		Est. par équations		Est. par simul.	Est. sans biais	Prob. de couverture	Eff. relative
H	ρ	σ^2_{EEO}	σ^2_{EAS}	$V\left(\bar{Y}_{\text{CC}}\right)$	$\hat{\sigma}^2_{\text{EEO}}$	$C\left(\bar{Y}_{\text{CC}}\right)$	ER_{CC}
2,0	0,50	-	0,462	0,432	0,433	0,851	1,070
3,0	0,50	-	0,307	0,263	0,263	0,868	1,164
4,0	0,50	-	0,229	0,189	0,190	0,882	1,208
5,0	0,50	-	0,182	0,141	0,141	0,889	1,296
2,0	0,75	-	0,462	0,413	0,413	0,852	1,119
3,0	0,75	-	0,307	0,240	0,238	0,868	1,276
4,0	0,75	-	0,229	0,171	0,170	0,878	1,337
5,0	0,75	-	0,182	0,129	0,129	0,884	1,415
2,0	1,00	0,389	0,462	0,387	0,386	0,839	1,195
3,0	1,00	0,228	0,307	0,225	0,227	0,852	1,364
4,0	1,00	0,154	0,229	0,155	0,155	0,857	1,479
5,0	1,00	0,113	0,182	0,113	0,113	0,862	1,614

6 Exemple

À la présente section, nous appliquons les estimateurs proposés à un ensemble de données sur une population de moutons élevés dans une ferme de recherche de l'Université Atatürk, à Erzurum, en Turquie. L'ensemble de données contient le poids à la naissance, le poids de la mère au moment de l'accouplement et le poids sept mois après la naissance pour 224 agneaux. L'ensemble de données complet est présenté dans Hollander, Wolfe et Chicken (2014, page 709). La variable d'intérêt est le poids (Y) sept mois après la naissance pour 224 agneaux. Nous utilisons le poids à la naissance (X_1) et le poids de la mère au moment de l'accouplement (X_2) comme variables auxiliaires pour effectuer le processus de classement. Les variables de classement sont corrélées positivement à la variable d'intérêt Y. Le coefficient de corrélation $(\rho = \text{corr}(X, Y))$ entre X_1, Y et X_2, Y est de 0,8425 et 0,5941, respectivement. L'histogramme de la variable d'intérêt, Y, est approximativement symétrique. La moyenne et la variance de Y sont $\overline{Y}_N$ = 28,125 kg et S_N^2 = 15,23 kg^2, respectivement, où $S_n^2 = \sum_{i=1}^{224} (Y_i - \overline{Y}_N)^2 \big/ 223$. Nous avons traité ces 224 agneaux comme une réalisation d'une superpopulation de moyenne μ et de variance σ^2 finies. Nous avons construit des échantillons selon un plan d'échantillonnage EEO en utilisant cette population finie. Les échantillons sont générés pour les combinaisons de tailles d'échantillon et d'ensemble (n, H), (10, 2), (15, 3), (20, 4), (25, 5). La taille de la simulation est de 50 000.

Dans cet exemple, nous intégrons le coût d'échantillonnage dans les plans d'échantillonnage EEO et EAS, avec classement concomitant dans le cas de l'EEO. Nous devons d'abord déterminer des coûts raisonnables associés aux divers aspects de l'échantillonnage EEO. Les mesures de poids sont obtenues pour des agneaux âgés de sept mois. Ces animaux sont très actifs et le coût de la mesure est élevé. Le processus de mesure nécessite habituellement trois personnes pour séparer l'agneau du troupeau, l'amener à la balance et le tenir fermement durant la mesure. Supposons que le salaire annuel des travailleurs employés par la ferme soit de 50 000 $. Cela correspond à un taux salarial d'environ 25 $ de l'heure par personne. Supposons que la mesure d'un agneau prend environ 5 minutes. Le coût de la mesure pour un agneau sera alors environ $c_{qy} = 3(25/12) \approx 6$ pour trois travailleurs. Le classement sera effectué en utilisant les variables auxiliaires X_1 et X_2. Ces variables sont tenues à jour dans la base de données pour d'autres usages. Le seul coût d'échantillonnage correspondra au coût en personnel pour le classement. Celui-ci sera effectué dans le bureau en sélectionnant des ensembles au hasard dans la base de données et en les classant en se basant sur les variables auxiliaires. Supposons que le classement d'un ensemble de taille H prend environ une demi-minute. Cela donne un coût de classement de Hc_{qx} = 0,21 $. Nous pouvons supposer que le coût associé à l'identification d'une unité dans la population est négligeable $(c_i = 0)$. Sous ces spécifications, le ratio des coûts de sélection et de mesure d'une unité sous EEO et sous EAS est donné par ratio $= (Hc_i + Hc_{qx} + c_{qy})/(c_i + c_{qy}) = (6 + 0{,}21)/6 = 1{,}035$. Puisque ce ratio est inférieur à toutes les entrées dans le tableau 4.1, nous nous attendons à ce que l'efficacité de $\overline{Y}_{\text{CC}}$ soit plus grande que celle de $\overline{Y}_{\text{EAS}}$.

Le tableau 6.1 présente l'EPQM estimée et l'efficacité relative de l'estimateur EEO, ainsi que la probabilité de couverture de l'intervalle de confiance de μ pour différentes combinaisons de ρ et de taille

d'échantillon. Il est clair que l'estimateur EEO donne de meilleurs résultats que l'estimateur EAS pour toutes les combinaisons de paramètres de simulation. Les estimations de l'EPQM et les probabilités de couverture présentent aussi des comportements similaires à ceux de la section 3. Les valeurs estimées de l'EPQM sont très proches des valeurs simulées. Les probabilités de couverture des intervalles de confiance basés sur l'approximation t semblent être très proches de la probabilité de couverture nominale de 0,950.

Tableau 6.1
Estimation de l'EPQM et efficacité relative de l'estimateur sur échantillon EEO, et probabilité de couverture d'un intervalle de confiance à 95 % de la moyenne de population d'une population de moutons de taille $N = 224$

		Est. par équations	Est. par simul.	Est. sans biais	Prob. de couverture	Eff. relative
H	ρ	σ^2_{EAS}	$V(\bar{Y}_{\text{CC}})$	$\hat{\sigma}^2_{\text{EEO}}$	$C(\bar{Y}_{\text{CC}})$	ER_{CC}
2,0	0,59	1,453	1,279	1,275	0,946	1,136
3,0	0,59	0,946	0,776	0,774	0,948	1,219
4,0	0,59	0,693	0,536	0,537	0,948	1,293
5,0	0,59	0,540	0,399	0,402	0,948	1,353
2,0	0,84	1,453	1,107	1,105	0,945	1,312
3,0	0,84	0,946	0,600	0,602	0,946	1,576
4,0	0,84	0,693	0,377	0,382	0,946	1,839
5,0	0,84	0,540	0,263	0,264	0,944	2,056

7 Observations finales

Nous avons établi une inférence statistique sous un modèle pour la moyenne et le total de population estimés sur des échantillons fondés sur des ensembles ordonnés en population finie, où les échantillons sont construits en utilisant un plan d'échantillonnage sans remise. Il est montré que la moyenne d'échantillon des échantillons EEO est sans biais par rapport au modèle et que son erreur de prédiction quadratique moyenne (EPQM) est plus petite que celle d'une moyenne d'échantillon aléatoire simple. Nous avons construit un estimateur sans biais de l'EPQM et l'intervalle de confiance de prédiction pour la moyenne de population. Une étude en simulation à petite échelle a montré que les estimateurs sont aussi bons, voire meilleurs, que les estimateurs EAS quand la qualité de l'information de classement dans l'échantillonnage EEO est faible ou élevée, respectivement, et que le ratio des coûts de l'obtention d'une unité sous EEO et sous EAS n'est pas trop élevé. Les probabilités de couverture des intervalles de prédiction sont également très proches des probabilités de couverture nominales. Les plans d'échantillonnage proposés et les procédures d'inférence sont appliqués à un ensemble de données sur une population de moutons élevée dans une ferme de recherche agricole.

Remerciements

Les travaux de recherche du deuxième auteur sont financés par *The Scientific and Technological Research Council* de la Turquie (TUBITAK). Les auteurs remercient le rédacteur en chef, le rédacteur associé et deux examinateurs de leurs suggestions et commentaires constructifs.

Annexe

Preuve du théorème 1 : Nous écrivons l'erreur de prédiction quadratique moyenne (EPQM) sous la forme

$$\text{EPQM}_M\left(\bar{Y}_R\right) = E_M\left\{\bar{Y}_R - \frac{1}{N}\sum_{i=1}^{N} Y_i\right\}^2 = E_M\left\{\frac{1}{dH}\sum_{h=1}^{H}\sum_{i=1}^{d} Y_{[h]i} - \frac{1}{N}\sum_{i=1}^{N} Y_i\right\}^2.$$

Soit Z_i, $i = 1, \ldots, N - nH$, les réponses sur $N - nH$ unités de population qui ne sont ni mesurées ni utilisées pour le classement dans aucun des ensembles de taille H sélectionnés aléatoirement dans la construction de l'échantillon EEO. Alors, l'EPQM peut s'écrire

$$\text{EPQM}_M\left(\bar{Y}_R\right) = E_M\left\{\frac{1}{dH}\sum_{h=1}^{H}\sum_{i=1}^{d} Y_{[h]i} - \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{d}\left[Y_{[h]i} + \sum_{h\neq h'}^{H} Y^*_{[h']i}\right] - \frac{1}{N}\sum_{i=1}^{N-nH} Z_i\right\}^2$$

où $Y^*_{[h']i}$, $h' \neq, h,$ sont les réponses sur les unités non mesurées qui sont utilisées pour classer les unités dans un ensemble. Donc, $Y^*_{[h']i}$ et $Y_{[h]i}$ sont corrélées entre elles, mais elles ne sont pas corrélées avec Z_i. Soit

$$c_{h,h'} = \begin{cases} \dfrac{N-n}{n} & h = h' \\ -1 & h \neq h'. \end{cases}$$

En utilisant la définition de $c_{h,h'}$, nous combinons $Y^*_{[h']i}$ et $Y_{[h]i}$ sous la même sommation et écrivons l'EPQM sous la forme

$$\begin{aligned}
\text{EPQM}_M\left(\bar{Y}_R\right) &= \frac{1}{N^2}\text{var}\left\{\sum_{h=1}^{H}\sum_{i=1}^{d}\sum_{h'=1}^{H} c_{h,h'} Y_{[h]i}\right\} + \text{var}\left\{\frac{1}{N}\sum_{i=1}^{N-nH} Z_i\right\} \\
&= \frac{1}{N^2}\sum_{h=1}^{H}\text{var}\left[\sum_{i=1}^{d}\sum_{h'=1}^{H} c_{h,h'} Y_{[h]i}\right] + \text{var}\left\{\frac{1}{N}\sum_{i=1}^{N-nH} Z_i\right\} \\
&= \frac{d}{N^2}\sum_{h=1}^{H}\sum_{h'=1}^{H}\left(c_{h,h'}\right)^2 \sigma^2_{[h']} + \frac{d}{N^2}\sum_{h=1}^{H}\left(\sum_{h'=1}^{H}\sum_{t\neq h'}^{H} c_{h,h'} c_{h,t}\sigma_{[h',t]}\right) \\
&\quad + \text{var}\left\{\frac{1}{N}\sum_{i=1}^{N-nH} Z_i\right\} \\
&= A + B + \frac{(N-nH)\sigma^2}{N^2}. \qquad \text{(A.1)}
\end{aligned}$$

L'expression A se réduit à

$$\begin{aligned}
A &= \frac{d}{N^2}\sum_{h=1}^{H}\left(c_{h,h}\right)^2 \sigma^2_{[h]} + \frac{d}{N^2}\sum_{h}\sum_{h'\neq h}^{H}\left(c_{h,h'}\right)^2 \sigma_{[h,h']} \\
&= \frac{d}{N^2}\left[\left(\frac{N-n}{n}\right)^2 + (H-1)\right]\sum_{h=1}^{H}\sigma^2_{[h]}.
\end{aligned}$$

De façon similaire, l'expression B se réduit à

$$
\begin{aligned}
B &= \frac{d}{N^2}\sum_{h=1}^{H}\left[\sum_{t\neq h',h}^{H}\left(\sum_{h'=1}^{H} c_{h,h'}c_{h,t}\sigma_{[h',t]} + c_{h,h}c_{h,t}\sigma_{[h,t]}\right) + \sum_{h'\neq h}^{H} c_{h,h'}c_{h,h}\sigma_{[h',h]}\right] \\
&= \frac{d}{N^2}\sum_{h=1}^{H}\left[\sum_{h'\neq h}^{H}\sum_{t\neq h',h}^{H}\sigma_{[h',t]} - 2\left(\frac{N-n}{n}\right)\left(\sum_{t=1}^{H}\sigma_{[h,t]} - \sigma_{[h,h]}\right)\right] \\
&= \frac{d}{N^2}\left[(H^2-2H)\sigma^2 - (H-2)\sum_{h=1}^{H}\sigma_{[h]}^2 - 2\left(\frac{N-n}{n}\right)\left(H\sigma^2 - \sum_{h=1}^{H}\sigma_{[h]}^2\right)\right].
\end{aligned}
$$

En insérant les expressions A et B dans l'équation (A.1), nous concluons que

$$
\begin{aligned}
\mathrm{EPQM}_M\left(\overline{Y}_R\right) &= \frac{d}{N^2}\left[\left(\frac{N-n}{n}\right)^2 + (H-1)\right]\sum_{h=1}^{H}\sigma_{[h]}^2 \\
&\quad + \frac{d}{N^2}\left[(H^2-2H)\sigma^2 - (H-2)\sum_{h=1}^{H}\sigma_{[h]}^2 - 2\left(\frac{N-n}{n}\right)\left(H\sigma^2 - \sum_{h=1}^{H}\sigma_{[h]}^2\right)\right] \\
&\quad + \left(\frac{N-nH}{N^2}\right)\sigma^2 \\
&= \left(\frac{N-n}{Nn}\right)\sigma^2 - \frac{1}{nH}\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2
\end{aligned}
$$

ce qui complète la preuve. Notons que, pour établir la dernière égalité, nous avons utilisé le fait que $\sigma^2 = \sum_{h=1}^{H}\sigma_{[h]}^2 \big/ H + \sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 \big/ H$.

Preuve du théorème 2 : Nous nous penchons d'abord sur les espérances de T_1^* et T_2^* sous le modèle de superpopulation de l'équation (2.1)

$$
\begin{aligned}
E\left(T_1^*\right) &= \frac{1}{H}\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 + \frac{H-1}{H^2}\sum_{h=1}^{H}\sigma_{[h]}^2 = \sigma^2 - \frac{1}{H^2}\sum_{h=1}^{H}\sigma_{[h]}^2 \\
E\left(T_2^*\right) &= \frac{1}{H^2}\sum_{h=1}^{H}\sigma_{[h]}^2.
\end{aligned}
$$

Il est maintenant facile d'établir que $E\left(T_1^* + T_2^*\right) = \sigma^2$. La preuve est alors achevée en insérant ces expressions dans l'équation (3.1).

Preuve du théorème 3 : Nous esquissons la preuve pour $\mathrm{ER}\left(\overline{Y}_{\mathrm{CC}}, \overline{Y}_{\mathrm{EAS}}\right)$. Partant de la fonction de coût total, nous écrivons

$$
n_S = \frac{C}{c_i + c_{qy}} \quad \text{et} \quad n_R = \frac{C}{Hc_i + Hc_{qx} + c_{qy}},
$$

où C est le coût total fixé. En utilisant ces expressions, nous avons

$$
\begin{aligned}
\mathrm{ER}\left(\overline{Y}_{\mathrm{CC}}, \mathrm{EAS}\right) &= \frac{n_R\left(N - n_S\right)}{n_S\left(N - n_r - ND\right)} \\
&= \frac{N\left(c_i + c_{qy}\right) - C}{N\left(Hc_i + Hc_{qx} + c_{qy}\right) - C - ND\left(Hc_i + Hc_{qx} + c_{qy}\right)}.
\end{aligned}
$$

Nous établissons maintenant que $\mathrm{ER}\left(\overline{Y}_{\mathrm{CC}}, \mathrm{EAS}\right) \geq 1$ si et seulement si

$$c_i + c_{qy} \geq \left(Hc_i + Hc_{qx} + c_{qy}\right)(1 - D) = \frac{Hc_i + Hc_{qx} + c_{qy}}{\mathrm{PR}}$$

$$\mathrm{PR} \geq \frac{Hc_i + Hc_{qx} + c_{qy}}{c_i + c_{qy}}$$

ce qui achève la preuve.

Bibliographie

Ahn, S., Lim, J. et Wang, X. (2014). The students's *t* approximation to distributions of pivotal statistics from ranked set samples. *Journal of Korean Statistical Society*, 43, 4, 643-652.

Al-Saleh, M.F., et Samawi, H.M. (2007). A note on inclusion probability in ranked set sampling and some of its variations. *Test*, 16, 1, 198-209.

Dell, T.R., et Clutter, J.L. (1972). Ranked-set sampling theory with order statistics background. *Biometrics*, 28, 2, 545-555.

Deshpande, J.V., Frey, J. et Ozturk, O. (2006). Nonparametric ranked set-sampling confidence intervals for a finite population. *Environmental and Ecological Statistics*, 13, 1, 25-40.

Frey, J. (2011). A note on ranked-set sampling using a covariate. *Journal of Statistical Planning and Inference*, 141, 2, 809-816.

Gokpinar, F., et Ozdemir, Y.A. (2010). Generalization of inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 39, 1, 89-95.

Hollander, M., Wolfe, D.A. et Chicken, E. (2014). *Nonparametric Statistical Methods, 3rd Edition*. Wiley, 709.

Nahhas, R.W., Wolfe, D.A. et Chen, H. (2002). Ranked set sampling: Cost and optimal set size. *Biometrics*, 58, 964-971.

Ozdemir, Y.A., et Gokpinar, F. (2007). A generalized formula for inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 36, 1, 89-99.

Ozdemir, Y.A., et Gokpinar, F. (2008). A new formula for inclusion probabilities in median ranked set sampling. *Communications in Statistics - Theory and Methods*, 37, 13, 2022-2033.

Ozturk, O. (2014). Estimation of population mean and total in finite population setting using multiple auxilary variables. *Journal of Agricultural, Biological and Environmental Statistics*, 19, 2, 161-184.

Ozturk, O. (2015). Distribution free two-sample methods for judgment-post stratified data. *Statistica Sinica*, 25, 1691-1712.

Ozturk, O. (2016a). Estimation of finite population mean and total using population ranks of sample units. *Journal of Agricultural, Biological and Environmental Statistics*, 21, 1, 181-202.

Ozturk, O. (2016b). Inférence statistique fondée sur des échantillons poststratifiés par choix raisonné en population finie. *Techniques d'enquête*, 42, 2, 255-280. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2016002/article/14664-fra.pdf.

Ozturk, O., et Jozani, M.J. (2013). Inclusion probabilities in partially rank ordered set sampling. *Computational Statistics & Data Analysis*, 69, 122-132.

Patil, G.P., Sinha, A.K. et Taillie, C. (1995). Finite population corrections for ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 47, 4, 621-636.

Patil, G.P., Sinha, A.K. et Taillie, C. (1997). Ranked set sampling, coherent rankings and size-biased permutations. *Journal of Statistical Planning and Inference*, 63, 2, 311-324.

Takahasi, K., et Futatsuya, M. (1998). Dependence between order statistics in samples from finite population and its application to ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 50, 1, 49-70.

Techniques d'enquête, juin 2018
Vol. 44, nº 1, pp. 19-44
Statistique Canada, nº 12-001-X au catalogue

Linéarisation contre Bootstrap pour estimer la variance de l'évolution de l'indice de Gini

Guillaume Chauvet et Camelia Goga[1]

Résumé

Le présent article traite de l'estimation de la variance par linéarisation ou par bootstrap pour l'indice de Gini, et pour l'évolution de cet indice entre deux périodes. Dans le cas d'un seul échantillon, nous adoptons l'approche de linéarisation par la fonction d'influence proposée par Deville (1999), la méthode du bootstrap sans remise proposée par Gross (1980) pour l'échantillonnage aléatoire simple sans remise, et la méthode de tirage avec remise des unités primaires d'écrite dans Rao et Wu (1988) pour l'échantillonnage à plusieurs degrés. Pour obtenir un estimateur de variance dans le cas de deux échantillons, nous utilisons la technique de linéarisation au moyen de fonctions d'influence partielles (Goga, Deville et Ruiz-Gazen, 2009). Nous élaborons aussi une extension des procédures bootstrap étudiées à l'échantillonnage bidimensionnel. Les deux approches sont comparées sur des données simulées.

Mots-clés : Estimateur composite; estimateur de Horvitz-Thompson; fonction d'influence; estimateur « intersection »; poids de rééchantillonnage; sondage à deux échantillons; plan d'échantillonnage bidimensionnel; estimateur « union »; estimation de la variance.

1 Introduction

L'indice de Gini (Gini, 1914) est l'une des mesures de concentration les plus connues dont l'utilisation est fréquente dans les études économiques. Si $\mathcal{Y}_1$ désigne une variable quantitative positive, telle que le revenu, et $F_1(\cdot)$, sa fonction de répartition définie sur $]-\infty, \infty[$, l'indice de Gini est donné par

$$G_1 = \frac{1}{2}\frac{\iint |v-u|\,dF(u)\,dF(v)}{\int u dF(u)},$$

en supposant que $\int u dF(u) \neq 0$. L'indice de Gini mesure la dispersion d'une variable quantitative positive à l'intérieur d'une population. Habituellement, les instituts de statistique se servent de l'indice de Gini pour évaluer les inégalités de revenu dans un pays à différentes périodes, ou entre différents pays à la même période. Au cours des dernières décennies, l'indice de Gini a également été utilisé dans les domaines de l'économie et de la sociodémographie (voir, par exemple, Navarro, Muntaner, Borrell, Benach, Quiroga, Rodriguez-Sanz, Vergès et Pasarin, 2006; Bhattacharya, 2007; Lai, Huang, Risser et Kapadia, 2008; Barrett et Donald, 2009), de la biologie (Graczyk, 2007), de l'environnement (Druckman et Jackson, 2008; Groves-Kirkby, Denman et Phillips, 2009) ou de l'astrophysique (Lisker, 2008).

Il existe une abondante littérature sur l'estimation de la variance de l'indice de Gini pour des observations tirées de données de sondage; pour une revue, voir Langel et Tillé (2013). Glasser (1962) et Sandström, Wretman et Waldèn (1985) ont étudié le cas de l'échantillonnage aléatoire simple. Sandström, Wretman et Waldèn (1988) ont dressé la liste d'estimateurs de variance possibles pour un plan d'échantillonnage général, y compris un estimateur de variance par le jackknife. Cette dernière approche a été étudiée plus en

1. Guillaume Chauvet, Université Rennes, ENSAI, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France. Courriel : chauvet@ensai.fr; Camelia Goga, Laboratoire des Mathématiques de Besançon, Université de Bourgogne Franche-Comté, UMR 6623, Besançon Cedex, France. Courriel : camelia.goga@univ-fcomte.fr.

détail par Yitzhaki (1991), Karagiannis et Kovačević (2000), et Berger (2008). Kovačević et Binder (1997) ont exploré l'estimation de la variance par linéarisation, et Berger (2008) a démontré l'équivalence entre la linéarisation et une technique de jackknife généralisée que Campbell (1980) avait été le premier à suggérer. Qin, Rao et Wu (2010) ont proposé des intervalles de confiance fondés sur le bootstrap et sur la vraisemblance empirique pour l'indice de Gini. Ils ont procédé à l'étude théorique ainsi qu'empirique de ces méthodes dans le cas particulier de l'échantillonnage aléatoire simple avec remise stratifié. En revanche, pour l'évolution de l'indice de Gini, l'estimation de la variance par bootstrap n'a pas été comparée à d'autres méthodes.

Dans le présent article, nous examinons la linéarisation comparativement au bootstrap pour estimer l'évolution de l'indice de Gini. La présentation de l'article est la suivante. À la section 2, nous examinons d'abord l'estimation de l'indice de Gini dans le cas d'un seul échantillon. La notation est définie à la section 2.1, et l'estimateur par substitution de l'indice de Gini est présenté à la section 2.2. L'estimateur de variance par linéarisation est donné à la section 2.3, avec une application à un plan d'échantillonnage aléatoire simple (SI) et à un plan d'échantillonnage à plusieurs degrés. Les principes fondamentaux du bootstrap pondéré sont passés brièvement en revue au début de la section 2.4, et le bootstrap sans remise (BWO pour *Bootstrap Without Replacement*) approprié pour l'échantillonnage SI est présenté à la section 2.4.1, tandis que le bootstrap avec remise des unités primaires d'échantillonnage (BWR pour *Bootstrap With Replacement*) approprié pour l'échantillonnage à plusieurs degrés est présenté à la section 2.4.2. À la section 3, nous considérons l'estimation de l'évolution de l'indice de Gini dans le cas de deux échantillons. La notation est définie à la section 3.1, et les principes de l'estimation composite qui est appliquée sont examinés brièvement à la section 3.1.1 pour le plan SI bidimensionnel (SI2), et à la section 3.1.2, pour un plan d'échantillonnage à deux degrés bidimensionnel (MULT2). L'estimateur composite de la variation de l'indice de Gini est présenté à la section 3.2. L'estimateur de variance par linéarisation au moyen des fonctions d'influence partielles est donné à la section 3.3, avec application au plan SI2 et au plan MULT2. Les extensions du BWO pour le plan SI2 et du BWR pour le plan MULT2 sont présentées à la section 3.4. La méthode de linéarisation et les méthodes bootstrap proposées sont comparées à la section 4 au moyen d'une étude en simulation. Les conclusions sont présentées à la section 5.

2 Le cas d'un seul échantillon

2.1 Notation

Soit U une population finie de taille N dont les unités peuvent être identifiées par les étiquettes $k = 1, \ldots, N$. Supposons que la variable $\mathcal{Y}_1$ est mesurée sur la population U, et soit $y_{11}, \ldots, y_{1N}$ les valeurs prises par $\mathcal{Y}_1$ sur les unités de la population. Soit $M_1 = \sum_{k \in U} \delta_{y_{1k}}$ la mesure discrète prenant une masse unitaire en tout point y_{1k} de la population et 0 ailleurs, où $\delta_{y_{1k}}$ est la masse de Dirac au point y_{1k}. La plupart des paramètres d'intérêt θ_1 étudiés dans les sondages peuvent s'écrire sous la forme d'une fonction T de M_1, à savoir $\theta_1 = T(M_1)$. Par exemple, le total $t_{y1} = \sum_{k \in U} y_{1k}$ est égal à $\int \mathcal{Y}_1 dM_1$. En pratique, un échantillon s (avec ou sans répétition) est sélectionné selon un plan d'échantillonnage $p(\cdot)$, et nous observons les valeurs y_{1k} pour $k \in s$ uniquement. Un principe de substitution est utilisé pour l'estimation

(voir Deville, 1999, et Goga, Deville et Ruiz-Gazen, 2009). Soit π_k l'espérance du nombre de tirages pour l'unité k dans l'échantillon; dans le cas d'un échantillonnage sans remise, il s'agit de la probabilité que l'unité k soit sélectionnée dans l'échantillon. Soit $\hat{M}_1 = \sum_{k\in s} w_k \delta_{y_{1k}}$ la mesure discrète prenant la masse w_k en tout point dans l'échantillon et 0 ailleurs, où $w_k = \pi_k^{-1}$ est le poids d'échantillonnage. La substitution de $\hat{M}_1$ dans θ_1 donne l'estimateur $\hat{\theta}_1 = T(\hat{M}_1)$.

Pour un plan d'échantillonnage sans remise, l'estimateur d'un total par substitution correspond à l'estimateur de Horvitz-Thompson (HT) $\hat{t}_{y1}^{\mathrm{HT}} = \sum_{k\in s} w_k y_{1k}$. L'estimateur de variance Horvitz-Thompson est

$$v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right) = \sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_{1k}}{\pi_k} \frac{y_{1l}}{\pi_l}, \tag{2.1}$$

où $\pi_{kl} = \Pr(k, l \in s)$ désigne la probabilité que les unités k et l soient sélectionnées conjointement dans l'échantillon, et $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Dans le cas particulier de l'échantillonnage aléatoire simple sans remise (SI) de taille n, nous avons $\hat{t}_{y1}^{\mathrm{HT}} = N\bar{y}_{1,s}$, avec $\bar{y}_{1,s} = n^{-1}\sum_{k\in s} y_{1k}$, et la formule (2.1) donne

$$v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right) S_{y_1,s}^2 \quad \text{où} \quad S_{y_1,s}^2 = \frac{1}{n-1}\sum_{k\in s}(y_{1k} - \bar{y}_{1,s})^2. \tag{2.2}$$

Pour un plan d'échantillonnage avec remise, l'estimateur par substitution d'un total correspond à l'estimateur de Hansen-Hurwitz (HH) $\hat{t}_{y1}^{\mathrm{HH}} = \sum_{k\in s} w_k y_{1k}$. Nous considérons le cas important de l'échantillonnage à plusieurs degrés, où les N unités sont groupées à l'intérieur de N_I unités primaires d'échantillonnage (UPE) disjointes $U_1, \ldots, U_{N_I}$, et où un échantillon s_I de taille m est sélectionné avec remise au premier degré. Soit π_{Ii} l'espérance du nombre de tirages pour l'UPE U_i dans s_I. Un échantillon de deuxième degré s_i est ensuite sélectionné à l'intérieur de toute unité $i \in s_I$ selon un plan d'échantillonnage $p_i(\cdot)$. Soit $\pi_{k|i}$ l'espérance du nombre de tirages pour l'unité k dans s_i. La mesure estimée est alors $\hat{M}_1 = \sum_{i\in s_I}\sum_{k\in s_i} \pi_{Ii}^{-1}\pi_{k|i}^{-1}\delta_{y_{1k}}$. Nous avons $\hat{t}_{y1}^{\mathrm{HH}} = \sum_{i\in s_I} \pi_{Ii}^{-1}\hat{Y}_i$ où $\hat{Y}_i = \sum_{k\in s_i} \pi_{k|i}^{-1} y_{1k}$, et un estimateur de variance sans biais pour $\hat{t}_{y1}^{\mathrm{HH}}$ est donné par

$$v^{\mathrm{HH}}\left(\hat{t}_{y1}^{\mathrm{HH}}\right) = \frac{m}{m-1}\sum_{i\in s_I}\left(\frac{\hat{Y}_i}{\pi_{Ii}} - \frac{\hat{t}_{y1}^{\mathrm{HH}}}{m}\right)^2. \tag{2.3}$$

2.2 Estimation de l'indice de Gini

Si la variable $\mathcal{Y}_1$ est mesurée sur la population U, l'indice de Gini est donné par

$$G_1 = \frac{1}{2}\frac{\sum_{k\in U}\sum_{l\in U}|y_{1k} - y_{1l}|}{N\sum_{k\in U} y_{1k}};$$

voir, par exemple, Nygård et Sandström (1985). Il s'ensuit que G_1 est nul si $\mathcal{Y}_1$ est constante sur la population, ce qui a lieu quand le total de $\mathcal{Y}_1$ est réparti uniformément entre toutes les unités de la population. Dans le cas opposé, quand un seul individu possède le montant total de $\mathcal{Y}_1$, G_1 est maximisé et vaut $1 - 1/N$: le total de $\mathcal{Y}_1$ est alors concentré en un seul point, ce qui signifie que l'inégalité entre les membres de la population est maximale.

Si les individus $k \neq l$ possèdent des valeurs différentes de la variable y_1, l'indice de Gini G_1 égale

$$G_1 = \frac{\sum_{k=1}^{N} y_{1(k)} (2k/N - 1)}{t_{y1}} - \frac{1}{N} = \frac{\sum_{k \in U} y_{1k} \{2F_{1N}(y_{1k}) - 1\}}{t_{y1}} - \frac{1}{N} \quad (2.4)$$

avec $y_{1(1)} \leq \cdots \leq y_{1(N)}$ les valeurs ordonnées et $F_{1N}(\cdot) = N^{-1} \sum_{k \in U} 1_{\{y_{1k} \leq \cdot\}}$ la fonction de répartition en population finie; voir Sandström, Wretman et Waldèn (1988) et Deville (1997) pour des renseignements plus détaillés sur l'obtention de (2.4). Nygård et Sandström (1985) appellent le terme $-1/N$ la correction en population finie de l'indice de Gini et donnent plusieurs raisons d'apporter cette correction, dont la non-négativité de la borne inférieure de G_1. Comme cela se fait fréquemment dans la littérature (voir, par exemple, Glasser, 1962), nous ignorons cette correction dans la suite de l'exposé. Nous redéfinissons l'indice de Gini sous la forme

$$G_1 = \frac{\sum_{k \in U} y_{1k} \{2F_{1N}(y_{1k}) - 1\}}{t_{y1}} = \frac{\int \{2F_{1N}(y) - 1\} y dM_1(y)}{\int y dM_1(y)} \quad (2.5)$$

où la fonction de répartition en population finie $F_{1N}(\cdot)$ est une famille de fonctionnelles

$$F_{1N}(y) = \frac{1}{\int dM_1(y)} \int 1_{\{\xi \leq y\}} dM_1(\xi) \quad (2.6)$$

indexée par y. L'introduction de $\hat{M}_1$ par substitution dans (2.5) et (2.6) donne l'estimateur

$$\hat{G}_1 = \frac{\int \{2\hat{F}_{1N}(y) - 1\} y d\hat{M}_1(y)}{\int y d\hat{M}_1(y)} = \frac{\sum_{k \in s} w_k \{2\hat{F}_{1N}(y_{1k}) - 1\} y_{1k}}{\sum_{k \in s} w_k y_{1k}}, \quad (2.7)$$

où

$$\hat{F}_{1N}(y) = \frac{1}{\int d\hat{M}_1(y)} \int 1_{\{\xi \leq y\}} d\hat{M}_1(\xi) = \frac{1}{\sum_{k \in s} w_k} \sum_{k \in s} w_k 1_{\{y_{1k} \leq y\}} \quad (2.8)$$

est l'estimateur par substitution de la fonction de répartition F_{1N}.

2.3 Estimation de la variance par linéarisation

Nous exposons brièvement ici la linéarisation par la fonction d'influence (LFI) (Deville, 1999), qui consiste à donner un développement d'ordre un de l'estimateur par substitution $\hat{\theta}_1 = T(\hat{M}_1)$ autour de la valeur réelle $\theta_1 = T(M_1)$, afin approximer l'erreur par un estimateur linéaire d'une *variable linéarisée* artificielle. Plus précisément, les dérivées premières de T par rapport à M_1 sont les fonctions d'influence

$$\mathrm{IT}(M_1; y) = \lim_{h \to 0} \frac{T(M_1 + h\delta_y) - T(M_1)}{h},$$

et $u_{1k} = \mathrm{IT}(M_1; y_{1k})$ est la variable linéarisée pour tout $k \in U$. Supposons que $T(\cdot)$ est homogène, c'est-à-dire qu'il existe un nombre positif β dépendant de T tel que $T(rM_1) = r^{\beta} T(M_1)$ pour tout réel $r > 0$. Supposons aussi que $\lim_{N \to \infty} N^{-\beta} T(M_1) < \infty$. Sous certaines hypothèses de régularité supplémentaires concernant $T(\cdot)$ et le plan d'échantillonnage (par exemple, Goga et Ruiz-Gazen, 2014), Deville (1999) établit que

$$\hat{\theta}_1 - \theta_1 = \left(\sum_{k \in s} w_k u_{1k} - \sum_{k \in U} u_{1k} \right) + o_p \left(N^{\beta} n^{-1/2} \right),$$

de sorte que l'erreur $\hat{\theta}_1 - \theta_1$ peut être approximée par l'erreur de l'estimateur HT pour le total de la variable linéarisée u_{1k}. Pour un plan d'échantillonnage sans remise, l'utilisation d'un estimateur fondé sur l'échantillon $\hat{u}_{1k}$ de la variable linéarisée u_{1k} dans l'estimateur HT de variance donne l'estimateur de variance

$$v_{\text{LIN}}^{\text{HT}} \left(\hat{\theta}_1 \right) = \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{u}_{1k}}{\pi_k} \frac{\hat{u}_{1l}}{\pi_l}, \tag{2.9}$$

où $\pi_{kl} = \Pr(k, l \in s)$ désigne la probabilité que les unités k et l soient sélectionnées conjointement dans l'échantillon, et $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Plusieurs résultats de normalité asymptotique ont été prouvés pour des plans d'échantillonnage particuliers; voir Hájek (1960, 1961, 1964), Rosén (1972), Sen (1980), Krewski et Rao (1981), Gordon (1983), Ohlsson (1986, 1989), Chen et Rao (2007), Brändén et Jonasson (2012), Saegusa et Wellner (2013) et Chauvet (2015), entre autres. Si le plan d'échantillonnage est tel que l'estimateur par substitution $\hat{\theta}_1$ satisfait un théorème central limite, un intervalle de confiance de niveau approximatif $(1 - 2\alpha)$ % est $\left[\hat{\theta}_1 - z_\alpha \sqrt{v_{\text{lin}}\left(\hat{\theta}_1\right)}, \hat{\theta}_1 + z_\alpha \sqrt{v_{\text{lin}}\left(\hat{\theta}_1\right)} \right]$, où z_α est le quantile d'ordre α pour la loi normale centrée réduite.

Dans le cas de l'indice de Gini, nous avons $\beta = 0$ et la variable linéarisée est

$$u_{1k} = 2F_{1N}(y_{1k}) \frac{y_{1k} - \bar{y}_{1k,U<}}{t_{y1}} - y_{1k} \frac{G_1 + 1}{t_{y1}} + \frac{1 - G_1}{N}, \tag{2.10}$$

où $\bar{y}_{1k,U<} = \left(\sum_{l \in U} 1_{\{y_{1l} < y_{1k}\}} \right)^{-1} \sum_{j \in U} y_{1j} 1_{\{y_{1j} < y_{1k}\}}$ désigne la moyenne des y_{1j} inférieurs à y_{1k}, voir Deville (1999). Kovačević et Binder (1997) ont obtenu la même expression en appliquant la méthode de linéarisation par les équations estimantes; l'approche de linéarisation de Demnati et Rao (2004) mène aussi au même résultat. La variable linéarisée estimée est

$$\hat{u}_{1k} = 2\hat{F}_{1N}(y_{1k}) \frac{y_{1k} - \bar{y}_{1k,s<}}{\hat{t}_{y1}} - y_{1k} \frac{\hat{G}_1 + 1}{\hat{t}_{y1}} + \frac{1 - \hat{G}_1}{\hat{N}} \tag{2.11}$$

où $\bar{y}_{1k,s<} = \left(\sum_{l \in s} w_l 1_{\{y_{1l} < y_{1k}\}} \right)^{-1} \sum_{j \in s} w_j y_{1j} 1_{\{y_{1j} < y_{1k}\}}$.

Dans le cas particulier de l'échantillonnage SI, l'estimateur de variance par linéarisation pour l'indice de Gini est

$$v_{\text{LIN}}^{\text{HT}} \left(\hat{G}_1 \right) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) S_{\hat{u}_1,s}^2 \quad \text{où} \quad S_{\hat{u}_1,s}^2 = \frac{1}{n-1} \sum_{k \in s} \left(\hat{u}_{1k} - \bar{\hat{u}}_{1,s} \right)^2, \tag{2.12}$$

et où $\bar{\hat{u}}_{1,s} = n^{-1} \sum_{k \in s} \hat{u}_{1k}$. Dans le cas particulier de l'échantillonnage à plusieurs degrés et de l'échantillonnage avec remise des UPE, l'estimateur de variance par linéarisation pour l'indice de Gini est

$$v_{\text{LIN}}^{\text{HH}} \left(\hat{G}_1 \right) = \frac{m}{m-1} \sum_{i \in s_I} \left(\frac{\hat{U}_{1i}}{\pi_{Ii}} - \frac{\hat{t}_{\hat{u}_1}^{\text{HH}}}{m} \right)^2 \quad \text{où} \quad \hat{U}_{1i} = \sum_{k \in s_i} \pi_{k|i}^{-1} \hat{u}_{1k}. \tag{2.13}$$

2.4 Estimation de la variance par bootstrap

L'utilisation de techniques bootstrap dans les sondages a fait l'objet d'un très grand nombre de publications. Les principales techniques bootstrap peuvent être vues comme des cas particuliers du bootstrap pondéré (Bertail et Combris, 1997; Antal et Tillé, 2011; Beaumont et Patak, 2012); consulter aussi Shao et Tu (1995, chapitre 6), Davison et Hinkley (1997, section 3.7), ainsi que Davison et Sardy (2007) pour des études détaillées. Sous une procédure de bootstrap pondéré, la mesure $\hat{M}_1 = \sum_s w_k \delta_{y_k}$ est estimée, conditionnellement à l'échantillon s, par la mesure bootstrap

$$\hat{M}_1^* = \sum_{k \in s} w_k D_k \delta_{y_k} \tag{2.14}$$

où $D = \{D_k\}_{k \in s}$ désigne un vecteur (aléatoire) de poids de rééchantillonnage. Nous notons E_* et V_* l'espérance et la variance par rapport au plan de rééchantillonnage. Dans le cas de l'échantillonnage sans remise, le vecteur D est généré de telle façon que

$$E_*\left(\sum_s w_k D_k y_k\right) \simeq \hat{t}_{y1}^{\mathrm{HT}} \quad \text{et} \quad V_*\left(\sum_s w_k D_k y_k\right) \simeq v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right) \tag{2.15}$$

correspondent approximativement aux deux premiers moments de l'estimateur HT. Dans le cas de l'échantillonnage avec remise, le vecteur D est généré de telle manière que

$$E_*\left(\sum_s w_k D_k y_k\right) \simeq \hat{t}_{y1}^{\mathrm{HH}} \quad \text{et} \quad V_*\left(\sum_s w_k D_k y_k\right) \simeq v^{\mathrm{HH}}\left(\hat{t}_{y1}^{\mathrm{HH}}\right) \tag{2.16}$$

correspondent approximativement aux deux premiers moments de l'estimateur HH.

Sous toute technique de bootstrap pondéré, l'estimateur par substitution (*plug-in*) de $\theta_1 = T(M_1)$ est $\hat{\theta}_1^* = T\left(\hat{M}_1^*\right)$, et la variance de $\hat{\theta}_1 = T\left(\hat{M}_1\right)$ est estimée par

$$V_*\left(\hat{\theta}_1^*\right) = E_*\left\{\hat{\theta}_1^* - E_*\left(\hat{\theta}_1^*\right)\right\}^2. \tag{2.17}$$

Puisque l'estimateur de variance (2.17) peut être difficile à calculer exactement, on peut le remplacer par un estimateur de variance basé sur des simulations. Plus précisément, nous générons C réalisations indépendantes $D_1, \ldots, D_C$ du vecteur D et nous désignons par $\hat{\theta}_{1c}^* = T\left(\hat{M}_{1c}^*\right)$ avec $\hat{M}_{1c}^*$ la mesure bootstrap associée au vecteur D_c. Alors, $V\left(\hat{\theta}_1\right)$ est estimée par

$$v_B\left(\hat{\theta}_1\right) = \frac{1}{C-1}\sum_{c=1}^{C}\left\{\hat{\theta}_{1c}^* - \frac{1}{C}\sum_{c'=1}^{C}\hat{\theta}_{1c'}^*\right\}^2. \tag{2.18}$$

Deux types d'intervalles de confiance sont habituellement calculés. La méthode des percentiles s'appuie sur les estimations bootstrap ordonnées $\hat{\theta}_{(1c)}^*$, $c = 1, \ldots, C$ pour former un intervalle de confiance à $(1 - 2\alpha)$ % noté $[\hat{\theta}_{(1L)}^*, \hat{\theta}_{(1U)}^*]$ avec $L = \alpha C$ et $U = (1-\alpha)C$. Le bootstrap $-t$ repose sur l'estimation de la statistique pivot $t = (\hat{\theta}_1 - \theta_1)\Big/\sqrt{v_{\mathrm{BWO}}(\hat{\theta}_1)}$ par son analogue bootstrap $t^* = (\hat{\theta}_1^* - \hat{\theta}_1)\Big/\sqrt{v_{\mathrm{BWO}}^*(\hat{\theta}_1^*)}$, où $v_{\mathrm{BWO}}^*(\hat{\theta}_1^*)$ s'obtient en appliquant la procédure bootstrap à la réplique de l'échantillon s^*. Le bootstrap $-t$ demande énormément de calcul puisqu'un double bootstrap est nécessaire, et est donc moins intéressant pour

l'utilisateur des données. Par conséquent, nous ne poursuivons pas cette approche et nous nous concentrons sur la méthode des percentiles.

Les méthodes de linéarisation fournissent des formules de variance applicables à des plans d'échantillonnage généraux, mais comprennent des calculs de dérivées qui peuvent être ardus pour les paramètres d'intérêt complexes, tels que l'indice de Gini. Contrairement à la linéarisation, le bootstrap évite le travail théorique grâce au calcul répété du système d'estimations existant. Les poids de rééchantillonnage sont fournis avec l'ensemble de données et peuvent être utilisés facilement pour produire des estimations de variance pour une grande gamme de statistiques. Cependant, les techniques bootstrap ne conviennent habituellement pas pour les plans d'échantillonnage généraux. Autrement dit, un plan d'échantillonnage particulier requiert habituellement un schéma de rééchantillonnage taillé sur mesure. Dans le présent article, nous nous concentrons sur deux techniques bootstrap particulières, qui seront généralisées à la section 3 au contexte de deux échantillons.

2.4.1 Bootstrap sans remise pour l'échantillonnage SI

Quand l'échantillon s est sélectionné par échantillonnage SI, nous considérons le bootstrap sans remise (BWO) introduit par Gross (1980). L'approche s'étend facilement à l'échantillonnage aléatoire simple stratifié (STSI) avec un nombre fini de strates. Supposons que N/n est un entier. Alors, le vecteur D s'obtient en créant d'abord une pseudopopulation U^* de taille N en reproduisant N/n fois chaque unité k dans l'échantillon original s, puis en sélectionnant une réplique d'échantillon SI s^* de taille n dans U^*.

La mesure bootstrap est donnée par (2.14), où le poids de rééchantillonnage D_k est le nombre de fois que l'unité $k \in s$ est sélectionnée dans s^*. La construction de U^* peut être évitée en notant que, sous la procédure BWO, le vecteur D suit une loi hypergéométrique multivariée. Par conséquent, les poids de rééchantillonnage peuvent être produits directement. On peut montrer que la procédure BWO mène à

$$E_*\left(\sum_s w_k D_k y_k\right) = \hat{t}_{y1}^{\mathrm{HT}} \quad \text{et} \quad V_*\left(\sum_s w_k D_k y_k\right) = \frac{1-n^{-1}}{1-N^{-1}} v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right), \tag{2.19}$$

où $v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right)$ est donné dans (2.2), de sorte que l'on obtient approximativement une concordance avec l'équation (2.15) pour une grande taille d'échantillon.

Plusieurs solutions ont été proposées pour traiter le cas où N/n n'est pas un entier; voir Chao et Lo (1985), Bickel et Freedman (1984), Sitter (1992b), Booth, Butler et Hall (1994), Presnell et Booth (1994), entre autres. La généralisation de l'estimation de la variance par la procédure BWO pour des plans d'échantillonnage avec probabilités inégales est examinée dans Särndal, Swensson et Wretman (1992) et Chauvet (2007).

2.4.2 Bootstrap avec remise pour l'échantillonnage à plusieurs degrés

Quand l'échantillon s est tiré par échantillonnage à plusieurs degrés et par échantillonnage à probabilités inégales avec remise des UPE, nous considérons le bootstrap des UPE (BWR) introduit par Rao et Wu (1988). Une réplique d'échantillon avec remise s_I^* de taille $m-1$ est sélectionnée par échantillonnage aléatoire simple avec remise (SIR) dans l'échantillon de premier degré original s_I. La mesure bootstrap est

$$\hat{M}_1^* = \frac{m}{m-1}\sum_{i\in s_I^*}\sum_{k\in s_i}\pi_{Ii}^{-1}\pi_{k|i}^{-1}\delta_{y1k} = \sum_{k\in s} w_k D_k \delta_{yk}, \tag{2.20}$$

où le poids de rééchantillonnage D_k égale $m(m-1)^{-1}$ multiplié par le nombre de fois que l'UPE contenant k est sélectionnée dans s_I^*.

La taille de rééchantillonnage $m-1$ est utilisée pour reproduire l'estimateur de variance sans biais habituel dans le cas linéaire (voir Rao et Wu, 1988). On peut montrer que la procédure BWR mène à

$$E_*\left(\sum_s w_k D_k y_k\right) = \hat{t}_{y1}^{\mathrm{HH}} \quad \text{et} \quad V_*\left(\sum_s w_k D_k y_k\right) = v^{\mathrm{HH}}\left(\hat{t}_{y1}^{\mathrm{HH}}\right), \tag{2.21}$$

où $v^{\mathrm{HH}}\left(\hat{t}_{y1}^{\mathrm{HH}}\right)$ est donné dans (2.3), de sorte que l'équation (2.16) est exactement reproduite. La procédure BWR est particulièrement simple, puisqu'elle comporte un rééchantillonnage pour le premier degré d'échantillonnage uniquement, et que les sous-échantillons d'unités d'échantillonnage secondaires (USE) ne sont pas modifiés à l'intérieur des UPE rééchantillonnées.

3 Le cas de deux échantillons

3.1 Notation et estimation composite

Supposons maintenant que deux variables $\mathcal{Y}_1$ et $\mathcal{Y}_2$ sont mesurées sur la population U, et soit $y_{d1},\ldots,y_{dN}$ les valeurs prises par $\mathcal{Y}_d, d=1,2,$ sur les unités de la population. Les variables $\mathcal{Y}_1$ et $\mathcal{Y}_2$ peuvent par exemple correspondre à une caractéristique d'intérêt observée à deux périodes différentes τ_1 et τ_2. Nous considérons l'estimation de paramètres $\Delta\theta$ qui peuvent s'écrire sous la forme d'une fonctionnelle $\Delta\theta = T(M_1, M_2)$, où $M_d = \sum_{k\in U}\delta_{\{y_{dk}\}}$. Par exemple, le cas linéaire $\Delta t = t_{y2} - t_{y1}$ correspond à la différence entre les totaux $t_{y2} = \sum_{k\in U} y_{2k}$ et $t_{y1} = \sum_{k\in U} y_{1k}$.

Soit s_1 et s_2 deux échantillons de tailles n_1 et n_2, respectivement, tirés de la même population U selon un plan d'échantillonnage bidimensionnel $p(\cdot,\cdot)$ (voir Goga, 2003). La variable $\mathcal{Y}_1$ est mesurée sur s_1, tandis que la variable $\mathcal{Y}_2$ est mesurée sur s_2. L'insertion des estimateurs fondés sur l'échantillon $\hat{M}_d$ dans $\Delta\theta$ donne l'estimateur par substitution $\widehat{\Delta\theta} = T\left(\hat{M}_1, \hat{M}_2\right)$. Contrairement au cas d'un seul échantillon, plusieurs estimateurs $\hat{M}_d$ sont possibles. Dans la suite de l'exposé, nous nous concentrons sur la classe générale d'*estimateurs composites* introduite par Goga, Deville et Ruiz-Gazen (2009). Nous notons $s_{1\bullet} = s_1 \setminus s_2$, $s_3 = s_1 \cap s_2$ et $s_{2\bullet} = s_2 \setminus s_1$. Pour $\lozenge \in \{1\bullet, 3, 2\bullet\}$, nous notons $\pi_{\lozenge,k}$ le nombre prévu de tirages de l'unité k dans $s_\lozenge$ et $\hat{M}_{d,\lozenge} = \sum_{k\in s_\lozenge} w_{\lozenge,k}\delta_{y_{dk}}$, où $w_{\lozenge,k} = \pi_{\lozenge,k}^{-1}$. Les estimateurs composites de M_1 et M_2 sont

$$\hat{M}_1^{\mathrm{co}}(a) = a\,\hat{M}_{1,1\bullet} + (1-a)\,\hat{M}_{1,3} \quad \text{et} \quad \hat{M}_2^{\mathrm{co}}(b) = b\,\hat{M}_{2,2\bullet} + (1-b)\,\hat{M}_{2,3}, \tag{3.1}$$

où a et b sont des constantes connues. Le choix $a = b = 0$ mène à l'*estimateur « intersection »* avec $\hat{M}_1^{\mathrm{int}} = \hat{M}_{1,3}$ et $\hat{M}_2^{\mathrm{int}} = \hat{M}_{2,3}$, où seul est utilisé l'échantillon « intersection » s_3 (correspondant à l'intersection).

Si l'on estime le paramètre $\Delta t = t_{y2} - t_{y1}$, l'estimateur composite est donné par

$$\widehat{\Delta t}^{\,\mathrm{co}}(a,b) \;=\; \hat{t}_{y_2}^{\,\mathrm{co}} - \hat{t}_{y_1}^{\,\mathrm{co}}, \tag{3.2}$$

où $\hat{t}_{y_1}^{\,\mathrm{co}} = \int y d\hat{M}_1^{\mathrm{co}}(y)$ et $\hat{t}_{y_2}^{\,\mathrm{co}} = \int y d\hat{M}_2^{\mathrm{co}}(y)$. Il peut se réécrire sous la forme

$$\widehat{\Delta t}^{\,\mathrm{co}}(a,b) \;=\; b\left(\hat{t}_{y_2,s_{2\bullet}} - \hat{t}_{y_2,s_3}\right) - a\left(\hat{t}_{y_1,s_{1\bullet}} - \hat{t}_{y_1,s_3}\right) + \left(\hat{t}_{y_2,s_3} - \hat{t}_{y_1,s_3}\right), \tag{3.3}$$

où $\hat{t}_{y_d,s_\diamond} = \sum_{k\in s_\diamond} w_{\diamond,k}\, y_{dk}$. La variance de l'estimateur composite est

$$V\left\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\right\} \;=\; (b,-a,1)\, V\left\{\left(\hat{t}_{y_2,s_{2\bullet}} - \hat{t}_{y_2,s_3}, \hat{t}_{y_1,s_{1\bullet}} - \hat{t}_{y_1,s_3}, \hat{t}_{y_2,s_3} - \hat{t}_{y_1,s_3}\right)^{\top}\right\}(b,-a,1)^{\top}. \tag{3.4}$$

Trouver le vecteur $(a_{\mathrm{opt}}, b_{\mathrm{opt}})^{\top}$ qui minimise la variance en (3.4) mène à l'*estimateur composite optimal* (Goga, Deville et Ruiz-Gazen, 2009, section 3.6). Notons qu'il ne s'agit pas d'un estimateur proprement dit, puisqu'il dépend de quantités inconnues qui doivent être estimées en pratique. Cependant, il représente une référence utile que nous utiliserons pour évaluer des estimateurs composites plus simples.

Un estimateur de variance s'obtient en substituant dans (3.4) un estimateur de la matrice de variance-covariance. L'obtention des estimateurs de variance est décrite en détail aux sections 3.1.1 et 3.1.2 pour deux exemples de plans d'échantillonnage bidimensionnels.

3.1.1 Plan SI bidimensionnel

Le plan SI bidimensionnel (SI2) de taille fixée $(n_{1\bullet}, n_3, n_{2\bullet})$ attribue des probabilités égales à tous les $s = (s_1, s_2)$ pour lesquels les sous-échantillons associés $s_{1\bullet}$, s_3 et $s_{2\bullet}$ possèdent les tailles requises $n_{1\bullet}$, n_3 et $n_{2\bullet}$, voir Goga (2003) ainsi que Qualité et Tillé (2008). Le plan SI2 a pour propriété intéressante que les échantillons marginaux $s_{1\bullet}$, s_3 et $s_{2\bullet}$ sont des échantillons SI provenant de la population U. De même, s_1 est un échantillon SI de taille $n_1 = n_{1\bullet} + n_3$, et s_2 est un échantillon SI de taille $n_2 = n_{2\bullet} + n_3$. Pour le plan d'échantillonnage SI2, l'estimateur composite en (3.3) donne

$$\widehat{\Delta t}^{\,\mathrm{co}}(a,b) \;=\; Nb\left(\bar{y}_{2,s_{2\bullet}} - \bar{y}_{2,s_3}\right) - Na\left(\bar{y}_{1,s_{1\bullet}} - \bar{y}_{1,s_3}\right) + N\left(\bar{y}_{2,s_3} - \bar{y}_{1,s_3}\right), \tag{3.5}$$

et la variance de l'estimateur composite s'exprime par

$$V\left\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\right\} \;=\; N^2\left\{c_1(a)\, S_{y_1,U}^2 - 2c_{12}(a,b)\, S_{y_1y_2,U} + c_2(b)\, S_{y_2,U}^2\right\}, \tag{3.6}$$

avec

$$\begin{aligned} c_1(a) &= \frac{(1-a)^2}{n_3} + \frac{a^2}{n_1 - n_3} - \frac{1}{N}, \\ c_2(b) &= \frac{(1-b)^2}{n_3} + \frac{b^2}{n_2 - n_3} - \frac{1}{N}, \\ c_{12}(a,b) &= \frac{(1-a)(1-b)}{n_3} - \frac{1}{N}, \end{aligned}$$

voir l'annexe pour une preuve.

Nous considérons deux exemples. Le choix $a = b = 0$ mène à l'estimateur « intersection »

$$\widehat{\Delta t}^{\,\text{int}} = \widehat{\Delta t}^{\,\text{co}}(0,0) = \frac{N}{n_3}\sum_{k\in s_3}(y_{2k} - y_{1k}), \tag{3.7}$$

et l'expression de la variance se simplifie en

$$V\left\{\widehat{\Delta t}^{\,\text{int}}\right\} = N^2\left(\frac{1}{n_3} - \frac{1}{N}\right)S^2_{y_2-y_1,U}. \tag{3.8}$$

Le choix $a = n_1^{-1}n_{1\bullet}$ et $b = n_2^{-1}n_{2\bullet}$ mène à l'*estimateur « union »*

$$\widehat{\Delta t}^{\,\text{uni}} = \widehat{\Delta t}^{\,\text{co}}(n_1^{-1}n_{1\bullet}, n_2^{-1}n_{2\bullet}) = \frac{N}{n_2}\sum_{k\in s_2} y_{2k} - \frac{N}{n_1}\sum_{k\in s_1} y_{1k} \tag{3.9}$$

où les échantillons complets sont utilisés, et la variance peut s'écrire sous la forme

$$V\left\{\widehat{\Delta t}^{\,\text{uni}}\right\} = N^2\left\{\left(\frac{1}{n_1} - \frac{1}{N}\right)S^2_{y_1,U} - 2\left(\frac{n_3}{n_1\,n_2} - \frac{1}{N}\right)S_{y_1y_2,U} + \left(\frac{1}{n_2} - \frac{1}{N}\right)S^2_{y_2,U}\right\}. \tag{3.10}$$

Les variances de l'estimateur « union » et de l'estimateur « intersection » ont été établies par Qualité et Tillé (2008), voir aussi Tam (1984).

Le choix de a et b revêt une importance pratique si l'on veut obtenir un estimateur composite efficace. Après un peu de calcul, le vecteur $(a_{\text{opt}}, b_{\text{opt}})^\top$ qui minimise la variance de $\widehat{\Delta t}^{\,\text{co}}(a,b)$ est donné par

$$(a_{\text{opt}}, b_{\text{opt}})^\top = A^{-1}X \tag{3.11}$$

avec

$$A = \begin{pmatrix} \dfrac{n_1}{n_1 - n_3} & -\dfrac{S_{y_1y_2,U}}{S^2_{y_1,U}} \\ -\dfrac{S_{y_1y_2,U}}{S^2_{y_2,U}} & \dfrac{n_2}{n_2 - n_3} \end{pmatrix} \quad \text{et} \quad X = \left(1 - \frac{S_{y_1y_2,U}}{S^2_{y_1,U}}, 1 - \frac{S_{y_1y_2,U}}{S^2_{y_2,U}}\right)^\top. \tag{3.12}$$

Pour deux variables $\mathcal{Y}_1$ et $\mathcal{Y}_2$ se rapportant à une même caractéristique observée à deux périodes différentes, $S_{y_1y_2,U}$ doit, en principe, être proche de $S^2_{y_1,U}$ et $S^2_{y_2,U}$. Le vecteur X dans (3.12) est, à son tour, proche du vecteur nul, et si la taille de l'échantillon « intersection » s_3 est comparable à celles de $s_{1\bullet}$ et $s_{2\bullet}$, nous obtenons $a_{\text{opt}} \simeq 0$ et $b_{\text{opt}} \simeq 0$. Par conséquent, l'utilisation de l'estimateur « intersection » où $a = b = 0$ paraît raisonnable en pratique. Au contraire, l'estimateur « union » peut être très inefficace; voir la section 4.2 pour un exemple. Ces conclusions concordent avec celles de Qualité et Tillé (2008), section 2.2.2.

