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

Techniques d’enquête

Techniques d’enquête
42-1

Date de diffusion : le 22 juin 2016

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)

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 2016

•

Volume 42

•

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*.

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 42, numéro 1, juin 2016

Table des matières

Articles réguliers

Sander Scholtus
Une généralisation du paradigme de Fellegi-Holt pour la localisation automatique des erreurs 1

Jae Kwang Kim, Emily Berg et Taesung Park
Appariement statistique par imputation fractionnaire 21

Michael A. Hidiroglou et Yong You
Comparaison d'estimateurs sur petits domaines au niveau de l'unité et au niveau du domaine 45

Susana Rubin-Bleuer et Yong You
Comparaison de certains estimateurs de variance positifs pour le modèle d'estimation sur petits domaines Fay-Herriot 69

Leo Pasquazzi et Lucio de Capitani
Une comparaison d'estimateurs non paramétriques pour les fonctions de répartition de populations finies 95

Michael A. Hidiroglou, Jae Kwang Kim et Christian Olivier Nambeu
Remarque concernant l'estimation par régression lorsque la taille de la population est inconnue 131

Jan A. van den Brakel
Échantillonnage fondé sur des registres pour les panels auprès des ménages 147

Ismael Flores Cervantes et J. Michael Brick
Ajustements pour la non-réponse dans les plans stratifiés assortis de modèles aux spécifications erronées 173

Communication brève

Linda Schulze Waltrup et Göran Kauermann
Note brève sur l'estimation fondée sur les quantiles et les expectiles dans les échantillons à probabilités inégales 191

Addendum 201
Autres revues 203

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

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

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 1-20
Statistique Canada, nº 12-001-X au catalogue

Une généralisation du paradigme de Fellegi-Holt pour la localisation automatique des erreurs

Sander Scholtus[1]

Résumé

La vérification automatique consiste en l'utilisation d'un ordinateur pour déceler et corriger sans intervention humaine les valeurs erronées dans un ensemble de données. La plupart des méthodes de vérification automatique actuellement employées aux fins de la statistique officielle sont fondées sur les travaux fondamentaux de Fellegi et Holt (1976). La mise en application de cette méthode dans la pratique révèle des différences systématiques entre les données vérifiées manuellement et celles qui sont vérifiées de façon automatisée, car l'humain est en mesure d'effectuer des opérations de vérification complexes. L'auteur du présent article propose une généralisation du paradigme de Fellegi-Holt qui permet d'intégrer de façon naturelle une grande catégorie d'opérations de vérification. Il présente aussi un algorithme qui résout le problème généralisé de localisation des erreurs qui en découle. Il est à espérer que cette généralisation puisse améliorer la pertinence des vérifications automatiques dans la pratique et ainsi accroître l'efficience des processus de vérification des données. Certains des premiers résultats obtenus à l'aide de données synthétiques sont prometteurs à cet égard.

Mots-clés : Vérification automatique; opérations de vérification; maximum de vraisemblance; données numériques; vérifications linéaires.

1 Introduction

Les données recueillies aux fins de la production de statistiques contiennent inévitablement des erreurs. Il est donc nécessaire de mettre en place un processus de vérification des données pour déceler et corriger ces erreurs, au moins dans la mesure où elles ont un effet appréciable sur la qualité des produits statistiques (Granquist et Kovar 1997). Traditionnellement, la vérification des données se faisait manuellement, idéalement par des vérificateurs spécialisés ayant une connaissance approfondie du sujet. Pour améliorer l'efficience, la rapidité et la reproductibilité de la vérification, beaucoup d'instituts de statistique ont tenté d'automatiser certains segments du processus (Pannekoek, Scholtus et van der Loo 2013). Il en a résulté des méthodes de correction déductive des *erreurs systématiques* et des algorithmes de localisation des erreurs pour les *erreurs aléatoires* (de Waal, Pannekoek et Scholtus 2011, chapitre 1). Le présent article est axé sur la vérification automatique des erreurs aléatoires.

Les méthodes pour l'exécution de cette tâche comprennent généralement un ajustement minimal de chaque enregistrement de données en fonction de certains critères d'optimisation, afin d'en assurer la cohérence avec un ensemble déterminé de contraintes que l'on appelle *règles de vérification*, ou simplement *contrôles*. Selon l'efficacité des critères d'optimisation et la puissance des contrôles, la vérification automatique peut remplacer en partie la vérification manuelle traditionnelle. Dans les faits, la vérification automatique est presque toujours jumelée à une forme quelconque de *vérification sélective*, ce qui signifie que les erreurs ayant les répercussions les plus importantes sont traitées manuellement (Hidiroglou et Berthelot 1986; Granquist 1995, 1997; Granquist et Kovar 1997; Lawrence et McKenzie 2000; Hedlin 2003; de Waal et coll. 2011).

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

La plupart des méthodes de vérification automatique actuellement utilisées pour la statistique officielle sont fondées sur le paradigme de Fellegi et Holt (1976) : pour chaque enregistrement, on trouve le plus petit sous-ensemble de variables erronées qui peuvent être imputées de sorte que l'enregistrement satisfasse aux contrôles. On peut obtenir une légère généralisation en attribuant ce qu'on appelle *poids de confiance* aux variables et en minimisant le poids total des variables imputées. Une fois résolu ce *problème de localisation des erreurs*, il faut trouver séparément de nouvelles valeurs qui conviennent pour les variables identifiées comme étant erronées. C'est ce qu'on appelle le *problème d'imputation cohérente*; voir à ce sujet de Waal et coll. (2011) et les ouvrages cités en référence. Le présent article est axé sur le problème de localisation des erreurs.

À *Statistics Netherlands*, la localisation des erreurs à l'aide du paradigme de Fellegi-Holt fait partie du processus de vérification des données relatives aux statistiques structurelles sur les entreprises (SSE) depuis plus de dix ans. Dans le cadre d'études d'évaluation, où les mêmes données sur les SSE ont été vérifiées à la fois automatiquement et manuellement, on a constaté un certain nombre de différences systématiques entre les deux processus. Bon nombre de ces différences pouvaient s'expliquer par le fait que les vérificateurs humains ont apporté certains types de correction qui ne sont pas optimaux selon le paradigme de Fellegi-Holt. Par exemple, les vérificateurs ont parfois interverti les valeurs de dépenses et de revenus associés ou transféré une partie des unités déclarées d'une variable à l'autre.

En pratique, le résultat de la vérification manuelle est généralement considéré comme étant la « norme de référence » pour évaluer la qualité de la vérification automatique. Une évaluation critique de cette hypothèse dépasse le cadre du présent article; toutefois, le lecteur intéressé pourra consulter EDIMBUS (2007, pages 34-35). On souligne simplement ici qu'en améliorant la capacité des méthodes de vérification automatique à reproduire les résultats de la vérification manuelle, on accroît leur utilité dans la pratique. Par ricochet, cela signifie que l'on peut accroître la part de la vérification automatique pour améliorer l'efficience du processus de vérification des données (Pannekoek et coll. 2013).

Dans une certaine mesure, les différences systématiques entre la vérification automatique et la vérification manuelle pourraient être éliminées par l'application judicieuse de poids de confiance. En règle générale, toutefois, les effets d'une modification des poids de confiance sur les résultats de la vérification automatique sont difficiles à prévoir. En outre, si les vérificateurs apportent un certain nombre de corrections différentes et complexes, il pourrait être impossible de toutes les modéliser sous le paradigme de Fellegi-Holt à l'aide d'un seul ensemble de poids de confiance. Une autre solution consiste à essayer de déceler les erreurs pour lesquelles on sait que le paradigme de Fellegi-Holt donne un résultat insatisfaisant dès les premières étapes du processus de vérification des données, c'est-à-dire durant la correction déductive des erreurs systématiques à l'aide de règles de correction automatique (de Waal et coll. 2011; Scholtus 2011). Cette méthode comporte toutefois des limites pratiques; elle peut notamment exiger un grand nombre de règles du type « si-alors », qui peuvent se révéler difficiles à concevoir et à tenir à jour au fil du temps (Chen, Thibaudeau et Winkler 2003). En outre, il n'est pas nécessairement aisé de trouver des règles de correction appropriées pour toutes les erreurs qui ne peuvent pas être traitées en vertu du paradigme de Fellegi-Holt.

Dans le présent article, on propose une autre approche : une nouvelle définition du problème de localisation des erreurs qui tient compte de la possibilité qu'une erreur puisse toucher plus d'une variable à la fois. On montre que ce problème contient la localisation des erreurs en vertu du paradigme original de Fellegi-Holt comme un cas particulier. Le présent article porte principalement sur les données numériques

et les règles de vérification linéaires; un élargissement possible aux données catégoriques et mixtes est présenté brièvement à la section 8.

Le reste de l'article se présente comme suit. La section 2 passe brièvement en revue les travaux antérieurs pertinents dans le domaine. À la section 3, on présente et on illustre le concept des opérations de vérification. Le nouveau problème de localisation des erreurs est formulé en termes de ces opérations à la section 4. La section 5 énonce une généralisation d'une méthode existante pour trouver des solutions au problème de localisation des erreurs fondé sur le paradigme de Fellegi-Holt, et le résultat est utilisé à la section 6 pour construire un algorithme possible pour la résolution du nouveau problème. Une étude par simulations de petite envergure est présentée à la section 7. Enfin, à la section 8, on énonce certaines conclusions et on formule des questions pour approfondir la recherche.

2 Contexte et travaux connexes

Soit $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ un enregistrement de p variables numériques. Supposons que cet enregistrement doive satisfaire à k règles de vérification, se présentant sous la forme du système d'(in)égalités linéaires suivant :

$$\mathbf{A}\mathbf{x} + \mathbf{b} \odot \mathbf{0}, \tag{2.1}$$

où $\mathbf{A} = (a_{rj})$ est une matrice $k \times p$ de coefficients et $\mathbf{b} = (b_1, \ldots, b_k)'$ est un vecteur de constantes. Ici comme ailleurs, $\mathbf{0}$ représente un vecteur de zéros de longueur appropriée; de même, $\odot$ représente un vecteur symbolique d'opérateurs de l'ensemble $\{\geq, \leq, =\}$.

Pour un enregistrement donné $\mathbf{x}$ qui ne satisfait pas à toutes les règles de vérification énoncées en (2.1), le problème de localisation des erreurs fondée sur le paradigme de Fellegi-Holt consiste à trouver la valeur minimale de

$$\sum_{j=1}^{p} w_j \delta_j, \tag{2.2}$$

où $w_j > 0$ est le poids de confiance de la variable x_j et $\delta_j \in \{0,1\}$, à condition qu'on puisse assurer la cohérence de l'enregistrement original avec les règles de vérification en imputant uniquement les variables x_j pour lesquelles $\delta_j = 1$ (de Waal et coll. 2011, page 66).

Fellegi et Holt (1976) ont aussi proposé une méthode de résolution du problème de localisation des erreurs ci-dessus fondée sur la production d'un ensemble suffisant de *vérifications implicites* (voir ci-dessous). Malheureusement, cette méthode exige souvent un très grand nombre de vérifications implicites. Au cours des dernières décennies, divers algorithmes spécialisés ont été élaborés pour le problème de localisation des erreurs, notamment par Schaffer (1987), Garfinkel, Kunnathur et Liepins (1988), Kovar et Whitridge (1990), Ragsdale et McKeown (1996), de Waal (2003), de Waal et Quere (2003), Riera-Ledesma et Salazar-González (2003, 2007), Bruni (2004), ainsi que de Jonge et van der Loo (2014). Les premiers algorithmes visaient principalement à renforcer la méthode originale de Fellegi et Holt (1976) en réduisant le nombre de vérifications implicites requises. Les algorithmes plus récents reposent sur le fait que le

problème de localisation des erreurs peut être rédigé sous forme de problème de programmation mixte en nombres entiers, ce qui permet l'application de techniques d'optimisation normalisées. Voir aussi de Waal et Coutinho (2005) ou de Waal et coll. (2011) pour une vue d'ensemble et une comparaison des divers algorithmes de localisation des erreurs.

Les vérifications implicites sont des contraintes qui découlent logiquement des règles de vérification originales (2.1). Dans le contexte qui nous occupe (données numériques, vérifications linéaires), toutes les vérifications implicites pertinentes peuvent être générées par une technique appelée *élimination de Fourier-Motzkin* (élimination FM; voir Williams 1986). L'élimination FM transforme un système de contraintes linéaires à p variables en un système de contraintes linéaires implicites à au plus $p-1$ variables; ainsi, au moins une des variables originales est éliminée. Pour les détails mathématiques, consultez l'annexe.

L'élimination FM est assortie de la propriété fondamentale suivante : le système de contraintes implicites est satisfait par les valeurs des variables non éliminées si et seulement s'il existe une valeur pour la variable éliminée qui, prise avec les autres valeurs, satisfait au système original de contraintes. Dans la localisation des erreurs en vertu du paradigme de Fellegi-Holt, on peut, en appliquant à répétition cette propriété fondamentale, vérifier si une combinaison particulière de variables peut être imputée pour obtenir un enregistrement cohérent, compte tenu des valeurs originales des autres variables. L'algorithme de localisation des erreurs de de Waal et Quere (2003) illustre bien cette utilisation de l'élimination FM.

Pour conclure cette section, il est intéressant d'examiner brièvement l'interprétation statistique du problème de localisation des erreurs. En fait, Fellegi et Holt (1976) n'ont fourni aucun argument statistique formel pour expliquer leur paradigme de localisation automatique des erreurs. Leur raisonnement était plutôt intuitif :

> *« Les données de chaque enregistrement doivent être corrigées afin de satisfaire à toutes les règles de vérification en changeant le moins d'éléments de données (champs) possible. Nous sommes d'avis que cette méthode respecte l'idée de garder telles quelles le plus grand nombre possible des données originales, compte tenu des contraintes des règles de vérification, et donc de modifier le moins de données possible. Parallèlement, si les erreurs sont relativement rares, il semble plus probable que l'on puisse identifier les champs réellement erronés. » (Fellegi et Holt 1976, page 18).* [Traduction]

Liepins (1980) ainsi que Liepins, Garfinkel et Kunnathur (1982), en se fondant sur les résultats antérieurs de Naus, Johnson et Montalvo (1972), ont formulé un argument statistique pour minimiser le nombre pondéré de variables imputées. Supposons que les erreurs se produisent selon un processus stochastique, chaque variable x_j étant erronée selon une probabilité p_j qui ne dépend pas de sa valeur réelle et les erreurs étant indépendantes d'une variable à l'autre. Supposons en outre que les poids de confiance sont définis comme suit :

$$w_j = -\log\left(\frac{p_j}{1-p_j}\right). \tag{2.3}$$

On peut alors montrer que la minimisation de l'expression (2.2) correspond approximativement à la maximisation de la vraisemblance de l'enregistrement exempt d'erreur non observé. Soulignons que ces auteurs supposent tacitement qu'une erreur affecte toujours une seule variable à la fois.

D'autres méthodes de localisation des erreurs reposant plus directement sur des modèles statistiques ont été proposées, notamment par Little et Smith (1987) et par Ghosh-Dastidar et Schafer (2006). Ces méthodes ont recours à des techniques de détection des valeurs aberrantes et exigent un modèle explicite pour les données réelles. Malheureusement, elles ne peuvent pas tenir compte de façon directe des règles de vérification comme celle qui est illustrée en (2.1).

3 Opérations de vérification

Poursuivons la notation de la section 2 en définissant une *opération de vérification* g comme une fonction affine de forme générale

$$g(\mathbf{x}) = \mathbf{T}\mathbf{x} + \mathbf{S}\boldsymbol{\alpha} + \mathbf{c}, \tag{3.1}$$

où $\mathbf{T}$ et $\mathbf{S}$ sont des matrices de coefficients connues de dimensions $p \times p$ et $p \times m$, respectivement, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ est un vecteur de paramètres libres qui peuvent se produire dans g, et $\mathbf{c}$ est un vecteur de p constantes connues. Dans le cas particulier où g ne comprend aucun paramètre libre $(m = 0)$, le second terme de l'équation (3.1) disparaît. Il est parfois utile d'imposer une ou plusieurs contraintes linéaires aux paramètres libres de g :

$$\mathbf{R}\boldsymbol{\alpha} + \mathbf{d} \odot \mathbf{0}, \tag{3.2}$$

où $\mathbf{R}$ est une matrice connue et $\mathbf{d}$, un vecteur de constantes connu. (Remarque : La notation matricielle-vectorielle est utilisée tout au long de l'article parce qu'elle permet de décrire avec concision les résultats; le recours à des matrices pour représenter les règles et les opérations de vérification ne constitue toutefois probablement pas le meilleur moyen de traiter ces résultats par ordinateur.)

Comme premier exemple, prenons l'opération qui remplace l'une des valeurs originales de $\mathbf{x}$ par une nouvelle valeur arbitraire (imputation), que nous appellerons *opération FH*, vu son rôle central dans la vérification automatique fondée sur le paradigme de Fellegi-Holt. Soit $\mathbf{I}$ la matrice identité $p \times p$ et $\mathbf{e}_i$ le i^{e} vecteur de base canonique de $\mathbb{R}^p$. L'opération FH qui impute la variable x_j est donnée par (3.1) où $\mathbf{T} = \mathbf{I} - \mathbf{e}_j\mathbf{e}_j'$, $\mathbf{S} = \mathbf{e}_j$ et $\mathbf{c} = \mathbf{0}$. On obtient : $g(\mathbf{x}) = \mathbf{x} + \mathbf{e}_j(\alpha - x_j) = (x_1, \ldots, x_{j-1}, \alpha, x_{j+1}, \ldots, x_p)'$, où $\alpha \in \mathbb{R}$, un paramètre libre représentant la valeur imputée. Soulignons que pour un enregistrement de p variables, on peut définir p opérations FH distinctes.

Pour mieux illustrer le concept d'opération de vérification, d'autres exemples sont présentés ci-dessous. Pour faciliter la notation, ces exemples sont limités au cas où $p = 3$.

- Opération de vérification qui change le signe d'une des variables :

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

- Opération de vérification qui intervertit les valeurs de deux éléments adjacents :

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix}.$$

- Opération de vérification qui transfère un nombre d'unités d'un élément à un autre, où le nombre d'unités transférées peut équivaloir à au plus K unités dans un sens ou dans l'autre :

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}\alpha + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} x_1 + \alpha \\ x_2 \\ x_3 - \alpha \end{pmatrix},$$

  où la contrainte suivante s'applique : $-K \leq \alpha \leq K.$

- Opération de vérification qui impute deux variables simultanément selon un ratio fixe :

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ x_3 \end{pmatrix},$$

  où la contrainte suivante s'applique : $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$ satisfait $10\alpha_1 - \alpha_2 = 0$.

Intuitivement, une opération de vérification est censée « renverser les effets » d'un type particulier d'erreur qui aurait pu se produire dans les données observées, c'est-à-dire que si l'erreur associée à l'opération de vérification g s'est réellement produite dans l'enregistrement $\mathbf{x}$ observé, alors $g(\mathbf{x})$ correspond à l'enregistrement que l'on aurait observé si cette erreur ne s'était pas produite. De façon un peu plus formelle, on présume ici que les erreurs qui surviennent dans les données peuvent être modélisées par un « processus de génération d'erreurs » stochastique $\mathcal{E}$, et que chaque opération de vérification joue le rôle de « correcteur » d'une erreur particulière qui peut se produire dans $\mathcal{E}$ (voir la remarque n° 4 à la section suivante).

Si l'opération de vérification g contient des paramètres libres, l'enregistrement $g(\mathbf{x})$ pourrait ne pas être déterminé de façon unique même lorsque les restrictions (2.1) et (3.2) sont prises en compte. Dans ce cas, il faut « imputer » des valeurs pour les paramètres libres de l'opération de vérification, ce qui signifie que certaines des variables de $\mathbf{x}$ sont imputées au moyen de la transformation affine donnée par (3.1). Comme pour la vérification traditionnelle reposant sur le paradigme de Fellegi-Holt, la recherche des « imputations » appropriées pour les paramètres libres n'est pas considérée ici comme faisant partie du problème de localisation des erreurs. En revanche, si g ne contient aucun paramètre libre, les valeurs imputées dans $g(\mathbf{x})$ découlent directement de l'opération de vérification elle-même et la distinction entre la localisation des erreurs et l'imputation devient floue.

Pour n'importe quelle application particulière, seul un petit sous-ensemble d'opérations de vérification possibles de la forme donnée en (3.1) pourrait être interprété de façon considérablement significative, au sens où l'on sait que les types d'erreur associés peuvent se produire. Dans ce qui suit, on présume qu'un

ensemble fini d'opérations de vérification spécifiques de la forme donnée en (3.1) a été déterminé comme étant pertinent pour une application particulière. Cet ensemble correspond aux *opérations de vérification autorisées* pour cette application. Des suggestions sur la façon de bâtir cet ensemble sont présentées à la section 8.

4 Un problème généralisé de localisation des erreurs

Soit $\mathcal{G}$ un ensemble fini d'opérations de vérification autorisées pour une application donnée de vérification automatique. De façon informelle, il est proposé ici de généraliser le problème de localisation des erreurs de Fellegi et Holt (1976) en remplaçant l'énoncé « le plus petit sous-ensemble de variables qui peuvent être imputées pour assurer la cohérence de l'enregistrement » par « la plus courte séquence d'opérations de vérification autorisées qui peut être appliquée pour assurer la cohérence de l'enregistrement ». Pour donner une définition formelle de ce problème généralisé de localisation des erreurs, il faut introduire de nouveaux éléments de notation et quelques concepts.

Supposons une séquence de points $\mathbf{x} = \mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t = \mathbf{y}$ appartenant à $\mathbb{R}^p$. Un *chemin* allant de $\mathbf{x}$ à $\mathbf{y}$ s'entend d'une séquence d'opérations de vérification *distinctes* $g_1, \ldots, g_t \in \mathcal{G}$ de sorte que $\mathbf{x}_n = g_n(\mathbf{x}_{n-1})$ pour tout $n \in \{1, \ldots, t\}$. (Remarque : Si g_n contient des paramètres libres, il faut interpréter cette égalité comme « il existe des valeurs des paramètres réalisables faisant en sorte que g_n établisse une correspondance entre $\mathbf{x}_{n-1}$ et $\mathbf{x}_n$ ».) Un chemin est désigné par $P = [g_1, \ldots, g_t]$. L'ensemble de tous les chemins possibles allant de $\mathbf{x}$ à $\mathbf{y}$ est désigné par $\mathcal{P}(\mathbf{x}, \mathbf{y})$. Cet ensemble peut être vide. Plus loin, on utilise $\mathcal{P}(\mathbf{x}; G)$ pour désigner, pour un sous-ensemble donné $G \subseteq \mathcal{G}$, l'ensemble de tous les chemins partant de $\mathbf{x}$ et correspondant aux opérations de vérification de G dans un certain ordre (sans spécifier les paramètres libres); si G contient t éléments, $\mathcal{P}(\mathbf{x}; G)$ contient $t!$ chemins.

Pour chaque opération de vérification $g \in \mathcal{G}$, on peut associer un poids $w_g > 0$ représentant le coût de l'application de l'opération de vérification g. Plus particulièrement, le poids d'une opération FH doit être égal au poids de confiance de la variable qu'elle impute. La *longueur* d'un chemin $P = [g_1, \ldots, g_t]$ peut donc être définie comme la somme des poids des opérations de vérification qui la constitue : $\ell(P) = \sum_{n=1}^{t} w_{g_n}$, où, par convention, le chemin vide a une longueur de zéro. La *distance* de $\mathbf{x}$ à $\mathbf{y}$ s'entend de la longueur du chemin le plus court reliant $\mathbf{x}$ à $\mathbf{y}$:

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} \min\{\ell(P) | P \in \mathcal{P}(\mathbf{x}, \mathbf{y})\} & \text{si } \mathcal{P}(\mathbf{x}, \mathbf{y}) \neq \varnothing, \\ \infty & \text{sinon.} \end{cases}$$

En règle générale, $d(\mathbf{x}, \mathbf{y})$ satisfait aux axiomes types d'un espace métrique, *sauf* qu'elle ne doit pas nécessairement être symétrique pour $\mathbf{x}$ et $\mathbf{y}$; il s'agit plutôt de ce qu'on appelle un *espace quasimétrique* (Scholtus 2014). En conséquence, $d(\mathbf{x}, \mathbf{y})$ représente « la distance de $\mathbf{x}$ à $\mathbf{y}$ » plutôt que « la distance entre $\mathbf{x}$ et $\mathbf{y}$ ».

La distance de $\mathbf{x}$ à n'importe quel sous-ensemble fermé non vide $D \subseteq \mathbb{R}^p$ s'entend de la distance jusqu'au plus proche $\mathbf{y} \in D$: $d(\mathbf{x}, D) = \min\{d(\mathbf{x}, \mathbf{y}) | \mathbf{y} \in D\}$. Aux fins de la localisation des erreurs, le

sous-ensemble fermé non vide de $\mathbb{R}^p$ présentant un intérêt particulier est l'ensemble D_0 de tous les points qui satisfont à (2.1).

On peut maintenant formuler le problème généralisé de localisation des erreurs.

Problème. Supposons un ensemble donné d'enregistrements cohérents D_0, un ensemble donné d'opérations de vérification autorisées $\mathcal{G}$ et un enregistrement donné $\mathbf{x}$. Si $d(\mathbf{x}, D_0) = \infty$, alors le problème de localisation des erreurs pour $\mathbf{x}$ est irréalisable. Sinon, n'importe quel chemin le plus court menant à un enregistrement $\mathbf{y} \in D_0$ de sorte que $d(\mathbf{x}, \mathbf{y}) < \infty$ correspond à une *solution réalisable* au problème de localisation des erreurs pour $\mathbf{x}$. Une solution réalisable est dite *optimale* si elle produit un enregistrement $\mathbf{x}^* \in D_0$ faisant en sorte que

$$d(\mathbf{x}, \mathbf{x}^*) = d(\mathbf{x}, D_0). \tag{4.1}$$

Le problème généralisé de la localisation des erreurs consiste donc officiellement à trouver un chemin optimal d'opérations de vérification.

Remarque n° 1. En règle générale, il peut y avoir un très grand nombre d'enregistrements $\mathbf{x}^*$ dans D_0 qui satisfont à (4.1) et qui peuvent être atteints par le même chemin d'opérations de vérification. Pour résoudre le problème de localisation des erreurs, il suffit de trouver un chemin optimal. La construction d'un enregistrement associé $\mathbf{x}^* \in D_0$ peut donc être considérée comme une généralisation du problème d'imputation cohérente (voir la discussion sur l'imputation à la fin de la section 3).

Remarque n° 2. Le problème de localisation des erreurs ci-dessus est irréalisable pour les enregistrements qui ne peuvent être mis en correspondance dans D_0 par aucune combinaison d'opérations de vérification distinctes de $\mathcal{G}$. Pour éviter cette situation, il faut définir un ensemble $\mathcal{G}$ suffisamment vaste pour que $d(\mathbf{x}, D_0) < \infty$ pour tout $\mathbf{x} \in \mathbb{R}^p$. Dans ce qui suit, on présume tacitement que $\mathcal{G}$ a cette propriété. Un moyen simple d'y arriver, mais pas nécessairement le seul, est de s'assurer que $\mathcal{G}$ contient au moins toutes les opérations FH. Cela est suffisant parce que deux points quelconques de $\mathbb{R}^p$ peuvent toujours être reliés par un chemin qui concatène les opérations FH associées aux coordonnées qui les différencient.

Remarque n° 3. Il n'est pas difficile de voir que le problème de localisation des erreurs ci-dessus réduit le problème original de Fellegi et Holt (1976) au cas particulier où $\mathcal{G}$ contient seulement les opérations FH.

Remarque n° 4. Comme pour le problème de localisation des erreurs original fondé sur le paradigme de Fellegi-Holt, on peut montrer que, en vertu de certaines hypothèses, la minimisation de $d(\mathbf{x}, \mathbf{y})$ pour tout $\mathbf{y} \in D_0$ pour un enregistrement observé donné $\mathbf{x}$ équivaut approximativement à la maximisation de la vraisemblance de l'enregistrement exempt d'erreur non observé associé. L'argument est sensiblement le même que celui de Kruskal (1983, pages 38-39) pour la distance dite de Levenshtein dans le contexte de l'appariement approximatif de chaînes. Pour cela, il faut d'abord que toutes les vérifications (2.1) soient des vérifications avec rejet, c'est-à-dire des vérifications auxquelles seules les valeurs erronées échouent. De plus, il faut présumer que le « processus de génération d'erreurs » stochastique $\mathcal{E}$ dont il est question à la section 3 a les propriétés suivantes :

- Il existe une correspondance biunivoque entre l'ensemble des erreurs qui peuvent se produire en vertu de $\mathcal{E}$ et l'ensemble des opérations de vérification autorisées $\mathcal{G}$ qui corrigent les erreurs.
- Les erreurs dans $\mathcal{E}$ se produisent de façon indépendante les unes des autres.
- L'erreur correspondant à l'opération g se produit selon une probabilité connue p_g.