Plusieurs estimateurs de variance peuvent être utilisés pour l'estimateur composite. L'estimation des dispersions sur l'échantillon « intersection » uniquement donne l'estimateur de variance sans biais

$$v_{\text{int}}^{\text{HT}}\left\{\widehat{\Delta t}^{\,\text{co}}(a,b)\right\} = N^2\left\{c_1(a)\,S^2_{y_1,s_3} - 2c_{12}(a,b)\,S_{y_1y_2,s_3} + c_2(b)\,S^2_{y_2,s_3}\right\}, \tag{3.13}$$

tandis qu'une estimation sur les échantillons entiers donne

$$v_{\text{uni}}^{\text{HT}}\left\{\widehat{\Delta t}^{\,\text{co}}(a,b)\right\} = N^2\left\{c_1(a)S^2_{y_{1,}s_1} - 2c_{12}(a,b)S_{y_1y_2,s_3} + c_2(b)S^2_{y_2,s_2}\right\}. \tag{3.14}$$

Berger (2004) a considéré l'estimation de la variance pour l'estimateur « union » sous un plan d'échantillonnage rotatif à entropie maximale en estimant séparément les trois composantes dans (3.6).

3.1.2 Plan à plusieurs degrés bidimensionnel

Considérons maintenant un plan d'échantillonnage à deux degrés bidimensionnel (MULT2). Nous supposons qu'un échantillon de premier degré s_I de taille m est d'abord sélectionné avec remise parmi les UPE $U_1,\ldots,U_{N_I}$. À l'intérieur de chaque UPE $i \in s_I$, on sélectionne ensuite un échantillon SI2 de taille $(n^i_{1\bullet}, n^i_{3}, n^i_{2\bullet})$. Ce type de plan d'échantillonnage se dégage en particulier dans le cas d'un plan à deux degrés autopondéré en deux vagues, avec à la deuxième vague un remplacement partiel des USE sélectionnées à la première vague. L'estimateur composite en (3.3) donne

$$\widehat{\Delta t}^{\,\text{co}}(a,b) = \sum_{i\in s_I}\pi_{Ii}^{-1}\widehat{\Delta t}^{\,i,\text{co}}(a,b) \tag{3.15}$$

où

$$\widehat{\Delta t}^{\,i,\text{co}}(a,b) = N_i b\left(\overline{y}_{2,s^i_{2\bullet}} - \overline{y}_{2,s^i_3}\right) - N_i a\left(\overline{y}_{1,s^i_{1\bullet}} - \overline{y}_{1,s^i_3}\right) + N_i\left(\overline{y}_{2,s^i_3} - \overline{y}_{1,s^i_3}\right), \tag{3.16}$$

où $\overline{y}_{d,s^i_\lozenge} = (n^i_\lozenge)^{-1}\sum_{k\in s^i_\lozenge} y_{\lozenge k}$, où $s^i_\lozenge = s_\lozenge \cap U_i$, et où N_i désigne le nombre d'USE dans l'UPE u_i.

Par exemple, en utilisant uniquement les échantillons communs à l'intérieur des UPE, on obtient l'estimateur « intersection »

$$\widehat{\Delta t}^{\,\text{int}} = \sum_{i\in s_I}\pi_{Ii}^{-1}\widehat{\Delta t}^{\,i,\text{int}} \quad \text{avec} \quad \widehat{\Delta t}^{\,i,\text{int}} = N_i\left(\overline{y}_{2,s^i_3} - \overline{y}_{1,s^i_3}\right). \tag{3.17}$$

En utilisant les échantillons complets à l'intérieur des UPE, on obtient l'estimateur « union »

$$\widehat{\Delta t}^{\,\text{uni}} = \sum_{i\in s_I}\pi_{Ii}^{-1}\widehat{\Delta t}^{\,i,\text{uni}} \quad \text{avec} \quad \widehat{\Delta t}^{\,i,\text{uni}} = N_i\left(\overline{y}_{2,s^i_2} - \overline{y}_{1,s^i_1}\right). \tag{3.18}$$

Nous notons que, pour tout vecteur de valeurs $(a,b)^\top$, la variance due au premier degré d'échantillonnage pour $\widehat{\Delta t}^{\,\text{co}}(a,b)$ est la même. Les estimateurs composites possibles diffèrent donc en ce qui concerne la variance de second degré uniquement. Compte tenu de la discussion de la section 3.1.1, nous nous attendons par conséquent à ce que l'estimateur « intersection » soit proche de l'estimateur composite optimal; voir la section 4.2 pour un exemple. Un estimateur de variance sans biais pour $\widehat{\Delta t}^{\,\text{co}}(a,b)$ est donné par

$$v^{\text{HH}}\left\{\widehat{\Delta t}^{\,\text{co}}(a,b)\right\} = \frac{m}{m-1}\sum_{i\in s_I}\left(\frac{\widehat{\Delta t}^{\,i,\text{co}}(a,b)}{\pi_{Ii}} - \frac{\widehat{\Delta t}^{\,\text{co}}(a,b)}{m}\right)^2. \tag{3.19}$$

3.2 Estimation de l'évolution de l'indice de Gini

L'évolution de l'indice de Gini $\Delta G = G_2 - G_1$ peut s'écrire sous la forme

$$\Delta G = \frac{\int \{2F_{2N}(y) - 1\}\, y dM_2(y)}{\int y dM_2(y)} - \frac{\int \{2F_{1N}(y) - 1\}\, y dM_1(y)}{\int y dM_1(y)} \tag{3.20}$$

où $F_{dN}(y) = N^{-1}\sum_{k\in U} 1_{\{y_{dk}\le y\}}$, $d = 1, 2$. L'utilisation de l'estimation composite mène à

$$\widehat{\Delta G}^{\text{co}}(a,b) = \frac{\int \left\{2\hat{F}_{2N}^{\text{co}}(y) - 1\right\} y d\hat{M}_2^{\text{co}}(y)}{\int y d\hat{M}_2^{\text{co}}(y)} - \frac{\int \left\{2\hat{F}_{1N}^{\text{co}}(y) - 1\right\} y d\hat{M}_1^{\text{co}}(y)}{\int y d\hat{M}_1^{\text{co}}(y)} \tag{3.21}$$

où $\hat{F}_{dN}^{\text{co}}(y) = \left\{\int d\hat{M}_d^{\text{co}}(y)\right\}^{-1} \int 1_{\{\xi\le y\}}\, d\hat{M}_d^{\text{co}}(\xi)$.

Habituellement, dans un cadre d'échantillonnage temporel, les échantillons s_1 et s_2 ne sont pas indépendants. Par conséquent, nos conditions diffèrent de l'estimation usuelle des fonctionnelles dépendantes des fonctions de répartition estimées sur des échantillons indépendants; voir, par exemple, Pires et Branco (2002) et Reid (1981), qui donnent le développement d'ordre un d'une fonctionnelle pour deux échantillons utilisant les fonctions d'influence partielles. Davison et Hinkley (1997, page 71) donnent des méthodes bootstrap sous un cadre similaire. Sous un plan d'échantillonnage bidimensionnel général $p(\cdot,\cdot)$, Goga, Deville et Ruiz-Gazen (2009) donnent une technique de linéarisation pour deux échantillons de fonctionnelles bivariées que nous utiliserons dans la suite de l'exposé.

3.3 Estimation de la variance par linéarisation

Pour obtenir la variance asymptotique de $\widehat{\Delta\theta}^{\text{co}}(a,b)$, nous adoptons le cadre asymptotique introduit par Goga, Deville et Ruiz-Gazen (2009), qui est une extension du cas à deux échantillons du cadre asymptotique d'Isaki et Fuller (1982). Définissons, quand elles existent, les *fonctions d'influence partielles* d'une fonctionnelle $T(M_1, M_2)$ au point y par

$$I_1T(M_1, M_2; y) = \lim_{h\to 0} \frac{T(M_1 + h\delta_y, M_2) - T(M_1, M_2)}{h},$$

$$I_2T(M_1, M_2; y) = \lim_{h\to 0} \frac{T(M_1, M_2 + h\delta_y) - T(M_1, M_2)}{h}.$$

Nous définissons les *variables linéarisées* $u_{dk} = I_dT(M_1, M_2; y_{dk})$ pour $d = 1, 2$ comme étant les fonctions d'influence partielles de T pour (M_1, M_2) et $y = y_{dk}$. Pour l'évolution de l'indice de Gini ΔG, nous pouvons calculer les variables linéarisées u_{dk} en utilisant (2.10), à savoir

$$u_{dk} = 2F_{dN}(y_{dk})\frac{y_{dk} - \bar{y}_{dk,U<}}{t_{y_d}} - y_{dk}\frac{G_d + 1}{t_{y_d}} + \frac{1 - G_d}{N}, \tag{3.22}$$

où $\bar{y}_{dk,U<} = \left(\sum_{l\in U} 1_{\{y_{dl} < y_{dk}\}}\right)^{-1} \sum_{j\in U} y_{dj} 1_{\{y_{dj} < y_{dk}\}}$. La variable linéarisée estimée est

$$\hat{u}_{dk} = 2\hat{F}_{dN}^{\text{co}}(y_{dk})\frac{y_{dk} - \overline{y}_{dk,s<}^{\text{co}}}{\hat{t}_{y1}^{\text{co}}} - y_{dk}\frac{\hat{G}_{d}^{\text{co}} + 1}{\hat{t}_{y1}^{\text{co}}} + \frac{1 - \hat{G}_{d}^{\text{co}}}{\hat{N}}. \tag{3.23}$$

3.3.1 Plan SI bidimensionnel

Dans le cas du plan SI2 présenté à la section 3.1.1, l'insertion des variables u_{dk} calculées en (3.22) dans la formule de variance (3.6) donne l'approximation de la variance

$$V\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} \simeq N^2\left\{c_1(a)S_{u_1,U}^2 - 2c_{12}(a,b)S_{u_1u_2,U} + c_2(b)S_{u_2,U}^2\right\},$$

voir le théorème 1 dans Goga, Deville et Ruiz-Gazen (2009). Pour obtenir un estimateur de variance, les variables linéarisées peuvent être estimées de plusieurs façons. Si l'on utilise seulement l'échantillon « intersection » s_3, les variables linéarisées estimées $\hat{u}_d$ s'obtiennent au moyen de (3.23) en prenant $\hat{M}_1^{\text{co}} = \hat{M}_{1,3}$ et $\hat{M}_2^{\text{co}} = \hat{M}_{2,3}$. Un estimateur de variance s'obtient alors en insérant ces variables linéarisées dans (3.13). Cela donne

$$v_{\text{int}}^{\text{HT}}\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} = N^2\left\{c_1(a)S_{\hat{u}_1,s_3}^2 - 2c_{12}(a,b)S_{\hat{u}_1\hat{u}_2,s_3} + c_2(b)S_{\hat{u}_2,s_3}^2\right\}. \tag{3.24}$$

Si les deux échantillons s_1 et s_2 sont utilisés, les variables linéarisées estimées $\hat{u}_d$ s'obtiennent au moyen de (3.23) en prenant $\hat{M}_1^{\text{co}} = \hat{M}_{1,1}$ et $\hat{M}_2^{\text{co}} = \hat{M}_{2,2}$. Un estimateur de variance s'obtient alors en insérant ces variables linéarisées dans (3.14). Cela donne

$$v_{\text{uni}}^{\text{HT}}\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} = N^2\left\{c_1(a)S_{\hat{u}_1,s_1}^2 - 2c_{12}(a,b)S_{\hat{u}_1\hat{u}_2,s_3} + c_2(b)S_{\hat{u}_2,s_2}^2\right\}. \tag{3.25}$$

3.3.2 Plan à plusieurs degrés bidimensionnel

Dans le cas du plan MULT2 présenté à la section 3.1.2, les variables linéarisées peuvent également être estimées de plusieurs façons. Pour simplifier, nous considérons l'utilisation de l'échantillon « intersection » s_3 seulement, de sorte que les variables linéarisées estimées $\hat{u}_d$ s'obtiennent au moyen de (3.23) en prenant $\hat{M}_1^{\text{co}} = \hat{M}_{1,3}$ et $\hat{M}_2^{\text{co}} = \hat{M}_{2,3}$. Un estimateur de variance s'obtient alors en insérant ces variables linéarisées dans (3.19). Cela donne

$$v^{\text{HH}}\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} = \frac{m}{m-1}\sum_{i\in s_I}\left(\frac{\widehat{\Delta u}^{i,\text{co}}(a,b)}{\pi_{Ii}} - \frac{\widehat{\Delta u}^{\text{co}}(a,b)}{m}\right)^2, \tag{3.26}$$

où $\widehat{\Delta u}^{\text{co}}(a,b)$ et $\widehat{\Delta u}^{i,\text{co}}(a,b)$ s'obtiennent à partir de (3.15) et (3.16), respectivement, en remplaçant y_{dk} par $\hat{u}_{dk}$.

3.4 Estimation de la variance par bootstrap

Les méthodes bootstrap n'ont pas encore été étudiées dans le cas de l'évolution de l'indice de Gini. Les principes des techniques de bootstrap pondéré peuvent être étendus au contexte de deux échantillons,

c'est-à-dire que chaque mesure $\hat{M}_{d,\Diamond}$ avec $d = 1, 2$ et $\Diamond \in \{1\bullet, 3, 2\bullet\}$ est estimée, conditionnellement aux échantillons sélectionnés au départ, par une mesure bootstrap pondérée $\hat{M}^*_{d,\Diamond}$ qui permet de reproduire, au moins approximativement, les deux premiers moments d'un estimateur sans biais dans le cas linéaire. À la section 3.4.1, nous examinons une généralisation du bootstrap sans remise (BWO) au plan SI2. À la section 3.4.2, nous proposons une généralisation du bootstrap avec remise (BWR) au plan MULT2.

3.4.1 Une généralisation du bootstrap sans remise au plan SI2

Nous considérons d'abord le plan SI2. La construction d'une pseudopopulation U^* est plus complexe dans le cas de deux échantillons, puisque les variables d'intérêt mesurées aux vagues τ_1 et τ_2 doivent être disponibles pour chaque unité dans U^*. Nous décrivons donc un algorithme bootstrap où seul l'échantillon « intersection » s_3 est utilisé pour construire la pseudopopulation U^*, dans l'esprit de l'estimateur de variance « intersection » en (3.24).

Supposons que N/n_3 est un entier. Les vecteurs $D_\Diamond$ s'obtiennent en créant d'abord une pseudopopulation U^* de taille N en dupliquant N/n_3 fois chaque unité k de l'échantillon original s_3. Une réplique d'échantillon SI2 $s^* = (s^*_{1\bullet}, s^*_3, s^*_{\bullet 2})$ de taille $(n_{1\bullet}, n_3, n_{2\bullet})$ est ensuite sélectionnée dans U^*. Les mesures bootstrap sont alors

$$\hat{M}^*_{d,\Diamond} = \sum_{k \in s_3} w_{\Diamond,k} D_{\Diamond,k} \delta_{y_{dk}}, \tag{3.27}$$

avec $D_{\Diamond,k}$ le nombre de fois que l'unité k est sélectionnée dans la réplique d'échantillon $s^*_\Diamond$. Dans le cas linéaire, l'estimateur bootstrap du paramètre Δt est alors donné par

$$\widehat{\Delta t}^{\,\mathrm{co}^*}(a,b) = b\left(\hat{t}_{y_2,s^*_{2\bullet}} - \hat{t}_{y_2,s^*_3}\right) - a\left(\hat{t}_{y_1,s^*_{1\bullet}} - \hat{t}_{y_1,s^*_3}\right) + \left(\hat{t}_{y_2,s^*_3} - \hat{t}_{y_1,s^*_3}\right), \tag{3.28}$$

où $\hat{t}_{y_d,s^*_\Diamond} = \sum_{k \in s_3} w_{\Diamond,k} D_{\Diamond,k} y_{dk}$. Après un peu de calcul, nous obtenons

$$E_*\left\{\widehat{\Delta t}^{\,\mathrm{co}^*}(a,b)\right\} = \widehat{\Delta t}^{\,\mathrm{int}} \quad \text{et} \quad V_*\left\{\widehat{\Delta t}^{\,\mathrm{co}^*}(a,b)\right\} = \frac{1 - n_3^{-1}}{1 - N^{-1}} v_{\mathrm{int}}^{\mathrm{HT}}\left\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\right\}, \tag{3.29}$$

où $\widehat{\Delta t}^{\,\mathrm{int}}$ est donné en (3.7), et $v_{\mathrm{int}}^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right)$ est donné en (3.13). La généralisation du bootstrap sans remise (BWO) permet donc de reproduire exactement l'estimateur « intersection » du premier moment et de reproduire approximativement l'estimateur « intersection » du deuxième moment pour une grande valeur de n_3.

La construction de U^* peut être évitée en notant que, sous la procédure BWO, chaque vecteur $D_\Diamond$ suit une loi hypergéométrique multivariée. Par conséquent, les poids de rééchantillonnage peuvent être produits directement. L'algorithme peut être adapté au cas général où N/n_3 n'est pas un entier en appliquant n'importe laquelle des techniques mentionnées à la section 2.4.

3.4.2 Une généralisation du bootstrap avec remise pour le plan à plusieurs degrés bidimensionnel

Nous considérons maintenant le plan d'échantillonnage à deux degrés bidimensionnel avec un échantillon de premier degré commun s_I présenté à la section 3.1.2. La procédure bootstrap proposée est similaire à celle décrite dans Rao et Wu (1988). Une réplique d'échantillon s_I^* de taille $m-1$ est tirée par échantillonnage aléatoire simple avec remise (SIR) dans l'échantillon de premier degré original s_I. Les mesures bootstrap sont alors

$$\hat{M}_{d,\Diamond}^* = \frac{m}{m-1}\sum_{i\in s_I^*}\sum_{k\in s_\Diamond^i}\pi_{Ii}^{-1}\pi_{\Diamond k|i}^{-1}\delta_{y_{dk}} \quad \text{où} \quad \pi_{\Diamond k|i} = \frac{n_\Diamond^i}{N_i}. \tag{3.30}$$

Celle-ci peut se réécrire sous la forme

$$\hat{M}_{d,\Diamond}^* = \sum_{k\in s_\Diamond} w_{\Diamond,k} D_{\Diamond,k}\delta_{y_{dk}}, \tag{3.31}$$

où $s_\Diamond$ est l'union des échantillons $s_\Diamond^i$ pour $i \in s_I$, et où le poids de rééchantillonnage $D_{\Diamond,k}$ est égal à $m(m-1)^{-1}$ multiplié par le nombre de fois que l'UPE contenant k est sélectionnée dans s_I^*.

Dans le cas linéaire, l'estimateur bootstrap du paramètre Δt est alors

$$\widehat{\Delta t}^{\,\mathrm{co}*}(a,b) = \frac{m}{m-1}\sum_{i\in s_I^*}\pi_{Ii}^{-1}\widehat{\Delta t}^{\,i,\mathrm{co}}(a,b) \tag{3.32}$$

où $\widehat{\Delta t}^{\,i,\mathrm{co}}(a,b)$ est défini en (3.16). Après un peu de calcul, nous obtenons

$$E_*\left\{\widehat{\Delta t}^{\,\mathrm{co}*}(a,b)\right\} = \widehat{\Delta t}^{\,\mathrm{co}}(a,b) \quad \text{et} \quad V_*\left\{\widehat{\Delta t}^{\,\mathrm{co}*}(a,b)\right\} = v^{\mathrm{HH}}\left\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\right\}, \tag{3.33}$$

où $\widehat{\Delta t}^{\,\mathrm{co}}(a,b)$ est donné en (3.15), et $v^{\mathrm{HH}}\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\}$ est donné en (3.19). La généralisation proposée du bootstrap avec remise permet donc de reproduire exactement l'estimateur composite du premier moment et l'estimateur associé au deuxième moment.

4 Étude par simulations

Pour commencer, cinq populations artificielles sont générées comme il est décrit à la section 4.1. À la section 4.2, l'estimateur « union » est comparé à l'estimateur « intersection » en ce qui concerne la variance asymptotique. Une expérience Monte Carlo est ensuite présentée à la section 4.3, et les performances de la linéarisation et du bootstrap sont comparées dans le cas d'un plan d'échantillonnage SI2. Une comparaison similaire est faite à la section 4.4 dans le cas du plan d'échantillonnage à deux degrés bidimensionnel.

4.1 Configuration de la simulation

Nous avons généré cinq populations finies de taille N = 40 000, contenant chacune deux variables étudiées y_1 et y_2. Les valeurs y_{1k} et les valeurs y_{2k} ont été produites conformément au modèle lognormal

$$y_{dk} = \exp(\alpha_d \varepsilon_k). \tag{4.1}$$

Les résidus ε_k ont été générés selon une loi normale centrée réduite. Les valeurs des indices de Gini pour les cinq populations sont présentées au tableau 4.1.

Tableau 4.1
Indices de Gini pour 5 populations

Population	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5
G_1	0,249	0,298	0,348	0,397	0,447
G_2	0,259	0,318	0,378	0,437	0,496
ΔG	0,010	0,020	0,030	0,040	0,049

Dans chacune des cinq populations, les unités ont été groupées en $M = 500$ grappes de taille égale $N_0 = 80$. Les grappes ont été construites de façon que le coefficient de corrélation intragrappe par rapport à la variable y_1 soit approximativement égal à 0,20 dans chaque population.

4.2 Comparaison des estimateurs « union » et « intersection »

À la présente section, nous comparons l'estimateur « union » à l'estimateur « intersection » pour la variation de l'indice de Gini en ce qui concerne la variance asymptotique. Nous considérons deux plans d'échantillonnage, à savoir le plan SI2 présenté à la section 3.1.1 avec $(n_{1\bullet}, n_3, n_{2\bullet})$ = (1 000; 1 000; 1 000), (1 000; 2 000; 1 000) ou (1 000; 4 000; 1 000); et le plan MULT2 présenté à la section 3.1.2 avec $m = 300$ et $(n^i_{1\bullet}, n^i_3, n^i_{2\bullet})$ = (10; 10; 10), (10; 20; 10) ou (10; 40; 10).

Pour chaque population, nous calculons la variance asymptotique $V_{\text{lin}}(\widehat{\Delta G}^{\text{uni}})$ de l'estimateur « union », et la variance asymptotique $V_{\text{lin}}(\widehat{\Delta G}^{\text{int}})$ de l'estimateur « intersection ». Afin de comparer ces variances, nous calculons l'efficacité relative définie comme étant

$$\text{ER}\left\{\widehat{\Delta G}^{\cdot}\right\} = \frac{V_{\text{lin}}\left\{\widehat{\Delta G}^{(\cdot)}\right\}}{V_{\text{lin}}\left\{\widehat{\Delta G}^{\text{opt}}\right\}}, \tag{4.2}$$

où $\widehat{\Delta G}^{\text{opt}}$ est l'estimateur optimal.

Les résultats sont présentés au tableau 4.2. L'estimateur « union » est très inefficace. Sa variance asymptotique est de 15 à 244 fois plus élevée que celle de l'estimateur « intersection » pour SI2, et de 2 à 44 fois plus élevée que celle de l'estimateur « intersection » pour MULT2. La différence entre les deux estimateurs a tendance à diminuer quand la taille de l'échantillon commun augmente ou quand ΔG augmente. Par ailleurs, l'estimateur « intersection » est légèrement moins efficace que l'estimateur optimal pour SI2, la valeur de ER variant de 1,33 à 2,46, et approximativement aussi efficace que l'estimateur optimal pour MULT2, la valeur de ER variant de 1,02 à 1,12. Ces constatations appuient le raisonnement heuristique de la section 3.1.1. Étant donné la médiocre performance de l'estimateur « union », et la bonne performance de l'estimateur « intersection », nous examinons uniquement ce dernier dans la suite de l'étude par simulations.

Tableau 4.2
Efficacité relative des estimateurs de variance « union » et « intersection » pour 5 populations

Plan	Taille d'échantillon	Pop. 1		Pop. 2		Pop. 3		Pop. 4		Pop. 5	
		$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$	$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$	$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$	$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$	$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$
SI2	n_3 = 1 000	600,22	2,46	200,23	2,27	96,72	2,10	58,73	1,96	39,35	1,85
	n_3 = 2 000	410,23	1,84	141,71	1,76	70,71	1,68	44,18	1,61	30,33	1,54
	n_3 = 4 000	250,02	1,47	88,40	1,43	45,17	1,40	28,86	1,36	20,23	1,33
MULT2	n_3^i = 10	49,10	1,12	19,89	1,13	11,83	1,14	8,84	1,15	7,28	1,16
	n_3^i = 20	23,08	1,05	9,75	1,05	6,08	1,05	4,73	1,06	4,04	1,07
	n_3^i = 40	9,15	1,02	4,25	1,02	2,90	1,02	2,41	1,02	2,16	1,02

4.3 Comparaison de la linéarisation et du bootstrap pour le plan SI2

À la présente section, nous comparons la linéarisation et le bootstrap pour l'estimation de la variance et pour la production des intervalles de confiance, dans le cas de l'estimateur « intersection », pour la variation de l'indice de Gini sous le plan d'échantillonnage SI2. Pour chaque population, nous avons tiré $B = 10\,000$ échantillons bidimensionnels selon le plan SI2 indexé par $(n_{1\bullet}, n_3, n_{2\bullet}) = (1\,000; 1\,000; 1\,000)$, $(n_{1\bullet}, n_3, n_{2\bullet}) = (1\,000; 2\,000; 1\,000)$ ou $(n_{1\bullet}, n_3, n_{2\bullet}) = (1\,000; 4\,000; 1\,000)$. Dans chaque échantillon, nous avons calculé l'estimateur « intersection » $\widehat{\Delta G}^{\text{int}}$ de la variation de l'indice de Gini. Pour cet estimateur, nous avons calculé i) l'estimateur de variance par linéarisation $v_{\text{int}}(\widehat{\Delta G}^{\text{int}})$ donné en (3.24), et ii) l'estimateur de variance par bootstrap $v_{\text{BWO}}(\widehat{\Delta G})$, en suivant la procédure bootstrap décrite à la section 3.4.1.

Pour mesurer le biais d'un estimateur de variance $v(\widehat{\Delta G})$, nous avons utilisé le biais relatif de Monte Carlo en pourcentage

$$\text{BR}\left\{v\left(\widehat{\Delta G}\right)\right\} = 100 \times \frac{B^{-1}\sum_{b=1}^{B} v\left(\widehat{\Delta G}_b\right) - \text{EQM}\left(\widehat{\Delta G}\right)}{\text{EQM}\left(\widehat{\Delta G}\right)}, \tag{4.3}$$

où $v(\widehat{\Delta G}_b)$ désigne l'estimateur $v(\widehat{\Delta G})$ dans le b^{e} échantillon, et $\text{EQM}(\widehat{\Delta G})$ est une approximation fondée sur la simulation de l'erreur quadratique moyenne réelle de $\widehat{\Delta G}$, obtenue d'après une exécution indépendante de 100 000 simulations. Comme mesure de stabilité de $v(\widehat{\Delta G})$, nous utilisons la stabilité relative

$$\text{SR}\left\{v\left(\widehat{\Delta G}\right)\right\} = \frac{\left[B^{-1}\sum_{b=1}^{B}\left\{v\left(\widehat{\Delta G}\right) - \text{EQM}\left(\widehat{\Delta G}\right)\right\}^2\right]^{1/2}}{\text{EQM}\left(\widehat{\Delta G}\right)}. \tag{4.4}$$

Enfin, nous avons comparé les taux de couverture de i) l'intervalle de confiance fondé sur l'hypothèse de normalité avec utilisation de l'estimateur de variance par linéarisation et ii) l'intervalle de confiance bootstrap par la méthode des percentiles. Les estimateurs de variance bootstrap et les intervalles de confiance bootstrap sont fondés sur $C = 1\,000$ répliques bootstrap. Les taux d'erreur des intervalles de confiance (avec un taux d'erreur nominal unilatéral de 2,5 % dans chaque queue) sont comparés. La comparaison avec le taux d'erreur nominal de 5 % n'a donné aucune différence qualitative et est donc omise.

Les résultats sont présentés au tableau 4.3. Les estimateurs de variance présentent tous deux un biais négatif. Ce biais est modéré (moins de 5 %) dans la plupart des cas, sauf pour la plus petite taille d'échantillon $n = 1\,000$, et pour la population U_5 possédant la valeur la plus élevée de ΔG. L'estimateur de variance par bootstrap présente systématiquement un biais légèrement plus important que l'estimateur de variance par linéarisation, mais la différence diminue à mesure que la taille d'échantillon augmente. Pour les deux estimateurs de variance, l'instabilité augmente avec ΔG. L'estimateur de variance par bootstrap est légèrement plus stable pour la taille d'échantillon la plus petite $n = 1\,000$, mais la situation est inversée quand la taille d'échantillon augmente. Pour ce qui est de la couverture des intervalles de confiance, les deux méthodes aboutissent à une sous-couverture qui concorde avec le biais négatif des deux estimateurs de variance. Les intervalles de confiance fondés sur l'hypothèse de normalité donnent une couverture un peu meilleure que les intervalles de confiance bootstrap par la méthode des percentiles. Pour les deux catégories d'intervalles de confiance, la sous-couverture est plus importante quand ΔG augmente, et diminue quand la taille d'échantillon augmente.

Tableau 4.3
Biais relatif, stabilité relative et taux d'erreur nominaux unilatéraux pour l'estimation de la variance par linéarisation et par bootstrap de l'estimateur « intersection » de la variation de l'indice de Gini pour 5 populations sous le plan d'échantillonnage SI2

Pop.	Linéarisation					Bootstrap				
	BR	SR	I	S	I+S	BR	SR	I	S	I+S
Taille d'échantillon $(n_{1\bullet}, n_3, n_{2\bullet})$ = (1 000; 1 000; 1 000)										
Pop. 1	-1,41	24,6	1,8	4,5	6,3	-1,83	24,6	1,8	4,9	6,7
Pop. 2	-1,98	32,4	1,6	5,2	6,8	-2,64	32,1	1,7	5,9	7,6
Pop. 3	-2,80	41,9	1,3	6,3	7,7	-3,83	40,9	1,3	7,0	8,3
Pop. 4	-4,00	52,5	1,0	7,7	8,7	-5,57	50,6	1,1	8,2	9,3
Pop. 5	-5,80	64,0	1,0	9,2	10,1	-8,11	60,6	0,8	9,9	10,7
Taille d'échantillon $(n_{1\bullet}, n_3, n_{2\bullet})$ = (1 000; 2 000; 1 000)										
Pop. 1	-1,38	17,3	1,6	3,7	5,3	-1,67	17,8	1,8	4,1	5,9
Pop. 2	-1,64	23,0	1,4	4,3	5,8	-2,05	23,2	1,4	4,7	6,1
Pop. 3	-1,99	30,1	1,2	5,0	6,2	-2,58	30,0	1,1	5,3	6,4
Pop. 4	-2,50	38,4	1,0	6,0	6,9	-3,38	37,9	1,0	6,3	7,3
Pop. 5	-3,30	47,9	0,7	7,2	7,9	-4,62	46,7	0,7	7,5	8,2
Taille d'échantillon $(n_{1\bullet}, n_3, n_{2\bullet})$ = (1 000; 4 000; 1 000)										
Pop. 1	-0,60	11,9	2,0	3,4	5,3	-0,68	12,8	2,1	3,4	5,5
Pop. 2	-0,67	15,9	1,8	3,7	5,6	-0,80	16,5	2,0	3,9	5,9
Pop. 3	-0,83	20,8	1,8	4,4	6,2	-1,03	21,3	1,9	4,4	6,3
Pop. 4	-1,13	26,7	1,5	5,0	6,6	-1,46	26,9	1,6	5,0	6,6
Pop. 5	-1,64	33,4	1,4	5,8	7,1	-2,18	33,5	1,4	5,8	7,1

4.4 Comparaison de la linéarisation et du bootstrap pour le plan MULT2

À la présente section, nous comparons la linéarisation et le bootstrap pour l'estimation de la variance et la production des intervalles de confiance, dans le cas de l'estimateur « intersection » pour la variation de l'indice de Gini sous le plan d'échantillonnage MULT2 présenté à la section 3.1.2. Pour chaque population, nous avons tiré $B = 10\,000$ échantillons à deux degrés bidimensionnels selon le plan MULT2 indexé par $m = 300$ et $(n_{1\bullet}^i, n_3^i, n_{2\bullet}^i) = (10; 10; 10)$, (10; 20; 10) ou (10; 40; 10). Dans chaque échantillon, nous avons calculé l'estimateur « intersection » $\widehat{\Delta G}^{\text{int}}$ de la variation de l'indice de Gini. Pour cet estimateur, nous avons calculé i) l'estimateur de variance par linéarisation $v^{\text{HH}}\{\widehat{\Delta G}^{\text{co}}(a,b)\}$ donné en (3.26), et

ii) l'estimateur de variance par bootstrap $v_{\mathrm{BWR}}(\widehat{\Delta G}^{\mathrm{int}})$, obtenu selon la procédure bootstrap décrite à la section 3.4.2.

Pour mesurer le biais d'un estimateur de variance $v(\widehat{\Delta G})$, nous avons utilisé le biais relatif Monte Carlo en pourcentage défini par l'équation (4.3), et la stabilité relative définie par l'équation (4.4). L'erreur quadratique moyenne réelle de $\widehat{\Delta G}$ a été obtenue par une exécution indépendante de 100 000 simulations. En outre, nous avons comparé les taux de couverture de i) l'intervalle de confiance fondé sur l'hypothèse de normalité avec utilisation de l'estimateur de variance par linéarisation et ii) l'intervalle de confiance bootstrap par la méthode des percentiles. Les estimateurs de variance par bootstrap et les intervalles de confiance par bootstrap sont fondés sur $C = 1\,000$ répliques bootstrap. Les taux d'erreur des intervalles de confiance (avec un taux d'erreur nominal unilatéral de 2,5 % dans chaque queue) sont comparés. La comparaison avec le taux d'erreur nominal de 5 % n'a produit aucune différence qualitative et est donc omise.

Les résultats sont présentés au tableau 4.4. Les estimateurs de variance sont tous deux approximativement sans biais pour les petites valeurs de ΔG, mais présentent un biais négatif modéré qui augmente avec ΔG. L'estimateur de variance par bootstrap présente un biais plus important que l'estimateur de variance par linéarisation. Pour les deux estimateurs de variance, l'instabilité augmente avec ΔG. L'estimateur de variance par bootstrap est légèrement plus stable que l'estimateur de variance par linéarisation. Les deux méthodes donnent lieu à une sous-couverture qui concorde avec le biais négatif des deux estimateurs de variance. Les intervalles de confiance fondés sur l'hypothèse de normalité donnent des résultats légèrement meilleurs. Pour les deux intervalles de confiance, la sous-couverture est plus importante quand ΔG augmente, et diminue quand la taille d'échantillon augmente.

Tableau 4.4
Biais relatif, stabilité relative et taux d'erreur nominaux unilatéraux pour l'estimation de la variance par linéarisation et par bootstrap de l'estimateur « intersection » de la variation de l'indice de Gini pour 5 populations sous le plan d'échantillonnage MULT2

Pop.	Linéarisation					Bootstrap				
	BR	SR	I	S	I+S	BR	SR	I	S	I+S
Tailles d'échantillon $m = 300$ et $(n_{1\bullet}^i, n_{3}^i, n_{2\bullet}^i) = (10; 10; 10)$										
Pop. 1	1,23	33,8	0,6	4,9	5,5	1,09	33,2	0,6	6,0	6,6
Pop. 2	0,64	41,1	0,8	5,5	6,3	-0,20	39,7	0,6	6,5	7,1
Pop. 3	-0,42	48,7	0,7	7,1	7,8	-2,05	46,6	0,7	8,4	9,1
Pop. 4	-2,07	56,4	0,8	8,4	9,2	-4,47	53,3	0,6	9,6	10,2
Pop. 5	-4,44	63,7	0,9	9,2	10,1	-7,56	59,5	0,4	10,3	10,7
Tailles d'échantillon $m = 300$ et $(n_{1\bullet}^i, n_{3}^i, n_{2\bullet}^i) = (10; 20; 10)$										
Pop. 1	1,70	32,6	1,5	4,9	6,4	-1,70	32,3	1,5	6,0	7,5
Pop. 2	1,10	39,0	1,4	5,4	6,8	-1,91	38,3	1,5	6,9	8,4
Pop. 3	0,17	45,6	1,2	7,4	8,6	-2,49	44,4	1,1	7,7	8,8
Pop. 4	-1,17	52,0	1,0	9,0	10,0	-3,58	50,3	0,8	9,7	10,5
Pop. 5	-3,03	57,9	0,9	10,4	11,3	-5,35	55,4	0,7	11,0	11,7
Tailles d'échantillon $m = 300$ et $(n_{1\bullet}^i, n_{3}^i, n_{2\bullet}^i) = (10; 40; 10)$										
Pop. 1	-0,99	32,1	1,2	6,1	7,3	-3,21	32,2	1,7	6,7	8,4
Pop. 2	-1,68	38,3	1,4	6,7	8,1	-3,70	38,3	1,4	7,6	9,0
Pop. 3	-2,58	44,6	1,3	7,5	8,8	-4,40	44,5	1,2	8,9	10,1
Pop. 4	-3,78	50,6	1,1	8,9	10,0	-5,50	50,1	0,9	10,6	11,5
Pop. 5	-5,39	55,9	0,8	10,9	11,7	-7,16	54,8	0,6	12,8	13,4

5 Conclusion

Dans le présent article, nous considérons l'estimation de l'évolution de l'indice de Gini. Nous avons présenté la classe d'estimateurs composites introduite par Goga et coll. (2009), et étudié plus particulièrement l'estimateur « intersection » qui utilise l'échantillon commun seulement, et l'estimateur « union » qui utilise les échantillons disponibles complets. Nous avons justifié tant de manière heuristique que par une étude en simulation décrite à la section 4.2 que l'estimateur « intersection » peut être proche de l'estimateur optimal, tandis que l'estimateur « union » présente des propriétés médiocres dans tous les scénarios considérés. L'estimateur « intersection » est également facile à calculer, tandis que l'estimateur optimal fait intervenir des quantités inconnues qui doivent être estimées en pratique. Par conséquent, nous recommandons l'utilisation de l'estimateur « intersection » pour estimer la variation de l'indice de Gini.

Nous avons également comparé les méthodes de linéarisation et de bootstrap pour estimer la variance et pour produire les intervalles de confiance. Dans les scénarios que nous avons considérés dans l'étude en simulation, la linéarisation donnait de meilleurs résultats que le bootstrap, les biais relatifs étant habituellement plus faibles pour l'estimateur de variance, et les taux de couverture étant meilleurs pour les intervalles de confiance fondés sur l'hypothèse de normalité que pour ceux fondés sur la méthode des percentiles. Les intervalles de confiance fondés sur le bootstrap $-t$ (non examinés dans l'étude en simulation) seraient une option intéressante, mais étant donné l'importance des calculs requis, ils sont moins attractifs pour un utilisateur des données. La linéarisation présente aussi l'avantage d'offrir une approche unifiée convenant pour tout plan d'échantillonnage, alors qu'un plan d'échantillonnage donné requiert habituellement une procédure bootstrap particulière, comme l'illustre l'application du bootstrap sans remise (BWO) pour l'échantillonnage SI et du bootstrap avec remise (BWR) pour l'échantillonnage à plusieurs degrés.

L'étude par simulations nous permet de constater que les taux de couverture ne sont parfois bien respectés ni dans le cas de la linéarisation ni dans celui du bootstrap, particulièrement dans le contexte de l'échantillonnage à plusieurs degrés et même si l'on utilise de grandes tailles d'échantillon. Des intervalles de confiance donnant de meilleurs taux de couverture tout en demandant un temps de calcul raisonnable sont nécessaires. Cette question devrait faire l'objet d'une étude plus approfondie.

Remerciements

Nous remercions Anne Ruiz-Gazen pour des discussions utiles. Nous remercions également deux examinateurs et un rédacteur associé pour leurs commentaires et suggestions constructifs qui nous ont permis d'améliorer considérablement le manuscrit.

Annexe

Preuve de l'équation (3.6)

Partant de (3.3), nous avons $\widehat{\Delta t}^{\,\mathrm{co}} = N(A^{\top}X)$, où $X = (\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3}, \overline{y}_{1,s_1\bullet} - \overline{y}_{1,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3})^{\top}$ et $A = (b, -a, 1)^{\top}$. Cela mène à

$$V\left\{\widehat{\Delta t}^{\,\mathrm{co}}\right\} = N^2\left\{A^\top V(X)\,A\right\}. \tag{A.1}$$

Nous calculons les éléments de $V(X)$ séparément. Nous avons

$$\begin{aligned} V(\overline{y}_{2,s_3} - \overline{y}_{1,s_3}) &= \left(\frac{1}{n_3} - \frac{1}{N}\right) S^2_{y_2-y_1,U} \\ &= \left(\frac{1}{n_3} - \frac{1}{N}\right)\left(S^2_{y_2,U} + S^2_{y_1,U} - 2S_{y_1y_2,U}\right). \end{aligned}$$

En outre, puisque $E(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3} \mid s_2) = 0,$ nous avons

$$\begin{aligned} V(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3}) &= \mathrm{EV}(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3} \mid s_2), \\ &= \mathrm{EV}\left(\frac{n_2}{n_{2\bullet}}\,\overline{y}_{2,s_2} - \frac{n_3}{n_{2\bullet}}\,\overline{y}_{2,s_3} - \overline{y}_{2,s_3} \mid s_2\right) \\ &= \left(1 + \frac{n_3}{n_{2\bullet}}\right)^2 \mathrm{EV}(\overline{y}_{2,s_3} \mid s_2) \\ &= \left(1 + \frac{n_3}{n_{2\bullet}}\right)^2 \left(\frac{1}{n_3} - \frac{1}{n_2}\right) S^2_{y_2,U} \\ &= \frac{n_2}{n_3\,(n_2 - n_3)}\,S^2_{y_2,U} \end{aligned}$$

et

$$\begin{aligned} \mathrm{Cov}(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3}) &= \mathrm{ECov}(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3} \mid s_2) \\ &= \mathrm{ECov}\left(\frac{n_2}{n_{2\bullet}}\,\overline{y}_{2,s_2} - \frac{n_3}{n_{2\bullet}}\,\overline{y}_{2,s_3} - \overline{y}_{2,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3} \mid s_2\right) \\ &= -\left(1 + \frac{n_3}{n_{2\bullet}}\right)\mathrm{ECov}(\overline{y}_{2,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3} \mid s_2) \\ &= \left(1 + \frac{n_3}{n_{2\bullet}}\right)\left(\frac{1}{n_3} - \frac{1}{n_2}\right)\left(S^2_{y_2,U} - S_{y_1y_2,U}\right) \\ &= -\frac{1}{n_3}\left(S^2_{y_2,U} - S_{y_1y_2,U}\right). \end{aligned}$$

Des arguments similaires mènent à

$$\begin{aligned} V(\overline{y}_{1,s_1\bullet} - \overline{y}_{1,s_3}) &= \frac{n_1}{n_3\,(n_1 - n_3)}\,S^2_{y_1,U}, \\ \mathrm{Cov}(\overline{y}_{1,s_1\bullet} - \overline{y}_{1,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3}) &= \frac{1}{n_3}\left(S^2_{y_1,U} - S_{y_1y_2,U}\right). \end{aligned}$$

Enfin, nous considérons $\mathrm{Cov}(\bar{y}_{2,s_2\bullet} - \bar{y}_{2,s_3}, \bar{y}_{1,s_1\bullet} - \bar{y}_{1,s_3})$. Nous commençons par calculer $\mathrm{Cov}(\bar{y}_{2,s_2\bullet}, \bar{y}_{1,s_1\bullet})$, qui peut s'écrire sous la forme

$$\begin{aligned}
\mathrm{Cov}(\bar{y}_{2,s_2\bullet}, \bar{y}_{1,s_1\bullet}) &= \mathrm{Cov}(E(\bar{y}_{2,s_2\bullet} \mid s_{1\bullet}), E(\bar{y}_{1,s_1\bullet} \mid s_{1\bullet})) \\
&= \mathrm{Cov}(\bar{y}_{2,U\setminus s_1\bullet}, \bar{y}_{1,s_1\bullet}) \\
&= \mathrm{Cov}\left(\frac{N}{N-n_{1\bullet}}\bar{y}_{2,U} - \frac{n_{1\bullet}}{N-n_{1\bullet}}\bar{y}_{2,s_1\bullet}, \bar{y}_{1,s_1\bullet}\right) \\
&= -\frac{n_{1\bullet}}{N-n_{1\bullet}}\mathrm{Cov}(\bar{y}_{2,s_1\bullet}, \bar{y}_{1,s_1\bullet}) \\
&= -\frac{n_{1\bullet}}{N-n_{1\bullet}}\left(\frac{1}{n_{1\bullet}} - \frac{1}{N}\right)S_{y_1y_2,U} \\
&= -\frac{1}{N}S_{y_1y_2,U}.
\end{aligned}$$

Des arguments similaires mènent à

$$\mathrm{Cov}(\bar{y}_{2,s_2\bullet}, \bar{y}_{1,s_3}) = \mathrm{Cov}(\bar{y}_{2,s_3}, \bar{y}_{1,s_1\bullet}) = -\frac{1}{N}S_{y_1y_2,U}.$$

Nous obtenons

$$\begin{aligned}
\mathrm{Cov}(\bar{y}_{2,s_2\bullet} - \bar{y}_{2,s_3}, \bar{y}_{1,s_1\bullet} - \bar{y}_{1,s_3}) &= \frac{1}{N}S_{y_1y_2,U} + \mathrm{Cov}(\bar{y}_{2,s_3}, \bar{y}_{1,s_3}) \\
&= \frac{1}{N}S_{y_1y_2,U} + \left(\frac{1}{n_3} - \frac{1}{N}\right)S_{y_1y_2,U} \\
&= \frac{1}{n_3}S_{y_1y_2,U}.
\end{aligned}$$

En résumé, nous obtenons

$$V(X) = \begin{pmatrix}
\frac{n_2}{n_3(n_2-n_3)}S^2_{y_2,U} & \frac{1}{n_3}S_{y_1y_2,U} & -\frac{1}{n_3}\left(S^2_{y_2,U} - S_{y_1y_2,U}\right) \\
 & \frac{n_1}{n_3(n_1-n_3)}S^2_{y_1,U} & \frac{1}{n_3}\left(S^2_{y_1,U} - S_{y_1y_2,U}\right) \\
 & & \left(\frac{1}{n_3} - \frac{1}{N}\right)\left(S^2_{y_2,U} + S^2_{y_1,U} - 2S_{y_1y_2,U}\right)
\end{pmatrix}$$

qui, avec (A.1), mène à (3.6).

Bibliographie

Antal, E., et Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106, 534-543.

Barrett, G.F., et Donald, S.G. (2009). Statistical inference with generalized Gini indices of inequality, poverty, and welfare. *Journal of Business and Economic Statistics*, 27, 1-17.

Beaumont, J.-F., et Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *Revue Internationale de Statistique*, 80, 127-148.

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

Berger, Y.G. (2008). A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *Journal of Official Statistics*, 24, 541-555.

Bertail, P., et Combris, P. (1997). Bootstrap généralisé d'un sondage. *Annales d'Économie et de Statistique*, 46, 49-83.

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics*, 137, 674-707.

Bickel, P.J., et Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

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

Brändén, P., et Jonasson, J. (2012). Negative dependence in sampling. *Scandinavian Journal of Statistics*, 39, 830-838.

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

Chao, M.-T., et Lo, S.-H. (1985). A Bootstrap method for finite population. *Sankhyā, Series A*, 47, 3, 399-405.

Chauvet, G. (2007). Méthodes de Bootstrap en population finie. Thèse de doctorat, Université Rennes 2.

Chauvet, G. (2015). Coupling methods for multistage sampling. *The Annals of Statistics*, 43(6), 2484-2506.

Chen, J., et Rao, J.N.K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, 17, 1047-1064.

Davison, A.C., et Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Davison, A.C., et Sardy, S. (2007). Resampling variance estimation in surveys with missing data. *Journal of Official Statistics*, 23, 3, 371-386.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 1, 17-27. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2004001/article/6991-fra.pdf.

Deville, J.-C. (1997). Estimation de la variance de l'indice de Gini mesurée par sondage. *Actes des Journées de Méthodologie Statistique*, *Insee Méthodes*.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 2, 219-230. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1999002/article/4882-fra.pdf.

Druckman, A., et Jackson, T. (2008). Measuring resource inequalities: The concepts and methodology for an area-based Gini coefficient. *Ecological Economics*, 65, 242-252.

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze Lettere ed Arti.*

Glasser, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

Goga, C. (2003). Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles nonparamétriques. Thèse de doctorat, Université Rennes 2.

Goga, C., et Ruiz-Gazen, A. (2014). Efficient estimation of nonlinear finite population parameters using nonparametrics. *Journal of the Royal Statistical Society B*, 76, 113-140.

Goga, C., Deville, J.-C. et Ruiz-Gazen, A. (2009). Composite estimation and linearization method for two-sample survey data. *Biometrika*, 96, 691-709.

Gordon, L. (1983). Successive sampling in large finite populations. *Annals of Statistics*, 11, 702-706.

Graczyk, P.P. (2007). Gini coefficient: A new way to express selectivity of kinase inhibitors against a family of Kinases. *Journal of Medicinal Chemistry*, 50, 5773-5779.

Gross, S.T. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, 181-184.

Groves-Kirkby, C.J., Denman, A.R. et Phillips, P.S. (2009). Lorenz Curve and Gini coefficient: Novel tools for analysing seasonal variation of environmental radon gas. *Journal of Environmental Management*, 90, 2480-2487.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Tud. Akad. Mat. Kutatò Int. Közl.*, 5, 361-374.

Hájek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *Annals of Mathematical Statistics*, 32, 506-523.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

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

Karagiannis, E., et Kovačević, M.S. (2000). A method to calculate the jackknife variance estimator for the Gini coefficient. *Oxford Bulletin of Economics and Statistics*, 62, 119-122.

Kovačević, M.S., et Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization - The estimating equation approach. *Journal of Official Statistics,* 13, 41-58.

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

Lai, D., Huang, J., Risser, J.M. et Kapadia, A.S. (2008). Statistical properties of generalized Gini coefficient with application to health inequality measurement. *Social Indicator Research,* 87, 249-258.

Langel, M., et Tillé, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society, Series A,* 176, 521-540.

Lisker, T. (2008). Is the Gini coefficient a stable measure on galaxy structure? *The Astrophysical Journal Supplement Series,* 179, 319-325.

Navarro, V., Muntaner, C., Borrell, C., Benach, J., Quiroga, A., Rodríguez-Sanz, M., Vergès, N. et Pasarín, M.I. (2006). Politics and health outcomes. *The Lancet*, 18, 1033-1037.

Nygård, F., et Sandström, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics*, 1, 4, 399-412.

Ohlsson, E. (1986). Asymptotic normality of the Rao-Hartley-Cochran estimator: An application of the martingale CLT. *Scandinavian Journal of Statistics*, 13, 17-28.

Ohlsson, E. (1989). Asymptotic normality for two-stage sampling from a finite population. *Probability Theory and Related Fields*, 81, 341-352.

Pires, A.M., et Branco, J.A. (2002). Partial influence functions. *Journal of Multivariate Analysis,* 83, 451-468.

Presnell, B., et Booth, J.G. (1994). *Resampling Methods for Sample Surveys*. Rapport technique.

Qin, Y., Rao, J.N.K. et Wu, C. (2010). Empirical likelihood confidence intervals for the Gini measure of income inequality. *Economic Modelling,* 27, 1429-1435.

Qualité, L., et Tillé, Y. (2008). Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée. *Techniques d'enquête*, 34, 2, 193-201. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2008002/article/10758-fra.pdf.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association,* 83, 231-241.

Reid, N. (1981). Influence functions for censored data. *The Annals of Statistics*, 9, 78-92.

Rosén, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. I, II. *Annals of Mathematical Statistics*, 43, 373-397, 748-776.

Saegusa, T., et Wellner, J.A. (2013). Weighted likelihood estimation under two-phase sampling. *The Annals of Statistics*, 41, 269-295.

Sandström, A., Wretman, J.H. et Waldèn, B. (1985). Variance estimators of the Gini coefficient - Simple random sampling. *Metron,* 43, 41-70.

Sandström, A., Wretman, J.H. et Waldèn, B. (1988). Variance estimators of the Gini coefficient - Probability sampling. *Journal of Business and Economic Statistics,* 6, 113-119.

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

Sen, P.K. (1980). Limit theorems for an extended coupon collector's problem and for successive subsampling with varying probabilities. *Calcutta Statistical Association Bulletin,* 29, 113-132.

Shao, J., et Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer.

Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association,* 87, 755-765.

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics,* 20, 135-154.

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

Yitzhaki, S. (1991). Calculating jackknife variance estimators for parameters of the Gini method. *Journal of Business and Economic Statistics,* 9, 235-239.

Techniques d'enquête, juin 2018
Vol. 44, nº 1, pp. 45-63
Statistique Canada, nº 12-001-X au catalogue

Étalonnage temporel avec préservation des taux de croissance (PTC) : Inconvénients et autres solutions

Jacco Daalmans, Tommaso Di Fonzo, Nino Mushkudiani et Reinier Bikker[1]

Résumé

L'étalonnage des séries mensuelles et trimestrielles à des données annuelles est une pratique courante adoptée par de nombreux instituts nationaux de statistique. Le problème de l'étalonnage se pose quand des données temporelles pour une même variable cible sont mesurées à différentes fréquences et qu'il est nécessaire d'éliminer les différences entre les sommes des valeurs infra-annuelles et les valeurs annuelles de référence. Plusieurs méthodes d'étalonnage sont décrites dans la littérature. La procédure d'étalonnage avec préservation des taux de croissance (PTC) est souvent considérée comme étant la meilleure. Il est généralement soutenu qu'elle a pour fondement un principe idéal de préservation du mouvement. Toutefois, nous montrons que l'étalonnage PTC présente des inconvénients appréciables qui importent pour les applications pratiques et qui ne sont pas décrits dans la littérature. Nous considérons d'autres modèles d'étalonnage qui ne souffrent pas de certains des effets indésirables de la PTC.

Mots-clés : Étalonnage; préservation des taux de croissance; réconciliation des données; macro-intégration.

1 Introduction

L'étalonnage des séries mensuelles et trimestrielles à des données annuelles est une pratique courante adoptée par de nombreux instituts nationaux de statistique. Par exemple, chaque année *Statistics Netherlands* met en concordance 12 tableaux trimestriels des ressources et des emplois avec les trois comptes annuels les plus récents (Eurostat, 2013, Annexe 8C).

Le problème de l'étalonnage se pose quand des données temporelles pour une même variable cible sont mesurées à différentes fréquences avec différents niveaux de précision. On pourrait s'attendre à ce qu'une relation d'agrégation temporelle entre ces séries soit satisfaite, par exemple, que la somme de quatre valeurs trimestrielles soit égale à une valeur annuelle, mais à cause de différences dans les sources de données et les méthodes de traitement, il est fréquent que ce ne soit pas le cas. L'étalonnage est le processus utilisé pour éliminer de telles différences. Dans ce processus, les valeurs provisoires sont ajustées afin d'établir une cohérence mathématique entre les séries temporelles de faible fréquence (par exemple, annuelles) et de haute fréquence (par exemple, trimestrielles ou mensuelles).

L'étalonnage est régi par deux grands principes. Premièrement, les valeurs de référence de faible fréquence sont fixées, parce que ces sources de données décrivent mieux les niveaux et les tendances de long terme que les sources de fréquence élevée. Deuxièmement, les mouvements de court terme des séries temporelles de haute fréquence sont préservés autant que possible, car ces séries de données sont les seules sources d'information sur les mouvements de court terme.

Plusieurs méthodes d'étalonnage sont décrites dans la littérature. Ces méthodes diffèrent par la façon dont les mouvements de court terme des séries de haute fréquence sont définis. On distingue des méthodes multiplicatives et des méthodes additives. Les méthodes multiplicatives visent à préserver les variations

1. Jacco Daalmans, Statistics Netherlands, CP 24500, 2490 HA La Haye, Pays-Bas. Courriel : j.daalmans@cbs.nl; Tommaso Di Fonzo, Département des sciences statistiques, Université de Padoue, Via Cesare Battisti, 241, 35121 Padoue PD, Italie. Courriel : difonzo@stat.unipd.it; Nino Mushkudiani, Statistics Netherlands, La Haye. Courriel : n.mushkudiani@cbs.nl; Reinier Bikker, Statistics Netherlands, La Haye, Courriel : r.bikker@cbs.nl.

relatives de la série temporelle de haute fréquence provisoire, tandis que les méthodes additives visent à préserver les variations absolues. Dans le présent article, nous nous penchons uniquement sur les variantes multiplicatives.

Deux méthodes multiplicatives bien connues sont celles des premières différences proportionnelles (PDP) de Denton (1971) et de la préservation des taux de croissance (PTC) de Causey et Trager (1981; voir aussi Trager, 1982 et Bozik et Otto, 1988).

Dans la littérature, il est généralement reconnu que la PTC est la méthode dont les fondements théoriques sont les plus solides (Bloem, Dippelsman et Maehle, 2001, page 100). Elle préserve explicitement les taux de variation d'une période à l'autre de la série provisoire. Cependant, la méthode PDP de Denton est d'usage plus répandu, parce qu'elle est techniquement plus facile à appliquer. Mathématiquement, la méthode de Denton comprend un problème d'optimisation quadratique sous des contraintes linéaires classiques, tandis que la méthode PTC requiert la résolution d'un problème non linéaire sous des contraintes linéaires plus difficile, dont la solution peut être obtenue efficacement au moyen d'un algorithme de points intérieurs (Di Fonzo et Marini, 2015).

Un certain nombre d'études en simulation ont montré que la méthode PDP de Denton et la méthode PTC donnent des résultats similaires ou presque similaires dans la grande majorité des cas (Harvill Hood, 2005; Titova, Findley et Monsell, 2010; Di Fonzo et Marini, 2012 et Daalmans et Di Fonzo, 2014). Par conséquent, la méthode PDP de Denton peut être utilisée comme une approximation de la méthode PTC.

L'objectif du présent article est de montrer que la PTC présente des inconvénients qui, autant que nous sachions, ne sont pas décrits dans la littérature. Un premier inconvénient est que le fait d'appliquer l'étalonnage en « avançant » ou en « reculant » dans le temps a de l'importance. Dans ce contexte, nous présenterons un lien avec la propriété de réversibilité du temps issue de la théorie des indices. Un deuxième inconvénient est que des singularités dans la fonction-objectif de la méthode PTC peuvent mener à des résultats indésirables.

Un deuxième objectif de l'article est de présenter d'autres méthodes d'étalonnage qui satisfont la propriété de réversibilité du temps. Le présent article peut être utile aux praticiens qui appliquent ou envisagent d'appliquer des techniques d'étalonnage.

À la section 2, nous commençons par décrire formellement les méthodes d'étalonnage PDP de Denton et PTC. À la section 3, nous décrivons les inconvénients de la méthode PTC. À la section 4, nous proposons deux nouvelles méthodes d'étalonnage qui peuvent être utilisées pour remplacer la méthode PTC. À la section 5, nous donnons les résultats d'une application effectuée sur des données réelles en guise d'illustration. Enfin, à la section 6, nous présentons nos conclusions.

2 Méthodes d'étalonnage des séries temporelles

La présente section explique les procédures d'étalonnage par la méthode PDP de Denton et la méthode PTC. Puisque les contraintes d'agrégation temporelle sont les mêmes pour les deux méthodes, nous commençons par les décrire. Puis, nous expliquons les procédures d'étalonnage par les méthodes PDP de Denton et PTC.

Nous nous concentrons sur les variantes univariées de ces méthodes, dans lesquelles la cohérence temporelle est la principale contrainte d'intérêt. Toutefois, les observations qui sont présentées dans la suite de l'article sont également valides pour le cas multivarié, où plusieurs séries temporelles sont réconciliées simultanément et des contraintes additionnelles entre les séries s'appliquent (voir Di Fonzo et Marini, 2011 et Bikker, Daalmans et Mushkudiani, 2013).

2.1 Notation générale et contraintes temporelles

En général, les contraintes d'agrégation temporelle peuvent être exprimées sous forme d'un système linéaire d'égalités $\mathbf{Ax} = \mathbf{b}$, où $\mathbf{x}$ est le vecteur cible de valeurs de haute fréquence, $\mathbf{b}$ est un vecteur de valeurs de faible fréquence, et $\mathbf{A}$ est une matrice d'agrégation temporelle convertissant les valeurs de haute fréquence en valeurs de faible fréquence.

La forme particulière de ces contraintes dépend de la nature des variables concernées. Pour les variables de flux, une somme de valeurs infra-annuelles, par exemple, quatre valeurs trimestrielles, doit habituellement être la même qu'une valeur annuelle. Pour les variables de stock, l'une des valeurs infra-annuelles, habituellement la première ou la dernière, doit être la même que la valeur annuelle pertinente. Par exemple, pour les variables de flux trimestrielles/annuelles, en supposant pour simplifier que l'intervalle de temps disponible débute au premier trimestre de la première année et se termine au quatrième trimestre de la dernière année observée, la matrice est

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \end{bmatrix}.$$

En désignant par $\mathbf{p}$ un vecteur de valeurs provisoires, en général la situation est $\mathbf{Ap} \neq \mathbf{b}$, sinon aucun ajustement ne serait nécessaire. Nous recherchons un vecteur d'estimations étalonnées $\mathbf{x}^*$, un résultat particulier pour $\mathbf{x}$, qui devrait être « aussi proche que possible » des valeurs provisoires et qui satisfait $\mathbf{Ax}^* = \mathbf{b}$.

Les périodes infra-annuelles ne doivent pas toutes être couvertes par une valeur repère. Donc, le nombre de lignes dans $\mathbf{A}$ peut être plus petit que le nombre total de périodes annuelles, voir par exemple, Dagum et Cholette (2006) pour plus de renseignements.

Dans un exercice d'étalonnage, les caractéristiques de la série originale $\mathbf{p}$ doivent être prises en considération. Par exemple, dans un cadre de séries temporelles économiques, la préservation de la dynamique temporelle (quelle que soit sa définition) des séries provisoires intéresse souvent beaucoup le praticien.

2.2 Préservation des taux de croissance (PTC) et premières différences proportionnelles (PDP) de Denton

La présente section donne une description formelle des méthodes PTC et PDP de Denton.

Causey et Trager (1981; voir aussi Monsour et Trager, 1979 et Trager, 1982) obtiennent les valeurs étalonnées $x_t^*,\ t = 1, \ldots,\ n$ en tant que solution du problème d'optimisation suivant :

$$\min_{x_t} f_{\mathrm{P}}^{\mathrm{PTC}}(\mathbf{x}) \quad \text{sous les contraintes} \quad \mathbf{Ax} = \mathbf{b}, \quad \text{où} \quad f_{\mathrm{P}}^{\mathrm{PTC}}(\mathbf{x}) = \sum_{t=2}^{n} \left(\frac{x_t}{x_{t-1}} - \frac{p_t}{p_{t-1}} \right)^2. \quad (2.1)$$

Le critère PTC à minimiser, $f_{\mathrm{P}}^{\mathrm{PTC}}(\mathbf{x})$, relie explicitement les taux de croissance : il minimise la somme des carrés des différences entre les taux de croissance des valeurs provisoires et des valeurs étalonnées. L'indice inférieur « P » dans la fonction de minimisation signifie « prospective »; plus loin dans l'article, nous définirons une fonction de minimisation « rétrospective ».

Denton (1971) a proposé une procédure d'étalonnage fondée sur les *premières différences proportionnelles* (PDP) entre la série cible et la série originale. Cholette (1984) a modifié légèrement le résultat de Denton, afin de traiter correctement les conditions de départ du problème. Les estimations étalonnées dans la méthode PDP sont donc obtenues en tant que solution du problème de minimisation quadratique sous contraintes

$$\min_{x_t} f_{\mathrm{P}}^{\mathrm{PDP}}(\mathbf{x}) \quad \text{sous les contraintes} \quad \mathbf{Ax} = \mathbf{b}, \quad \text{où} \quad f_{\mathrm{P}}^{\mathrm{PDP}}(\mathbf{x}) = \sum_{t=2}^{n} \left(\frac{x_t}{p_t} - \frac{x_{t-1}}{p_{t-1}} \right)^2. \quad (2.2)$$

Le critère PDP de Denton à minimiser, $f_{\mathrm{P}}^{\mathrm{PDP}}(\mathbf{x})$, est une somme de termes linéaires au carré, qui est plus facile à manipuler que la fonction-objectif PTC non linéaire.

3 Deux problèmes associés à l'étalonnage PTC

3.1 Réversibilité du temps

La réversibilité du temps signifie que cela n'a pas d'importance qu'une méthode soit appliquée prospectivement ou rétrospectivement. Cette propriété peut être intéressante dans de nombreux domaines d'application.

En physique, cela signifie que si le temps se déroulait à reculons, tous les mouvements seraient inversés. En théorie des indices, la réversibilité du temps a été introduite dans un ouvrage classique de Fisher (1922, page 64). Il y est dit que [Traduction] « si, en prenant 1913 comme base et en avançant vers 1918, nous constatons que, en moyenne, les prix ont doublé, alors, en procédant dans la direction inverse, nous devons trouver que le niveau des prix de 1913 est la moitié de celui de 1918 ». L'élément qui motive l'application de ce principe est que la direction du temps peut être considérée comme étant arbitraire, c'est-à-dire qu'il n'existe aucune direction privilégiée naturellement.

Le principe de la réversibilité du temps peut également être appliqué dans le contexte de l'étalonnage. Cela signifie que, si nous inversons une série temporelle, appliquons l'étalonnage et inversons de nouveau la série étalonnée, nous obtiendrons exactement les mêmes résultats que pour l'étalonnage de la série originale. Autrement dit, les résultats étalonnés ne permettent pas de voir si l'étalonnage a été appliqué prospectivement ou rétrospectivement.

L'étalonnage d'une série temporelle inversée, conformément aux méthodes PTC et PDP de Denton, respectivement, équivaut à minimiser les fonctions-objectif suivantes

$$f_{\mathrm{R}}^{\mathrm{PTC}}(\mathbf{x}) = \sum_{t=2}^{n}\left(\frac{x_{t-1}}{x_t} - \frac{p_{t-1}}{p_t}\right)^2 \tag{3.1}$$

et

$$f_{\mathrm{R}}^{\mathrm{PDP}}(\mathbf{x}) = \sum_{t=2}^{n}\left(\frac{x_{t-1}}{p_{t-1}} - \frac{x_t}{p_t}\right)^2, \tag{3.2}$$

où l'indice inférieur « R » signifie « rétrospective ». Ces fonctions-objectif sont obtenues à partir des fonctions-objectif prospectives en interchangeant t et $t-1$. À partir d'ici, la minimisation de (3.1) ou (3.2) sera appelée « étalonnage rétrospectif », à l'opposé de l'étalonnage prospectif, classique.

Comme nous venons de le mentionner, une méthode d'étalonnage satisfait la propriété de réversibilité du temps si les étalonnages prospectif et rétrospectif aboutissent aux mêmes résultats. Il est facile de voir que $f_{\mathrm{P}}^{\mathrm{PTC}}(\mathbf{x}) \neq f_{\mathrm{R}}^{\mathrm{PTC}}(\mathbf{x})$, tandis que $f_{\mathrm{P}}^{\mathrm{PDP}}(\mathbf{x}) = f_{\mathrm{R}}^{\mathrm{PDP}}(\mathbf{x})$. Il découle de cela que la méthode PDP de Denton satisfait la propriété de réversibilité du temps, tandis que la méthode PTC ne le fait pas.

Plus concrètement, dans de nombreux processus de production, on applique un étalonnage « prospectif », par exemple pour réconcilier les tableaux des ressources et des emplois des Pays-Bas (Bikker et coll., 2013). Cependant, après une révision, la série temporelle révisée peut être construite « rétrospectivement », en utilisant des fonctions-objectif rétrospectives. Il est fortement indésirable qu'existent dans les résultats des différences attribuables purement à une différence de « direction temporelle ». Les praticiens qui ne sont pas conscients de la propriété de réversibilité du temps peuvent appliquer des méthodes d'étalonnage prospective et rétrospective, et supposer erronément que les deux méthodes mènent aux mêmes résultats.

Même s'il est vrai que toute application d'étalonnage peut être restreinte à la préservation des taux de croissance prospectifs, il est indésirable que les résultats soient affectés par la propriété non pertinente de direction du temps. Par conséquent, toute méthode d'étalonnage devrait, de préférence, satisfaire le principe de réversibilité du temps. De surcroît, la sous-section 3.3 illustre le fait qu'une méthode d'étalonnage qui n'est pas symétrique dans le temps peut modifier le moment où ont lieu les événements économiques les plus importants, par exemple, les pics et les creux qui marquent le début et la fin d'une crise.

3.2 Singularité

Un second problème que pose la PTC découle de la singularité de sa fonction-objectif. Si x_{t-1} tend vers zéro dans le cas d'un étalonnage prospectif (ou x_t dans le cas d'un étalonnage rétrospectif), la valeur de la fonction-objectif tend vers l'infini, ce qui cause plusieurs difficultés.

L'une des complications est que le problème d'optimisation devient instable, une faible variation dans les valeurs provisoires pouvant donner lieu à un important changement dans les valeurs étalonnées.

Conséquemment, des révisions indésirablement grandes peuvent être obtenues quand on effectue l'étalonnage de données mises à jour.

Une autre complication est que, puisqu'une correction des valeurs proches de zéro peut être lourdement pénalisée, les taux de croissance de ces valeurs sont fortement préservés. Néanmoins, cela peut avoir pour prix des corrections relativement importantes des autres taux de croissance. Par ailleurs, on peut soutenir que les taux de croissance contiennent peu d'information pour les valeurs extrêmement petites (proches de zéro). D'où, la préservation des taux de croissance peut être jugée inappropriée dans ce cas. La sous-section 5.3 offre un exemple réel de ce problème.

Une troisième complication est que, comme les valeurs étalonnées proches de zéro peuvent entraîner une grande valeur de la fonction-objectif, les méthodes PTC ont tendance à éviter de telles valeurs, ce qui peut donner des profils de correction irréguliers. En particulier, un problème dans lequel toutes les valeurs provisoires sont positives peut donner des valeurs étalonnées négatives. Considérons un exemple dans lequel deux valeurs consécutives égalent 100. Alors, un ajustement de la première valeur de 100 à -100 est moins onéreux en ce qui concerne la valeur de la fonction-objectif de la PTC qu'une correction allant de 100 à 30. Les valeurs correspondantes de la fonction-objectif sont $((100/-100) - (100/100))^2 = 4$ et $((100/30) - (100/100))^2 = 5{,}44$. Toutefois, une valeur qui passe d'un grand chiffre positif à un grand chiffre négatif ne sera habituellement pas considérée comme une bonne préservation du mouvement. Donc, l'exemple montre aussi qu'il est contestable d'utiliser les taux de croissance en présence de valeurs positives et négatives.

Par conséquent, il peut être indiqué d'éviter les résultats négatifs grâce à des contraintes de non-négativité, voir la sous-section 4.1 pour des renseignements plus détaillés. La méthode PDP de Denton est moins susceptible de produire des valeurs négatives. Dans l'exemple précédent, un ajustement de 100 à 30 est préféré à un ajustement de 100 à -100. Un exemple réel de ce problème est présenté à la sous-section 5.3.

Même si la singularité de la fonction-objectif de la méthode PTC peut donner lieu à des valeurs étalonnées négatives, il ne s'agit pas de la seule cause. La méthode PDP de Denton peut aussi produire des valeurs négatives. En général, des valeurs étalonnées négatives risquent d'être obtenues quand la variation (relative) d'une valeur repère à une autre diffère de manière significative de la variation (relative) par rapport aux valeurs provisoires annualisées sous-jacentes.

Une quatrième complication liée à la fonction-objectif singulière de la méthode PTC est que des pics et des creux irréguliers peuvent apparaître dans une série temporelle étalonnée. Cela tient au fait que, dans la méthode PTC classique, une correction d'une grande valeur positive pour passer à une valeur proche de zéro est moins coûteuse en ce qui concerne la valeur de la fonction-objectif qu'une correction opposée pour passer d'une valeur proche de zéro à une grande valeur positive. Autrement dit, une correction d'un taux de croissance g par un facteur c, où $c > 1$, correspond à une plus grande valeur de la fonction-objectif qu'une correction par $1/c$, surtout si c est grand. Les valeurs de la fonction-objectif sont $((c-1)\,g)^2$ et $\left(\frac{(c-1)}{c}\,g\right)^2$, respectivement. Puisque de grandes corrections à la hausse à partir d'une valeur proche de zéro sont relativement coûteuses, elles sont évitées dans la mesure du possible. Donc, les valeurs étalonnées par la méthode PTC évoluent plus graduellement à partir d'une valeur proche de zéro que les résultats de la méthode de Denton. Pour corriger cette situation, de plus grands pics peuvent être nécessaires aux périodes suivantes pour satisfaire la contrainte d'agrégation temporelle. Puisque l'étalonnage vise habituellement à

produire des corrections aussi lisses que possible au cours du temps, les pics irréguliers peuvent être jugés indésirables. À la relativement faible croissance à partir d'une valeur proche de zéro est associé le fait que les pics ont tendance à apparaître plus tard dans le temps que dans le cas d'une méthode temporellement symétrique telle que la méthode PDP de Denton. Dans le cas de la variante rétrospective de la méthode PTC, l'opposé se produit, les séries temporelles étalonnées évoluant assez rapidement à partir d'une valeur proche de zéro, ce qui produit des pics relativement tôt. L'exemple de la sous-section 3.3 illustre ce problème.

3.3 Exemple

Suit un exemple qui illustre les problèmes que posent les méthodes PTC. Dans cet exemple, une série temporelle comprenant 15 mois est réconciliée à cinq valeurs trimestrielles. La série mensuelle est constante, chaque valeur mensuelle étant égale à 10. Les valeurs trimestrielles sont 80, 250, 80, 400 et 100, respectivement. La figure 3.1 permet de comparer les résultats de la méthode PDP de Denton, la méthode PTC prospective (PTCP) et la méthode PTC rétrospective (PTCR).

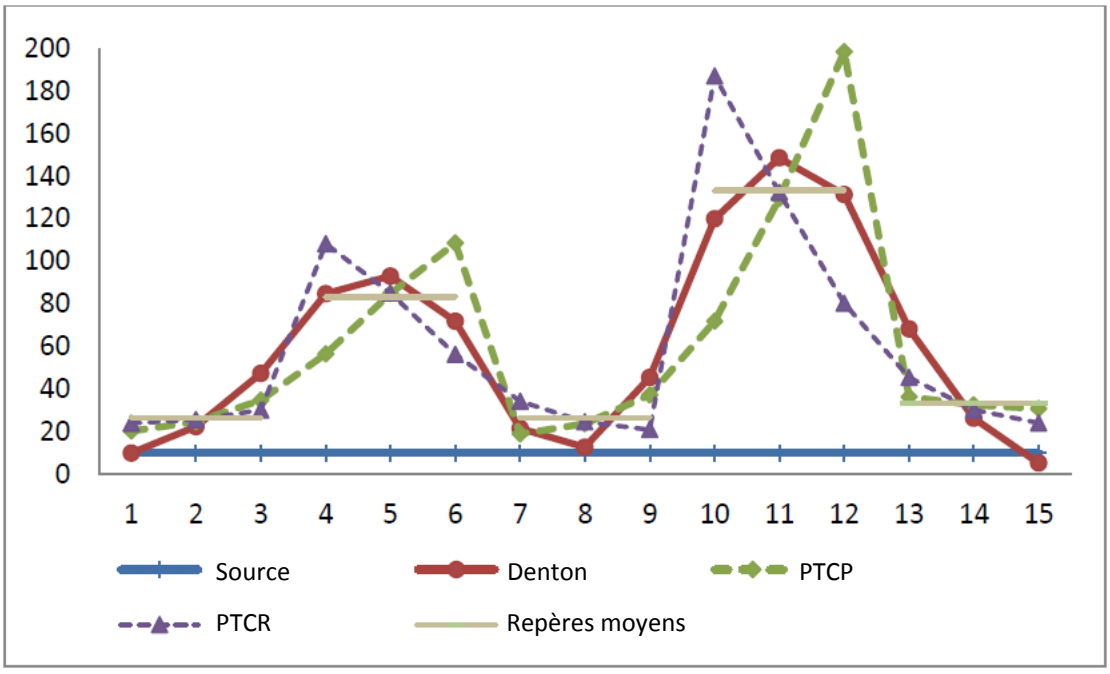

Figure 3.1 Exemple : Résultats de trois méthodes d'étalonnage. « Repère moyen » désigne le niveau moyen des valeurs mensuelles qui concordent avec les repères trimestriels et qui est calculé comme étant le tiers de son équivalent trimestriel.

Vu que les différences les plus importantes s'observent entre les deux méthodes PTC, le principe de réversibilité du temps n'est manifestement pas satisfait. Les points le plus élevé et le plus bas apparaissent à des mois différents. L'exemple montre clairement que l'utilisation d'une méthode d'étalonnage différente peut donner lieu à des conclusions considérablement différentes.

Comme il est indiqué à la sous-section 3.2, la méthode PTCP mène à des pics relativement tardifs, c'est-à-dire au dernier mois de chaque trimestre, tandis que la méthode PTCR produit des pics tôt, c'est-à-dire le premier mois de chaque trimestre. Les résultats de la méthode PDP de Denton se situent entre les deux, les pics et les creux se manifestant durant le mois du milieu de chaque trimestre.

Il convient toutefois de souligner que l'exemple ne peut pas être considéré comme étant représentatif des applications réelles. En général, les méthodes d'étalonnage ne sont pas destinées à être utilisées pour réconcilier des écarts importants ni pour des séries infra-annuelles constantes. Pour expliquer ce dernier point, une hypothèse importante de la méthode PDP de Denton est que les séries infra-annuelles fournissent de l'information sur les variations de court terme. Toutefois, une série constante ne peut pas être considérée comme très informative. Néanmoins, la question de la réconciliation de séries temporelles constantes se pose dans des problèmes étroitement liés à l'étalonnage, comme l'interpolation et la calendrialisation (voir par exemple, Dagum et Cholette, 2006 et Boot, Feibes et Lisman, 1967). Le présent exemple est choisi pour des raisons purement éducatives. Il donne un bon aperçu des propriétés des différents types de fonctions-objectif. Le lecteur est invité à consulter la sous-section 5.3 pour des exemples plus réalistes.

4 Autres techniques d'étalonnage

À la section 3, nous avons décrit deux problèmes liés aux méthodes PTC. À la présente section, nous considérons deux techniques d'étalonnage de rechange qui résolvent la question de la propriété d'irréversibilité du temps.

4.1 Préservation simultanée des taux de croissance

Ici, nous proposons deux autres fonctions-objectif pour la PTC. La première est une variante « temporellement symétrique » de la PTC, définie par

$$f_S^{\mathrm{PTC}}(\mathbf{x}) = \frac{1}{2}\sum_{t=2}^{n}\left(\frac{x_t}{x_{t-1}} - \frac{p_t}{p_{t-1}}\right)^2 + \frac{1}{2}\sum_{t=2}^{n}\left(\frac{x_{t-1}}{x_t} - \frac{p_{t-1}}{p_t}\right)^2, \tag{4.1}$$

où l'indice inférieur « S » signifie « simultanée ». La méthode sera appelée PTCS dans la suite de l'exposé. La fonction-objectif de la méthode PTCS préserve à la fois les taux de croissance prospectifs et rétrospectifs. Autant que sachent les auteurs, cette méthode n'a été mentionnée nulle part ailleurs dans la littérature. Il est facile de voir que la PTCS satisfait le principe de réversibilité du temps : la permutation de t et $t-1$ ne modifie pas la fonction-objectif.

Cependant, le second problème mentionné à la section 3 (singularité de la fonction-objectif) ne doit pas être considéré. L'une des conséquences, les valeurs étalonnées négatives, peut être évitée en imposant une borne inférieure nulle aux valeurs étalonnées. Cela peut se faire par une technique bien établie qui consiste à inclure des contraintes d'inégalité dans un problème d'optimisation (par exemple, Nocedal et Wright, 2006). Les autres problèmes associés à la singularité peuvent néanmoins encore se produire.

4.2 Préservation logarithmique des taux de croissance

Une autre variante « temporellement symétrique » de la méthode PTC est donnée par la forme logarithmique :

$$f_{\mathrm{L}}^{\mathrm{PTC}}(\mathbf{x}) = \sum_{t=2}^{n}\left[\log\left(\frac{x_t}{x_{t-1}}\right) - \log\left(\frac{p_t}{p_{t-1}}\right)\right]^2. \tag{4.2}$$

Cette fonction a été considérée pour la première fois par Helfand, Monsour et Trager (1977). Il est immédiatement vérifié que la fonction (4.2) satisfait la propriété d'inversion du temps également. La fonction-objectif peut être considérée comme la version logarithmique de la méthode PTC, aussi bien que la version logarithmique de la méthode PDP de Denton. Elle sera notée PTCL dans la suite de l'exposé, où « L » signifie « logarithmique ».

Notons que (4.2) ne peut être utilisée que pour des valeurs provisoires strictement positives, et qu'elle produit des valeurs étalonnées qui sont plus grandes que zéro également. Cela ne semble pas être une limite importante, puisque nous avons déjà mentionné à la section 3 que la préservation des taux de croissance peut être jugée inappropriée pour les problèmes contenant des valeurs positives et négatives. Néanmoins, une solution possible pour les séries temporelles contenant des valeurs négatives consiste à ajouter une constante suffisamment grande à la série avant l'étalonnage et à soustraire cette constante de la série étalonnée. Un inconvénient éventuel de cette solution est que l'ajout d'une constante fausse les taux de croissance initiaux. Donc, on ne sait pas clairement si les taux de croissance provisoires sont effectivement préservés. Une étude plus approfondie est nécessaire pour mieux comprendre les implications de cette solution.

Même si la méthode PTCL donne nécessairement des valeurs positives, d'autres problèmes associés à une fonction-objectif singulière mentionnés à la section 3.2 peuvent encore se produire.

4.3 Comparaison

Lorsque l'on compare les méthodes PTCS et PTCL, on peut s'attendre à ce que la PTCL se comporte davantage comme la méthode PDP de Denton. Voici deux raisons pour lesquelles il en est ainsi.

Premièrement, en raison des propriétés asymptotiques de la fonction logarithmique, le problème d'évitement des valeurs proches de zéro est moins grave dans le cas de la PTCL que dans celui de la PTCS. Les valeurs proches de zéro sont associées à des ajustements importants des taux de croissance. Les très grands ajustements des taux de croissance sont moins pénalisés dans la PTCL que dans la PTCS, puisque la fonction-objectif de la PTCS croît plus rapidement quand les corrections sont importantes.

Deuxièmement, la linéarisation de Taylor d'ordre un de la fonction-objectif de la PTCL correspond à la fonction de la PDP de Denton, tandis que l'approximation de la PTCS mène à un résultat différent. La linéarisation des termes au carré de la fonction-objectif dans les valeurs provisoires est donnée par $\left(\frac{x_t}{p_t}\right) - \left(\frac{x_{t-1}}{p_{t-1}}\right)$ et $\left(\frac{p_t}{P_{t-1}} + \frac{p_{t-1}}{p_t}\right)\left\{\left(\frac{x_t}{p_t}\right) - \left(\frac{x_{t-1}}{p_{t-1}}\right)\right\}$ pour la PTCL et la PTCS, respectivement.

4.4 Exemple

Afin d'explorer les propriétés des méthodes PTCL et PTCS, nous reprenons l'exemple de la sous-section 3.3. La figure 4.1 donne la comparaison des méthodes symétriques f_S^{PTC}, f_L^{PTC} et $f^{PDP}(\mathbf{x})$.

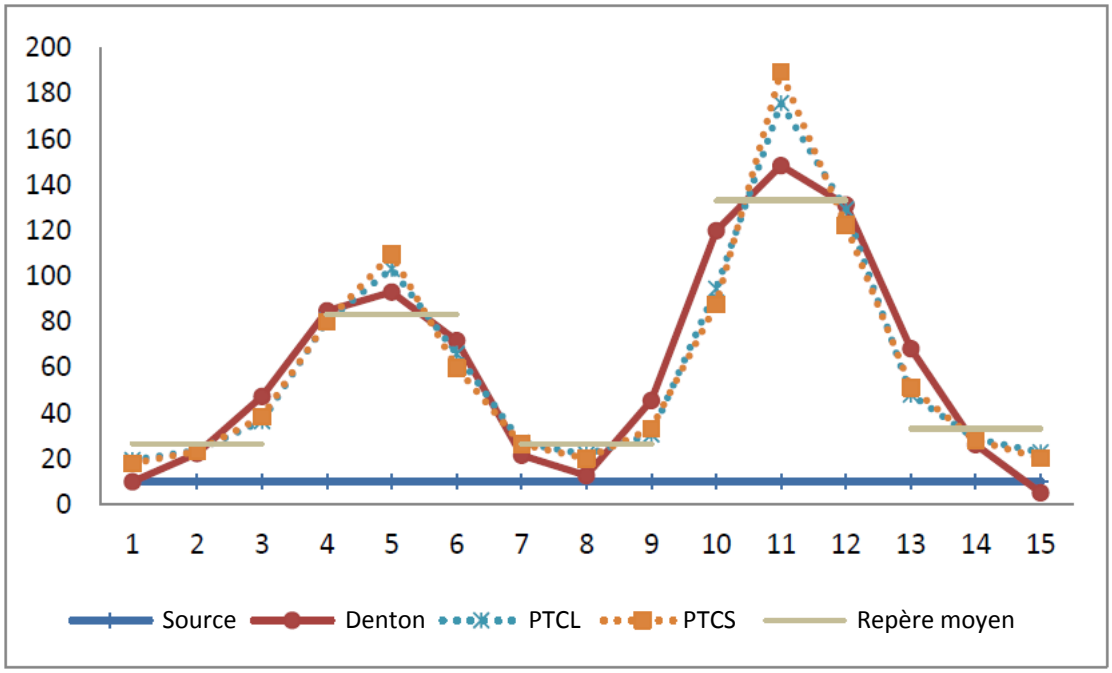

Figure 4.1 Exemple : Résultats de trois méthodes d'étalonnage symétriques. « Repère moyen » désigne le niveau moyen des valeurs mensuelles qui concorde avec les repères trimestriels et qui est calculé comme étant le tiers de son équivalent trimestriel.

Premièrement, on peut constater que les pics et les creux se produisent aux mêmes périodes pour toutes les méthodes temporellement symétriques.

Deuxièmement, certains inconvénients associés à la singularité de la fonction-objectif se produisent encore. Quand on les compare à la méthode PDP de Denton, les méthodes PTC ont tendance à éviter les valeurs proches de zéro, à s'écarter relativement lentement des faibles valeurs (dans les deux directions) et à aboutir à des grands pics irréguliers.

Troisièmement, comme il est mentionné à la sous-section 3.3, la méthode PTCL ressemble davantage à la méthode PDP de Denton que la méthode PTCS, ce qui découle des pics légèrement plus bas de la méthode PTCL.

5 Test empirique

À la présente section, nous effectuons un exercice illustratif sur des données réelles, afin de découvrir si les problèmes mentionnés à la section 3 se produisent dans le cas d'une application pratique réaliste.

5.1 Ensembles de données

L'ensemble de données utilisé pour l'illustration provient des données trimestrielles et annuelles sur le commerce publiées sur le site Web de l'Organisation des Nations Unies (ONU).

La *Commodity Trade Statistics Database* des Nations Unies (UN Comtrade) contient des données transmises par les autorités statistiques des pays déclarants, ou reçues par l'entremise d'organisations partenaires comme l'Organisation de coopération et de développement économiques (OCDE). Les données sur le commerce total des Nations Unies (UN Tottrade) proviennent principalement de la base de données *International Financial Statistics* (IFS), publiée mensuellement par le Fonds monétaire international (FMI). Les différences entre les deux sources sont dues à des différences entre les méthodes de collecte des données et les fins auxquelles les données sont destinées (Nations Unies, 2017). Toutes les données sont mises à la disposition du public à l'adresse http://comtrade.un.org/.

Nous utilisons UN Tottrade comme source des données trimestrielles, et UN Tottrade ainsi que UN Comtrade comme sources des données annuelles. Les deux sources de données comprennent les importations et les exportations pour environ 200 États membres de l'ONU.

Pour notre application, nous avons sélectionné toutes les séries qui contiennent trois totaux annuels et 12 valeurs trimestrielles pour la période allant de 2002 à 2004. Les variables d'intérêt sont les importations et les exportations totales. Les séries dont les valeurs trimestrielles ou annuelles étaient inférieures à 0,1 million de dollars ont été supprimées, car les méthodes d'étalonnage multiplicatives ne peuvent pas être considérées comme appropriées pour les valeurs nulles ou quasi nulles (voir la sous-section 3.2). Puisque les séries sont exprimées en millions de dollars, la valeur seuil exclut uniquement les cas « extrêmes » et retient certains cas réels de problème de singularité.

Nous nous retrouvons ainsi avec 238 séries temporelles pour Comtrade et 253 séries pour Tottrade. L'écart moyen entre les taux de croissance d'une année à l'autre des séries trimestrielles annualisées et de leurs séries annuelles repères est de 5,9 points de pourcentage et 2,7 points de pourcentage pour les repères Comtrade et Tottrade, respectivement. Pour la majorité des séries, l'écart peut être considéré comme étant faible. Le pourcentage de séries pour lesquelles l'écart maximal est inférieur à 5 points de pourcentage est de 79 % et 87 %, respectivement.

5.2 Résultats

Notre premier objectif est d'évaluer la performance globale. Nous comparerons le degré de préservation des valeurs provisoires et leur taux de croissance pour les diverses méthodes discutées dans le présent article.

Le tableau 5.1 montre, pour les cinq méthodes, les valeurs médianes sur toutes les séries, pour les fonctions $f_{\mathrm{P}}^{\mathrm{PTC}}$, $f_{\mathrm{R}}^{\mathrm{PTC}}$, $f_{\mathrm{S}}^{\mathrm{PTC}}$ pour la préservation du mouvement prospective, rétrospective et simultanée et f^{Niveau} pour la préservation de la valeur provisoire. La dernière fonction mesure le total des carrés des ajustements relatifs, défini par

$$f^{\mathrm{Niveau}}(\mathbf{x}) := \sum_{t=1}^{n}\left(\frac{x_t}{p_t} - 1\right)^2. \tag{5.1}$$

Tableau 5.1
Valeurs médianes des critères dans (2.1), (3.1), (4.1) et (5.1)

	Ensemble de données COM				Ensemble de données TOT			
	f_P^{PTC}	f_R^{PTC}	f_S^{PTC}	f^{Niveau}	f_P^{PTC}	f_R^{PTC}	f_S^{PTC}	f^{Niveau}
PDP de Denton	0,87	0,88	0,88	26,42	0,33	0,41	0,37	2,07
PTCP	0,84	0,98	0,93	26,43	0,27	0,48	0,45	2,06
PTCR	1,00	0,82	0,91	26,47	0,48	0,28	0,45	2,07
PTCS	0,87	0,89	0,88	26,41	0,34	0,38	0,36	2,07
PTCL	0,87	0,88	0,88	26,42	0,33	0,41	0,37	2,07

Les valeurs pour les ensembles de données COM et TOT sont $*10^{-2}$ et $*10^{-5}$, respectivement.

Le tableau 5.1 montre que la méthode PTCP, qui est conçue pour préserver les taux de croissance prospectifs, aboutit à une préservation relativement médiocre du mouvement rétrospectif. L'opposé est également vrai, à savoir que la PTCR ne préserve pas très bien les mouvements prospectifs. Au vu de ces résultats, nous pouvons conclure que la réversibilité du temps importe. Le tableau 5.1 montre aussi que les méthodes temporellement symétriques, c'est-à-dire les méthodes PDP de Denton, PTCS et PTCL, donnent de bons résultats sur toutes les mesures et que les différences entre ces méthodes ne sont que légères.

Pour évaluer la préservation prospective, rétrospective et simultanée des taux de croissance, nous utilisons un critère relatif qui consiste à comparer les valeurs des fonctions-objectif $f_P^{PTC}(\mathbf{x})$, $f_R^{PTC}(\mathbf{x})$ et $f_S^{PTC}(\mathbf{x})$ avec leurs valeurs optimales, qui sont obtenues par les méthodes PTCP, PTCR et PTCS, respectivement. De façon analogue aux normes de référence dans Di Fonzo et Marini (2012), la préservation du mouvement est jugée acceptable si la fonction se situe dans un intervalle de plus ou moins 10 % par rapport à la valeur optimale. Autrement dit, si $f^{méthode}(\mathbf{x}) / f^{optimum}(\mathbf{x}) \leq 1{,}1$, où f est l'une des fonctions-objectif mentionnées plus haut.

Pour les cinq méthodes examinées, le tableau 5.2 donne le pourcentage de séries temporelles pour lesquelles la préservation du mouvement prospective, rétrospective ou simultanée est acceptable.

Tableau 5.2
Pourcentage de séries temporelles pour lesquelles la préservation du mouvement est acceptable

	Ensemble de données COM			Ensemble de données TOT		
	Prospective	Rétrospective	Simultanée	Prospective	Rétrospective	Simultanée
PDP de Denton	79,4	78,6	95,8	79,4	79,4	96,0
PTCP	100,0	48,7	81,5	100,0	47,8	82,6
PTCR	47,1	100,0	76,9	44,3	100,0	75,1
PTCS	82,4	77,3	100,0	80,6	79,4	100,0
PTCL	79,8	79,0	96,6	79,4	79,4	96,0

Pour la méthode PDP de Denton, un degré acceptable de préservation simultanée du mouvement est observé dans plus de 95 % des cas. Donc, nous concluons que cette méthode peut être considérée comme

une très bonne approximation pour la méthode PTCS optimale; l'approximation est encore meilleure que dans le cas des méthodes PTCP et PTCR, pour lesquelles des résultats acceptables sont observés dans environ 80 % des cas.

Jusqu'à présent, nous nous sommes concentrés sur la performance pour la série temporelle complète. Nous allons maintenant examiner l'application de grandes ou extrêmement grandes corrections de réconciliation à des valeurs ou taux de croissance individuels.

Pour mesurer les corrections apportées aux taux de croissance, nous utilisons la différence absolue $|g_{it}(\mathbf{x}) - g_{it}(\mathbf{p})| * 100$ %, où g_{it} est un taux de croissance pour la série i et la période t. Les tableaux 5.3 et 5.4 comparent l'existence de grandes ou extrêmement grandes corrections apportées aux taux de croissance prospectifs, rétrospectifs et simultanés.

Tableau 5.3
Pourcentage de grandes corrections des taux de croissance (différence > 10 points de pourcentage)

	Ensemble de données COM			Ensemble de données TOT		
	Prospective	**Rétrospective**	**Simultanée**	**Prospective**	**Rétrospective**	**Simultanée**
PDP de Denton	2,0	2,1	1,9	0,8	0,6	0,6
PTCP	1,9	2,4	2,3	0,6	0,9	0,7
PTCR	2,3	1,5	2,0	1,1	0,3	0,8
PTCS	1,9	1,9	1,8	0,8	0,6	0,6
PTCL	2,0	1,9	1,9	0,8	0,6	0,5

Tableau 5.4
Pourcentage de corrections extrêmes des taux de croissance (différence > 50 points de pourcentage)

	Ensemble de données COM			Ensemble de données TOT		
	Prospective	**Rétrospective**	**Simultanée**	**Prospective**	**Rétrospective**	**Simultanée**
PDP de Denton	0,3	0,2	0,4	0,1	0,1	0,1
PTCP	0,2	0,2	0,2	0,0	0,1	0,1
PTCR	0,3	0,0	0,2	0,2	0,0	0,1
PTCS	0,2	0,1	0,2	0,1	0,0	0,1
PTCL	0,2	0,1	0,2	0,1	0,0	0,1

Ces tableaux montrent des différences mineures entre les méthodes.

De petites différences entre les méthodes sont également observées au tableau 5.5, qui présente les grandes ou extrêmement grandes corrections apportées aux valeurs provisoires, telles que mesurées par le critère relatif $(x_{it} / p_{it}) * 100$ %.

Donc, on peut conclure que les problèmes causés par la singularité ne se traduisent pas par une fréquence plus élevée des grandes corrections.

Tableau 5.5
Pourcentage de grandes corrections apportées aux valeurs provisoires

	Ensemble de données COM			Ensemble de données TOT		
	Grandes (> 10 %)	Extrêmes (> 100 %)	Négatives (< 0 %)	Grandes (> 10 %)	Extrêmes (> 100 %)	Négatives (< 0 %)
PDP de Denton	13,2	1,0	0,0	5,8	0,4	0,0
PTCP	13,0	1,0	0,0	5,8	0,3	0,1
PTCR	13,1	0,9	0,0	5,6	0,3	0,0
PTCS	13,1	0,9	0,0	5,8	0,4	0,0
PTCL	13,0	0,9	0,0	5,8	0,4	0,0

L'élément le plus remarquable dans le tableau 5.5 est l'existence de valeurs étalonnées négatives pour la méthode PTCP dans les données TOT. Un exemple de cette situation est donné à la figure 5.3.

Malgré la similarité des résultats des cinq méthodes d'étalonnage dans les tableaux 5.3 à 5.5, il existe des différences nettes à savoir à quel point les corrections de réconciliation sont lisses ou pas. Pour le démontrer, nous utiliserons l'indicateur de lissage (Temurshoev, 2012).

$$\text{Lissage} \quad = \quad \sum_{t=2}^{n-2} \left[\mathrm{BI}_t - \overline{\mathrm{BI}_t} \right]^2, \tag{5.2}$$

où BI_t est ce qu'il est convenu d'appeler le ratio série étalonnée/indicateur (BI pour *Benchmarked to Indicator*), c'est-à-dire x_t / p_t et $\overline{\mathrm{BI}_t}$ est la moyenne mobile de 5 termes $\frac{1}{5} \sum_{k=t-2}^{k=t+2} \mathrm{BI}_k$.

Selon cet indicateur, nous constatons au tableau 5.6 que les résultats les plus lisses sont donnés par les méthodes PDP de Denton et PTCL. Inversement, les méthodes asymétriques PTCP et PTCR produisent les corrections les plus irrégulières. Il s'ensuit que les méthodes temporellement symétriques PTCS, mais encore davantage PTCL, souffrent moins de la singularité que les méthodes asymétriques PTCP et PTCR. Ces résultats illustrent clairement les problèmes associés à la singularité de la fonction-objectif de la méthode PTC qui ont été décrits à la sous-section 3.2.

Tableau 5.6
Valeurs de l'indicateur de lissage (5.2), sommées sur toutes les séries

	Ensemble de données COM	Ensemble de données TOT
PDP de Denton	3,4	0,3
PTCP	9,8	39,0
PTCR	8,2	2,9
PTCS	4,3	1,1
PTCL	3,3	0,5

5.3 Exemples

Nous donnons ci-après deux exemples pour montrer que les problèmes décrits à la section 3 se manifestent dans une application réelle.

Le premier exemple, dans les figures 5.1 et 5.2, illustre que les méthodes PTC non symétriques peuvent modifier le moment où se produisent les événements économiques les plus importants. Si nous considérons les neuf premiers trimestres, les deux valeurs les plus élevées se produisent à des périodes différentes. Les périodes de pic pour la PTCP sont les 6e et 7e trimestres et celles pour la PTCR sont les 5e et 6e trimestres.

Étroitement associé à cela est le fait que la PTCP s'écarte relativement lentement des valeurs proches de zéro du 1er au 4e trimestre.

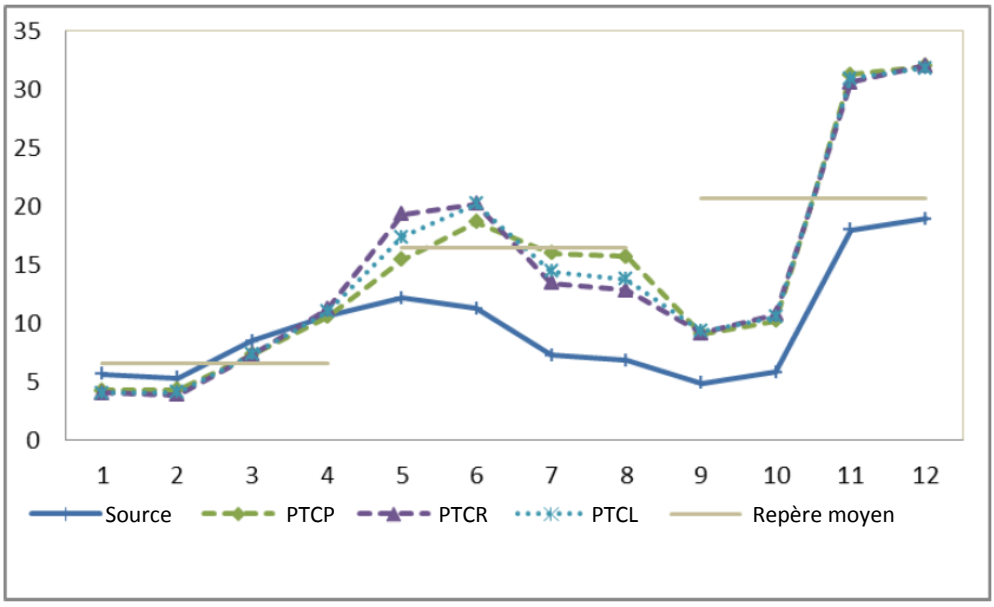

Figure 5.1 Exportations du Burundi, Comdata, 2002 à 2004, millions de dollars US. « Repère moyen » désigne le niveau moyen des données trimestrielles qui concorde avec les repères annuels et qui est calculé comme étant le quart de son équivalent annuel.

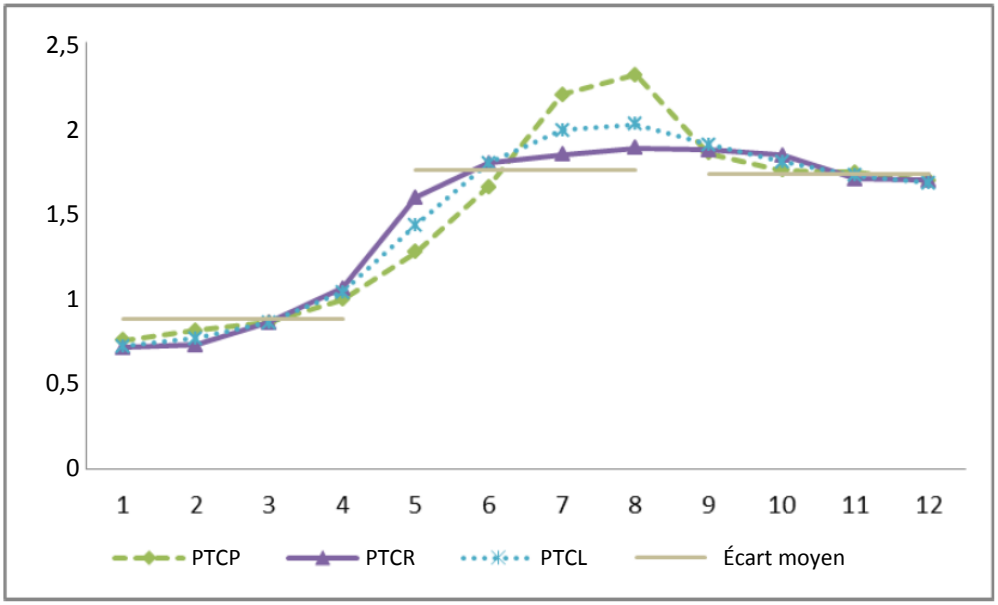

Figure 5.2 Ratios série étalonnée/indicateur, exportations du Burundi, 2002 à 2004. « Écart moyen » désigne le ratio BI annuel, c'est-à-dire le ratio d'un repère annuel et de la somme des indicateurs trimestriels sous-jacents.

Le deuxième exemple illustre les complications dues à une fonction-objectif singulière. Comme le montre la figure 5.4, la méthode PTCP préserve étroitement les taux de croissance du 6e au 10e trimestre. Cela a cependant pour prix un pic irrégulier au 5e trimestre et des valeurs étalonnées négatives aux 11e et 12e trimestres.

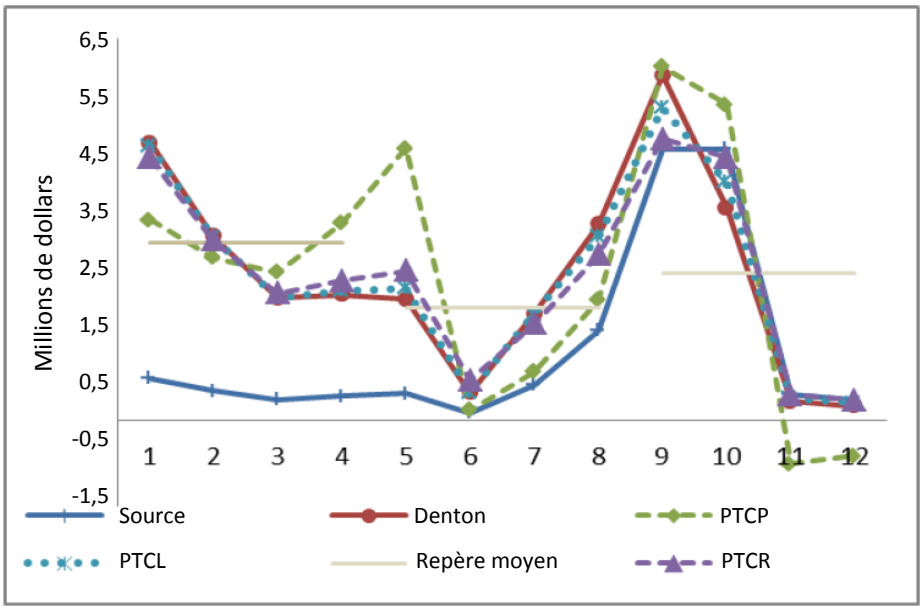

Figure 5.3 Exportations de la Gambie, Totdata, 2002 à 2004, millions de dollars US. « Repère moyen » désigne le niveau moyen des données trimestrielles qui concorde avec les repères annuels et qui est calculé comme étant le quart de son équivalent annuel.

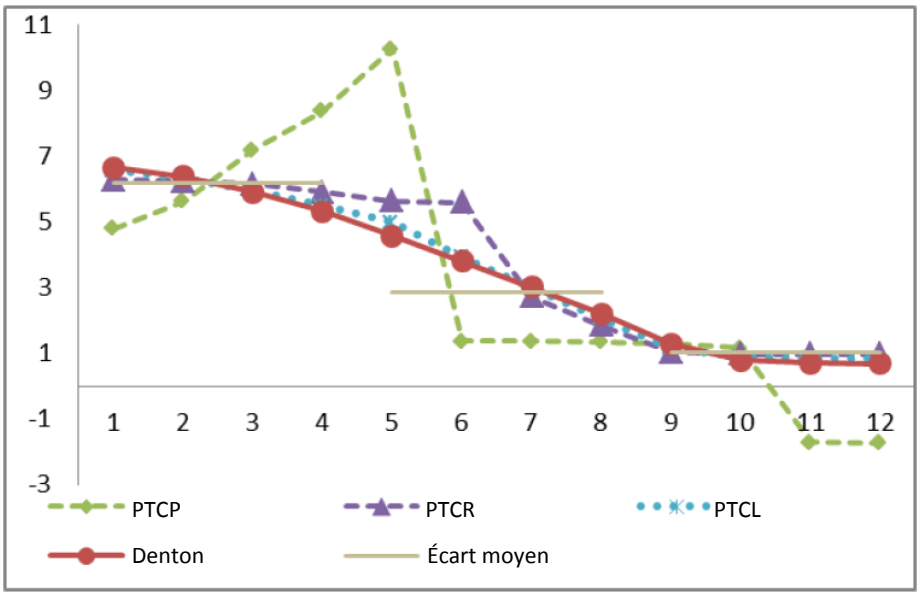

Figure 5.4 Exportations de la Gambie, Totdata, 2002 à 2004, ratio série étalonnée/indicateur. « Écart moyen » désigne le ratio BI annuel, c'est-à-dire le ratio d'un repère annuel et de la somme des indicateurs trimestriels sous-jacents.

6 Conclusion

La méthode des premières différences proportionnelles (PDP) de Denton et la méthode de préservation des taux de croissance (PTC) sont deux méthodes d'étalonnage multiplicatives bien connues. Il est généralement reconnu que la méthode PTC possède le fondement théorique le plus solide. Cette méthode préserve mieux les taux de croissance initiaux que la méthode PDP de Denton. Cependant, d'un point de vue technique, la méthode la plus facile à appliquer est celle de Denton. Pour cette raison, et parce qu'elle est souvent une bonne approximation de la méthode PTC, la méthode PDP de Denton est d'usage plus répandu.

Le présent article montre que la méthode PTC présente deux inconvénients qui, autant que le sachent les auteurs, n'ont pas été mentionnés ailleurs.

Le premier inconvénient est que la méthode PTC ne satisfait pas le principe de réversibilité du temps. Selon ce principe, le fait que l'on procède à la préservation des taux de croissance prospectifs ou rétrospectifs ne devrait pas avoir d'incidence sur les résultats. Autrement dit, l'étalonnage d'une série temporelle originale, $t = 1, \ldots, n,$ ou d'une série temporelle « inversée », $t = n, \ldots, 1$ devrait aboutir à la même série étalonnée. Puisque, dans toute application d'étalonnage, la direction du temps est sans pertinence, toute méthode d'étalonnage devrait de préférence satisfaire au principe de réversibilité du temps. De surcroît, une méthode d'étalonnage qui ne satisfait pas à ce principe peut, selon la direction temporelle choisie, donner des conclusions entièrement différentes quant aux moments des événements économiques. Pour ces raisons, il est préférable de décourager l'usage des méthodes PTC prospective et rétrospective.

Dans le présent article, deux méthodes PTC de rechange qui satisfont le principe de réversibilité du temps sont présentées. La première, qui est une nouvelle méthode appelée PTC simultanée ou PTCS, préserve les taux de croissance prospectifs ainsi que rétrospectifs. L'autre, qui est une méthode existante appelée PTC logarithmique ou PTCL, préserve les logarithmes des taux de croissance prospectifs.

Un deuxième inconvénient de toutes les méthodes PTC examinées dans le présent article est la singularité de leurs fonctions-objectif. Les complications qui en découlent sont l'évitement des résultats proches de zéro, ainsi que la présence de pics irréguliers dans les résultats et de valeurs négatives non nécessaires dans les résultats étalonnés.

Ces problèmes se manifestent effectivement dans une application illustrative à des données réelles. Même si des valeurs négatives non nécessaires ne se produisent qu'occasionnellement, les corrections de réconciliation sont beaucoup plus irrégulières que dans le cas de la méthode PDP de Denton. Puisque le caractère lisse des corrections de réconciliation (ratios BI) est souvent le principal objet de l'étalonnage, l'usage des méthodes PTC asymétriques peut être déconseillé pour de nombreuses applications.

Tandis que la méthode PDP de Denton est considérée dans la littérature comme une « bonne approximation » de la méthode PTC idéale, notre principale conclusion est que la méthode PDP de Denton est même plus appropriée que la méthode PTC classique pour de nombreuses applications. La méthode de Denton comprend des calculs plus faciles, et ne souffre pas de problèmes liés à l'irréversibilité du temps et à une fonction-objectif singulière. En outre, les résultats de la méthode PDP de Denton sont une approximation encore meilleure pour les variantes temporellement symétriques de la PTC que pour la PTC classique.

Cependant, si la préservation des taux de croissance est le principal point d'intérêt, une version temporellement symétrique de la méthode PTC peut également être un bon choix, en particulier la PTCL. Les méthodes temporellement symétriques préservent un peu mieux les taux de croissance que la méthode PDP de Denton, satisfont le principe de réversibilité du temps et souffrent moins gravement des inconvénients d'une fonction-objectif singulière que la PTC classique.

Remerciements

Les auteurs remercient trois examinateurs et le rédacteur associé de leurs commentaires utiles qui leur ont permis d'améliorer considérablement l'article.

Bibliographie

Bikker, R.P., Daalmans, J.A. et Mushkudiani, N. (2013). Benchmarking large accounting frameworks: A generalized multivariate model. *Economic Systems Research*, 25, 390-408.

Bloem, A., Dippelsman, R. et Maehle, N. (2001). Quarterly National Accounts Manual. Concepts, Data Sources, and Compilation, International Monetary Fund, Washington DC.

Boot, J.C.G., Feibes, W. et Lisman, J.H.C. (1967). Further methods of derivation of quarterly figures from annual data. *Applied Statistics*, 16, 65-75.

Bozik, J.E., et Otto, M.C. (1988). Benchmarking: Evaluating methods that preserve month-to-month changes. Bureau of the Census - Statistical Research Division, RR-88/07, accessible au http://www.census.gov/srd/papers/pdf/rr88-07.pdf.

Causey, B., et Trager, M.L. (1981). Derivation of Solution to the Benchmarking Problem: Trend Revision. Notes de recherche non publiées, U.S. Census Bureau, Washington D.C. Disponible en annexe dans Bozik et Otto (1988).

Cholette, P.A. (1984). L'ajustement des séries infra-annuelles aux répères annuels. *Techniques d'enquête*, 10, 1, 39-53. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1984001/article/14348-fra.pdf.

Daalmans, J.A., et Di Fonzo, T. (2014). Denton PFD and GRP benchmarking are friends. An empirical evaluation on Dutch Supply and Use Tables. Document présenté à la conférence de l'IIOA, Lisboa 14 au 18 juillet.

Dagum, E.B., et Cholette, P. (2006). *Benchmarking, Temporal Distribution and Reconciliation Methods for Time Series Data*. New York: Springer-Verlag.

Denton, F. (1971). Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association*, 66, 99-102.

Di Fonzo, T., et Marini, M. (2011). Simultaneous and two-step reconciliation of systems of time series: Methodological and practical issues. *Journal of the Royal Statistical Society C*, 60, 143-164.

Di Fonzo, T., et Marini, M. (2012). Benchmarking time series according to a growth rates preservation principle. *Journal of Economic and Social Measurement*, 37, 225-252.

Di Fonzo, T., et Marini, M. (2015). Reconciliation of systems of time series according to a growth rates preservation principle. *Statistical Methods & Applications*, 24, 651-669.

Eurostat (2013). *Handbook on Quarterly National Accounts – 2013 Edition*. http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-13-004.

Fisher, I. (1922). *The Making of Index Numbers*, Houghton-Mifflin, Boston.

Harvill Hood, C.C. (2005). An empirical comparison of methods for benchmarking seasonally adjusted series to annual totals. Document présenté à l'Atelier the Frontiers in Benchmarking Techniques and their Applications to Official Statistics, Luxembourg, 7 au 8 avril 2005.

Helfand, S.D., Monsour, N.J. et Trager, M.L. (1977). Historical Revision of Current Business Survey Estimates. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 246-250.

Monsour, N.J., et Trager, M.L. (1979). Revision and benchmarking of business time series. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 333-337.

Nations Unies (2017). *2016 International Trade Statistics Yearbook. Volume I Trade by Countries*. New York: Nations Unies. https://comtrade.un.org/pb/downloads/2016/ITSY2016VolI.pdf.

Nocedal, J., et Wright, S. (2006). *Numerical Optimization, 2nd Edition*. New York: Springer.

Temurshoev, U. (2012). *Entropy-Based Benchmarking Methods*. University of Groningen. http://www.ggdc.net/publications/memorandum/gd122.pdf.

Titova, N., Findley, D. et Monsell, B.C. (2010). Comparing the Causey-Trager method to the multiplicative Cholette-Dagum regression-based method of benchmarking sub-annual data to annual benchmarks. *Joint Statistical Meetings Proceedings, Business and Economic Statistics Section*, 3007-3021.

Trager, M.L. (1982). Derivation of Solution to the Benchmarking Problem: Relative Revision. Notes de recherche non-publiées, U.S. Census Bureau, Washington D.C. Disponible en annexe dans Bozik et Otto (1988).

Techniques d'enquête, juin 2018
Vol. 44, nº 1, pp. 65-79
Statistique Canada, nº 12-001-X au catalogue

Étude de divers estimateurs de la prévalence de la maladie mentale grave fondés sur un échantillon à deux phases

Phillip S. Kott, Dan Liao, Jeremy Aldworth, Sarra L. Hedden, Joseph C. Gfroerer, Jonaki Bose et Lisa Colpe[1]

Résumé

Un processus à deux phases a été utilisé par la *Substance Abuse and Mental Health Services Administration* pour estimer la proportion d'Américains adultes atteints d'une maladie mentale grave (MMG). La première phase correspondait à la *National Survey on Drug Use and Health* (NSDUH) réalisée annuellement, tandis que la seconde phase consistait en un sous-échantillon aléatoire d'adultes ayant répondu à la NSDUH. Les personnes qui ont répondu à la deuxième phase d'échantillonnage ont été soumises à une évaluation clinique visant à déceler les maladies mentales graves. Un modèle de prédiction logistique a été ajusté à ce sous-échantillon en prenant la situation de MMG (oui ou non) déterminée au moyen de l'instrument de deuxième phase comme variable dépendante, et les variables connexes recueillies dans la NSDUH auprès de tous les adultes comme variables explicatives du modèle. Des estimations de la prévalence de la MMG chez l'ensemble des adultes et au sein de sous-populations d'adultes ont ensuite été calculées en attribuant à chaque participant à la NSDUH une situation de MMG établie en comparant sa probabilité estimée d'avoir une MMG avec un seuil diagnostique choisi sur la distribution des probabilités prédites. Nous étudions d'autres options que cet estimateur par seuil diagnostique classique, dont l'estimateur par probabilité. Ce dernier attribue une probabilité estimée d'avoir une MMG à chaque participant à la NSDUH. La prévalence estimée de la MMG est la moyenne pondérée de ces probabilités estimées. Au moyen des données de la NSDUH et de son sous-échantillon, nous montrons que, même si l'estimateur par probabilité donne une plus petite erreur quadratique moyenne quand on estime la prévalence de la MMG parmi l'ensemble des adultes, il a une plus grande tendance que l'estimateur par seuil diagnostique classique à présenter un biais au niveau de la sous-population.

Mots-clés : Biais; estimateur corrigé du biais; domaine; théorie des sondages; asymptotique.

1 Introduction

La maladie mentale grave est définie comme le fait de présenter à l'heure actuelle, ou d'avoir présenté au cours de la dernière année, un trouble mental, comportemental ou émotionnel diagnostiquable (excepté les troubles du développement et les toxicomanies) d'une durée suffisante pour satisfaire les critères diagnostiques spécifiés dans le Manuel diagnostique et statistique des troubles mentaux, 4e édition (American Psychiatric Association, 1994). La *National Survey on Drug Use and Health* (NSDUH), parrainée par la *Substance Abuse and Mental Health Services Administration* (SAMHSA), fournit des estimations nationales et au niveau de l'État de la consommation de produits du tabac, d'alcool et de drogues illicites parmi la population civile des États-Unis âgée de 12 ans et plus non placée en établissement. La *Mental Health Surveillance Study* (MHSS) est une étude de suivi qui était réalisée après l'interview principale de la NSDUH en vue d'estimer la prévalence de la maladie mentale grave (MMG) chez les adultes de 18 ans et plus au niveau national et pour certaines sous-populations d'intérêt. En raison de contraintes de budget et de temps, il n'était pas faisable en pratique d'administrer les interviews cliniques à l'échantillon complet de la NSDUH, qui compte environ 46 000 adultes par année, en vue d'obtenir les estimations de la

1. Phillip S. Kott, Dan Liao et Jeremy Aldworth, RTI International. Courriel : pkott@rti.org; Sarra L. Hedden, Joseph C. Gfroerer et Jonaki Bose, Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality (CBHSQ), 5600 Fishers Lane, Rockville, MD 20857, État-Unis; Lisa Colpe, National Institute of Mental Health, 6001 Executive Blvd., Bethesda, MD 20892.

prévalence de la MMG. Donc, on a procédé à une deuxième phase d'échantillonnage, comme dans le cas de la *National Comorbidity Survey Replication* (voir Kessler, Chiu, Demler et Walters, 2005). De 2008 jusqu'à 2012, un sous-échantillon aléatoire (avant la non-réponse) d'adultes ayant répondu à la NSDUH a été soumis à une interview clinique dans les quatre semaines après l'interview principale de la NSDUH. Pour plus de renseignements au sujet de la NSDUH et du sous-échantillon clinique, le lecteur est invité à consulter le rapport du *Center for Behavioral Health Statistics and Quality* (CBHSQ), 2014.

Les évaluations cliniques ont été traitées comme la « norme de référence », malgré la possibilité d'erreurs humaines (ce sujet est abordé plus en détail à la section 4). Un modèle de prédiction logistique a été ajusté au sous-échantillon de répondants en traitant les évaluations cliniques de MMG (oui ou non) comme la variable dépendante et les variables contenues dans la NSDUH, comme les variables explicatives du modèle. Les variables de la NSDUH incluses dans le modèle de prédiction, obtenues auprès de 46 000 adultes participant à l'enquête principale, étaient des mesures de détresse psychologique et de déficience fonctionnelle dérivées des réponses aux questions de la NSDUH, une variable d'âge, l'existence autodéclarée d'un épisode dépressif majeur l'année précédente, et l'existence de pensées suicidaires graves l'année précédente. Le modèle de prédiction ajusté a ensuite été appliqué à l'échantillon complet de la NSDUH pour prédire, pour chaque adulte participant, la probabilité d'avoir une MMG en fonction de ses réponses à l'interview principale de la NSDUH.

Un participant adulte à la NSDUH dont la probabilité estimée d'avoir une MMG était égale ou supérieure à un certain seuil diagnostique s'est vu attribuer une valeur prédite de MMG de 1 (c'est-à-dire présente une MMG); sinon, on lui a attribué une valeur prédite de MMG de 0 (c'est-à-dire ne présente pas de MMG). Les taux de prévalence de la MMG pour l'ensemble des adultes et pour diverses sous-populations ont été estimés en se servant de ces valeurs prédites. Le seuil diagnostique a été établi de façon que, dans le sous-échantillon de la MHSS, les fractions pondérées de résultats faussement positifs (répondants du sous-échantillon dont le modèle prédisait qu'ils avaient une MMG, mais dont le diagnostic clinique indiquait qu'ils n'en avaient *pas*) et les résultats faussement négatifs (répondants du sous-échantillon dont le modèle prédisait qu'ils n'avaient pas de MMG, mais dont le diagnostic clinique indiquait qu'ils en avaient une) soient aussi proches de l'égalité que possible. L'égalité stricte était habituellement impossible parce que les probabilités prédites d'avoir une MMG ne prenaient qu'un nombre limité de valeurs selon les covariables incluses dans le modèle.