Enfin, comme pour (2.3), les poids w_g doivent être choisis comme suit :

$$w_g = -\log\left(\frac{p_g}{1-p_g}\right). \tag{4.2}$$

En vertu de ces hypothèses, Scholtus (2014) a adapté l'argument de Kruskal (1983) pour montrer que la solution optimale au problème de localisation des erreurs (4.1) peut être justifiée comme un estimateur approximatif du maximum de vraisemblance. [Remarque : Le calcul présenté par Scholtus (2014) suppose en outre que $p_g \ll 1$ pour toutes les valeurs, auquel cas $w_g \approx -\log p_g$. Cette hypothèse n'est pas nécessaire; voir Liepins (1980).]

5 Vérifications implicites pour les opérations de vérification générales

Dans la présente section, on dérive un résultat qui détermine si un chemin donné d'opérations de vérification de la forme (3.1) peut être utilisé pour assurer la cohérence d'un enregistrement particulier avec un système de règles de vérification donné (c'est-à-dire s'il correspond à une solution réalisable au problème de localisation des erreurs). Pour obtenir ce résultat, on fait appel à la technique d'élimination FM présentée à la section 2.

Soit $\mathbf{x}$ un enregistrement donné et $\mathbf{y}_t$ un enregistrement quelconque obtenu en appliquant séquentiellement les opérations de vérification $g_1,\ldots,g_t$ à $\mathbf{x}$:

$$\mathbf{y}_t = g_t \circ g_{t-1} \circ \cdots \circ g_1(\mathbf{x}). \tag{5.1}$$

Écrivons $g_n(\mathbf{x}) = \mathbf{T}_n\mathbf{x} + \mathbf{S}_n\boldsymbol{\alpha}_n + \mathbf{c}_n$, pour $n \in \{1,\ldots,t\}$. De l'équation (5.1) ci-dessus, il découle par induction que :

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{T}_1\mathbf{x} + \mathbf{S}_1\boldsymbol{\alpha}_1 + \mathbf{c}_1,\\ \mathbf{y}_2 &= \mathbf{T}_2\mathbf{T}_1\mathbf{x} + \mathbf{S}_2\boldsymbol{\alpha}_2 + \mathbf{c}_2 + \mathbf{T}_2(\mathbf{S}_1\boldsymbol{\alpha}_1 + \mathbf{c}_1),\end{aligned}$$

et, en général,

$$\mathbf{y}_t = \mathbf{T}_t\cdots\mathbf{T}_1\mathbf{x} + \mathbf{S}_t\boldsymbol{\alpha}_t + \mathbf{c}_t + \sum_{n=2}^{t}\mathbf{T}_t\cdots\mathbf{T}_n(\mathbf{S}_{n-1}\boldsymbol{\alpha}_{n-1} + \mathbf{c}_{n-1}), \tag{5.2}$$

où la somme pour tous les n est nulle lorsque $t=1$. En outre, tous les termes comprenant $\mathbf{S}_n\boldsymbol{\alpha}_n$ disparaissent lorsque g_n ne contient aucun paramètre libre.

Le chemin des opérations de vérification $P = [g_1,\ldots,g_t]$ peut être appliqué à $\mathbf{x}$ pour obtenir un enregistrement cohérent avec les règles de vérification énoncées en (2.1) si et seulement s'il existe une valeur $\mathbf{y}_t$ de la forme (5.2) qui satisfait $\mathbf{A}\mathbf{y}_t + \mathbf{b} \odot \mathbf{0}$ et toutes les restrictions supplémentaires pertinentes de la forme (3.2) s'appliquant à $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_t$. À l'aide de (5.2), on peut écrire $\mathbf{A}\mathbf{y}_t + \mathbf{b} \odot \mathbf{0}$ comme suit :

$$(\mathbf{A}\mathbf{T}_t\cdots\mathbf{T}_1)\mathbf{x} + (\mathbf{A}\mathbf{S}_t)\boldsymbol{\alpha}_t + \sum_{n=2}^{t}(\mathbf{A}\mathbf{T}_t\cdots\mathbf{T}_n\mathbf{S}_{n-1})\boldsymbol{\alpha}_{n-1} + \mathbf{b}_t \odot \mathbf{0}, \tag{5.3}$$

où $\mathbf{b}_t = \mathbf{b} + \mathbf{A}\mathbf{c}_t + \sum_{n=2}^{t}\mathbf{A}\mathbf{T}_t\cdots\mathbf{T}_n\mathbf{c}_{n-1}$ correspond à un vecteur de constantes.

Fait intéressant, (5.3) et les restrictions supplémentaires possibles de la forme (3.2) constituent un système linéaire de la forme (2.1) appliqué à l'enregistrement élargi $(\mathbf{x}', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_t')'$. En conséquence, l'élimination FM peut servir à retirer tous les paramètres libres du système; on obtient alors un système de contraintes implicites pour $\mathbf{x}$. De plus, l'application répétée de cette propriété fondamentale de l'élimination FM établit que $\mathbf{x}$ satisfait au système de règles de vérification implicites si et seulement s'il existe des valeurs pour les paramètres $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$ qui, avec $\mathbf{x}$, satisfont (5.3) et (3.2). Il s'ensuit que le chemin des opérations de vérification $P = [g_1, \ldots, g_t]$ peut déboucher sur un enregistrement cohérent pour $\mathbf{x}$ si et seulement si $\mathbf{x}$ satisfait le système de règles de vérification implicites obtenues par l'élimination de $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$ de (5.3) et, s'il y a lieu, les restrictions supplémentaires de la forme (3.2).

Exemple. Supposons les règles de vérification suivantes dans x_1 et x_2 :

$$x_1 \geq 0, \tag{5.4}$$

$$x_2 \geq 0, \tag{5.5}$$

$$x_1 + x_2 \leq 5. \tag{5.6}$$

Soit g l'opération de vérification qui transfère un nombre d'au plus quatre unités entre x_1 et x_2, dans l'une ou l'autre direction : $g((x_1, x_2)') = (x_1 + \alpha, x_2 - \alpha)'$ où $-4 \leq \alpha \leq 4$. Pour cette opération de vérification unique, le système des règles de vérification transformées (5.3) est le suivant :

$$x_1 + \alpha \geq 0, \tag{5.7}$$

$$x_2 - \alpha \geq 0, \tag{5.8}$$

$$x_1 + x_2 \leq 5. \tag{5.9}$$

On ajoute aussi les restrictions suivantes à la forme (3.2) pour α :

$$\alpha \geq -4, \tag{5.10}$$

$$\alpha \leq 4. \tag{5.11}$$

On obtient cinq contraintes linéaires (5.7)−(5.11) pour x_1, x_2 et α, desquelles α peut être retirée par élimination FM pour obtenir :

$$x_1 \geq -4, \tag{5.12}$$

$$x_2 \geq -4, \tag{5.13}$$

$$x_1 + x_2 \geq 0, \tag{5.14}$$

$$x_1 + x_2 \leq 5. \tag{5.15}$$

En théorie, tout enregistrement $(x_1, x_2)'$ qui satisfait (5.12)−(5.15) peut être rendu cohérent avec les règles de vérification originales (5.4)−(5.6) en transférant un certain nombre d'unités $-4 \leq \alpha \leq 4$ entre x_1 et x_2. L'enregistrement $(x_1, x_2)' = (-2, 3)'$ donné en exemple n'est pas cohérent avec les règles de vérification originales (5.4)−(5.6), mais satisfait (5.12)−(5.15). Cela signifie qu'on peut rendre l'enregistrement cohérent avec les règles de vérification originales en appliquant g. On peut facilement constater que cette affirmation est vraie; n'importe quelle valeur $2 \leq \alpha \leq 3$ fera l'affaire.

Il est intéressant de souligner que, dans le cas particulier où P correspond à l'opération FH unique qui impute x_j, on obtient le système transformé de règles de vérification (5.3) en remplaçant chaque occurrence de x_j des règles de vérification originales par un paramètre non restreint α. L'élimination de α de (5.3)

équivaut dans ce cas à l'élimination directe de x_j des règles de vérification originales. En ce sens, le résultat ci-dessus généralise la propriété fondamentale de l'élimination FM pour les opérations FH à toutes les opérations de vérification de la forme (3.1).

En général, l'ensemble d'enregistrements défini par l'expression (5.2) dépend de l'ordre d'exécution des opérations de vérification. Ainsi, deux chemins composés du même ensemble d'opérations de vérification exécutées dans un ordre différent ne donnent pas nécessairement la même solution au problème de localisation des erreurs. À cet égard, les opérations de vérification générales diffèrent des opérations FH (Scholtus 2014).

6 Algorithme de localisation des erreurs

La section qui suit propose un algorithme relativement simple pour résoudre le problème de localisation des erreurs de la section 4 à l'aide du résultat théorique obtenu à la section 5.

Étape 0.	Soit $\mathbf{x}$ un enregistrement donné et $\mathcal{G}$ un ensemble donné d'opérations de vérification autorisées. Initialiser : $\mathcal{L} := \emptyset$; $\mathcal{B}_0 := \{\emptyset\}$; $W := \infty$; et $t := 1$.
Étape 1.	Déterminer tous les sous-ensembles $G \subseteq \mathcal{G}$ de cardinalité t qui satisfont aux conditions suivantes : 1. Chaque sous-ensemble de $t-1$ éléments de G appartient à $\mathcal{B}_{t-1}$. 2. Il est vérifié que $\sum_{g \in G} w_g \leq W$.
Étape 2.	Pour chaque G obtenu à l'étape 1, construire $\mathcal{P}(\mathbf{x}; G)$ et, pour chaque chemin $P \in \mathcal{P}(\mathbf{x}; G)$, déterminer s'il est possible d'obtenir un enregistrement cohérent. Dans l'affirmative : • si $\ell(P) < W$, définir $\mathcal{L} := \{P\}$ et $W := \ell(P)$; • si $\ell(P) = W$, définir $\mathcal{L} := \mathcal{L} \cup \{P\}$. Si *aucun* des chemins $P \in \mathcal{P}(\mathbf{x}; G)$ ne permet d'obtenir un enregistrement cohérent, ajouter G à $\mathcal{B}_t$.
Étape 3.	Si $t < R$ et $\mathcal{B}_t \neq \emptyset$, définir $t := t + 1$ et revenir à l'étape 1.

Figure 6.1 Algorithme pour trouver tous les chemins optimaux des opérations de vérification relatives au problème (4.1).

Dans le cadre des applications pratiques de localisation des erreurs aux fins de la statistique officielle, il arrive souvent que les enregistrements contiennent plus de 100 variables. Pour formuler un problème dont les calculs sont réalisables, les applications actuelles de vérification automatique fondée sur le paradigme de Fellegi-Holt définissent généralement une borne supérieure M au nombre de variables qui peuvent être imputées dans un même enregistrement (par exemple $M = 12$ ou $M = 15$). de Waal et Coutinho (2005) soutiennent que l'introduction d'une telle borne supérieure est raisonnable parce qu'un enregistrement qui exige plus de quinze imputations, par exemple, ne devrait pas être considéré admissible à la vérification automatique de toute manière. Selon cette convention, on peut aussi fixer une borne supérieure R au nombre d'opérations de vérification distinctes qui peuvent être appliquées à un même enregistrement. Même en appliquant cette restriction supplémentaire, l'espace de recherche des solutions possibles à (4.1) est

généralement trop vaste dans la pratique pour trouver une solution optimale au moyen d'une recherche exhaustive.

La figure 6.1 présente un résumé de l'algorithme de localisation des erreurs proposé. La formulation de base est inspirée de l'*algorithme apriori* établi par Agrawal et Srikant (1994) pour l'exploration de données. L'exécution de l'algorithme donne un ensemble $\mathcal{L}$ contenant tous les chemins d'opérations de vérification autorisées qui correspondent à une solution optimale à (4.1), ainsi que la longueur du chemin optimal W. [Remarque : Un problème de localisation des erreurs peut avoir plusieurs solutions optimales, et il peut être utile de les trouver toutes (Giles 1988; de Waal et coll. 2011, pages 66-67).]

Après la définition initiale à l'étape 0, l'algorithme passe par les étapes 1, 2 et 3 au plus R fois. À l'étape 1 de l'algorithme, l'espace de recherche est limité par ce qui suit : si G comprend un sous-ensemble approprié $H \subset G$ pour lequel $\mathcal{P}(\mathbf{x}; H)$ contient un chemin menant à un enregistrement cohérent, alors $\mathcal{P}(\mathbf{x}; G)$ ne peut contenir que des solutions sous-optimales. Ainsi, tout ensemble G contenant un tel sous-ensemble peut être ignoré par l'algorithme. De même, G peut aussi être ignoré lorsque le poids total des opérations de vérification contenues dans G est supérieur à la longueur du chemin de la meilleure solution réalisable déjà trouvée.

Durant la t^{e} itération, le nombre de sous-ensembles G obtenus à l'étape 1 de l'algorithme est égal à $\binom{N}{t}$. Pour chacun de ces sous-ensembles, les conditions de l'étape 1 doivent être vérifiées. Si un sous-ensemble de G réussit les vérifications, à l'étape 2 tous les chemins $t!$ de $\mathcal{P}(\mathbf{x}; G)$ sont évalués selon la théorie exposée à la section 5. L'algorithme apriori repose sur le principe suivant : à mesure que t augmente, la majorité des sous-ensembles échouent aux vérifications de la première étape, de sorte que le nombre total de calculs à effectuer demeure limité. Dans le contexte de l'exploration de données, ce comportement souhaitable a effectivement été observé dans les faits. La question de savoir s'il se produit aussi dans le contexte de la localisation des erreurs reste à déterminer.

Il est possible d'améliorer l'algorithme si l'on observe que l'ordre dans lequel les opérations de vérification sont exécutées n'a pas toujours d'importance. Il arrive que deux chemins de $\mathcal{P}(\mathbf{x}; G)$ soient *équivalents*, c'est-à-dire qu'un enregistrement qu'on peut atteindre à partir de $\mathbf{x}$ en empruntant le premier chemin peut aussi être atteint par le second chemin, et vice versa. Cette propriété définit une relation d'équivalence dans $\mathcal{P}(\mathbf{x}; G)$. Soit $\tilde{\mathcal{P}}(\mathbf{x}; G)$ un ensemble contenant un représentant de chaque catégorie d'équivalence de $\mathcal{P}(\mathbf{x}; G)$ en vertu de cette relation. Il est clair que l'algorithme de la figure 6.1 demeure correct si à l'étape 2 la recherche est limitée à $\tilde{\mathcal{P}}(\mathbf{x}; G)$ plutôt qu'à $\mathcal{P}(\mathbf{x}; G)$. Scholtus (2014) présente une méthode simple pour construire $\tilde{\mathcal{P}}(\mathbf{x}; G)$ à partir de $\mathcal{P}(\mathbf{x}; G)$.

Un exemple détaillé illustrant l'algorithme ci-dessus est présenté dans Scholtus (2014).

7 Étude par simulations

Pour mettre à l'essai l'utilité potentielle de la nouvelle méthode de localisation des erreurs, on a mené une étude par simulations de petite envergure dans l'environnement R pour calcul statistique (R Development Core Team 2015). Une mise en œuvre prototype de l'algorithme de la figure 6.1 a été créée dans R. Dans le cadre de cet exercice, on a largement utilisé la fonctionnalité de vérification automatique

fondée sur le paradigme de Fellegi-Holt du progiciel `editrules` (van der Loo et de Jonge 2012; de Jonge et van der Loo 2014). Le programme n'était pas optimisé pour assurer l'efficacité du calcul, mais il s'est révélé suffisamment rapide pour les problèmes de localisation des erreurs d'envergure relativement petite de l'étude par simulations. (Remarque : L'auteur peut fournir le code R utilisé sur demande.)

L'étude par simulations a été réalisée à l'aide d'enregistrements contenant cinq variables numériques qui devaient satisfaire les neuf règles de vérification linéaires suivantes :

$$\begin{aligned} x_1 + x_2 &= x_3, \\ x_3 - x_4 &= x_5, \\ x_j &\geq 0, \qquad j \in \{1,2,3,4\}, \\ x_1 &\geq x_2, \\ x_5 &\geq -0{,}1x_3, \\ x_5 &\leq 0{,}5x_3. \end{aligned}$$

On trouve généralement ce genre de règles de vérification pour les SSE, dans le cadre d'un ensemble de règles de vérification beaucoup plus vaste (Scholtus 2014).

Un ensemble aléatoire de données exempt d'erreurs contenant 2 000 enregistrements a été bâti à partir d'une distribution normale multivariée (à l'aide du progiciel `mvtnorm`) selon les paramètres suivants :

$$\boldsymbol{\mu} = \begin{pmatrix} 500 \\ 250 \\ 750 \\ 600 \\ 150 \end{pmatrix} \quad \text{et} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 10\,000 & -1\,250 & 8\,750 & 7\,500 & 1\,250 \\ -1\,250 & 5\,000 & 3\,750 & 4\,000 & -250 \\ 8\,750 & 3\,750 & 12\,500 & 11\,500 & 1\,000 \\ 7\,500 & 4\,000 & 11\,500 & 11\,750 & -250 \\ 1\,250 & -250 & 1\,000 & -250 & 1\,250 \end{pmatrix}.$$

Seuls les enregistrements satisfaisant à toutes les règles de vérification susmentionnées ont été inclus dans l'ensemble de données. Soulignons que $\boldsymbol{\Sigma}$ est une matrice singulière de covariances comprenant les deux règles de vérification fondées sur une égalité. Techniquement, les données obtenues suivent une distribution normale singulière multivariée tronquée; voir de Waal et coll. (2011, pages 318ff) ou Tempelman (2007).

Les neuf opérations de vérification autorisées retenues dans le cadre de l'étude sont présentées au tableau 7.1. Soulignons que les cinq premières lignes correspondent aux opérations FH pour cet ensemble de données. Comme il est précisé dans le tableau, chaque opération de vérification est associée à un type d'erreur. Un ensemble de données synthétiques à vérifier a été créé par l'ajout aléatoire d'erreurs de ces types à l'ensemble de données exempt d'erreur susmentionné. La probabilité de chaque type d'erreur est indiquée dans la quatrième colonne du tableau 7.1. Le poids « idéal » qui y est associé selon (4.2) est précisé dans la dernière colonne.

Pour restreindre l'ampleur des calculs à effectuer, seuls les enregistrements exigeant trois opérations de vérification ou moins ont été pris en compte. Les enregistrements ne contenant aucune erreur ont aussi été retirés. Il restait donc 1 025 enregistrements à vérifier, chacun contenant une, deux ou trois des erreurs énumérées au tableau 7.1.

Tableau 7.1
Opérations de vérification autorisées aux fins de l'étude par simulations

nom	opération	type d'erreur associé	p_g	w_g
FH1	imputer x_1	valeur erronée de x_1	0,10	2,20
FH2	imputer x_2	valeur erronée de x_2	0,08	2,44
FH3	imputer x_3	valeur erronée de x_3	0,06	2,75
FH4	imputer x_4	valeur erronée de x_4	0,04	3,18
FH5	imputer x_5	valeur erronée de x_5	0,02	3,89
IC34	intervertir x_3 et x_4	valeurs réelles de x_3 et x_4 interverties	0,07	2,59
TF21	transférer une partie de x_2 à x_1	partie de la valeur réelle de x_1 déclarée comme faisant partie de x_2	0,09	2,31
CS4	changer le signe de x_4	erreur de signe dans x_4	0,11	2,09
CS5	changer le signe de x_5	erreur de signe dans x_5	0,13	1,90

Plusieurs méthodes de localisation des erreurs ont été appliquées à l'ensemble de données. On a tout d'abord utilisé la méthode de localisation des erreurs fondée sur le paradigme de Fellegi-Holt (c'est-à-dire à l'aide des opérations de vérification FH1–FH5 uniquement) et sur le nouveau paradigme (c'est-à-dire à l'aide de toutes les opérations de vérification du tableau 7.1). Les deux méthodes ont été mises à l'essai une fois à l'aide des poids « idéaux » indiqués dans le tableau 7.1 et une fois à l'aide de poids tous fixés à 1 (« aucun poids »). Ce dernier cas simule une situation où les opérations de vérification pertinentes sont connues, mais pas leurs fréquences respectives. Enfin, pour vérifier la robustesse de la nouvelle méthode de localisation des erreurs en cas de manque d'information à propos des opérations de vérification pertinentes, la méthode a aussi été appliquée en retirant l'une des opérations non-FH du tableau 7.1 de l'ensemble des opérations de vérification autorisées.

La qualité de la localisation des erreurs a été évaluée de deux façons. Tout d'abord, on a évalué dans quelle mesure les chemins optimaux des opérations de vérification trouvés par l'algorithme correspondaient à la distribution réelle des erreurs, en utilisant le tableau de contingences ci-dessous pour toutes les $1\,025 \times 9 = 9\,225$ combinaisons possibles des enregistrements et des opérations de vérification :

Tableau 7.2
Tableau de contingences des erreurs et des opérations de vérification suggérées par l'algorithme

	opération de vérification suggérée	opération de vérification non suggérée
l'erreur associée s'est produite	VP	FN
l'erreur associée ne s'est pas produite	FP	VN

À partir de ce tableau, on a calculé des indicateurs mesurant la proportion de faux négatifs (FN), de faux positifs (FP) et de l'ensemble des mauvaises décisions, respectivement :

$$\alpha = \frac{FN}{VP + FN}; \quad \beta = \frac{FP}{FP + VN}; \quad \delta = \frac{FN + FP}{VP + FN + FP + VN}$$

Des indicateurs similaires sont présentés dans de Waal et coll. (2011, pages 410-411). On a aussi calculé $\overline{\rho} = 1 - \rho$, où ρ correspond à la fraction des enregistrements de l'ensemble de données pour lesquels l'algorithme de localisation des erreurs a trouvé exactement la bonne solution. Un bon algorithme de localisation des erreurs devrait donner des notes faibles pour les quatre indicateurs.

Il importe de souligner que les indicateurs de qualité ci-dessus désavantagent la méthode originale de Fellegi-Holt, qui ne fait pas appel à toutes les opérations de vérification énumérées au tableau 7.1. On a donc aussi calculé un deuxième ensemble d'indicateurs de qualité α, β, δ et $\overline{\rho}$ portant sur les valeurs erronées plutôt que sur les opérations de vérification. Dans ce cas, α mesure la proportion des valeurs de l'ensemble de données comportant des erreurs, mais non modifiées par la solution optimale au problème de localisation des erreurs, et de même pour les autres mesures.

Le tableau 7.3 présente les résultats de l'étude par simulations pour les deux ensembles d'indicateurs de qualité. Dans les deux cas, on constate une amélioration notable de la qualité des résultats de la localisation des erreurs de la méthode faisant appel à toutes les opérations de vérification, comparativement à la méthode utilisant uniquement les opérations FH. En outre, le fait d'omettre une seule opération de vérification pertinente de l'ensemble des opérations de vérification autorisées compromettait la qualité de la localisation des erreurs. Dans certains cas, cet effet était assez important – particulièrement en ce qui concerne les opérations de vérification utilisées –, mais les résultats de la nouvelle méthode de localisation des erreurs demeurent considérablement supérieurs à ceux de la méthode de Fellegi-Holt. Contrairement aux attentes, le fait de ne pas utiliser des poids de confiance différents a contribué à améliorer légèrement la qualité des résultats de la localisation des erreurs pour cet ensemble de données selon la méthode de Fellegi-Holt (pour les deux ensembles d'indicateurs) et aussi, dans une certaine mesure, selon la nouvelle méthode (second ensemble d'indicateurs seulement). Enfin, il semble que l'utilisation de toutes les opérations de vérification ait contribué à accroître le temps de calcul nécessaire par rapport à l'utilisation des opérations FH uniquement, mais pas de façon spectaculaire.

Tableau 7.3
Qualité de la localisation des erreurs en fonction des opérations de vérification utilisées et des valeurs erronées recensées; temps de calcul requis

	indicateurs de qualité (opérations de vérification)				indicateurs de qualité (valeurs erronées)				
méthode	α	β	δ	$\overline{\rho}$	α	β	δ	$\overline{\rho}$	**temps***
Fellegi-Holt (avec poids)	74 %	12 %	23 %	80 %	19 %	10 %	13 %	32 %	46
Fellegi-Holt (sans poids)	70 %	12 %	21 %	74 %	13 %	8 %	9 %	24 %	33
toutes les opérations (avec poids)	14 %	3 %	5 %	24 %	10 %	5 %	7 %	17 %	98
sauf IC34	29 %	5 %	9 %	35 %	15 %	9 %	11 %	29 %	113
sauf TF21	34 %	5 %	10 %	37 %	10 %	5 %	7 %	18 %	80
sauf CS4	28 %	6 %	9 %	39 %	10 %	5 %	7 %	17 %	80
sauf CS5	35 %	7 %	10 %	47 %	11 %	6 %	7 %	18 %	82
toutes les opérations (sans poids)	27 %	5 %	8 %	36 %	6 %	4 %	5 %	13 %	99

* Temps total de calcul (en secondes) sur un ordinateur portatif doté d'un processeur à 2,5 GHz sous Windows 7.

8 Conclusion

Le présent article propose une nouvelle formulation du problème de localisation des erreurs dans le contexte de la vérification automatique. On suggère de trouver le nombre minimal (pondéré) d'opérations

de vérification nécessaires pour assurer la cohérence d'un enregistrement observé avec les règles de vérification. Le nouveau problème de localisation des erreurs peut être considéré comme une généralisation du problème proposé dans l'article fondamental de Fellegi et Holt (1976), parce que l'opération qui impute une nouvelle valeur à une seule variable à la fois constitue un important cas particulier d'une opération de vérification.

L'objectif principal était de mettre au point la théorie mathématique sur laquelle repose le nouveau problème de localisation des erreurs. Il ressort que l'élimination FM, une technique utilisée par le passé pour résoudre le problème de localisation des erreurs fondé sur le paradigme de Fellegi-Holt, peut aussi être appliquée dans le contexte du nouveau problème (voir la section 5). Néanmoins, la résolution du problème de localisation des erreurs demeure une tâche difficile du point de vue du calcul, du moins pour les quantités de variables, de règles de vérification et d'opérations de vérification qui entrent en jeu dans les applications pratiques au sein des instituts de statistique. Un algorithme de localisation des erreurs possible est proposé à la section 6. D'autres algorithmes plus efficients pourraient et devraient probablement être mis au point. Comme pour l'élimination FM, il pourrait être possible d'adapter d'autres idées mises en œuvre pour résoudre le problème fondé sur le paradigme de Fellegi-Holt au problème général étudié ici.

Le présent article ne porte que sur les données numériques et les règles de vérification linéaires. Le paradigme original de Fellegi-Holt a aussi été appliqué à des données catégoriques et mixtes. Plusieurs auteurs, dont Bruni (2004) et de Jonge et van der Loo (2014), ont montré qu'une grande catégorie de règles de vérification s'appliquant à des données mixtes peuvent être reformulées en fonction de données numériques et de règles de vérification linéaires, sous réserve de la restriction supplémentaire que certaines variables doivent avoir une valeur entière. En principe, cela signifie que les résultats présentés dans l'article pourraient aussi s'appliquer à des données mixtes. Pour tenir compte du fait que certaines variables ont une valeur entière, on pourrait utiliser l'élargissement de l'élimination FM aux nombres entiers proposé par Pugh (1992); voir aussi de Waal et coll. (2011) pour en savoir davantage à propos de cette technique d'élimination élargie dans le contexte de la localisation des erreurs fondée sur le paradigme de Fellegi-Holt. Il reste à déterminer si les calculs nécessaires en vertu de cette approche sont réalisables.