L'estimateur par seuil diagnostique classique est dérivé de la théorie de la fonction d'efficacité du récepteur ou courbe ROC (de l'anglais *Receiver Operating Characteristics*). Voir, par exemple, Fawcett (2006). À la section 2, nous faisons appel à la théorie des sondages fondée sur l'échantillonnage probabiliste (également dite « fondée sur le plan de sondage ») pour proposer d'autres options pour remplacer cet estimateur. L'une de celles-ci est l'estimateur par probabilité, qui attribue simplement à chaque participant à la NSDUH une probabilité estimée d'avoir une MMG, sans aucune désignation dichotomique. Nous présentons aussi les versions corrigées du biais de l'estimateur par seuil diagnostique classique et de

l'estimateur par probabilité. Ces versions sont similaires aux estimateurs par la régression généralisée (GREG) classiques (voir, par exemple, Särndal, Swensson et Wretman, 1989) et sont presque sans biais sous la théorie des sondages, que le modèle logistique tienne ou non. Dans chaque cas, le terme de correction du biais sert de fondement à une statistique de test pour déterminer si l'estimateur basé sur le modèle associé – estimateur par seuil diagnostique ou par probabilité – présente un biais significatif.

À la section 3, nous nous servons des données du sous-échantillon NSDUH/MHSS recueillies entre 2008 et 2012 pour évaluer les estimateurs proposés basés sur le modèle logistique. Nous montrons que l'estimateur par probabilité donne une plus petite erreur quadratique moyenne quand on estime la prévalence de la MMG chez l'ensemble des adultes, mais qu'il a une plus grande tendance que l'estimateur par seuil diagnostique classique à présenter un biais au niveau de la sous-population. Cela nous mène à proposer un estimateur par seuil diagnostique hybride qui est (au moins) plus efficace que l'estimateur par seuil diagnostique classique pour l'ensemble des adultes, tout en n'ayant pas tendance à être biaisé au niveau de la sous-population.

À la section 4, nous formulons quelques observations finales. Il importe de rappeler que la SAMHSA prévoyait se servir du modèle logistique ajusté sur les sous-échantillons de 2008 à 2012 pour calculer les estimations annuelles de la prévalence de la MMG chez les adultes d'après les réponses à la NSDUH au-delà de 2012, sans tirer de nouveaux sous-échantillons cliniques.

2 Quelques estimateurs

2.1 Sur l'ensemble des adultes

Soit S l'échantillon pertinent de répondants à la NSDUH (adultes de 18 ans et plus) de 2008 jusqu'à 2012, et w_k le poids d'analyse de la NSDUH (première phase) pour un individu $k \in S$. Soit S' le sous-échantillon de S dont les membres répondent à une évaluation clinique de leur situation de MMG. Soit $y_k = 1$ quand k obtient un diagnostic de maladie mentale grave, et $y_k = 0$ quand k obtient un diagnostic d'absence de maladie mentale grave. Soit ω_k le poids de deuxième phase d'échantillonnage d'un individu k dans S'. Par souci de commodité, nous fixons ω_k à 0 pour les individus présents dans S mais non dans S'.

Dans la pratique, les deux ensembles de poids ont été ajustés pour tenir compte de la non-réponse et de la sous-couverture, et pour augmenter leur efficacité, mais nous ignorerons ce fait ici pour simplifier. Nous supposerons plutôt que $1/w_k$ est la probabilité de sélection pour un répondant à la NSDUH, $1/\omega_k$, la probabilité de sélection pour un répondant du sous-échantillon de la MHSS, et donc w_k/ω_k, la probabilité de sélection conditionnelle d'un répondant du sous-échantillon sachant qu'il était un répondant à la NSDUH. Un estimateur presque sans biais de la prévalence de la MMG chez les adultes entre 2008 et 2012 basé sur l'échantillon à deux phases est $\bar{y}_U = \sum_{S'} \omega_k y_k \Big/ \sum_{S'} \omega_k$, « presque » sans biais parce que le dénominateur peut contenir une erreur d'échantillonnage.

Supposons que l'on exécute une régression logistique pondérée par ω_k sur les répondants du sous-échantillon d'adultes de la MHSS dans S', avec y_k comme variable dépendante et en utilisant un vecteur raisonnable de covariables explicatives, $\mathbf{x}_k$, disponibles pour chaque répondant dans l'échantillon d'adultes de la NSDUH. La façon exacte dont les covariables sont choisies dépasse le cadre de la présente étude (pour cela, le lecteur est invité à consulter le rapport du Center for Behavioral Health Statistics and Quality, 2015; chapitre 4). Posons que le prédicteur pour y_k issu de cette régression logistique pondérée est $p_k = p(\mathbf{x}_k'\mathbf{b}) = [1 + \exp(-\mathbf{x}_k'\mathbf{b})]^{-1}$.

L'utilisation de poids pour ajuster le modèle de régression logistique protège contre la possibilité d'une corrélation entre les résidus du modèle et les probabilités de sélection. Cela est également en harmonie avec la façon dont la prévalence de la MMG a été estimée; cette estimation résultait de la régression pondérée de y_k sur la constante 1 sans aucune covariable.

En triant le sous-échantillon en fonction des valeurs de p_k, on peut trouver la valeur du seuil diagnostique p_C telle que l'égalité

$$\sum_{\substack{k \in S' \\ p_k \geq p_C}} \omega_k = \sum_{k \in S'} \omega_k y_k \tag{2.1}$$

soit vérifiée exactement ou aussi exactement que possible. Autrement dit, le nombre estimé d'adultes dans la population ayant une valeur de p_k égale ou supérieure au seuil diagnostique est approximativement égal au nombre estimé d'adultes ayant une MMG. Soit une variable indicatrice aléatoire c_k qui vaut 1 quand $p_k \geq p_C$ et 0 autrement. Un seuil diagnostique déterminé en utilisant l'équation (2.1) égalise aussi autant que possible les nombres pondérés de faux positifs $\left(\sum_{S':c_k=1} \omega_k (1 - y_k)\right)$ et de faux négatifs $\left(\sum_{S':c_k=0} \omega_k y_k\right)$ dans S'.

Deux estimateurs de la prévalence de la MMG chez les adultes sont l'estimateur *par seuil diagnostique* et l'estimateur *par probabilité* basés sur le modèle :

$$\overline{y}_C = \frac{\sum_S w_k c_k}{\sum_S w_k}, \tag{2.2}$$

et

$$\overline{y}_P = \frac{\sum_S w_k p_k}{\sum_S w_k}, \tag{2.3}$$

qui sont calculés en utilisant l'échantillon complet de la NSDUH plutôt que le sous-échantillon plus petit de la MHSS comme cela est le cas de $\overline{y}_U$.

Nous supposons maintenant que l'une des covariables dans le modèle logistique est 1 ou l'équivalent ($\mathbf{x}_k'\boldsymbol{\gamma} = 1$ pour une certaine valeur de $\boldsymbol{\gamma}$). Sous cette hypothèse, l'estimateur par probabilité de la prévalence de la MMG est exactement égal à un estimateur par probabilité *corrigé du biais* donné ci-après :

$$\begin{aligned} \overline{y}_{P-\mathrm{BC}} &= \frac{\sum_{S'} \omega_k y_k}{\sum_{S'} \omega_k} + \left(\frac{\sum_S w_k p_k}{\sum_S w_k} - \frac{\sum_{S'} \omega_k p_k}{\sum_{S'} \omega_k} \right) \\ &= \frac{\sum_S w_k p_k}{\sum_S w_k} + \frac{\sum_{S'} \omega_k (y_k - p_k)}{\sum_{S'} \omega_k}. \end{aligned} \tag{2.4}$$

L'égalité entre $\overline{y}_P$ et $\overline{y}_{P-\mathrm{BC}}$ résulte du fait que le numérateur du *terme de correction du biais* à la deuxième ligne de l'équation (2.4), $\sum_{S'} \omega_k (y_k - p_k) \big/ \sum_{S'} \omega_k$, égale zéro. L'ajustement d'une régression logistique force $\sum_{S'} \omega_k (y_k - p_k)\, \mathbf{x}_k = \mathbf{0}$, et nous avons supposé que $\mathbf{x}_k$ contient 1 ou l'équivalent.

Puisque l'espérance du terme entre parenthèses à la *première* ligne de l'équation (2.4) est presque nulle sous des conditions faibles, $\overline{y}_P = \overline{y}_{P-\mathrm{BC}}$, comme $\overline{y}_U$, est presque sans biais sous la théorie des sondages. *Cela est vrai, que le modèle utilisé pour déterminer les* p_k *soit correct ou non* à condition que $\mathbf{b}$ dans $p_k = p\left(\mathbf{x}_k'\mathbf{b}\right) = \left[1 + \exp\left(-\mathbf{x}_k'\mathbf{b}\right)\right]^{-1}$ converge vers *quelque chose* à mesure que les tailles du sous-échantillon de la MHSS et de l'échantillon de la NSDUH deviennent arbitrairement grandes.

L'estimateur $\overline{y}_{P-\mathrm{BC}}$ est analogue à l'estimateur GREG bien connu. Il suit Lehtonen et Veijanen (1998), et calcule les p_k au moyen d'un modèle logistique au lieu du modèle linéaire de l'estimateur GREG.

Un estimateur par seuil diagnostique corrigé du biais est donné par

$$\begin{aligned} \overline{y}_{C-\mathrm{BC}} &= \frac{\sum_{S'} \omega_k y_k}{\sum_{S'} \omega_k} + \left(\frac{\sum_{S} w_k c_k}{\sum_{S} w_k} - \frac{\sum_{S'} \omega_k c_k}{\sum_{S'} \omega_k} \right) \\ &= \frac{\sum_{S} w_k c_k}{\sum_{S} w_k} + \frac{\sum_{S'} \omega_k (y_k - c_k)}{\sum_{S'} \omega_k}. \end{aligned} \tag{2.5}$$

En suivant le même raisonnement que plus haut, cet estimateur est également presque sans biais sous des conditions faibles. Il est proche de l'estimateur par seuil diagnostique basé sur le modèle puisque le terme de correction du biais, $\sum_{S'} \omega_k (y_k - c_k) \big/ \sum_{S'} \omega_k$, est presque nul. Le terme de correction du biais serait exactement nul s'il existait un seuil diagnostique p_C qui satisfait exactement l'équation (2.1).

2.2 Estimation par domaine

Examinons maintenant une sous-population de l'ensemble d'adultes, par exemple les hommes ou tous les adultes qui ont été traités pour une maladie mentale (ou tous les adultes qui vivent dans un État particulier). Ce genre de population est appelé « domaine » d'intérêt. Pour estimer la prévalence de la MMG dans un domaine, nous pouvons simplement insérer un indicateur d'appartenance au domaine, d_k, qui vaut 1 quand k est dans le domaine, et 0 autrement, dans toutes nos estimations :

$$\overline{y}_{U(d)} = \frac{\sum_{S'} \omega_k y_k d_k}{\sum_{S'} \omega_k d_k} \tag{2.6}$$

$$\overline{y}_{P(d)} = \frac{\sum_{S} w_k p_k d_k}{\sum_{S} w_k d_k} \tag{2.7}$$

$$\overline{y}_{C(d)} = \frac{\sum_{S} w_k c_k d_k}{\sum_{S} w_k d_k} \tag{2.8}$$

$$\begin{aligned} \overline{y}_{P-\mathrm{BC}(d)} &= \overline{y}_{U(d)} + \left(\frac{\sum_{S} w_k p_k d_k}{\sum_{S} w_k d_k} - \frac{\sum_{S'} \omega_k p_k d_k}{\sum_{S'} \omega_k d_k} \right) \\ &= \frac{\sum_{S} w_k p_k d_k}{\sum_{S} w_k d_k} + \frac{\sum_{S'} \omega_k (y_k - p_k) d_k}{\sum_{S'} \omega_k d_k} \end{aligned} \tag{2.9}$$

$$\begin{aligned}\overline{y}_{C-\mathrm{BC}(d)} &= \overline{y}_{U(d)} + \left(\frac{\sum_S w_k c_k d_k}{\sum_S w_k d_k} - \frac{\sum_{S'} \omega_k c_k d_k}{\sum_{S'} \omega_k d_k} \right) \\ &= \frac{\sum_S w_k c_k d_k}{\sum_S w_k d_k} + \frac{\sum_{S'} \omega_k (y_k - c_k) d_k}{\sum_{S'} \omega_k d_k}. \end{aligned} \tag{2.10}$$

C'est ici que les termes de correction du biais jouent un rôle important. Si le modèle logistique, qui a été ajusté sur le sous-échantillon de *l'ensemble* des adultes, tient dans le domaine, alors $\sum_{S'} \omega_k d_k (y_k - p_k) \big/ \sum_{S'} \omega_k d_k$ sera une estimation de zéro, et l'estimateur par probabilité basé sur le modèle, $\overline{y}_{P(d)}$ dans l'équation (2.7), sera presque sans biais. Si le modèle ne tient pas dans le domaine (par exemple, si les hommes sont plus susceptibles d'avoir une MMG que ne le prédit le modèle), alors l'estimateur par probabilité basé sur le modèle peut être considérablement biaisé.

L'ajout de la correction du biais $\sum_{S'} \omega_k d_k (y_k - p_k) \big/ \sum_{S'} \omega_k d_k$ à $\overline{y}_{P(d)}$ produit un estimateur qui est presque sans biais sous la théorie des sondages. Toutefois, quand le modèle tient dans le domaine, l'application de la correction résultera presque certainement en une diminution de la précision. Un argument similaire peut être fait concernant le caractère approprié de l'ajout du terme $\sum_{S'} \omega_k d_k (y_k - c_k) \big/ \sum_{S'} \omega_k d_k$ figurant dans l'équation (2.10) à l'estimateur par seuil diagnostique $\overline{y}_{C(d)}$, donné par l'équation (2.8).

Les équations (2.4) et (2.5) peuvent être considérées comme des cas particuliers de (2.9) et (2.10), respectivement, avec $d_k \equiv 1$.

3 Le sous-échantillon de la MHSS

3.1 À propos du sous-échantillon de la MHSS

La NSDUH est une enquête réalisée selon un plan d'échantillonnage probabiliste à plusieurs degrés stratifié. De 2008 à 2012, le sous-échantillon de la MHSS a été tiré chaque année par échantillonnage de Poisson parmi les adultes ayant répondu à la NSDUH cette année-là. Les probabilités de sélection dans le sous-échantillon ont été déterminées chaque année en utilisant un algorithme qui avait tendance à suréchantillonner les adultes présentant un niveau élevé de détresse psychologique. L'algorithme variait d'une année à l'autre. Voir le rapport du Center for Behavioral Health Statistics and Quality (2014, chapitre 3) pour des renseignements plus détaillés.

La taille cible du sous-échantillon de répondants était d'environ 750 pour 2008, tandis qu'elle était de 500 pour 2009 et 2010, respectivement, et de 1 500 pour 2011 et 2012, respectivement. Un ensemble de données regroupant tous les répondants pour la période de 2008 à 2012 a été créé pour modéliser la MMG. Les poids pour la modélisation ont été établis en supposant que le même modèle était vérifié pour toutes les années. Par conséquent, un poids plus important a été donné aux échantillons de 2011 et de 2012 qu'aux échantillons des années antérieures (Center for Behavioral Health Statistics and Quality, 2014; chapitre 5).

Pour les besoins de notre étude, nous avons traité ces poids de sous-échantillon et les poids de la NSDUH associés comme étant donnés et basés sur la théorie des sondages. Nous avons également traité les strates et deux unités primaires d'échantillonnage (UPE) de variance pour chacune des 50 strates de variance

établies pour l'estimateur de variance pour le sous-échantillon de la MHSS comme s'il s'agissait des strates de variance et des UPE de variance pour la NSDUH. Enfin, nous avons traité les UPE de la NSDUH comme si elles étaient sélectionnées avec remise.

3.2 Estimation de la variance sous la théorie des sondages

Puisque les estimations des totaux de domaine corrigées du biais dans les équations (2.9) et (2.10) sont presque sans biais sous la théorie des sondages, on peut faire appel à la linéarisation pour estimer leurs variances. Dans la suite de l'exposé, nous utilisons des variantes des estimateurs corrigés du biais donnés par les équations (2.9) et (2.10) pour simplifier l'estimation de la variance.

En nous souvenant que $\omega_k = 0$ quand $k \notin S'$, un estimateur de variance pour la moyenne d'échantillon

$$\overline{y}_{z(d)} = \frac{\sum_S w_k z_k d_k}{\sum_S w_k d_k}, \tag{3.1}$$

sous échantillonnage à plusieurs degrés stratifié, où $z_k = p_k + (\omega_k / w_k)(y_k - p_k)$, est donné par

$$v(\overline{y}_{z(d)}) = \frac{\sum_{h=1}^{50} \left[\sum_{k \in S_{h1}} w_k d_k (z_k - \overline{y}_{z(d)}) - \sum_{k \in S_{h2}} w_k d_k (z_k - \overline{y}_{z(d)}) \right]^2}{\left(\sum_S w_k d_k \right)^2}, \tag{3.2}$$

où S_{hj} représente les répondants dans la j^{e} UPE de variance et la strate de variance h. Il est également un estimateur de variance pour la variante asymptotiquement identique de l'estimateur par probabilité corrigé du biais qui suit :

$$\overline{y}_{P-\text{BC2}(d)} = \frac{\sum_S w_k p_k d_k}{\sum_S w_k d_k} + \frac{\sum_S \omega_k (y_k - p_k) d_k}{\sum_S w_k d_k}. \tag{3.3}$$

Il en est ainsi parce que le sous-échantillon de la MHSS est poissonnien (et donc indépendant sur les adultes ainsi que les UPE) et que le premier degré de l'échantillon de la NSDUH est traité comme si le tirage avait été effectué avec remise.

De même, en redéfinissant $z_k = c_k + (\omega_k / w_k)(y_k - c_k)$, un estimateur de la variance de la moyenne d'échantillon de l'équation (3.1) est également un estimateur de variance pour cette variante de l'estimateur corrigé du biais :

$$\overline{y}_{C-\text{BC2}(d)} = \frac{\sum_S w_k c_k d_k}{\sum_S w_k d_k} + \frac{\sum_S \omega_k (y_k - c_k) d_k}{\sum_S w_k d_k}. \tag{3.4}$$

L'approche d'estimation de la variance susmentionnée suppose que les tailles des sous-échantillons de répondants dans les domaines sont telles que p_k / P_k et c_k / C_k peuvent être traités comme une valeur unitaire, où P_k et C_k sont les limites de p_k et c_k, respectivement, à mesure que le sous-échantillon (simultanément à l'échantillon et à la population de la NSDUH) devient arbitrairement grand. En fait, tous ces ratios sont supposés être $1 + \text{O}_P(1/\sqrt{n})$, où n est la taille du sous-échantillon de la MHSS.

Considérons maintenant un terme de correction du biais calculé, disons $\sum_S \omega_k (y_k - p_k) d_k \Big/ \sum_S w_k d_k$ ou $\sum_S \omega_k (y_k - c_k) d_k \Big/ \sum_S w_k d_k$. Pour déterminer si le terme diffère significativement de zéro, on peut créer une statistique t asymptotique de la manière habituelle, en divisant le terme par son erreur-type.

Pour évaluer les estimateurs à la section 3.3, nous utiliserons plutôt les équivalents asymptotiques :

$$Mesure\ du\ biais(\overline{y}_{P(d)}) = \sum_{S'} \omega_k (y_k - p_k) d_k \Big/ \sum_{S'} \omega_k d_k, \tag{3.5}$$

et

$$Mesure\ du\ biais(\overline{y}_{C(d)}) = \sum_{S'} \omega_k (y_k - c_k) d_k \Big/ \sum_{S'} \omega_k d_k \tag{3.6}$$

pour créer une statistique t asymptotique pour l'évaluation des biais au niveau du domaine, afin que la procédure DESCRIPT dans SUDAAN (RTI International, 2012) puisse être employée pour traiter les p_k et c_k comme étant fixés (similairement, que l'estimateur de variance pour (3.3) dans l'équation (3.2) puisse être calculé en utilisant DESCRIPT). En outre, puisque pratiquement toute l'erreur d'échantillonnage dans les termes de correction du biais provient de la phase de sous-échantillonnage pour la MHSS (même en 2011 et en 2012, le sous-échantillon ne représentait que 3 % de l'échantillon d'adultes de la NSDUH), nous traitons les erreurs-types des mesures du biais comme si elles étaient calculées pour un échantillon de Poisson avec des fractions d'échantillonnage si petites qu'elles peuvent être ignorées, ce qui équivaut à un échantillon d'éléments sélectionné avec remise aux fins de l'estimation de la variance. Par exemple, pour l'estimateur de variance de $Mesure\ du\ biais(\overline{y}_{P(d)})$, nous calculons (en utilisant la procédure DESCRIPT dans SUDAAN) : $v[Mesure\ du\ biais(\overline{y}_{P(d)})] = \frac{n}{n-1} \sum_{S'} [\omega_k \{[y_k - p_k] - Mesure\ du\ biais(\overline{y}_{P(d)})\} d_k]^2 \Big/ \left(\sum_{S'} \omega_k d_k\right)^2$, où n est la taille d'échantillon de S'.

3.3 Évaluation des estimateurs

Le modèle utilisé par la SAMHSA pour prédire la prévalence de la MMG d'après les réponses des adultes à la NSDUH était un modèle logistique à cinq variables (Center for Behavioral Health Statistics and Quality, 2014; chapitre 7). Deux des variables étaient des scores totaux rééchelonnés, obtenus au moyen de courts questionnaires visant à mesurer la détresse psychologique et la déficience fonctionnelle due à la détresse. La troisième était une variable dichotomique $(0/1)$ créée à partir des réponses à une série de questions visant à déterminer si la personne avait vécu un épisode dépressif majeur l'année précédente. La quatrième était également une variable dichotomique $0/1$ qui indiquait si la personne avait sérieusement envisagé de se suicider l'année précédente, et la cinquième était une fonction de l'âge, linéaire de 18 à 30 ans et constante après 30 ans. Des renseignements sur la sélection de ce modèle figurent dans le rapport du Center for Behavioral Health Statistics and Quality (2015, chapitre 4).

Nous nous sommes servis de ce modèle afin de créer un ensemble d'estimations par seuil diagnostique et par probabilité au niveau du domaine à partir des ensembles de données combinés pour 2008 à 2012 et d'évaluer le biais potentiel de ces estimations. Certains résultats sont présentés aux tableaux 3.1 et 3.2. Ces tableaux contiennent des estimations pour des domaines fondés sur des caractéristiques personnelles plutôt que sur l'État de résidence, parce qu'il semblait plus probable que d'importants biais soient observés pour ces caractéristiques que pour les États. En outre, les tailles d'échantillon pour les caractéristiques tendaient à être plus grandes que pour les États.

Le tableau 3.1 montre que l'utilisation de la probabilité corrigée du biais dans l'équation (2.9) est habituellement un peu plus efficace (donne une erreur-type plus petite) que l'estimateur direct $\overline{y}_{U(d)} = \sum_{S'} \omega_k d_k y_k \Big/ \sum_{S'} \omega_k d_k$. L'estimateur par seuil diagnostique corrigé du biais dans l'équation (2.10) est

plus efficace que l'estimateur direct dans certains cas et dans d'autres, non. Dans le tableau 3.1, les erreurs-types sont les racines carrées des estimateurs de variance par linéarisation pour l'estimateur direct $\overline{y}_{U(d)}$ susmentionné ou pour l'estimateur corrigé du biais de l'équation (3.1) avec la variable non aléatoire z_k définie comme il convient, chacun calculé en supposant un échantillon avec remise stratifié d'unités primaires d'échantillonnage et un sous-échantillon probabiliste d'individus à l'intérieur de chaque UPE; c'est-à-dire, avec l'équation (3.2). Pour $v\left(\overline{y}_{U(d)}\right)$, le terme $z_k - \overline{y}_{z(d)}$ est remplacé par $y_k - \overline{y}_{U(d)}$.

Tableau 3.1
Estimateurs presque sans biais avec leurs erreurs-types

	Direct (éq. 2.6)		Seuil diagnostique corrigé du biais (éq. 2.10)		Probabilité corrigée du biais (éq. 2.9)	
	Estimation	E.-t.	Estimation	E.-t.*	Estimation	E.-t.*
Ensemble des adultes	3,93	0,29	3,96	0,26	3,91	0,23
Hommes	2,96	0,34	2,91	0,39	3,01	0,31
Femmes	4,84	0,46	4,93	0,39	4,74	0,36
Âge : 18 à 25 ans	3,77	0,62	3,97	0,48	3,66	0,52
Âge : 26 à 34 ans	4,35	0,68	4,29	0,61	4,37	0,57
Âge : 35 à 49 ans	5,74	0,57	6,15	0,52	5,87	0,50
Âge : 50 ans et plus	2,74	0,40	2,47	0,47	2,60	0,36
Blanc, non hispanique	4,43	0,35	4,47	0,30	4,34	0,27
Noir, non hispanique	3,28	0,54	3,62	0,42	3,38	0,40
Autre, non hispanique	4,09	1,25	4,27	1,10	4,33	1,12
Hispanique	2,02	0,71	1,68	0,88	2,11	0,70
Nord-est	2,80	0,51	3,59	0,49	3,25	0,47
Centre nord	4,17	0,49	3,99	0,53	4,13	0,37
Sud	3,74	0,49	3,93	0,51	3,65	0,45
Ouest	5,04	0,84	4,26	0,57	4,62	0,57
Occupé à temps plein	2,36	0,29	2,36	0,28	2,32	0,25
Occupé à temps partiel	4,34	0,71	3,82	0,55	3,91	0,46
Chômeur	5,64	1,22	6,57	0,92	6,13	0,90
Autre situation d'emploi	6,21	0,66	6,22	0,64	6,15	0,55
Pas de diplôme d'études secondaires	5,69	0,99	4,44	0,77	4,72	0,71
Diplôme d'études secondaires	4,05	0,57	4,08	0,57	4,14	0,44
Études collégiales partielles	4,14	0,57	4,31	0,44	4,18	0,40
Diplôme collégial	2,88	0,52	3,27	0,46	3,01	0,46
Métropolitain	3,78	0,45	3,96	0,39	3,74	0,37
Petite métropole	4,15	0,47	3,60	0,44	3,96	0,29
Non métropolitain	3,99	0,47	4,63	0,54	4,36	0,48
Assurance maladie : Oui	3,57	0,31	3,83	0,26	3,65	0,24
Assurance maladie : Non	5,73	0,94	4,65	0,93	5,24	0,74
< 100 % du seuil de pauvreté	9,01	1,30	9,00	1,23	8,62	1,05
100 % à 199 % du seuil de pauvreté	5,61	0,85	4,72	0,63	4,88	0,52
100 % du seuil de pauvreté	2,59	0,28	2,64	0,28	2,61	0,23
Traitement de santé mentale reçu : Oui	18,84	1,57	19,69	1,29	19,00	1,32
Traitement de santé mentale reçu : Non	1,54	0,18	1,42	0,20	1,46	0,15

* L'erreur-type est égale à la racine carrée de l'estimation de variance calculée en utilisant l'équation (3.2) avec z_k définie de manière appropriée.

Le tableau donne à penser que la correction du modèle apporte peu d'amélioration et nous renvoie aux estimateurs par probabilité ou par seuil diagnostique basés sur le modèle des équations (2.7) et (2.8), à moins que ces estimateurs présentent des biais systématiques. Le tableau 3.2 (avec les mesures du biais et leurs erreurs-types calculées comme il est décrit à la section 3.2) suggère fortement que l'estimateur par

probabilité, quoique sans biais si l'on estime la prévalence de la MMG pour l'ensemble des adultes, peut être très biaisé au niveau du domaine. Par contre, l'estimateur par seuil diagnostique ne présente un biais important au seuil de signification de 0,1 que dans deux domaines, et n'en présente jamais au seuil de signification de 0,05. Puisque nous avons calculé les valeurs p bilatérales pour 32 domaines, deux domaines dont les valeurs p sont inférieures à 0,1 est à peu près ce que l'on devrait s'attendre à observer sous l'hypothèse nulle selon laquelle l'estimateur par seuil diagnostique est sans biais au niveau du domaine.

Tableau 3.2
Estimations basées sur le modèle et leurs mesures du biais

	Seuil diagnostique classique (éq. 2.8)			Probabilité (éq. 2.7)		
	Estimation	Mesure du biais	E.-t. de la mesure du biais	Estimation	Mesure du biais	E.-t. de la mesure du biais
Ensemble des adultes	3,95	-0,01	0,27	3,91	0,00	0,23
Hommes	2,99	0,08	0,42	3,18	0,17	0,34
Femmes	4,84	-0,10	0,34	4,58	-0,16	0,31
Âge : 18 à 25 ans	3,94	-0,02	0,55	3,59	-0,07	0,49
Âge : 26 à 34 ans	5,03	0,69	0,66	4,64	0,26	0,51
Âge : 35 à 49 ans	5,08	-1,10*	0,57	4,77	-1,15**	0,55
Âge : 50 ans et plus	2,84	0,37	0,42	3,21	0,61*	0,32
Blanc, non hispanique	4,31	-0,17	0,33	4,18	-0,16	0,28
Noir, non hispanique	3,14	-0,48	0,45	3,38	0,00	0,46
Autre, non hispanique	3,14	-1,14	1,13	3,47	-0,86	1,08
Hispanique	3,31	1,63*	0,85	3,28	1,17	0,65
Nord-est	3,55	-0,04	0,39	3,62	0,33	0,35
Centre nord	4,16	0,16	0,60	4,02	-0,10	0,40
Sud	3,80	-0,13	0,52	3,86	0,22	0,44
Ouest	4,28	0,02	0,56	4,10	-0,56	0,55
Occupé à temps plein	2,76	0,38	0,33	3,09	0,75**	0,28
Occupé à temps partiel	4,19	0,39	0,59	4,05	0,15	0,47
Chômeur	6,61	0,03	0,75	5,48	-0,57	0,70
Autre situation d'emploi	5,33	-0,93	0,66	4,91	-1,30	0,56
Pas de diplôme d'études secondaires	4,34	-0,11	0,90	4,15	-0,64	0,83
Diplôme d'études secondaires	4,09	0,01	0,59	3,92	-0,22	0,46
Études collégiales partielles	4,50	0,18	0,37	4,35	0,17	0,31
Diplôme collégial	3,09	-0,16	0,46	3,36	0,33	0,40
Métropolitain	3,63	-0,34	0,38	3,68	-0,06	0,35
Petite métropole	4,35	0,73	0,49	4,20	0,23	0,35
Non métropolitain	4,24	-0,38	0,59	4,09	-0,27	0,51
Assurance maladie : Oui	3,67	-0,16	0,27	3,72	0,07	0,24
Assurance maladie : Non	5,39	0,72	0,86	4,89	-0,34	0,68
< 100 % du seuil de pauvreté	7,21	-2,07	1,27	6,13	-2,88**	1,16
100 % à 199 % du seuil de pauvreté	4,83	0,12	0,62	4,53	-0,38	0,55
100 % du seuil de pauvreté	2,98	0,32	0,28	3,24	0,61***	0,21
Traitement de santé mentale reçu : Oui	18,33	-1,37	1,31	13,97	-5,07***	1,26
Traitement de santé mentale reçu : Non	1,62	0,20	0,23	2,28	0,81***	0,17

* La mesure du biais diffère significativement de zéro au seuil de signification de 0,1.
** La mesure du biais diffère significativement de zéro au seuil de signification de 0,05.
*** La mesure du biais diffère significativement de zéro au seuil de signification de 0,01.

Un résultat curieux mérite d'être mentionné brièvement. L'estimateur par seuil diagnostique appliqué à l'ensemble des adultes présente un biais très faible (-0,01), de sorte que la racine carrée de son erreur quadratique moyenne estimée est égale à l'erreur-type de l'estimateur par seuil diagnostique corrigé du biais

après arrondissement (0,26). Étrangement, cette valeur est plus petite que l'erreur-type de sa mesure du biais (0,27). Une raison possible de la différence entre les deux erreurs-types est que nous avons utilisé $\sum_S \omega_k (y_k - c_k) d_k \Big/ \sum_S w_k d_k$ comme terme de correction du biais et $\sum_S \omega_k (y_k - c_k) d_k \Big/ \sum_S \omega_k d_k$ comme mesure du biais dans un domaine; l'ensemble des adultes étant le cas particulier où $d_k \equiv 1$. Notre analyse (non présentée) est que la différence entre les dénominateurs a très peu d'effet.

L'élément qui a un effet plus important est de ne pas tenir compte de la stratification et de la mise en grappes de l'échantillon de la NSDUH pour calculer les erreurs-types des mesures du biais. Étonnamment, ne pas tenir compte de la mise en grappes a effectivement tendance à augmenter les erreurs-types. Cela pourrait tenir au fait que la mise en grappes dans la NSDUH n'a pratiquement aucun effet mesurable sur la variance, de sorte que toute différence entre les estimations de l'erreur-type calculées avec et sans mise en grappes est attribuable à un bruit aléatoire ou à des biais asymptotiques qui ne sont effectivement pas ignorables dans les estimations finies.

3.4 Un seuil diagnostique hybride

Considérons l'hybride qui suit des estimateurs par probabilité et par seuil diagnostique classique. Supposons que nous ayons trié l'échantillon de la NSDUH plutôt que simplement le sous-échantillon de la MHSS en fonction des valeurs de p_k prédites, et que nous ayons établi un seuil diagnostique hybride p_H tel que l'expression

$$\sum_{\substack{k \in S \\ p_k \geq p_H}} w_k = \sum_{k \in S} w_k p_k \tag{3.7}$$

soit vérifiée le plus étroitement possible. En posant que $h_k = 1$ quand $p_k > p_H$ et 0 autrement, l'estimateur par seuil diagnostique hybride de la prévalence de la MMG dans un domaine est

$$\bar{y}_{H(d)} = \frac{\sum_S w_k h_k d_k}{\sum_S w_k d_k}. \tag{3.8}$$

Il n'est pas difficile de voir que, pour l'ensemble des adultes, si l'on peut trouver une valeur de p_H qui satisfait l'équation (3.7), alors l'estimateur par seuil diagnostique hybride sera exactement égal à l'estimateur par probabilité. Sinon, l'estimateur par seuil diagnostique hybride pour l'ensemble des adultes aura un léger biais, qui pourra être mesuré, élevé au carré, puis ajouté à l'erreur-type de l'estimateur par probabilité pour égaler la racine carrée de son erreur quadratique moyenne. Dans ces conditions, l'estimation hybride de la prévalence de la MMG pour l'ensemble des adultes s'arrondit à 3,89. La racine carrée de son erreur quadratique moyenne s'arrondit à la même valeur que l'erreur-type de l'estimateur par probabilité (0,23).

Le tableau 3.3 répète en grande partie le tableau 3.2 pour l'estimateur par seuil diagnostique classique, mais donne aussi les résultats pour l'estimateur hybride

$$(\textit{Mesure du biais}(\bar{y}_{H(d)})) = \sum_S \omega_k (y_k - h_k) d_k \Big/ \sum_S \omega_k d_k. \tag{3.9}$$

L'erreur-type de ce dernier est calculée de manière analogue à celles de $\bar{y}_{C(d)}$ et $\bar{y}_{P(d)}$. Les deux ensembles de résultats fondés sur les seuils diagnostiques sont similaires, mais la mesure du biais pour l'estimateur hybride diffère de manière significative de zéro au seuil de signification de 0,05 dans deux domaines (tous

deux avec une valeur p de 0,043). Puisque 32 domaines sont analysés, ce résultat demeure concordant avec l'hypothèse nulle d'absence de biais au niveau du domaine.

Tableau 3.3
Estimateurs par seuil diagnostique et leurs mesures du biais

	Seuil diagnostique classique (éq. 2.8)			Seuil diagnostique hybride (éq. 3.8)		
	Estimation	Mesure du biais	E.-t. de la mesure du biais	Estimation	Mesure du biais	E.-t. de la mesure du biais
Ensemble des adultes	3,95	-0,01	0,27	3,89	-0,10	0,27
Hommes	2,99	0,08	0,42	2,94	0,03	0,42
Femmes	4,84	-0,10	0,34	4,78	-0,21	0,33
Âge : 18 à 25 ans	3,94	-0,02	0,55	3,89	-0,03	0,55
Âge : 26 à 34 ans	5,03	0,69	0,66	4,97	0,68	0,66
Âge : 35 à 49 ans	5,08	-1,10*	0,57	5,02	-1,16**	0,57
Âge : 50 ans ou plus	2,84	0,37	0,42	2,79	0,22	0,41
Blanc, non hispanique	4,31	-0,17	0,33	4,24	-0,22	0,33
Noir, non hispanique	3,14	-0,48	0,45	3,10	-0,48	0,45
Autre, non hispanique	3,14	-1,14	1,13	3,11	-1,14	1,13
Hispanique	3,31	1,63*	0,85	3,25	1,30	0,79
Nord-est	3,55	-0,04	0,39	3,50	-0,05	0,39
Centre nord	4,16	0,16	0,60	4,12	0,07	0,59
Sud	3,80	-0,13	0,52	3,74	-0,29	0,51
Ouest	4,28	0,02	0,56	4,23	0,01	0,56
Occupé à temps plein	2,76	0,38	0,33	2,71	0,36	0,33
Occupé à temps partiel	4,19	0,39	0,59	4,16	0,37	0,59
Chômeur	6,61	0,03	0,75	6,43	-0,27	0,69
Autre situation d'emploi	5,33	-0,93	0,66	5,27	-1,09*	0,65
Pas de diplôme d'études secondaires	4,34	-0,11	0,90	4,21	-0,14	0,90
Diplôme d'études secondaires	4,09	0,01	0,59	4,03	-0,26	0,56
Études collégiales partielles	4,50	0,18	0,37	4,45	0,18	0,37
Diplôme collégial	3,09	-0,16	0,46	3,07	-0,17	0,46
Métropolitain	3,63	-0,34	0,38	3,58	-0,36	0,38
Petite métropole	4,35	0,73	0,49	4,27	0,58	0,47
Non métropolitain	4,24	-0,38	0,59	4,19	-0,53	0,58
Assurance maladie : Oui	3,67	-0,16	0,27	3,62	-0,20	0,27
Assurance maladie : Non	5,39	0,72	0,86	5,31	0,44	0,82
< 100 % du seuil de pauvreté	7,21	-2,07	1,27	7,12	-2,44**	1,21
100 % à 199 % du seuil de pauvreté	4,83	0,12	0,62	4,78	-0,01	0,61
> 200 % du seuil de pauvreté	2,98	0,32	0,28	2,93	0,30	0,28
Traitement de santé mentale reçu : Oui	18,33	-1,37	1,31	18,19	-1,46	1,31
Traitement de santé mentale reçu : Non	1,62	0,20	0,23	2,28	0,81	1,17

* La mesure du biais diffère significativement de zéro au seuil de signification de 0,1.
** La mesure du biais diffère significativement de zéro au seuil de signification de 0,05.

4 Quelques observations finales

Les attributs de la santé mentale de la population ont été estimés en utilisant des données provenant d'un sous-échantillon de personnes sélectionnées parmi les participants à une grande enquête générale et soumises à une évaluation diagnostique clinique pour élaborer des modèles prévisionnels qui sont ensuite appliqués à l'échantillon complet (voir, par exemple, Kessler, Abelson, Demler, Escobar, Gibbon, Guyer, Howes, Jin, Vega, Walters, Wang, Zaslavsky et Zheng, 2004). Il s'agit de la méthodologie utilisée par la

SAMHSA pour la NSDUH annuelle et le sous-échantillon d'adultes de la MHSS pour la période de 2008 à 2012.

Nous avons montré au moyen des données du sous-échantillon NSDUH/MHSS que l'estimation de la prévalence de la MMG pour l'ensemble de la population, ici les adultes, en utilisant les probabilités directement estimées d'avoir une MMG plutôt que la méthodologie du seuil diagnostique classique produit une erreur-type plus faible. Néanmoins, l'estimateur de la prévalence de la MMG que nous avons convenu d'appeler estimateur par probabilité peut souvent présenter un biais important au niveau du domaine, contrairement à l'estimateur par seuil diagnostique classique.

Nous avons également étudié les versions corrigées du biais des deux estimateurs fondés sur la théorie des sondages. Malheureusement, ces estimateurs étaient à peine plus efficaces que le simple calcul direct des estimations à partir du sous-échantillon de la MHSS, surtout pour les sous-populations.

Nous avons évalué un estimateur par seuil diagnostique hybride qui était légèrement plus efficace que l'estimateur par seuil diagnostique classique pour estimer la prévalence de la MMG chez les adultes. Il ne présentait pas les biais importants au niveau du domaine dont souffrait l'estimateur par probabilité, mais la question de savoir s'il est dépourvu de biais au niveau du domaine n'est pas entièrement éclaircie.

En 2013, la SAMHSA a mis un terme à la réalisation des évaluations cliniques. Néanmoins, l'organisme a continué de calculer les estimations de la prévalence de la MMG chez les adultes en se basant sur le modèle et les seuils diagnostiques établis en se servant du sous-échantillon de la MHSS pour la période de 2008 à 2012. L'estimateur par seuil diagnostique classique qui affichait des biais au niveau du domaine plus faibles dans nos tableaux continue aussi d'être utilisé pour l'estimation de la prévalence de la MMG au niveau de l'État, ainsi que pour les domaines analysés ici.

Une question troublante est celle de savoir comment un estimateur par seuil diagnostique classique pour la MMG dérivé d'un modèle logistique ajusté aux données du sous-échantillon de la MHSS peut être moins sujet au biais qu'un estimateur par probabilité fondé sur un modèle ajusté à l'échantillon complet. Nous soupçonnons que la régression logistique produit un classement raisonnablement bon des probabilités relatives que les adultes aient une MMG, mais non une estimation raisonnablement bonne des probabilités individuelles, spécialement dans les queues de la distribution. Les domaines où les prévalences sont habituellement élevées ou faibles, comme les adultes recevant (ou ne recevant pas) un traitement pour une maladie mentale, se sont avérés particulièrement sujets à des estimations de probabilité biaisées. Il se pourrait que l'application de la théorie asymptotique ne soit pas appropriée dans les queues de la distribution.

Le lecteur que la question intéresse voudra vraisemblablement savoir comment on peut calculer les erreurs-types pour les estimateurs par seuil diagnostique au niveau du domaine. Parce qu'un estimateur par seuil diagnostique n'est pas continu, nous avons essayé de calculer les erreurs-types par la méthode des répliques répétées équilibrées ou méthode BRR (de l'anglais *balanced repeated replications*) de Fay. Malheureusement, comme il est expliqué dans le rapport du Center for Behavioral Health Statistics and Quality (2015, chapitre 2.4.2.), nos estimateurs de variance pour les estimations de domaine n'étaient pas satisfaisants dans une modeste expérience par simulation.

Peu de répondants du sous-échantillon de la MHSS se situaient dans les queues de la distribution (et un nombre encore plus faible de valeurs réalisées pour la probabilité d'avoir une MMG sachant les covariables

du modèle). Cela a contrarié nos efforts visant à améliorer l'estimateur par probabilité (pour lequel les erreurs-types au niveau du domaine peuvent être mesurées par linéarisation) par ajustement des probabilités dans les queues (lesquelles n'étaient pas très bien prédites sous le modèle logistique). Nous ne discuterons pas de ces efforts plus en détail ici. Davantage de travaux de recherche sont manifestement nécessaires pour élaborer de bonnes mesures de l'erreur-type pour les estimations par seuil diagnostique classique ou de bonnes estimations de la prévalence de la MMG pour lesquelles les erreurs-types peuvent être mesurées raisonnablement.

Un dernier commentaire au sujet de la « norme de référence » s'impose. Pour les besoins de notre étude, nous avons traité un diagnostic clinique de MMG chez un adulte comme équivalant au fait que la personne était réellement atteinte d'une maladie mentale grave diagnostiquable. En fait, les diagnostics sont plus fluides que cela. Ils peuvent varier selon le clinicien ou l'humeur de la personne qui répond aux questions du clinicien. Le rapport du CBHSQ (2014, chapitres 5 et 6) décrit l'effort déployé en vue d'éliminer autant de variation que possible des diagnostics cliniques de la MHSS. Nous avons supposé ici que chaque diagnostic clinique était effectivement sans biais, c'est-à-dire que la probabilité qu'un diagnostic aléatoire pour une personne dans un domaine soit un faux positif égalait la probabilité qu'il soit un faux négatif.

Bibliographie

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* (4ième Éd.). Washington, DC: American Psychiatric Association.

Center for Behavioral Health Statistics and Quality (2015). *Estimating Mental Illness among Adults in the United States: Revisions to the 2008 Estimation Procedures*. Substance Abuse and Mental Health Services Administration, Rockville, MD: http://www.samhsa.gov/data/sites/default/files/NSDUH-N8-EstimatingMI-2012.pdf.

Center for Behavioral Health Statistics and Quality (2014). *2012 National Survey on Drug Use and Health: Methodological Resource Book (Section 16a, 2012 Mental Health Surveillance Study: Design and Estimation Report)*. Substance Abuse and Mental Health Services Administration, Rockville, MD: http://www.samhsa.gov/data/sites/default/files/NSDUH2012MRB-Ammended/NSDUHmrbMHSS-DesignEst2012.pdf.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874.

Kessler, R.C., Abelson, J., Demler, O., Escobar, J., Gibbon, M., Guyer, M., Howes, M., Jin, R., Vega, W., Walters, E., Wang, P., Zaslavsky, A. et Zheng, H. (2004). Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMHCIDI). *International Journal of Methods in Psychiatric Research*, 13, 122-139.

Kessler, R.C., Chiu, W., Demler, O. et Walters, E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62, 617-627.

Lehtonen, R., et Veijanen, A. (1998). Estimateurs de régression généralisés logistiques. *Techniques d'enquête*, 24, 1, 53-58. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1998001/article/3909-fra.pdf.

RTI International (2012). *SUDAAN Language Manual, Volumes 1 and 2, Release 11*. Research Triangle Park, NC: Research Triangle Institute.

Särndal, C.-E., Swensson, B. et Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.

Techniques d'enquête, juin 2018
Vol. 44, nº 1, pp. 81-107
Statistique Canada, nº 12-001-X au catalogue

Stratégies de sous-échantillonnage des non-répondants pour les programmes économiques

Katherine Jenny Thompson, Stephen Kaputa et Laura Bechtel[1]

Résumé

Le *U.S. Census Bureau* étudie des stratégies de sous-échantillonnage des non-répondants en prévision de l'*Economic Census* de 2017. Les contraintes imposées au plan de sondage comprennent une borne inférieure obligatoire pour le taux de réponse totale, ainsi que des taux de réponse cibles par industrie. Le présent article expose la recherche sur les procédures de répartition de l'échantillon pour le sous-échantillonnage des non-répondants conditionnellement à ce que ce sous-échantillonnage soit systématique. Nous considérons deux approches, à savoir 1) l'échantillonnage avec probabilités égales et 2) la répartition optimisée avec contraintes sur les taux de réponse totale et la taille d'échantillon, avec pour objectif la sélection de plus grands échantillons dans les industries qui, au départ, affichent des taux de réponse plus faibles. Nous présentons une étude en simulation qui examine le biais relatif et l'erreur quadratique moyenne pour les répartitions proposées, en évaluant la sensibilité de chaque procédure à la taille du sous-échantillon, aux propensions à répondre et à la procédure d'estimation.

Mots-clés : Programme quadratique; taux de réponse totale; correction de la non-réponse; échantillonnage systématique; répartition optimale; échantillonnage à deux phases.

1 Introduction

De nombreux programmes fédéraux sont confrontés simultanément à une baisse des taux de réponse et à des restrictions budgétaires. Parallèlement, ces programmes sont tenus de maintenir des niveaux de fiabilité prédéterminés, et sont invités à recueillir un nombre accru d'éléments de données et à publier davantage de statistiques. Naturellement, à mesure qu'augmente la non-réponse, la précision des estimations d'enquête diminue par rapport aux niveaux originaux prévus par le plan de sondage et peut être sensible au biais de non-réponse. Par conséquent, de nombreux organismes fédéraux étudient des stratégies avec plan de collecte adaptatif, le terme « plan de collecte » faisant référence au(x) protocole(s) de collecte des données.

Dans le cas des enquêtes auprès des entreprises, le plan de collecte peut varier selon le type d'unité. Ces populations sont généralement très asymétriques; la plus grande part d'un total de tableau pour une industrie donnée est souvent attribuable à un petit nombre de grandes entreprises. Comme les statistiques publiées sont généralement des totaux ou des variations en pourcentage au niveau de l'industrie, les données manquantes pour les entreprises les plus grandes peuvent induire un biais de non-réponse important dans les totaux, tandis que les données manquantes pour les entreprises plus petites (même celles pour lesquelles les poids d'échantillonnage sont grands) ont souvent peu d'effet *apparent* sur les niveaux des totaux (Thompson et Washington, 2013). Donc, les stratégies de prise de contact sont conçues de manière à s'assurer que les entreprises les plus grandes fournissent des réponses valides. La figure 1.1 illustre des procédures de suivi des cas de non-réponse (SCNR) différentes selon la classification par taille des unités

1. Katherine Jenny Thompson, Economic Statistical Methods Division, U.S. Census Bureau, Washington, DC 20233. Courriel : Katherine.J.Thompson@census.gov; Stephen Kaputa, Economic Statistical Methods Division, U.S. Census Bureau, Washington, DC 20233. Courriel : Stephen.Kaputa@census.gov; Laura Bechtel, Economic Statistical Methods Division, U.S. Census Bureau, Washington, DC 20233. Courriel : Laura.Bechtel@census.gov.

propre à l'enquête, les plans de collecte ayant tous deux des calendriers de collecte fixes et un budget de SCNR fixe.

Pour la catégorie des grandes unités, les procédures de SCNR deviennent de plus en plus coûteuses (par unité), à l'exception de la dernière tentative de prise de contact. Par contre, pour les unités plus petites, les procédures de SCNR ne comportent pas de prise de contact en personne et sont par conséquent moins coûteuses.

La sélection d'un sous-échantillon probabiliste de non-répondants est une caractéristique stratégique de nombreux plans de collecte dynamiques et adaptatifs (Tourangeau, Brick, Lohr et Li, 2016). Évidemment, il ne s'agit pas d'une nouveauté. En effet, le sous-échantillonnage des non-répondants est pratiqué dans les enquêtes depuis la première discussion à ce sujet dans Hansen et Hurwitz (1946). En fait, la configuration de l'approche d'échantillonnage à deux phases présentée dans l'article de Hansen et Hurwitz (1946) est fort semblable à celle des enquêtes auprès des entreprises discutées ici, à savoir l'envoi par la poste « peu coûteux » d'un questionnaire à toutes les unités échantillonnées (cf. le « plan du XXI^{e} siècle » qui comprend l'envoi par la poste d'une lettre contenant une URL, un nom d'utilisateur et un mot de passe), suivi par des interviews sur place « coûteuses » des non-répondants sous-échantillonnés (cf. appels téléphoniques personnels ou lettres de rappel certifiées). Les procédures de répartition optimales que ces auteurs proposent ne sont pas entièrement différentes non plus, les répartitions finales dépendant fortement du fait que les taux de réponse pour chaque mode de collecte sont connus ou estimés en utilisant des données auxiliaires, plutôt que les réponses recueillies antérieurement.

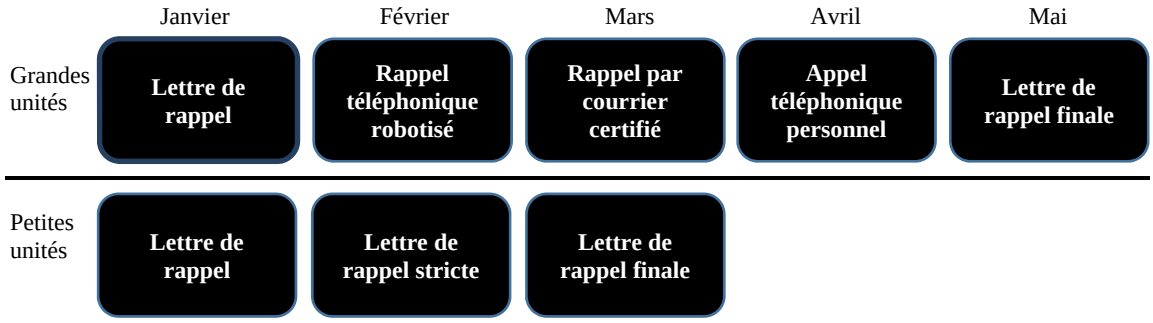

Figure 1.1 Procédures de suivi des cas de non-réponse pour différents types d'entreprises dans une enquête fictive.

Établir le sous-échantillonnage des non-répondants dans un cadre de collecte dynamique ou adaptative est un exercice simple. Comme l'ont proposé au départ Groves et Herringa (2006), les plans de collecte dynamiques requièrent au moins deux phases distinctes de collecte, la seconde étant souvent un sous-échantillon probabiliste de non-répondants lorsqu'est atteinte la « capacité de phase », point où les estimations d'enquête ne varient plus, donnant la preuve que le protocole de collecte existant n'est plus rentable. Schouten, Calinescu et Luiten (2013) caractérisent les plans de collecte dynamiques comme un cas particulier des plans de collecte adaptatifs. Sous un plan de collecte adaptatif, les procédures de collecte des

données peuvent changer (*s'adapter*) durant la période de collecte. Les paradonnées et les données d'échantillon sont utilisées pour déterminer s'il faut modifier les procédures courantes. Le budget global est fixe, mais la mise en œuvre d'une stratégie donnée dépend 1) de l'échantillon de répondants réalisé à un point dans le temps, 2) des données informatives obtenues durant la collecte des données au sujet des répondants et des non-répondants, et 3) de l'information connue sur les unités d'enquête extraite de la base de sondage. Par conséquent, la sélection d'un échantillon probabiliste de non-répondants pour le SCNR – au lieu d'essayer de prendre contact avec tous les non-répondants – rentre dans la classe des plans de collecte adaptatifs, les paradonnées (surtout le statut de réponse) étant utilisées pour déterminer la base de sondage et les données de la base de sondage (par exemple, taille de l'unité et classification industrielle) qui serviront de base pour le plan d'échantillonnage.

Le *U.S. Census Bureau* étudie des stratégies de sous-échantillonnage des non-répondants pour l'*Economic Census* (EC) de 2017. Bien qu'il s'agisse d'un programme unique, l'EC emploie différents plans d'échantillonnage selon le secteur (probabilité proportionnelle à la taille pour le secteur de la construction, échantillonnage avec seuil pour les secteurs de la fabrication et des mines pour les collectes de données antérieures à 2017, dénombrement complet pour le secteur du commerce de gros, et échantillonnage aléatoire simple stratifié sans remise (EAS-SR) pour les autres secteurs). En outre, comme il est habituel pour de nombreux programmes visant les entreprises, il s'agit d'une collecte polyvalente, durant laquelle les éléments statistiques généraux sont recueillis auprès de toutes les unités sondées dans un secteur; par exemple, mais sans s'y limiter, recettes/livraisons, paie annuelle et du premier trimestre, et effectif total. De surcroît, le programme de l'EC recueille des renseignements sur les ventes de produits, dont les types diffèrent selon le secteur et, souvent, selon l'industrie. Les procédures d'imputation varient selon l'élément de donnée, de même que les estimateurs. Par conséquent, le plan de sous-échantillonnage doit être robuste à l'échantillonnage et à l'estimateur dans la mesure du possible. Nous considérons un *échantillonnage systématique* des non-répondants triés selon une mesure de taille, c'est-à-dire un plan d'échantillonnage que l'on sait être aussi efficace qu'un échantillonnage aléatoire simple stratifié sans remise (EAS-SR) en moyenne, si la liste est en ordre aléatoire, et plus efficace si la liste augmente ou diminue de façon monotone (Zhang, 2008; Lohr, 2010, chapitre 2, pages 50 et 51).

Idéalement, la procédure de répartition du sous-échantillon de non-répondants devrait s'appuyer sur les propriétés de l'échantillon de répondants *durant* la période de collecte. Évidemment, si le programme est conçu pour recueillir un ou deux éléments de données clés, les procédures de répartition doivent (au moins essayer d') intégrer directement l'information sur le plan de sondage et la procédure d'estimation, ainsi que les données de coût détaillées, comme l'ont proposé Hansen et Hurwitz (1946) il y a longtemps. Dans ce cas, on doit utiliser une procédure de répartition optimale qui minimise les coûts sous des contraintes (estimables) de fiabilité. Voir Harter, Mach, Chaplin et Wolken (2007) ainsi que Beaumont, Bocci et Haziza (2014) pour des exemples.

Ce genre d'optimisation est difficile à réaliser dans les conditions d'enquêtes polyvalentes considérées, surtout si des variables auxiliaires fortement corrélées ne sont pas disponibles pour tous les éléments de données. Cependant, les normes statistiques de l'OMB pour les enquêtes fédérales requièrent que (le plan) d'enquête produise les taux de réponse les plus élevés possibles, proportionnés à l'importance des usages

de l'enquête, au fardeau de réponse et aux coûts de collecte des données, et rendent obligatoires des analyses du biais de non-réponse pour les programmes qui n'atteignent pas ces taux (Federal Register Notice, 2006). Pour le sous-échantillonnage des non-répondants effectué *durant* le cycle de collecte des données, imposer des bornes inférieures obligatoires sur le taux de réponse au niveau du programme et pour des domaines particuliers (par exemple strates d'échantillonnage ou d'autres poststrates, telles que le code d'industrie ou le type de gouvernement) est par conséquent une contrainte naturelle qui doit être incluse dans la procédure de répartition.

Dans le présent article, nous explorons les approches de répartition qui tiennent compte de ce genre de contraintes, l'objectif global étant de sélectionner de grands sous-échantillons systématiques dans les domaines pour lesquels les taux de réponse sont inférieurs aux taux cibles. Nous introduisons deux procédures de répartition optimisée, toutes deux formulées sous forme de programme quadratique et résolues au moyen de progiciels classiques : l'une qui minimise les écarts entre les taux de réponse totale de domaine et l'autre qui minimise les écarts entre les intervalles de sous-échantillonnage de domaine. Notre étude de cas vise à comparer les propriétés statistiques des sous-échantillons obtenus au moyen de chaque méthode de répartition proposée avec trois estimateurs distincts, à savoir deux estimateurs par le ratio utilisés couramment par les programmes d'enquêtes auprès des entreprises, ainsi que le simple estimateur par expansion (Horvitz-Thompson). Ce dernier n'est pas l'estimateur le plus précis lorsque l'on dispose de données auxiliaires fortement corrélées, mais il donne une « borne supérieure » pour l'augmentation de la variance due au sous-échantillonnage. Les estimateurs par le ratio ont été choisis pour illustrer le fait que la composante de variance de sous-échantillonnage peut être réduite en intégrant des données auxiliaires corrélées à l'étape de l'estimation.

Notons que les procédures de répartition présentées sont conçues spécialement pour les enquêtes auprès des entreprises et qu'elles supposent implicitement que les unités les plus grandes sont exclues du sous-échantillonnage. Dans ce cas, la réduction globale du coût peut ne pas être importante, parce que la majorité du budget de SCNR d'un programme sera vraisemblablement affectée à l'obtention de réponses auprès des grandes unités désignées. Cependant, la qualité des estimations peut être améliorée. En égalisant les taux de réponse dans les domaines pris en considération, nous espérons réduire le biais des estimations grâce à l'obtention d'un ensemble de répondants qui ressemble à l'échantillon parent. De surcroît, égaliser les intervalles de sous-échantillonnage devrait aider à éviter d'accroître indûment la variance d'échantillonnage en raison de la deuxième phase de sélection, un effet secondaire contrariant de l'étape supplémentaire d'échantillonnage susceptible de compenser entièrement toute réduction du biais obtenue au moyen du sous-échantillon probabiliste (Biemer, 2010). En outre, il pourrait être possible de réduire encore davantage le biais de non-réponse ainsi que la variance de sous-échantillonnage au moyen d'une procédure améliorée d'estimation par le ratio ou par la régression, si des covariables reliées sont disponibles.

À la section 2, nous donnons le contexte, décrivons brièvement les estimateurs étudiés, et présentons nos procédures de répartition. À la section 3, nous présentons une étude en simulation conçue pour comparer les propriétés statistiques des estimateurs étudiés pour chaque répartition réalisée. Nous concluons, à la section 4, par des recommandations et des suggestions de futurs travaux de recherche.

2 Méthodologie

2.1 Plan de sondage et estimation

Le cadre général de notre recherche est le plan d'échantillonnage à deux phases illustré à la figure 2.1. La première phase est le tirage d'un échantillon probabiliste stratifié dont la taille totale est n à partir d'une population finie (base de sondage) de taille N, effectué *avant* que débute la collecte des données. L'enquête est réalisée, et les unités sélectionnées répondent ou non. Durant la collecte des données, les taux de réponse sont surveillés dans H domaines, où les domaines ne sont pas nécessairement égaux aux strates d'échantillonnage. Par exemple, les taux de réponse totale pourraient être surveillés par classe d'industries au niveau à trois chiffres, même si les strates d'échantillonnage des industries sont ventilées par catégorie de taille. En outre, les domaines pourraient être indépendants des strates d'échantillonnage originales, par exemple, correspondre à des catégories de race ou de sexe (ressemblant à des poststrates). Dans la suite de l'exposé, le terme « domaine » désigne les strates de sous-échantillonnage des non-répondants, désignées par l'indice $h\,(h = 1, 2, \ldots, H)$.