La remarque n° 4 de la section 4 laisse entrevoir une analogie entre la localisation des erreurs dans les microdonnées statistiques et le domaine de l'appariement approximatif de chaînes. Dans l'appariement approximatif de chaînes, des chaînes de caractères sont comparées en vertu de l'hypothèse qu'elles pourraient avoir été partiellement corrompues (Navarro 2001). Diverses fonctions de distance ont été proposées pour cette tâche. La distance de Hamming, qui correspond au nombre de positions où deux chaînes diffèrent, peut être considérée comme analogue à la fonction cible fondée sur le paradigme de Fellegi-Holt (2.2). Le problème généralisé de localisation des erreurs défini dans le présent article peut quant à lui être considéré comme la contrepartie de l'utilisation de la distance de Levenshtein, ou « distance d'édition », pour l'appariement approximatif de chaînes. Il pourrait être intéressant d'explorer plus avant cette analogie. Plus particulièrement, des algorithmes efficients ont été mis au point pour calculer les distances d'édition entre deux chaînes; il pourrait être possible d'appliquer certaines idées sous-jacentes au problème généralisé de localisation des erreurs.

Le nouvel algorithme de localisation des erreurs a été appliqué avec succès à un petit ensemble synthétique de données (section 7). Globalement, les résultats de l'étude par simulations indiquent que la nouvelle méthode de localisation des erreurs pourrait améliorer considérablement la qualité de la vérification automatique par rapport à la méthode actuellement mise en œuvre. Il faut toutefois disposer

d'information suffisante pour déterminer toutes – ou à tout le moins la majorité – des opérations de vérification pertinentes pour une application particulière. Les gains possibles en termes de qualité de la localisation des erreurs doivent aussi être pondérés dans la pratique par rapport aux exigences supérieures de calcul du problème généralisé de localisation des erreurs.

Les SSE constituent un candidat parfait pour l'application de cette nouvelle méthode dans la pratique. Toutefois, des recherches plus poussées sont nécessaires avant que la méthode puisse être utilisée dans un contexte de production courante. Pour appliquer la méthode dans un contexte particulier, il faut d'abord préciser les opérations de vérification pertinentes. Idéalement, chaque opération de vérification doit correspondre à une combinaison de modifications aux données que les vérificateurs humains considèrent comme une correction d'une erreur en particulier. De plus, un ensemble approprié de poids w_g doit être déterminé pour ces opérations de vérification. Pour ce faire, il faut disposer d'information sur les fréquences relatives des types de modification les plus courants durant la vérification manuelle. Ces deux aspects pourraient être déterminés à partir des données historiques avant et après la vérification manuelle, des instructions de vérification et des autres sources de référence utilisées par les vérificateurs, ainsi qu'à partir d'entrevues avec des vérificateurs et des superviseurs des opérations de vérification.

Sur un plan plus fondamental, il faut encore répondre à la question de la démarcation entre les méthodes de correction déductives et la vérification automatique en vertu du nouveau problème de localisation des erreurs. En principe, bon nombre de types d'erreur connus pourraient être résolus soit par des règles de correction automatique, soit par la localisation des erreurs au moyen d'opérations de vérification. Chaque méthode présente ses propres avantages et inconvénients (Scholtus 2014). Il est probable qu'un compromis entre les deux donnera les meilleurs résultats, certaines erreurs étant traitées de façon déductive et d'autres, au moyen d'opérations de vérification. La meilleure façon d'établir un tel compromis dans la pratique demeure toutefois difficile à déterminer.

En fin de compte, la nouvelle méthode proposée dans l'article vise à accroître l'utilité de la vérification automatique dans la pratique. Les résultats obtenus à ce jour sont prometteurs.

Remerciements

Les opinions exprimées dans le présent article sont celles de l'auteur et ne reflètent pas forcément les politiques de *Statistics Netherlands*. L'auteur tient à remercier Jeroen Pannekoek, Ton de Waal et Mark van der Loo pour leurs commentaires à propos des premières versions de l'article, ainsi que le rédacteur en chef adjoint et deux évaluateurs anonymes.

Annexe

Élimination de Fourier-Motzkin

Soit un système de contraintes linéaires (2.1) et x_f, la variable à éliminer. Supposons d'abord que x_f ne participe qu'à des inégalités. Pour faciliter l'explication, supposons que les règles de vérification sont normalisées de sorte que toutes les inégalités utilisent l'opérateur $\geq$. La méthode d'élimination FM considère toutes les paires (r,s) d'inégalités pour lesquelles les coefficients de x_f ont des signes opposés, c'est-à-dire $a_{rf}a_{sf} < 0.$ Supposons, sans perte de généralité, que $a_{rf} < 0$ et $a_{sf} > 0.$ À partir de la paire

originale de règles de vérification, on dérive la contrainte implicite suivante :

$$\sum_{j=1}^{p} a_j^* x_j + b^* \geq 0, \tag{A.1}$$

où $a_j^* = a_{sf} a_{rj} - a_{rf} a_{sj}$ et $b^* = a_{sf} b_r - a_{rf} b_s$. Soulignons que $a_f^* = 0,$ de sorte que x_f ne participe pas à (A.1). Une inégalité de la forme (A.1) est dérivée de chacune des paires (r, s) susmentionnées. La totalité du système de contraintes résultant de l'élimination FM est maintenant composée de ces contraintes dérivées, ainsi que de toutes les contraintes originales dans lesquelles x_f n'intervient pas.

S'il y a des égalités linéaires où intervient x_f, on pourrait appliquer la technique ci-dessus après avoir remplacé chaque égalité linéaire par deux inégalités linéaires équivalentes. de Waal et Quere (2003) ont proposé une autre solution plus efficiente pour ce cas. Supposons que la r^{e} contrainte en (2.1) soit une égalité dans laquelle intervient x_f. On peut réécrire cette contrainte comme suit :

$$x_f = \frac{-1}{a_{rf}} \left(b_r + \sum_{j \neq f} a_{rj} x_j \right). \tag{A.2}$$

En remplaçant x_f par le terme de droite de l'équation (A.2) pour toutes les autres contraintes, on obtient de nouveau un système implicite de contraintes dans lesquelles x_f n'intervient pas et qui peuvent être réécrites comme en (2.1).

Pour consulter une preuve que l'élimination FM possède la propriété fondamentale énoncée à la section 2, voir entre autres de Waal et coll. (2011, pages 69-70).

Bibliographie

Agrawal, R., et Srikant, R. (1994). *Fast Algorithms for Mining Association Rules*. Rapport technique, IBM Almaden Research Center, San José, Californie.

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

Chen, B., Thibaudeau, Y. et Winkler, W.E. (2003). *A Comparison Study of ACS If-Then-Else, NIM, DISCRETE Edit and Imputation Systems Using ACS Data*. Document de travail n° 7, UN/ECE Work Session on Statistical Data Editing, Madrid.

de Jonge, E., et van der Loo, M. (2014). *Error Localization as a Mixed Integer Problem with the Editrules Package*. Document de discussion 2014-07, Statistics Netherlands, La Haye. Disponible au : http://www.cbs.nl.

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

de Waal, T., et Coutinho, W. (2005). Automatic editing for business surveys: An assessment for selected algorithms. *Revue Internationale de Statistique*, 73, 73-102.

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

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

EDIMBUS (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Manuel préparé par ISTAT, Statistics Netherlands, et SFSO.

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

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

Ghosh-Dastidar, B., et Schafer, J.L. (2006). Outlier detection and editing procedures for continuous multivariate data. *Journal of Official Statistics*, 22, 487-506.

Giles, P. (1988). A model for generalized edit and imputation of survey data. *The Canadian Journal of Statistics*, 16, 57-73.

Granquist, L. (1995). Improving the traditional editing process. Dans *Business Survey Methods*, (Éds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott), John Wiley & Sons, Inc., 385-401.

Granquist, L. (1997). The new view on editing. *Revue Internationale de Statistique*, 65, 381-387.

Granquist, L., et Kovar, J. (1997). Editing of survey data: How much is enough? Dans *Survey Measurement and Process Quality*, (Éds., L.E. Lyberg, P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwartz et D. Trewin), John Wiley & Sons, Inc., 415-435.

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

Hidiroglou, M.A., et Berthelot, J.-M. (1986). Contrôle statistique et imputation dans les enquêtes-entreprises périodiques. *Techniques d'enquête*, 12, 1, 79-89.

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

Kruskal, J.B. (1983). An overview of sequence comparison. Dans *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, (Éds., D. Sankoff et J.B. Kruskal), Addison-Wesley, 1-44.

Lawrence, D., et McKenzie, R. (2000). The general application of significance editing. *Journal of Official Statistics*, 16, 243-253.

Liepins, G.E. (1980). *A Rigorous, Systematic Approach to Automatic Data Editing and its Statistical Basis*. Rapport ORNL/TM-7126, Oak Ridge National Laboratory.

Liepins, G.E., Garfinkel, R.S. et Kunnathur, A.S. (1982). Error localization for erroneous data: A survey. *TIMS/Studies in the Management Sciences*, 19, 205-219.

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

Naus, J.I., Johnson, T.G. et Montalvo, R. (1972). A probabilistic model for identifying errors in data editing. *Journal of the American Statistical Association*, 67, 943-950.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33, 31-88.

Pannekoek, J., Scholtus, S. et van der Loo, M. (2013). Automated and manual data editing: A view on process design and methodology. *Journal of Official Statistics*, 29, 511-537.

Pugh, W. (1992). The omega test: A fast and practical integer programming algorithm for data dependence analysis. *Communications of the ACM*, 35, 102-114.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienne, Autriche : R Foundation for Statistical Computing. URL: http://www.R-project.org/.

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

Riera-Ledesma, J., et Salazar-González, J.J. (2003). *New Algorithms for the Editing and Imputation Problem*. Document de travail n^{o} 5, UN/ECE Work Session on Statistical Data Editing, Madrid.

Riera-Ledesma, J., et Salazar-González, J.J. (2007). A branch-and-cut algorithm for the continuous error localization problem in data cleaning. *Computers & Operations Research*, 34, 2790-2804.

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

Scholtus, S. (2011). Algorithms for correcting sign errors and rounding errors in business survey data. *Journal of Official Statistics*, 27, 467-490.

Scholtus, S. (2014). *Error Localisation using General Edit Operations*. Document de discussion 2014-14, Statistics Netherlands, La Haye. Disponible au : http://www.cbs.nl.

Tempelman, D.C.G. (2007). *Imputation of Restricted Data*. Thèse de doctorat, University of Groningen. Disponible au : http://www.cbs.nl.

van der Loo, M., et de Jonge, E. (2012). *Automatic Data Editing with Open Source R*. Document de travail n^{o} 33, UN/ECE Work Session on Statistical Data Editing, Oslo.

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

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 21-43
Statistique Canada, nº 12-001-X au catalogue

Appariement statistique par imputation fractionnaire

Jae Kwang Kim, Emily Berg et Taesung Park[1]

Résumé

L'appariement statistique est une technique permettant d'intégrer deux ou plusieurs ensembles de données lorsque les renseignements nécessaires pour apparier les enregistrements des participants individuels dans les ensembles de données sont incomplets. On peut considérer l'appariement statistique comme un problème de données manquantes en vertu duquel on souhaite effectuer une analyse conjointe de variables qui ne sont jamais observées ensemble. On utilise souvent une hypothèse d'indépendance conditionnelle pour créer des données imputées aux fins d'appariement statistique. Nous examinons une approche générale de l'appariement statistique faisant appel à l'imputation fractionnaire paramétrique de Kim (2011) pour créer des données imputées en vertu de l'hypothèse que le modèle spécifié est entièrement identifié. La méthode proposée ne produit pas une séquence espérance-maximisation (EM) convergente si le modèle n'est pas identifié. Nous présentons aussi des estimateurs de variance convenant à la procédure d'imputation. Nous expliquons comment la méthode s'applique directement à l'analyse des données obtenues à partir de plans de sondage à questionnaire scindé et aux modèles d'erreur de mesure.

Mots-clés : Combinaison de données; fusion de données; imputation hot deck; plan de sondage à questionnaire scindé; modèle d'erreur de mesure.

1 Introduction

L'échantillonnage d'enquête est un outil scientifique permettant de faire des inférences à propos de la population cible. Toutefois, il arrive souvent que toutes les données nécessaires ne soient pas recueillies dans le cadre d'une même enquête, à cause de contraintes de temps et de coût. Dans ce cas, on souhaite exploiter le plus possible les données existantes provenant d'autres sources portant sur la même population cible. L'appariement statistique, que l'on appelle parfois « fusion de données » (Baker, Harris et O'Brien 1989) ou « combinaison de données » (Ridder et Moffit 2007), vise à intégrer deux ou plusieurs ensembles de données lorsque les renseignements nécessaires pour apparier les enregistrements des participants individuels dans les ensembles de données sont incomplets. D'Orazio, Zio et Scanu (2006) ainsi que Leulescu et Agafitei (2013) présentent un bon aperçu des techniques d'appariement statistique dans l'échantillonnage d'enquête.

L'appariement statistique peut être considéré comme un problème de données manquantes en vertu duquel on souhaite effectuer une analyse conjointe de variables qui ne sont jamais observées ensemble. Moriarity et Scheuren (2001) proposent un cadre théorique pour l'appariement statistique en vertu d'une hypothèse de normalité multivariée. Rässler (2002) a mis au point des techniques d'imputation multiple pour l'appariement statistique à l'aide de valeurs prédéterminées pour les paramètres non identifiables. Lahiri et Larsen (2005) traitent de l'analyse par régression à l'aide de données couplées. Ridder et Moffit (2007) présentent un traitement rigoureux des hypothèses et des approches pour l'appariement statistique dans le domaine de l'économétrie.

L'appariement statistique vise à construire des fichiers de données entièrement augmentées pour effectuer des analyses conjointes statistiquement valides. Pour simplifier la mise en situation, supposons

1. Jae Kwang Kim, Département de statistique, Iowa State University, Ames, IA 50011, États-Unis. Courriel : jkim@iastate.edu; Emily Berg, Département de statistique, Iowa State University, Ames, Iowa, États-Unis. Courriel : emilyb@iastate.edu; Taesung Park, Département de statistique, Université nationale de Séoul, Séoul, Corée. Courriel : taesungp@gmail.com.

que deux enquêtes, l'enquête A et l'enquête B, offrent des données partielles à propos de la population, et que l'on observe x et y_1 dans l'échantillon de l'enquête A et x et y_2 dans l'échantillon de l'enquête B. Le tableau 1.1 illustre une structure de données simple pour l'appariement. Si l'échantillon de l'enquête B (échantillon B) est un sous-ensemble de l'échantillon de l'enquête A (échantillon A), on peut employer les techniques de couplage d'enregistrements (Herzog, Scheuren et Winkler 2007) pour obtenir les valeurs de y_1 pour l'échantillon de l'enquête B. Toutefois, dans de nombreux cas, un tel appariement parfait n'est pas possible (par exemple, parce que les échantillons peuvent contenir des sous-ensembles non chevauchants); on dépend alors d'une méthode probabiliste d'identification des « jumeaux statistiques » de l'autre échantillon, c'est-à-dire que l'on doit créer y_1 pour chaque élément de l'échantillon B en trouvant son plus proche voisin dans l'échantillon A. L'imputation par la méthode du plus proche voisin a été examinée par de nombreux auteurs, dont Chen et Shao (2001) et Beaumont et Bocci (2009), dans le contexte des réponses manquantes.

Tableau 1.1
Structure de données simple pour l'appariement

	$\boldsymbol{X}$	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$
Échantillon A	o	o	
Échantillon B	o		o

La détermination du plus proche voisin repose souvent sur la « proximité » en fonction de la valeur de x seulement. Ainsi, dans de nombreux cas, l'appariement statistique est fondé sur l'hypothèse que y_1 et y_2 sont indépendants, conditionnellement à x, c'est-à-dire

$$y_1 \perp y_2 \,|\, x. \tag{1.1}$$

L'hypothèse (1.1) est souvent appelée « hypothèse d'indépendance conditionnelle (IC) » et est très utilisée dans la pratique.

Dans le présent article, nous examinons une autre approche, qui ne repose pas sur l'hypothèse d'IC. Nous présentons les hypothèses à la section 2, puis les méthodes proposées à la section 3. Nous examinons en outre deux extensions de l'approche, l'une aux plans de sondage à questionnaire scindé (section 4) et l'autre aux modèles d'erreur de mesure (section 5). Les résultats de deux études par simulation sont présentés à la section 6. La section 7 conclut l'article.

2 Scénario de base

Pour simplifier la présentation, nous considérons deux enquêtes indépendantes réalisées auprès de la même population cible consistant en N éléments. Comme il est précisé à la section 1, supposons que l'échantillon A comporte des données uniquement à propos de x et de y_1 et que l'échantillon B comporte des données uniquement à propos de x et de y_2.

Pour illustrer cette idée, supposons pour l'instant que les variables (x, y_1, y_2) sont générées à partir d'une distribution normale comme suit :

$$\begin{pmatrix} x \\ y_1 \\ y_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{1x} & \sigma_{2x} \\ & \sigma_{11} & \sigma_{12} \\ & & \sigma_{22} \end{pmatrix}\right].$$

Selon la structure de données présentée dans le tableau 1.1, il est clair que le paramètre σ_{12} ne peut pas être estimé à partir des échantillons. Il découle de l'hypothèse d'indépendance conditionnelle énoncée en (1.1) que $\sigma_{12} = \sigma_{1x}\sigma_{2x} / \sigma_{xx}$ et que $\rho_{12} = \rho_{1x}\rho_{2x}$, c'est-à-dire que σ_{12} est entièrement déterminé à partir d'autres paramètres, plutôt qu'estimé directement à partir des échantillons réalisés.

Dans ce cas, l'imputation de données synthétiques en vertu de l'hypothèse d'indépendance conditionnelle peut se faire en deux étapes :

[Étape 1] Estimer $f(y_1 \mid x)$ à partir de l'échantillon A, et désigner l'estimation $\hat{f}_a(y_1 \mid x)$.

[Étape 2] Pour chaque élément i de l'échantillon B, utiliser la valeur de x_i pour générer des valeurs imputées de y_1 à partir de $\hat{f}_a(y_1 \mid x_i)$.

Comme les valeurs de y_1 ne sont jamais observées dans l'échantillon B, des valeurs synthétiques de y_1 sont créées pour tous les éléments de l'échantillon B, ce qui donne lieu à une imputation synthétique. Haziza (2009) présente un bon examen des publications relatives à la méthodologie d'imputation. Kim et Rao (2012) présentent une approche assistée par modèle pour l'imputation synthétique lorsque seul x est disponible dans l'échantillon B. Une telle imputation synthétique ne tient absolument pas compte des données observées pour y_2 dans l'échantillon B.

L'appariement statistique fondé sur l'indépendance conditionnelle suppose que $\mathrm{Cov}(y_1, y_2 \mid x) = 0$. Ainsi, la régression de y_2 sur x et y_1 à partir des données imputées issues de l'imputation synthétique ci-dessus estimera un coefficient de régression nul pour y_1. Autrement dit, l'estimation $\hat{\beta}_2$ pour

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y_1$$

donnera une valeur nulle. De telles analyses peuvent être trompeuses si l'IC n'est pas vérifiée. Pour comprendre pourquoi, posons un problème de régression à variable omise :

$$\begin{aligned} y_1 &= \beta_0^{(1)} + \beta_1^{(1)} x + \beta_2^{(1)} z + e_1 \\ y_2 &= \beta_0^{(2)} + \beta_1^{(2)} x + \beta_2^{(2)} z + e_2 \end{aligned}$$

où les variables z, e_1, e_2 sont indépendantes et non observées. Sauf si $\beta_2^{(1)} = \beta_2^{(2)} = 0$, la variable latente z est un facteur de confusion non observable qui explique pourquoi $\mathrm{Cov}(y_1, y_2 \mid x) \neq 0$. Ainsi, le coefficient de y_1 dans la régression de la population de y_2 sur x et y_1 n'est pas nul.

Soulignons que l'hypothèse d'IC concerne l'identification du modèle. L'hypothèse de variable instrumentale (VI) constitue une autre hypothèse d'identification, décrite ci-dessous.

Remarque 2.1 *Nous présentons une description formelle de l'hypothèse de VI. D'abord, présumons que* x *se décompose en* $x = (x_1, x_2)$ *de sorte que*

(*i*) $f(y_2 | x_1, x_2, y_1) = f(y_2 | x_2, y_1)$

(*ii*) $f(y_1 | x_2, x_1 = a) \neq f(y_1 | x_2, x_1 = b)$

pour $a \neq b$. *Ainsi,* x_1 *est conditionnellement indépendante de* y_2 *sachant* x_2 *et* y_1, *mais* x_1 *est corrélée avec* y_1 *sachant* x_2. *Soulignons que* x_2 *peut être nulle ou avoir une distribution dégénérée, par exemple une ordonnée à l'origine. La variable* x_1 *satisfaisant aux deux conditions ci-dessus est souvent appelée une variable instrumentale (VI) pour* y_1. *Le graphe acyclique orienté de la figure 2.1 illustre la structure de dépendance d'un modèle assorti d'une variable instrumentale. Ridder et Moffit (2007) ont utilisé des « contraintes d'exclusion » pour décrire l'hypothèse de variable instrumentale. Les enquêtes répétées sont un exemple de cas où il est raisonnable de poser une hypothèse de variable instrumentale. Supposons une enquête répétée où* y_t *est la variable étudiée à l'année* t *et vérifie la propriété de Markov*

$$P(y_{t+1} | y_1, \ldots, y_t) = P(y_{t+1} | y_t),$$

où $P(y_t)$ *désigne une fonction de distribution cumulative. Dans ce cas,* y_{t-1} *est une variable instrumentale pour* y_t. *En fait, la dernière observation de* y_s $(s \leq t)$, *quelle qu'elle soit, est la variable instrumentale pour* y_t.

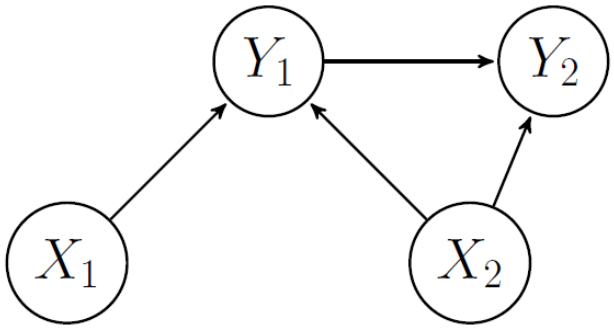

Figure 2.1 Structure de dépendance pour un modèle où x_1 est une variable instrumentale pour y_1 et où x_2 est une covariable supplémentaire dans les modèles pour y_2 et y_1.

En vertu de l'hypothèse de variable instrumentale, on peut utiliser une régression en deux étapes pour estimer les paramètres de régression d'un modèle linéaire. L'exemple suivant présente les concepts de base.

Exemple 2.1 *Prenons la structure des données de deux échantillons présentée dans le tableau 1.1. On présume le modèle de régression linéaire suivant :*

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + \beta_2 x_{2i} + e_i, \tag{2.1}$$

où $e_i \sim (0, \sigma_e^2)$ *et* e_i *est indépendante de* (x_{1j}, x_{2j}, y_{1j}) *pour toutes les valeurs de* i, j. *Dans ce cas, on peut obtenir un estimateur convergent de* $\beta = (\beta_0, \beta_1, \beta_2)'$ *à l'aide de la méthode des moindres carrés en deux étapes (MC2E) comme suit :*

1. *À partir de l'échantillon A, on ajuste le « modèle de travail » suivant pour* y_1

$$y_{1i} = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i, \quad u_i \sim (0, \sigma_u^2) \tag{2.2}$$

pour obtenir un estimateur convergent de $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$ *défini par*

$$\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)' = (X'X)^{-1} X'Y_1$$

où $X = [X_0, X_1, X_2]$ *est une matrice dont la* i^e *ligne est* $(1, x_{1i}, x_{2i})$ *et* Y_1 *est un vecteur dont* y_{1i} *est la* i^e *composante.*

2. *On obtient un estimateur convergent de* $\beta = (\beta_0, \beta_1, \beta_2)'$ *à l'aide de la méthode des moindres carrés pour la régression de* y_{2i} *sur* $(1, \hat{y}_{1i}, x_{2i})$ *où* $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$.

La question de l'absence asymptotique de biais de l'estimateur par les MC2E en vertu de l'hypothèse de variable instrumentale est abordée à l'annexe A. La méthode des MC2E n'est pas applicable directement si le modèle de régression (2.1) n'est pas linéaire. En outre, bien que la méthode des MC2E permette d'estimer les paramètres de régression, elle ne fournit pas des estimateurs convergents pour les paramètres plus généraux comme $\theta = \Pr(y_2 < 1 | y_1 < 3)$. L'imputation stochastique peut constituer une solution pour estimer une classe plus générale de paramètres. Nous expliquons comment modifier l'imputation fractionnaire paramétrique de Kim (2011) pour effectuer une estimation générale dans le contexte d'un problème d'appariement statistique.

3 Imputation fractionnaire

Nous allons maintenant décrire les méthodes d'imputation fractionnaire aux fins d'appariement statistique sans avoir recours à l'hypothèse d'IC. L'utilisation de l'imputation fractionnaire pour l'appariement statistique a été présentée pour la première fois dans le chapitre 9 de Kim et Shao (2013) en vertu de l'hypothèse de VI. Dans le présent article, nous présentons la méthodologie sans recourir à l'hypothèse de VI. Nous présumons seulement que le modèle spécifié est entièrement identifié. L'identifiabilité du modèle spécifié peut facilement être vérifiée dans le calcul de la procédure proposée.

Pour expliquer l'idée, rappelons que la variable y_1 est absente de l'échantillon B et que le but est de générer y_1 à partir de la distribution conditionnelle de y_1 sachant les observations. Autrement dit, nous voulons générer y_1 à partir de

$$f(y_1 | x, y_2) \propto f(y_2 | x, y_1) f(y_1 | x). \tag{3.1}$$

Pour ce faire, on peut utiliser la stratégie d'imputation en deux étapes suivante :

1. Générer y_1^* à partir de $\hat{f}_a(y_1 | x)$.

2. Accepter y_1^* si $f\left(y_2 \mid x, y_1^*\right)$ est suffisamment grande.

Soulignons que la première étape est la méthode habituelle en vertu de l'hypothèse d'IC. La deuxième étape intègre l'information dans y_2. Pour déterminer si $f\left(y_2 \mid x, y_1^*\right)$ est suffisamment grande à l'étape 2, on applique souvent une méthode Monte Carlo par chaîne de Markov (MCMC), par exemple l'algorithme de Metropolis-Hastings (Chib et Greenberg 1995). Soit $y_1^{(t-1)}$ la valeur courante de y_1 dans la chaîne de Markov; on accepte alors y_1^* selon la probabilité

$$R\left(y_1^*, y_1^{(t-1)}\right) = \min\left\{1, \frac{f\left(y_2 \mid x, y_1^*\right)}{f\left(y_2 \mid x, y_1^{(t-1)}\right)}\right\}.$$

De tels algorithmes peuvent devenir fastidieux à calculer, à cause de la convergence lente de l'algorithme MCMC.

L'imputation fractionnaire paramétrique de Kim (2011) permet de générer les valeurs imputées en (3.1) sans recourir à la méthode MCMC. On peut utiliser l'algorithme espérance-maximisation (EM) par imputation fractionnaire suivant :

1. Pour tout $i \in B$, générer m valeurs imputées de y_{1i}, désignées par $y_{1i}^{*(1)}, \ldots, y_{1i}^{*(m)}$, à partir de $\hat{f}_a(y_1 \mid x_i)$, où $\hat{f}_a(y_1 \mid x)$ correspond à la densité estimée pour la distribution conditionnelle de y_1 sachant x obtenue à partir de l'échantillon A.
2. Soit $\hat{\theta}_t$ la valeur courante du paramètre θ dans $f(y_2 \mid x, y_1)$. Pour la j^{e} valeur imputée $y_{1i}^{*(j)}$, affecter le poids fractionnaire

   $$w_{ij(t)}^* \propto f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right)$$

   de sorte que $\sum_{j=1}^m w_{ij}^* = 1$.
3. Résoudre l'équation de score obtenue par imputation fractionnaire pour θ

   $$\sum_{i \in B} w_{ib} \sum_{j=1}^m w_{ij(t)}^* S\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0 \tag{3.2}$$

   pour obtenir $\hat{\theta}_{t+1}$, où $S(\theta; x, y_1, y_2) = \partial \log f(y_2 \mid x, y_1; \theta) / \partial \theta$, et w_{ib} correspond au poids d'échantillonnage de l'unité i dans l'échantillon B.
4. Reprendre à l'étape 2 et poursuivre jusqu'à la convergence.