La deuxième phase d'échantillonnage probabiliste a lieu à un point prédéterminé du cycle de collecte des données, où nous sélectionnons un sous-échantillon global de 1 unité sur K de taille m_1 pour les m non-répondants (un échantillon à deux phases); ce point prédéterminé peut correspondre à une date de l'année civile fixe ou être établi au moyen d'un protocole de collecte dynamique/adaptative. La valeur de K est déterminée par les gestionnaires du programme, en tenant compte du budget global affecté au SCNR (supposé fixe), des mesures de performance exigées (par exemple, taux de réponse, exigences concernant les coefficients de variation), et d'autres considérations opérationnelles, telles que la longueur de la période de collecte et les ressources disponibles. Notre procédure de répartition détermine le sous-échantillon systématique de 1 unité sur K_h de taille m_{1h} pour les m_h non-répondants dans chaque domaine. Seules les m_{1h} unités échantillonnées font l'objet du SCNR.

Notre objectif consiste à estimer Y, le total de population de la caractéristique y. Cette estimation est $\hat{Y} = \hat{Y}_{R1} + \hat{Y}_{R2}$ où $\hat{Y}_{R1}$ est estimé d'après les r_{1h} répondants de l'échantillon de première phase et $\hat{Y}_{R2}$ est estimé d'après les r_{2h} répondants de l'échantillon de seconde phase (voir la figure 2.1). Les ajustements pour la non-réponse apportés aux r_{2h} unités sous-échantillonnées (répondantes) supposent un mécanisme de réponses manquant au hasard (MAR), traité comme un échantillon de Bernoulli (Särndal, Swensson et Wretman, 1992, chapitre 15; Kott, 1994). Nous considérons trois estimateurs par ajustement des poids de sondage (repondération) distincts de $\hat{Y}_{R2}$ (Kalton et Flores-Cervantes, 2003), à savoir l'estimateur par double expansion (DE) repondéré (Binder, Babyak, Brodeur, Hidiroglou et Wisner, 2000; Shao et Thompson, 2009; Haziza, Thompson et Yung, 2010), un estimateur par ratio distinct (SR pour *separate ratio*), qui corrige la non-réponse totale en utilisant une covariable fortement corrélée à la propension à répondre ainsi qu'à la caractéristique d'intérêt de l'enquête (Shao et Thompson, 2009; Haziza et coll., 2010), et un estimateur par ratio combiné (CR pour *combined ratio*) (Binder et coll., 2000). Les formules sont données en annexe.

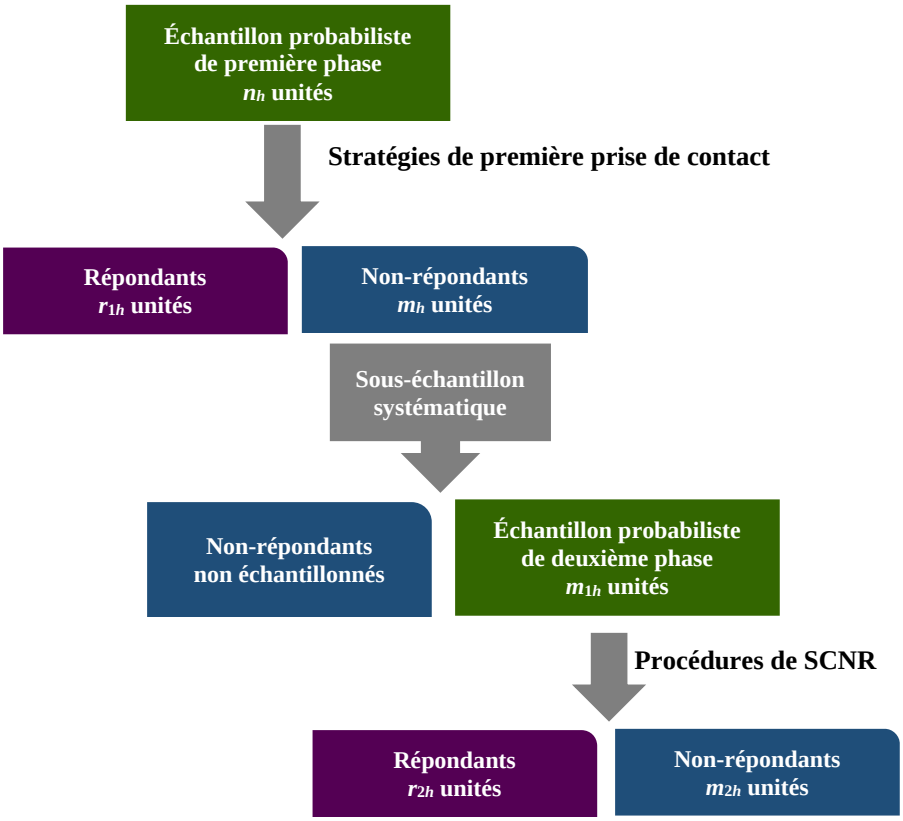

Figure 2.1 Sous-échantillon de non-répondants tiré de l'échantillon probabiliste, sélectionné durant la collecte des données (plan d'échantillonnage à deux phases). Les non-répondants non échantillonnés ne sont pas visés par le SCNR.

Ces estimateurs requièrent un *minimum* de $r_{2h} = 1$ unité dans chaque domaine et un minimum de $r_{2h} = 2$ pour l'estimation de la variance. Ces conditions minimales peuvent ne pas être vérifiées pour plusieurs raisons. Durant les premiers stades de la collecte du SCNR, un nombre insuffisant d'unités sous-échantillonnées pourraient répondre dans un domaine donné. Ou, la procédure de répartition pourrait déterminer qu'aucun sous-échantillonnage n'est requis dans un ou plusieurs domaines. Enfin, la procédure de répartition pourrait nécessiter un suivi total (*toutes* les unités sous-échantillonnées) dans certains domaines; donc, nous désignons le suivi total/sans sous-échantillonnage comme étant un « suivi complet ». Dans ces cas, la procédure d'estimation ignore la dernière phase d'échantillonnage comme si celle-ci n'avait pas eu lieu et produit des estimations pour le domaine h en utilisant la formule de l'estimateur regroupé fournie à l'annexe.

2.2 Stratégies de répartition

Si *tous* les cas non répondants sont soumis au SCNR, les stratégies de prise de contact avec les répondants sont axées sur l'amélioration des taux de réponse globaux. Les analystes pourraient se concentrer

principalement sur l'obtention de réponses auprès des cas de refus non catégorique qui, à leur avis, ont des caractéristiques similaires à celles des répondants précédents (« succès rapides »), quoique ce phénomène soit plus probable quand la collecte des données est effectuée sur le terrain, comme dans le cas des enquêtes auprès des ménages ou des enquêtes agricoles, et peut-être moins quand la collecte se fait par Internet ou par la poste. Dans le cas des enquêtes auprès des entreprises, la taille de l'unité est un facteur qui joue un rôle dans les procédures de SCNR, comme il est discuté à la section 1.

Notre objectif est d'obtenir un ensemble réalisé de répondants qui est approximativement un sous-échantillon aléatoire de l'échantillon sélectionné au départ au moyen d'un échantillon probabiliste de non-répondants. Si l'échantillon est probabiliste, les cas ciblés constituent un ensemble représentatif de la population de non-répondants. En concentrant les efforts de prise de contact sur le sous-échantillon, nous espérons diminuer les effets du biais de non-réponse sur les totaux estimés en obtenant des données auprès de tous les types d'unités non répondantes. En outre, les méthodes de pondération ou d'imputation pourraient réduire plus efficacement les effets du biais de non-réponse dans le cas d'un sous-échantillon probabiliste de non-répondants (Brick, 2013). Même s'ils ne font pas l'objet d'un SCNR supplémentaire, les non-répondants non échantillonnés peuvent fournir des réponses plus tard durant le cycle de collecte. Le cas échéant, une procédure d'estimation sans biais n'inclurait pas les réponses tardives non échantillonnées dans l'estimation finale, en supposant que toutes les unités sous-échantillonnées répondent, car ces unités sont représentées par les cas sous-échantillonnés. Cependant, cette procédure déplaît beaucoup à de nombreux gestionnaires d'enquête. Donc, nous incluons plutôt les données de ces unités dans les totalisations comme si elles avaient répondu *avant* le sous-échantillonnage. Cela induit un biais dans l'estimation. En pratique, nous veillons à ce que cette situation se produise peu fréquemment en effectuant le sous-échantillonnage à un stade tardif du cycle de collecte des données.

Dans le cas d'une enquête auprès des entreprises où ne sont suivies que peu de données démographiques, voire aucune, la plupart de l'information sur les non-répondants, telles l'industrie et la taille de l'unité (par exemple, paie totale, recettes totales) est tirée de la base de sondage. La répartition des non-répondants dans des domaines préalablement spécifiés par taille d'unité et la sélection d'un échantillon systématique devraient produire un sous-échantillon qui ressemble à l'échantillon de départ en ce qui a trait à sa composition en tailles d'unité. Cela est particulièrement important dans le cas des enquêtes auprès des entreprises où l'on a tendance à obtenir les réponses auprès des grandes unités (Thompson et Washington, 2013). Le choix du domaine de sous-échantillonnage est déterminé par les objectifs globaux de l'enquête, tels que les niveaux de publication, ou par la conception des cellules d'ajustement (par exemple, cellules de pondération ou classes d'imputation), quoique les calculs soient considérablement simplifiés quand le domaine d'intérêt coïncide avec la strate d'échantillonnage originale. Dans l'EC, l'industrie est le domaine d'intérêt.

Nous considérons deux approches de répartition, à savoir 1) l'échantillonnage avec probabilités égales, et 2) la répartition optimisée avec contraintes sur les taux de réponse totale et la taille d'échantillon dans des domaines prédéterminés. L'échantillonnage avec probabilités égales est facile à mettre en œuvre et devrait

présenter la variance d'échantillonnage la plus faible parmi les stratégies de répartition du sous-échantillon de non-répondants envisagées, puisque le facteur d'ajustement des poids de sous-échantillonnage aura une valeur constante dans tous les domaines. Cependant, puisque la même proportion de non-répondants est échantillonnée dans chaque domaine, le sous-échantillon pourrait ne pas être suffisamment grand pour compenser les effets du biais de non-réponse sur les totaux dans les domaines à faible réponse. Nous désignons les répartitions obtenues par échantillonnage avec probabilités égales par le terme Constant $-K$, où K représente l'intervalle d'échantillonnage global $(1 \text{ sur } K)$.

Nos méthodes de répartition optimisée tiennent compte de la préoccupation susmentionnée en concentrant les efforts de SCNR dans les domaines où les taux de réponse sont faibles, en essayant de sélectionner suffisamment de cas pour obtenir les valeurs de performance de référence. Cette stratégie peut réduire le biais de non-réponse dans les totaux si le mécanisme de réponse est de type MAR, conditionnellement aux variables auxiliaires utilisées pour définir les domaines; voir Wagner (2012). Cependant, elle peut accroître la variance, car les intervalles de sous-échantillonnage différeront et les poids deviendront plus variables. Afin de réduire au minimum la variance d'échantillonnage additionnelle causée par des intervalles d'échantillonnage différents, les intervalles de *sous-échantillonnage des non-répondants* dans les domaines devraient être proches de l'intervalle global de sous-échantillonnage des non-répondants. Pour contrôler les coûts, la répartition ne devrait pas aboutir à la sélection d'un plus grand nombre d'unités qu'il n'a été budgété pour le SCNR. Rappelons que, dans le contexte des enquêtes fédérales, des taux de réponse cibles doivent être atteints ou quasi atteints, ce qui rend tous les domaines « également » importants du point de vue de la collecte des données.

Pour décrire les procédures de répartition, nous introduisons une notation supplémentaire :

taux de réponse totale : $$\text{TRT} = \frac{\sum_h (r_{1h} + r_{2h})}{\sum_h n_h}$$

taux de réponse cible : $$\text{TRT}^C = \frac{\sum_h r_{1h} + (q_h m_h / K)}{\sum_h n_h}$$

taux de réponse cible de domaine : $$\text{TRT}_h^C = \frac{r_{1h} + (q_h m_h / K_h)}{n_h}$$

avec r_{1h} unités des n_h unités échantillonnées au départ répondant *avant* le sous-échantillonnage, ce qui laisse m_h unités disponibles pour le sous-échantillonnage dans chaque domaine. Le taux de réponse totale (TRT) est la proportion *réelle* d'unités échantillonnées répondantes (Thompson et Oliver, 2012) et *n*'inclut *pas* d'ajustement pour le sous-échantillonnage. Le taux de réponse cible (TRT^C) utilisé pour la répartition est le TRT réalisable maximal prévu pour un taux de sous-échantillonnage donné K, avec q_h représentant la probabilité conditionnelle de répondre en dernière analyse au recensement/à l'enquête dans le domaine h, sachant que l'unité n'a pas répondu avant le sous-échantillonnage. Dans la procédure de répartition, q_h

peut être modélisé d'après les données historiques s'il en existe, ou supposé constant pour une nouvelle enquête ou pour des analyses de sensibilité.

Nous formulons la répartition optimale comme un programme quadratique et considérons deux fonctions objectif différentes. Le premier programme quadratique minimise le carré de l'écart du taux de réponse cible dans chaque domaine TRT_h^C par rapport au taux de réponse totale cible global TRT^C, sous des contraintes linéaires sur la taille de l'échantillon de non-répondants. Cette fonction objectif est analogue au numérateur du test d'adéquation du khi carré de Pearson.

Le deuxième programme quadratique minimise le carré de l'écart des intervalles d'échantillonnage de domaine par rapport à l'*intervalle d'échantillonnage global* (K) sous des contraintes linéaires sur les taux de réponse totale dans chaque domaine et sur le nombre de non-répondants échantillonnés. Donc, même si la procédure d'optimisation permet que les intervalles d'échantillonnage varient selon le domaine, le programme essaie d'éviter des accroissements éventuellement grands de la variance causés par les « fractions d'échantillonnage disproportionnées » introduites délibérément mentionnées dans Kish (1992). Nous dénommons les répartitions obtenues au moyen de ces programmes quadratiques $\text{Min}-\text{TRT}$ et $\text{Min}-K$, respectivement.

Les deux programmes quadratiques sont principalement déterministes. Cependant, rappelons qu'à l'étape de la répartition, nous devons estimer le nombre d'unités sous-échantillonnées qui finiront par répondre dans chaque domaine. Les programmes quadratiques utilisent tous deux les contraintes (1) à (3) dans le tableau 2.1. La contrainte (4) est incluse dans la répartition $\text{Min}-K$ pour s'assurer que la solution d'optimisation ne soit pas $K_h = K$ pour tout domaine h. Il existe deux scénarios limites (conditions préalables) que nous abordons avant l'optimisation $\text{Min}-K$. Premièrement, les domaines pour lesquels $\text{TRT}_h^C \geq \text{TRT}^C$ *avant* le sous-échantillonnage doivent être éliminés du problème d'optimisation $(K_h = \infty)$. Deuxièmement, si le taux de réponse totale estimé ne peut pas être atteint dans un domaine donné pour une valeur q_h supposée, alors toutes les unités dans le domaine sont sélectionnées pour le SCNR $(K_h = 1)$. L'optimisation $\text{Min}-K$ est appliquée aux domaines restants, ce qui requiert que les TRT de ces domaines sous-échantillonnés égalent ou dépassent les TRT cibles.

En utilisant des données d'échantillon contenant les répondants et les non-répondants, ainsi que différentes valeurs spécifiées de q_h, nous nous servons de la procédure PROC NLP de SAS® (L'analyse des données du présent article a été effectuée en utilisant le logiciel SAS. Tous droits réservés, SAS Institute Inc. SAS et tous les autres noms de produit ou service du SAS Institute Inc. sont des marques de commerce déposées ou des marques de commerce du SAS Institute Inc., Cary, NC, USA.) pour résoudre les programmes quadratiques (en obtenant l'ensemble de K_h). Les répartitions réalisées ne sont pas des valeurs entières, et les valeurs réelles des intervalles (K_h) ont été entrées dans la procédure PROC SURVEYSELECT de SAS® pour sélectionner les sous-échantillons systématiques stratifiés de non-répondants. Comme l'a mentionné un examinateur, cela donne une solution qui est arrondie aléatoirement, mais contrainte à la taille d'échantillon globale requise, ce qui pourrait avoir une incidence sur la fiabilité en raison de l'erreur d'arrondissement. Ces genres d'effets n'ont pas été étudiés dans le présent article.

Tableau 2.1
Programmes quadratiques des répartitions optimisées

		Min−TRT	Min−K	But
Fonction objectif		$\min \sum_h (\mathrm{TRT}_h^C - \mathrm{TRT}^C)^2$	$\min \sum_h (K_h - K)^2 = \min \sum_h \left(\frac{m_h}{m_{1h}} - K \right)^2$	
Contraintes	(1)	$K \leq \sum_h m_h \Big/ \sum_h m_{1h}$		La taille de l'échantillon sélectionnée ne peut pas dépasser la taille globale de l'échantillon de 1 sur K.
	(2)	$m_h / m_{1h} \geq 1$		La taille du sous-échantillon de domaine ne peut pas dépasser le nombre de non-répondants dans la strate.
	(3)	$m_{1h} \geq 0$		Contrainte de non-négativité
	(4)	Sans objet	$\frac{r_{1h} + q_h m_h}{n_h} < \mathrm{TRT}^C \quad K_h = 1$ $\frac{r_{1h}}{n_h} \geq \mathrm{TRT}^C \quad K_h = \infty$ $\mathrm{TRT}_h^C \geq \mathrm{TRT}^T \quad$ autrement	Assure que le TRT cible est atteint dans tous les domaines dans la mesure du possible.

3 Étude de cas

À la présente section, nous présentons les résultats d'une étude en simulation effectuée pour évaluer les procédures de répartition considérées sur des données d'échantillon empiriques provenant des cycles de collecte des données de 2010 et de 2011 de l'*Annual Survey of Manufactures* (ASM). Pour plus de renseignements sur l'ASM, consulter http://www.census.gov/manufacturing/asm.

L'ASM est une enquête auprès des établissements conçue pour estimer sur échantillon des statistiques pour tous les établissements manufacturiers comptant un ou plusieurs employés rémunérés (http://www.census.gov/manufacturing/asm/); il s'agit d'un échantillon Pareto-PPT d'environ 50 000 établissements sélectionnés à partir d'un univers de 328 500 établissements. Environ 20 000 établissements sont tirés avec certitude, et les autres sont sélectionnés avec probabilité proportionnelle à une mesure composite de taille. Les unités sélectionnées restent dans l'échantillon pendant les quatre années entre les recensements. Les strates d'échantillonnage sont définies en utilisant le code d'industrie à six chiffres du Système de classification des industries de l'Amérique du Nord.

Le programme de l'ASM estime les totaux au moyen d'un estimateur par différence (Särndal et coll., 1992). Afin de réduire le fardeau de réponse, les unités qui se trouvent sous un certain seuil sont éliminées complètement de la base de sondage. En lieu et place, leurs données sont imputées en utilisant les valeurs de données administratives pour certains éléments de données, et des modèles de régression au niveau de l'industrie pour les éléments restants. De même, des enregistrements complets sont imputés pour les cas de non-réponse totale. Consulter http://www.census.gov/manufacturing/asm/ pour des renseignements supplémentaires sur la méthodologie de l'ASM.

Comme les éléments de données recueillis au moyen du questionnaire de l'ASM représentent un sous-ensemble des éléments de données recueillis pour le secteur manufacturier durant l'EC, l'ASM est souvent utilisée pour l'essai de nouvelles procédures de traitement ou de collecte des données de l'EC. Dans le cas

de l'ASM et de l'EC, la mise en œuvre d'un sous-échantillon probabiliste de non-répondants pour le SCNR représente un changement de procédure important. Les procédures de SCNR de l'ASM sont très semblables à celles de l'EC. Comme une entreprise donnée peut être constituée de plusieurs établissements, l'ensemble d'établissements correspondant à une entreprise à unités multiples (UM) peut être désigné pour un suivi téléphonique, ainsi que pour d'autres vérifications de la complétude des données pour l'entreprise. Par contre, les procédures de SCNR diffèrent pour les établissements d'entreprise à unité unique (UU), c'est-à-dire les établissements dont l'entreprise mère ne compte qu'un seul emplacement. Les établissements UU les plus grands sont inclus avec certitude (échantillonnés avec une probabilité = 1) et peuvent recevoir un appel téléphonique personnel dans les domaines sélectionnés. Les établissements UU échantillonnés (« établissements UU sélectionnés sans certitude ») reçoivent certains rappels, mais sont très peu susceptibles de recevoir un appel téléphonique personnel.

Notre étude en simulation porte sur l'un des 14 éléments de données clés de l'ASM et fait appel à l'estimation par double expansion et aux deux estimateurs par le ratio décrits en annexe, mais non à l'estimateur par différence utilisé pour produire les estimations de l'ASM. Par conséquent, nos résultats ne doivent pas être extrapolés à l'ASM.

3.1 Plan de l'étude en simulation

Notre étude en simulation a pour objet de comparer les propriétés statistiques des estimations des livraisons totales obtenues selon les trois plans de sous-échantillonnage des non-répondants pris en considération sur des échantillons répétés, en utilisant trois estimateurs différents. Notre *base de sondage des non-répondants* est établie à partir de l'échantillon de l'ASM de 2011 entièrement imputé et est limitée aux établissements UU sélectionnés sans certitude de sorte que soient maintenues les exigences globales de fiabilité des données publiées de l'ASM. Dans les estimateurs par le ratio, nous employons les valeurs fondées sur échantillon de la paye annuelle comme variable auxiliaire. Cette variable est fortement corrélée aux livraisons totales, mais peut être imputée. Notons que nous utilisons l'échantillon *complet* de l'ASM (tous les établissements UM et UU) pour les répartitions, mais nous présentons les résultats du biais relatif et de l'EQM pour les domaines sous-échantillonnés (établissements UU sélectionnés sans certitude) uniquement.

Pour les établissements UU sélectionnés sans certitude, la première tentative de SCNR – consistant en une lettre de rappel – est habituellement très efficace, de sorte que la sélection du sous-échantillon de non-répondants a lieu avant la deuxième tentative de SCNR. Celle-ci est généralement plus coûteuse (habituellement, un nouvel envoi du questionnaire, quoique l'on utilisera l'envoi de lettres de rappel par courrier certifié pour les futurs cycles de collecte). Le sous-échantillonnage des non-répondants pour les établissements UU sélectionnés sans certitude a lieu *après* la deuxième tentative de prise de contact (c'est-à-dire après la première tentative de SCNR).

Pour exécuter la simulation, nous avons supprimé tous les établissements UM et tous les établissements UU sélectionnés avec certitude des données de l'échantillon de l'ASM pour créer une base de sondage, puis nous avons répété indépendamment la procédure qui suit 5 000 fois pour chaque procédure de répartition :

1. En utilisant les propensions à répondre estimées données au tableau 3.1, induire aléatoirement une non-réponse dans l'échantillon en utilisant un mécanisme de réponse MAR.
2. Trier les non-répondants induits selon le poids d'échantillonnage.

3. Tirer un échantillon systématique stratifié en utilisant les taux de sous-échantillonnage des non-répondants des domaines pour une stratégie de répartition donnée.
4. Simuler la réponse totale pour chaque cycle de SCNR. Le tableau 3.1 donne les propensions à répondre conditionnelles utilisées pour chaque phase distincte de prise de contact du SCNR. Ces statistiques se fondent sur des paradonnées provenant des cycles de collecte de 2010 et de 2011 de l'ASM (Fink et Lineback, 2013). Dans la suite de l'exposé, nous désignons ces probabilités conditionnelles par « taux de conversion des non-répondants ». Si l'unité a répondu, le mode de réponse est affecté aléatoirement en utilisant les fréquences historiques fournies par les spécialistes du domaine. Après affectation du statut de réponse/mode de réponse à chaque unité, calculer le coût de collecte cumulé, le TRT et les estimations.
5. Pour chaque répartition, répéter l'étape 4 jusqu'à ce que dix cycles de suivi soient réalisés ou que le budget total soit dépensé. Si le budget est épuisé en un cycle, le SCNR s'arrête. Étant donné que le budget fixe suppose qu'une fraction de $1/K$ de l'ensemble original de non-répondants fera l'objet d'un SCNR, le budget peut être épuisé en cas de suivi complet. En cas de sous-échantillonnage des non-répondants, le budget total n'est jamais dépensé avant que dix cycles de SCNR soient réalisés, car le coût par unité de l'envoi d'une lettre de rappel par la poste est assez faible. Le choix d'un maximum de dix cycles de SCNR dans la simulation était subjectif; l'objectif était de montrer que le sous-échantillonnage faciliterait des efforts supplémentaires de prise de contact sans coût additionnel.

Tableau 3.1
Taux de conversion des non-répondants pour les établissements de type unité unique sélectionnés sans certitude selon le cycle de SCNR utilisé pour la simulation

Domaine	Probabilité de réponse initiale	Taux de conversion des non-répondants pour un cycle donné de suivi des cas de non-réponse									
		1	2	3	4	5	6	7	8	9	10
1	0,31	0,27	0,15	0,17	0,24	0,12	0,06	0,03	0,03	0,03	0,03
2	0,44	0,32	0,24	0,15	0,36	0,18	0,09	0,05	0,05	0,05	0,05
3	0,39	0,28	0,24	0,18	0,11	0,06	0,03	0,02	0,02	0,02	0,02
4	0,35	0,36	0,17	0,19	0,18	0,09	0,05	0,02	0,02	0,02	0,02
5	0,25	0,19	0,13	0,10	0,17	0,09	0,04	0,02	0,02	0,02	0,02
6	0,27	0,13	0,29	0,02	0,02	0,02	0,02	0,02	0,02	0,02	0,02
7	0,44	0,34	0,23	0,20	0,25	0,13	0,06	0,03	0,03	0,03	0,03
8	0,38	0,45	0,12	0,33	0,25	0,13	0,06	0,03	0,03	0,03	0,03
9	0,39	0,30	0,23	0,13	0,25	0,13	0,06	0,03	0,03	0,03	0,03
10	0,75	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
11	0,28	0,23	0,12	0,18	0,15	0,07	0,04	0,02	0,02	0,02	0,02
12	0,36	0,30	0,21	0,15	0,31	0,16	0,08	0,04	0,04	0,04	0,04
13	0,39	0,22	0,19	0,13	0,23	0,12	0,06	0,03	0,03	0,03	0,03
14	0,37	0,36	0,16	0,06	0,45	0,22	0,11	0,06	0,06	0,06	0,06
15	0,41	0,32	0,22	0,19	0,26	0,13	0,06	0,03	0,03	0,03	0,03
16	0,40	0,34	0,22	0,23	0,32	0,16	0,08	0,04	0,04	0,04	0,04
17	0,34	0,26	0,18	0,10	0,21	0,11	0,05	0,03	0,03	0,03	0,03
18	0,40	0,31	0,18	0,10	0,18	0,09	0,04	0,02	0,02	0,02	0,02
19	0,37	0,29	0,20	0,19	0,23	0,11	0,06	0,03	0,03	0,03	0,03
20	0,40	0,28	0,21	0,15	0,18	0,09	0,04	0,02	0,02	0,02	0,02
21	0,36	0,27	0,20	0,14	0,23	0,11	0,06	0,03	0,03	0,03	0,03

Dans la majorité des domaines, les taux de conversion des non-répondants suivent la même tendance, à savoir une probabilité de réponse décroissante, suivie par une légère augmentation durant le quatrième cycle attribuable à une plus longue période de collecte. Le taux ne suit pas cette tendance pour le domaine 10; celui-ci ne contenait que quatre unités qui ont toutes répondu avant que débute le sous-échantillonnage. Après le quatrième cycle de SCNR, les taux de conversion des non-répondants sont réduits de moitié, jusqu'à ce qu'ils atteignent la valeur permise minimale de 0,02. La tendance reflète les constatations d'Olson et Groves (2012) (ces derniers postulent que les propensions à répondre évoluent au cours du cycle de collecte, surtout à mesure que sont modifiés les protocoles de collecte des données. Dans le cas de l'ASM, les lettres de rappel deviennent plus strictes à chaque phase de prise de contact du SCNR. De même, les auteurs montrent que les propensions à répondre diminuent au cours de la phase de collecte quand un protocole de collecte de données stable est utilisé, comme en témoignent les taux de conversion des non-répondants). Les estimations des propensions à répondre par la poste et par téléphone ont été fournies par les spécialistes du domaine, de même que les coûts approximatifs par mode de collecte et un chiffre budgétaire global.

Pour évaluer les propriétés statistiques de chaque méthode de répartition pour chaque estimateur, nous avons calculé le biais relatif et l'erreur quadratique moyenne. Le biais relatif (BR) pour chaque estimation des livraisons totales à la phase t de SCNR pour un intervalle global d'échantillonnage (K), une méthode de répartition a (Constant$-K$, Min$-K$, Min$-$TRT), une probabilité de réponse éventuelle q, et un estimateur e (DE, SR, CR) donnés est

$$\mathrm{BR}\,(Y)^{e}_{Kaqt} = 100 * \left[\left(\frac{\sum_{s=1}^{5000} \hat{Y}^{e}_{Kaqts}}{5000} \Big/ Y \right) - 1 \right]$$

où $\hat{Y}^{e}_{Kaqts}$ est le total estimé et Y est la valeur des livraisons totales dans la population.

L'erreur quadratique moyenne à la phase t du SCNR pour un intervalle d'échantillonnage, une méthode de répartition et un estimateur donnés est

$$\mathrm{EQM}\,(Y)^{e}_{Kaqt} = \left[\sum_{s=1}^{5000} \left(\hat{Y}^{e}_{Kaqts} - Y \right)^{2} \right] \Big/ 5000 .$$

Puisque notre simulation induit une réponse MAR, les estimations DE devraient être approximativement sans biais sur des échantillons répétés, tandis que les deux estimations par le ratio ne devraient pas l'être. Cependant, les estimations DE auront, en principe, une grande variance; l'utilisation des estimateurs par le ratio avec une variable auxiliaire positivement corrélée devrait réduire cette variance (c'est-à-dire accroître la précision). Donc, l'examen de l'EQM renseigne sur le compromis entre le biais et la variance.

3.2 Répartition

L'étude en simulation est réalisée sur des données provenant du cycle de collecte de 2011 de l'ASM. Les paramètres en entrée pour la répartition ont été estimés d'après les données du cycle de collecte de 2010 de l'ASM. Rappelons que le TRT cible s'applique à l'ensemble du programme de l'ASM et n'est pas limité aux domaines de sous-échantillonnage – dans notre cas, les établissements UU sélectionnés sans certitude. Par conséquent, les nombres d'unités UU et UM sélectionnées avec certitude obtenus d'après les données

de l'ASM de 2010 sont incluses dans les programmes de répartition dans les r_{1h} en tant que constantes; le reste des r_{1h} représente le nombre estimé d'établissements UU sélectionnés sans certitude répondant après l'achèvement du premier cycle de SCNR. Pour être certains que chaque domaine d'échantillonnage des non-répondants contenait un nombre suffisant d'unités pour obtenir une solution faisable, nous avons utilisé le code d'industrie à trois chiffres comme domaine d'échantillonnage pour le SCNR au lieu du code d'industrie à six chiffres utilisé pour le plan de sondage de l'ASM. [Nota : Le domaine de sous-échantillonnage a été déterminé en collaboration avec les gestionnaires du programme de l'ASM et les méthodologistes.]

Les programmes quadratiques requièrent tous deux une estimation de la probabilité de réponse éventuelle au suivi (q_h) pour calculer le $\mathrm{TRT}^{\mathrm{C}}$ (global et par domaine). Pour évaluer la sensibilité de la procédure de répartition, nous avons testé dix valeurs constantes différentes $(q_h = 0{,}10; 0{,}20; \ldots; 1)$, en gardant la valeur constante sur l'ensemble des domaines. Une approche similaire peut être adoptée quand on ne dispose pas de paradonnées historiques. En outre, nous estimons les probabilités q_h directement d'après les données de l'ASM de 2010. Ces estimations varient de 20 % au niveau du code d'industrie à trois chiffres. Cependant, la médiane de ces estimations est presque de 50 %. Par conséquent, la répartition obtenue en utilisant les valeurs estimées (données historiques) de q_h est très semblable à celle obtenue avec $q = 0{,}50$.

Environ 21 000 $ étaient affectés au SCNR des établissements UU sélectionnés sans certitude après le sous-échantillonnage. Dans le cas d'un suivi complet, le taux de réponse totale final attendu était d'environ 79 %. En utilisant des données provenant de l'EC de 2007, Bechtel et Thompson (2013) ont constaté que les taux de réponse totale cibles au niveau de l'industrie de 70 % ne pouvaient être obtenus qu'avec un sous-échantillon de 1 unité sur 3 si le taux de réponse totale moyen dans la majorité des industries visées par l'EC était égal ou supérieur à 60 % *avant* le début du suivi. Dans le cas de l'ASM, le taux de réponse avant le sous-échantillonnage était de 57 % environ. Nous avons préféré sélectionner un sous-échantillon global de 1 unité sur 2, ce qui épargnerait environ 50 % du budget octroyé après cinq cycles achevés de SCNR, au prix d'une diminution du taux de réponse prévu (69 %). Les cinq cycles supplémentaires de SCNR ont ajouté environ 4 000 $ au coût total, sans accroissement proportionnel du taux de réponse (70 %). Un grand sous-échantillon serait préférable pour ce qui est de la qualité, mais n'est pas rentable.

Pour la répartition, nous obtenons le $\mathrm{TRT}^{\mathrm{C}}$, en permettant aux q_h de varier par domaine. Le TRT maximal est toujours atteint avec le programme quadratique $\mathrm{Min{-}TRT}$. Le tableau 3.2 donne les TRT cibles et les taux de sous-échantillonnage pour la répartition obtenue au moyen du programme quadratique $\mathrm{Min{-}TRT}$. Un tiret (-) indique qu'aucun sous-échantillon n'est sélectionné pour le SCNR (un intervalle d'échantillonnage égal à ∞). Si $K = 1$, toutes les unités présentes dans le domaine sont sélectionnées pour le SCNR (suivi complet). Une étiquette $q =$ <valeur> indique que la probabilité éventuelle de réponse est la même valeur constante dans tous les domaines; les valeurs estimées d'après les données historiques sont étiquetées $q_h =$ Est. Rappelons que le $\mathrm{TRT}^{\mathrm{C}}$ englobe toutes les unités répondantes dans l'échantillon de l'ASM, et non simplement les unités uniques sélectionnées sans certitude qui sont admissibles pour le sous-échantillonnage. Par conséquent, les domaines sélectionnés ont atteint leurs TRT cibles *avant* le sous-échantillonnage et ne sont pas considérés comme des candidats au sous-échantillonnage dans les programmes de répartition.

À mesure qu'augmente la probabilité de réponse éventuelle, cette répartition a tendance à sélectionner de plus petits sous-échantillons dans des nombres croissants de domaines. Quand la probabilité d'une réponse éventuelle (q_h) est faible (20 % ou moins), les répartitions tendent judicieusement vers l'absence d'un sous-échantillonnage ou suivi complet, en se concentrant sur l'obtention d'un échantillon à partir des quelques domaines où les taux de réponse sont les plus mauvais. À mesure que la probabilité de réponse éventuelle augmente, la quantité de sous-échantillonnage a tendance à augmenter également. À 70 %, près de la moitié des domaines se voient attribuer au moins une unité échantillonnée, ce qui étale l'échantillon visé par la répartition sur plusieurs domaines au lieu de se concentrer sur quelques domaines qui ont des taux de réponse exceptionnellement bas. Notons que des taux inférieurs à 20 % sont (espérons-le) irréalistes, de même que des taux supérieurs à 70 %. Le domaine 10 présente quoi qu'il en soit des taux d'échantillonnage très variables; comme les quatre unités ont répondu avant le sous-échantillonnage, le programme quadratique a sélectionné *n'importe quel* taux d'échantillonnage, parce que, en fait, il sous-échantillonne toujours zéro cas.

Tableau 3.2
Répartitions Min−TRT (intervalles d'échantillonnage) (niveau du programme $K = 2$)

	Min−TRT										
Domaine	**q = 10**	**q = 20**	**q = 30**	**q = 40**	**q = 50**	**q = 60**	**q = 70**	**q = 80**	**q = 90**	**q = 100**	**q_h = Est**
1	-	-	-	-	-	-	-	81,63	9,23	5,40	-
2	-	-	-	-	-	-	3,88	2,26	1,71	1,44	-
3	-	-	9,32	3,40	2,58	2,19	1,98	1,86	1,77	1,71	2,12
4	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,06	1,13	1,00
5	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	-
7	-	-	-	-	-	-	-	14,95	9,26	7,10	-
8	-	-	-	-	-	-	-	-	-	-	-
9	1,00	1,00	1,22	1,44	1,61	1,76	1,89	2,01	2,12	2,22	1,62
10	1,03	30,26	30,37	30,26	30,46	29,90	30,51	29,04	1,00	10,03	10,04
11	-	-	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-	-	-
13	-	-	-	-	5,00	2,94	2,29	2,01	1,88	1,78	2,91
14	-	-	-	-	-	-	-	-	-	-	-
15	7,86	4,45	3,22	2,57	2,42	2,32	2,28	2,30	2,35	2,40	2,38
16	1,00	1,35	1,46	1,46	1,49	1,51	1,53	1,57	1,62	1,66	1,66
17	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
18	-	-	-	-	-	-	-	-	-	37,95	-
19	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
20	1,00	1,00	1,00	1,16	1,34	1,49	1,63	1,75	1,87	1,97	1,35
21	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
TRT[c]	72,5 %	72,9 %	73,3 %	73,7 %	74,1 %	74,4 %	74,8 %	75,2 %	75,6 %	76,0 %	74,3 %

Contrairement au programme quadratique Min−TRT, le programme quadratique Min−K n'a pas toujours obtenu une solution pour un TRT cible donné, en raison des contraintes au niveau du domaine sur

les TRT cibles. Le cas échéant, nous avons réduit progressivement le taux de réponse cible jusqu'à ce qu'il soit possible d'obtenir une solution pratique. Le tableau 3.3 donne les TRT cibles et les répartitions obtenues au moyen du programme quadratique Min$-K$.

Les deux méthodes de répartition ont tendance à désigner les mêmes domaines pour l'absence de sous-échantillonnage ou pour le suivi complet. Cependant, les deux méthodes produisent des répartitions très différentes pour les *mêmes* q_h dans les domaines *sous-échantillonnés*. Les répartitions Min$-K$ évitent le sous-échantillonnage dans les domaines où est déjà atteint le TRT cible estimé maximal, quelle que soit la probabilité d'obtention éventuelle d'une réponse, avec de 40 % à 50 % des domaines non sous-échantillonnés quand $0{,}30 \leq q_h \leq 0{,}50$. Sinon, le sous-échantillonnage a tendance à être divisé entre le suivi complet (toutes les unités sélectionnées) et le sous-échantillonnage à un taux d'environ 1 unité sur 2. Brièvement, les répartitions Min$-$TRT donnent des intervalles de sous-échantillonnage des domaines qui peuvent différer considérablement de l'intervalle global, car la répartition vise à égaliser le TRT cible dans chaque domaine. La variabilité résultante des intervalles d'échantillonnage peut aboutir à de grandes augmentations de la variance d'échantillonnage. Comme la fonction objectif de Min$-K$ vise à égaliser les intervalles d'échantillonnage, les intervalles de sous-échantillonnage de domaine ont tendance à être moins variables et sont généralement proches de l'intervalle d'échantillonnage global.

Tableau 3.3
Répartitions Min$-K$ (intervalles d'échantillonnage) (niveau de programme $K = 2$)

	Min$-K$ (cible $K = 2$)										
Domaine	q = **10**	q = **20**	q = **30**	q = **40**	q = **50**	q = **60**	q = **70**	q = **80**	q = **90**	q = **100**	q = **Est**
1	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	2,00	2,00	-
3	-	-	-	-	1,99	2,00	2,00	2,00	2,00	2,01	1,99
4	1,00	1,00	1,00	1,00	1,00	1,00	1,01	1,10	1,18	1,26	1,00
5	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
7	-	-	-	-	-	-	-	-	-	2,06	-
8	-	-	-	-	-	-	-	-	-	-	-
9	1,00	1,32	1,44	1,72	1,90	1,99	1,97	1,96	1,96	2,09	1,90
10	-	-	-	-	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-	-	-
13	-	-	-	-	-	1,99	1,99	1,98	1,98	2,04	-
14	-	-	-	-	-	-	-	-	-	-	-
15	2,52	2,23	2,36	1,90	1,76	1,97	1,92	1,90	1,90	2,27	1,76
16	2,17	2,08	1,71	1,83	1,90	1,97	1,97	1,96	1,96	2,09	1,90
17	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
18	-	-	-	-	-	-	-	-	-	-	-
19	-	2,01	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
20	1,00	1,00	1,11	1,36	1,57	1,75	1,90	1,97	1,97	2,06	1,59
21	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
TRT[c]	71,0 %	71,4 %	72,3 %	72,7 %	73,1 %	73,4 %	73,8 %	74,2 %	74,6 %	75,0 %	73,3 %

3.3 Résultats

Notre base de référence reproduit fidèlement les procédures de SCNR utilisées pour le SCNR de l'ASM de 2012, à savoir quatre phases de suivi complet $(K = 1)$, mais peut comprendre un cinquième cycle de suivi incomplet supplémentaire quand le budget prévu n'a pas été épuisé, afin de maintenir la cohérence de programmation. Pour d'autres valeurs de K, le SCNR est achevé après dix cycles, indépendamment des fonds restants.

Le tableau 3.4 donne le biais relatif (BR) des estimations et l'erreur quadratique moyenne (EQM) obtenus pour le SCNR complet et la répartition Constant$-K$ pour chaque estimateur pris en considération. Dans tous les cas, l'estimateur par double expansion donne des estimations sans biais, tandis que les estimateurs par le ratio sont légèrement biaisés comme prévu. Avec le sous-échantillonnage, le biais relatif des estimateurs par le ratio augmente, tandis que l'estimateur DE demeure sans biais. Quel que soit l'estimateur, l'étape supplémentaire de sous-échantillonnage accroît la variance d'échantillonnage et, par conséquent, l'EQM; le biais a tendance à ne pas être affecté, parce que les unités sous-échantillonnées forment un sous-échantillon représentatif à chaque cycle de suivi.

Dans le cas du sous-échantillonnage avec probabilités égales (Constant$-K$), un sous-échantillon peut contenir quelques cas échantillonnés dans un ou plusieurs domaines. Même si l'ajustement de la pondération du sous-échantillonnage n'est pas variable, les facteurs d'ajustement de la non-réponse peuvent être assez grands. Les répartitions optimales sont conçues en vue d'égaliser les taux de réponse dans les divers domaines, ce qui peut donner lieu à un « suréchantillonnage » occasionnel dans les domaines à faible taux de réponse. Le tableau 3.5 donne le BR et l'EQM pour les répartitions optimales Min$-$TRT, en utilisant trois valeurs constantes différentes de q $(q = 0{,}30; 0{,}50; 0{,}70)$ et les taux propres au domaine estimés d'après des données historiques (q_h = Estimé). Dans tous les scénarios, les estimations DE sont sans biais, les estimations CR présentent un léger biais, et les estimations SR présentent le biais le plus important. Cela reproduit la tendance du BR observée pour les résultats de la répartition Constant$-K$. De plus, les estimations du BR ne semblent pas être exagérément sensibles aux valeurs de q_h utilisées dans la répartition. De nouveau, même avec les cycles supplémentaires de SCNR, le biais des estimations sur les sous-échantillons est plus grand que celui obtenu pour le suivi complet des non-répondants. Dans tous les cas, les valeurs de l'EQM des estimations obtenues d'après les répartitions optimales sont plus faibles que celles obtenues pour les répartitions Constant$-K$.

Indépendamment de l'estimateur, le biais diminue quand la probabilité de réponse éventuelle est faible. Cela semble un peu contre-intuitif, mais il s'agit en fait d'une conséquence directe de la procédure de répartition du sous-échantillon. Quand la probabilité d'obtenir une réponse éventuelle est faible, la répartition Min$-$TRT a tendance à sous-échantillonner toutes les unités dans un domaine ou à n'en échantillonner aucune. Dans le cas du suivi complet, toutes les unités répondantes figurant dans le même domaine reçoivent le même ajustement pour la non-réponse. Dans le cas d'un sous-échantillon, *seulement* les poids des unités sous-échantillonnées répondantes sont ajustés pour la non-réponse et le sous-échantillonnage, ce qui à son tour crée parfois des poids très variables à l'intérieur d'un domaine. À mesure que la probabilité d'une réponse éventuelle augmente, la répartition optimale produit l'échantillonnage d'un plus grand nombre de domaines, et des ajustements plus fins sont possibles. Cela étant dit, les estimateurs CR ont tendance à produire les EQM les plus faibles, indépendamment de la méthode de répartition.

Tableau 3.4
Sommaire du biais relatif en pourcentage de l'estimation et de l'EQM pour les répartitions Constant−*K* en *x*10[12]

Constant−*K* – Biais relatif de l'estimation						
Pourcentage	***K* = 1 (suivi complet)**			***K* = 2**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	0,01 %	0,03 %	0,10 %	0,00 %	0,51 %	1,43 %
3	0,00 %	0,03 %	0,08 %	-0,01 %	0,29 %	0,77 %
4	0,00 %	0,01 %	0,06 %	-0,02 %	0,14 %	0,40 %
5	0,01 %	0,02 %	0,04 %	-0,01 %	0,12 %	0,32 %
6				-0,01 %	0,11 %	0,29 %
7				0,00 %	0,11 %	0,28 %
8				0,00 %	0,11 %	0,27 %
9				0,00 %	0,10 %	0,25 %
10				0,00 %	0,10 %	0,25 %
Constant−*K* – Erreur quadratique moyenne						
x10^12	***K* = 1 (suivi complet)**			***K* = 2**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	4,96	2,60	5,56	37,53	26,34	70,49
3	3,67	1,96	4,17	19,82	13,80	28,88
4	2,55	1,39	3,03	11,75	8,30	14,87
5	2,48	1,39	2,87	9,94	7,10	12,12
6				9,36	6,75	11,16
7				9,09	6,63	10,63
8				8,80	6,48	10,23
9				8,51	6,32	9,95
10				8,27	6,19	9,74

Tableau 3.5
Sommaire du biais relatif de l'estimation et de l'EQM pour les répartitions optimales Min−TRT

Min TRT - BR (cible *K* = 2)												
Pourcentage	***q* = 0,30**			***q* = 0,50**			***q* = 0,70**			***q* = Estimé**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	-0,01 %	0,06 %	0,20 %	0,01 %	0,07 %	0,36 %	0,01 %	0,08 %	0,31 %	-0,01 %	0,08 %	0,32 %
3	0,00 %	0,05 %	0,16 %	0,01 %	0,05 %	0,26 %	0,01 %	0,07 %	0,23 %	0,01 %	0,07 %	0,23 %
4	0,00 %	0,05 %	0,14 %	0,01 %	0,05 %	0,18 %	0,01 %	0,06 %	0,19 %	0,01 %	0,06 %	0,17 %
5	0,00 %	0,05 %	0,14 %	0,01 %	0,05 %	0,18 %	0,01 %	0,05 %	0,17 %	0,00 %	0,06 %	0,15 %
6	0,01 %	0,05 %	0,13 %	0,02 %	0,05 %	0,17 %	0,01 %	0,05 %	0,17 %	0,00 %	0,06 %	0,15 %
7	0,01 %	0,05 %	0,13 %	0,01 %	0,05 %	0,17 %	0,01 %	0,05 %	0,16 %	0,00 %	0,06 %	0,15 %
8	0,01 %	0,05 %	0,13 %	0,01 %	0,05 %	0,16 %	0,01 %	0,05 %	0,16 %	0,00 %	0,05 %	0,14 %
9	0,01 %	0,05 %	0,13 %	0,01 %	0,05 %	0,16 %	0,01 %	0,05 %	0,16 %	0,00 %	0,05 %	0,14 %
10	0,00 %	0,05 %	0,13 %	0,01 %	0,05 %	0,15 %	0,01 %	0,05 %	0,15 %	0,00 %	0,05 %	0,14 %
Min TRT − EQM (cible *K* = 2)												
x10^12	***q* = 0,30**			***q* = 0,50**			***q* = 0,70**			***q* = Estimé**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	12,55	6,77	16,03	14,35	7,48	17,60	14,31	7,44	17,15	14,39	7,79	17,05
3	8,88	5,13	10,87	9,80	5,43	11,32	9,57	5,42	11,36	9,75	5,47	10,98
4	7,00	4,28	8,15	7,61	4,45	8,28	7,31	4,43	8,44	7,53	4,44	8,05
5	6,61	4,07	7,41	7,02	4,17	7,44	6,80	4,13	7,60	6,94	4,14	7,42
6	6,45	3,97	7,16	6,78	4,09	7,18	6,62	4,03	7,25	6,75	4,05	7,15
7	6,37	3,92	7,05	6,68	4,06	7,08	6,55	3,97	7,07	6,67	4,02	7,03
8	6,34	3,90	6,94	6,57	4,01	6,97	6,45	3,93	6,95	6,57	3,98	6,93
9	6,28	3,87	6,86	6,50	3,98	6,89	6,39	3,90	6,86	6,47	3,94	6,84
10	6,23	3,85	6,78	6,40	3,91	6,76	6,35	3,87	6,75	6,42	3,89	6,73

La procédure de répartition Min−*K* est conçue pour réduire la variabilité des poids d'ajustement des unités sous-échantillonnées. Le tableau 3.6 donne le biais relatif de l'estimation et l'EQM pour la méthode de répartition optimale Min−*K*. Les estimateurs Min−*K* présentent la même tendance que précédemment. Les estimations DE sont sans biais, les estimations CR sont presque sans biais et les estimations SR présentent un léger biais.

Les estimations de l'EQM pour la méthode Min−K suivent une tendance similaire à celle observée pour la méthode Min−TRT, comme prévu étant donné les similarités entre les répartitions Min−TRT et Min−K correspondantes. Ces résultats semblent être assez peu sensibles à la probabilité de réponse éventuelle prévue (q). Les taux de conversion estimés d'après les données historiques produisent des résultats similaires à une valeur supposée $q = 0{,}50$. Dans de nombreux cas, la méthode Min−TRT produit les estimations présentant le biais le plus faible. Cependant, le biais n'est qu'une des composantes de l'EQM, et les répartitions Min−TRT ont tendance à donner un plus petit nombre prévu de répondants dans plusieurs strates que les répartitions Min−K correspondantes. En outre, les variances d'échantillonnage des répartitions Min−K sont plus petites par conception, ce qui donne en dernière analyse des estimations dont les EQM sont plus faibles que celles observées pour les répartitions Min−TRT correspondantes.

Les figures 3.1 et 3.2 donnent le tracé des valeurs du BR et de l'EQM observées à chaque cycle de SCNR pour l'estimateur CR (notre « meilleur » estimateur) en utilisant les valeurs de q_h obtenues d'après les données historiques, pour chacune des méthodes de répartition optimale considérées, ainsi que les valeurs de référence (étiquetées « suivi complet »). À la figure 3.1, les estimations de référence sont celles dont le biais est le plus faible. Cependant, ce biais très faible est en partie la conséquence de notre modèle de non-réponse, qui est uniforme à l'intérieur du domaine et de la phase de SCNR. Les estimations obtenues par les répartitions optimales n'atteignent aucune les niveaux des estimations de référence, mais elles deviennent très proches après sept cycles de SCNR, et les BR des estimations Min−TRT et Min−K sont inférieurs à *un dixième* de 1 % (0,06 % et 0,05 %, respectivement). Brièvement, le sous-échantillonnage suivant l'une ou l'autre des stratégies de répartition optimale produit une augmentation négligeable du biais par rapport au suivi complet.

Tableau 3.6
Sommaire du biais relatif de l'estimation et de l'EQM pour les répartitions optimales Min−*K*

Min−*K* - BR (cible *K* = 2)												
Pourcentage	***q* = 0,30**			***q* = 0,50**			***q* = 0,70**			***q* = Estimé**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	0,03 %	0,08 %	0,24 %	0,03 %	0,09 %	0,31 %	0,00 %	0,08 %	0,33 %	0,01 %	0,07 %	0,30 %
3	0,03 %	0,05 %	0,20 %	0,03 %	0,08 %	0,22 %	0,00 %	0,05 %	0,22 %	0,01 %	0,06 %	0,21 %
4	0,02 %	0,04 %	0,16 %	0,03 %	0,07 %	0,18 %	0,00 %	0,05 %	0,17 %	0,01 %	0,05 %	0,17 %
5	0,02 %	0,04 %	0,15 %	0,03 %	0,06 %	0,17 %	0,01 %	0,05 %	0,16 %	0,01 %	0,05 %	0,16 %
6	0,02 %	0,05 %	0,14 %	0,02 %	0,06 %	0,16 %	0,01 %	0,05 %	0,15 %	0,00 %	0,05 %	0,15 %
7	0,02 %	0,05 %	0,14 %	0,02 %	0,05 %	0,16 %	0,01 %	0,05 %	0,15 %	0,01 %	0,05 %	0,15 %
8	0,02 %	0,05 %	0,14 %	0,02 %	0,05 %	0,16 %	0,01 %	0,05 %	0,15 %	0,01 %	0,04 %	0,14 %
9	0,02 %	0,05 %	0,14 %	0,02 %	0,05 %	0,15 %	0,01 %	0,05 %	0,14 %	0,01 %	0,04 %	0,14 %
10	0,02 %	0,05 %	0,14 %	0,02 %	0,05 %	0,16 %	0,01 %	0,05 %	0,15 %	0,01 %	0,04 %	0,14 %
Min−*K* - EQM (cible *K* = 2)												
x10^12	***q* = 0,30**			***q* = 0,50**			***q* = 0,70**			***q* = Estimé**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	12,86	7,19	15,85	13,81	7,42	16,80	15,09	8,34	18,00	13,43	7,19	16,07
3	8,74	5,04	10,26	9,32	5,38	10,82	10,45	5,89	11,38	9,25	5,30	10,69
4	6,92	4,07	7,65	7,26	4,26	7,92	7,84	4,60	8,19	7,22	4,33	7,93
5	6,50	3,85	7,07	6,77	4,05	7,33	7,23	4,28	7,47	6,65	4,06	7,21
6	6,32	3,80	6,80	6,57	3,94	7,02	7,02	4,19	7,28	6,45	3,95	6,91
7	6,23	3,76	6,69	6,49	3,88	6,91	6,90	4,15	7,16	6,31	3,91	6,78
8	6,21	3,73	6,61	6,39	3,84	6,82	6,78	4,10	7,06	6,23	3,87	6,68
9	6,16	3,70	6,54	6,35	3,79	6,71	6,68	4,05	6,93	6,15	3,83	6,57
10	6,10	3,66	6,43	6,24	3,74	6,62	6,60	3,98	6,87	6,11	3,80	6,48

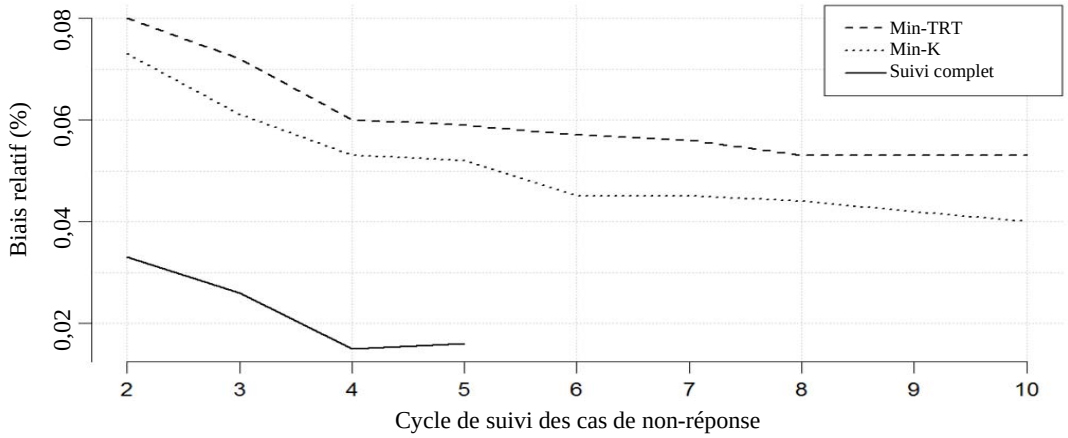

Figure 3.1 Biais relatif des estimations (q_h historique) pour l'estimateur CR.

La figure 3.2 donne le tracé des valeurs de l'EQM en fonction du cycle de SCNR lorsqu'on utilise l'estimateur CR. La stratégie d'échantillonnage ciblé de la non-réponse utilisée pour la répartition Min$-K$ semble réduire l'erreur globale. Nous pensons que cela tient à deux facteurs. Premièrement, sous la procédure de répartition Min$-K$, de plus grandes proportions de non-répondants sont échantillonnées dans les domaines à faible réponse que sous les répartitions Min$-$TRT. Deuxièmement, la formule quadratique pour la répartition Min$-K$ contient une contrainte sur les taux de réponse de domaine, ce qui fait baisser la réponse cible globale, mais réduit la variabilité de la proportion de répondants par domaine. En fin de compte, cette approche donne des taux de réponse similaires dans les divers domaines d'échantillonnage, ce qui témoigne d'un échantillon représentatif (Wagner, 2012; Schouten, Cobben et Bethlehem, 2009). Notons que l'accroissement de l'EQM observé pour le sous-échantillonnage des non-répondants n'est pas négligeable, même si l'on utilise une procédure d'ajustement qui bénéficie d'une forte covariable dans la procédure d'ajustement par le ratio. Il s'agit d'un prix reconnu du sous-échantillonnage des non-répondants (Biemer, 2010). Cependant, cette composante de variance additionnelle est mesurable. Si la composante mesurée est trop grande, les gestionnaires du programme peuvent réduire le sous-échantillonnage (utiliser un K plus grand).

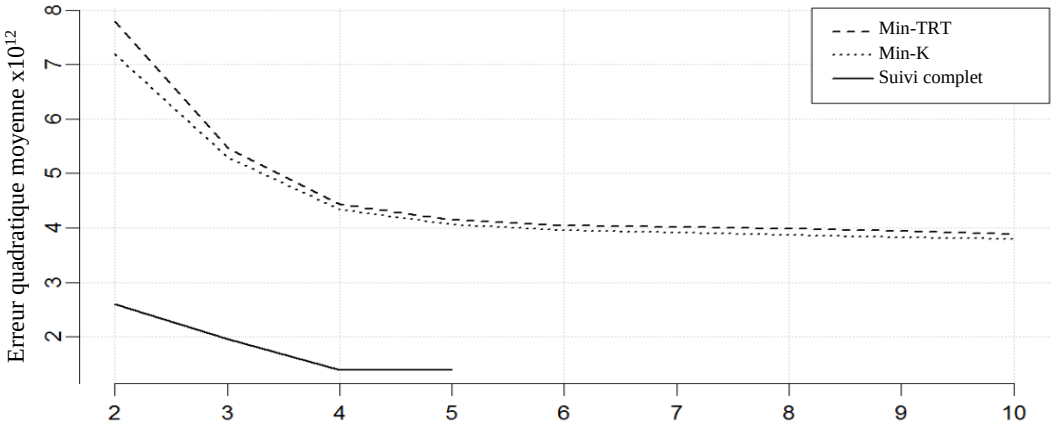

Figure 3.2 Erreur quadratique moyenne (q_h historique) pour l'estimateur CR.

3.4 Discussion

Étant donné une méthode de répartition sophistiquée, un estimateur par le ratio employant une variable auxiliaire fortement corrélée, et un sous-échantillon assez grand, la présente étude de cas montre que le sous-échantillonnage des non-répondants ne pénalise pas indûment la qualité au profit du coût. L'étape supplémentaire d'échantillonnage a augmenté l'EQM pour la variable étudiée, mais le niveau a été réduit par le choix judicieux de l'estimateur. Naturellement, nous ne considérons qu'une seule variable dans notre simulation, et cette variable pourrait ou non « se comporter » de la même façon que d'autres variables d'enquête. Un examinateur a proposé d'utiliser un indicateur R (Schouten et coll., 2009) ou un indicateur d'équilibre (Särndal et Lundquist, 2014) pour évaluer la représentativité globale de l'ensemble de répondants dans les conditions d'une enquête sur le terrain. Cela pourrait être utile à des étapes ultérieures de la collecte des données (après le sous-échantillonnage des non-répondants et durant le SCNR), mais ne fournirait aucun autre renseignement sur le degré de réduction du biais pour tout élément de donnée recueilli, comme nous pouvons le faire dans les présentes conditions de simulation.

Des trois méthodes de répartition prises en considération, la méthode $\text{Constant}-K$ affichait la pire performance, donnant souvent lieu à la sélection d'un très petit sous-échantillon probabiliste quand cela n'était pas nécessaire et augmentant conséquemment la variance d'échantillonnage sans réduire le biais. Des trois méthodes de répartition considérées, la répartition $\text{Min}-K$ était la plus efficace pour ce qui est de réaliser des taux de réponse acceptables et de produire des estimations fiables; le biais le plus important causé par les intervalles d'échantillonnage de domaine variables est généralement compensé par la variance d'échantillonnage réduite. Toutefois, la répartition $\text{Min}-K$ peut être plus difficile à mettre en œuvre que la répartition $\text{Min}-\text{TRT}$.

Pour les deux procédures de répartition optimale, nous avons testé quatre probabilités de réponse éventuelle différentes afin d'évaluer la sensibilité des procédures de répartition à ces entrées. En comparant les répartitions obtenues avec une valeur en entrée supposée constante à celles obtenues en utilisant des estimations empiriques, nous avons constaté que les répartitions réalisées pouvaient comprendre un sous- ou un suréchantillonnage dans certains domaines, et que les taux de réponse de domaines pouvaient varier plus que prévu quand les valeurs réelles (d'enquête) étaient assez différentes des valeurs en entrée. Par conséquent, nous recommandons d'utiliser des valeurs estimées d'après des paradonnées historiques dans la mesure du possible.

Si l'objectif global est de réduire le coût, nous notons alors que les tentatives supplémentaires de prise de contact du SCNR au-delà du cinquième contact n'amélioraient ni le biais ni l'EQM des estimations avec sous-échantillonnage dans notre étude de cas. Naturellement, si l'économie réalisée pour un sous-échantillon de 1 unité sur 2 avec jusqu'à dix tentatives de prise de contact durant le SCNR est acceptable, les fonds affectés à ces tentatives finales de prise de contact pourraient être dépensés à meilleur escient à un stade plus précoce des cycles de collecte en utilisant d'autres stratégies de prise de contact.

4 Conclusion

En général, les procédures de SCNR pour les programmes économiques menés par le *U.S. Census Bureau* suivent un calendrier d'exécution. Le budget est lié à l'exercice, et les stratégies de prise de contact

sont budgétées en conséquence. Puisque les populations économiques sont très asymétriques et que les statistiques d'intérêt sont les totaux, une fraction importante du budget de SCNR est affectée aux grandes unités. Les unités plus petites sont considérées comme étant homogènes, du moins en taille. Cependant, il est difficile de valider cette opinion en l'absence de données recueillies auprès des répondants. Étant donné que les procédures de SCNR dépendent de l'obtention de données de réponse auprès des grandes unités, les taux de réponse des unités plus petites ont tendance à être nettement plus faibles. Il est fort probable que l'ensemble réalisé de répondants ne soit pas équilibré, c'est-à-dire un échantillon sélectionné possédant des caractéristiques identiques ou presque identiques à celles de l'ensemble de la population pour certains éléments de données (Särndal, 2011), ni représentatif en ce qui concerne l'échantillon si les propensions à répondre ρ_i sont les mêmes pour toutes les unités dans la population (Schouten et coll., 2009). L'accent mis sur l'obtention de réponses auprès des grandes unités au prix de la réponse des unités plus petites crée, à son tour, un biais dans les estimations, car les valeurs imputées ou ajustées pour les petites unités ressemblent aux valeurs des grandes unités (Thompson et Washington, 2013).

En limitant le domaine cible pour le sous-échantillonnage des non-répondants aux unités plus petites, nous pouvons réduire ce biais non mesurable. Notre méthode de répartition accroît la possibilité d'obtenir un échantillon équilibré et représentatif en ciblant les domaines à faible réponse qui, normalement, ne feraient l'objet d'aucun traitement spécial. Elle peut être mise en œuvre à n'importe quelle étape du processus de collecte des données et sous n'importe quel plan de sondage, ce qui la rend assez flexible, quoique pas nécessairement optimale pour des plans de sondage et des estimateurs particuliers. Il s'agit d'une approche « sûre » pour une enquête polyvalente, vraisemblablement conçue en vue d'obtenir des estimations fiables pour une gamme d'éléments. En outre, la sélection d'un sous-échantillon systématique à partir d'une liste triée selon une mesure de taille d'unité évite l'existence d'un biais de non-réponse supplémentaire résultant de la concentration des efforts de SCNR sur les cas à propension à répondre élevée (Tourangeau et coll., 2016; Beaumont et coll., 2014). Nous reconnaissons que l'accroissement de la variabilité des poids de sondage et la réduction des taux de réponse sont des effets moins que désirables du sous-échantillonnage. Cependant, ces effets peuvent être atténués grâce au choix de l'estimateur, comme le montre l'amélioration de nos résultats en utilisant un estimateur par le ratio. Des estimateurs par calage plus évolués ou d'autres estimateurs regroupés (*collapsed estimators*) pourraient aussi être pris en considération à l'étape de l'estimation.

Sans sous-échantillonnage probabiliste, la thèse selon laquelle l'ensemble réalisé de petites entreprises répondantes demeure un échantillon probabiliste est discutable. Plusieurs discussions du rapport sommaire du groupe de travail sur l'échantillonnage non probabiliste de l'AAPOR (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau, 2013) soulèvent spécifiquement la question de savoir si un échantillon probabiliste avec couverture incomplète et non-réponse élevée devrait encore être considéré comme un échantillon probabiliste. Cette question est certainement pertinente dans le contexte de notre étude, où les plus petites unités échantillonnées « optent » vraiment de répondre. Sélectionner un sous-échantillon probabiliste de non-répondants et indiquer aux analystes de l'enquête de restreindre la prise de contact du SCNR à ces cas pourraient limiter ce phénomène. En outre, pour un sous-échantillon probabiliste, on peut utiliser des mesures de qualité reconnues, telles que l'erreur d'échantillonnage ou les taux de réponse pour l'évaluation.

Tous les résultats présentés pour notre étude de cas reposent sur l'hypothèse que les stratégies de prise de contact du SCNR existantes sont utilisées avec les plans de sous-échantillonnage. Cependant, le sous-échantillonnage des non-répondants sans modifier la procédure de collecte des données pourrait n'avoir que des effets concrets minimes outre la réduction du coût. L'inverse est également vrai : par exemple, Kirgis et Lepkowski (2013) présentent des résultats témoignant d'une amélioration des données de réponse pour de petits domaines cibles obtenues avec des échantillons probabilistes et des stratégies de prise de contact révisées.

Tourangeau et coll. (2016) notent qu'il n'est pas toujours évident de savoir comment intervenir pour obtenir que les cas, particulièrement ceux dont la propension à répondre sous-jacente est faible, répondent. Cette remarque est particulièrement pertinente dans le contexte des enquêtes auprès des entreprises. Ces enquêtes peuvent s'appuyer sur une abondance d'études cognitives sur les stratégies de collecte des données pour les grandes entreprises (voir Paxson, Dillman et Tarnai, 1995; Tuttle, Morrison et Willimack, 2010; Willimack et Nichols, 2010; Snijkers, Haraldsen, Jones et Willimack, 2013). En revanche, les entreprises plus petites ne font l'objet que de très peu de prises de contact personnelles (voire aucune) et la recherche cognitive sur laquelle s'appuyer pour choisir les stratégies de prise de contact préférables est limitée. Cela dit, la littérature donne à penser qu'il existe des différences de qualité entre les données recueillies auprès des grandes et des petites entreprises; voir Thompson et Washington (2013), Willimack et Nichols (2010), Bavdaž (2010), Torres van Grinsven, Bolko et Bavdaž (2014), et Thompson, Oliver et Beck (2015). D'autres études cognitives visant les petits établissements conjuguées à des essais sur le terrain pourraient aboutir à de meilleures stratégies de prise de contact. Le sous-échantillonnage des non-répondants associé à une nouvelle stratégie de prise de contact pour ces établissements « difficiles à joindre » créerait une approche vraiment adaptative pour toutes les unités, plutôt que simplement les grandes. À cet égard, en réponse aux analyses que nous avons présentées, le *Census Bureau* a réalisé une expérience sur le terrain intégrée pour tester diverses stratégies de SCNR pour certaines petites unités durant l'ASM de 2014 (Thompson et Kaputa, 2017). Cette étude a abouti à la mise en œuvre d'un nouveau protocole de SCNR pour l'ASM de 2015 et à une deuxième expérience sur le terrain intégrée qui associait notre plan proposé de sous-échantillonnage des non-répondants avec les procédures de suivi les plus efficaces cernées durant l'essai de 2014 (Kaputa, Thompson et Beck, 2017).

Remerciements

Les opinions exprimées sont celles des auteurs et ne reflètent pas nécessairement celles du *U.S. Census Bureau*. Les auteurs remercient Eric Fink, Xijian Liu, Jared Martin, Edward Watkins III, Hannah Thaw, le rédacteur associé et deux examinateurs pour leur révision d'une version antérieure du manuscrit, David Haziza pour sa discussion constructive de l'article, et Barry Schouten pour ses suggestions utiles concernant les problèmes d'optimisation.

Annexe

Notre objectif est d'estimer Y, le total de population de la caractéristique y, à partir de l'échantillon réalisé de répondants. Soit

S_{hi} = 1 si l'unité i dans le domaine h était dans l'échantillon original; 0 sinon.

θ_{hi} = la probabilité d'échantillonner l'unité i dans le domaine h dans l'échantillon original $(w_{hi} = 1/\theta_{hi})$.

R_{hi} = 1 si l'unité i dans le domaine h a fourni une réponse avant le moment t du sous-échantillonnage (valeur pour y); 0 sinon.

I_{hi} = 1 si l'unité i dans le domaine h a été sélectionnée pour le SCNR (c'est-à-dire était un non-répondant sous-échantillonné); 0 sinon.

J_{hi} = 1 si l'unité i dans le domaine h répond, sachant qu'elle est sélectionnée dans le sous-échantillon de non-répondants; 0 sinon.

f_{hi} = facteur d'ajustement pour le sous-échantillonnage des non-répondants et la non-réponse totale après le SCNR.

y_{hi} = valeur de la caractéristique y pour l'unité i dans le domaine h, disponible seulement pour les répondants.

x_{hi} = valeur de la caractéristique x pour l'unité i dans le domaine h, disponible pour toutes les unités échantillonnées prises en considération pour le sous-échantillonnage des non-répondants (c'est-à-dire la base de sous-échantillonnage des non-répondants). Alors $\hat{Y} = \sum_h \sum_i w_{hi} y_{hi} S_{hi} R_{hi} + \sum_h \sum_i w_{hi}\ f_{hi} y_{hi}\ S_{hi}\ (1 - R_{hi})\, I_{hi} J_{hi} = \hat{Y}_{R1} + \hat{Y}_{R2}$.

Nous considérons trois estimateurs par repondération (ajustement des poids d'échantillonnage) de $\hat{Y}_{R2}$:

Double Expansion (DE) : $\hat{Y}_{R2}^{\mathrm{DE}} = \sum_h \sum_{i \in h} w_{hi} K_h \left(\frac{m_{1h}}{r_{2h}}\right) y_{hi} S_{hi}\ (1 - R_{hi})\, I_{hi} J_{hi}$

Ratio distinct (SR) : $\hat{Y}_{R2}^{\mathrm{SR}} = \sum_h \sum_{i \in h} w_{hi} K_h \left(\frac{\sum_{i \in m_{1h}} x_{hi}}{\sum_{i \in r_{2h}} x_{hi}}\right) y_{hi} S_{hi}\ (1 - R_{hi})\, I_{hi} J_{hi}$

Ratio combiné (CR) : $\hat{Y}_{R2}^{\mathrm{CR}} = \sum_h \sum_{i \in h} w_{hi} K_h \left(\frac{m_{1h}}{r_{2h}}\right) \left(\frac{\sum_{i \in m_{1h}} w_{hi} K_h x_{hi}}{\sum_{i \in r_{2h}} w_{hi} K_h \left(\frac{m_{1h}}{r_{2h}}\right) x_{hi}}\right) y_{hi} S_{hi}\ (1 - R_{hi})\, I_{hi} J_{hi}.$

Notons que les estimateurs DE et CR sont des variations de la procédure de repondération recommandée décrite dans Brick (2013) et sont discutés dans Binder et coll. (2000), entre autres. L'estimateur DE est l'estimateur InfoS présenté dans Särndal et Lundström (2005), étudié dans Shao et Thompson (2009), entre autres; l'estimateur SR est une variation de l'estimateur InfoP présenté dans Särndal et Lundström (2005), qui traitent l'échantillon réalisé comme étant la « population ». Les poids d'échantillonnage n'ont pas été inclus dans l'estimateur SR de sorte que l'ajustement se réduit à l'ajustement DE quand $x_{hi} \equiv 1 \forall\, i \in h$; notons que cet ajustement du taux de réponse non pondéré est recommandé dans Little et Vartivarian (2005). L'estimateur CR est présenté dans Binder et coll. (2000), et est également étudié dans Shao et Thompson (2009). Dans notre étude de cas, un meilleur choix aurait pu être l'estimateur par quasi-randomisation de

Oh et Scheuren (1983), qui intègre les poids d'échantillonnage dans le facteur d'ajustement, et réduit leur variabilité.

Des estimateurs regroupés (*collapsed estimators*) sont utilisés dans trois scénarios : 1) toutes les unités dans le domaine font l'objet d'un SCNR (pas de sous-échantillonnage); 2) aucune unité dans le domaine ne fait l'objet d'un SCNR parce que les taux de réponse cibles ont été atteints (pas de sous-échantillonnage); 3) une seule unité sous-échantillonnée a répondu au SCNR (sous-échantillonnage). Les estimateurs regroupés analogues sont exprimés comme il suit :

DE regroupé : $$\hat{Y}_h^{\mathrm{DE,C}} = \sum_{i \in h} w_{hi} \left(\frac{n_h}{r_{1h} + r_{2h}} \right) y_{hi} S_{hi} R_{hi}$$

SR regroupé : $$\hat{Y}_h^{\mathrm{SR,C}} = \sum_{i \in h} w_{hi} \left(\frac{\sum_{i \in n_h} x_{hi}}{\sum_{i \in r_{1h}+r_{2h}} x_{hi}} \right) y_{hi} S_{hi} \left(1 - R_{hi}\right) I_{hi} J_{hi}$$

CR regroupé : $$\hat{Y}_h^{\mathrm{CR,C}} = \sum_{i \in h} w_{hi} \left(\frac{n_h}{r_{1h} + r_{2h}} \right) \left(\frac{\sum_{i \in n_h} w_{hi} x_{hi}}{\sum_{i \in r_{1h}+r_{2h}} w_{hi} \left(\frac{n_h}{r_{1h} + r_{2h}} \right) x_{hi}} \right) y_{hi} S_{hi} R_{hi}.$$

Bibliographie

Baker, R., Brick, J.M., Bates, N., Battaglia, M., Couper, M., Dever, J., Gile, K. et Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling – Report and rejoinder. *Journal of Survey Statistics and Methodology*, 1, 90-137.

Bavdaž, M. (2010). Modèle multidimensionnel intégral de réponse aux enquêtes auprès des entreprises. *Techniques d'enquête*, 36, 1, 89-102. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2010001/article/11245-fra.pdf.

Beaumont, J.-F., Bocci, C. et Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30(4), 607-621.

Bechtel, L., et Thompson, K.J. (2013). Optimizing unit nonresponse adjustment procedures after subsampling nonrespondents in the Economic Census. *Proceedings of the Federal Committee on Statistical Methods Research Conference*, https://nces.ed.gov/FCSM/index.asp.

Biemer, P. (2010). Total survey error: Design, implementation, and evaluation. *The Public Opinion Quarterly*, 74(5), 817-848.

Binder, D., Babyak, C., Brodeur, M., Hidiroglou, M. et Wisner, J. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 751-764.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329-353.

Federal Register Notice (2006). OMB Standards and Guidelines for Statistical Surveys, Washington, DC.

Fink, E., et Lineback, J.F. (2013). Using paradata to understand business survey reporting patterns. *Proceedings of the Federal Committee on Statistical Methods Research Conference*, https://nces.ed.gov/FCSM/index.asp.

Groves, R., et Herringa, S. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A*, 169(3), 439-57.

Hansen, M.H., et Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Harter, R.M., Mach, T.L., Chaplin, J.F. et Wolken, J.D. (2007). Determining subsampling rates for nonrespondents. *Proceedings of the Third International Conference on Establishment Surveys*, American Statistical Association.

Haziza, D., Thompson, K.J. et Yung, W. (2010). L'effet des ajustements pour la non-réponse sur l'estimation de la variance. *Techniques d'enquête*, 36, 1, 39-48. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2010001/article/11246-fra.pdf.

Kalton, G., et Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 2, 81-97.

Kaputa, S., Thompson, K.J. et Beck, J. (2017). An embedded experiment for targeted nonresponse follow-up in establishment surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Kirgis, N., et Lepkowski, J. (2013). Design and management strategies for paradata-driven responsive design: Illustrations for the 2006-2010 National Survey of Family Growth. *Improving Surveys with Paradata*, (Éd., Frauke Kreuter). Hoboken, NJ: John Wiley & Sons, Inc.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8(2), 183-200.

Kott, P. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89, 693-696.

Little, R.J., et Vartivarian, S. (2005). La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage ? *Techniques d'enquête*, 31, 2, 175-183. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2005002/article/9046-fra.pdf.

Lohr, S.L. (2010). *Sampling: Design and Analysis*. Boston: Brooks/Cole.

Oh, H.L., et Scheuren, F.J. (1983). Weighting adjustment of unit nonresponse. *Incomplete Data in Sample Surveys*. New York: Academic Press, 20, 143-184.

Olson, K., et Groves, R.M. (2012). An examination of within-person variation in response propensity over the data collection field period. *Journal of Official Statistics*, 28, 29-51.

Paxson, M.C., Dillman, D.A. et Tarnai, J. (1995). Improving response to business mail surveys. Dans *Business Survey Methods*, (Éds., B.G. Cox, D. Binder, B. Nanajamma Chinnappa, M. Colledge et P. Kott). New York: John Wiley & Sons, Inc.

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

Särndal, C., et Lundquist, P. (2014). Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, 2(4), 361-387.

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

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

Schouten, B., Calinescu, M. et Luiten, A. (2013). Optimiser la qualité de la réponse au moyen de plans de collecte adaptatifs. *Techniques d'enquête*, 39, 1, 33-66. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11824-fra.pdf.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 107-121. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10887-fra.pdf.

Shao, J., et Thompson, K.J. (2009). Estimation de la variance en présence de non-répondants et de strates à tirage complet. *Techniques d'enquête*, 35, 2, 235-246. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2009002/article/11043-fra.pdf.

Snijkers, G., Haraldsen, G., Jones, J. et Willimack, D.K. (2013). *Designing and Conducting Business Surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

Thompson, K.J., et Kaputa, S. (2017). Investigating adaptive nonresponse follow-up strategies for small businesses through embedded experiments. *Journal of Official Statistics*, 33(3), 1-23.

Thompson, K.J., et Oliver, B. (2012). Response rates in business surveys: Going beyond the usual performance measure. *Journal of Official Statistics*, 27, 221-237.

Thompson, K.J., Oliver, B. et Beck, J. (2015). An analysis of the mixed collection modes for two business surveys conducted by the US Census Bureau. *Public Opinion Quarterly*, 79(3), 769-789.

Thompson, K.J., et Washington, K.T. (2013). Challenges in the treatment of unit nonresponse for selected business surveys: A case study. *Survey Methods: Insights from the Field*. Récupéré de http://surveyinsights.org/?p=2991.

Torres van Grinsven, V., Bolko, I. et Bavdaž, M. (2014). In search of motivation for the business survey response task. *Journal of Official Statistics*, 30(4), 579-606.

Tourangeau, R., Brick, J.M., Lohr, S. et Li, J. (2016). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society A*, 180, 203-223.

Tuttle, A., Morrison, R. et Willimack, D. (2010). From start to pilot: A multi-method approach to the comprehensive redesign of an economic survey questionnaire. *Journal of Official Statistics*, 26, 87-103.

Wagner, J. (2012). Research synthesis: A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, 76(3), 555-575.

Willimack, D., et Nichols, E. (2010). A hybrid response process model for business surveys. *Journal of Official Statistics*, 26, 3-24.

Zhang, L.C. (2008). On some common practices of systematic sampling. *Journal of Official Statistics*, 24, 557-569.

Techniques d'enquête, juin 2018
Vol. 44, nº 1, pp. 109-124
Statistique Canada, nº 12-001-X au catalogue

Estimation bayésienne robuste sur petits domaines

Malay Ghosh, Jiyoun Myung et Fernando A.S. Moura[1]

Résumé

Les modèles pour petits domaines conçus pour traiter les données au niveau du domaine reposent habituellement sur l'hypothèse de normalité des effets aléatoires. Cette hypothèse ne tient pas toujours. L'article présente un nouveau modèle pour petits domaines dont les effets aléatoires suivent une loi *t*. En outre, la modélisation conjointe des moyennes et des variances de petit domaine est examinée. Il est montré que cette approche donne de meilleurs résultats que les autres méthodes.

Mots-clés : Modèle à effets aléatoires; loi *t* de Student; lois a priori non subjectives; MCMC; échantillonnage de Gibbs; algorithme de Metropolis-Hastings.

1 Introduction

Depuis près de quatre décennies, l'article classique de Fay et Herriot (1979) est devenu une pierre angulaire de la recherche au sujet de l'estimation sur petits domaines. Le modèle de Fay-Herriot est un modèle à effets aléatoires qui s'appuie sur une hypothèse de normalité des effets aléatoires, ainsi que des erreurs. En outre, les variances d'erreur sont supposées connues. Cette dernière hypothèse est presque impérative en raison d'un problème d'identifiabilité. Si l'on dispose uniquement d'estimations sur petits domaines directes au niveau du domaine et que l'on ne possède pas de microdonnées, toute modélisation efficace des variances d'erreur est quasiment impossible.

De vaillantes tentatives en vue de remédier à ce problème ont été faites par W.R. Bell et ses collègues au *US Census Bureau* (Bell et Huang, 2006; Bell, 2008) pour le traitement de certaines données de recensement, mais des questions persistent quant à l'application universelle de leur approche. En outre, l'absence de microdonnées pour les utilisateurs secondaires tient surtout à des problèmes de confidentialité, principalement pour les organismes fédéraux. Quand des microdonnées deviennent disponibles, les modèles au niveau de l'unité conviennent mieux que ceux au niveau du domaine. L'article souvent cité de Battese, Harter et Fuller (1988) en est un exemple classique. Toutefois, les modèles au niveau du domaine sont d'usage très répandu en raison de la simplicité de leur mise en œuvre dans un contexte d'enquête complexe comparativement aux modèles au niveau de l'unité.

À mesure que s'est développée la branche de l'estimation sur petits domaines et que davantage de données ont été analysées, les chercheurs ont constaté que les hypothèses de normalité et de connaissance des variances d'erreur étaient inappropriées. Comme il est mentionné au paragraphe précédent, la dernière hypothèse est difficile à corriger sans aucune information supplémentaire. L'une des premières tentatives à cet égard est attribuable à Lahiri et Rao (1995) qui ont remplacé l'hypothèse de normalité des effets aléatoires par la finitude de leur huitième moment. Datta et Lahiri (1995) ont considéré pour les effets aléatoires un mélange général de lois normales incluant la loi t. Certains articles font entièrement abstraction de la normalité, mais maintiennent la linéarité du modèle, et font appel à des estimateurs

1. Malay Ghosh, Department of Statistics, University of Florida, 102 Griffin-Floyd Hall, P.O. Box 118545, Gainesville, Floride, États-Unis, 32611. Courriel : ghoshm@stat.ufl.edu; Jiyoun Myung, Department of Statistics et Applied Probability, University of California, Santa Barbara, 5513 South Hall, Santa Barbara, Californie, États-Unis, 93106. Courriel : myung@pstat.ucsb.edu; Fernando A.S. Moura, Instituto de Matemática, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, CEP: 21941-909, RJ, Brésil. Courriel : fmoura@im.ufrj.br.

ANOVA des variances. On peut citer Butar et Lahiri (2002), ainsi que Jiang, Lahiri et Wan (2002) qui ont calculé les mesures d'incertitude correspondantes par les méthodes du jackknife ou du bootstrap. Bell et Huang (2006) ont utilisé des lois t pour les effets aléatoires ou les erreurs d'échantillonnage afin de diminuer les effets des valeurs aberrantes.

L'objectif du présent article est d'aborder ces deux questions importantes dans le contexte de l'estimation sur petits domaines. En premier lieu, nous considérons la modélisation des moyennes ainsi que des variances de population pour des petits domaines. Cette modélisation est possible grâce à l'existence de données additionnelles destinées à estimer les variances d'erreur. En deuxième lieu, afin d'induire une certaine robustesse dans notre procédure, nous considérons des lois a priori t pour les effets aléatoires.

L'ensemble de données utilisé dans le présent article provenait d'un essai de recensement démographique réalisé dans une municipalité brésilienne comptant 140 districts de recensement, que nous appellerons petits domaines. La variable étudiée était le revenu moyen des chefs de ménage pour chaque petit domaine, et l'objectif était de faire des prédictions pour les 140 moyennes de population des revenus des chefs de ménage. Les variables auxiliaires étaient les moyennes de population de petit domaine respectives du niveau d'études des chefs de ménage, et les moyennes de population respectives du nombre de pièces dans les logements des ménages pour chaque petit domaine. Seules des données au niveau du domaine nous ont été fournies.

Nous proposons une analyse bayésienne non subjective complète du problème général d'estimation sur petits domaines, où nous modélisons les moyennes ainsi que les variances de population. Au départ, l'idée était d'utiliser la loi a priori découlant de la règle générale de Jeffreys, en traitant tous les paramètres, y compris le nombre de degrés de liberté de la loi t de Student, comme étant inconnus. Cependant, la loi a priori résultante donnait une loi a posteriori impropre, qui a mené à une loi a priori de Jeffreys modifiée résultant en une loi a posteriori propre.

La présentation des autres sections de l'article est la suivante. À la section 2, nous décrivons le modèle, la matrice d'information de Fisher, ainsi que la loi a priori de Jeffreys et sa modification. L'impropriété de la loi a posteriori sous la loi a priori de Jeffreys et sa propriété sous sa version modifiée sont également traitées dans cette section. À la section 3, nous présentons une analyse de données réelles, ainsi qu'une étude en simulation. Enfin, à la section 4, nous formulons certaines observations finales.

Le caractère vraiment aléatoire des variances d'erreur est reconnu depuis longtemps. Les travaux d'Otto et Bell (1995), d'Arora et Lahiri (1997), de Wang et Fuller (2003), de Rivest et Vandal (2003) et d'autres ont tenté de tenir compte de ce fait de différentes façons. Slud et Maiti (2006), Dass, Maiti, Ren et Sinha (2012) et Maiti, Ren et Sinha (2014) ont adopté une approche bayésienne empirique à cette fin en estimant les hyperparamètres. Une analyse entièrement bayésienne en utilisant des méthodes bayésiennes hiérarchiques avec normalité des effets au niveau du domaine a été examinée dans You et Chapman (2006) et Sugasawa, Tamae et Kubokawa (2017). Au moyen d'une analyse des données et de simulations, nous montrerons que les lois a priori t pour les effets aléatoires donnent souvent de meilleurs résultats que les méthodes présentées dans ces deux derniers articles.

L'utilisation de lois a priori t pour les erreurs dans les modèles de régression normaux centrés réduits, mais non dans les modèles à effets mixtes, a été proposée dans Lange, Little et Taylor (1989), Fernandez et

Steel (1998), Vrontos, Dellaportas et Politis (2000), Jacquier, Polson et Rossi (2004), et Fonseca, Ferreira et Migon (2008), principalement en guise de protection contre les valeurs aberrantes. Toutefois, dans certaines situations, la normalité des erreurs est une hypothèse raisonnable, en raison du théorème central limite. En outre, il existe des techniques de diagnostic de modèle pour le vérifier. Malheureusement, l'hypothèse de normalité des effets aléatoires ne fonctionne pas toujours bien. Pour les données brésiliennes dont nous disposons, la modélisation conjointe des moyennes ainsi que des variances d'échantillon ainsi que des lois a priori t pour les effets aléatoires donne de meilleurs résultats que certains autres modèles au niveau du domaine. Comme l'a suggéré un examinateur, nous avons utilisé les données pour calculer les résidus en ajustant une régression avec des moyennes de domaine et des covariables réelles afin d'étudier la distribution des effets aléatoires pour cette application. Voir la section 3.

2 Le modèle

Un modèle au niveau du domaine type est donné par $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + u_i + e_i$, $(i = 1, \ldots, m)$, où m désigne le nombre de petits domaines, $\mathbf{x}_1, \ldots, \mathbf{x}_m$ est le vecteur de covariables de dimension $p\,(< m)$, et $\boldsymbol{\beta}\,(p \times 1)$ est le vecteur de coefficients de régression. Les distributions des effets aléatoires u_i et des erreurs d'échantillonnage e_i sont supposées indépendantes avec $u_i \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$ et $e_i \overset{\text{ind}}{\sim} N(0, v_i)$. Autrement dit, le modèle au niveau du domaine classique est

$$y_i \,|\, \theta_i \overset{\text{ind}}{\sim} N(\theta_i, v_i),$$

$$\theta_i \,|\, \boldsymbol{\beta}, \sigma_u^2 \overset{\text{ind}}{\sim} N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma_u^2), \quad i = 1, \ldots, m.$$

Les v_i sont supposées connues afin d'éviter la non-identifiabilité. L'hypothèse que les v_i sont connues devient presque obligatoire pour les utilisateurs secondaires des données qui n'ont accès à aucunes microdonnées pour les modéliser. Cependant, en réalité les variances sont aléatoires, basées sur des données échantillonnées. Dans les situations où l'on dispose de données additionnelles pour modéliser les v_i, ces dernières peuvent être estimées efficacement. En outre, dans de telles situations, il est possible d'obtenir des estimateurs à rétrécissement des moyennes de petit domaine θ_i, ainsi que des variances v_i.

Nous nous penchons sur les problèmes d'estimation sur petits domaines pour lesquels nous disposons de données additionnelles pour modéliser les v_i. En outre, pour la robustification, nous supposons que les effets aléatoires suivent une loi t plutôt qu'une loi normale. Nous énonçons notre modèle comme il suit,

$$y_i \,|\, \theta_i, v_i \overset{\text{ind}}{\sim} N(\theta_i, v_i),\ s_i^2 \,|\, v_i \overset{\text{ind}}{\sim} G\left(\frac{n_i - 1}{2}, \frac{1}{2v_i}\right)$$

$$\theta_i \,|\, \boldsymbol{\beta}, \sigma_\delta^2, \nu \overset{\text{ind}}{\sim} t_\nu(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma_\delta), \quad i = 1, \ldots, m, \tag{2.1}$$

où n_i est la taille de l'échantillon dans le i^{e} domaine, $t_\nu(\mu, \sigma)$ désigne la loi t de Student avec paramètres de position μ, d'échelle σ et de degrés de liberté ν, et $G(c, d)$ désigne la loi gamma avec la densité estimée par noyau $x^{c-1}\exp(-dx)$ pour $x > 0$.

Pour une analyse bayésienne complète, notre objectif est de trouver la loi a posteriori de $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$, sachant $\mathbf{y} = (y_1, \ldots, y_m)^T$ et $\mathbf{s}^2 = (s_1^2, \ldots, s_m^2)^T$. Pour cela, nous devons d'abord

trouver les lois a priori de tous les hyperparamètres, $\boldsymbol{\beta}$, $\mathbf{v} = (v_1, \ldots, v_m)^T$, σ_δ^2, et ν. Le premier essai habituel est la loi a priori de Jeffreys qui est proportionnelle à la racine carrée positive du déterminant de la matrice d'information de Fisher. Dans notre cas, cette matrice est

$$\mathbf{I}(\boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu) = \begin{bmatrix} \dfrac{(\nu+1)}{\sigma_\delta^2(\nu+3)}\mathbf{X}^T\mathbf{X} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dfrac{m\nu}{2(\sigma_\delta^2)^2(\nu+3)} & \dfrac{-m}{\sigma_\delta^2(\nu+1)(\nu+3)} \\ \mathbf{0} & \mathbf{0} & \dfrac{-m}{\sigma_\delta^2(\nu+1)(\nu+3)} & mg(\nu) \end{bmatrix},$$

où $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^T$, $\mathbf{D} = \mathrm{Diag}\left(\frac{n_1}{2v_1^2}, \ldots, \frac{n_m}{2v_m^2}\right)$ et $g(\nu) = \{\Psi'(\frac{\nu}{2}) - \Psi'(\frac{\nu+1}{2})\}/4 - (\nu+5)/\{2\nu(\nu+1)(\nu+3)\}$, avec $\Psi(z) = \Gamma'(z)/\Gamma(z)$ et $\Psi'(z) = d\Psi(z)/dz$ qui sont des fonctions digamma et trigamma. Donc, la loi a priori de Jeffreys est

$$\boldsymbol{\pi}_J(\boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu) \propto (\sigma_\delta^2)^{-\frac{p}{2}-1} |\mathbf{X}^T\mathbf{X}|^{\frac{1}{2}} |\mathbf{D}|^{\frac{1}{2}} \left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}} \left[\frac{\nu g(\nu)}{2(\nu+3)} - \frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}}. \qquad (2.2)$$

Cependant, la loi a priori de Jeffreys mène à une loi a posteriori impropre en raison du facteur $(\sigma_\delta^2)^{-\frac{p}{2}-1}$ dans (2.2).

Théorème 1. *La loi a priori de Jeffreys (2.2) mène à une loi a posteriori impropre.*

Preuve. Soit $\boldsymbol{\pi}_J \equiv \boldsymbol{\pi}_J(\boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu \,|\, \mathbf{y}, \mathbf{s}^2)$ la densité a posteriori avec la loi a priori de Jeffreys (2.2). En considérant les termes qui contiennent σ_δ^2 dans $\boldsymbol{\pi}_J$ et en prenant la transformation $w_i = (\theta_i - \mathbf{x}_i^T\boldsymbol{\beta})/\sigma_\delta$, c'est-à-dire $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma_\delta w_i$, nous avons

$$\int_0^\infty (\sigma_\delta^2)^{-\frac{p}{2}-1} \exp\left[-\frac{1}{2}\sum_{i=1}^m \frac{1}{v_i}(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \sigma_\delta w_i)^2\right] d\sigma_\delta^2$$

$$\geq \int_0^\infty (\sigma_\delta^2)^{-\frac{p}{2}-1} \exp\left[-\sum_{i=1}^m \left\{\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{v_i} + \frac{\sigma_\delta^2 w_i^2}{v_i}\right\}\right] d\sigma_\delta^2$$

$$\geq \exp\left\{-\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{v_i}\right\} \int_0^k (\sigma_\delta^2)^{-\frac{p}{2}-1} \exp\left(-\sum_{i=1}^m \frac{\sigma_\delta^2 w_i^2}{v_i}\right) d\sigma_\delta^2 \text{ pour une constante } k > 0,$$

$$\geq \exp\left[-\sum_{i=1}^m \left\{\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{v_i} + \frac{k w_i^2}{v_i}\right\}\right] \int_0^k (\sigma_\delta^2)^{-\frac{p}{2}-1} d\sigma_\delta^2 = \infty.$$

Donc, la loi a priori de Jeffreys mène à une loi a posteriori impropre.

Cependant, une fois que la composante $(\sigma_\delta^2)^{-\frac{p}{2}-1}$ dans (2.2) est remplacée par $(\sigma_\delta^2)^{-\frac{p}{2}-1}\exp(-a/2\sigma_\delta^2)$ pour une valeur $a > 0$, cette version modifiée de la loi a priori de Jeffreys mène à une loi a posteriori propre sous la contrainte $\min(n_1, \ldots, n_m) > p$. Par conséquent, nous proposons pour notre modèle une loi a priori de Jeffreys modifiée comme il suit :

$$\boldsymbol{\pi}_{\mathrm{MJ}}(\boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu) \propto (\sigma_\delta^2)^{-\frac{p}{2}-1}\exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^{m}\frac{1}{v_i}\right)\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}$$

$$\times\left[\frac{\nu g(\nu)}{2(\nu+3)} - \frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}} \text{ où } a > 0. \tag{2.3}$$

En combinant la vraisemblance de (2.1) et la loi a priori de Jeffreys modifiée (2.3), la loi a posteriori complète des paramètres, sachant les données, est

$$\boldsymbol{\pi}_{\mathrm{MJ}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu \,|\, \mathbf{y}, \mathbf{s}^2) \propto (\sigma_\delta^2)^{-\frac{p+m}{2}-1}\exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^{m} v_i^{-\frac{n_i}{2}-1}\right)\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\frac{1}{v_i}(y_i-\theta_i)^2\right]$$

$$\times\exp\left(-\frac{1}{2}\sum_{i=1}^{m}\frac{s_i^2}{v_i}\right)\left[\prod_{i=1}^{m}\left\{1+\frac{(\theta_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{\nu\sigma_\delta^2}\right\}^{-\frac{\nu+1}{2}}\right]\left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\right]^m$$

$$\times\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{\nu g(\nu)}{2(\nu+3)} - \frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}}. \tag{2.4}$$

Théorème 2. *Sous le modèle (2.1), la loi a posteriori* $\boldsymbol{\pi}_{\mathrm{MJ}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu \,|\, \mathbf{y}, \mathbf{s}^2)$ *dans (2.4) est propre, à condition que* $\min(n_1, \ldots, n_m) > p$.

Preuve. Voir l'annexe A.

Le théorème 2 montre que la loi a priori de Jeffreys modifiée (2.3) mène à une loi a posteriori propre (2.4). L'idée essentielle est que nous avons besoin d'une loi a priori pour σ_δ^2 telle que $\int_0^\infty \boldsymbol{\pi}(\sigma_\delta^2)(\sigma_\delta^2)^{-\frac{p}{2}-1}\,d\sigma_\delta^2 < \infty$.

Remarque 1. $\boldsymbol{\pi}_{\mathrm{MJ}}(\boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu)$ *peut être factorisé en quatre lois a priori indépendantes pour chaque paramètre.*

$$\boldsymbol{\pi}_{\mathrm{MJ}}(\boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu) \propto \boldsymbol{\pi}(\boldsymbol{\beta})\boldsymbol{\pi}(\mathbf{v})\boldsymbol{\pi}(\sigma_\delta^2)\boldsymbol{\pi}(\nu)$$

où

$$\boldsymbol{\pi}(\boldsymbol{\beta}) \propto 1,\ \boldsymbol{\pi}(v_i) \propto \frac{1}{v_i} \quad \text{pour} \quad i = 1, \ldots, m,$$

$$\boldsymbol{\pi}(\sigma_\delta^2) \sim \mathrm{GI}\left(\frac{p}{2}, \frac{a}{2}\right)$$

et

$$\boldsymbol{\pi}(\nu) \propto \left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}} \left[\frac{\nu g(\nu)}{2(\nu+3)} - \frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}}.$$

Ici, $\mathrm{GI}(c, d)$ désigne la loi gamma inverse avec la densité estimée par noyau $x^{-c-1}\exp(-d/x)$ pour $x > 0$.

Les lois conditionnelles complètes pour appliquer la méthode Monte Carlo par chaîne de Markov (MCMC) sont données à l'annexe B. Pour générer les échantillons, nous utilisons l'échantillonnage de Gibbs avec l'algorithme de Metropolis-Hastings, où la loi conditionnelle d'un paramètre est connue uniquement jusqu'à une constante multiplicative. Nous expliquons de façon détaillée comment appliquer un résultat de Chib et Greenberg (1995) pour l'algorithme de Metropolis-Hastings afin de générer les échantillons.

3 Application

3.1 Analyse de données réelles

L'ensemble de données est sélectionné par échantillonnage aléatoire de 10 % des ménages dans chaque domaine provenant d'un test de recensement démographique effectué dans une municipalité au Brésil. La municipalité compte 38 740 ménages dans 140 petits domaines au total, et le nombre de ménages par domaine dans la population varie de 57 à 588. Donc, les tailles d'échantillon de domaine dans l'ensemble de données varient de 6 à 59. Nous souhaitons estimer les 140 moyennes de population des revenus des chefs de ménage. La variable de réponse y_i est le revenu moyen des chefs de ménages dans le i^{e} domaine.

Cet ensemble de données comprend deux covariables auxiliaires centrées qui sont les moyennes de population de petit domaine respectives du niveau d'études des chefs de ménages (échelle ordinale de 0 à 5) et du nombre de pièces dans les logements des ménages (1 à 11+). Enfin, l'ensemble de données contient les variances d'échantillonnage respectives qui sont calculées de la manière habituelle. Puisque nous n'avons obtenu que des données au niveau du domaine et que les moyennes de domaine réelles sont connues, nous pouvons comparer les 140 prédictions sur petits domaines aux moyennes de domaine réelles, respectivement. L'analyse laisse entendre que notre modèle donne de meilleurs résultats que les autres modèles où les effets aléatoires sont fondés sur la loi normale. Aux fins de comparaison, nous utilisons trois autres modèles.

Le premier est le modèle de Fay-Herriot, noté FH, avec variances d'échantillonnage connues.

$$y_i \mid \theta_i, v_i \overset{\text{ind.}}{\sim} N(\theta_i, v_i), \quad \theta_i \mid \boldsymbol{\beta}, \sigma_\delta^2 \overset{\text{ind.}}{\sim} N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma_\delta^2)$$

$$\pi(\boldsymbol{\beta}) \propto 1, \quad \pi(\sigma_\delta^2) \sim \mathrm{GI}(a_0, b_0),$$

où a_0 et b_0 sont choisis égaux à 0,0001 (une petite constante) pour traduire la vague connaissance de σ_δ^2.

Le deuxième modèle, proposé par You et Chapman (2006), et noté YC, est un modèle bayésien hiérarchique donné par

$$y_i \,|\, \theta_i, v_i \overset{\text{ind.}}{\sim} N(\theta_i, v_i), \quad \theta_i \,|\, \boldsymbol{\beta}, \sigma_\delta^2 \overset{\text{ind.}}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma_\delta^2)$$

$$s_i^2 \,|\, v_i \overset{\text{ind}}{\sim} G\left(\frac{n_i - 1}{2}, \frac{1}{2v_i}\right)$$

$$\pi(v_i) \sim \text{GI}(a_i, b_i), \quad \pi(\boldsymbol{\beta}) \propto 1, \quad \pi(\sigma_\delta^2) \sim \text{GI}(a_0, b_0),$$

où a_i, b_i, a_0 et b_0 sont également choisis égaux à 0,0001.

Le troisième modèle est un modèle bayésien à plusieurs degrés pour petits domaines proposé par Sugasawa et coll. (2017), désigné par STK. Le modèle STK produit une estimation à rétrécissement des moyennes ainsi que des variances.

$$y_i \,|\, \theta_i, v_i \overset{\text{ind.}}{\sim} N(\theta_i, v_i), \quad \theta_i \,|\, \boldsymbol{\beta}, \sigma_\delta^2 \overset{\text{ind.}}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma_\delta^2)$$

$$s_i^2 \,|\, v_i \overset{\text{ind}}{\sim} G\left(\frac{n_i - 1}{2}, \frac{1}{2v_i}\right), \quad v_i \,|\, \gamma \sim \text{GI}(a_i, b_i\gamma),$$

$$\pi(\boldsymbol{\beta}, \sigma_\delta^2, \gamma) = 1,$$

où $a_i = 2$ et $b_i = 1/n_i$ comme l'ont proposé les auteurs pour un choix raisonnable.

Nous comparons les moyennes de petit domaine prédites par les modèles FH, YC, STK et par notre modèle, ci-après dénommé modèle RTS. Pour l'exécution de la méthode MCMC, nous générons une chaîne d'une longueur de rodage (*burn-in*) de 50 000 et la taille d'échantillonnage de $G = 50\,000$. Les estimations des θ_i sont données par

$$\hat{\theta}_i = \frac{1}{G}\sum_{g=1}^{G}\left(\gamma_i^{(g)} y_i + \left(1 - \gamma_i^{(g)}\right)\mathbf{x}_i^T \boldsymbol{\beta}^{(g)}\right)$$

où

$$\gamma_i^{(g)} = \frac{v_i^{-1(g)}}{v_i^{-1(g)} + \sigma_\delta^{-2(g)}\eta_i^{(g)}}.$$

Les critères de comparaison sont la moyenne des écarts quadratiques (MEQ), le biais absolu moyen (BAM), le biais relatif quadratique moyen (BRQM) et le biais relatif moyen (BRM). Ils sont définis comme il suit :

$$\text{MEQ} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\theta}_i - \theta_i\right)^2, \quad \text{BAM} = \frac{1}{m}\sum_{i=1}^{m}\left|\hat{\theta}_i - \theta_i\right|, \quad \text{BRQM} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right)^2,$$

et

$$\text{BRM} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right|,$$

où $\hat{\theta}_i$ et θ_i sont les valeurs estimées et réelles, respectivement, dans le i^{e} domaine. Le tableau 3.1 donne une comparaison des quatre modèles. Rappelons que la loi a priori de σ_δ^2 est $\boldsymbol{\pi}(\sigma_\delta^2) \sim \text{GI}\left(\frac{p}{2}, \frac{a}{2}\right)$. Avec le paramètre de forme, $\frac{p}{2} = 1$, nous considérons plusieurs valeurs de a. Si nous choisissons une valeur de a proche de p, le modèle RTS est mieux ajusté que les autres sous les quatre critères. Si nous choisissons $a = 1$, le modèle RTS est celui qui donne les meilleurs résultats. Le modèle YC donne de moins bons

résultats que les trois autres modèles. Si nous choisissons une très petite valeur de a, telle que 0,01 ou 0,001, le modèle RTS est celui qui donne les pires résultats.

Tableau 3.1
Comparaison entre le modèle RTS, le modèle FH, le modèle YC et le modèle STK

Modèle	MEQ	BAM	BRQM	BRM
Modèle RTS (a = 0,0001)	57,297	6,152	0,395	0,589
Modèle RTS (a = 0,01)	16,546	2,741	0,090	0,244
Modèle RTS (a = 0,5)	3,244	1,249	0,020	0,118
Modèle RTS (a = 0,2)	4,185	1,439	0,025	0,133
Modèle RTS (a = 1)	2,745	1,164	0,019	0,113
Modèle RTS (a = 2)	3,080	1,231	0,020	0,117
Modèle RTS (a = 3)	3,079	1,229	0,020	0,117
Modèle RTS (a = 5)	2,994	1,213	0,019	0,116
Modèle RTS (a = 10)	3,377	1,278	0,020	0,119
Modèle RTS (a = 50)	2,905	1,180	0,018	0,112
Modèle RTS (a = 100)	2,799	1,154	0,018	0,109
Modèle FH	4,484	1,448	0,026	0,133
Modèle YC	4,983	1,543	0,029	0,141
Modèle STK	3,199	1,257	0,021	0,121

De plus, comme l'a suggéré un examinateur, nous calculons les résidus en ajustant une régression à l'aide des moyennes de domaine et de covariables réelles pour visualiser la distribution des effets aléatoires pour ces données réelles. La figure 3.1 montre que la distribution s'écarte de la distribution normale.

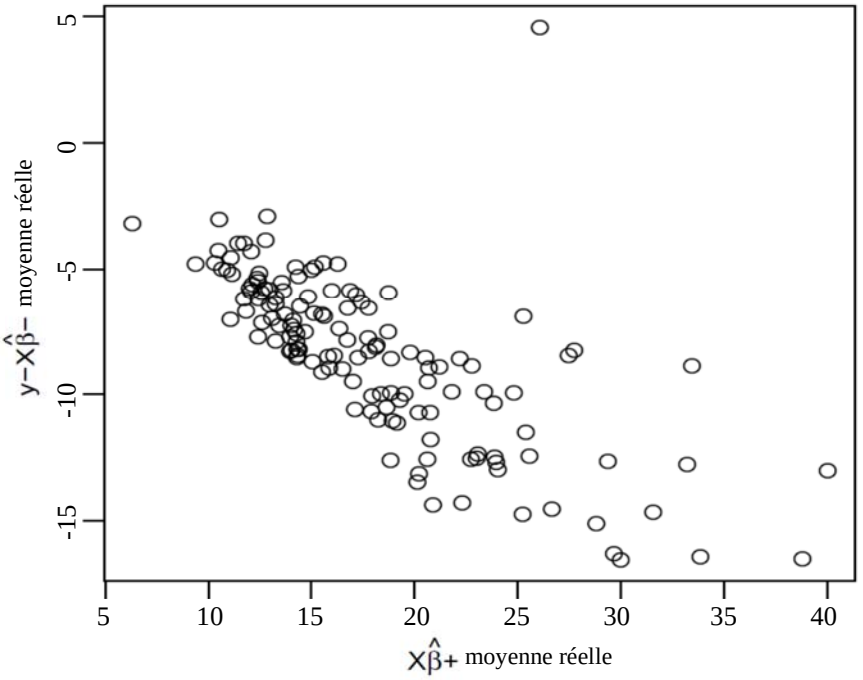

Figure 3.1 Résidus en ajustant une régression avec les moyennes et covariables réelles.

3.2 Étude en simulation

À la présente section, nous configurons une simulation proche de celle décrite dans Maiti et coll. (2014) (ou Sugasawa et coll. (2017)) pour comparer la précision de nos estimateurs à d'autres, dont ceux de You et Chapman (2006) et de Sugasawa et coll. (2017). Nous générons des observations pour chaque petit domaine à partir du modèle

$$y_{ij} = \beta_0 + \beta_1 x_i + u_i + e_{ij}, \quad j = 1, \ldots, n_i, \; i = 1, \ldots, m,$$

où $u_i \sim t_\nu(0, \sigma_\delta)$ et $e_{ij} \sim N(0, n_i v_i)$. Alors, le modèle des effets aléatoires pour la moyenne de petit domaine est

$$y_i = \beta_0 + \beta_1 x_i + u_i + e_i, \quad i = 1, \ldots, m,$$

où $y_i = \bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ et $e_i = n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$. Donc, $y_i \,|\, \theta_i \sim N(\theta_i, v_i)$, où $\theta_i = \beta_0 + \beta_1 x_i + u_i$; $\theta_i \sim t_\nu(\beta_0 + \beta_1 x_i, \sigma_\delta)$, et $e_i \sim N(0, v_i)$. Le paramètre d'intérêt est la moyenne θ_i, pour le i^{e} petit domaine. En outre, l'estimateur direct de v_i est

$$\frac{1}{n_i (n_i - 1)} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

Nous prenons $m = 30$ et $n_i = 7$ pour tous les domaines, et générons les covariables x_i à partir de la loi uniforme sur (2, 8). Les valeurs réelles des paramètres sont fixées à $\beta_0 = 0{,}5$; $\beta_1 = 0{,}8$; $\sigma_\delta = 1$; $\nu = 3$ et $v_i \sim \text{GI}(10, 5\exp(0{,}3x_i))$. En outre, nous choisissons $a = 3$ pour toutes les simulations.

Pour l'exécution de la méthode MCMC, nous avons généré 5 000 échantillons a posteriori après avoir écarté les 1 000 premiers pour $R = 2\,000$ simulations. Le tableau 3.2 donne une comparaison des quatre modèles. Les critères de comparaison sont la MEQ, le BAM et le BIAIS, ce dernier étant défini comme étant

$$\text{BIAIS} = \frac{1}{mR} \sum_{i=1}^{m} \left| \sum_{r=1}^{R} \left(\hat{\theta}_i^{(r)} - \theta_i^{(r)} \right) \right|.$$

Tableau 3.2
Résultat des simulations pour des effets aléatoires de loi t avec $\nu = 3$

Modèle	Moyenne			Variance		
	MEQ	BAM	BIAIS	MEQ	BAM	BIAIS
RTS	1,393	0,895	0,021	5,048	1,608	1,324
STK	1,821	0,933	0,025	5,042	1,514	1,367
FH	1,540	0,942	0,022			
YC	2,165	0,974	0,030	5,970	1,803	1,689

Le modèle RTS donne de meilleurs résultats que les autres sous les critères MEQ, BAM et BIAIS pour la moyenne. Tandis que le modèle RTS présente de petites améliorations par rapport aux autres modèles pour les critères BAM et BIAIS, il produit des améliorations d'environ 23,5 %, 10 % et 35,7 % par rapport

aux modèles STK, FH et YC, respectivement, pour le critère MEQ. Pour la variance, les modèles RTS et STK sont meilleurs que le modèle YC.

Les deux tableaux qui suivent donnent les résultats des simulations lorsque l'on fixe le nombre de degrés de liberté à $\nu = 2$ et $\nu = 4$.

Tableau 3.3
Résultat des simulations pour des effets aléatoires de loi t avec $\nu = 2$

Modèle	Moyenne			Variance		
	MEQ	BAM	BIAIS	MEQ	BAM	BIAIS
RTS	1,617	0,949	0,020	6,569	1,610	1,070
STK	7,566	1,107	0,038	11,144	1,441	0,996
FH	1,921	1,035	0,022			
YC	9,063	1,187	0,038	7,072	1,685	1,340

Tableau 3.4
Résultat des simulations pour des effets aléatoires de loi t avec $\nu = 4$

Modèle	Moyenne			Variance		
	MEQ	BAM	BIAIS	MEQ	BAM	BIAIS
RTS	1,265	0,862	0,019	4,876	1,619	1,428
STK	1,322	0,874	0,019	5,077	1,577	1,489
FH	1,350	0,894	0,020			
YC	1,509	0,905	0,022	6,201	1,869	1,802

Avec $\nu = 2$, le modèle RTS donne de meilleurs résultats que les autres sous les critères MEQ, BAM et BIAIS pour la moyenne. Dans cette simulation, les valeurs de MEQ pour les modèles STK et YC sont très grandes comparativement aux modèles RTS et FH. Le modèle RTS affiche des améliorations d'environ 78,6 % par rapport au modèle STK, 82,2 % par rapport au modèle YC, et 15,8 % par rapport au modèle FH. Pour les critères BAM et BIAIS, les valeurs du modèle RTS sont plus petites que celles des autres modèles. Si l'on examine la variance, le modèle RTS donne la valeur la plus petite pour MEQ.

Avec $\nu = 4$, le modèle RTS donne aussi de meilleurs résultats que les autres. En particulier, les valeurs de MEQ et de BIAIS indiquent que le modèle RTS améliore les résultats comparativement aux modèles STK et YC.

Les deux tableaux qui suivent représentent la situation lorsqu'on émet l'hypothèse de normalité des effets aléatoires. Ici, le modèle RTS donne des résultats un peu moins bons que les autres modèles.

Tableau 3.5
Résultat des simulations pour des effets aléatoires normaux avec $N(0, 5^2)$

Modèle	Moyenne			Variance		
	MEQ	BAM	BIAIS	MEQ	BAM	BIAIS
RTS	2,896	1,305	0,038	6,036	1,512	0,514
STK	2,560	1,229	0,051	1,851	0,961	0,114
FH	2,597	1,240	0,036			
YC	2,735	1,259	0,048	3,674	1,305	0,463

Tableau 3.6
Résultat des simulations pour des effets aléatoires normaux avec $N(0,10^2)$

Modèle	Moyenne			Variance		
	MEQ	BAM	BIAIS	MEQ	BAM	BIAIS
RTS	3,007	1,316	0,032	10,117	1,895	1,202
STK	2,784	1,272	0,031	2,221	1,038	0,155
FH	2,765	1,272	0,048			
YC	2,873	1,285	0,033	9,166	1,798	1,129

4 Observations finales

L'article décrit des modèles pour petits domaines en vue de traiter des données au niveau du domaine. L'élément nouveau de cet article est la modélisation à la fois des moyennes et des variances de petit domaine, parallèlement à l'utilisation d'une loi t pour les effets aléatoires. Il est montré par analyse de données et par simulation que la méthode proposée donne essentiellement de meilleurs résultats que les modèles de You et Chapman (2006) et de Sugasawa et coll. (2017) dans la plupart des situations.

Remerciements

Les travaux de recherche de Ghosh ont été financés en partie par la subvention SES-1327359 de la NSF.

Annexe A

Preuve

Théorème 2. Sous le modèle (2.1) avec la loi a priori de Jeffreys modifiée (2.3), la loi a posteriori (2.4) est propre, à condition que $\min(n_1, \ldots, n_m) > p$.

Preuve. Rappelons la loi a posteriori (2.4),

$$\boldsymbol{\pi}_{\mathrm{MJ}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu \mid \mathbf{y}, \mathbf{s}^2) \propto (\sigma_\delta^2)^{-\frac{p+m}{2}-1} \exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left[-\frac{1}{2}\sum_{i=1}^m \frac{1}{v_i}(y_i-\theta_i)^2\right]$$

$$\times \exp\left(-\frac{1}{2}\sum_{i=1}^m \frac{s_i^2}{v_i}\right)\left[\prod_{i=1}^m \left\{1+\frac{(\theta_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{\nu\sigma_\delta^2}\right\}^{-\frac{\nu+1}{2}}\right]\left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\right]^m$$

$$\times\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{\nu g(\nu)}{2(\nu+3)}-\frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}}$$

où $g(\nu) = \left\{\Psi'\left(\frac{\nu}{2}\right) - \Psi'\left(\frac{\nu+1}{2}\right)\right\}\Big/4 - (\nu+5)\Big/\{2\nu(\nu+1)(\nu+3)\}$.

Avant tout, puisque $\log\left[\Gamma\left\{\frac{(\nu+1)}{2}\right\}\Big/\Gamma\left(\frac{\nu}{2}\right)\right] \doteq \{-\log(2e) + \nu\log(\nu+1) - (\nu-1)\log\nu\}/2$ en vertu de l'approximation de Stirling, nous avons

$$\frac{1}{2}\left[\Psi\left(\frac{\nu+1}{2}\right)-\Psi\left(\frac{\nu}{2}\right)\right] \doteq \frac{d}{d\nu}\log\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} = \frac{1}{2}\left[\log(\nu+1)-\log\nu\right]+\frac{1}{2\nu(\nu+1)}$$

et

$$\frac{1}{4}\left[\Psi'\left(\frac{\nu+1}{2}\right)-\Psi'\left(\frac{\nu}{2}\right)\right] \doteq \frac{1}{2}\left(\frac{1}{\nu+1}-\frac{1}{\nu}\right)-\frac{1}{2\nu^2}+\frac{1}{2(\nu+1)^2}.$$

Cela mène à

$$g(\nu) \doteq \frac{5\nu+3}{2\nu^2(\nu+1)^2(\nu+3)}$$

et

$$\left[\frac{\nu g(\nu)}{2(\nu+3)}-\frac{1}{(\nu+1)^2(\nu+3)^2}\right] \doteq \frac{1}{4\nu(\nu+1)^2(\nu+3)}.$$

Donc, cette approximation simplifie le dernier terme dans (2.4). La loi a posteriori correspondante est

$$\begin{aligned}\boldsymbol{\pi}_{\mathrm{MJ}}(\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{v},\sigma_\delta^2,\nu\,|\,\mathbf{y},\,\mathbf{s}^2) \propto{}& (\sigma_\delta^2)^{-\frac{p+m}{2}-1}\exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left[-\frac{1}{2}\sum_{i=1}^m\frac{1}{v_i}(y_i-\theta_i)^2\right]\\ &\times\exp\left(-\frac{1}{2}\sum_{i=1}^m\frac{s_i^2}{v_i}\right)\left[\prod_{i=1}^m\left\{1+\frac{(\theta_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{\nu\sigma_\delta^2}\right\}^{-\frac{\nu+1}{2}}\right]\left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\right]^m\\ &\times\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{1}{\nu(\nu+1)^2(\nu+3)}\right]^{\frac{1}{2}}. \qquad \text{(A.1)}\end{aligned}$$

D'abord, éliminons $\boldsymbol{\beta}$ par intégration. En prenant $w_i = (\theta_i - \mathbf{x}_i^T\boldsymbol{\beta})/\sigma_\delta$, c'est-à-dire $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma_\delta w_i$, nous avons

$$\begin{aligned}\boldsymbol{\pi}_{\mathrm{MJ}}(\mathbf{w},\mathbf{v},\sigma_\delta^2,\nu\,|\,\mathbf{y},\,\mathbf{s}^2) \propto{}& (\sigma_\delta^2)^{-\frac{p}{2}-1}\exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left(-\frac{1}{2}\sum_{i=1}^m\frac{s_i^2}{v_i}\right)\\ &\times\left[\prod_{i=1}^m\left(1+\frac{w_i^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right]\left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\right]^m\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{1}{\nu(\nu+1)^2(\nu+3)}\right]^{\frac{1}{2}}\left|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right|^{-\frac{1}{2}}\\ &\times\exp\left[-\frac{1}{2}(\mathbf{Y}-\sigma_\delta\mathbf{w})^T\left\{\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\right\}(\mathbf{Y}-\sigma_\delta\mathbf{w})\right].\end{aligned}$$

où $\mathbf{w} = (w_1, \ldots, w_m)^T$. Notons que $\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$ est définie non négative. En utilisant $K(>0)$ comme une constante générique qui ne dépend que des données $(\mathbf{y}, \mathbf{s}^2)$, et en éliminant par intégration $\mathbf{w}$ et σ_δ^2 par ordre respectivement, nous obtenons

$$\boldsymbol{\pi}_{\mathrm{MJ}}(\mathbf{v}, \sigma_\delta^2, \nu \,|\, \mathbf{y}, \mathbf{s}^2) \leq K(\sigma_\delta^2)^{-\frac{p}{2}-1} \exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)$$

$$\times \exp\left(-\frac{1}{2}\sum_{i=1}^m \frac{s_i^2}{v_i}\right)\left[\frac{1}{\nu(\nu+1)^2(\nu+3)}\right]^{\frac{1}{2}} |\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{-\frac{1}{2}},$$

et puis

$$\boldsymbol{\pi}_{\mathrm{MJ}}(\mathbf{v}, \nu \,|\, \mathbf{y}, \mathbf{s}^2) \leq K\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left(-\frac{1}{2}\sum_{i=1}^m \frac{s_i^2}{v_i}\right) \times \left[\frac{1}{\nu(\nu+1)^2(\nu+3)}\right]^{\frac{1}{2}} |\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{-\frac{1}{2}}.$$

Notons que

$$\int_0^\infty \nu^{-\frac{1}{2}}(\nu+1)^{-1}(\nu+3)^{-\frac{1}{2}}\,d\nu = \int_0^c \nu^{-\frac{1}{2}}(\nu+1)^{-1}(\nu+3)^{-\frac{1}{2}}\,d\nu + \int_c^\infty \nu^{-\frac{1}{2}}(\nu+1)^{-1}(\nu+3)^{-\frac{1}{2}}\,d\nu$$

$$\leq \int_0^c \nu^{-\frac{1}{2}}d\nu + \int_c^\infty \nu^{-2}d\nu = 2c^{\frac{1}{2}} + c^{-1} < \infty \text{ pour tout } c > 0.$$

Après élimination de ν par intégration, nous avons

$$\boldsymbol{\pi}_{\mathrm{MJ}}(\mathbf{v} \,|\, \mathbf{y}, \mathbf{s}^2) \leq K\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left(-\frac{1}{2}\sum_{i=1}^m \frac{s_i^2}{v_i}\right)|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{-\frac{1}{2}}.$$

Enfin, puisque $|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{-\frac{1}{2}} \leq (v_{\max})^{\frac{p}{2}}|\mathbf{X}^T\mathbf{X}|^{-\frac{1}{2}}$, où $v_{\max} = \max(v_1, \ldots, v_m)$, si $\min(n_1, \ldots, n_m) > p$, la loi a priori de Jeffreys modifiée mène à une loi a posteriori propre.