Une fois le modèle identifié, la séquence EM obtenue à partir de la méthode d'imputation fractionnaire paramétrique ci-dessus converge. Si le modèle spécifié n'est pas identifiable, c'est qu'il n'y a pas de solution unique pour maximiser la vraisemblance observée et la séquence EM ci-dessus ne converge pas. Soulignons qu'en (3.2), pour une valeur suffisamment grande de m,

$$\begin{aligned}\sum_{j=1}^m w_{ij(t)}^* S\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) &\cong \frac{\int S(\theta; x_i, y_1, y_{2i})\, f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a(y_1 \mid x_i)\, dy_1}{\int f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a(y_1 \mid x_i)\, dy_1} \\ &= E\left\{S(\theta; x_i, Y_1, y_{2i}) \mid x_i, y_{2i}; \hat{\theta}_t\right\}.\end{aligned}$$

Si y_{i1} est catégorique, le poids fractionnaire peut être établi par la probabilité conditionnelle correspondant à la valeur imputée réalisée (Ibrahim 1990). On a recours à l'étape 2 pour intégrer les données observées de y_{i2} dans l'échantillon B. Précisons que l'étape 1 n'est pas répétée pour chaque itération. Seules les étapes 2 et 3 sont reprises jusqu'à ce qu'il y ait convergence. Comme l'étape 1 n'est pas répétée, la convergence est garantie et la vraisemblance observée augmente, à condition que le modèle soit identifiable (voir le théorème 2 de Kim [2011]).

Remarque 3.1 *À la section 2, il est question de VI uniquement parce que c'est la façon la plus répandue d'assurer l'identifiabilité. La méthode proposée ici ne dépend pas de cette hypothèse. Pour illustrer une situation où il est possible d'identifier le modèle sans l'hypothèse de VI, supposons le modèle suivant :*

$$\begin{aligned} y_2 &= \beta_0 + \beta_1 x + \beta_2 y_1 + e_2 \\ y_1 &= \alpha_0 + \alpha_1 x + e_1 \end{aligned}$$

où $e_1 \sim N\left(0, x^2\sigma_1^2\right)$ *et* $e_2 \,|\, e_1 \sim N\left(0, \sigma_2^2\right)$. *Alors*

$$f\left(y_2 \,|\, x\right) = \int f\left(y_2 \,|\, x, y_1\right) f\left(y_1 \,|\, x\right) dy_1$$

est aussi une distribution normale de moyenne $\left(\beta_0 + \beta_2\alpha_0\right) + \left(\beta_1 + \beta_2\alpha_1\right)x$ *et de variance* $\sigma_2^2 + \beta_2^2\sigma_1^2x^2$. *En vertu de la structure de données présentée dans le tableau 1.1, on peut identifier ce modèle sans poser l'hypothèse de VI. L'hypothèse d'absence d'interaction entre* y_1 *et* x *dans le modèle pour* y_2 *est essentielle pour s'assurer que le modèle est identifiable.*

Au lieu de générer $y_{1i}^{*(j)}$ à partir de $\hat{f}_a\left(y_1 \,|\, x_i\right)$, on peut utiliser une méthode d'imputation fractionnaire hot deck (IFHD), en vertu de laquelle toutes les valeurs observées de y_{1i} dans l'échantillon A sont utilisées comme valeurs imputées. Dans ce cas, les poids fractionnaires de l'étape 2 sont donnés par

$$w_{ij}^*\left(\hat{\theta}_t\right) \propto w_{ij0}^* f\left(y_{2i} \,|\, x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right),$$

où

$$w_{ij0}^* = \frac{\hat{f}_a\left(y_{1j} \,|\, x_i\right)}{\sum_{k\in A} w_{ka}\,\hat{f}_a\left(y_{1j} \,|\, x_k\right)}. \tag{3.3}$$

Le poids fractionnaire initial w_{ij0}^* en (3.3) est calculé par l'application d'une pondération préférentielle à l'aide de

$$\hat{f}_a\left(y_{1j}\right) = \int \hat{f}_a\left(y_{1j} \,|\, x\right) \hat{f}_a\left(x\right) dx \propto \sum_{i\in A} w_{ia}\,\hat{f}_a\left(y_{1j} \,|\, x_i\right)$$

comme densité proposée pour y_{1j}. L'étape de maximisation est la même que pour l'imputation fractionnaire paramétrique. Pour en savoir davantage à propos de l'IFHD, voir Kim et Yang (2014). Dans la pratique, on peut utiliser une seule valeur imputée pour chaque unité. Dans ce cas, les poids fractionnaires peuvent être utilisés comme probabilité de sélection dans l'échantillonnage avec probabilité proportionnelle à la taille (PPT) de taille $m = 1$.

Pour estimer la variance, on peut utiliser une méthode de linéarisation ou une méthode de ré-échantillonnage. On examine d'abord l'estimation de la variance pour l'estimateur du maximum de vraisemblance (EMV) de θ. Si on a recours à un modèle paramétrique $f(y_1|x) = f(y_1|x;\theta_1)$ et $f(y_2|x,y_1;\theta_2)$, on obtient l'EMV de $\theta = (\theta_1,\theta_2)$ en résolvant

$$\left[S_1(\theta_1), \bar{S}_2(\theta_1,\theta_2)\right] = (0,0), \tag{3.4}$$

où $S_1(\theta_1) = \sum_{i\in A} w_{ia} S_{i1}(\theta_1)$, $S_{i1}(\theta_1) = \partial \log f(y_{1i}|x_i;\theta_1)/\partial\theta_1$ est la fonction de score de θ_1,

$$\bar{S}_2(\theta_1,\theta_2) = E\{S_2(\theta_2)|X,Y_2;\theta_1,\theta_2\},$$

$S_2(\theta_2) = \sum_{i\in B} w_{ib} S_{i2}(\theta_2)$, et $S_{i2}(\theta_2) = \partial \log f(y_{2i}|x_i,y_{1i};\theta_2)/\partial\theta_2$ est la fonction de score de θ_2. Soulignons qu'on peut écrire $\bar{S}_2(\theta_1,\theta_2) = \sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2)|x_i,y_{2i};\theta\}$. Ainsi,

$$\begin{aligned}
\frac{\partial}{\partial\theta_1'}\bar{S}_2(\theta) &= \sum_{i\in B} w_{ib}\frac{\partial}{\partial\theta_1'}\left[\frac{\int S_{i2}(\theta_2) f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}{\int f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}\right] \\
&= \sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2) S_{i1}(\theta_1)'|x_i,y_{2i};\theta\} \\
&\quad -\sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2)|x_i,y_{2i};\theta\} E\{S_{i1}(\theta_1)'|x_i,y_{2i};\theta\}
\end{aligned}$$

et

$$\begin{aligned}
\frac{\partial}{\partial\theta_2'}\bar{S}_2(\theta) &= \sum_{i\in B} w_{ib}\frac{\partial}{\partial\theta_2'}\left[\frac{\int S_{i2}(\theta_2) f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}{\int f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}\right] \\
&= \sum_{i\in B} w_{ib} E\left\{\frac{\partial}{\partial\theta_2'} S_{i2}(\theta_2)|x_i,y_{2i};\theta\right\} \\
&\quad +\sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2) S_{i2}(\theta_2)'|x_i,y_{2i};\theta\} \\
&\quad -\sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2)|x_i,y_{2i};\theta\} E\{S_{2i}(\theta_2)'|x_i,y_{2i};\theta\}.
\end{aligned}$$

Maintenant, on peut estimer de manière convergente $\partial\bar{S}_2(\theta)/\partial\theta_1'$ par

$$\hat{B}_{21} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* S_{2ij}^*\left(\hat{\theta}_2\right)\left\{S_{1ij}^*\left(\hat{\theta}_1\right) - \bar{S}_{1i}^*\left(\hat{\theta}_1\right)\right\}', \tag{3.5}$$

où $S_{1ij}^*\left(\hat{\theta}_1\right) = S_1\left(\hat{\theta}_1; x_i, y_{1i}^{*(j)}\right)$, $S_{2ij}^*\left(\hat{\theta}_2\right) = S_2\left(\hat{\theta}_2; x_i, y_{1i}^{*(j)}, y_{2i}\right)$, et $\bar{S}_{1i}^*\left(\hat{\theta}_1\right) = \sum_{j=1}^{m} w_{ij}^* S_1\left(\hat{\theta}_1; x_i, y_{1i}^{*(j)}\right)$. De plus, on peut estimer $\partial\bar{S}_2(\theta)/\partial\theta_2'$ de manière convergente par

$$-\hat{I}_{22} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* \dot{S}_{2ij}^*\left(\hat{\theta}_2\right) - \hat{B}_{22} \tag{3.6}$$

où

$$\hat{B}_{22} = \sum_{i\in B} w_{ib} \sum_{j=1}^{m} w_{ij}^{*} S_{2ij}^{*}\left(\hat{\theta}_2\right)\left\{S_{2ij}^{*}\left(\hat{\theta}_2\right) - \overline{S}_{2i}^{*}\left(\hat{\theta}_2\right)\right\}',$$

$\dot{S}_{2ij}^{*}(\theta_2) = \partial S_2\left(\theta_2; x_i, y_{1i}^{*(j)}, y_{2i}\right)\big/\partial\theta_2'$ et $\overline{S}_{2i}^{*}(\theta_2) = \sum_{j=1}^{m} w_{ij}^{*} S_{2ij}^{*}(\theta_2)$.

En effectuant un développement en série de Taylor par rappport à θ_1,

$$\begin{aligned}\overline{S}_2\left(\hat{\theta}_1, \theta_2\right) &\cong \overline{S}_2(\theta_1, \theta_2) - E\left\{\frac{\partial}{\partial\theta_1'}\overline{S}_2(\theta)\right\}\left[E\left\{\frac{\partial}{\partial\theta_1'}S_1(\theta_1)\right\}\right]^{-1} S_1(\theta_1)\\ &= \overline{S}_2(\theta) + KS_1(\theta_1),\end{aligned}$$

on peut écrire

$$V\left(\hat{\theta}_2\right) \doteq \left\{E\left(\frac{\partial}{\partial\theta_2'}\overline{S}_2\right)\right\}^{-1} V\left\{\overline{S}_2(\theta) + KS_1(\theta_1)\right\}\left\{E\left(\frac{\partial}{\partial\theta_2'}\overline{S}_2\right)\right\}^{-1'}.$$

En écrivant

$$\overline{S}_2(\theta) = \sum_{i\in B} w_{ib}\overline{s}_{2i}(\theta),$$

où $\overline{s}_{2i}(\theta) = E\left\{S_{i2}(\theta_2)\,\middle|\, x_i, y_{2i}; \theta\right\}$, on peut obtenir un estimateur convergent de $V\left\{\overline{S}_2(\theta)\right\}$ en appliquant un estimateur de la variance convergent par rapport au plan à $\sum_{i\in B} w_{ib}\hat{s}_{2i}$ où $\hat{s}_{2i} = \sum_{j=1}^{m} w_{ij}^{*} S_{2ij}^{*}\left(\hat{\theta}_2\right)$. En vertu d'un échantillonnage aléatoire simple pour l'échantillon B, on obtient

$$\hat{V}\left\{\overline{S}_2(\theta)\right\} = n_B^{-2}\sum_{i\in B}\hat{s}_{2i}\hat{s}_{2i}'.$$

De plus, on peut estimer de manière convergente $V\left\{KS_1(\theta_1)\right\}$ par

$$\hat{V}_2 = \hat{K}\hat{V}(S_1)\hat{K}',$$

où $\hat{K} = \hat{B}_{21}\hat{I}_{11}^{-1}$, $\hat{B}_{21}$ est défini selon l'équation (3.5), et $\hat{I}_{11} = -\partial S_1(\theta_1)\big/\partial\theta_1'$ est évalué à $\theta_1 = \hat{\theta}_1$. Comme les deux termes $\overline{S}_2(\theta)$ et $S_1(\theta_1)$ sont indépendants, on peut estimer la variance par

$$\hat{V}\left(\hat{\theta}\right) \doteq \hat{I}_{22}^{-1}\left[\hat{V}\left\{\overline{S}_2(\theta)\right\} + \hat{V}_2\right]\hat{I}_{22}^{-1'},$$

où $\hat{I}_{22}$ est défini selon l'équation (3.6).

De façon plus générale, on pourrait considérer l'estimation d'un paramètre η défini comme une racine de l'équation d'estimation de recensement $\sum_{i=1}^{N} U(\eta; x_i, y_{1i}, y_{2i}) = 0$. L'estimation de la variance de l'estimateur par imputation fractionnaire de η calculée à partir de $\sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^{*} U\left(\eta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0$ est présentée à l'annexe B.

4 Plan de sondage à questionnaire scindé

À la section 3, on examine le cas où l'échantillon A et l'échantillon B sont deux échantillons indépendants de la même population cible. Nous allons maintenant examiner un autre cas, celui d'un plan de sondage à questionnaire scindé en vertu duquel l'échantillon initial S est sélectionné à partir d'une population cible, puis l'échantillon A et l'échantillon B sont sélectionnés au hasard de sorte que $A \cup B = S$ et $A \cap B = \phi$. On observe (x, y_1) dans l'échantillon A et (x, y_2) dans l'échantillon B. On souhaite créer des données entièrement augmentées avec observation de (x, y_1, y_2) dans S.

De tels plans de sondage à questionnaire scindé gagnent en popularité parce qu'ils réduisent le fardeau de réponse (Raghunathan et Grizzle 1995; Chipperfield et Steel 2009). Des plans de sondage à questionnaire scindé ont notamment été explorés dans le cadre de la *Consumer Expenditure Survey* (Gonzalez et Eltinge 2008) et de la *National Assessment of Educational Progress (NAEP) Survey* aux États-Unis. Les analystes qui utilisent les résultats des enquêtes à questionnaire scindé peuvent s'intéresser à des paramètres multiples, comme la moyenne pour y_1 et la moyenne pour y_2, en plus du coefficient de la régression de y_2 sur y_1.

Nous avons examiné un plan de sondage où l'échantillon initial S est divisé en deux sous-échantillons : A et B. On suppose que x_i est observé pour $i \in S$, que y_{1i} est recueilli pour $i \in A$ et que y_{2i} est recueilli pour $i \in B$. La probabilité de sélection dans A ou B peut dépendre de x_i mais ne dépend pas de y_{1i} ni de y_{2i}. En conséquence, le plan de sondage utilisé pour sélectionner les sous-échantillons A et B est non informatif pour le modèle spécifié (Fuller 2009, chapitre 6). Soit w_i le poids d'échantillonnage associé à l'échantillon complet S. On suppose qu'il existe une procédure pour estimer la variance d'un estimateur de la forme $\hat{Y} = \sum_{i \in S} w_i y_i$, et on désigne l'estimateur de la variance par $\hat{V}_s \left(\sum_{i \in S} w_i y_i \right)$.

Décrivons maintenant une procédure pour obtenir un ensemble de données entièrement imputées. D'abord, on utilise la méthode décrite à la section 3 pour obtenir les valeurs imputées $\{y_{1i}^{*(j)} : i \in B, j = 1, \ldots, m\}$ et une estimation $\hat{\theta}$ du paramètre de la distribution $f(y_2 \mid y_1, x; \theta)$. On obtient l'estimation $\hat{\theta}$ en résolvant

$$\sum_{i \in B} w_i \sum_{j=1}^{m} w_{ij}^* S_2 \left(\theta; x_i, y_{1i}^{*(j)}, y_{2i} \right) = 0, \tag{4.1}$$

où $S_2(\theta; x, y_1, y_2) = \partial \log f(y_2 \mid y_1, x; \theta) / \partial \theta$. Sachant $\hat{\theta}$, on génère les valeurs imputées $y_{2i}^{*(j)} \sim f \left(y_2 \mid y_{1i}, x_i; \hat{\theta} \right)$, pour $i \in A$ et $j = 1, \ldots, m$.

Si l'on suppose que le modèle est identifié, l'estimateur de paramètre $\hat{\theta}$ généré par la résolution de (4.1) est entièrement efficace au sens où la valeur imputée de y_{2i} pour l'échantillon A ne donne lieu à aucun gain d'efficacité. Pour le voir, notons que l'équation de score utilisant la valeur imputée de y_{2i} se calcule comme suit :

$$\sum_{i \in A} w_i m^{-1} \sum_{j=1}^{m} S_2 \left(\theta; x_i, y_{1i}, y_{2i}^{*(j)} \right) + \sum_{i \in B} w_i \sum_{j=1}^{m} w_{ij}^* S_2 \left(\theta; x_i, y_{1i}^{*(j)}, y_{2i} \right) = 0. \tag{4.2}$$

Comme $y_{2i}^{*(1)},\ldots,y_{2i}^{*(m)}$ sont générées à partir de $f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$,

$$p\lim_{m\to\infty} \sum_{i\in A} w_i m^{-1} \sum_{j=1}^{m} S_2\left(\theta; x_i, y_{1i}, y_{2i}^{*(j)}\right) = \sum_{i\in A} w_i E\left\{S_2\left(\theta; x_i, y_{1i}, Y_2\right) \mid y_{1i}, x_i; \hat{\theta}\right\}.$$

Ainsi, en vertu de la propriété de la fonction de score, le premier terme de (4.2) évalué à $\theta = \hat{\theta}$ est proche de zéro et la solution de l'équation (4.2) est essentiellement la même que celle de l'équation (4.1), c'est-à-dire qu'on ne gagne pas en efficience en utilisant la valeur imputée de y_{2i} pour calculer l'EMV pour θ dans $f(y_2 \mid y_1, x; \theta)$.

Toutefois, les valeurs imputées de y_{2i} peuvent améliorer l'efficience des inférences pour les paramètres de la distribution conjointe de (y_{1i}, y_{2i}). À titre d'exemple simple, prenons l'estimation de μ_2, la moyenne marginale de y_{2i}. En vertu d'un échantillonnage aléatoire simple, l'estimateur imputé de $\mu = E(Y_2)$ est

$$\hat{\mu}_{I,m} = \frac{1}{n}\left\{\sum_{i\in A}\left(m^{-1}\sum_{j=1}^{m} y_{2i}^{*(j)}\right) + \sum_{i\in B} y_{2i}\right\}, \tag{4.3}$$

où les valeurs de $y_{2i}^{*(1)},\ldots,y_{2i}^{*(m)}$ sont générées à partir de $f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$. Pour des valeurs de m, suffisamment grandes, on peut écrire

$$\begin{aligned}\hat{\mu}_{I,\infty} &= \frac{1}{n}\left\{\sum_{i\in A} \hat{y}_{2i} + \sum_{i\in B} y_{2i}\right\} \\ &= \frac{1}{n}\left\{\sum_{i\in A} E\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right) + \sum_{i\in B} y_{2i}\right\}.\end{aligned}$$

En vertu du scénario de l'exemple 2.1, on peut écrire $\hat{y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 y_{1i} + \hat{\beta}_2 x_{2i}$ où $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)$ satisfont

$$\sum_{i\in B}\left(y_{2i} - \hat{\beta}_0 - \hat{\beta}_{1i}\hat{y}_{1i} - \hat{\beta}_2 x_{2i}\right) = 0$$

et $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$ où $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ satisfont $\sum_{i\in A}(y_{1i} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i}) = 0$. Ainsi, en ignorant les termes d'ordre plus faible, on obtient

$$V\left(\hat{\mu}_{I,\infty}\right) = \frac{1}{n}V(y_2) + \left(\frac{1}{n_b} - \frac{1}{n}\right)V(y_2 - \hat{y}_2),$$

qui est inférieure à la variance de l'estimateur direct $\hat{\mu}_b = n_b^{-1}\sum_{i\in B} y_{2i}$.

5 Modèles d'erreur de mesure

Examinons maintenant l'application d'un appariement statistique au problème des modèles d'erreur de mesure. Supposons que l'on s'intéresse au paramètre θ de la distribution conditionnelle $f(y_2 \mid y_1; \theta)$. Dans l'échantillon initial, au lieu d'observer (y_{1i}, y_{2i}), on observe (x_i, y_{2i}), où x_i est une version contaminée

de y_{1i}. Comme il est possible que l'inférence pour θ fondée sur (x_i, y_{2i}) soit biaisée, d'autres renseignements sont nécessaires. L'une des façons courantes d'obtenir ces renseignements supplémentaires est de recueillir (x_i, y_{1i}) dans le cadre d'une étude de calage externe. Dans ce cas, on observe (x_i, y_{1i}) dans l'échantillon A et (x_i, y_{2i}) dans l'échantillon B, l'échantillon A étant l'échantillon de calage et l'échantillon B, l'échantillon principal. Guo et Little (2011) présentent une application d'un calage externe.

Le cadre de calage externe peut s'exprimer sous forme de problème d'appariement statistique. Le tableau 5.1 établit de façon explicite le lien entre l'appariement statistique et le calage externe. Une hypothèse de variable instrumentale permet l'inférence pour θ en fonction de données selon la structure présentée dans le tableau 1.1. Dans la notation du modèle d'erreur de mesure, l'hypothèse de variable instrumentale est

$$f(y_{2i} \mid y_{1i}, x_i) = f(y_{2i} \mid y_{1i}) \quad \text{et} \quad f(y_{1i} \mid x_i = a) \neq f(y_{1i} \mid x_i = b), \tag{5.1}$$

pour certains $a \neq b$. L'hypothèse de variable instrumentale peut être considérée raisonnable dans les applications relatives à l'erreur dans les covariables parce que le modèle d'intérêt en question est $f(y_{2i} \mid y_{1i})$, et x_i est une version contaminée de y_{1i} ne contenant aucun renseignement supplémentaire à propos de y_{2i} sachant y_{1i}.

Tableau 5.1
Structure de données pour le modèle d'erreur de mesure

	x_i	y_{1i}	y_{2i}
Enquête A (étude de calage)	o	o	
Enquête B (étude principale)	o		o

Dans le cas où $f(y_{2i} \mid y_{1i})$ et $f(y_{1i} \mid x_i)$ sont entièrement paramétriques, on peut utiliser l'imputation fractionnaire paramétrique pour exécuter l'algorithme EM. Cette méthode exige une évaluation de l'espérance conditionnelle de la fonction de score des données complètes sachant les valeurs observées. Pour évaluer l'espérance conditionnelle par imputation fractionnaire, on écrit d'abord la distribution conditionnelle de y_1 sachant (x, y_2) comme suit :

$$f(y_1 \mid x, y_2) \propto f(y_1 \mid x) f(y_2 \mid y_1). \tag{5.2}$$

Soit un estimateur $\hat{f}_a(y_{1i} \mid x_i)$ de $f(y_{1i} \mid x_i)$ provenant de l'échantillon de calage (échantillon A). La mise en œuvre de l'algorithme EM par imputation fractionnaire se déroule comme suit :

1. Pour chaque $i \in B$, générer $y_{1i}^{*(j)}$ à partir de $\hat{f}_a(y_1 \mid x_i)$, pour $j = 1, \ldots, m$.
2. Calculer les poids fractionnaires

$$w_{ij(t)}^{*} \propto f\left(y_{2i} \mid y_{1i}^{*(j)}; \hat{\theta}_t\right)$$

avec $\sum_{j=1}^{m} w_{ij(t)}^{*} = 1$.

3. Mettre à jour θ en résolvant

$$\sum_{i \in B} w_{ib} \sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; y_{1i}^{*(j)}, y_{2i}\right) = 0,$$

où $S(\theta; y_1, y_2) = \partial \log f(y_2 \mid y_1; \theta) / \partial \theta$.

4. Reprendre à l'étape 2 jusqu'à la convergence.

Cette méthode exige que l'on génère des données à partir de $f(y_1 \mid x)$. Dans le cas de certains modèles non linéaires ou de modèles assortis de variances non constantes, la simulation à partir de la distribution conditionnelle de y_1 sachant x peut exiger le recours à des méthodes Monte Carlo comme l'acceptation-rejet ou l'algorithme de Metropolis-Hastings. La simulation présentée à la section 6.2 est un bon exemple d'une simulation dans laquelle la distribution conditionnelle de $y_1 \mid x$ n'a pas d'expression de forme explicite. Dans ce cas, on peut envisager une autre solution plus simple à calculer. Pour décrire cette solution, posons $h(y_1 \mid x)$ comme distribution conditionnelle « de travail », par exemple la distribution normale, à partir de laquelle les échantillons peuvent être facilement générés. On présume que les estimations $\hat{f}_a(y_1 \mid x)$ et $\hat{h}_a(y_1 \mid x)$ de $f(y_1 \mid x)$ et $h(y_1 \mid x)$, respectivement, peuvent être obtenues à partir de l'échantillon A. La mise en œuvre de l'algorithme EM par imputation fractionnaire s'effectue ensuite comme suit :

1. Pour chaque $i \in B$, générer $x_i^{*(j)}$ à partir de $\hat{h}_a(y_1 \mid x_i)$, pour $j = 1, \ldots, m$.
2. Calculer les poids fractionnaires

$$w_{ij(t)}^{*} \propto f\left(y_{2i} \mid y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a\left(y_{1i}^{*(j)} \mid x_i\right) \Big/ \hat{h}_a\left(y_{1i}^{*(j)} \mid x_i\right) \tag{5.3}$$

avec $\sum_{j=1}^{m} w_{ij(t)}^{*} = 1$.

3. Mettre à jour θ en résolvant

$$\sum_{i \in B} w_{ib} \sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; y_{1i}^{*(j)}, y_{2i}\right) = 0.$$

4. Reprendre à l'étape 2 jusqu'à la convergence.

L'estimation de la variance est une application directe de la méthode de linéarisation présentée à la section 3. La méthode d'imputation fractionnaire hot deck décrite à la section 3 assortie des poids fractionnaires définis en (3.3) s'applique aussi directement au contexte de l'erreur de mesure.

6 Études par simulation

Pour mettre à l'essai notre théorie, nous présentons deux études par simulation limitées. La première porte sur la combinaison de deux enquêtes indépendantes avec observation partielle afin d'effectuer une analyse conjointe. La deuxième porte sur la définition de modèles d'erreur de mesure avec calage externe.

6.1 Première simulation

Pour comparer les méthodes proposées avec les méthodes actuelles, nous avons généré 5 000 échantillons Monte Carlo comprenant (x_i, y_{1i}, y_{2i}) et de taille $n = 400$, où

$$\begin{pmatrix} y_{1i} \\ x_i \end{pmatrix} \sim N\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0{,}7 \\ 0{,}7 & 1 \end{bmatrix} \right),$$

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + e_i, \tag{6.1}$$

$e_i \sim N(0, \sigma^2)$, et $\beta = (\beta_0, \beta_1, \sigma^2)' = (1,1,1)'$. Soulignons que dans ce scénario, on trouve $f(y_2 \mid x, y_1) = f(y_2 \mid y_1)$; la variable x joue donc le rôle de variable instrumentale pour y_1.

Au lieu d'observer (x_i, y_{1i}, y_{2i}) conjointement, on présume que seuls (y_1, x) sont observés dans l'échantillon A, et que seuls (y_2, x) sont observés dans l'échantillon B; l'échantillon A est obtenu par sélection des $n_a = 400$ premiers éléments de l'échantillon initial, et l'échantillon B, par sélection des $n_b = 400$ éléments qui restent. On veut estimer quatre paramètres : trois paramètres de régression $\beta_0, \beta_1, \sigma^2$ et $\pi = P(y_1 < 2, y_2 < 3)$, la proportion de $y_1 < 2$ et de $y_2 < 3$. Quatre méthodes sont envisagées pour estimer ces paramètres :

1. Estimation de l'échantillon complet (EEC) : Utiliser toutes les observations de (y_{1i}, y_{2i}) de l'échantillon B.
2. Imputation par régression stochastique (IRS) : Utiliser la régression de y_1 sur x dans l'échantillon A pour obtenir $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\sigma}_1^2)$, où le modèle de régression est $y_1 = \alpha_0 + \alpha_1 x + e_1$ avec $e_1 \sim (0, \sigma_1^2)$. Pour chaque $i \in B$, $m = 10$ valeurs imputées sont générées par $y_{1i}^{*(j)} = \hat{\alpha}_0 + \hat{\alpha}_1 x_i + e_i^{*(j)}$ où $e_i^{*(j)} \sim N(0, \hat{\sigma}_1^2)$.
3. Imputation fractionnaire paramétrique (IFP) avec $m = 10$ selon l'hypothèse de variable instrumentale.
4. Imputation fractionnaire hot deck (IFHD) avec $m = 10$ selon l'hypothèse de variable instrumentale.

Le tableau 6.1 présente les moyennes et les variances Monte Carlo des estimateurs ponctuels des quatre paramètres d'intérêt. La méthode IRS comporte des biais importants pour tous les paramètres étudiés, parce qu'elle repose sur l'hypothèse d'indépendance conditionnelle. Les méthodes IFP et IFHD fournissent des estimateurs presque sans biais pour tous les paramètres. Les estimateurs obtenus à l'aide de la méthode IFHD sont légèrement plus efficients que ceux obtenus par la méthode IFP, parce que la démarche en deux étapes de la méthode IFHD utilise l'ensemble complet des répondants à la première étape. La variance asymptotique théorique de $\hat{\beta}_1$ calculée à partir de la méthode IFP est

$$V(\hat{\beta}_1) \doteq \frac{1}{(0{,}7)^2}\frac{1}{400} 2\left(1 - \frac{0{,}7^2}{2}\right) + \frac{1}{(0{,}7)^2}\frac{1}{400}(1 - 0{,}7^2) \doteq 0{,}0103$$

ce qui correspond au résultat de simulation présenté dans le tableau 6.1. En plus de l'estimation ponctuelle, on calcule aussi les estimateurs de variance pour les méthodes IFP et IFHD. Les estimateurs de variance

montrent de faibles biais relatifs (moins de 5 % en valeur absolue) pour tous les paramètres. Les résultats de l'estimation de la variance ne sont pas présentés ici par souci de concision.

Tableau 6.1
Moyennes et variances Monte Carlo des estimateurs ponctuels pour la première simulation. (EEC : estimation de l'échantillon complet; IRS : imputation par régression stochastique; IFP : imputation fractionnaire paramétrique; IFHD : imputation fractionnaire hot deck)

Paramètre	Méthode	Moyenne	Variance
β_0	EEC	1,00	0,0123
	IRS	1,90	0,0869
	IFP	1,00	0,0472
	IFHD	1,00	0,0465
β_1	EEC	1,00	0,00249
	IRS	0,54	0,01648
	IFP	1,00	0,01031
	IFHD	1,00	0,01026
σ^2	EEC	1,00	0,00482
	IRS	1,73	0,01657
	IFP	0,99	0,02411
	IFHD	0,99	0,02270
π	EEC	0,374	0,00058
	IRS	0,305	0,00255
	IFP	0,375	0,00059
	IFHD	0,375	0,00057

La méthode proposée repose sur l'hypothèse de variable instrumentale. Pour déterminer la sensibilité de la méthode d'imputation fractionnaire proposée aux violations de l'hypothèse de variable instrumentale, nous avons réalisé une étude par simulation supplémentaire. Au lieu de générer y_{2i} à partir de (6.1), on utilise

$$y_{2i} = 0{,}5 + y_{1i} + \rho(x_i - 3) + e_i, \tag{6.2}$$

où $e_i \sim N(0,1)$ et ρ peuvent prendre des valeurs non nulles. Nous avons utilisé trois valeurs de ρ, $\rho \in \{0; 0{,}1; 0{,}2\}$, pour l'analyse de sensibilité et nous avons employé la même procédure d'IFP et d'IFHD fondée sur l'hypothèse que x est une variable instrumentale pour y_1. Cette hypothèse est satisfaite pour $\rho = 0$, mais elle est légèrement enfreinte pour $\rho = 0{,}1$ ou $\rho = 0{,}2$. À l'aide des données obtenues par imputation fractionnaire dans l'échantillon B, nous avons estimé trois paramètres, $\theta_1 = E(Y_1)$, θ_2 la pente de la régression simple de y_2 sur y_1, et $\theta_3 = P(y_1 < 2, y_2 < 3)$, la proportion de $y_1 < 2$ et $y_2 < 3$. Le tableau 6.2 présente les moyennes et les variances Monte Carlo des estimateurs ponctuels pour les trois paramètres en vertu des trois différents modèles. Dans le tableau 6.2, on constate que les valeurs absolues de l'écart entre l'estimateur obtenu par imputation fractionnaire et l'estimateur obtenu à partir de l'échantillon complet augmente avec la valeur de ρ, ce qui est normal puisque l'hypothèse de variable instrumentale est plus gravement enfreinte pour les valeurs élevées de ρ, mais les écarts sont relativement faibles dans tous les cas. Plus particulièrement, l'estimateur de θ_1 n'est pas touché par la dérogation à

l'hypothèse de variable instrumentale, parce que l'estimateur par imputation obtenu en vertu du modèle d'imputation inexact fournit quand même un estimateur non biaisé pour la moyenne de population, à condition que le modèle d'imputation par régression contienne un terme d'ordonnée à l'origine (Kim et Rao 2012). Ainsi, cette analyse de sensibilité limitée indique que la méthode proposée semble produire des estimations comparables lorsque l'hypothèse de variable instrumentale est légèrement enfreinte.

Tableau 6.2
Moyennes et variances Monte Carlo des deux estimateurs ponctuels pour l'analyse de sensibilité de la première simulation (EEC : estimation de l'échantillon complet; IFP : imputation fractionnaire paramétrique; IFHD : imputation fractionnaire hot deck)

Modèle	Paramètre	Méthode	Moyenne	Variance
$\rho = 0$	θ_1	EEC	2,00	0,00235
		IFP	2,00	0,00352
		IFHD	2,00	0,00249
	θ_2	EEC	1,00	0,00249
		IFP	1,00	0,01031
		IFHD	1,00	0,01026
	θ_3	EEC	0,43	0,00061
		IFP	0,43	0,00059
		IFHD	0,43	0,00057
$\rho = 0,1$	θ_1	EEC	2,00	0,00235
		IFP	2,00	0,00353
		IFHD	2,00	0,00250
	θ_2	EEC	1,07	0,00248
		IFP	1,14	0,01091
		IFHD	1,14	0,01081
	θ_3	EEC	0,44	0,00061
		IFP	0,45	0,00062
		IFHD	0,45	0,00059
$\rho = 0,2$	θ_1	EEC	2,00	0,00235
		IFP	2,00	0,00353
		IFHD	2,00	0,00250
	θ_2	EEC	1,14	0,00250
		IFP	1,28	0,01115
		IFHD	1,28	0,01102
	θ_3	EEC	0,44	0,00061
		IFP	0,46	0,00066
		IFHD	0,46	0,00062

6.2 Deuxième simulation

Dans la deuxième étude par simulation, on examine une variable de réponse binaire y_{2i}, où

$$y_{2i} \sim \text{Bernoulli}(p_i), \tag{6.3}$$

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 y_{1i},$$

et $y_{1i} \sim N\left(\mu_1, \sigma_1^2\right)$. Dans l'échantillon principal, désigné par la lettre B, au lieu d'observer (y_{1i}, y_{2i}), on observe (x_i, y_{2i}), où

$$x_i = \beta_0 + \beta_1 y_{1i} + u_i, \tag{6.4}$$

et $u_i \sim N\left(0, \sigma^2 \left| y_{1i} \right|^{2\alpha}\right)$. On observe (x_i, y_{1i}), $i = 1, \ldots, n_A$ dans un échantillon de calage, désigné par la lettre A. Aux fins de la simulation, $n_A = n_B = 800$, $\gamma_0 = 1$, $\gamma_1 = 1$, $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2 = 0{,}25$, $\alpha = 0{,}4$, $\mu_1 = 0$, et $\sigma_1^2 = 1$. Le but principal est d'estimer γ_1 et de mettre à l'essai l'hypothèse nulle que $\gamma_1 = 1$. La taille de l'échantillon Monte Carlo (MC) est 1 000.

On compare les estimateurs IFP de γ_1 aux trois autres estimateurs. Comme la distribution conditionnelle de y_{1i} sachant x_i n'est pas standard, on utilise les poids de (5.3) pour réaliser l'IFP, où la distribution proposée $\hat{h}_a(y_{1i} \mid x_i)$ est une estimation de la distribution marginale de y_{1i} fondée sur les données de l'échantillon A. On considère les quatre estimateurs suivants :

1. *IFP :* Pour l'IFP, la distribution proposée pour générer $y_{1i}^{*(j)}$ est une distribution normale de moyenne $\hat{\mu}_1$ et de variance $\hat{\sigma}_1^2$, où $\hat{\mu}_1$ et $\hat{\sigma}_1^2$ sont les estimations du maximum de vraisemblance de μ_1 et σ_1^2, respectivement, en fonction de l'échantillon A. Les poids fractionnaires définis en (5.3) se présentent sous la forme

$$w_{ij}^* \propto \hat{p}_{ij}^{\,y_{2i}} \left(1 - \hat{p}_{ij}\right)^{1-y_{2i}} \hat{f}_a\left(y_{1i}^{*(j)} \mid x_i\right), \tag{6.5}$$

   où $\hat{p}_{ij} = \left\{1 + \exp\left(-\hat{\gamma}_0 - \hat{\gamma}_1 y_{1i}^{*(j)}\right)\right\}^{-1}$ et $\hat{f}_a(y_{1i} \mid x_i)$ est l'estimation de $f(y_{1i} \mid x_i)$ fondée sur l'estimation du maximum de vraisemblance à partir des données de l'échantillon A. La taille de la classe d'imputation est $m = 800$.

2. Estimateur *naïf :* Un estimateur *naïf* est l'estimateur de la pente dans la régression logistique de y_{2i} sur x_i pour $i \in B$.

3. Estimateur *bayésien :* On utilise l'approche de Guo et Little (2011) pour définir un estimateur bayésien. Le modèle de notre simulation diffère de celui de Guo et Little (2011) par le fait que la réponse d'intérêt est binaire. On établit un échantillonnage de Gibbs à l'aide du programme JAGS (Plummer 2003), en précisant les distributions a priori diffuses appropriées pour les paramètres du modèle. Soit

$$\theta_1 = \left(\log\left(\sigma^2\right), \log\left(\sigma_1^2\right), \mu_1, \beta_0, \beta_1, \gamma_0, \gamma_1\right);$$

   on suppose a priori que $\theta_1 \sim N\left(0, 10^6 I_7\right)$, où I_7 est une matrice identité de dimensions 7×7 et où la notation $N(0, V)$ désigne une distribution normale de moyenne 0 et de matrice de covariances V. La distribution a priori pour la puissance α est uniforme sur l'intervalle $[-5, 5]$.

   Pour évaluer la convergence, on examine les tracés des courbes et les facteurs de réduction d'échelle possibles définis par Gelman, Carlin, Stern et Rubin (2003) pour 10 ensembles de données simulées préliminaires. On produit trois chaînes MCMC, chacune d'une longueur de 1 500, à partir de valeurs initiales aléatoires, et on rejette les 500 premières itérations, considérées comme faisant partie du rodage. Les facteurs de réduction d'échelle possibles dans les

10 ensembles de données simulées vont de 1,0 à 1,1, et les tracés des courbes indiquent que les chaînes se combinent bien. Pour réduire le temps de calcul, on utilise 1 000 itérations d'une seule chaîne pour la simulation principale, après rejet des 500 premières itérations de rodage.

4. Un estimateur par *calage par régression pondérée (CRP)*. Cet estimateur par CRP est une modification de l'estimateur par calage par régression pondérée défini par Guo et Little (2011) pour une variable de réponse binaire. On calcule l'estimateur par calage par régression pondérée comme suit :

   (i) À l'aide des moindres carrés ordinaires (MCO), effectuer une régression de y_{1i} sur x_i pour l'échantillon de calage.

   (ii) Effectuer une régression du logarithme des résidus quadratiques obtenus à l'étape (i) sur le logarithme de x_i^2 pour l'échantillon de calage. Soit $\hat{\lambda}$ la pente estimée de la régression.

   (iii) À l'aide des moindres carrés pondérés (MCP) avec le poids $|x_i|^{2\hat{\lambda}}$, effectuer la régression de y_{1i} sur x_i pour l'échantillon de calage. Soient $\hat{\eta}_0$ et $\hat{\eta}_1$ l'ordonnée à l'origine et la pente estimées, respectivement, de la régression des MCP.

   (iv) Pour chaque unité i de l'échantillon principal, posons $\hat{y}_{1i} = \hat{\eta}_0 + \hat{\eta}_1 x_i$.

   (v) L'estimation de (γ_0, γ_1) est obtenue à partir de la régression logistique de y_{2i} sur $\hat{y}_{1i}$ dans l'échantillon principal.

Le tableau 6.3 indique le biais, la variance et l'EQM Monte Carlo des quatre estimateurs de γ_1. L'estimateur naïf a un biais négatif parce que x_i est une version contaminée de y_{1i}. L'estimateur par IFP est supérieur à l'estimateur bayésien et à l'estimateur par CRP.

On calcule une estimation de la variance des estimateurs par IFP de γ_1 à l'aide de l'expression de la variance fondée sur l'approximation linéaire. On définit le biais relatif MC comme étant le ratio de la différence entre la moyenne MC de l'estimateur de variance et la variance MC de l'estimateur à la variance MC de l'estimateur. Le biais relatif MC des estimateurs de variance pour l'IFP est négligeable (moins de 2 % en valeur absolue).

Tableau 6.3
Moyennes, variances et erreurs quadratiques moyennes Monte Carlo des estimateurs ponctuels de γ_1 pour la deuxième simulation. (IFP : imputation fractionnaire paramétrique; CRP : calage par régression pondérée; MC : Monte Carlo; EQM : erreur quadratique moyenne)

Méthode	Biais MC	Variance MC	EQM MC
IFP	0,0239	0,0386	0,0392
Estimateur naïf	-0,2241	0,0239	0,0742
Estimateur bayésien	0,0406	0,0415	0,0432
Estimateur par CRP	0,112	0,0499	0,0625

7 Conclusion

Nous considérons l'appariement statistique comme un problème de données manquantes et proposons la méthode d'IFP pour obtenir des estimateurs convergents et des estimateurs de variance correspondants. En vertu de l'hypothèse que le modèle spécifié est entièrement identifié, la méthode proposée permet d'obtenir les estimateurs du pseudo maximum de vraisemblance des paramètres du modèle.

Pour qu'un modèle puisse être considéré comme identifiable, il suffit qu'il comporte une variable instrumentale. Le cadre d'erreur de mesure énoncé aux sections 5 et 6.2, en vertu duquel un calage externe fournit une mesure indépendante de la vraie covariable d'intérêt, représente une situation dans laquelle le plan de sondage peut être considéré comme soutenant l'hypothèse de variable instrumentale. La méthodologie proposée peut être appliquée sans hypothèse de variable instrumentale, à condition que le modèle soit identifié. Si le modèle n'est pas identifiable, l'algorithme EM pour la méthode d'IFP proposée ne converge pas nécessairement. Dans la pratique, on peut considérer le modèle spécifié comme étant identifié si la séquence EM converge. Autrement dit, tant que la séquence EM converge, l'analyse connexe est convergente sous le modèle spécifié. C'est là l'un des nombreux avantages du recours à la méthode axée sur les fréquences par rapport à la méthode bayésienne. Avec la méthode bayésienne, il est possible d'obtenir les valeurs a posteriori même avec des modèles non identifiés et dans ce cas, l'analyse qui en découle peut être trompeuse.

En vertu de la structure de données présentée dans le tableau 1.1, il est beaucoup plus difficile, voire impossible, de déterminer si l'hypothèse de VI se vérifie dans les données disponibles. Compte tenu du modèle spécifié, on ne peut que vérifier si les paramètres du modèle peuvent être entièrement estimés. En revanche, il en va autrement de la détermination du caractère approprié du modèle spécifié par rapport aux données disponibles. Le diagnostic de modèles et la sélection d'un modèle parmi les différents modèles identifiables constituent d'importants sujets qu'il conviendrait d'approfondir dans le cadre de recherches futures.

L'appariement statistique peut aussi servir à évaluer les effets de traitements multiples dans les études d'observation. En utilisant adéquatement les techniques d'appariement statistique, on peut créer un fichier de données augmentées sur les résultats possibles afin d'étudier l'inférence causale (Morgan et Winship 2007). De telles utilisations seront présentées ailleurs.

Remerciements

Nous remercions le professeur Yanyuan Ma, un examinateur anonyme et le rédacteur adjoint pour leurs commentaires très constructifs. Les travaux de recherche de Jae Kwang Kim ont été en partie financés par le programme *Brain Pool* (131S-1-3-0476) de la *Korean Federation of Science and Technology Society* et par une subvention de la NSF (MMS-121339). Les travaux de recherche d'Emily Berg ont été financés en vertu d'une entente de coopération entre le *US Department of Agriculture Natural Resources Conservation Service* et la *Iowa State University*. Les travaux de recherche de Taesung Park ont été financés dans le cadre du *Bio-Synergy Research Project* (2013M3A9C4078158) du *Ministry of Science, ICT and Future Planning*, par l'entremise de la *National Research Foundation* de Corée.

Annexe

A. Absence asymptotique de biais de l'estimateur par les MC2E

Supposons que l'on observe (y_1, x) dans l'échantillon A et (y_2, x) dans l'échantillon B. Par souci de rigueur, on peut écrire (y_{1a}, x_a) pour désigner l'observation de (y_1, x) dans l'échantillon A, et (y_{2b}, x_b) pour désigner les observations dans l'échantillon B. Dans ce cas, le modèle peut s'écrire

$$y_{1a} = \phi_0 1_a + \phi_1 x_{1a} + \phi_2 x_{2a} + e_{1a}$$
$$y_{2b} = \beta_0 1_b + \beta_1 y_{1b} + \beta_2 x_{2b} + e_{2b}$$

avec $E(e_{1a} \mid x_a) = 0$ et $E(e_{2b} \mid x_b, y_{1b}) = 0$. Soulignons que y_{1b} n'est pas observée dans l'échantillon. On utilise plutôt $\hat{y}_{1b}$ produite grâce à l'estimation par les MCO obtenue à partir de l'échantillon A.

En écrivant $X_a = [1_a, x_a]$ et $X_b = [1_b, x_b]$, on obtient $\hat{y}_{1b} = X_b (X_a' X_a)^{-1} X_a' y_{1a} = X_b \hat{\phi}_a$. L'estimateur par les MC2E de $\beta = (\beta_0, \beta_1, \beta_2)'$ est donc

$$\hat{\beta}_{\text{MC2E}} = (Z_b' Z_b)^{-1} Z_b' y_{2b}$$

où $Z_b = [1_b, \hat{y}_{1b}, x_{2b}]$. Ainsi, on obtient

$$\begin{aligned} \hat{\beta}_{\text{MC2E}} - \beta &= (Z_b' Z_b)^{-1} Z_b' (y_{2b} - Z_b \beta) \\ &= (Z_b' Z_b)^{-1} Z_b' \{\beta_1 (y_{1b} - \hat{y}_{1b}) + e_{2b}\}. \end{aligned} \tag{A.1}$$

On peut écrire

$$y_{1b} = \phi_0 1_b + \phi_1 x_b + e_{1b} = X_b \phi + e_{1b}$$

où $E(e_{1b} \mid x_b) = 0$. Comme

$$\begin{aligned} \hat{y}_{1b} &= X_b (X_a' X_a)^{-1} X_a' y_{1a} \\ &= X_b (X_a' X_a)^{-1} X_a' (X_a \phi + e_{1a}) \\ &= X_b \phi + X_b (X_a' X_a)^{-1} X_a' e_{1a}, \end{aligned}$$

on obtient

$$y_{1b} - \hat{y}_{1b} = e_{1b} - X_b (X_a' X_a)^{-1} X_a' e_{1a}$$

et (A.1) devient

$$\hat{\beta}_{\text{MC2E}} - \beta = (Z_b' Z_b)^{-1} Z_b' \left\{\beta_1 e_{1b} - \beta_1 X_b (X_a' X_a)^{-1} X_a' e_{1a} + e_{2b}\right\}. \tag{A.2}$$

Supposons que les deux échantillons sont indépendants. Ainsi, $E(e_{1b} \mid x_a, x_b, y_{1a}) = 0$. De plus, $E\left\{(Z_b' Z_b)^{-1} Z_b' e_{2b} \,\middle|\, x_a, x_b, y_{1a}, y_{1b}\right\} = 0$. Ainsi,

$$E\left\{\hat{\beta}_{\text{MC2E}} - \beta \,\middle|\, x_a, x_b, y_{1a}\right\} = E\left\{-\beta_1 (Z_b' Z_b)^{-1} Z_b' X_b (X_a' X_a)^{-1} X_a' e_{1a} \,\middle|\, x_a, x_b, y_{1a}\right\}$$

et

$$\begin{aligned} (Z_b' Z_b)^{-1} Z_b' X_b (X_a' X_a)^{-1} X_a' e_{1a} &= (Z_b' Z_b)^{-1} Z_b' \left\{X_b (X_a' X_a)^{-1} X_a' (y_{1a} - X_a \phi)\right\} \\ &= (Z_b' Z_b)^{-1} Z_b' X_b \left(\hat{\phi}_a - \phi\right). \end{aligned}$$

Ce terme a une espérance nulle asymptotiquement parce que $n_b^{-1} Z_b' Z_b$ et $n_b^{-1} Z_b' X_b$ sont bornés en probabilité et que $\left(\hat{\phi}_a - \phi\right)$ converge vers zéro.

B. Estimation de la variance

Soit le paramètre d'intérêt défini par la solution de $U_N(\eta) = \sum_{i=1}^{N} U(\eta; y_{1i}, y_{2i}) = 0$. On suppose que $\partial U_N(\eta)/\partial\theta = 0$. Ainsi, le paramètre η est indépendant a priori de θ, qui est le paramètre de la distribution de production de données de (x, y_1, y_2).

En vertu du scénario de la section 3, on pose $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ comme l'EMV de $\theta = (\theta_1, \theta_2)$ obtenu par la résolution de (3.4). De plus, posons $\hat{\eta}$ comme la solution de $\bar{U}(\eta|\hat{\theta}) = 0$ où

$$\bar{U}(\eta|\theta) = \sum_{i\in B}\sum_{j=1}^{m} w_{ib} w_{ij}^{*} U\left(\eta; y_{1i}^{*(j)}, y_{2i}\right),$$

et

$$w_{ij}^{*} \propto f\left(y_{1i}^{*(j)} \middle| x_i; \hat{\theta}_1\right) f\left(y_{2i} \middle| y_{1i}^{*(j)}; \hat{\theta}_2\right) \Big/ h\left(y_{1i}^{*(j)} \middle| x_i\right)$$

avec $\sum_{j=1}^{m} w_{ij}^{*} = 1$. Ici, $h(y_1|x)$ est la distribution proposée pour la génération des valeurs imputées de y_1 dans l'imputation fractionnaire paramétrique. En introduisant la distribution proposée h, on peut sans risque ignorer la dépendance des valeurs imputées $y_{1i}^{*(j)}$ à la valeur de paramètre estimée $\hat{\theta}_1$.

En effectuant une linéarisation par série de Taylor,

$$\bar{U}(\eta|\hat{\theta}) \cong \bar{U}(\eta|\theta) + (\partial\bar{U}/\partial\theta_1')(\hat{\theta}_1 - \theta_1) + (\partial\bar{U}/\partial\theta_2')(\hat{\theta}_2 - \theta_2)$$

on constate que

$$\hat{\theta}_1 - \theta_1 \cong \{I_1(\theta_1)\}^{-1} S_1(\theta_1)$$

où $I_1(\theta_1) = -\partial S_1(\theta_1)/\partial\theta_1'$. De plus,

$$\hat{\theta}_2 - \theta_2 \cong \left\{-\frac{\partial}{\partial\theta_2'}\bar{S}_2(\theta)\right\}^{-1} \bar{S}_2(\theta)$$

où

$$\bar{S}_2(\theta) = \sum_{i\in B}\sum_{j=1}^{m} w_i w_{ij}^{*}(\theta) S_2\left(\theta_2; y_{1i}^{*(j)}, y_{2i}\right).$$

Ainsi, on peut établir

$$\bar{U}(\eta|\hat{\theta}) \cong \bar{U}(\eta|\theta) + K_1 S_1(\theta_1) + K_2 \bar{S}_2(\theta),$$

où $K_1 = D_{21} I_{11}^{-1}$ et $K_2 = D_{22} I_{22}^{-1}$ avec $I_{11} = -E(\partial S_1/\partial\theta_1')$, $I_{22} = -E(\partial\bar{S}_2/\partial\theta_2')$, $D_{21} = E\{U(\eta) S_1(\theta_1)'\}$ et $D_{22} = E\{U(\eta) S_2(\theta_2)'\}$, on obtient

$$V\left\{\bar{U}(\eta|\hat{\theta})\right\} = \tau^{-1}\{V_1 + V_2\}\tau^{-1'}$$

où $\tau = -E\{\partial \bar{U}(\eta|\theta)/\partial \eta'\}$,

$$V_1 = V\left\{\sum_{i\in B} w_i\left(\bar{u}_i^* + K_2 S_{2i}^*\right)\right\},$$

$\bar{u}_i^* = E\left[U(\hat{\eta}; y_{1i}, y_{2i})\,|\,y_{2i}; \hat{\theta}\right]$, et $V_2 = V\left\{K_1 \sum_{i\in A} w_i S_{1i}\right\}$. Un estimateur convergent de chaque composante peut être élaboré selon la technique décrite à la section 3.

Bibliographie

Baker, K.H., Harris, P. et O'Brien, J. (1989). Data fusion: An appraisal and experimental evaluation. *Journal of the Market Research Society*, 31, 152-212.

Beaumont, J.-F., et Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 37, 3, 400-416.

Chen, J., et Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American Statistical Association*, 96, 453, 260-269.

Chib, S., et Greenberg, E. (1995). Jackknife variance estimation for nearest neighbor imputation. *The American Statistician*, 46, 327-333.

Chipperfield, J.O., et Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25, 2, 227-244.

D'Orazio, M., Zio, M.D. et Scanu, M. (2006). *Statistical Matching: Theory and Practice.* Chichester, R.U.: Wiley.

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

Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (2003). *Bayesian Data Analysis*, Chapman and Hall Texts in Statistical Science. Chapman and Hall/CRC, deuxième édition.

Gonzalez, J., et Eltinge, J. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2081-2088.

Guo, Y., et Little, R.J. (2011). Regression analysis with covariates that have heteroskedastic measurement error. *Statistics Medicine*, 30, 18, 2278-2294.

Haziza, D. (2009). Imputation and inference in the presence of missing data. Dans *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference*, (Éds., C.R. Rao et D. Pfeffermann), 215-246.

Herzog, T.N., Scheuren, F.J. et Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques.* New York: Springer.

Ibrahim, J.G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85, 765-769.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119-132.

Kim, J.K., et Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99, 85-100.

Kim, J.K., et Shao, J. (2013). *Statistical Methods in Handling Incomplete Data*, Chapman and Hall/CRC.

Kim, J.K., et Yang, S. (2014). Imputation fractionnaire hot deck pour une inférence robuste sous un modèle de non-réponse partielle en échantillonnage. *Techniques d'enquête*, 40, 2, 235-256.

Lahiri, P., et Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 1265-1275.

Leulescu, A., et Agafitei, M. (2013). Statistical matching: A model based approach for data integration. *Eurostat Methodologies and Working Papers.*

Morgan, S.L., et Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York, USA: Cambridge University Press.

Moriarity, C., et Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17, 407-422.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Dans *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Raghunathan, T.E., et Grizzle, J.E. (1995). A split questionnaire design. *Journal of the American Statistical Association*, 90, 54-63.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer-Verlag.

Ridder, S., et Moffit, R. (2007). The econometrics of data combination. *Handbook of Econometrics*, 5470-5544.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 45-67
Statistique Canada, nº 12-001-X au catalogue

Comparaison d'estimateurs sur petits domaines au niveau de l'unité et au niveau du domaine

Michael A. Hidiroglou et Yong You[1]

Résumé

Les auteurs comparent les estimateurs EBLUP et pseudo-EBLUP pour l'estimation sur petits domaines en vertu d'un modèle de régression à erreur emboîtée, ainsi que trois autres estimateurs fondés sur un modèle au niveau du domaine à l'aide du modèle de Fay-Herriot. Ils réalisent une étude par simulations fondée sur un plan de sondage pour comparer les estimateurs fondés sur un modèle pour des modèles au niveau de l'unité et au niveau du domaine sous un échantillonnage informatif et non informatif. Ils s'intéressent particulièrement aux taux de couverture des intervalles de confiance des estimateurs au niveau de l'unité et au niveau du domaine. Les auteurs comparent aussi les estimateurs sous un modèle dont la spécification est inexacte. Les résultats de la simulation montrent que les estimateurs au niveau de l'unité sont plus efficaces que les estimateurs au niveau du domaine. L'estimateur pseudo-EBLUP donne les meilleurs résultats à la fois au niveau de l'unité et au niveau du domaine.

Mots-clés : Intervalle de confiance; convergence sous le plan de sondage; modèle de Fay-Herriot; échantillonnage informatif; spécification inexacte du modèle; modèle de régression à erreur emboîtée; racine de l'erreur quadratique moyenne relative (REQMR); poids d'enquête.