Annexe B

Lois conditionnelles complètes

La loi a posteriori complète des paramètres sachant les données est spécifiée en (A.1). Pour l'exécution de la méthode MCMC, il est commode d'utiliser les paramètres latents η_i $(i = 1, \ldots, m)$ tels que

$$\theta_i \,|\, \eta_i \overset{\text{ind}}{\sim} \mathrm{N}(\mathbf{x}_i^T\boldsymbol{\beta}, \eta_i^{-1}\sigma_\delta^2)$$

$$\eta_i \overset{\text{iid}}{\sim} \mathrm{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right).$$

Soit $\boldsymbol{\Lambda} = \mathrm{Diag}(\eta_1, \ldots, \eta_m)$. Les lois conditionnelles complètes sont

I. $[\sigma_\delta^2 \,|\, \boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{v}, \nu, \mathbf{y}, \mathbf{s}^2] \sim \mathrm{GI}\left[\frac{p+m}{2}, \frac{1}{2}\left\{a + \sum_{i=1}^m \eta_i(\theta_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\right\}\right]$;

II. $[v_i \,|\, \boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{\beta}, \sigma_\delta^2, \nu, \mathbf{y}, \mathbf{s}^2] \overset{\text{ind}}{\sim} \mathrm{GI}\left[\frac{n_i}{2}, \frac{1}{2}\left\{(y_i - \theta_i)^2 + s_i^2\right\}\right]$;

III. $[\boldsymbol{\beta} \,|\, \boldsymbol{\theta}, \boldsymbol{\Lambda}, \mathbf{v}, \sigma_\delta^2, \nu, \mathbf{y}, \mathbf{s}^2] \sim \mathrm{N}\left[(\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}\boldsymbol{\theta}, \sigma_\delta^2(\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X})^{-1}\right]$;

IV. $[\eta_i \,|\, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu, \mathbf{y}, \mathbf{s}^2] \overset{\text{ind}}{\sim} \mathrm{G}\left[\frac{\nu+1}{2}, \frac{1}{2}\left\{\nu + \frac{(\theta_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{\sigma_\delta^2}\right\}\right]$;

$$\text{V.} \quad [\theta_i \,|\, \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu, \mathbf{y}, \mathbf{s}^2] \overset{\text{ind}}{\sim} \mathrm{N}\left[\left(v_i^{-1} + \sigma_\delta^{-2}\eta_i\right)^{-1}\left(v_i^{-1} y_i + \sigma_\delta^{-2}\eta_i \mathbf{x}_i^T \boldsymbol{\beta}\right), \left(v_i^{-1} + \sigma_\delta^{-2}\eta_i\right)^{-1}\right];$$

et

$$\text{VI.} \quad [\nu \,|\, \boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \mathbf{y}, \mathbf{s}^2] \propto \nu^{\frac{m-1}{2}} \exp\left\{-\tfrac{\nu}{2}\sum_{i=1}^{m}(\eta_i - \log\eta_i - 1)\right\}$$

$$\times \left\{\frac{\nu^{\frac{m\nu-m}{2}}\exp\left(-\frac{m\nu}{2}\right)}{2^{\frac{m\nu}{2}}\Gamma^m\left(\frac{\nu}{2}\right)}\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}(\nu+1)^{-1}(\nu+3)^{-\frac{1}{2}}\right\}. \tag{B.1}$$

Toutes les lois sauf (VI) requièrent la génération d'échantillons à partir de lois standards. Alors que nous utilisons la méthode d'échantillonnage de Gibbs pour les lois (I) à (V), nous utilisons l'algorithme de Metropolis-Hastings pour générer les échantillons à partir de (VI) comme il est décrit dans Chib et Greenberg (1995).

Application du résultat de Chib et Greenberg (1995) à (IV) dans (B.1)

Si la densité cible $\pi(t)$ peut s'écrire sous la forme $\pi(t) \propto \psi(t)\,h(t)$, où $h(t)$ est une densité qui peut être échantillonnée et $\psi(t)$ est uniformément bornée, alors on peut prendre $h(t)$ comme une densité candidate pour le tirage d'échantillons et utiliser $\psi(t)$ dans $\alpha(x, y) = \min\{\psi(y)/\psi(x), 1\}$ qui est la probabilité de mouvement.

Rappelons que la loi conditionnelle complète de ν est

$$\boldsymbol{\pi}(\nu \,|\, \cdot) \propto \nu^{\frac{m-1}{2}}\exp\left\{-\frac{\nu}{2}\sum_{i=1}^{m}(\eta_i - \log\eta_i - 1)\right\} \times \left\{\frac{\nu^{\frac{m\nu-m}{2}}\exp\left(-\frac{m\nu}{2}\right)}{2^{\frac{m\nu}{2}}\Gamma^m\left(\frac{\nu}{2}\right)}\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}(\nu+1)^{-1}(\nu+3)^{-\frac{1}{2}}\right\}.$$

Puisque $\Gamma\left(\frac{\nu}{2}\right) \geq \sqrt{2\pi}\,\exp\left(-\frac{\nu}{2}\right)\left(\frac{\nu}{2}\right)^{\frac{\nu-1}{2}}$, le deuxième terme $\{\cdot\}$ ci-dessus est borné par $\left(\sqrt{2\pi}\right)^{-m}(\nu+1)^{\frac{p}{2}-1}(\nu+3)^{-\frac{p}{2}-\frac{1}{2}} \leq \left(\sqrt{2\pi}\right)^{-m}\Big/\sqrt{3}$. Donc, nous pouvons appliquer la troisième méthode de la section 5 de Chib et Greenberg (1995) avec

$$h(\nu) \sim \mathrm{G}\left(\frac{m+1}{2}, \frac{1}{2}\sum_{i=1}^{m}(\eta_i - \log\eta_i - 1)\right)$$

et

$$\psi(\nu) = \frac{\nu^{\frac{m\nu-m}{2}}\exp\left(-\frac{m\nu}{2}\right)}{2^{\frac{m\nu}{2}}\Gamma^m\left(\frac{\nu}{2}\right)}\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}(\nu+1)^{-1}(\nu+3)^{-\frac{1}{2}}.$$

Ici, $h(\nu)$ est une densité génératrice de candidats, et $\psi(\nu)$ est uniformément bornée.

Bibliographie

Arora, V., et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Stastica Sinica*, 7, 1053-1063.

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error component model for prediction of mean crop areas using survey and satellite data. *Journal of the American Statistical Association*, 95, 28-36.

Bell, W.R. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. Dans *Proceedings of the Survey Research Methods Section,* American Statistical Association, 327-333.

Bell, W.R., et Huang, E.T. (2006). Utilisation de la loi t pour traiter les valeurs aberrantes dans l'estimation pour de petits domaines. Recueil : *Symposium 2006, Enjeux méthodologiques reliés à la mesure de la santé des populations,* Statistique Canada.

Butar, F., et Lahiri, P. (2002). On the measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference,* 112, 63-76.

Chib, S., et Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician,* 49, 327-335.

Dass, S.C., Maiti, T., Ren, H. et Sinha, S. (2012). Estimation des intervalles de confiance des paramètres de petit domaine avec rétrécissement des moyennes et des variances. *Techniques d'enquête,* 38, 2, 187-203. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2012002/article/11756-fra.pdf.

Datta, G.S., et Lahiri, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of outliers. *Journal of Multivariate Analysis,* 54, 310-328.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association,* 74, 269-277.

Fernandez, C., et Steel, M.F.J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association,* 93, 359-371.

Fonseca, T.C.O., Ferreira, M.A.R. et Migon, H.S. (2008). Objective Bayesian analysis for the student $-t$ regression model. *Biometrika,* 95, 325-333.

Jacquier, E., Polson, N.G. et Rossi, P.E. (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics,* 122, 185-212.

Jiang, J., Lahiri, P. et Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Annals of Statistics,* 30, 1782-1810.

Lahiri, P., et Rao, J.N.K. (1995). Robust estimation of mean square error of small area estimators. *Journal of the American Statistical Association,* 90, 758-766.

Lange, K.L., Little, R.J.A. et Taylor, J.M.G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association,* 84, 881-896.

Maiti, T., Ren, H. et Sinha, A. (2014). Prediction error of small area predictors shrinking both means and variances. *Scandinavian Journal of Statistics,* 41, 775-790.

Otto, M.C., et Bell, W.R. (1995). Sampling error modelling of poverty and income statistics for states. Dans *Proceedings of the Section on Government Statistics,* American Statistical Association, 160-165.

Rivest, L.-P., et Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. Dans *Proceedings of the International Conference on Recent Advances in Survey Sampling.*

Slud, E.V., et Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of Royal Statistical Society, Series B*, 68, 239-257.

Sugasawa, S., Tamae, H. et Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150-167.

Vrontos, I.D., Dellaportas, P. et Politis, D.N. (2000). Full Bayesian inference for GARCH and EGARCH models. *Journal of Business and Economic Statistics*, 18, 187-198.

Wang, J., et Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 1, 107-114. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2006001/article/9263-fra.pdf.

Techniques d'enquête, juin 2018
Vol. 44, nº 1, pp. 125-155
Statistique Canada, nº 12-001-X au catalogue

Calage assisté par un modèle pour des données de sondage non probabiliste en utilisant le LASSO adaptatif

Jack Kuang Tsung Chen, Richard L. Valliant et Michael R. Elliott[1]

Résumé

Le cadre fondé sur l'échantillonnage probabiliste a joué un rôle dominant en recherche par sondage, parce qu'il fournit des outils mathématiques précis pour évaluer la variabilité d'échantillonnage. Toutefois, en raison de la hausse des coûts et de la baisse des taux de réponse, l'usage d'échantillons non probabilistes s'accroît, particulièrement dans le cas de populations générales, pour lesquelles le tirage d'échantillons à partir d'enquêtes en ligne devient de plus en plus économique et facile. Cependant, les échantillons non probabilistes posent un risque de biais de sélection dû à des différences d'accès et de degrés d'intérêt, ainsi qu'à d'autres facteurs. Le calage sur des totaux statistiques connus dans la population offre un moyen de réduire éventuellement l'effet du biais de sélection dans les échantillons non probabilistes. Ici, nous montrons que le calage assisté par un modèle en utilisant le LASSO adaptatif peut donner un estimateur convergent d'un total de population à condition qu'un sous-ensemble des variables explicatives réelles soit inclus dans le modèle de prédiction, permettant ainsi qu'un grand nombre de covariables possibles soit incluses sans risque de surajustement. Nous montrons que le calage assisté par un modèle en utilisant le LASSO adaptatif produit une meilleure estimation, pour ce qui est de l'erreur quadratique moyenne, que les méthodes concurrentes classiques, tels les estimateurs par la régression généralisée (GREG), quand un grand nombre de covariables sont nécessaires pour déterminer le modèle réel, sans vraiment qu'il y ait perte d'efficacité par rapport à la méthode GREG quand de plus petits modèles suffisent. Nous obtenons aussi des formules analytiques pour les estimateurs de variance des totaux de population, et comparons le comportement de ces estimateurs aux estimateurs bootstrap. Nous concluons par un exemple réel en utilisant des données provenant de la *National Health Interview Survey*.

Mots-clés : Estimateurs de type LASSO adaptatif; estimateur par la régression généralisée; échantillon non représentatif; surajustement; sélection des variables; propriété d'oracle.

1 Introduction

L'échantillonnage probabiliste a joué un rôle prépondérant en recherche par sondage pendant la plus grande partie du siècle dernier (Stephan, 1948; Frankel et Frankel, 1987). Étant donné des mesures complètes d'unités échantillonnées dont les probabilités de sélection sont connues, la théorie de la randomisation élimine le biais de sélection en générant des échantillons représentatifs de la population cible. Par ailleurs, les échantillons non probabilistes produits sans connaître les probabilités de sélection posent automatiquement un risque de biais de sélection, car les échantillons peuvent différer de la population cible en ce qui concerne des statistiques importantes (Groves, 2006). Les échecs bien documentés des sondages réalisés pendant les campagnes électorales présidentielles de 1936 et de 1948 mettent en relief les erreurs qui peuvent être commises lorsqu'on fait des inférences sur une population à partir d'échantillons non probabilistes (Mosteller, 1949).

Bien que le cadre d'échantillonnage probabiliste fournisse aux spécialistes des enquêtes des outils mathématiques précis pour évaluer et corriger les erreurs d'échantillonnage, la baisse des taux de réponse aux enquêtes s'appuyant sur les méthodes traditionnelles de collecte de données suscite des inquiétudes

1. Jack Kuang Tsung Chen, statisticien, Survey Monkey Inc., Palo Alto, CA. Courriel : jjkkttcc@gmail.com; Richard L. Valliant, professeur-chercheur, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. Courriel : valliant@umich.edu; Michael R. Elliott, professeur-chercheur, Survey Research Center, Institute for Social Research, et professeur, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI. Courriel : mrelliot@umich.edu.

quant au biais de non-réponse potentiellement élevé des échantillons probabilistes. Le *Pew Research Center* a indiqué que les taux de réponse à leurs sondages téléphoniques sont passés de 36 % en 1997 à 9 % en 2012 (Kohut, Keeter, Doherty, Dimock et Christian, 2012), ce qui fait penser que l'échantillonnage probabiliste par téléphone pourrait ne plus être une méthode viable pour les enquêtes auprès des populations générales. En outre, l'obtention de données sans exercer un grand contrôle sur l'ensemble d'unités auprès desquelles elles sont recueillies est souvent un exercice moins coûteux et plus rapide que l'échantillonnage probabiliste. Ces raisons font que l'échantillonnage non probabiliste connaît à l'heure actuelle une sorte de renaissance (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau, 2013; Elliott et Valliant, 2017). La collecte de données en ligne, une plateforme sans base de sondage universelle pour effectuer un échantillonnage probabiliste, représentait près de la moitié des dépenses en recherche par sondage aux États-Unis en 2012 (Terhanian et Bremer, 2012), et il est presque certain qu'elle a pris de l'essor depuis.

Pour de nombreux organismes d'enquête, l'ajustement des poids de sondage sur des données auxiliaires connues est l'étape finale et la plus cruciale du processus de construction des pondérations. Les approches classiques comprennent la poststratification, qui consiste à ajuster les poids de manière que la distribution des variables auxiliaires catégoriques dans l'échantillon pondéré concorde avec leur distribution dans la population, ainsi que son extension à l'estimation par la régression généralisée (GREG), qui fait en sorte que, pour chaque variable auxiliaire (continue ou catégorique), la somme des valeurs soit égale au total correspondant dans la population (Deville et Särndal, 1992). Le calage joue un rôle important en statistique officielle, parce qu'il peut produire des poids tels que les estimations démographiques pondérées obtenues d'après différentes enquêtes concordent.

Sous échantillonnage probabiliste, si les poids de sondage sont égaux à l'inverse des probabilités de sélection, les estimations pondérées des totaux dans l'échantillon sont des estimations sans biais sous le plan de sondage du total de population. Le calage ajuste les poids de sondage de façon minimale afin que les totaux pour les variables auxiliaires dans l'échantillon pondéré concordent avec leurs totaux de population connus (Särndal, Swensson et Wretman, 1992). Dans le contexte d'échantillonnage probabiliste, le calage est introduit pour réduire la variance ou corriger le biais grâce à un ajustement pour tenir compte du sous-dénombrement ou du surdénombrement de sous-groupes de l'échantillon. Dans le cas de grands échantillons, les poids calés finaux peuvent être appliqués à toutes les variables de l'enquête, parce qu'ils retiennent approximativement la propriété d'absence de biais des poids de sondage originaux. Par contre, dans le cas de l'échantillonnage non probabiliste, il n'existe pas de probabilités de sélection pour construire les poids de sondage initiaux qui peuvent produire des estimations sans biais. Donc, il n'est pas garanti que les poids calés habituels puissent fonctionner pour toutes les variables dans l'échantillon non probabiliste. Afin de faire des inférences à partir d'échantillons non probabilistes, une approche pratique consiste à construire un ensemble de poids qui peuvent réduire la racine carrée de l'erreur quadratique moyenne (REQM ou RMSE, de l'anglais *root mean square error*) des estimations pondérées pour un résultat d'intérêt particulier. Le calage assisté par un modèle fournit le cadre pour construire des poids calés ciblant une variable de résultat, étant donné un modèle pouvant approximer les valeurs espérées du résultat (Wu et

Sitter, 2001). La clé d'un calage assisté par un modèle réussi est un modèle possédant de fortes propriétés prédictives, c'est-à-dire que les paramètres du modèle estimés d'après un échantillon peuvent être utilisés pour prédire fiablement les valeurs dans un échantillon différent de la même population. Évidemment, de telles variables explicatives ne sont pas toujours disponibles; Tourangeau, Conrad et Couper (2013) donnent un exemple où le manque de covariables prédictives empêche que les ajustements des pondérations produisent de bons résultats. Cependant, Tourangeau et coll. (2013) s'intéressaient aux enquêtes auprès des ménages. Les variables explicatives peuvent être plus puissantes dans le cas d'enquêtes auprès d'établissements ou d'institutions, ou de certaines enquêtes spécialisées, comme les sondages électoraux. Par exemple, Wang, Rothschild, Goel et Gelman (2015) s'appuient sur l'affiliation politique et le candidat élu lors de l'élection précédente pour faire des prédictions exactes du résultat de l'élection présidentielle américaine de 2012 en se basant sur un échantillon non probabiliste dont la répartition différait considérablement de celle de l'ensemble des électeurs.

Manifestement, on pourrait donc s'attendre à ce que le calage assisté par un modèle soit le plus efficace lorsqu'il existe un ensemble relativement riche de covariables de population auxiliaires et, en conséquence, un très grand ensemble de modèles à prendre en considération. Dans ces conditions, il peut être difficile d'obtenir un équilibre entre la structure – pour minimiser l'erreur de spécification du modèle et donc le biais –, et la parcimonie – pour stabiliser les estimations et donc minimiser la variance. Le *Least Angle Shrinkage and Selection Operator,* ou LASSO, est une régression régularisée qui permet d'effectuer à la fois la sélection des variables et l'estimation des paramètres (Tibshirani, 1996). Une vaste gamme d'applications ont montré que le LASSO prévient bien le surajustement du modèle grâce à la sélection automatique de modèles plus précis et parcimonieux. Kamarianakis, Shen et Wynter (2012) ont réussi, à l'aide du LASSO, à prédire la vitesse de circulation moyenne en présence de multicollinéarités importantes dues à des variables explicatives agrégées au niveau de la région. Kohannim, Hibar, Stein, Jahanshad, Hua, Rajagopalan, Toga, Jack Jr, Weiner, de Zubicaray et McMahon (2012) ont appliqué la régression LASSO pour déterminer des sous-ensembles de polymorphismes de nucléotide unique (SNP) de grandes dimensions et corrélés qui sont associés aux mesures de la structure cérébrale. Dans le cadre d'une revue des défis liés à l'analyse écologique avec des covariables linéaires, Dormann, Elith, Bacher, Buchmann, Carl, Carre, Marquez, Gruber, Lafourcade, Leitao et Mnkemller (2013) ont constaté que le LASSO est l'une des méthodes qui produisent systématiquement de faibles valeurs de la racine carrée de l'erreur quadratique moyenne. Dans le domaine de la génétique et de la finance, le LASSO a été utilisé efficacement en modélisation prédictive comprenant des centaines ou des milliers de variables explicatives (Wu, Chen, Hastie, Sobel et Lange, 2009).

Des formes stabilisantes du calage classique sont prises en considération dans diverses publications. Park et Yang (2008) examinent pour un estimateur par la régression généralisée une forme de type régression ridge comprenant un terme de pénalité pour stabiliser les estimateurs par calage, qui s'avère converger sous le plan de sondage et réduire la variance dans des études en simulation. Goga, Muhammad-Shehzad et Vanheuverzwyn (2011), et Cardot, Goga et Shehzad (2017) ont examiné le calage sur les composantes

principales des totaux de population plutôt que les totaux de population proprement dit, ce qui permet de regrouper un grand nombre de variables auxiliaires en un sous-ensemble de dimension raisonnable. Aspect peut-être le plus pertinent pour les présents travaux, McConville (2011) et McConville, Breidt, Lee et Moisen (2017) ont élaboré, de nouveau sous calage classique, le cadre théorique pour montrer l'absence de biais et la convergence approximatives sous le plan de l'estimateur d'un total par calage avec le LASSO, sachant les estimations des paramètres de la régression LASSO. Bien que le calage assisté par un modèle LASSO soit très prometteur quant à la construction d'un ensemble de poids qui peut résulter en une petite REQM des estimations pondérées pour une variable de résultat dans un échantillon non probabiliste, il n'existe aucun cadre théorique établi pour les propriétés de biais et de convergence des estimateurs par calage assistés par un modèle LASSO pour un échantillon non probabiliste.

Donc, les principaux objectifs du présent article sont :

1) élaborer le cadre théorique pour le calage assisté par un modèle LASSO pour les variables de résultat continues ainsi que binaires, c'est-à-dire calculer l'estimation ponctuelle du total, son espérance asymptotique et l'estimation asymptotique de la variance théorique;
2) étudier la performance relative, en ce qui concerne la racine carrée de l'erreur quadratique moyenne, du calage avec le LASSO par rapport au calage classique sous différents types de résultats, de plans d'échantillonnage, de tailles d'échantillon et de structures de covariance des variables de calage.

Alors que notre développement de la théorie asymptotique repose sur l'hypothèse que les poids de sondage sont connus, une constatation clé est que le calage avec le LASSO donne des estimateurs convergents d'un total de population, que les poids de sondage soient ou non spécifiés correctement, à condition que le modèle de régression contienne tous les paramètres de superpopulation sous forme d'un sous-ensemble des paramètres dans le modèle. Donc, dans les études en simulation, nous nous concentrons sur l'estimation dans des conditions d'échantillonnage non probabiliste, où les poids de sondage initiaux sont considérés comme étant les mêmes que pour l'échantillonnage aléatoire simple (EAS), $d_i = N/n$ pour des tailles de population et d'échantillon N et n, indépendamment de la façon dont les échantillons sont formés (qui, en pratique, serait inconnue). Nous appliquons également le calage avec le LASSO à l'estimation du nombre total d'adultes ayant reçu un diagnostic de cancer dans la population américaine, en utilisant des données sur l'incidence du cancer provenant de la *National Health Interview Survey* (NHIS) de 2013 et des données auxiliaires de population provenant de l'*American Community Survey* du *US Census Bureau*, en ne tenant pas compte des poids de sondage pour approximer un échantillon non probabiliste et comparer les résultats aux estimations entièrement pondérées (représentatives).

La présentation de l'article est la suivante. À la section 2, nous donnons la définition et la notation pour le calage et la régression LASSO. À la section 3, nous développons l'estimateur par calage avec le LASSO d'un total de population, son espérance sous le modèle et ses variances asymptotiques. À la section 4, nous décrivons la simulation et les résultats de l'évaluation des estimations de la racine carrée de l'erreur

quadratique moyenne et de la variance de l'estimateur calé par la méthode LASSO. À la section 5, nous présentons l'exemple de la NHIS. Enfin, à la section 6, nous résumons nos constatations en guise de conclusion.

2 Calage

2.1 Calage classique

Pour un échantillon analytique s_A (l'échantillon qui nécessite un calage des poids) de taille n tiré selon le plan de sondage $\mathcal{A}$ avec les poids de sondage $\underset{n\times 1}{\mathbf{d}}$ et la matrice diagonale des poids de sondage $\mathbf{D}$, les poids calés $\underset{n\times 1}{\mathbf{w}}$ minimisent une mesure de distance

$$E_{\mathcal{A}}\left[\sum_{i\in s_A} g(w_i, d_i)\Big/ q_i\right] \tag{2.1}$$

sous la contrainte :

$$\sum_{i\in s_A} w_i \mathbf{x}_i^T = \mathbf{T}^{\mathbf{X}} \tag{2.2}$$

où $E_{\mathcal{A}}$ est l'espérance sous le plan analytique (probabiliste), $g(w_i, d_i)$ est une fonction différentiable par rapport à w_i, strictement convexe sur un intervalle contenant d_i, et $g(d_i, d_i) = 0$, et où $\mathbf{T}^X$ est un vecteur ligne de totaux de population connus de variables de calage d'échantillon $\mathbf{X}$ (Deville et Särndal, 1992). La constante q_i est indépendante du poids de sondage d_i. L'estimateur par la régression généralisée (GREG) utilisé habituellement fait appel à la distance du khi-deux : $g(w_i, d_i) = (w_i - d_i)^2 \big/ d_i$ avec $q_i = 1$. Sous cette mesure de distance :

$$\mathbf{w}^{\text{GREG}} = \mathbf{d} + \mathbf{DX}\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)^T. \tag{2.3}$$

L'estimation du total de population du résultat $\mathbf{y}$ est fondée sur les poids calés :

$$\begin{aligned}\hat{T}_y^{\text{GREG}} &= \mathbf{w}^{(\text{GREG})T}\mathbf{y} \\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy} \\ &= \hat{T}_y^{\text{HT}} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}}\end{aligned} \tag{2.4}$$

où $\hat{T}_y^{\text{HT}} = \sum_{i\in s_A} d_i y_i$ est l'estimateur fondé sur le plan de sondage (pondéré) classique, $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy}$ est l'estimation des moindres carrés pondérés de la régression linéaire $E_\xi\left[y_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}\right] = \mathbf{x}_i^T\boldsymbol{\beta}$, sachant les poids $\mathbf{D}$. (Cela correspond à l'estimateur poststratifié quand $\mathbf{X}$ consiste entièrement en totaux de cellule pour des variables catégoriques.) Les poids calés définis dans l'équation (2.3) ne dépendent d'aucune variable de résultat. Donc, le même ensemble de poids peut être appliqué à toutes les variables de l'enquête. Notons que la régression GREG suppose un modèle de travail linéaire. Même si $\hat{T}_y^{\text{GREG}}$ est

asymptotiquement sans biais sous le plan pour T_y, quand la relation entre $\mathbf{y}$ et $\mathbf{X}$ est non linéaire, comme cela est le cas quand $\mathbf{y}$ est binaire, la variance sous le plan de $\hat{T}_y^{\text{GREG}}$ peut être plus grande que la variance sous le plan de $\hat{T}_y^{\text{HT}}$.

2.2 Calage assisté par un modèle

Les estimateurs par calage assisté par un modèle peuvent présenter un avantage important par rapport à l'estimateur $\hat{T}_y^{\text{GREG}}$, parce que le calage assisté par un modèle permet d'utiliser des modèles non linéaires pour aider à construire les poids calés. Dans le calage assisté par un modèle (noté MC, de l'anglais *model-assisted calibration*), nous supposons qu'il existe une relation entre un résultat $\mathbf{y}$ et $\mathbf{X}$ par l'intermédiaire des deux premiers moments (Wu et Sitter, 2001) :

$$E_\xi(y_i \mid \mathbf{x}_i) = \mu(\mathbf{x}_i, \boldsymbol{\beta}), V_\xi(y_i \mid \mathbf{x}_i) = v_i^2\sigma^2 \tag{2.5}$$

où $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ et σ sont les paramètres de superpopulation inconnus, $\mu(x_i, \boldsymbol{\beta})$ est une fonction connue de $\mathbf{x}_i$ et $\boldsymbol{\beta}$, et v_i est une fonction connue de $\mathbf{x}_i$ ou $\mu(\mathbf{x}_i, \boldsymbol{\beta})$. E_ξ et V_ξ sont l'espérance et la variance sous le modèle ξ. Soit $\mathbf{B}$ l'estimation en population finie (ou recensement) de $\boldsymbol{\beta}$ (c'est-à-dire l'estimateur de la quasi-vraisemblance de $\boldsymbol{\beta}$ fondé sur la population finie entière), et $\hat{\mu}_i = \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right)$, où $\hat{\mathbf{B}}$ est l'estimation sur échantillon de $\mathbf{B}$. Les poids calés avec assistance d'un modèle $\mathbf{w}$ minimisent alors une mesure de distance $E_{\mathcal{A}}\left[\sum_{i\in s_A} g(w_i, d_i)/q_i\right]$ sous les contraintes $\sum_{i\in s_A} w_i = N$ et $\sum_{i\in s_A} w_i\hat{\mu}_i = \sum_i^N \hat{\mu}_i$. La principale différence conceptuelle entre le calage classique et le calage assisté par un modèle est que, dans ce dernier, les contraintes sont fondées sur deux quantités, à savoir 1) la taille de la population et 2) le total de population des valeurs prédites $\hat{\mu}_i$. Dans le calage classique, la contrainte est un vecteur de totaux de population de $\mathbf{X}$ (voir l'équation (2.2)). Sous la mesure de distance du khi-deux avec $q_i = 1$, les poids calés avec assistance d'un modèle sont :

$$\mathbf{w}^{\text{MC}} = \mathbf{d} + \mathbf{DM}\left(\mathbf{M}^T\mathbf{DM}\right)^{-1}\left(\mathbf{T}^M - \mathbf{d}^T\mathbf{M}\right)^T \tag{2.6}$$

où $\mathbf{T}^M = \left[N, \sum_i^N \hat{\mu}_i\right]$ et $\mathbf{M} = \left[\mathbf{d}, (\hat{\mu}_i)_{i\in s_A}\right]$. (Dans le contexte non probabiliste, le vecteur des poids de sondage $\mathbf{d}$ peut être remplacé par $(N/n)\,\mathbf{1}$.) L'estimation du total de population basée sur les poids calés avec assistance d'un modèle est alors :

$$\begin{aligned}\hat{T}_y^{\text{MC}} &= \left(\mathbf{w}^{\text{MC}}\right)^T\mathbf{y} \\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{M}} - \mathbf{d}^T\mathbf{M}\right)\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy} \\ &= \hat{T}_y^{\text{HT}} + \left(\sum_i^N \hat{\mu}_i - \sum_{i\in s_A} d_i\hat{\mu}_i\right)\hat{B}^{\text{MC}}\end{aligned} \tag{2.7}$$

où $\hat{B}^{\text{MC}}$ est la pente de calage qui satisfait les contraintes de calage (différente des estimations des paramètres du modèle $\hat{\mathbf{B}}$) :

$$\hat{B}^{\mathrm{MC}} = \frac{\sum_{i\in s_A} d_i \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)(y_i - \bar{y})}{\sum_{i\in s_A} d_i \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)^2}, \quad \hat{\bar{\mu}} = \sum_{i\in s_A} d_i \hat{\mu}_i \Big/ \sum_{i\in s_A} d_i, \quad \bar{y} = \sum_{i\in s_A} d_i y_i \Big/ \sum_{i\in s_A} d_i .$$

L'absence de biais et la petite variance de $\hat{T}_y^{\mathrm{MC}}$ dépendent toutes deux de la mesure dans laquelle les $\hat{\mu}_i$ sont de bonnes approximations de la vraie valeur espérée de y_i.

3 Sélection du modèle et calage robuste avec le LASSO adaptatif

3.1 Contexte du LASSO adaptatif

3.1.1 Définition et paramètres

Les coefficients de la régression de type LASSO adaptatif s'obtiennent en résolvant une équation de régression pénalisée. Pour la régression linéaire pénalisée par le LASSO adaptatif (Zou, 2006) :

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{i\in s_A} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} \alpha_j^{\gamma} \mid \beta_j \mid \right) \tag{3.1}$$

où α_j^{γ} est un poids ajustable et λ_n est une pénalité utilisée pour optimiser une mesure de l'adéquation du modèle. Similairement, pour la régression logistique pénalisée par le LASSO adaptatif :

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{i\in s_A} [-y_i (\mathbf{x}_i^T \beta) + \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}))] + \lambda_n \sum_{j=1}^{p} \alpha_j^{\gamma} \mid \beta_j \mid \right). \tag{3.2}$$

Sachant λ_n et γ, nous pouvons calculer $\hat{\boldsymbol{\beta}}$ par des procédures itératives. Le package *glmnet* de R calcule les estimateurs LASSO adaptatif linéaire ainsi que logistique (Friedman, Hastie et Tibshirani, 2010).

Le rôle du paramètre de pondération, α_j, est d'empêcher le LASSO de sélectionner des covariables dont la taille de l'effet est grande en faveur d'une réduction de l'erreur de prédiction quand la taille d'échantillon est petite. Donc, les poids sont inversement proportionnels aux tailles des effets des paramètres de régression : $\alpha_j \propto 1 / |\beta_j|$. Un choix fréquent de α_j est $1 / |\hat{\beta}_j^{\mathrm{EMV}}|$, où $\hat{\beta}_j^{\mathrm{EMV}}$ est l'estimation du maximum de vraisemblance de β_j. La puissance du paramètre de poids, γ, est une constante plus grande que 0 qui interagit avec α_j pour empêcher le LASSO de sélectionner ou d'exclure des paramètres. Par exemple, si nous souhaitons quand même que le LASSO favorise les covariables dont l'effet est grand quand la taille d'échantillon est petite, nous devons choisir une petite valeur de γ. Si nous voulons accorder encore moins d'importance aux tailles des effets, nous devons choisir une grande valeur de γ.

3.1.2 Propriété d'oracle

La mesure de la performance d'une méthode de sélection de modèles et d'estimation s'appuie sur un concept important appelé la « propriété d'oracle ». La méthode optimale sélectionne les bonnes variables et

fournit des estimations sans biais des paramètres choisis. Supposons que, dans un modèle de régression complet, les paramètres ont à la fois des composantes nulles et non nulles. Sans perte de généralité, considérons que les p premières sont non nulles et que les q dernières sont nulles :

$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}^{(1)}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} = \mathbf{0} \end{pmatrix}.$$

Un modèle de régression possède la propriété d'oracle s'il satisfait les conditions suivantes (Fan et Li, 2001) :

- la probabilité d'obtenir l'estimation 0 pour des paramètres dont la valeur est nulle tend vers un : $\Pr\left(\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}\right) \to 1$ quand $n \to \infty$;
- les estimations des paramètres non nuls sont aussi bonnes que si le sous-modèle réel était connu : $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}\right) \to N(\mathbf{0}, \mathbf{C})$ où $\mathbf{C} = \Sigma\left(\boldsymbol{\beta}^{(1)}\right)$ est la matrice de covariance de $\boldsymbol{\beta}^{(1)}$ sous un modèle linéaire, et $\mathbf{C} = I^{-1}\left(\boldsymbol{\beta}^{(1)}\right)$ est l'inverse de la matrice d'information de Fisher de $\boldsymbol{\beta}^{(1)}$ sous le modèle linéaire généralisé.

Pour l'inférence en population finie, supposons que l'indice ν désigne une population de taille N_ν, représentons par $\mathbf{B}_\nu$ les estimations de la quasi-vraisemblance de $\boldsymbol{\beta}$ dans la population ν, et par $\hat{\mathbf{B}}_\nu$ l'estimation de $\mathbf{B}_\nu$ basée sur un échantillon de taille $n_\nu \le N_\nu$. Nous supposons que $N_\nu \to \infty$, $n_\nu \to \infty$ et $n_\nu / N_\nu \to 0$ quand $\nu \to \infty$. L'équivalent en population finie de la propriété d'oracle est alors :

$$\begin{aligned} \Pr\left(\hat{\mathbf{B}}^{(2)}_\nu = \mathbf{0}\right) &\to 1 \\ \sqrt{n_\nu}\left(\hat{\mathbf{B}}^{(1)}_\nu - \mathbf{B}^{(1)}_\nu\right) &\to N_\nu(\mathbf{0}, \mathbf{C}_\nu) \\ \mathbf{B}_\nu &\to \boldsymbol{\beta} \\ \text{quand } \nu \to \infty & \end{aligned}$$

où $\mathbf{C}_\nu = \Sigma\left(\mathbf{B}^{(1)}_\nu\right)$ est la matrice de covariance de $\mathbf{B}^{(1)}_\nu$ si le modèle est linéaire, et $\mathbf{C}_\nu = I^{-1}\left(\mathbf{B}^{(1)}_\nu\right)$ est l'inverse de la matrice d'information de Fisher de $\mathbf{B}^{(1)}_\nu$ sous le modèle linéaire généralisé.

Zou (2006) a montré que si $\lambda_n \Big/ \left(\sqrt{n} \Big/ \left(\sqrt{n}\right)^\gamma\right) \to \infty$ et $\lambda_n / \sqrt{n} \to 0$, alors le LASSO adaptatif satisfait la propriété d'oracle. Les conditions requièrent que λ_n augmente au moins à la vitesse de $\sqrt{n} \Big/ \left(\sqrt{n}\right)^\gamma$, mais pas plus rapidement que $\sqrt{n}$. Le choix de λ_n et γ, et le code R pour l'implémenter sont discutés en annexe.

3.2 Calage avec le LASSO

À la présente section, nous établissons les formules analytiques pour un estimateur LASSO du total, son espérance sous le modèle et les estimateurs asymptotiques des variances sous le plan de sondage. Nous faisons les hypothèses suivantes :

1. Les échantillons sont tirés selon un plan d'échantillonnage à un seul degré $\mathcal{A}$, en permettant des probabilités de sélection inégales. La probabilité de sélection de l'unité i est notée π_i^A, et la probabilité de sélection conjointe des unités i et j est notée π_{ij}^A. Nous désignons le poids de sondage pour l'unité i par $d_i^A = 1/\pi_i^A$, le vecteur des poids de sondage par $\mathbf{d}^A$, et la matrice diagonale des poids de sondage par $\mathbf{D}^A$.
2. Les données auxiliaires au niveau de la population sont connues et notées $\mathbf{X} = (\mathbf{x}_i^T)$, $i = 1, \ldots, N$.
3. Un modèle de superpopulation est considéré, comme il est décrit à la section 2.2 :

$$E_\xi(y_i \mid \mathbf{x}_i) = \mu(\mathbf{x}_i, \boldsymbol{\beta})$$
$$V_\xi(y_i \mid \mathbf{x}_i) = v_i^2 \sigma^2.$$

4. Les paramètres de superpopulation réels sont un sous-ensemble du modèle de régression complet pour le LASSO :

$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} \end{pmatrix}.$$

5. La probabilité d'observer dans l'échantillon analytique l'étendue complète de $\mathbf{X}$ dans la population est non nulle.

3.2.1 Estimation ponctuelle : $\hat{T}_y^{\text{LASSO}}$

L'estimation par calage avec le LASSO du total peut être obtenue par les étapes suivantes :

1. Obtenir les coefficients de régression LASSO $\hat{\mathbf{B}}$ comme il est décrit à l'annexe.
2. Utiliser $\hat{\mathbf{B}}$ pour calculer $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\mathbf{B}})$ dans la population.
3. Définir $\mathbf{T}^M = \left(N, \sum_i^N \hat{\mu}_i\right)$ et $\mathbf{M} = \left[\mathbf{d}^A, \sum_{i\in s_A} \hat{\mu}_i\right]$ sous la mesure de distance du khi-deux avec $q_i = 1$:

$$\mathbf{w}^{\text{LASSO}} = \mathbf{d}^A + \mathbf{D}^A\mathbf{M}(\mathbf{M}^T\mathbf{D}^A\mathbf{M})^{-1}\left(\mathbf{T}^M - (\mathbf{d}^A)^T\mathbf{M}\right)^T. \tag{3.3}$$

4. Déterminer l'estimateur par calage avec le LASSO du total :

$$\begin{aligned} \hat{T}_y^{\text{LASSO}} &= (\mathbf{w}^{\text{LASSO}})^T\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{M}} - (\mathbf{d}^A)^T\mathbf{M}\right)(\mathbf{X}^T\mathbf{D}^A\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}^A\mathbf{y} \\ &= (\mathbf{d}^A)^T\mathbf{y} + \left(\sum_i^N \hat{\mu}_i - \sum_{i\in s_A} d_i^A\hat{\mu}_i\right)\hat{B}^{\text{MC}} \end{aligned} \tag{3.4}$$

où $\hat{B}^{\text{MC}}$ est la pente de calage qui satisfait les contraintes de calage :

$$\hat{B}^{\mathrm{MC}} = \frac{\sum_{i\in s_A} d_i^A \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)(y_i - \bar{y})}{\sum_{i\in s_A} d_i^A \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)^2},\ \hat{\bar{\mu}} = \sum_{i\in s_A} d_i^A \hat{\mu}_i \Big/ \sum_{i\in s_A} d_i^A,\ \bar{y} = \sum_{i\in s_A} d_i^A y_i \Big/ \sum_{i\in s_A} d_i^A.$$

3.2.2 Comportement asymptotique de $\hat{T}_y^{\mathrm{LASSO}}$

Wu et Sitter (2001) ont établi les conditions pour calculer un estimateur asymptotique par calage assisté par un modèle. Nous énonçons les conditions ici en modifiant légèrement les notations par souci de cohérence avec les présents travaux de recherche. Soit $\boldsymbol{\beta}$ le paramètre de superpopulation réel pour le modèle défini en l'équation (2.5), et $\mathbf{B}$ l'estimateur de la quasi-vraisemblance en population finie de $\boldsymbol{\beta}$. Les conditions qui suivent sont utilisées pour obtenir estimateur asymptotique par calage avec le LASSO du total :

1. $\hat{\mathbf{B}} = \mathbf{B} + O_p\left(1/\sqrt{n}\right)$, où $\mathbf{B}$ est la pente de régression en population finie de $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$.
2. Pour chaque $\mathbf{x}_i$, $\partial\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}$ est continue en $\mathbf{t}$, et $\max_i \left|\partial\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}\right| \le h(\mathbf{x}_i, \boldsymbol{\beta})$ pour $\mathbf{t}$ dans un voisinage de $\boldsymbol{\beta}$, et $N^{-1}\sum_{i\in U} h(\mathbf{x}_i, \boldsymbol{\beta}) = O(1)$.
3. Pour chaque $\mathbf{x}_i$, $\partial^2\mu(\mathbf{x}_i, \mathbf{t})/\partial\mathbf{t}\partial\mathbf{t}^T$ est continue en $\mathbf{t}$, et $\max_{j,k} \left|\partial^2\mu(\mathbf{x}_i, \mathbf{t})/\partial t_j \partial t_k\right| \le k(\mathbf{x}_i, \boldsymbol{\beta})$ pour $\mathbf{t}$ dans un voisinage de $\boldsymbol{\beta}$, et $N^{-1}\sum_{i\in U} k(\mathbf{x}_i, \boldsymbol{\beta}) = O(1)$.
4. Les estimateurs de Horvitz-Thompson (HT) de certaines moyennes de population suivent asymptotiquement une loi normale (Fuller, 2009; pages 47 à 57).
5. $\lambda_n \Big/ \left(\sqrt{n}\Big/\left(\sqrt{n}\right)^{\gamma}\right) \to \infty$ et $\lambda_n/\sqrt{n} \to 0$.

Lemme 1 : Supposons que le modèle de superpopulation (2.5) est vérifié. Soit $\mathbf{B}$ l'estimation de la quasi-vraisemblance en population finie de $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$. Sous les conditions (1) à (5), l'estimateur asymptotique assisté par un modèle du total de population est :

$$\hat{T}_y^{\mathrm{MC}} = \sum_{i\in s_A} d_i^A \left(y_i - \mu_i B^{\mathrm{MC}}\right) + \sum_{i=1}^{N} \mu_i B^{\mathrm{MC}} + o_p\left(\frac{N}{\sqrt{n}}\right) \tag{3.5}$$

où

$$\mu_i = \mu(\mathbf{x}_i, \mathbf{B})$$

$$B^{\mathrm{MC}} = \frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2}.$$

Preuve. Voir l'annexe.

Étant donné le lemme 1, dans le théorème 1, nous obtenons $\hat{T}_y^{\mathrm{LASSO}}$, l'estimateur LASSO asymptotique du total. Dans le théorème 2, nous montrons que $\hat{T}_y^{\mathrm{LASSO}}$ est sans biais sous le modèle pour le total de population. Enfin, dans le théorème 3, nous déterminons les estimations de la variance pour l'estimateur d'un total par calage avec le LASSO.

Théorème 1 : Supposons que les paramètres d'un modèle de régression complet comprennent des composantes nulles ainsi que non nulles. Sans perte de généralité, posons que les p premières sont non nulles et que les q dernières sont nulles :

$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}^{(1)}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} \end{pmatrix}, \quad \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta} \quad \text{et} \quad \boldsymbol{\beta}^{(2)} = \mathbf{0}_{(q\times 1)},$$

sous les conditions (1) à (5), l'estimateur asymptotique du total par calage avec le LASSO est :

$$\hat{T}_y^{\text{LASSO}} = \sum_{i\in s_A} d_i^A \left(y_i - \mu_i B^{\text{MC}}\right) + \sum_{i=1}^{N} \mu_i B^{\text{MC}} + o_p\left(\frac{N}{\sqrt{n}}\right).$$

Preuve. Voir l'annexe.

Théorème 2 : $\hat{T}_y^{\text{LASSO}}$ est sans biais sous le modèle, c'est-à-dire que $E_\xi(\hat{T}_y^{\text{LASSO}}) = T$.

Preuve. Voir l'annexe.

Donc, à condition que les paramètres de la régression LASSO incluent les paramètres de superpopulation, $\hat{T}_y^{\text{LASSO}}$ est sans biais sous le modèle, quels que soient les poids de sondage. (Notons qu'il s'agit d'une qualité que $\hat{T}_y^{\text{GREG}}$ partage avec $\hat{T}_y^{\text{LASSO}}$. Cependant, $\hat{T}_y^{\text{LASSO}}$ peut accepter des modèles contenant de beaucoup plus grands nombres de covariables que $\hat{T}_y^{\text{GREG}}$). Cette propriété est essentielle dans le cas des échantillons non probabilistes, pour lesquels il n'existe pas de poids de sondage initiaux pour garantir l'absence de biais.

Théorème 3 : L'estimateur asymptotique de la variance de $\hat{T}_y^{\text{LASSO}}$ est donné par

$$\begin{aligned} v_{\mathcal{A}}\left(\hat{T}_y^{\text{LASSO}}\right) &= \sum_{i\in s_A}\left(\frac{y_i - \hat{\mu}_i\hat{B}^{\text{MC}}}{\pi_i}\right)^2 (1-\pi_i) \\ &\quad + \sum_{i\in s_A}\sum_{j\neq i} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \frac{\left(y_i - \hat{\mu}_i\hat{B}^{\text{MC}}\right)}{\pi_i} \frac{\left(y_j - \hat{\mu}_j\hat{B}^{\text{MC}}\right)}{\pi_j}. \end{aligned} \tag{3.6}$$

Preuve. La variance sous le plan théorique de l'estimateur LASSO est

$$\begin{aligned} V_{\mathcal{A}}\left(\hat{T}_y^{\text{LASSO}}\right) &= V_{\mathcal{A}}\left(\sum_{i\in s_A} d_i^A\left(y_i - \mu_i B^{\text{MC}}\right) + \sum_{i=1}^{N}\mu_i B^{\text{MC}}\right) \\ &= V_{\mathcal{A}}\left(\sum_{i\in s_A} d_i^A\left(y_i - \mu_i B^{\text{MC}}\right)\right) \\ &= \sum_{i\in U}\left(\frac{y_i - \mu_i B^{\text{MC}}}{\pi_i}\right)^2 \pi_i(1-\pi_i) \\ &\quad + \sum_{i\in U}\sum_{j\neq i}(\pi_{ij} - \pi_i\pi_j)\frac{\left(y_i - \mu_i B^{\text{MC}}\right)}{\pi_i}\frac{\left(y_j - \mu_j B^{\text{MC}}\right)}{\pi_j} \end{aligned} \tag{3.7}$$

qui découle de l'équation (3.30) obtenue pour la variance de l'estimateur par calage avec le LASSO classique d'un total dans McConville (2011). L'équation (3.6) s'obtient en remplaçant les quantités de population par les estimations.

Une autre estimation de la variance, proposée par Särndal, Swensson et Wretman (1989), consiste à multiplier $(y_i - \hat{\mu}_i \hat{B}^{\mathrm{MC}})$ par les poids g, qui sont les ratios des poids calés aux poids de sondage originaux :

$$\mathbf{g} = \mathbf{1}_{(n\times 1)} + \mathbf{M}\left(\mathbf{M}^T \mathbf{D}^A \mathbf{M}\right)^{-1}\left(\mathbf{T}^M - \left(\mathbf{d}^A\right)^T \mathbf{M}\right)^T$$

$$v.g_{\mathcal{A}}\left(\hat{T}_y^{\mathrm{LASSO}}\right) = \sum_{i\in s_A}\left(\frac{g_i\left(y_i - \hat{\mu}_i\hat{B}^{\mathrm{MC}}\right)}{\pi_i}\right)^2 (1-\pi_i) + \sum_{i\in s_A}\sum_{j\neq i}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{g_i\left(y_i - \hat{\mu}_i\hat{B}^{\mathrm{MC}}\right)}{\pi_i}\frac{g_j\left(y_j - \hat{\mu}_j\hat{B}^{\mathrm{MC}}\right)}{\pi_j}. \quad (3.8)$$

Pour simplifier la notation, nous représentons $v_{\mathcal{A}}(\hat{T}_y^{\mathrm{LASSO}})$ par v^{LASSO} et $v.g_{\mathcal{A}}(\hat{T}_y^{\mathrm{LASSO}})$ par v_g^{LASSO}.

4 Étude en simulation

Nous concevons une simulation pour évaluer les propriétés en échantillon fini de $\hat{T}_y^{\mathrm{LASSO}}$ et les estimations asymptotiques de la variance de $\hat{T}_y^{\mathrm{LASSO}}$, v^{LASSO} et v_g^{LASSO}. Nous considérons aussi un estimateur bootstrap naïf $v_{\mathrm{boot}}^{\mathrm{LASSO}}$, obtenu en tirant avec remise 500 échantillons de chaque échantillon simulé, en tant qu'estimateur de rechange de la variance de $\hat{T}_y^{\mathrm{LASSO}}$.

Afin de simuler des échantillons non probabilistes, nous générons des échantillons avec probabilités de sélection inégales, mais fixons les poids de sondage à $\mathbf{d}^A = N/n$. Nous considérons aussi $\hat{T}_y^{\mathrm{GREG}}$ (estimateur par calage classique) et $\hat{T}_y^{\mathrm{HT}}$ (estimateur de Horvitz-Thompson purement fondé sur le plan de sondage). Puisque $\hat{T}_y^{\mathrm{LASSO}}$ effectue à la fois la sélection de variables et l'estimation, nous appliquons une sélection pas à pas descendante (*backward stepwise selection*) pour sélectionner le modèle de travail pour la régression GREG. Bien qu'il n'existe aucune justification théorique de l'utilisation de la sélection pas-à-pas des variables, Skinner et Silva (1997) ont montré que, étant donné deux variables auxiliaires, une procédure pas à pas peut aboutir à une plus grande efficacité de l'estimateur GREG. Nous voulons connaître la performance de chaque estimateur sous 1) des populations dont les rapports signal-bruit (RSB) diffèrent, 2) des plans d'échantillonnage indépendants, informatifs et biaisés, et 3) des petites et grandes tailles d'échantillon. Le ratio signal-bruit est calculé selon les définitions données dans Czanner, Sarma, Eden et Brown (2008). Nous prenons deux niveaux de corrélation (faible/forte) entre les covariables, recoupés par deux niveaux de taille de l'effet (petite/grande) des covariables. Nous configurons les populations faible/grande et forte/petite de manière qu'elles aient le même RSB, afin de comprendre l'influence de la corrélation et de la taille de l'effet sur la performance de l'estimateur, étant donné le même RSB. Trois plans d'échantillonnage sont utilisés pour tirer les échantillons, à savoir l'échantillonnage aléatoire simple sans

remise, l'échantillonnage aléatoire simple (EAS), l'échantillonnage de Poisson avec probabilités de sélection proportionnelles aux covariables, $\text{POI}(\text{X})$, et l'échantillonnage de Poisson avec probabilités de sélection proportionnelles aux covariables et au résultat, $\text{POI}(\text{X+Y})$. L'échantillonnage $\text{POI}(\text{X+Y})$ simule le biais d'autosélection des échantillons non probabilistes, où la propension d'un répondant à participer à une étude est reliée à la variable d'analyse. Nous considérons deux tailles d'échantillon, soit 250 et 1 000. Donc, nous avons un total de $2 \times 2 \times 3 \times 2 = 24$ groupes expérimentaux.

4.1 Population

Pour créer une collinéarité entre les covariables, nous suivons une structure d'autocorrélation décroissante fréquemment utilisée dans les simulations liées au LASSO (Tibshirani, 1996) : $\text{cor}(X_i, X_j) = \rho^{|i-j|}$, $i = 1, \ldots, p$. Nous générons une population de taille $N = 100\,000$ à partir d'une loi normale multivariée de moyenne $\mathbf{0}_{(p\times 1)}$ et de covariance Σ^{ρ}, $p = 40$. La variable de résultat continue est générée par le modèle de régression :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{40} x_{i40} + N(0, 3).$$

La variable de résultat binaire est générée par le modèle de régression logistique :

$$\begin{aligned} \phi_i &= \text{expit}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{40} x_{i40}), \qquad \text{expit}(u) = (1 + \exp(u))^{-1} \\ y_i &= \text{bernoulli}(\phi_i). \end{aligned}$$

Nous prenons $\rho = 0{,}15$ pour la population avec faible corrélation, et $\rho = 0{,}73$ pour la population avec forte corrélation. Pour les variables de résultat continues ainsi que binaires :

$$\begin{aligned} &\text{Petite taille de l'effet} \quad \boldsymbol{\beta}^{(1)} := \beta_{12} \ldots \beta_{19}, \beta_{32} \ldots \beta_{39} = 0{,}45 \\ &\text{Grande taille de l'effet} \quad \boldsymbol{\beta}^{(1)} := \beta_{12} \ldots \beta_{19}, \beta_{32} \ldots \beta_{39} = 0{,}74. \end{aligned}$$

Pour $\mathbf{y}$ continue : $\beta_0 = 1$, pour $\mathbf{y}$ binaire : $\beta_0 = 0{,}4$. Le reste des $\beta_i = 0$. Sur 41 paramètres de régression, 16 sont non nuls et 25 sont nuls.

4.2 Plans d'échantillonnage

Trois plans d'échantillonnage sont utilisés pour générer l'échantillon :

1. Échantillonnage aléatoire simple (EAS) : probabilités de sélection $= n/N$.
2. Échantillonnage de Poisson avec probabilités proportionnelles à $\mathbf{X}$, $\text{POI}(\text{X})$

$$\begin{cases} \mathbf{y} \text{ continue :} & \pi_i \propto 0{,}4 + 0{,}4x_{i5} + 0{,}4x_{i15} + 0{,}4x_{i25} + 0{,}4x_{i35} \\ \mathbf{y} \text{ binaire :} & \text{logit}(\pi_i) = 0{,}4 + 0{,}4x_{i5} + 0{,}4x_{i15} + 0{,}4x_{i25} + 0{,}4x_{i35}. \end{cases}$$

3. Échantillonnage de Poisson avec probabilités proportionnelles à $\mathbf{X}$ et $\mathbf{y}$, $\text{POI}(\text{X+Y})$

$$\begin{cases} \mathbf{y} \text{ continue :} & \pi_i \propto 0{,}4 + 0{,}4x_{i5} + 0{,}4x_{i15} + 0{,}4x_{i25} + 0{,}4x_{i35} + 0{,}5y_i \\ \mathbf{y} \text{ binaire :} & \text{logit}(\pi_i) \propto 0{,}4 + 0{,}4x_{i5} + 0{,}4x_{i15} + 0{,}4x_{i25} + 0{,}4x_{i35} + y_i. \end{cases}$$

4.3 Mesures d'évaluation

Nous évaluons le biais, la variance et la REQM empiriques pour chaque estimateur du total. Nous évaluons les estimations asymptotiques de la variance et les estimations bootstrap de la variance d'après la couverture nominale de 95 % et le biais en pourcentage par rapport à la variance empirique. Nous utilisons l'approximation normale pour générer les intervalles de confiance. Nous calculons le biais en pourcentage comme étant $\text{biais en \%} = 100[v - \text{var}(\hat{T}_y^{\text{LASSO}})] / \text{var}(\hat{T}_y^{\text{LASSO}})$, où $\text{var}(\hat{T}_y^{\text{LASSO}})$ est la variance empirique obtenue à partir des échantillons simulés.

4.4 Résultats des simulations

Les résultats des simulations portent sur $S = 1\,000$ échantillons simulés par groupe expérimental. Le tableau 4.1 donne les résultats numériques pour le biais, la variance et la racine carrée de l'erreur quadratique moyenne de chaque estimateur sous différents plans expérimentaux en vue d'estimer le total d'une variable de résultat continue. Le tableau 4.2 donne les résultats numériques pour l'estimation du total d'une variable de résultat binaire.

4.4.1 Racine carrée de l'erreur quadratique moyenne

Sous EAS, tous les estimateurs sont sans biais, et les estimateurs LASSO et GREG donnent des résultats approximativement aussi bon que l'estimateur HT. Les plans $\text{POI}(\text{X})$ et $\text{POI}(\text{X+Y})$ donnent des échantillons biaisés en raison de la sélection avec probabilités plus élevées de cas présentant de plus grandes valeurs de covariables. Sous $\text{POI}(\text{X+Y})$, la sélection favorise aussi les cas présentant de plus grandes valeurs de la variable de résultat. Le biais absolu de l'estimateur LASSO diminue relativement à celui de l'estimateur GREG à mesure qu'augmente le RSB. Cette amélioration est plus spectaculaire dans le cas binaire que dans le cas continu, surtout pour $\text{POI}(\text{X+Y})$. Pour ce qui est de la REQM, l'estimateur LASSO offre une amélioration insignifiante par rapport à l'estimateur GREG pour l'estimation des totaux de variables de résultat continues. L'amélioration est légèrement notable, environ 3 %, lorsque le modèle contient des variables explicatives fortement corrélées. Dans le cas de variables binaires, l'estimateur LASSO donne lieu à une amélioration importante de l'EQM par rapport à l'estimateur GREG à mesure qu'augmente le RSB, la réduction étant de 20 % pour le plan $\text{POI}(\text{X})$ et de près de 50 % pour le plan $\text{POI}(\text{X+Y})$ quand le RSB est grand. En particulier, sous les types de population faible/grande et forte/petite, le RSB est le même, de sorte que la différence de performance entre les estimateurs LASSO et GREG est attribuée à la corrélation ou à la taille de l'effet. Le LASSO donne de meilleurs résultats en ce qui concerne le biais et la REQM pour le type de population forte/petite, ce qui donne à penser que l'avantage de l'estimateur LASSO par rapport à l'estimateur GREG est plus important lorsque le modèle contient des variables explicatives fortement corrélées. Il semble donc qu'en présence d'une multicollinéarité, la capacité de sélectionner les variables du LASSO est meilleure que celle de la procédure pas-à-pas utilisée dans la méthode GREG.