1 Introduction

Ces dernières années, les chercheurs ont eu largement recours à des estimateurs sur petits domaines fondés sur des modèles pour obtenir des estimations indirectes fiables sur de petits domaines. Les estimateurs fondés sur des modèles reposent sur des modèles explicites qui établissent des liens avec de petits domaines connexes grâce à des données supplémentaires, comme des données de recensement et des données administratives. On peut classer les modèles d'estimation sur petits domaines en deux grandes catégories : (i) les modèles au niveau de l'unité, qui établissent des liens entre les valeurs unitaires de la variable étudiée et des variables auxiliaires propres à l'unité et (ii) les modèles au niveau du domaine, qui établissent des liens entre les estimateurs directs de la variable étudiée du petit domaine et les variables auxiliaires propres au domaine correspondantes. En général, les modèles au niveau du domaine servent à améliorer les estimateurs directs lorsqu'il n'y a pas de données disponibles au niveau de l'unité. L'échantillonnage est établi selon la méthode de Rao (2003), c'est-à-dire qu'un univers U de taille N est divisé en m petits domaines non chevauchants U_i de taille N_i, où $i = 1,\ldots,m$. L'échantillonnage est réalisé dans chaque petit domaine selon un mécanisme probabiliste pour produire des échantillons s_i de taille n_i. La probabilité de sélection associée à chaque élément $j = 1,\ldots,n_i$ sélectionné dans l'échantillon s_i est désignée par p_{ij}. Les poids de sondage qui en découlent sont donnés par $w_{ij} = n_i^{-1} p_{ij}^{-1}$. En pratique, ces poids peuvent être ajustés pour tenir compte de la non-réponse et de données auxiliaires. Les poids obtenus correspondent aux poids de l'enquête. Dans le présent article, on présume une réponse totale à l'enquête et aucun ajustement pour tenir compte de données auxiliaires. Les estimations directes au niveau du domaine pour chaque domaine sont obtenues à partir des poids de l'enquête et des unités observées dans

1. Michael A. Hidiroglou, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario), K1A 0T6, Canada. Courriel : hidirog@yahoo.ca; Yong You, Division de la coopération internationale et des méthodes statistiques institutionnelles, Statistique Canada, Ottawa (Ontario), K1A 0T6, Canada. Courriel : yong.you@canada.ca.

le domaine. Le plan d'enquête peut être intégré de différentes manières aux modèles d'estimation sur petits domaines. Au niveau du domaine, les estimateurs directs fondés sur le plan de sondage sont modélisés directement et la variance de sondage de l'estimateur direct connexe est intégrée au modèle au moyen des erreurs fondées sur le plan de sondage. Au niveau de l'unité, les observations peuvent être pondérées à l'aide du poids de l'enquête. Un certain nombre de facteurs influent sur l'efficacité des estimateurs. Deux facteurs importants sont l'exactitude lors de la spécification du modèle et la corrélation entre la variable d'intérêt et les probabilités de sélection associées au processus d'échantillonnage, c'est-à-dire le caractère informatif du processus d'échantillonnage. Dans le présent article, les auteurs comparent, au moyen d'une étude par simulations, l'incidence de la spécification inexacte du modèle et du caractère informatif du plan d'échantillonnage pour deux méthodes de base utilisées pour l'estimation sur petits domaines au niveau de l'unité et au niveau du domaine en termes de biais, d'erreur quadratique moyenne estimée et de taux de couverture des intervalles de confiance. Un plan d'échantillonnage est informatif si les probabilités de sélection p_{ij} demeurent reliées à la variable d'intérêt y_{ij} même après conditionnement sur les covariables $\mathbf{x}_{ij}$. Dans un tel cas, on dit que l'échantillonnage est informatif parce que le modèle de population n'est plus vérifié pour l'échantillon. Pfeffermann et Sverchkov (2007) tiennent compte de cette possibilité en ajustant la méthode d'estimation sur petits domaines. Verret, Rao et Hidiroglou (2015) ont simplifié la méthode. Dans le présent article, les méthodes d'estimation sur petits domaines ne sont pas ajustées en fonction du caractère informatif; on étudie plutôt leur impact.

La présentation de l'article est la suivante. Les estimateurs ponctuels et les estimateurs de l'erreur quadratique moyenne associés pour les modèles d'estimation au niveau de l'unité et au niveau du domaine sont décrits à la section 2 et à la section 3 respectivement. La simulation et les résultats sont présentés à la section 4. La simulation calcule les estimateurs ponctuels et les erreurs quadratiques moyennes associées pour un plan d'échantillonnage avec probabilités proportionnelles à la taille avec remise (PPTAR) en faisant varier les deux facteurs suivants : (a) le modèle supposé est exact ou inexact, et (b) le plan de sondage est présumé non informatif ou très informatif. La section 5 présente un exemple d'utilisation des données de Battese, Harter et Fuller (1988) pour comparer les estimations au niveau de l'unité et au niveau du domaine. Enfin, les conclusions des travaux sont exposées à la section 6.

2 Modèle d'estimation au niveau de l'unité

L'un des modèles de base pour l'estimation sur petits domaines au niveau de l'unité est le modèle de régression à erreur emboîtée (Battese et coll. 1988) donné par $y_{ij} = \mathbf{x}'_{ij}\beta + v_i + e_{ij}$, $j = 1,\ldots,N_i$, $i = 1,\ldots,m$, où y_{ij} est la variable d'intérêt pour la j^{e} unité de population du i^{e} petit domaine, $\mathbf{x}_{ij} = \left(x_{ij1},\ldots,x_{ijp}\right)'$ est un vecteur $p \times 1$ de variables auxiliaires où $x_{ij1} = 1$, $\boldsymbol{\beta} = \left(\beta_0,\ldots,\beta_{p-1}\right)'$ est un vecteur $p \times 1$ de paramètres de régression et N_i est le nombre d'unités de population dans le i^{e} petit domaine. Les effets aléatoires v_i sont présumés indépendants et identiquement distribués $(i.i.d.)$ $N\left(0,\sigma_v^2\right)$ et indépendants des erreurs au niveau de l'unité e_{ij}, qui sont présumées $i.i.d.$ $N\left(0,\sigma_e^2\right)$. À supposer que N_i est grand, le paramètre d'intérêt correspond à la moyenne pour le i^{e} domaine, $\bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$, qui peut être approximée par :

$$\theta_i = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + v_i, \tag{2.1}$$

où $\bar{\mathbf{X}}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij} \Big/ N_i$ est le vecteur des moyennes de population connues de $\mathbf{x}_{ij}$ pour le i^{e} domaine. On présume que les échantillons sont tirés indépendamment dans chaque petit domaine selon un plan d'échantillonnage spécifié. Sous un échantillonnage non informatif, les données d'échantillon $\left(y_{ij}, \mathbf{x}_{ij}\right)$ sont présumées obéir au modèle de population, c'est-à-dire

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1,\ldots,n_i, \quad i = 1,\ldots,m, \tag{2.2}$$

où w_{ij} est le poids de sondage de base associé à l'unité (i, j) et n_i est la taille de l'échantillon dans le i^{e} petit domaine.

2.1 Estimation EBLUP

Selon le modèle de régression à erreur emboîtée (2.2), l'estimateur de la meilleure prédiction linéaire sans biais (BLUP) de la moyenne d'un petit domaine, $\theta_i = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + v_i$, est donné par

$$\tilde{\theta}_i = r_i \bar{y}_i + \left(\bar{\mathbf{X}}_i - r_i \bar{\mathbf{x}}_i\right)' \tilde{\boldsymbol{\beta}}, \tag{2.3}$$

où $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \Big/ n_i$, $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} \Big/ n_i$, $r_i = \sigma_v^2 \Big/ \left(\sigma_v^2 + \sigma_e^2 \Big/ n_i\right)$, et

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \bar{\mathbf{x}}_i' \mathbf{V}_i^{-1} \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{m} \bar{\mathbf{x}}_i' \mathbf{V}_i^{-1} y_i\right) \equiv \tilde{\boldsymbol{\beta}}\left(\sigma_e^2, \sigma_v^2\right), \tag{2.4}$$

où $\mathbf{x}_i' = \left(x_{i1},\ldots,x_{in_i}\right)$, $\mathbf{V}_i = \sigma_e^2 \mathbf{I}_{n_i} + \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}'$, $y_i = \left(y_{i1},\ldots,y_{in_i}\right)'$, $i = 1,\ldots,m$. Les deux estimations $\tilde{\theta}_i$ et $\tilde{\boldsymbol{\beta}}$ dépendent des paramètres de variance inconnus σ_e^2 et σ_v^2. On peut utiliser la méthode d'ajustement des constantes pour estimer σ_e^2 et σ_v^2; les estimateurs résultants sont $\hat{\sigma}_e^2 = (n - m - p + 1)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2$ et $\hat{\sigma}_v^2 = \max\left(\tilde{\sigma}_v^2, 0\right)$, où $\tilde{\sigma}_v^2 = n_*^{-1}\left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{u}_{ij}^2 - (n-p)\hat{\sigma}_e^2\right]$, $n_* = n - \text{tr}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1} \sum_{i=1}^{m} n_i^2 \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'\right]$, $\mathbf{X}' = \left(x_1',\ldots,x_m'\right)$ et $n = \sum_{i=1}^{m} n_i$.

Les résidus $\left\{\hat{\varepsilon}_{ij}\right\}$ sont obtenus par la régression par les moindres carrés ordinaires (MCO) de $y_{ij} - \bar{y}_i$ sur $\left\{\mathbf{x}_{ij1} - \bar{\mathbf{x}}_{i\cdot 1},\ldots,\mathbf{x}_{ijp} - \bar{\mathbf{x}}_{i\cdot p}\right\}$ et les résidus $\left\{\hat{u}_{ij}\right\}$, par la régression par les MCO de y_{ij} sur $\left\{x_{ij1},\ldots,x_{ijp}\right\}$. Plus pour de détails, voir Rao (2003, page 138).

En remplaçant σ_e^2 et σ_v^2 par les estimateurs $\hat{\sigma}_e^2$ et $\hat{\sigma}_v^2$ dans l'équation (2.3), on obtient l'estimateur EBLUP de la moyenne de petit domaine θ_i suivant :

$$\hat{\theta}_i^{\text{EBLUP}} = r_i \bar{y}_i + \left(\bar{\mathbf{X}}_i - \hat{r}_i \bar{\mathbf{x}}_i\right)' \hat{\boldsymbol{\beta}}, \tag{2.5}$$

où $\hat{r}_i = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \Big/ n_i\right)$ et $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$. L'erreur quadratique moyenne (EQM) de l'estimateur EBLUP $\hat{\theta}_i^{\text{EBLUP}}$ est donnée par

$$\text{EQM}\left(\hat{\theta}_i^{\text{EBLUP}}\right) \approx g_{1i}\left(\sigma_e^2, \sigma_v^2\right) + g_{2i}\left(\sigma_e^2, \sigma_v^2\right) + g_{3i}\left(\sigma_e^2, \sigma_v^2\right)$$

voir Prasad et Rao (1990). Les termes g sont

$$\begin{aligned} g_{1i}\left(\sigma_e^2,\sigma_v^2\right) &= (1-r_i)\,\sigma_v^2, \\ g_{2i}\left(\sigma_e^2,\sigma_v^2\right) &= \left(\overline{\mathbf{X}}_i - r_i\overline{\mathbf{x}}_i\right)'\left(\sum_{i=1}^m \mathbf{x}_i'\mathbf{V}_i^{-1}\mathbf{x}_i\right)^{-1}\left(\overline{\mathbf{X}}_i - r_i\overline{\mathbf{x}}_i\right) \end{aligned}$$

et

$$g_{3i}\left(\sigma_e^2,\sigma_v^2\right) = n_i^{-2}\left(\sigma_v^2 + \sigma_e^2 n_i^{-1}\right)^{-3} h\left(\sigma_e^2,\sigma_v^2\right),$$

où $h\left(\sigma_e^2,\sigma_v^2\right) = \sigma_e^4 V\left(\tilde{\sigma}_v^2\right) - 2\sigma_e^2\sigma_v^2\,\mathrm{cov}\left(\hat{\sigma}_e^2,\tilde{\sigma}_v^2\right) + \sigma_v^4 V\left(\hat{\sigma}_e^2\right)$. Les variances et la covariance de $\hat{\sigma}_e^2$ et $\tilde{\sigma}_v^2$ sont données par

$$\begin{aligned} V\left(\hat{\sigma}_e^2\right) &= 2\left(n-m-p+1\right)^{-1}\sigma_e^4 \\ V\left(\tilde{\sigma}_v^2\right) &= 2n_*^{-2}\left[\left(n-m-p+1\right)^{-1}(m-1)(n-p)\,\sigma_e^4 + 2n_*\sigma_e^2\sigma_v^2 + n_{**}\sigma_v^4\right], \end{aligned}$$

et

$$\mathrm{cov}\left(\hat{\sigma}_e^2,\tilde{\sigma}_v^2\right) = -(m-1)\,n_*^{-1}V\left(\hat{\sigma}_e^2\right),$$

où $n_{**} = \mathrm{tr}\left(\mathbf{Z}'\mathbf{M}\mathbf{Z}\right)^2$, $\mathbf{M} = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$,– $\mathbf{Z} = \mathrm{diag}\left(\mathbf{1}_{n_1},\ldots,\mathbf{1}_{n_m}\right)$.

Un estimateur de deuxième ordre sans biais de l'EQM (Prasad et Rao 1990) est donné par

$$\mathrm{eqm}\left(\hat{\theta}_i^{\mathrm{EBLUP}}\right) = g_{1i}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right) + 2g_{3i}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right). \tag{2.6}$$

Soulignons que l'estimateur EBLUP $\hat{\theta}_i^{\mathrm{EBLUP}}$ donné par (2.5) dépend du modèle d'estimation au niveau de l'unité (2.2). Il est sans biais par rapport au modèle, mais il n'est pas convergent par rapport au plan de sondage sauf si ce dernier repose sur un échantillonnage aléatoire simple. Si le modèle (2.2) n'est plus vérifié pour les données échantillonnées, l'estimateur EBLUP $\hat{\theta}_i^{\mathrm{EBLUP}}$ peut alors être biaisé, c'est-à-dire qu'il comprend un biais additionnel attribuable à la spécification inexacte du modèle.

2.2 Estimation pseudo-EBLUP

You et Rao (2002) ont proposé un estimateur pseudo-EBLUP de la moyenne de petit domaine θ_i combinant les poids de l'enquête et le modèle d'estimation au niveau de l'unité (2.2) afin d'atteindre la convergence par rapport au plan. Soient w_{ij} les poids associés à chaque unité (i,j). Un estimateur direct fondé sur le plan de sondage de la moyenne de petit domaine est donné par

$$\overline{y}_{iw} = \frac{\sum_{j=1}^{n_i} w_{ij}y_{ij}}{\sum_{j=1}^{n_i} w_{ij}} = \sum_{j=1}^{n_i}\tilde{w}_{ij}y_{ij}, \tag{2.7}$$

où $\tilde{w}_{ij} = w_{ij}\Big/\sum_{j=1}^{n_i} w_{ij} = w_{ij}/w_{i\cdot}$ et $\sum_{j=1}^{n_i}\tilde{w}_{ij} = 1$. L'estimateur pondéré $\overline{y}_{iw}$ est aussi appelé « estimateur pondéré de Hájek ». En combinant l'estimateur direct (2.7) et le modèle d'estimation au niveau de l'unité (2.2), on peut obtenir le modèle au niveau du domaine agrégé (pondéré par les poids d'enquête) suivant :

$$\overline{y}_{iw} = \overline{\mathbf{x}}_{iw}'\boldsymbol{\beta} + v_i + \overline{e}_{iw}, \quad i = 1, \ldots, m, \tag{2.8}$$

où $\overline{e}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} e_{ij}$ avec $E(\overline{e}_{iw}) = 0$, $V(\overline{e}_{iw}) = \sigma_e^2 \sum_{j=1}^{n_i} \tilde{w}_{ij}^2 \equiv \delta_i^2$ et $\overline{\mathbf{x}}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} \mathbf{x}_{ij}$. Soulignons que le paramètre de régression $\boldsymbol{\beta}$ et les composantes de variance σ_e^2 et σ_v^2 ne sont pas connus dans le modèle (2.8). Selon le modèle (2.8), en supposant que les paramètres $\boldsymbol{\beta}$, σ_e^2 et σ_v^2 sont connus, l'estimateur BLUP de θ_i est donné par

$$\tilde{\theta}_{iw} = r_{iw}\overline{y}_{iw} + \left(\overline{\mathbf{X}}_i - r_{iw}\overline{\mathbf{x}}_{iw}\right)' \boldsymbol{\beta} = \tilde{\theta}_{iw}\left(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2\right), \tag{2.9}$$

où $r_{iw} = \sigma_v^2 / \left(\sigma_v^2 + \delta_i^2\right)$. L'estimateur BLUP $\tilde{\theta}_{iw}$ dépend de $\boldsymbol{\beta}$, σ_e^2 et σ_v^2. Pour estimer le paramètre de régression, You et Rao (2002) ont proposé une méthode d'équation d'estimation pondérée, qui permet d'obtenir un estimateur de $\boldsymbol{\beta}$ comme suit :

$$\tilde{\boldsymbol{\beta}}_w = \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\left(\mathbf{x}_{ij} - r_{iw}\overline{\mathbf{x}}_{iw}\right)'\right]^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\left(\mathbf{x}_{ij} - r_{iw}\overline{\mathbf{x}}_{iw}\right) y_{ij}\right] \equiv \tilde{\boldsymbol{\beta}}_w\left(\sigma_e^2, \sigma_v^2\right).$$

$\tilde{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w\left(\sigma_e^2, \sigma_v^2\right)$ dépend de σ_e^2 et σ_v^2. En remplaçant σ_e^2 et σ_v^2 dans $\tilde{\boldsymbol{\beta}}_w$ par les estimateurs d'ajustement des constantes $\hat{\sigma}_e^2$ et $\hat{\sigma}_v^2$, on obtient $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$; voir Rao (2003, page 149). En remplaçant $\boldsymbol{\beta}$, σ_e^2 et σ_v^2 dans (2.9) par $\hat{\boldsymbol{\beta}}_w$, $\hat{\sigma}_e^2$ et $\hat{\sigma}_v^2$, l'estimateur pseudo-EBLUP de la moyenne de petit domaine θ_i est donné par

$$\hat{\theta}_i^{P-\mathrm{EBLUP}} \triangleq \hat{\theta}_{iw} = \hat{r}_{iw}\overline{y}_{iw} + \left(\overline{\mathbf{X}}_i - \hat{r}_{iw}\overline{\mathbf{x}}_{iw}\right)' \hat{\boldsymbol{\beta}}_w. \tag{2.10}$$

À mesure que la taille de l'échantillon n_i augmente, l'estimateur $\hat{\theta}_i^{P-\mathrm{EBLUP}}$ devient convergent par rapport au plan de sondage. Il a aussi une propriété d'autocalage lorsque les poids w_{ij} sont ajustés de façon à correspondre au total de population connu. Ainsi, si $\sum_{j=1}^{n_i} w_{ij} = N_i$, $\sum_{i=1}^{m} N_i \hat{\theta}_i^{P-\mathrm{EBLUP}}$ correspond à l'estimateur direct de régression du total global

$$\sum_{i=1}^{m} N_i \hat{\theta}_i^{P-\mathrm{EBLUP}} = \hat{Y}_w + \left(\mathbf{X} - \hat{\mathbf{X}}_w\right)' \hat{\boldsymbol{\beta}}_w,$$

où $\hat{Y}_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij} y_{ij}$, et $\hat{\mathbf{X}}_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}$. Pour plus de détails, voir You et Rao (2002).

L'EQM de $\hat{\theta}_i^{P-\mathrm{EBLUP}}$ est donnée par

$$\mathrm{EQM}\left(\hat{\theta}_i^{P-\mathrm{EBLUP}}\right) \approx g_{1iw}\left(\sigma_e^2, \sigma_v^2\right) + g_{2iw}\left(\sigma_e^2, \sigma_v^2\right) + g_{3iw}\left(\sigma_e^2, \sigma_v^2\right),$$

où $g_{1iw}\left(\sigma_e^2, \sigma_v^2\right) = (1 - r_{iw})\sigma_v^2$ et $g_{2iw}\left(\sigma_e^2, \sigma_v^2\right) = \left(\overline{X}_i - r_{iw}\overline{x}_{iw}\right)'\Phi_w\left(\overline{X}_i - r_{iw}\overline{x}_{iw}\right)$. Le terme Φ_w est

$$\Phi_w\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}z_{ij}'\right)^{-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{z}_{ij}\mathbf{z}_{ij}'\right)\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}_{ij}'\right)^{-1}\right]'\sigma_e^2$$

$$+ \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}_{ij}'\right)^{-1}\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)'\right]\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}\mathbf{z}_{ij}\right)^{-1}\right]'\sigma_v^2,$$

où $\mathbf{z}_{ij} = w_{ij}\left(\mathbf{x}_{ij} - r_{iw}\overline{\mathbf{x}}_{iw}\right)$ et $g_{3iw}\left(\sigma_e^2, \sigma_v^2\right) = r_{iw}\left(1 - r_{iw}\right)^2 \sigma_e^{-4}\sigma_v^{-2} h\left(\sigma_e^2, \sigma_v^2\right)$. Le facteur $h\left(\sigma_e^2, \sigma_v^2\right)$ correspond à la même fonction que pour l'EQM de l'estimateur EBLUP donné à la section 2.1. Un estimateur de deuxième ordre presque sans biais de l'EQM peut s'écrire

$$\text{eqm}\left(\hat{\theta}_i^{P-\text{EBLUP}}\right) = g_{1iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right). \tag{2.11}$$

(Voir Rao 2003, page 150 et You et Rao 2002, page 435). Soulignons que l'estimateur de l'EQM (2.11) ne tient pas compte des termes du produit vectoriel. Torabi et Rao (2010) ont obtenu un estimateur de deuxième ordre de l'EQM exact tenant compte des termes du produit vectoriel à l'aide des méthodes de linéarisation et de « bootstrap ». Le produit vectoriel compte deux termes. Le premier est simple et a une forme explicite. Bien que la méthode de linéarisation fonctionne bien, la forme explicite du deuxième terme du produit vectoriel est très longue; de plus, les formules fondées sur la méthode de linéarisation ne sont pas fournies dans l'article de Torabi et Rao (2010). La méthode « bootstrap » sous-estime toujours l'EQM réelle. Pour obtenir un estimateur non biaisé de l'EQM, il faut appliquer une méthode bootstrap double exigeant beaucoup de calculs. L'estimateur de l'EQM (2.11) se comporte comme l'estimateur par linéarisation de Torabi et Rao (2010) lorsque la variation des poids d'enquête est faible. Dans le cas de l'autopondération à l'intérieur des domaines, l'un des termes du produit vectoriel est zéro et l'autre est de l'ordre $o\left(m^{-1}\right)$. En conséquence, l'estimateur de l'EQM (2.11) est presque sans biais; d'autres détails sont présentés dans Torabi et Rao (2010). C'est pour ces raisons que les termes du produit vectoriel n'ont pas été inclus dans l'estimateur de l'EQM donné en (2.11) dans le cadre de l'étude.

Soulignons qu'en vertu du modèle (2.2), l'estimateur pseudo-EBLUP $\hat{\theta}_i^{P-\text{EBLUP}}$ est légèrement moins efficient que l'estimateur EBLUP $\hat{\theta}_i^{\text{EBLUP}}$. Toutefois, cet estimateur pseudo-EBLUP est convergent par rapport au plan et est donc plus robuste à une spécification inexacte du modèle. L'efficacité des estimateurs EBLUP et pseudo-EBLUP a été évaluée à l'aide d'une étude par simulations.

3 Modèle d'estimation au niveau du domaine

Le modèle de Fay-Herriot (Fay et Herriot 1979) est un modèle d'estimation au niveau du domaine de base couramment utilisé pour l'estimation sur petits domaines afin d'améliorer les estimations d'enquête directes. Le modèle de Fay-Herriot a deux composantes, soit un modèle d'échantillonnage pour les estimations d'enquête directes et un modèle de lien pour les paramètres d'intérêt du petit domaine. Le modèle d'échantillonnage suppose que pour une taille d'échantillon de domaine spécifique $n_i > 1$, il existe un estimateur d'enquête direct $\hat{\theta}_i^{\text{DIR}}$. Cet estimateur d'enquête direct est sans biais sous le plan pour le paramètre de petit domaine θ_i. Le modèle d'échantillonnage est donné par

$$\hat{\theta}_i^{\text{DIR}} = \theta_i + e_i,\ i = 1, \ldots, m, \tag{3.1}$$

où e_i est l'erreur d'échantillonnage associée à l'estimateur direct $\hat{\theta}_i^{\text{DIR}}$ et m est le nombre de petits domaines. Dans la pratique, il est courant de supposer que les variables e_i sont des variables aléatoires normales indépendantes de moyenne $E\left(e_i\right) = 0$ et de variance d'échantillonnage $\text{var}\left(e_i\right) = \sigma_i^2$. Le modèle de lien est obtenu en supposant que le paramètre de petit domaine d'intérêt θ_i est lié aux variables auxiliaires au niveau du domaine $\mathbf{z}_i = \left(\text{z}_{i1}, \ldots, \text{z}_{ip}\right)'$ par le modèle de régression linéaire suivant :

$$\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i,\ i = 1,\ldots,m, \tag{3.2}$$

où $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)'$ est un vecteur $p \times 1$ de coefficients de régression et où les termes v_i sont des effets aléatoires propres au domaine présumés être $i.i.d.$, avec $E(v_i) = 0$ et $\text{var}(v_i) = \sigma_v^2$. On émet aussi généralement une hypothèse de normalité, même si elle est plus difficile à justifier. Une telle hypothèse est nécessaire pour obtenir l'estimation de l'EQM. La variance du modèle σ_v^2 est inconnue et doit être estimée à partir des données. L'effet aléatoire au niveau du domaine v_i rend compte de l'hétérogénéité non structurée entre les domaines que n'expliquent pas les variances d'échantillonnage. La combinaison des modèles (3.1) et (3.2) produit un modèle linéaire mixte au niveau du domaine donné par

$$\hat{\theta}_i^{\text{DIR}} = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i. \tag{3.3}$$

Le modèle (3.3) comprend des erreurs aléatoires fondées sur le plan e_i et des effets aléatoires fondés sur le modèle v_i. Aux fins du modèle de Fay-Herriot, la variance d'échantillonnage σ_i^2 est présumée être connue dans le modèle (3.3). Il s'agit d'une hypothèse très forte. On utilise généralement des estimateurs lissés des variances d'échantillonnage dans le modèle de Fay-Herriot; les paramètres σ_i^2 sont ensuite considérés comme connus. Toutefois, si des estimateurs directs des variances d'échantillonnage sont utilisés dans le modèle de Fay-Herriot, il faut ajouter un terme à l'estimateur de l'EQM pour tenir compte de la variation additionnelle (Wang et Fuller 2003).

Si l'on suppose que la variance du modèle σ_v^2 est connue, le meilleur prédicteur linéaire sans biais (BLUP) du paramètre de petit domaine θ_i peut s'écrire

$$\tilde{\theta}_i = \gamma_i \hat{\theta}_i^{\text{DIR}} + (1-\gamma_i)\,\mathbf{z}_i'\tilde{\boldsymbol{\beta}}_{\text{MCP}}, \tag{3.4}$$

où $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$, et $\tilde{\boldsymbol{\beta}}_{\text{MCP}}$ est l'estimateur des moindres carrés pondérés (MCP) de $\boldsymbol{\beta}$ donné par

$$\tilde{\boldsymbol{\beta}}_{\text{MCP}} = \left[\sum_{i=1}^{m} (\sigma_i^2 + \sigma_v^2)^{-1}\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\left[\sum_{i=1}^{m} (\sigma_i^2 + \sigma_v^2)^{-1}\mathbf{z}_i y_i\right] = \left[\sum_{i=1}^{m} \gamma_i \mathbf{z}_i\mathbf{z}_i'\right]^{-1}\left[\sum_{i=1}^{m} \gamma_i \mathbf{z}_i y_i\right].$$

Il existe plusieurs méthodes pour estimer la variance du modèle inconnue σ_v^2; You (2010) présente une vue d'ensemble de ces méthodes. Les auteurs ont choisi la méthode du maximum de vraisemblance restreint (méthode REML) mise au point par Cressie (1992) pour estimer la variance du modèle en vertu du modèle de Fay-Herriot. À l'aide de l'algorithme de score, on obtient l'estimateur REML $\hat{\sigma}_v^2$ suivant :

$$\sigma_v^{2(k+1)} = \sigma_v^{2(k)} + \left[I_R\left(\sigma_v^{2(k)}\right)\right]^{-1} S_R\left(\sigma_v^{2(k)}\right),\ \text{pour}\ k = 1,2,\ldots,$$

où $I_R(\sigma_v^2) = 1/2\,\text{tr}[\mathbf{PP}]$, et $S_R(\sigma_v^2) = 1/2\,\mathbf{y}'\mathbf{PPy} - 1/2\,\text{tr}[\mathbf{P}]$, et $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}$. Si on utilise une valeur supposée pour $\sigma_v^{2(1)}$ comme valeur de départ, l'algorithme converge très rapidement.

En remplaçant σ_v^2 dans l'équation (3.4) par l'estimateur REML $\hat{\sigma}_v^2$, on obtient l'estimateur EBLUP du paramètre de petit domaine θ_i fondé sur le modèle de Fay-Herriot suivant :

$$\hat{\theta}_i^{\text{FH}} = \hat{\gamma}_i \hat{\theta}_i^{\text{DIR}} + (1-\hat{\gamma}_i)\,\mathbf{z}_i'\hat{\boldsymbol{\beta}}_{\text{MCP}}, \tag{3.5}$$

où $\hat{\gamma}_i = \hat{\sigma}_v^2 \big/ \left(\hat{\sigma}_v^2 + \sigma_i^2\right)$. L'estimateur de l'EQM de $\hat{\theta}_i^{\mathrm{FH}}$ est donné par (voir Rao 2003)

$$\mathrm{eqm}\left(\hat{\theta}_i^{\mathrm{FH}}\right) = g_{1i} + g_{2i} + 2g_{3i}, \tag{3.6}$$

où g_{1i} est le terme principal, g_{2i} rend compte de la variabilité attribuable à l'estimation du paramètre de régression β, et g_{3i} est attribuable à l'estimation de la variance du modèle. Ces termes g sont définis comme suit :

$$g_{1i} = \hat{\gamma}_i \sigma_i^2, g_{2i} = (1-\hat{\gamma}_i)^2 \mathbf{z}_i' \mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{MCP}}\right) \mathbf{z}_i = \hat{\sigma}_v^2 (1-\hat{\gamma}_i)^2 \mathbf{z}_i' \left(\sum_{i=1}^{m} \hat{\gamma}_i \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \mathbf{z}_i$$

et $g_{3i} = \left(\sigma_i^2\right)^2 \left(\hat{\sigma}_v^2 + \sigma_i^2\right)^{-3} \mathrm{var}\left(\hat{\sigma}_v^2\right)$.

La variance estimée de $\hat{\sigma}_v^2$ est donnée par $\mathrm{var}\left(\hat{\sigma}_v^2\right) = 2\left(\sum_{i=1}^{m} \left(\hat{\sigma}_v^2 + \sigma_i^2\right)^{-2}\right)^{-1}$; voir Datta et Lahiri (2000).

Jusqu'à maintenant, on a supposé que la variance d'échantillonnage σ_i^2 est présumée connue sous le modèle de Fay-Herriot (3.3). Il s'agit d'une hypothèse très forte. En règle générale, on connaît un estimateur d'enquête direct, disons s_i^2, de la variance d'échantillonnage σ_i^2. Comme ces variances estimées peuvent être très variables, elles sont lissées au moyen de modèles externes et de fonctions généralisées de variance; ces variances lissées sont désignées $\tilde{s}_i^2$. Les estimations lissées de la variance d'échantillonnage $\tilde{s}_i^2$ sont utilisées dans le modèle de Fay-Herriot et considérées comme connues. La valeur $\mathrm{eqm}\left(\hat{\theta}_i^{\mathrm{FH}}\right)$ associée est obtenue en remplaçant σ_i^2 par $\tilde{s}_i^2$ dans l'équation (3.6). Rivest et Vandal (2003) et Wang et Fuller (2003) ont étudié l'estimation de petit domaine sous le modèle de Fay-Herriot à l'aide des estimations directes de la variance d'échantillonnage s_i^2 en vertu de l'hypothèse que les estimateurs s_i^2 sont indépendants des estimateurs d'enquête directs y_i et $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, où $d_i = n_i - 1$ et n_i est la taille de l'échantillon pour le i^{e} domaine. Quand on utilise l'estimation directe de la variance d'échantillonnage s_i^2 au lieu de la variance d'échantillonnage réelle σ_i^2, un terme additionnel rendant compte de l'incertitude liée à l'utilisation de s_i^2 doit être intégré à l'estimateur de l'EQM (3.6); ce terme, désigné par g_{4i}, est donné par

$$g_{4i} = \frac{4}{n_i - 1} \frac{\hat{\sigma}_v^4 s_i^4}{\left(\hat{\sigma}_v^2 + s_i^2\right)^3};$$

voir Rivest et Vandal (2003) et Wang et Fuller (2003) pour les détails.

Pour appliquer le modèle de Fay-Herriot, il faut obtenir les estimations directes au niveau du domaine et les estimations de la variance d'échantillonnage correspondantes, qui serviront de valeurs d'entrée au modèle de Fay-Herriot. On tient compte de trois estimateurs directs au niveau du domaine, soit l'estimateur direct de la moyenne de l'échantillon en supposant un échantillonnage aléatoire simple (EAS), l'estimateur de Horvitz-Thompson (HT) et l'estimateur pondéré de Hájek (HA). L'estimateur pondéré de Hájek est aussi utilisé dans l'estimateur pseudo-EBLUP pour le modèle d'estimation au niveau de l'unité désigné par $\bar{y}_{iw}$ dans l'équation (2.7). Le tableau 3.1 présente ces trois estimateurs directs au niveau du domaine et les estimateurs de la variance d'échantillonnage correspondants.

Tableau 3.1
Estimateurs directs au niveau du domaine et variances d'échantillonnage

	Estimateur ponctuel	Estimateur de la variance d'échantillonnage
Moyenne directe (EAS)	$\hat{\theta}_i^{\text{EAS}} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$	$\text{var}\left(\hat{\theta}_i^{\text{EAS}}\right) = \frac{1}{n_i\,(n_i-1)}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{\theta}_i^{\text{EAS}}\right)^2$
Estimateur de Horvitz-Thompson (HT)	$\hat{\theta}_i^{\text{HT}} = \frac{1}{N_i}\sum_{j=1}^{n_i} w_{ij}y_{ij} = \frac{1}{N_i}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i p_{ij}}$	$\text{var}\left(\hat{\theta}_i^{\text{HT}}\right) = \frac{1}{N_i^2 n_i\,(n_i-1)}\sum_{j=1}^{n_i}\left(\frac{y_{ij}}{p_{ij}} - N_i\hat{\theta}_i^{\text{HT}}\right)^2$
Estimateur pondéré de Hájek (HA)	$\hat{\theta}_i^{\text{HA}} = \frac{\sum_{j=1}^{n_i} w_{ij}y_{ij}}{\sum_{j=1}^{n_i} w_{ij}} = \frac{1}{\hat{N}_i}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i p_{ij}}$	$\text{var}\left(\hat{\theta}_i^{\text{HA}}\right) = \frac{1}{\hat{N}_i^2 n_i\,(n_i-1)}\sum_{j=1}^{n_i}\left(\frac{y_{ij}-\hat{\theta}_i^{\text{HA}}}{p_{ij}}\right)^2$

Ces estimateurs au niveau du domaine servent de valeurs d'entrée au modèle de Fay-Herriot. Les trois estimateurs fondés sur le modèle au niveau du domaine sont désignés comme suit : FH-EAS, FH-HT et FH-HA. On remplace donc $\hat{\theta}_i^{\text{DIR}}$ par $\hat{\theta}_i^{\text{EAS}}$, $\hat{\theta}_i^{\text{HT}}$ ou $\hat{\theta}_i^{\text{HA}}$ dans (3.5) pour obtenir l'estimateur fondé sur le modèle correspondant $\hat{\theta}_i^{\text{FH-EAS}}$, $\hat{\theta}_i^{\text{FH-HT}}$ ou $\hat{\theta}_i^{\text{FH-HA}}$. L'estimateur direct sous EAS $\hat{\theta}_i^{\text{EAS}}$ ne tient pas compte du plan d'échantillonnage et ne converge pas par rapport au plan, sauf si ce dernier repose sur un échantillonnage aléatoire simple. Soulignons que les estimateurs $\hat{\theta}_i^{\text{HT}}$ et $\hat{\theta}_i^{\text{HA}}$ convergent par rapport au plan. Il s'ensuit que les estimateurs fondés sur le modèle correspondants $\hat{\theta}_i^{\text{FH-HT}}$ et $\hat{\theta}_i^{\text{FH-HA}}$ convergent par rapport au plan à mesure que la taille de l'échantillon augmente. En outre, cela signifie que ces estimateurs sont robustes à la spécification inexacte du modèle.

Dans la section qui suit, on compare le modèle au niveau de l'unité avec le modèle de Fay-Herriot au moyen d'une étude par simulations. Les paramètres statistiques utilisés pour ces comparaisons sont le biais, la racine de l'EQM relative et les intervalles de confiance des estimateurs fondés sur le modèle.

4 Étude par simulations

4.1 Génération des données

Pour comparer les estimateurs sur petits domaines au niveau de l'unité et au niveau du domaine, les auteurs ont réalisé une étude par simulations fondée sur le plan. À partir des conditions de simulation de You, Rao et Kovacevic (2003), deux populations finies ont été établies. Chaque population finie comptait $m = 30$ domaines, et chaque domaine consistait en $N_i = 200$ unités de population. Chacune des populations finies a été produite à l'aide du modèle d'estimation au niveau de l'unité $y_{ij} = \beta_0 + x_{1ij}\beta_1 + v_i + e_{ij}$. La variable auxiliaire x_{1ij} a été générée selon une loi exponentielle de moyenne 4 et de variance 8, et les composantes aléatoires ont été générées selon une loi normale avec $v_i \sim N\left(0, \sigma_v^2\right)$ et $e_{ij} \sim N\left(0, \sigma_e^2\right)$, où $\sigma_v^2 = 100$ et $\sigma_e^2 = 225$. Pour la première population, les effets fixes de régression ont été établis à $\beta_0 = 50$ et $\beta_1 = 10$ pour les 30 domaines. Pour la deuxième population, des effets fixes de différentes valeurs ont été utilisés : $\beta_0 = 50$ et $\beta_1 = 10$ pour les domaines $m = 1,\ldots,10$; $\beta_0 = 75$ et $\beta_1 = 15$ pour les domaines $m = 11,\ldots,20$; et $\beta_0 = 100$ et $\beta_1 = 20$ pour les domaines $m = 21,\ldots,30$. Il y avait trois moyennes différentes pour les effets fixes $\beta_0 + x_{1ij}\beta_1$ dans la deuxième population, alors qu'il n'y en avait qu'une dans la première population. Les échantillons PPTAR dans chaque domaine ont été tirés

indépendamment de chaque population construite. Pour mettre en œuvre l'échantillonnage PPTAR, on a d'abord défini une mesure de taille z_{ij} pour une unité donnée (i, j). À l'aide de ces valeurs z_{ij}, on a calculé les probabilités de sélection $p_{ij} = z_{ij} \Big/ \sum_j z_{ij}$ pour chaque unité (i, j), qui ont ensuite servi à sélectionner des échantillons PPTAR de tailles égales $n_i = n$. Dans chaque population générée, on a sélectionné des échantillons de taille $n = 10$ et 30. Le poids de sondage de base est donné par $w_{ij} = n_i^{-1} p_{ij}^{-1}$, de sorte que le poids normalisé correspond à $\tilde{w}_{ij} = p_{ij}^{-1} \Big/ \sum_j p_{ij}^{-1}$. On a choisi la mesure de taille z_{ij} comme une combinaison linéaire de la variable auxiliaire x_{1ij} et des données produites selon une loi exponentielle de moyenne 4 et de variance 16. Le coefficient de corrélation ρ entre y_{ij} et la probabilité de sélection p_{ij} dans chaque domaine variait entre 0,02 et 0,95. La fourchette des probabilités p_{ij} va de la sélection non informative $(\rho = 0{,}02)$ à la sélection fortement informative $(\rho = 0{,}95)$ des échantillons PPTAR. L'échantillonnage est non informatif lorsque la corrélation entre y_{ij} et la probabilité de sélection p_{ij} est très faible, ce qui signifie que l'échantillon et le modèle de population coïncident. Si la probabilité de sélection p_{ij} est fortement corrélée avec l'observation y_{ij}, l'échantillonnage est informatif, et le modèle de population n'est peut-être plus vérifié pour l'échantillon. Pour chaque population, le processus d'échantillonnage PPTAR a été répété $R = 3\,000$ fois. Comme dans Prasad et Rao (1990), l'étude par simulations est fondée sur le plan, puisque les deux populations ont été générées une seule fois, et que des échantillons répétés ont été produits à partir de la même population.

Pour la modélisation au niveau de l'unité, on a ajusté le modèle de régression à erreur emboîtée en fonction des données d'échantillonnage PPTAR générées à partir de chaque population. On a obtenu les estimations EBLUP et pseudo-EBLUP correspondantes ainsi que les estimations de l'EQM à l'aide des formules énoncées à la section 2. On a ensuite établi les estimations des intervalles de confiance en calculant la racine carrée des estimations de l'EQM; les détails du calcul sont présentés à la section 4.2.3. Pour la modélisation au niveau du domaine, on a d'abord calculé les estimations directes au niveau du domaine $\hat{\theta}_i^{\mathrm{EAS}}$, $\hat{\theta}_i^{\mathrm{HT}}$ et $\hat{\theta}_i^{\mathrm{HA}}$ ainsi que les variances d'échantillonnage correspondantes. On a ensuite appliqué le modèle de Fay-Herriot pour obtenir les estimateurs fondés sur le modèle $\hat{\theta}_i^{\mathrm{FH\text{-}EAS}}$, $\hat{\theta}_i^{\mathrm{FH\text{-}HT}}$ et $\hat{\theta}_i^{\mathrm{FH\text{-}HA}}$. La moyenne de population de la variable auxiliaire x_{1ij} de chaque domaine a été utilisée comme variable auxiliaire dans le modèle de Fay-Herriot. On a additionné g_{4i} à l'estimateur de l'EQM pour tenir compte du recours à des variances d'échantillonnage non lissées dans le modèle de Fay-Herriot. Les intervalles de confiance correspondants ont été obtenus de façon similaire pour les estimateurs EBLUP et pseudo-EBLUP au niveau de l'unité.

L'ajustement du modèle aux deux niveaux (unité et domaine) est fondé sur deux scénarios. Le premier (scénario I) suppose que la modélisation est exacte; les données ont été générées à partir de la première population et les modèles d'ajustement étaient le modèle au niveau de l'unité (2.2) et le modèle au niveau du domaine (3.3), tous deux avec le même vecteur $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Le deuxième (scénario II) suppose que la modélisation est inexacte; les données ont été générées à partir de la deuxième population avec des moyennes différentes pour les effets fixes et les mêmes modèles d'ajustement que pour le scénario I, avec un même vecteur $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Soulignons qu'en vertu du scénario I, l'échantillonnage n'est pas informatif lorsque le niveau d'unité exact (2.2) est ajusté en fonction des données de l'échantillon pour obtenir l'estimateur EBLUP; cela est vrai pour tous les coefficient de corrélation ρ entre y_{ij} et p_{ij}.

4.2 Résultats

Dans la section qui suit, on compare certaines données statistiques des estimations au niveau de l'unité et au niveau du domaine en vertu du scénario I (modélisation exacte) et du scénario II (modélisation inexacte).

4.2.1 Comparaison à l'intérieur de chaque petit domaine

À la figure 4.1, on compare les moyennes de population avec les estimations au niveau de l'unité et au niveau du domaine lorsque $n = 10$ pour le scénario I. Les résultats sont fondés sur un plan d'échantillonnage fortement informatif où le coefficient de corrélation entre y_{ij} et la probabilité de sélection p_{ij} est $\rho = 0{,}88$. Les estimations fondées sur le modèle reposent sur la moyenne de $R = 3\,000$ simulations. Les résultats présentés à la figure 4.1 indiquent clairement que les estimateurs EBLUP (équation 2.5) et pseudo-EBLUP (équation 2.10) au niveau de l'unité sont presque sans biais. Les résultats montrent que si la modélisation est exacte, l'échantillonnage n'est pas informatif pour le modèle au niveau de l'unité (2.2) et l'estimateur EBLUP est sans biais. L'estimateur FH-EAS au niveau du domaine surestime systématiquement la moyenne de population, ce qui entraîne un biais important. L'estimateur FH-HT au niveau du domaine sous-estime généralement la moyenne de population et entraîne un biais légèrement plus grand que celui de l'estimateur FH-HA. On obtient des résultats similaires pour $n = 30$.

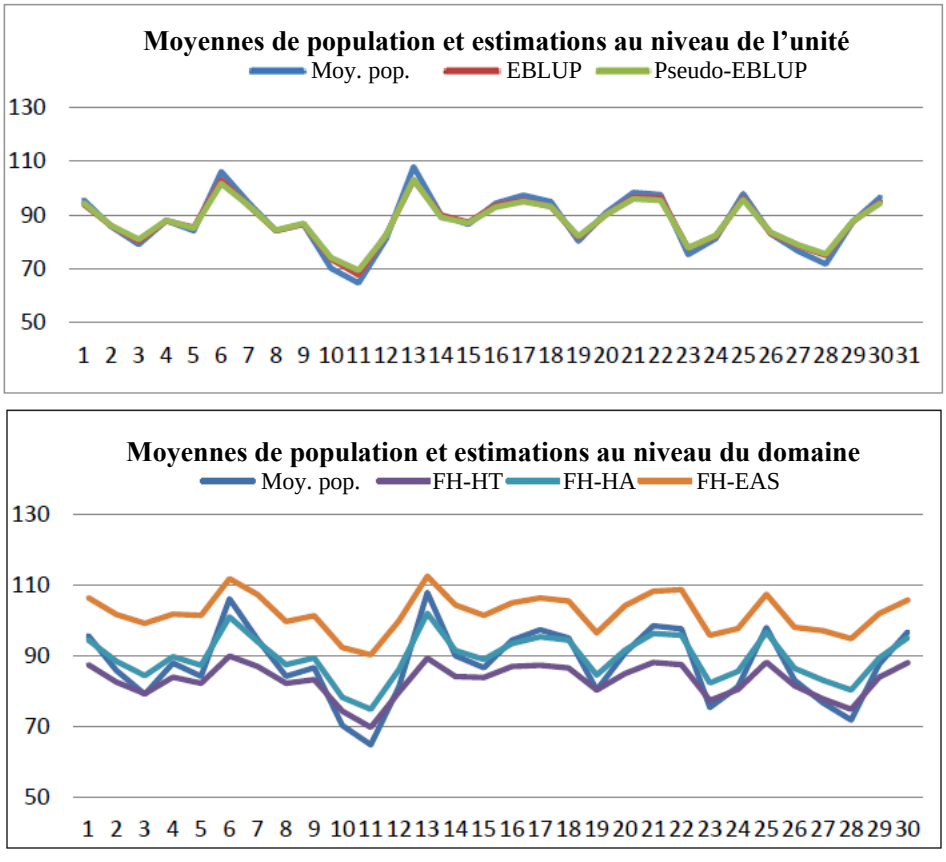

Figure 4.1 Comparaison des moyennes en vertu du scénario I pour $n = 10$.

À la figure 4.2, on compare la racine carrée moyenne de l'eqm pour les estimateurs au niveau de l'unité et au niveau du domaine pour le scénario I lorsque $n = 10$ et $n = 30$. La racine de l'eqm correspond à la racine carrée de l'EQM estimée donnée aux sections 2 et 3 pour les estimateurs au niveau de l'unité et au niveau du domaine. Il est clair que la racine de l'eqm des estimateurs EBLUP et pseudo-EBLUP est beaucoup plus faible que celle des estimateurs FH au niveau du domaine pour $n = 10$ et $n = 30$. Comme on s'y attendait (You et Rao 2002), l'estimateur EBLUP a la plus petite racine de l'eqm et l'estimateur pseudo-EBLUP, une racine de l'eqm légèrement supérieure. Les estimateurs FH-EAS au niveau du domaine ont une racine de l'eqm élevée et affichent des variations importantes. Les estimateurs FH-HT et FH-HA ont en moyenne à peu près la même racine de l'eqm, mais, comme l'illustrent les deux figures, l'estimateur FH-HT est plus variable que l'estimateur FH-HA, particulièrement lorsque $n = 10$. Lorsque $n = 30$, la variabilité de la racine de l'eqm pour les estimateurs FH-HT et FH-HA est considérablement réduite, mais il est clair que l'estimateur FH-HA est plus stable que l'estimateur FH-HT.

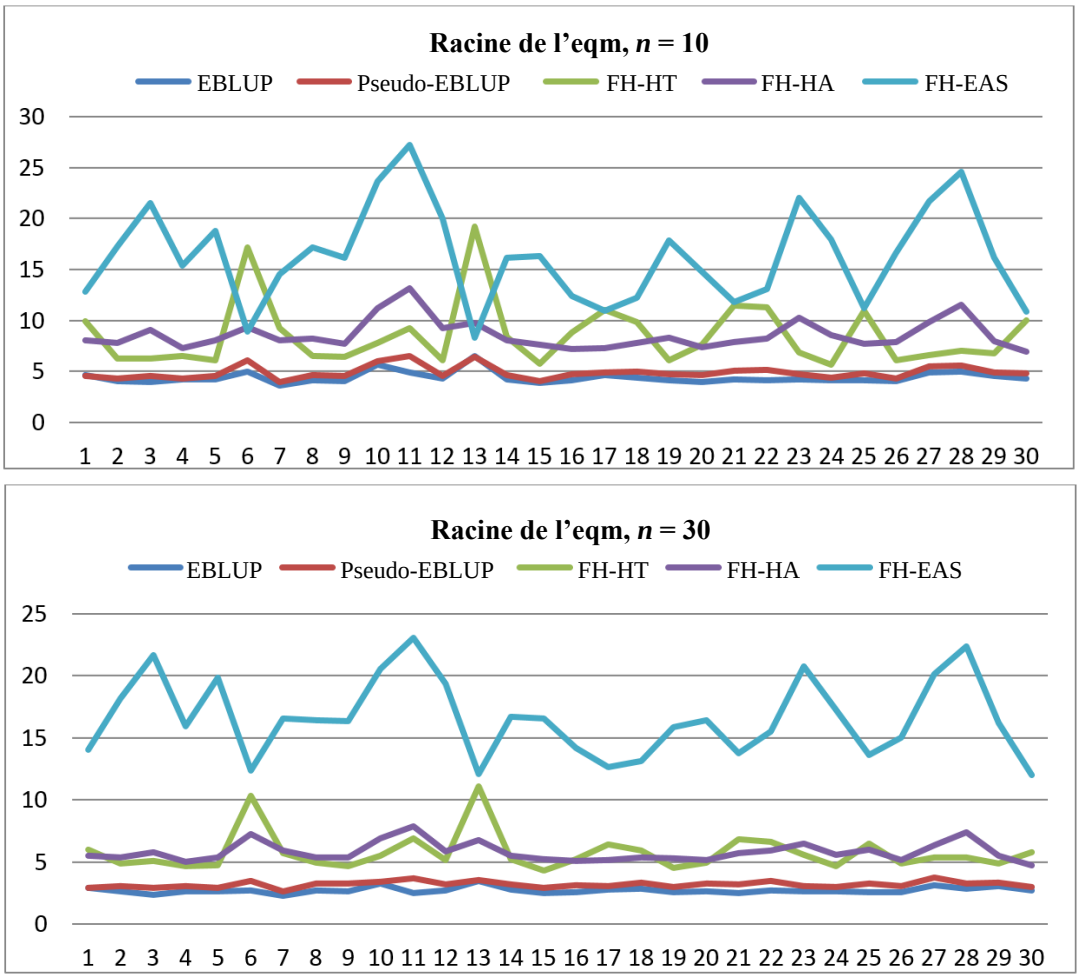

Figure 4.2 Comparaison de la racine de l'eqm en vertu du scénario I pour $n = 10$ et $n = 30$.

À la figure 4.3, on compare les estimations au niveau de l'unité et au niveau du domaine avec les moyennes de population lorsque $n = 10$ en vertu du scénario II. Dans le cas des modèles au niveau de

l'unité, il est clair que l'estimateur EBLUP sous-estime et surestime à la fois la moyenne de population lorsque la spécification du modèle est inexacte, alors que l'estimateur pseudo-EBLUP est sans biais (les estimations pseudo-EBLUP et les moyennes de population sont superposées à la figure 4.3). En ce qui concerne les estimateurs au niveau du domaine, l'estimateur FH-EAS surestime systématiquement les moyennes réelles, alors que l'estimateur FH-HT sous-estime davantage les valeurs que l'estimateur FH-HA lorsque la spécification du modèle est inexacte.

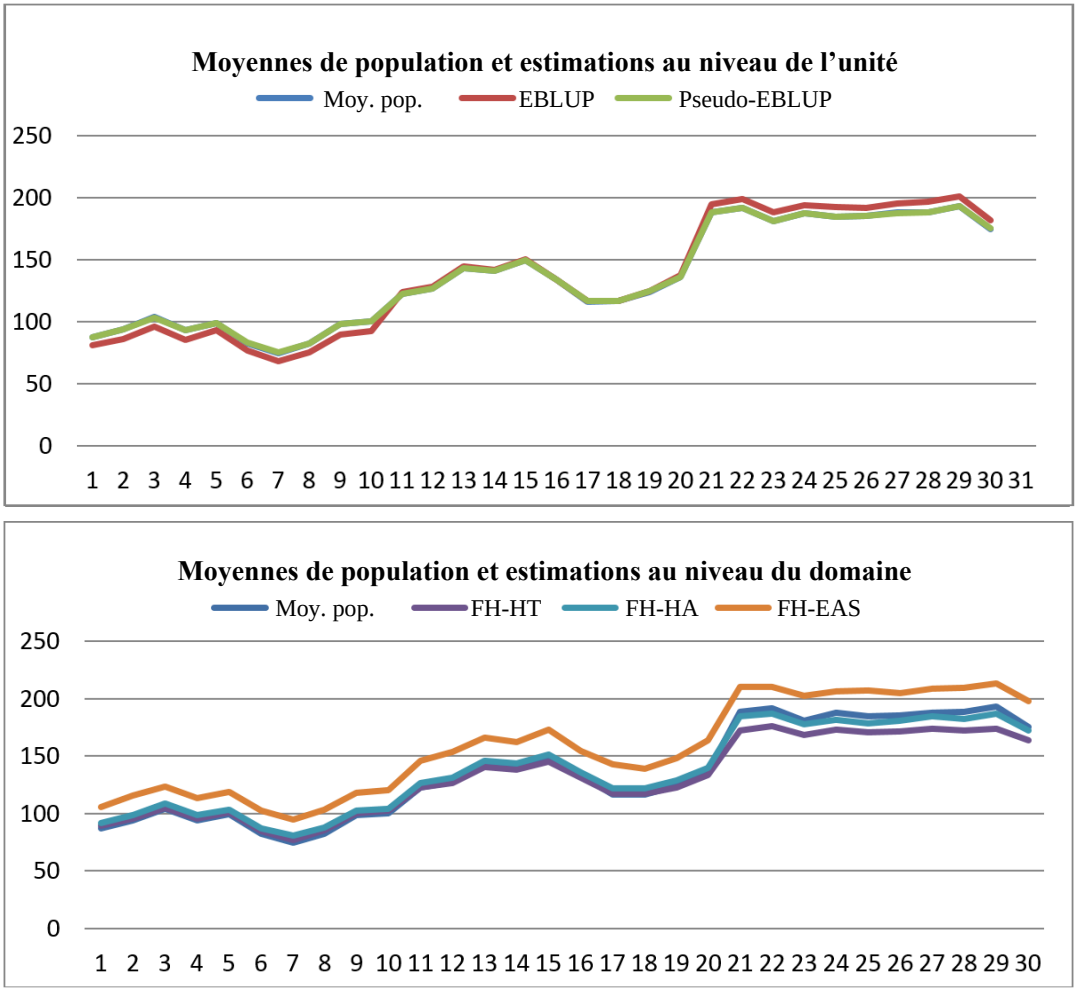

Figure 4.3 Comparaison des moyennes en vertu du scénario II sous une modélisation inexacte pour $n = 10$.