Tableau 4.1
Résumé des simulations pour un résultat continu : total, biais et REQM × 10^3; variance × 10^6

Population	n	Plan d'échantillonnage	HT			GREG			LASSO		
			biais	var	reqm	biais	var	reqm	biais	var	reqm
Faible/petite T = 100,8 RSB = 0,47	250	EAS	0,5	546	23,3	0,9	425	20,6	0,9	428	20,7
		POI(X)	12,4	525	26,0	-0,6	446	21,1	-0,4	441	21,0
		POI(X+Y)	19,4	519	29,9	4,6	443	21,5	4,7	431	21,3
	1 000	EAS	0,2	129	11,4	0,3	94	9,6	0,3	94	9,7
		POI(X)	12,6	129	17,0	-0,1	91	9,5	-0,2	92	9,6
		POI(X+Y)	19,7	128	22,7	4,9	91	10,7	5,0	91	10,7
Faible/grande T = 101,4 RSB = 1,26	250	EAS	0,4	849	29,1	0,9	415	20,4	1,0	417	20,4
		POI(X)	21,1	818	35,6	-1,3	434	20,9	-1,0	432	20,8
		POI(X+Y)	31,7	817	42,7	3,7	427	21,0	4,0	427	21,1
	1 000	EAS	0,0	200	14,1	0,3	94	10,0	0,3	93	9,7
		POI(X)	21,1	199	25,4	-0,1	91	9,6	-0,2	90	9,6
		POI(X+Y)	31,7	196	34,6	4,9	91	10,7	4,8	89	10,6
Forte/petite T = 101,8 RSB = 1,26	250	EAS	0,1	941	30,7	1,0	421	20,6	1,0	399	20,0
		POI(X)	50,2	895	58,5	-0,7	434	20,8	-1,6	402	20,1
		POI(X+Y)	57,8	872	64,9	4,0	435	21,2	3,0	399	20,2
	1 000	EAS	0,0	218	14,8	0,3	94	9,7	0,3	93	9,6
		POI(X)	50,6	210	53,0	-0,1	93	9,7	-0,5	91	9,6
		POI(X+Y)	58,2	209	59,9	4,7	95	10,8	4,2	92	10,5
Forte/grande T = 103,1 RSB = 3,41	250	EAS	-0,4	1 897	43,6	0,8	436	20,9	1,0	407	20,2
		POI(X)	83,3	1 826	93,7	-0,8	435	20,9	-1,5	406	20,2
		POI(X+Y)	96,4	1 779	105,3	3,7	428	21,0	3,0	404	20,3
	1 000	EAS	-0,2	444	21,0	0,3	93	9,7	0,3	93	9,7
		POI(X)	83,6	424	86,1	-0,2	93	9,7	-0,5	91	9,6
		POI(X+Y)	96,9	423	99,0	4,4	94	10,6	4,1	92	10,4

Tableau 4.2
Résumé des simulations pour un résultat binaire : total, biais et REQM × 10^3; variance × 10^6

Population	n	Plan d'échantillonnage	HT			GREG			LASSO		
			biais	var	reqm	biais	var	reqm	biais	var	reqm
Faible/petite T = 56,2 RSB = 0,51	250	EAS	0,0	10,2	3,2	0,0	7,2	2,7	0,0	7,0	2,7
		POI(X)	2,6	10,0	4,1	0,2	8,0	2,8	0,1	7,8	2,8
		POI(X+Y)	4,9	9,8	5,8	2,0	8,1	3,5	1,8	7,8	3,3
	1 000	EAS	-0,0	2,7	1,6	0,0	1,7	1,3	0,0	1,6	1,3
		POI(X)	2,5	2,4	2,9	0,0	1,8	1,3	-0,0	1,7	1,3
		POI(X+Y)	4,7	2,3	5,0	1,8	1,8	2,2	1,6	1,7	2,1
Faible/grande T = 54,4 RSB = 1,10	250	EAS	-0,0	10,8	3,3	0,0	6,1	2,5	0,1	5,4	2,3
		POI(X)	3,0	10,2	4,4	0,1	6,1	2,5	0,1	5,8	2,4
		POI(X+Y)	5,3	9,8	6,2	1,6	6,2	2,9	1,3	5,8	2,8
	1 000	EAS	-0,0	2,7	1,6	0,0	1,3	1,1	0,0	1,1	1,0
		POI(X)	2,9	2,4	3,3	0,0	1,4	1,2	-0,1	1,2	1,1
		POI(X+Y)	5,2	2,2	5,4	1,4	1,4	1,8	1,1	1,2	1,6
Forte/petite T = 54,2 RSB = 1,10	250	EAS	-0,0	10,3	3,2	0,0	5,8	2,4	0,1	4,9	2,2
		POI(X)	6,6	9,6	7,3	0,3	6,2	2,5	-0,2	4,8	2,2
		POI(X+Y)	8,6	9,3	9,1	1,8	6,3	3,1	0,9	4,9	2,4
	1 000	EAS	-0,0	2,5	1,6	0,0	1,2	1,1	0,0	1,0	1,0
		POI(X)	6,6	2,2	6,7	0,2	1,4	1,2	-0,2	1,1	1,1
		POI(X+Y)	8,5	2,1	8,7	1,6	1,4	2,0	1,0	1,0	1,4
Forte/grande T = 52,8 RSB = 2,75	250	EAS	-0,1	10,2	3,1	-0,0	5,2	2,3	0,1	3,8	1,9
		POI(X)	7,1	9,8	7,8	0,3	5,7	2,4	-0,2	3,6	1,9
		POI(X+Y)	9,1	9,4	9,6	1,5	5,7	2,8	0,5	3,7	2,0
	1 000	EAS	-0,1	2,5	1,6	-0,0	1,1	1,0	0,0	0,6	0,8
		POI(X)	7,1	2,2	7,2	0,2	1,3	1,1	-0,2	0,7	0,9
		POI(X+Y)	9,1	2,2	9,2	1,4	1,2	1,8	0,5	0,7	1,0

4.4.2 Estimations de la variance de l'estimateur LASSO

Les tableaux 4.3 et 4.4 donnent les résultats pour la couverture nominale de 95 % des IC et le biais en pourcentage pour les deux estimateurs de variance asymptotiques de forme analytique établis dans le cadre de la présente étude, ainsi que l'estimateur de variance bootstrap naïf pour l'estimateur par calage avec le LASSO.

Pour les variables de résultat continues, la couverture des variances bootstrap est systématiquement proche de 95 % sous les plans d'échantillonnage EAS et $\mathrm{POI}(\mathrm{X})$ pour les deux tailles d'échantillon. Sous le plan $\mathrm{POI}(\mathrm{X+Y})$, on note une sous-couverture très modeste dans le tableau 4.3. La couverture des variances de forme analytique est sensible à la taille d'échantillon ainsi qu'au plan d'échantillonnage, les petits échantillons ayant tendance à produire une sous-couverture, particulièrement sous le plan $\mathrm{POI}(\mathrm{X+Y})$. La différence de couverture des estimations de la variance entre les échantillons de petite et de grande taille est prévisible, puisque les estimations de la variance sont asymptotiques et s'améliorent pour les grands échantillons. Pour ce qui est du biais des estimateurs de variance, il apparaît qu'il diminue quand le RSB augmente. Pour un même RSB, tant les variances asymptotiques de forme analytique que les variances bootstrap ont un plus petit biais, étant donné des variables explicatives avec forte corrélation relativement à des variables explicatives avec grande taille de l'effet. Les variances de forme analytique ont tendance à sous-estimer la variance empirique, surtout quand la taille d'échantillon est petite. Dans l'ensemble, la différence est faible entre les deux estimations de variance de forme analytique. La variance bootstrap a tendance à surestimer la variance empirique, mais le biais absolu est généralement plus faible que celui des estimations de variance de forme analytique.

Pour les variables de résultat binaires, tant les estimations asymptotiques de variance de forme analytique que les estimations de variance bootstrap sont sensibles à la taille de l'échantillon, au plan d'échantillonnage et au RSB. Pour la variance bootstrap, la couverture est systématiquement proche de 95 % sous les plans EAS et $\mathrm{POI}(\mathrm{X})$ pour les deux tailles d'échantillon et tous les types de population, mais elle varie de 75 % à 94 % sous le plan $\mathrm{POI}(\mathrm{X+Y})$. Sous ce plan $\mathrm{POI}(\mathrm{X+Y})$, la couverture de la variance bootstrap est meilleure pour la taille d'échantillon de 250 que pour celle de 1 000 quand le biais représente une plus grande part de la REQM, et meilleure pour les populations avec forte corrélation que pour celles avec faible corrélation. Pour ce qui est de la couverture, les variances de forme analytique présentent une tendance similaire aux variances bootstrap sous le plan $\mathrm{POI}(\mathrm{X+Y})$, c'est-à-dire une meilleure couverture pour les petites que pour les grandes tailles d'échantillon, et une meilleure couverture pour les populations avec forte corrélation que pour les populations avec faible corrélation. Sous EAS et $\mathrm{POI}(\mathrm{X})$, la couverture pour les variances de forme analytique s'améliore à mesure qu'augmente la taille de l'échantillon. En ce qui concerne le biais, tant la variance bootstrap que les variances de forme analytique ont un biais plus faible pour les plus grandes tailles d'échantillon. Si la taille d'échantillon est maintenue constante, le biais des estimations de variance de forme analytique augmente quand le RSB augmente. La même tendance ne s'observe pas pour les estimations de variance bootstrap. Comme pour les variables de résultat continues, les estimateurs de variance de forme analytique ont tendance à sous-estimer la variance empirique, surtout quand la taille

d'échantillon est petite. Contrairement aux constatations pour les variables de résultat continues, il s'avère que les estimations de variance de forme analytique pondérées par les poids g ont de meilleures propriétés de biais que les estimations de variance de forme analytique non pondérées. La variance bootstrap a tendance à surestimer la variance empirique. Cependant, les biais sont nettement plus petits que pour les estimations de variance de forme analytique.

Tableau 4.3
Couverture nominale de 95 % et biais en % des estimations de variance pour le LASSO

Variable de résultat continue			Couverture			Biais en %		
Population	**n**	**Plan**	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$
Faible/petite	250	EAS	91,7 %	91,8 %	95,4 %	-22,6 %	-22,3 %	2,9 %
		POI(X)	91,2 %	91,2 %	96,1 %	-25,1 %	-24,5 %	5,7 %
		POI(X+Y)	89,6 %	89,9 %	95,4 %	-23,5 %	-22,8 %	7,9 %
	1 000	EAS	93,2 %	93,2 %	93,8 %	-7,3 %	-7,2 %	-0,3 %
		POI(X)	94,0 %	93,9 %	95,5 %	-5,7 %	-5,3 %	6,6 %
		POI(X+Y)	90,0 %	90,1 %	92,1 %	-4,9 %	-4,4 %	7,9 %
Faible/grande	250	EAS	91,5 %	91,5 %	95,7 %	-22,6 %	-22,3 %	6,2 %
		POI(X)	90,9 %	91,2 %	96,4 %	-25,4 %	-24,9 %	8,8 %
		POI(X+Y)	90,0 %	90,2 %	95,1 %	-24,5 %	-23,7 %	9,9 %
	1 000	EAS	93,4 %	93,5 %	94,3 %	-6,6 %	-6,5 %	-0,1 %
		POI(X)	94,1 %	94,2 %	95,9 %	-4,0 %	-3,5 %	7,6 %
		POI(X+Y)	90,7 %	90,7 %	92,7 %	-2,9 %	-2,3 %	9,6 %
Forte/petite	250	EAS	92,3 %	92,2 %	95,4 %	-17,4 %	-17,1 %	2,0 %
		POI(X)	92,5 %	92,6 %	95,8 %	-17,9 %	-16,1 %	6,4 %
		POI(X+Y)	91,2 %	91,8 %	96,5 %	-17,4 %	-15,4 %	7,1 %
	1 000	EAS	93,5 %	93,5 %	94,4 %	-6,5 %	-6,4 %	-0,9 %
		POI(X)	94,1 %	94,0 %	95,4 %	-5,0 %	-3,1 %	5,7 %
		POI(X+Y)	91,9 %	92,3 %	93,4 %	-6,0 %	-3,9 %	5,0 %
Forte/grande	250	EAS	92,3 %	92,3 %	95,2 %	-19,6 %	-19,3 %	2,2 %
		POI(X)	92,0 %	92,3 %	96,1 %	-19,6 %	-17,8 %	7,4 %
		POI(X+Y)	91,2 %	91,8 %	95,6 %	-19,1 %	-16,9 %	8,3 %
	1 000	EAS	93,4 %	93,4 %	94,5 %	-6,5 %	-6,4 %	-0,7 %
		POI(X)	94,0 %	94,5 %	95,6 %	-4,7 %	-2,8 %	6,7 %
		POI(X+Y)	92,2 %	92,4 %	93,4 %	-5,6 %	-3,3 %	6,1 %

Tableau 4.4
Couverture nominale de 95 % et biais en % des estimations de variance pour le LASSO

Variable de résultat binaire			Couverture			Biais en %		
Population	**n**	**Plan**	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$
Faible/petite	250	EAS	89,8 %	90,0 %	95,9 %	-28,1 %	-27,8 %	9,2 %
		POI(X)	88,1 %	88,6 %	96,7 %	-37,3 %	-35,3 %	9,2 %
		POI(X+Y)	79,0 %	79,9 %	91,2 %	-38,7 %	-35,9 %	8,0 %
	1 000	EAS	92,8 %	92,8 %	93,5 %	-11,9 %	-11,8 %	-3,5 %
		POI(X)	92,0 %	92,8 %	95,7 %	-17,9 %	-15,5 %	1,0 %
		POI(X+Y)	68,6 %	69,6 %	74,6 %	-18,5 %	-14,9 %	0,5 %
Faible/grande	250	EAS	86,8 %	87,0 %	94,9 %	-37,7 %	-37,3 %	11,3 %
		POI(X)	85,4 %	86,1 %	95,5 %	-42,9 %	-41,2 %	14,4 %
		POI(X+Y)	78,7 %	80,1 %	92,6 %	-44,0 %	-41,3 %	14,4 %
	1 000	EAS	94,4 %	94,3 %	95,2 %	-5,5 %	-5,4 %	5,8 %
		POI(X)	91,8 %	92,1 %	94,9 %	-20,5 %	-18,6 %	-1,8 %
		POI(X+Y)	76,8 %	77,8 %	82,9 %	-20,4 %	-16,9 %	-1,3 %

Tableau 4.4 (suite)
Couverture nominale de 95 % et biais en % des estimations de variance pour le LASSO

Variable de résultat binaire			Couverture			Biais en %		
Population	**n**	**Plan**	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$
Forte/petite	250	EAS	89,2 %	89,1 %	94,4 %	-28,5 %	-28,1 %	0,4 %
		POI(X)	89,0 %	90,1 %	95,5 %	-31,9 %	-25,3 %	12,7 %
		POI(X+Y)	85,7 %	88,4 %	93,8 %	-33,9 %	-25,4 %	10,9 %
	1 000	EAS	93,9 %	93,9 %	95,6 %	-6,3 %	-6,2 %	3,5 %
		POI(X)	92,6 %	93,4 %	94,8 %	-16,5 %	-9,2 %	1,9 %
		POI(X+Y)	83,3 %	85,4 %	88,1 %	-15,0 %	-5,0 %	5,2 %
Forte/grande	250	EAS	82,8 %	82,8 %	93,8 %	-44,6 %	-44,3 %	-6,4 %
		POI(X)	83,6 %	85,5 %	95,1 %	-44,3 %	-39,4 %	3,8 %
		POI(X+Y)	82,9 %	85,1 %	93,8 %	-45,1 %	-38,4 %	4,6 %
	1 000	EAS	94,3 %	94,4 %	96,1 %	-7,8 %	-7,6 %	6,3 %
		POI(X)	91,3 %	92,2 %	94,0 %	-20,0 %	-13,8 %	0,2 %
		POI(X+Y)	86,3 %	88,6 %	91,5 %	-18,1 %	-9,2 %	2,8 %

5 Application à la National Health Interview Survey (NHIS)

5.1 Données de la NHIS et l'ACS

Nous appliquons maintenant le calage avec le LASSO à la *National Health Interview Survey* (NHIS) de 2013 pour estimer le nombre total d'adultes (de 18 ans et plus) ayant reçu un diagnostic de cancer dans la population. La NHIS est réalisée auprès d'un échantillon représentatif de la population nationale de ménages de civils ne vivant pas en établissement tiré selon un plan d'échantillonnage aréolaire probabiliste à plusieurs degrés (*Centers for Disease Control and Prevention*, 2005). Chaque mois, des données liées à la santé sont recueillies par interviews sur place auprès d'un échantillon transversal de personnes appartenant aux ménages sélectionnés. Les données fournissent des pseudo unités primaires d'échantillonnage (UPE), des pseudo strates et des poids d'échantillonnage qui permettent de produire des estimations pondérées sous un plan de sondage complexe. En plus de mesures liées à la santé, la NHIS recueille aussi des données démographiques. Notre objectif est d'évaluer notre estimateur LASSO en traitant l'échantillon non pondéré de la NHIS comme représentant un échantillon non probabiliste et d'étudier comment les estimateurs par calage GREG et LASSO se comparent à l'estimateur pondéré selon le plan de sondage.

Pour caler les données de la NHIS sur un ensemble de variables démographiques et liées au revenu, nous utilisons les microdonnées de l'*American Community Survey* (ACS) de 2013 comme données de référence. L'échantillon de l'ACS est constitué de ménages sélectionnés par échantillonnage aréolaire probabiliste à plusieurs degrés dans 3 143 comtés des États-Unis. L'ACS est conçue pour améliorer les estimations sur petits domaines produites entre les échantillons pour le questionnaire détaillé du recensement décennal. Environ trois millions de ménages sont sélectionnés chaque année, et des mesures sont recueillies sur les types de ménages, ainsi que les caractéristiques démographiques des membres des ménages. L'ACS recueille aussi des données sur les logements collectifs, qui sont exclues de la présente analyse. Pour l'ACS de 2013, la taille de l'échantillon d'adultes était de 2 317 301. La taille de l'échantillon de la NHIS de 2013 était de 34 201 après élimination de 242 cas pour lesquels des valeurs manquaient pour les variables

démographiques. Pour les besoins de la présente analyse, nous traitons les estimations pondérées provenant de l'ACS comme les totaux connus de population, une hypothèse raisonnable étant donné les différences de taille d'échantillon.

5.2 Estimateurs

La variable de résultat étudiée est celle de savoir si une personne a reçu ou non un diagnostic de cancer. Définissons l'indicateur binaire pour la variable de résultat :

$$y_i = \begin{cases} 1 : \text{ si la personne } i \text{ a reçu un diagnostic de cancer} \\ 0 : \text{ autrement.} \end{cases}$$

Nous commençons par utiliser les poids de sondage de la NHIS de 2013, $\mathbf{w}^{\text{NHIS}}$, et les variables du plan de sondage pour obtenir une estimation sans biais du total de population $T_y = \sum_{i=1}^{N} y_i$. Puis, nous supposons que l'échantillon de la NHIS de 2013 est tiré selon un plan d'échantillonnage aléatoire simple, avec les poids de sondage initiaux $\mathbf{d}^A = N/n$, où N est le total de population obtenu d'après l'ACS, et n est la taille d'échantillon de la NHIS. Nous calons les poids de sondage $\mathbf{d}^A$ sur un ensemble de variables démographiques et de variables de revenu par calage GREG classique et par calage avec le LASSO. Enfin, à titre de compromis entre les méthodes GREG et LASSO, nous considérons le calage assisté par un modèle sur un modèle linéaire pour y_i au lieu du LASSO en utilisant (2.7); notons que, quand $\hat{\mu}_i$ est calculé en utilisant le même modèle linéaire que dans l'estimateur GREG, les estimations ponctuelles du total correspondent, même si les poids de calage diffèrent. Donc, nous produisons sept estimations :

1. $\hat{T}_y^{\text{NHIS}} = \sum_{i \in s_A} w_i^{\text{NHIS}} y_i$: estimation obtenue avec les poids de la NHIS.
2. $\hat{T}_y^{\text{HTEAS}} = \sum_{i \in s_A} (N/n)\, y_i$: estimation obtenue avec les poids $\mathbf{d}^A = N/n$.
3. $\hat{T}_y^{\text{GREG1}} = \sum_{i \in s_A} w_i^{\text{GREG1}} y_i$: estimation obtenue par calage de $\mathbf{d}^A$ par la régression généralisée (GREG) en utilisant toutes les variables de calage.
4. $\hat{t}_y^{\text{GREG1MC}} = \sum_{i \in s_A} w_i^{\text{GREG1MC}} y_i$: estimation obtenue par calage assisté par un modèle (MC) sur un modèle linéaire utilisant les variables explicatives dans GREG1.
5. $\hat{T}_y^{\text{GREG2}} = \sum_{i \in s_A} w_i^{\text{GREG2}} y_i$: estimation obtenue par calage de $\mathbf{d}^A$ par la méthode GREG en utilisant uniquement les variables de calage choisies par sélection pas-à-pas descendante.
6. $\hat{t}_y^{\text{GREG2MC}} = \sum_{i \in s_A} w_i^{\text{GREG2MC}} y_i$: estimation obtenue par calage assisté par un modèle (MC) sur un modèle linéaire utilisant les variables explicatives dans GREG2.
7. $\hat{T}_y^{\text{LASSO}} = \sum_{i \in s_A} w_i^{\text{LASSO}} y_i$: estimation obtenue par calage assisté par un modèle en utilisant le LASSO.

La variance de $\hat{T}_y^{\text{NHIS}}$ est l'estimation de la variance du total par linéarisation, tenant compte de la strate d'échantillonnage, des unités primaires d'échantillonnage et des poids de sondage dans l'échantillon de la NHIS de 2013. Les variances des estimateurs HTEAS, GREG1 et GREG2 sont les estimations de la variance

par linéarisation avec les poids $\mathbf{d}^A$, $\mathbf{w}^{\text{GREG1}}$ et $\mathbf{w}^{\text{GREG2}}$, respectivement. Nous obtenons la variance de l'estimateur LASSO par le bootstrap naïf.

5.3 Modèles de travail

Le tableau 5.1 donne les noms, les étiquettes, les valeurs et les répartitions des variables de calage utilisées dans la présente analyse. La première colonne donne la répartition non pondérée des variables dans l'échantillon de la NHIS. La deuxième colonne contient les répartitions des variables dans l'échantillon de la NHIS, pondérées par les poids au niveau de la personne $\mathbf{w}^{\text{NHIS}}$. La troisième colonne donne la répartition des variables dans la population obtenue à partir des données de référence de l'ACS. La catégorie « Données manquantes » pour le revenu est incluse en tant que catégorie distincte pour saisir la différence entre les profils de valeurs manquantes dans la NHIS et l'ACS. L'ajout d'une catégorie « Données manquantes » nous permet aussi de maintenir la taille de l'échantillon analytique. Comparativement à l'ACS, l'échantillon de la NHIS non pondéré contient de plus fortes proportions de femmes et de personnes veuves/divorcées/séparées, et une plus faible proportion de personnes blanches non hispaniques. Après pondération, les répartitions de la NHIS selon le sexe et la race sont proches des données de référence, et seules les catégories de l'état matrimonial présentent certaines différences.

Nous utilisons un modèle linéaire non pondéré avec sélection pas-à-pas descendante des variables pour déterminer le modèle de travail pour GREG2. Les variables finales incluses dans le modèle pour GREG2 sont l'âge, le niveau d'études, la race, la situation d'emploi (oui/non) et le revenu familial. Pour le calage GREG classique et le calage avec le LASSO, nous utilisons toutes les variables disponibles.

Tableau 5.1
Variables de calage

			NHIS	ACS
		Pas de poids	Poids au niveau de la personne	Poids au niveau de la personne
Région	Nord-est	16 %	18 %	18 %
	Midwest	20 %	23 %	21 %
	Sud	37 %	37 %	37 %
	Ouest	26 %	23 %	23 %
Âge	18 à 29 ans	19 %	21 %	21 %
	30 à 39 ans	17 %	17 %	17 %
	40 à 49 ans	16 %	18 %	18 %
	50 à 59 ans	17 %	18 %	18 %
	60 à 69 ans	15 %	14 %	14 %
	70 à 79 ans	9 %	8 %	8 %
	80 ans et plus	6 %	4 %	5 %
Sexe	Hommes	45 %	48 %	48 %
	Femmes	55 %	52 %	52 %
Niveau d'études	Inférieur au diplôme d'études secondaires	16 %	14 %	13 %
	Diplôme d'études secondaires ou moins élevé	26 %	26 %	28 %
	Études collégiales partielles	20 %	20 %	23 %
	Diplôme d'études collégiales	29 %	30 %	25 %
	Études de cycle supérieur	10 %	10 %	10 %
Race/ethnicité	Blanche non hispanique	60 %	66 %	66 %
	Noire non hispanique	15 %	12 %	12 %
	Hispanique	17 %	15 %	15 %
	Autre	8 %	7 %	7 %

Tableau 5.1 (suite)
Variables de calage

			NHIS	ACS
		Pas de poids	Poids au niveau de la personne	Poids au niveau de la personne
État matrimonial	Marié(e)/partenariat	49 %	60 %	52 %
	Veuf(ve)/divorcé(e)/séparé(e)	27 %	18 %	20 %
	Jamais marié(e)	24 %	22 %	28 %
A un emploi	Oui	35 %	33 %	39 %
	Non	65 %	67 %	61 %
Revenu	1er quartile	22 %	15 %	19 %
	2e quartile	20 %	17 %	20 %
	3e quartile	21 %	22 %	20 %
	4e quartile	21 %	28 %	19 %
	Données manquantes	17 %	19 %	22 %

5.4 Résultats

Le tableau 5.2 donne les estimations, les erreurs-types (e.-t.), la racine carrée de l'erreur quadratique moyenne en traitant la valeur de la NHIS correctement pondérée comme étant la vraie valeur (REQM), l'écart en pourcentage par rapport à l'estimation de la NHIS : écart en % = $100\,(\hat{T} - \hat{T}_y^{\text{NHIS}}) / \hat{T}_y^{\text{NHIS}}$ et l'écart-type, ainsi que le minimum et le maximum, des poids associés à un estimateur donné. Nous traitons l'estimation de la NHIS comme étant l'estimation sans biais parce qu'elle est calculée avec les poids de sondage probabilistes fournis par la NHIS. Sans ajustement de la pondération, l'estimateur HTEAS présente un biais positif de 5,9 %. L'estimateur GREG2 réduit ce biais, le faisant passer de 5,9 % à 2,0 %, l'estimateur GREG1 réduit le biais à 1,8 %, tandis que l'estimateur LASSO réduit le biais à 0,9 %. Par définition, l'emploi de l'estimateur assisté par un modèle utilisant des variables explicatives linéaires donnera le même estimateur que le modèle GREG; cependant, la variabilité est réduite considérablement. Dans la présente analyse, si l'échantillon de la NHIS était non probabiliste, sans ajustement de la pondération, nous aurions surestimé de 1,18 million le nombre d'adultes atteints d'un cancer. Sous le calage classique, l'erreur est réduite à une surestimation de 365 000 (sans sélection des variables) ou de 392 000 (avec sélection des variables). Le calage avec le LASSO réduit encore davantage la surestimation qui se chiffre à 175 000.

Tableau 5.2
Résultats pour l'estimation du nombre total de personnes atteintes d'un cancer. L'écart en % est la différence par rapport à l'estimation de la NHIS divisée par l'estimation de la NHIS

Estimateur	$\hat{T}$	e.-t.	REQM	Écart en % par rapport à la NHIS	É.-T. (min, max) des poids
NHIS	19 889 327	492 263	492 263	0,00 %	5 913 (168; 93 244)
HTEAS	21 070 498	362 883	1 235 657	5,94 %	0 (6 866; 6 866)
GREG1	20 254 449	375 064	523 438	1,84 %	2 474 (-2 409; 16 679)
GREG1 MC	20 254 449	349 100	505 158	1,84 %	269 (6 181; 7 326)
GREG2	20 281 603	367 900	537 802	1,97 %	2 039 (-626; 13 947)
GREG2 MC	20 281 603	349 552	525 421	1,97 %	260 (6 215; 7 291)
LASSO	20 064 671	347 586	389 309	0,88 %	323 (5 786; 7 168)

Comme prévu, l'estimation de la NHIS présente la plus grande erreur-type, car elle intègre convenablement le plan de sondage complexe. Si le modèle de calage de travail saisit correctement la relation entre la variable de résultat et les variables de calage, nous nous attendons à ce que les erreurs-types de l'estimateur par calage soient plus petites que celles de l'estimateur HTEAS. Ce n'est le cas pour aucun des deux estimateurs GREG, pour lesquels l'erreur-type est plus grande que celle de l'estimateur HTEAS, quoique la REQM soit plus petite en raison de la réduction du biais. En outre, l'erreur-type de l'estimateur GREG classique est environ 2,0 % plus grande que celle de l'estimateur GREG avec sélection descendante des variables, caractéristique qui est compensée par la réduction estimée de 6,6 % du biais (bien que cela ne suffise pas à réduire la REQM); l'utilisation de l'estimateur GREG assisté par un modèle (MC) réduit l'erreur-type et la racine carrée de l'erreur quadratique moyenne d'environ 5 à 7 % et 2 à 3 %, respectivement, comparativement aux estimations GREG classiques. Pour le calage avec le LASSO, nous observons une plus petite erreur-type que pour l'estimateur HTEAS, même avec l'estimation bootstrap de la variance qui a tendance à produire une surestimation. Si l'on n'utilise pas les poids de sondage corrects, le calage avec le LASSO produit l'estimation la plus précise d'un total de population, tout en donnant l'erreur-type la plus petite parmi les estimateurs considérés dans la présente application. Et cela, malgré le fait que la variabilité des poids calés avec le LASSO n'était qu'environ le septième de celle des poids GREG, ce que reflète la plus petite erreur-type de l'estimateur proprement dit et la REQM fortement réduite.

6 Conclusion

Dans le présent article, nous avons élaboré un estimateur par calage avec le LASSO des totaux de population, $\hat{T}_y^{\text{LASSO}}$, étant donné des données auxiliaires de population. Nous avons également calculé des estimations de variance de forme analytique pour $\hat{T}_y^{\text{LASSO}}$. Les résultats de simulation montrent que les estimations ponctuelles sont approximativement sans biais sous échantillonnage aléatoire simple et sous échantillonnage informatif. Pour la sélection d'échantillons qui sont reliés aux variables d'analyse, le LASSO a permis de réduire de manière significative le biais d'échantillon, même sans utiliser les poids de sondage corrects. Le LASSO a tendance à donner de meilleurs résultats que les modèles de travail obtenus par sélection pas à pas quand les covariables sont fortement colinéaires. Pour l'analyse portant sur de nombreuses variables catégoriques, pour lesquelles il existe des corrélations naturelles entre les catégories, l'estimateur par calage avec le LASSO peut donner de meilleurs résultats que les estimateurs par calage classiques, même si les tailles des effets sont petites. L'amélioration est modeste dans le cas des variables continues, mais considérable quand le résultat d'intérêt est binaire, comme en témoignent les simulations et l'exemple portant sur les données de la NHIS. Nous avons montré théoriquement et au moyen de simulations que le calage avec le LASSO est très prometteur pour ce qui est de faire des inférences sans biais sur les totaux de population à partir d'échantillons non probabilistes. Même si les estimations asymptotiques de variance de forme analytique n'ont pas produit une couverture nominale très précise, le bootstrap naïf est une approche de rechange viable. Dans une application en vue d'estimer le total de population des personnes ayant reçu un diagnostic de cancer, sans utiliser les poids de sondage corrects, l'estimateur par calage avec le LASSO a produit l'estimation qui était la plus proche de l'estimation fondée sur les poids de sondage

corrects. De tous les estimateurs pris en considération, l'estimateur par calage avec le LASSO possédait aussi la plus petite erreur-type, quoique l'estimation bootstrap de la variance qui a été utilisée ne tienne pas compte entièrement de la mise en grappe dans la NHIS, ce qui accroît généralement les erreurs-types. L'application montre que le calage avec le LASSO peut générer une inférence pour la population dans le cas d'une variable de résultat particulière, et que l'inférence est à la fois plus exacte et précise que lorsqu'on utilise les estimateurs par calage classiques.

La question qui se pose est celle de savoir quand le calage assisté par un modèle LASSO doit être utilisé au lieu des méthodes de calage classiques telles que la méthode GREG. Les résultats tant théoriques qu'empiriques décrits dans le présent article donnent à penser qu'il n'y a pas grand-chose à perdre en termes d'efficacité statistique lorsqu'on utilise le calage assisté par un modèle LASSO, mais sa mise en œuvre requiert un effort supplémentaire de la part de l'analyste. Bien que nous ne puissions pas donner des seuils précis, notre analyse laisse entendre que cet effort vaut la peine quand a) il existe un grand nombre de variables de calage possibles, b) un bon nombre de ces variables de calage sont vraisemblablement fortement corrélées, et c) la variable de résultat est binaire plutôt que continue. Nous pensons qu'au moins les conditions a) et b) se présenteront de plus en plus probablement dans des contextes non probabilistes, où les ensembles de données administratives pourraient fournir ces types de variables de calage et que des sous-ensembles de données obtenus par divers moyens contiendront la variable d'intérêt.

Même si la mise en œuvre du LASSO est particulièrement commode et rapide, il existe évidemment d'autres méthodes de régression modernes qui pourraient être considérées en plus du LASSO pour élaborer des modèles de régression pénalisée pour la régression assistée par un modèle de grande dimensionnalité, y compris les approches telles que la régression ridge, l'analyse en composantes principales, ou les arbres de régression additifs bayésiens (Chipman, George et McCulloch, 2010). Ces approches offrent l'occasion de poursuivre la recherche dans ce domaine.

Enfin, nous soulignons que les présents travaux ne représentent qu'une partie d'une littérature plus vaste et qui s'enrichit rapidement sur l'inférence à partir de sondages non probabilistes. En plus des travaux de McConville et coll. (2017), l'approche « Mr. P » (*multi-level regression and poststratification* ou MRP) de Wang et coll. (2015) fait aussi appel à des covariables de grande dimensionnalité pour ajuster les échantillons non probabilistes, en utilisant un modèle hiérarchique plutôt qu'une régression pénalisée. La quasi randomisation (Elliott, 2009; Elliott, Resler, Flannagan et Rupp, 2010; Elliott et Valliant, 2017) et l'appariement d'échantillons (Rivers, 2007; Vavreck et Rivers, 2008) fournissent aussi des solutions de rechange s'appuyant sur des données provenant de quantités de population connues ou d'estimations issues d'un échantillonnage probabiliste pour traiter les problèmes de biais de sélection dans les échantillons non probabilistes. Chacune possède des forces et des faiblesses par rapport aux autres et par rapport à l'assistance par un modèle LASSO. L'approche MRP s'appuie sur des hypothèses de distribution qui pourraient améliorer l'efficacité, mais aussi réduire la robustesse, et elle n'est pas simple à mettre en œuvre dans sa forme entièrement bayésienne. La quasi randomisation perd le lien avec une variable de résultat particulière, ce qui rend les poids qu'elle produit d'usage plus général, mais vraisemblablement moins efficaces, tandis que l'appariement d'échantillons requiert une intervention à l'étape du plan de sondage pour échantillonner

des éléments provenant de la base non probabiliste qui concordent avec les éléments provenant de la population, comme l'échantillonnage par quotas. La décision d'utiliser le calage assisté par un modèle LASSO doit être prise dans le contexte de ces compromis.

Remerciements

Les auteurs remercient le professeur Fred M. Feinberg et la chercheuse scientifique associée Sunghee Lee de leur révision et suggestions utiles, ainsi que le rédacteur, le rédacteur associé et trois examinateurs de leurs suggestions qui ont beaucoup contribué à l'amélioration du présent article.

Annexe

Détermination des estimations pour le LASSO adaptatif

En pratique, nous n'observons pas le taux théorique de croissance de λ_n, qui optimise les mesures d'adéquation du modèle telles que l'AIC ou le BIC, à moins que nous n'ayons obtenu de nombreux échantillons de la même population ayant diverses tailles. Étant donné un échantillon, les choix de λ_n et γ dépendent du modélisateur. Dans la mise en œuvre de la fonction *glmnet* en R (Friedman et coll., 2010), une gamme de valeurs de λ_n est déterminée selon le schéma suivant :

1. Prendre $\gamma = 0$.
2. Déterminer $\lambda_n^{\max}$ en trouvant la plus petite valeur de λ_n qui fixe tous les coefficients à 0.
3. Si la taille d'échantillon n est plus grande que le nombre de paramètres dans le modèle de régression, prendre $\lambda_n^{\min} = 0{,}0001\lambda_n^{\max}$. Si la taille d'échantillon n est plus petite que le nombre de paramètres, prendre $\lambda_n^{\min} = 0{,}01\lambda_n^{\max}$ (pour fixer les paramètres à 0 plus rapidement).
4. Générer une grille de valeurs de λ_n, habituellement 100 points espacés également entre $\lambda_n^{\min}$ et $\lambda_n^{\max}$.

La gamme initiale de valeurs de λ_n est déterminée indépendamment de γ. Les choix de γ sont moins dictés par les données. Certains modélisateurs choisissent l'une des valeurs de $\gamma = 0{,}1; 0{,}5; 1; 2$. Ici, nous déterminons (λ_n, γ) par validation croisée comme il suit :

1. Obtenir $\alpha_j = 1 / |\hat{\beta}_j^{\text{EMV}}|$.
2. Déterminer 100 valeurs également espacées de λ_n basées sur l'exécution de *glmnet* en R.
3. Pour chaque paire (λ_n, γ), λ_n provenant de l'étape 2, et $\gamma = 0{,}1; 0{,}5; 1; 2$, scinder les données en cinq parties (folds). Utiliser quatre parties pour obtenir $\hat{\boldsymbol{\beta}}$.
4. Appliquer $\hat{\boldsymbol{\beta}}$ à la dernière partie non utilisée pour estimer $\hat{\boldsymbol{\beta}}$ et calculer une mesure. Pour une variable $\mathbf{y}$ continue, nous calculons l'erreur absolue moyenne (EAM), $\sum_{i \in s_{A(k)}} |\hat{\mu}_i - y_i|$. Pour une variable $\mathbf{y}$ binaire, nous calculons l'aire sous la courbe (ASC) (au moyen de la fonction *glmnet* :: *auc* en R).

5. Calculer la moyenne des cinq mesures pour chaque paire (λ_n, γ), et choisir la paire possédant la meilleure mesure moyenne, c'est-à-dire l'EAM minimum pour une variable $\mathbf{y}$ continue, l'ASC maximum pour une variable $\mathbf{y}$ binaire.

Les estimations des coefficients de la régression LASSO adaptatif sont alors obtenues en résolvant les équations (3.1) ou (3.2) de la section 3.1 sachant la paire (λ_n, γ) sélectionnée. Le code R utilisé pour effectuer la validation croisée est fourni dans le matériel complémentaire en ligne.

Absence de biais asymptotique et variance de l'estimateur par calage assisté par un modèle LASSO d'un total de population

Lemme 1 : *Considérons le modèle de superpopulation :*

$$E_\xi(y_k \mid \mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\beta}), V_\xi(y_k \mid \mathbf{x}_k) = v_k^2 \sigma^2.$$

Soit $\mathbf{B}$ *l'estimation de la quasi-vraisemblance en population finie de* $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$. *Sous les conditions (1) à (5) de la section 3.2, l'estimateur asymptotique assisté par un modèle du total de population est :*

$$\hat{T}_y^{MC} = \sum_{i \in s_A} d_i^A (y_i - \mu_i B^{MC}) + \sum_{i=1}^{N} \mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right) \tag{A.1}$$

où

$$\mu_i = \mu(\mathbf{x}_i, \mathbf{B})$$

$$B^{MC} = \frac{\sum_{i=1}^{N} (\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^{N} (\mu_i - \bar{\mu})^2}.$$

Preuve. La preuve est adaptée et étendue en partant de la preuve du théorème 1 dans Wu et Sitter (2001), avec de légères modifications de la notation pour qu'elle concorde avec celle du présent article. Nous commençons par dériver l'estimateur asymptotique assisté par un modèle pour une moyenne de population, $\hat{\bar{y}}^{\text{MC}} = N^{-1}\hat{T}_y^{\text{MC}}$ *(voir l'équation (2.7)). Étant donné les conditions (2) et (3), le développement en série de Taylor d'ordre deux de* $\mu(\mathbf{x}_i, \hat{\mathbf{B}})$ *autour de* $\mathbf{B}$ *est :*

$$\mu(\mathbf{x}_i, \hat{\mathbf{B}}) = \mu(\mathbf{x}_i, \mathbf{B}) + \left\{\frac{\partial \mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}}\right\}^T (\hat{\mathbf{B}} - \mathbf{B}) + (\hat{\mathbf{B}} - \mathbf{B})^T \left\{\frac{\partial^2 \mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t} \partial \mathbf{t}^T}\Big|_{\mathbf{t}=\mathbf{B}^*}\right\}(\hat{\mathbf{B}} - \mathbf{B}) \tag{A.2}$$

pour $\mathbf{B}^* \in (\hat{\mathbf{B}}, \mathbf{B})$ *ou* $(\mathbf{B}, \hat{\mathbf{B}})$. *Soit*

$$\mathbf{h}(\mathbf{x}_i, \mathbf{B}) = \frac{\partial \mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t}}\Big|_{\mathbf{t}=\mathbf{B}}$$

$$\mathbf{k}(\mathbf{x}_i, \mathbf{B}^*) = \frac{\partial^2 \mu(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t} \partial \mathbf{t}^T}\bigg|_{\mathbf{t}=\mathbf{B}^*}$$

Notons que $\mathbf{h}$ *est un vecteur de longueur* m *et* $\mathbf{k}$ *est une matrice de dimension* $m \times m$, *où* m *est le nombre de paramètres dans* $\boldsymbol{\beta}$. *Étant donné les conditions (2) et (3),*

$$\max_i \left| \mathbf{h}(\mathbf{x}_i, \mathbf{B}) \right| \leq h(\mathbf{x}_i, \mathbf{B}) \tag{A.3}$$

$$\max_{k,\, j} \left| \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*) \right| \leq k(\mathbf{x}_i, \mathbf{B}^*). \tag{A.4}$$

Les conditions (1) et (3) impliquent que

$$\mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) = \mu(\mathbf{x}_i, \mathbf{B}) + O_p\left(1/\sqrt{n}\right) \tag{A.5}$$

$$\equiv \mu_i + O_p\left(1/\sqrt{n}\right). \tag{A.6}$$

L'équation (2.2) de la section 2.1 et les conditions d'existence de bornes (boundedness) de (2) et (3) de la section 3.2.2 impliquent que

$$\begin{aligned} N^{-1}\sum_{i\in s_A} d_i^A \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) &= N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T \left(\hat{\mathbf{B}} - \mathbf{B}\right) \\ &\quad + \left(\hat{\mathbf{B}} - \mathbf{B}\right)^T N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)\right)\left(\hat{\mathbf{B}} - \mathbf{B}\right) \\ &= N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T \left(\hat{\mathbf{B}} - \mathbf{B}\right) + O_p\left(\frac{1}{n}\right). \end{aligned} \tag{A.7}$$

Étant donné les conditions (1), (4) et l'équation (A.7) :

$$\begin{aligned} N^{-1}\sum_{k=1}^{N} \mu\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) &- N^{-1}\sum_{i\in s_A} d_i^A \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) \\ &= N^{-1}\sum_{k=1}^{N} \mu(\mathbf{x}_i, \mathbf{B}) - N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{A.8}$$

En utilisant les conditions (1) et (3),

$$\begin{aligned} \bar{\hat{\mu}} &= \sum_{i\in s_A} d_i^A \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) \Big/ \sum_{i\in s_A} d_i^A \\ &= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})\left(\hat{\mathbf{B}} - \mathbf{B}\right) + \left(\hat{\mathbf{B}} - \mathbf{B}\right)^T \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)\left(\hat{\mathbf{B}} - \mathbf{B}\right)\right) \\ &= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})\left(\hat{\mathbf{B}} - \mathbf{B}\right)\right) + O_p(1/n) \\ &= \bar{\mu} + \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})\left(\hat{\mathbf{B}} - \mathbf{B}\right) + O_p(1/n) \\ &\quad (\textit{selon les conditions } (1) \textit{ et } (18)) \\ &= \bar{\mu} + O_p\left(1/\sqrt{n}\right) + O_p(1/n) \\ &= \bar{\mu} + O_p\left(1/\sqrt{n}\right) \end{aligned} \tag{A.9}$$

pour $\bar{\mu} = \sum_{i\in s_A} d_i^A \mu_i \Big/ \sum_{i\in s_A} d_i^A$.

Alors, partant de (A.2) et (A.9) et en utilisant les conditions (1) à (3), nous avons

$$\begin{aligned}
N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right) &= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu(\mathbf{x}_i,\mathbf{B})+\mathbf{h}^T(\mathbf{x}_i,\mathbf{B})\left(\hat{\mathbf{B}}-\mathbf{B}\right)+\left(\hat{\mathbf{B}}-\mathbf{B}\right)^T\mathbf{k}\left(\mathbf{x}_i,\mathbf{B}^*\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)-\bar{\mu}\right)\\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)+N^{-1}\sum_{i\in s_A} d_i^A\mathbf{h}^T\left(\mathbf{x}_i,\mathbf{B}\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)\\
&\quad + N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mathbf{B}}-\mathbf{B}\right)^T\mathbf{k}\left(\mathbf{x}_i,\mathbf{B}^*\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)-O_p\left(1/\sqrt{n}\right)\\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)+O_p\left(1/\sqrt{n}\right)+O_p\left(1/n\right)-O_p\left(1/\sqrt{n}\right)\\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)+O_p\left(1/\sqrt{n}\right). \qquad \text{(A.10)}
\end{aligned}$$

Similairement,

$$N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)^2 = N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)^2+O_p\left(1/n\right). \qquad \text{(A.11)}$$

Partant de (A.10) et (A.11), nous avons :

$$\begin{aligned}
\hat{B}^{MC} &= \frac{\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)\left(y_i-\bar{y}\right)}{\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)^2}=\frac{N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)\left(y_i-\bar{y}\right)}{N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)^2}\\
&= \frac{\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)\left(y_i-\bar{y}\right)+O_p\left(1/\sqrt{n}\right)}{\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)^2+O_p\left(1/n\right)}\\
&\to B^{MC} \qquad \textit{quand } n\to\infty. \qquad \text{(A.12)}
\end{aligned}$$

Donc, $\hat{B}^{MC} = B^{MC}+o_p(1)$, *et nous avons :*

$$\begin{aligned}
\hat{\bar{y}}^{MC} &= N^{-1}\hat{T}_y^{MC}\\
&= N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\hat{\mathbf{B}}\right)+\sum_{i\in s_A} N^{-1}d_i^A\mu\left(\mathbf{x}_i,\hat{\mathbf{B}}\right)\right)\hat{B}^{MC}\\
&= N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\mathbf{B}\right)-N^{-1}\sum_{i\in s_A} d_i^A\mu\left(\mathbf{x}_i,\mathbf{B}\right)+O_p\left(\frac{1}{\sqrt{n}}\right)\right)\left(B^{MC}+o_p(1)\right)\\
&= N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\mathbf{B}\right)-N^{-1}\sum_{i\in s_A} d_i^A\mu\left(\mathbf{x}_i,\mathbf{B}\right)\right)B^{MC}+o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Puisque $N = O_p(N)$, *nous avons* $N\cdot o_P\left(1/\sqrt{n}\right) = O_p(N)\,o_p\left(1/\sqrt{n}\right) = o_p\left(N/\sqrt{n}\right)$. *Donc,*

$$\begin{aligned}\hat{T}_y^{MC} &= N\hat{\bar{y}}^{MC} = N\left(N^{-1}\mathbf{d}^A\mathbf{y} + \left(N^{-1}\sum_{k=1}^{N}\mu(\mathbf{x}_k,\mathbf{B}) - N^{-1}\sum_{i\in s_A}\mu(\mathbf{x}_i,\mathbf{B})\right)B^{MC} + o_p\left(\frac{1}{\sqrt{n}}\right)\right)\\ &= \mathbf{d}^A\mathbf{y} + \left(\sum_{k=1}^{N}\mu(\mathbf{x}_k,\mathbf{B}) - \sum_{i\in s_A}\mu(\mathbf{x}_i,\mathbf{B})\right)B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right)\\ &= \sum_{iins_A} d_i^A(y_i - \mu_i B^{MC}) + \sum_{i=1}^{N}\mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right). \qquad \text{(A.13)}\end{aligned}$$

Théorème 2 : *Supposons que les paramètres d'un modèle de régression complet possèdent à la fois des composantes nulles et non nulles, et sans perte de généralité, posons que les p premières sont non nulles et que les q dernières sont nulles :*

$$\boldsymbol{\beta}^F = \begin{pmatrix}\boldsymbol{\beta}^{(1)}_{(p\times 1)}\\ \boldsymbol{\beta}^{(2)}_{(q\times 1)}\end{pmatrix},\ \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta} \quad et \quad \boldsymbol{\beta}^{(2)} = \mathbf{0}_{(q\times 1)}.$$

Sous les conditions (1) à (5), l'estimateur asymptotique du total par calage avec le LASSO est :

$$\hat{T}_y^{LASSO} = \sum_{i\in s_A} d_i^A(y_i - \mu_i B^{MC}) + \sum_{i=1}^{N}\mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right). \qquad \text{(A.14)}$$

Preuve. Sous la condition (5), la régression LASSO adaptatif satisfait la propriété d'oracle par l'entremise des théorèmes 1 et 4 dans Zou (2006) :

$$\begin{aligned}Pr(\mathbf{B}^{(2)} = \mathbf{0}) &\to 1\\ \sqrt{n}\left(\hat{\mathbf{B}}^{(1)} - \mathbf{B}\right) &\to N(\mathbf{0},\mathbf{C})\\ \mathbf{B} &\to \boldsymbol{\beta}\end{aligned}$$

où $\mathbf{C} = \Sigma(\mathbf{B})$ est la matrice de covariance de $\mathbf{B}^{(1)}$ sous le modèle linéaire, et $\mathbf{C} = I^{-1}(\mathbf{B})$ est l'inverse de la matrice d'information de Fisher de $\mathbf{B}^{(1)}$ sous le modèle linéaire généralisé. D'après le théorème de Slutsky, la propriété d'oracle implique que $\hat{\mathbf{B}}^{(1)} = \mathbf{B} + O_p\left(1/\sqrt{n}\right)$. Selon la condition (1) et le lemme 1 :

$$\begin{aligned}\hat{T}_y^{LASSO} &\approx \hat{T}_y^{MC}\\ &= \sum_{iins_A} d_i^A(y_i - \mu_i B^{MC}) + \sum_{i=1}^{N}\mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right).\end{aligned}$$

Théorème 3 : *$\hat{T}_y^{LASSO}$ est sans biais sous le modèle.*

Preuve. Sous l'hypothèse de notre cadre théorique, les paramètres de superpopulation sont un sous-ensemble des paramètres de la régression LASSO complète, et nous pouvons prouver l'absence de biais sous le modèle de $\hat{T}_y^{LASSO}$ en prenant les espérances par rapport au modèle ξ. Nous notons d'abord que :

$$E_\xi[B^{MC}] = E_\xi\left[\frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2}\right] = \frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2} = 1.$$

Donc,

$$E_\xi\left[\hat{T}_y^{LASSO} - T\right] \approx E_\xi\left[\sum_{i\in s_A} d_i^A\left(y_i - \mu_i B^{MC}\right) + \sum_{i=1}^N \mu_i B^{MC} - \sum_{i=1}^N y_i\right]$$

$$= \sum_{i\in s_A} d_i^A\left(\mu_i - \mu_i\right) + \sum_{i=1}^N \mu_i - \sum_{i=1}^N \mu_i \qquad \left(Puisque\ E_\xi\left[B^{MC}\right] = 1\right)$$

$$= 0.$$

Donc, à condition que les paramètres de la régression LASSO comprennent les paramètres de superpopulation, $\hat{T}_y^{\text{LASSO}}$ est sans biais sous le modèle quels que soient les poids de sondage. Cette propriété est essentielle en échantillonnage non probabiliste, où il n'existe pas de poids de sondage initiaux pour garantir l'absence de biais.

Bibliographie

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K. et Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

Chipman, H.A., George, E.I. et McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4, 266-298.

Cardot, H., Goga, C. et Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27, 243-260.

Centers for Disease Control and Prevention (2005). *2004 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description*. National Center for Health Statistics: Hyattsville, Maryland. www.cdc.gov/nchs/data/nhis/srvydesc.pdf.

Czanner, G., Sarma, S.V., Eden, U.T. et Brown, E.N. (2008). A signal-to-noise ratio estimator for generalized linear model systems. *Proceedings of the World Congress on Engineering*, vol. 2.

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

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J.R.G., Gruber, B., Lafourcade, B., Leitao, P.J. et Mnkemller, T. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecology*, 36, 27-46.

Elliott, M.R. (2009). Combining data from probability and nonprobability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott, M.R, Resler, A., Flannagan, C. et Rupp, J. (2010). Combining data from probability and non-probability samples using pseudo-weights. *Accident Analysis and Prevention*, 42, 530-539.

Elliott, M.R., et Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.

Fan, J., et Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Friedman, J., Hastie, T. et Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1-22.

Frankel, M.R., et Frankel, L.R. (1987). Fifty years of survey sampling in the United States. *Public Opinion Quarterly*, S127-S138.

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

Goga, C., Muhammad-Shehzad, A. et Vanheuverzwyn, A. (2011). Principal component regression with survey data: Application on the French media audience. *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, 3847-3852.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Kamarianakis, Y., Shen, W. et Wynter, L. (2012). Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO. *Applied Stochastic Models in Business and Industry*, 28, 297-315.

Kohannim, O., Hibar, D.P., Stein, J.L., Jahanshad, N., Hua, X., Rajagopalan, P., Toga, A., Jack Jr, C.R., Weiner, M.W., de Zubicaray, G.I. et McMahon, K.L. (2012). Discovery and replication of gene influences on brain structure using LASSO regression. *Frontiers in Neuroscience*, 6, 115.

Kohut, A., Keeter, S., Doherty, C., Dimock, M. et Christian, L. (2012). Assessing the representativeness of public opinion surveys. *Pew Research Center for The People & The Press*. http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/.

Mosteller, F. (1949). *The Pre-Election Polls of 1948: The Report to the Committee on Analysis of Pre-Election Polls and Forecasts*, vol. 60, Social Science Research Council.

McConville, K. (2011). *Improved Estimation for Complex Surveys Using Modern Regression Techniques*. Thèse de doctorat non-publiée, Colorado State University.

McConville, K., Breidt, F.J., Lee, T.M. et Moisen, G.G. (2017). Model-assisted survey regression estimation with the LASSO. *Journal of Survey Statistics and Methodology*, 5, 131-158.

Park, M., et Yang, M. (2008). Ridge regression estimation for survey samples. *Communication in Statistics - Theory and Methods*, 37, 532-543.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Joint Statistical Meetings*, American Statistical Association.

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

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

Skinner, C., et Silva, P. (1997). Variable selection for regression estimation in the presence of nonresponse. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 76-81.

Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association,* 43, 12-39.

Terhanian, G., et Bremer, J. (2012). A smarter way to select respondents for surveys? *International Journal of Market Research,* 54, 751-780.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, 58, 267-288.

Tourangeau, R., Conrad, F.G. et Couper, M.P. (2013). *The Science of Web Surveys.* Oxford University Press, Oxford, Royaume-Uni.

Vavreck, L., et Rivers, D. (2008). The 2006 Cooperative Congressional Election Study. *Journal of Elections, Public Opinion, and Parties,* 355-366.

Wang, W., Rothschild, D., Goel, S. et Gelman, A. (2015). Forecasting elections with non-representative Polls. *International Journal of Forecasting*, 31, 980-991.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E. et Lange, K. (2009). Genome-wide association analysis by LASSO penalized logistic regression. *Bioinformatics*, 25, 714-721.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.

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 33, No. 4, December 2017

The Morris Hansen Lecture
Bates, Nancy 873

Discussion
Edwards, Brad 887
Jacobsen, Linda A. 891

Adaptive Intervention Methodology for Reduction of Respondent Contact Burden in the American Community Survey
Ashmead, Robert / Slud, Eric / Hughes, Todd 901

Estimating Classification Errors Under Edit Restrictions in Composite Survey-Register Data Using Multiple Imputation Latent Class Modelling (MILC)
Boeschoten, Laura / Oberski, Daniel / de Waal, Ton 921

How to Obtain Valid Inference under Unit Nonresponse?
Boeschoten, Laura / Vink, Gerko / Hox, Joop J.C.M. 963

The Effects of the Frequency and Implementation Lag of Basket Updates on the Canadian CPI
Huang, Ning / Wimalaratne, Waruna / Pollard, Brent 979

Multiply-Imputed Synthetic Data: Advice to the Imputer
Loong, Bronwyn / Rubin, Donald B. 1005

Estimating Cross-Classified Population Counts of Multidimensional Tables: An Application to Regional Australia to Obtain Pseudo-Census Counts
Suesse, Thomas / Namazi-Rad, Mohammad-Reza / Mokhtarian, Payam / Barthélemy, Johan 1021

Figuring Figures: Exploring Europeans' Knowledge of Official Economic Statistics
Vicente, Maria R. / López, Ana J. 1051

Book Review
Cruze, Nathan 1087
Earp, Morgan. 1091

Editorial Collaborators 1093
Index to Volume 33, 2017 1101

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 34, No. 1, March 2018

Household Classification Using Smart Meter Data
Carroll, Paula / Murphy, Tadhg / Hanley, Michael / Dempsey, Daniel / Dunne, John 1

Constraint Simplification for Data Editing of Numerical Variables
Daalmans, Jacco 27

Design-Based Estimation with Record-Linked Administrative Files and a Clerical Review Sample
Dasylva, Abel 41

Administrative Data Quality: Investigating Record-Level Address Accuracy in the Northern Ireland Health Register
Foley, Brian / Shuttleworth, Ian / Martin, David 55

Typology and Representation of Alterations in Territorial Units: A Proposal
Goerlich, Francisco / Ruiz, Francisco 83

Calibration Weighting for Nonresponse with Proxy Frame Variables (So that Unit Nonresponse Can Be Not Missing at Random)
Kott, Phillip S. / Liao, Dan 107

Optimal Stratification and Allocation for the June Agricultural Survey
Lisic, Jonathan / Sang, Hejian / Zhu, Zhengyuan / Zimmer, Stephanie 121

Components of Gini, Bonferroni, and Zenga Inequality Indexes for EU Income Data
Pasquazzi, Leo / Zenga, Michele 149

Using Social Network Information for Survey Estimation
Suesse, Thomas / Chambers, Ray 181

An Analysis of Interviewer Travel and Field Outcomes in Two Field Surveys
Wagner, James / Olson, Kristen 211

An Overview of Population Size Estimation where Linking Registers Results in Incomplete Covariates, with an Application to Mode of Transport of Serious Road Casualties
Heijden, Peter G.M. van der / Smith, Paul A. / Cruyff, Maarten / Bakker, Bart 239

Factor Structural Time Series Models for Official Statistics with an Application to Hours Worked in Germany
Weigand, Roland / Wanger, Susanne / Zapf, Ines 265

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 46, No. 1, March/mars 2018

Richard Lockhart
Special issue on big data and the statistical sciences: Guest editor's introduction 4

Mary E. Thompson
Dynamic data science and official statistics 10

Shelley B. Bull, Irene L. Andrulis, Andrew D. Paterson
Statistical challenges in high-dimensional molecular and genetic epidemiology 24

Jonathan Taylor, Robert Tibshirani
Post-selection inference for ℓ_1-penalized likelihood models 41

R. Dennis Cook, Liliana Forzani
Big data and partial least-squares prediction 62

Bing Li
Linear operator-based statistical analysis: A useful paradigm for big data 79

Rui Nie, Douglas P. Wiens, Zhichun Zhai
Minimax robust active learning for approximately specified regression models....104

Chun Wang, Ming-Hui Chen, Jing Wu, Jun Yan, Yuping Zhang, Elizabeth Schifano
Online updating method with new variables for big data streams 123

Reihaneh Entezari, Radu V. Craiu, Jeffrey S. Rosenthal
Likelihood inflating sampling algorithm 147

Qiong Li, Xin Gao, Hélène Massam
Approximate Bayesian estimation in large coloured graphical Gaussian models 176

Acknowledgement of referees' services: Remerciements aux lecteurs critiques 204

CONTENTS TABLE DES MATIÈRES

Volume 46, No. 2, June/juin 2018

Song Cai, Jiahua Chen
Empirical likelihood inference for multiple censored samples 212

Qihua Wang, Xuan Wang
Analysis of censored data under heteroscedastic transformation regression models with unknown transformation function 233

Irène Gijbels, Mohammed A. Ibrahim, Anneleen Verhasselt
Testing the heteroscedastic error structure in quantile varying coefficient models 246

Hyonho Chun, Myung Hee Lee, Sung-Ho Kim, Jihwan Oh
Robust precision matrix estimation via weighted median regression with regularization 265

Paul Kabaila
On the minimum coverage probability of model averaged tail area confidence intervals 279

Jean-François Plante, Jean-Baptiste Débordès
Combining ROC curves using MAMSE weighted distributions 298

Yuki Kawakubo, Shonosuke Sugasawa, Tatsuya Kubokawa
Conditional Akaike information under covariate shift with application to small area estimation 316

Yi Zhai, Zhide Fang
Locally optimal designs for binary dose-response models 336

Ibrahim Abdelrazeq, B. Gail Ivanoff, Rafal Kulik
Goodness-of-fit tests for Lévy-driven Ornstein-Uhlenbeck processes 355

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique au rédacteur en chef (statcan.smj-rte.statcan@canada.ca). Avant de soumettre l'article, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 39, n° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word et MathType pour les expressions mathématiques. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$, etc.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées par un chiffre arabe à la droite si l'auteur y fait référence plus loin. Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, l'équation (4.2) est la deuxième équation importante de la section 4.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0 ; l, 1).

3.6 Si possible, éviter l'emploi de caractères gras dans les formules.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux). Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, le tableau 3.1 est le premier tableau de la section 3.

4.2 Une description textuelle détaillée des figures pourrait être requise à des fins d'accessibilité si le message transmis par l'image n'est pas suffisamment expliqué dans le texte.

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple : Cochran (1977, page 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

6. Communications brèves

6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.