À la figure 4.4, on compare les racines de l'eqm des estimateurs au niveau de l'unité et au niveau du domaine pour les échantillons de taille $n = 10$ et $n = 30$ en vertu d'une modélisation inexacte. On voit bien à la figure 4.4 que l'estimateur pseudo-EBLUP a la plus petite racine de l'eqm lorsque la modélisation est inexacte. L'estimateur EBLUP a une très grande racine de l'eqm lorsque la spécification du modèle est inexacte; de fait, pour les domaines 1 à 10 et 21 à 30, la racine moyenne de l'eqm est 10,01, alors que pour l'estimateur pseudo-EBLUP, la racine correspondante de l'eqm est 7,38 lorsque l'échantillon est de taille $n = 10$. Lorsque l'échantillon est de taille $n = 30$, la racine moyenne de l'eqm est 8,85 pour l'estimateur

EBLUP et seulement 4,38 pour l'estimateur pseudo-EBLUP lorsque le modèle est inexact. En résumé, les résultats montrent que l'estimateur EBLUP donne lieu à des estimations biaisées et à une racine de l' eqm élevée lorsque la modélisation est inexacte.

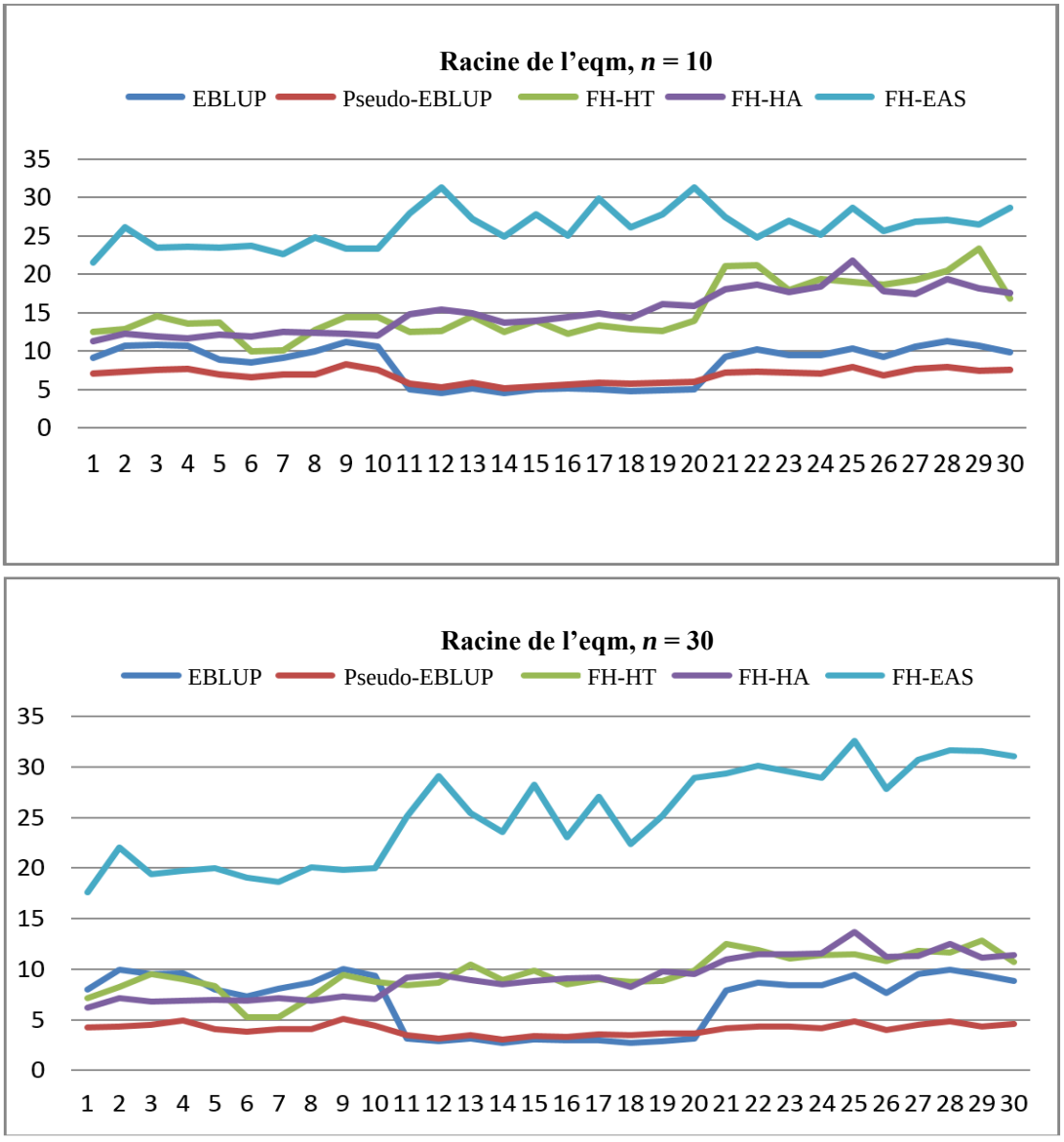

Figure 4.4 Comparaison de la racine de l' eqm en vertu du scénario II pour $n = 10$ et $n = 30$.

4.2.2 Comparaison entre petits domaines

Pour comparer les estimateurs entre domaines, on a examiné le biais relatif absolu (BRA) moyen pour un estimateur spécifié $\hat{\theta}_i$ de la moyenne de population simulée $\bar{Y}_i$ calculé comme suit : $\overline{\text{BRA}} = \left(\sum_{i=1}^{m} \text{BRA}_i\right)\Big/ m$, où

$$\text{BRA}_i = \left| \frac{1}{R} \sum_{r=1}^{R} \frac{\left(\hat{\theta}_i^{(r)} - \bar{Y}_i \right)}{\bar{Y}_i} \right|,$$

et $\hat{\theta}_i^{(r)}$ est l'estimation fondée sur le r^{e} échantillon simulé, $R = 3\,000, m = 30$. Le tableau 4.1 présente le pourcentage du biais relatif absolu moyen $\overline{\text{BRA}}$ des estimateurs au niveau de l'unité et au niveau du domaine pour les 30 domaines en vertu du scénario I. Les résultats sont fondés sur des échantillons sélectionnés de taille 10 et 30 respectivement dans chaque domaine.

Tableau 4.1
Biais relatif absolu moyen $\overline{\text{BRA}}$ en pourcentage en vertu du scénario I

Type	Estimateur	$n = 10$	$n = 30$
Au niveau de l'unité	EBLUP	1,71	0,75
	Pseudo-EBLUP	2,14	0,86
Au niveau du domaine	FH-EAS	17,51	18,64
	FH-HT	6,02	3,12
	FH-HA	4,33	2,59

Dans le cas des modèles au niveau de l'unité, il est clair que si le modèle est exact, l'échantillon devient non informatif en ce qui concerne le modèle au niveau de l'unité (2.2), et les estimateurs EBLUP et pseudo-EBLUP sont sans biais. Le biais relatif absolu moyen $\overline{\text{BRA}}$ pour l'estimateur EBLUP est de 1,71 % lorsque l'échantillon est de taille $n = 10$ et de 0,75 % lorsque l'échantillon est de taille $n = 30$. Pour l'estimateur pseudo-EBLUP, le $\overline{\text{BRA}}$ est de 2,14 % lorsque $n = 10$ et de 0,86 % lorsque $n = 30$. L'estimateur pseudo-EBLUP est associé à un biais légèrement plus élevé que celui de l'estimateur EBLUP. Dans le cas des modèles au niveau du domaine, l'estimateur FH-EAS surestime grandement les moyennes, le $\overline{\text{BRA}}$ atteignant jusqu'à 17,51 % lorsque $n = 10$ et 18,6 % lorsque $n = 30$. Les deux estimateurs au niveau du domaine FH-HT et FH-HA donnent des estimations raisonnables : (i) le $\overline{\text{BRA}}$ pour l'estimateur FH-HT est de 6,02 % lorsque $n = 10$ et de 3,12 % lorsque $n = 30$; (ii) le $\overline{\text{BRA}}$ pour l'estimateur FH-HA est de 4,33 % lorsque $n = 10$ et de 2,59 % lorsque $n = 30$. L'estimateur FH-HA donne de meilleurs résultats que l'estimateur FH-HT. Le biais relatif absolu pour les estimateurs au niveau du domaine est plus grand que celui qui est associé aux estimateurs au niveau de l'unité.

Le tableau 4.2 présente le $\overline{\text{BRA}}$ de divers estimateurs en vertu du scénario II. Il est clair que l'estimateur pseudo-EBLUP est assorti d'un $\overline{\text{BRA}}$ beaucoup plus faible que celui de l'estimateur EBLUP lorsque la modélisation est inexacte. Les $\overline{\text{BRA}}$ de l'estimateur EBLUP en vertu d'une modélisation inexacte sont de 4,31 % $(n = 10)$ et de 4,52 % $(n = 30)$. Pour l'estimateur pseudo-EBLUP, les $\overline{\text{BRA}}$ s'établissent à seulement 0,25 % $(n = 10)$ et 0,12 % $(n = 30)$. Les deux estimateurs FH-HT et FH-HA donnent de très bons résultats. Leurs $\overline{\text{BRA}}$ sont de 3,91 % et 3,48 % respectivement lorsque $n = 10$ et diminuent à 1,51% et à 1,47 % lorsque $n = 30$. L'estimateur FH-EAS donne des résultats médiocres. Les deux estimateurs au niveau du domaine FH-HT et FH-HA donnent de bons résultats, en plus de converger par rapport au plan. Encore une fois, l'estimateur FH-HA est légèrement plus intéressant que l'estimateur FH-HT en termes de $\overline{\text{BRA}}$. Les résultats montrent que le recours à des poids d'enquête dans la modélisation au niveau de l'unité joue un rôle très important lorsque la spécification du modèle au niveau de l'unité est inexacte. L'estimateur

pseudo-EBLUP est sans biais même lorsque la spécification du modèle est inexacte. Il s'agit du meilleur estimateur lorsque le modèle est inexact.

Tableau 4.2
Biais relatif absolu moyen $\overline{\text{BRA}}$ en pourcentage en vertu du scénario II

Type	Estimateur	$n = 10$	$n = 30$
Au niveau de l'unité	EBLUP	4,31	4,52
	Pseudo-EBLUP	0,25	0,12
Au niveau du domaine	FH-EAS	17,11	17,87
	FH-HT	3,91	1,51
	FH-HA	3,48	1,47

On a ensuite comparé la racine de l'EQM relative (REQMR) pour tous les estimateurs. On a notamment calculé la REQMR réelle pour la simulation et la REQMR estimée à partir des estimateurs de l'EQM. La REQMR réelle moyenne pour la simulation est calculée comme suit : $\overline{\text{REQMR}} = \left(\sum_{i=1}^{m} \text{REQMR}_i\right) \Big/ m$, où

$$\text{REQMR}_i = \frac{\sqrt{\text{EQM}_i}}{\bar{Y}_i}, \quad \text{et} \quad \text{EQM}_i = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_i^{(r)} - \bar{Y}_i\right)^2.$$

La racine de l'EQM relative estimée moyenne est calculée comme suit : $\overline{\text{ReqmR}} = \left(\sum_{i=1}^{m} \text{ReqmR}_i\right) \Big/ m$, où

$$\text{ReqmR}_i = \frac{\sqrt{\text{eqm}_i}}{\hat{\theta}_i}, \quad \text{et} \quad \text{eqm}_i = \frac{1}{R}\sum_{r=1}^{R}\text{eqm}_i^{(r)}, \quad \text{et} \quad \hat{\theta}_i = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_i^{(r)}.$$

Le paramètre $\text{eqm}_i^{(r)}$ correspond à l'EQM estimée de $\hat{\theta}_i^{(r)}$ pour le i^{e} domaine. Il est calculé à l'aide des formules énoncées aux sections 2 et 3.

Le tableau 4.3 indique la $\overline{\text{REQMR}}$ et la $\overline{\text{ReqmR}}$ pour les 30 petits domaines. Quand l'échantillon est de taille $n = 10$, la $\overline{\text{REQMR}}$ est de 4,98 % pour l'estimateur EBLUP et de 5,49 % pour l'estimateur pseudo-EBLUP. Comme prévu (You et Rao 2002), l'estimateur pseudo-EBLUP a une REQMR légèrement supérieure à celle de l'estimateur EBLUP. Au niveau de l'unité, les deux estimateurs EBLUP et pseudo-EBLUP ont une REQMR beaucoup plus faible qu'au niveau du domaine. Dans le cas des modèles au niveau du domaine, les estimateurs FH-HT et FH-HA ont un rendement similaire, la REQMR réelle moyenne correspondante s'établissant à 9,72 % et à 9,68 % respectivement lorsque $n = 10$. L'estimateur FH-EAS ne donne pas de bons résultats sous un échantillonnage informatif, la REQMR réelle moyenne s'établissant à 18,89 % lorsque $n = 10$. Même lorsque $n = 30$, la REQMR moyenne pour l'estimateur FH-EAS peut atteindre jusqu'à 18,62 %. Soulignons que la $\overline{\text{ReqmR}}$ est très proche de sa valeur réelle.

En résumé, les résultats présentés au tableau 4.3 montrent que les estimateurs EBLUP et pseudo-EBLUP au niveau de l'unité donnent de meilleurs résultats que les estimateurs FH-HT et FH-HA au niveau du domaine lorsque la modélisation est exacte. Les deux estimateurs FH-HT et FH-HA au niveau du domaine donnent des résultats raisonnablement satisfaisants sous un échantillonnage informatif. Comme prévu, l'estimateur FH-EAS ne donne pas de bons résultats.

Tableau 4.3
REQMR moyenne en pourcentage en vertu du scénario I

Type	Estimateur	$n = 10$		$n = 30$	
		$\overline{\text{REQMR}}$	$\overline{\text{ReqmR}}$	$\overline{\text{REQMR}}$	$\overline{\text{ReqmR}}$
Au niveau de l'unité	EBLUP	4,98	5,09	3,01	3,13
	Pseudo-EBLUP	5,49	5,66	3,58	3,67
Au niveau du domaine	FH-EAS	18,89	17,53	18,62	16,34
	FH-HT	9,72	10,25	6,67	6,69
	FH-HA	9,68	9,71	6,51	6,63

Le tableau 4.4 présente les résultats de la REQMR moyenne en vertu du scénario II. L'estimateur pseudo-EBLUP est le plus robuste et a les plus faibles $\overline{\text{REQMR}}$: les $\overline{\text{REQMR}}$ sont de 5,42 % et de 3,21 % pour $n = 10$ et $n = 30$ respectivement. Les estimateurs au niveau du domaine FH-HT et FH-HA ont un rendement similaire, alors que l'estimateur FH-EAS ne donne pas de bons résultats. Lorsque $n = 10,$ la $\overline{\text{REQMR}}$ est de 11,68 % pour l'estimateur FH-HT et de 11,21 % pour l'estimateur FH-HA. Lorsque $n = 30,$ la $\overline{\text{REQMR}}$ est de 7,24 % pour l'estimateur FH-HT et de 6,79 % pour l'estimateur FH-HA. Comme prévu, l'estimateur FH-EAS a une $\overline{\text{REQMR}}$ élevée sous un échantillonnage informatif. L'estimateur pseudo-EBLUP donne les meilleurs résultats en termes de biais, d'erreur type et de REQMR lorsque la spécification du modèle est inexacte. L'estimateur FH-HA donne des résultats légèrement meilleurs que ceux de l'estimateur FH-HT. La $\overline{\text{ReqmR}}$ estimée est très proche de la $\overline{\text{REQMR}}$ réelle pour tous les estimateurs.

Tableau 4.4
REQMR moyenne en pourcentage en vertu du scénario II

Type	Estimateur	$n = 10$		$n = 30$	
		$\overline{\text{REQMR}}$	$\overline{\text{ReqmR}}$	$\overline{\text{REQMR}}$	$\overline{\text{ReqmR}}$
Au niveau de l'unité	EBLUP	6,78	6,94	5,62	5,81
	Pseudo-EBLUP	5,42	5,45	3,21	3,26
Au niveau du domaine	FH-EAS	19,76	17,43	19,06	16,24
	FH-HT	11,68	11,78	7,24	7,26
	FH-HA	11,21	11,27	6,79	6,91

4.2.3 Comparaison des intervalles de confiance

On a ensuite comparé les intervalles de confiance associés aux estimateurs au niveau de l'unité et au niveau du domaine. L'intervalle de confiance se présente sous la forme $\text{estimateur} \pm z_{\alpha/2}\sqrt{\text{eqm}}$, où $z_{\alpha/2}$ correspond au $100(1-\alpha/2)\%$ centile de la distribution normale centrée réduite. Par exemple, l'intervalle de confiance à 95 % de l'estimateur EBLUP $\hat{\theta}_i^{\text{EBLUP}}$ est obtenu par $\hat{\theta}_i^{\text{EBLUP}} \pm 1{,}96\sqrt{\text{eqm}\left(\hat{\theta}_i^{\text{EBLUP}}\right)}$, où $\text{eqm}\left(\hat{\theta}_i^{\text{EBLUP}}\right)$ est donnée par (2.6). Les intervalles de confiance sont calculés comme ci-dessous. Pour un estimateur donné $\hat{\theta}_i^{(r)},\ r = 1,\ldots,R,\ i = 1,\ldots,m,$ la variable indicatrice $I_i^{(r)}$ est définie comme suit :

$$I_i^{(r)} = \begin{cases} 1 \text{ si } \theta_i \subseteq \left(\hat{\theta}_i^{(r)} - 1{,}96\sqrt{\text{eqm}\left(\hat{\theta}_i^{(r)}\right)}, \hat{\theta}_i^{(r)} + 1{,}96\sqrt{\text{eqm}\left(\hat{\theta}_i^{(r)}\right)}\right) \\ 0 \text{ sinon} \end{cases}.$$

Le taux de couverture des intervalles de confiance correspond à la moyenne des variables $I_i^{(r)}$ pour l'ensemble des $R = 3\,000$ simulations. Les tableaux 4.5 et 4.6 présentent les taux de couverture des intervalles de confiance à 95 % pour les estimateurs au niveau de l'unité et au niveau du domaine en vertu du scénario I. Le coefficient de corrélation ρ entre les probabilités de sélection p_{ij} et y_{ij} est présenté dans la première colonne pour refléter le degré du caractère informatif de l'échantillonnage PPT.

Tableau 4.5
Taux de couverture des intervalles de confiance en vertu du scénario I pour $n = 10$

Coefficient de corrélation (ρ)	EBLUP	Pseudo-EBLUP	FH-EAS	FH-HT	FH-HA
0,95	0,932	0,946	0,618	0,898	0,911
0,88	0,945	0,948	0,649	0,882	0,908
0,75	0,948	0,948	0,705	0,863	0,911
0,51	0,944	0,949	0,825	0,845	0,916
0,28	0,947	0,951	0,901	0,822	0,917
0,12	0,948	0,949	0,924	0,778	0,893
0,02	0,948	0,951	0,925	0,595	0,886
Taux moyen	*0,945*	*0,949*	*0,792*	*0,812*	*0,906*

Discutons d'abord des propriétés de couverture associées aux estimateurs au niveau de l'unité EBLUP et pseudo-EBLUP. Les tableaux montrent que, lorsque le modèle est exact, les taux de couverture des estimateurs EBLUP et pseudo-EBLUP sont assez stables : l'estimateur pseudo-EBLUP a un taux de couverture légèrement meilleur que celui de l'estimateur EBLUP. Lorsque l'échantillon est de taille $n = 10$, le taux de couverture moyen est de 94,5 % pour l'estimateur EBLUP et de 94,9 % pour l'estimateur pseudo-EBLUP. Lorsque l'échantillon est de taille $n = 30$, il est de 93,4 % pour l'estimateur EBLUP et de 94,8 % pour l'estimateur pseudo-EBLUP. Lorsque la taille de l'échantillon passe de $n = 10$ à $n = 30$, les taux de couverture de l'estimateur EBLUP se détériorent légèrement plus que ceux de l'estimateur pseudo-EBLUP. L'estimateur pseudo-EBLUP n'est pas aussi influencé par l'importance du caractère informatif découlant de l'échantillonnage PPT. Les taux de couverture relativement stables de l'estimateur EBLUP montrent que l'échantillon n'est pas informatif en ce qui concerne le modèle au niveau de l'unité exact. Toutefois, lorsque $n = 30$, l'estimateur EBLUP est associé à un taux de couverture légèrement inférieur.

Tableau 4.6
Taux de couverture des intervalles de confiance en vertu du scénario I pour $n = 30$

Coefficient de corrélation (ρ)	EBLUP	Pseudo-EBLUP	FH-EAS	FH-HT	FH-HA
0,95	0,905	0,946	0,265	0,932	0,926
0,88	0,938	0,948	0,286	0,915	0,921
0,75	0,941	0,949	0,377	0,911	0,924
0,51	0,940	0,951	0,625	0,895	0,931
0,28	0,941	0,950	0,806	0,874	0,929
0,12	0,939	0,945	0,923	0,866	0,922
0,02	0,937	0,948	0,937	0,772	0,917
Taux moyen	*0,934*	*0,948*	*0,603*	*0,881*	*0,924*

Passons maintenant aux taux de couverture associés aux estimateurs au niveau du domaine. Comme prévu, les estimateurs FH-EAS ont des taux de couverture faibles lorsque l'échantillonnage est informatif; le taux de couverture augmente à mesure que le plan d'échantillonnage devient non informatif. L'estimateur FH-HA a un meilleur taux de couverture que l'estimateur FH-HT. Le taux de couverture de l'estimateur FH-HT diminue à mesure que le plan d'échantillonnage devient non informatif. Par exemple, lorsque l'échantillon est de taille $n = 10,$ le taux de couverture pour l'estimateur FH-HT n'est que de 59,5 % lorsque l'échantillonnage est non informatif, comparativement à 88,6 % pour l'estimateur FH-HA. À mesure que la taille de l'échantillon augmente, le taux de couverture pour les estimateurs FH-HT et FH-HA s'améliore. Le taux de couverture moyen pour l'estimateur FH-HA est de 90,6 % lorsque $n = 10$ et de 92,4 % lorsque $n = 30.$ L'estimateur FH-HT a un taux de couverture inférieur à celui de l'estimateur FH-HA. Le taux de couverture moyen n'est que de 81,2 % pour l'estimateur FH-HT lorsque $n = 10.$ Le taux de couverture pour l'estimateur FH-EAS est très faible, soit 61,8 % sous un échantillonnage informatif lorsque $n = 10$ et 26,5 % lorsque $n = 30.$ À mesure que la taille de l'échantillon augmente, le taux de couverture diminue pour l'estimateur FH-EAS sous un échantillonnage informatif. Comme prévu, le taux de couverture augmente graduellement pour l'estimateur FH-EAS à mesure que l'échantillonnage devient non informatif. De tous les estimateurs, que l'échantillon soit de taille $n = 10$ ou $n = 30,$ l'estimateur pseudo-EBLUP a le meilleur taux de couverture, et l'estimateur FH-HA, le deuxième meilleur taux de couverture.

Les tableaux 4.7 et 4.8 présentent les taux de couverture en vertu du scénario II. Les résultats montrent que l'estimateur EBLUP a un faible taux de couverture sous un échantillonnage informatif, alors que l'estimateur pseudo-EBLUP a des taux de couverture très stables et élevés (tous autour de 95 % et plus) sous un échantillonnage informatif ou non informatif. Par exemple, lorsque $n = 10,$ l'estimateur EBLUP a un taux de couverture de 84,6 % sous un échantillonnage informatif (coefficient de corrélation de 0,95), qui diminue à 62,9 % lorsque la taille de l'échantillon augmente à $n = 30.$ Sous une modélisation inexacte, le taux de couverture moyen de l'estimateur EBLUP est de 90,4 % pour $n = 10$ et de 79,6 % pour $n = 30.$ Les résultats montrent que l'estimateur EBLUP est sensible à la modélisation lorsque l'échantillonnage est informatif, ce qui s'explique par le fait que cet estimateur repose entièrement sur le modèle et ne dépend pas du tout du plan d'échantillonnage.

Tableau 4.7
Taux de couverture des intervalles de confiance en vertu du scénario II pour $n = 10$

Coefficient de corrélation (ρ)	EBLUP	Pseudo-EBLUP	FH-EAS	FH-HT	FH-HA
0,95	0,846	0,965	0,701	0,865	0,896
0,88	0,855	0,964	0,729	0,887	0,893
0,75	0,881	0,962	0,787	0,873	0,898
0,51	0,921	0,961	0,872	0,848	0,898
0,28	0,936	0,961	0,912	0,843	0,887
0,12	0,945	0,955	0,917	0,765	0,867
0,02	0,943	0,951	0,913	0,592	0,838
Taux moyen	*0,904*	*0,959*	*0,833*	*0,811*	*0,883*

Tableau 4.8
Taux de couverture des intervalles de confiance en vertu du scénario II pour $n = 30$

Coefficient de corrélation (ρ)	EBLUP	Pseudo-EBLUP	FH-EAS	FH-HT	FH-HA
0,95	0,629	0,969	0,239	0,913	0,923
0,88	0,638	0,965	0,275	0,895	0,919
0,75	0,708	0,964	0,406	0,908	0,923
0,51	0,829	0,963	0,701	0,923	0,926
0,28	0,902	0,964	0,854	0,911	0,921
0,12	0,931	0,958	0,921	0,884	0,912
0,02	0,937	0,953	0,918	0,778	0,894
Taux moyen	*0,796*	*0,962*	*0,616*	*0,887*	*0,918*

Des trois estimateurs au niveau du domaine, c'est l'estimateur FH-HA qui donne les meilleurs résultats. Le taux de couverture pour l'estimateur FH-HA est très stable; le taux de couverture moyen est de 88,3 % lorsque $n = 10$ et de 91,8 % lorsque $n = 30$. L'estimateur FH-HT a le taux de couverture le plus faible lorsque l'échantillonnage est très peu informatif, particulièrement lorsque l'échantillon est de taille $n = 10$. Le taux de couverture moyen pour l'estimateur FH-HT n'est que de 81,1 % lorsque $n = 10$ et de 88,7 % lorsque $n = 30$. Les résultats montrent que l'estimateur FH-HA est supérieur à l'estimateur FH-HT. Quant à l'estimateur FH-EAS, il donne des résultats médiocres lorsque l'échantillonnage est informatif, particulièrement lorsque l'échantillon est de taille $n = 30$. Toutefois, l'estimateur FH-EAS donne des résultats relativement bons lorsque l'échantillonnage devient non informatif. Le taux de couverture moyen pour l'estimateur FH-EAS est de 83,3 % lorsque $n = 10$, mais seulement de 61,6 % lorsque l'échantillon est de taille $n = 30$.

Il est clair que l'estimateur pseudo-EBLUP a un taux de couverture très élevé et stable sous une modélisation inexacte. L'estimateur FH-HA a aussi un taux de couverture très stable, mais légèrement inférieur. Les taux de couverture des estimateurs EBLUP et FH-EAS diminuent à mesure que la taille de l'échantillon augmente, particulièrement lorsque l'échantillonnage est informatif.

5 Application aux données réelles

Dans la section qui suit, on compare les estimations au niveau de l'unité et au niveau du domaine au moyen d'une analyse de données réelles. L'ensemble de données étudié est celui présenté par Battese et coll. (1988) dans le cadre d'une étude estimant le nombre moyen d'hectares consacrés à la culture du maïs et du soja par segment dans douze comtés du Centre-Nord de l'Iowa. De ces douze comtés, trois ne comportaient qu'un seul segment échantillonné. Aux fins de la présente étude, les données de ces trois comtés ont été regroupées en un seul, ce qui donne un ensemble de données contenant 10 comtés dont la taille d'échantillon n_i varie entre 2 et 5 dans chaque comté. Le nombre total de segments N_i (taille de la population) dans chaque comté allait de 402 à 1 505. Suivant la méthode de You et Rao (2002), on a présumé un échantillonnage aléatoire simple (EAS) dans chaque comté, et le poids d'enquête de base a été calculé comme suit : $w_{ij} = N_i / n_i$. Pour la modélisation au niveau de l'unité, y_{ij} correspond au nombre d'hectares

de maïs (ou de soja) dans le j^{e} segment du i^{e} comté, les variables auxiliaires étant le nombre de pixels classés comme étant du maïs ou du soja selon Battese et coll. (1988). On a appliqué le modèle au niveau de l'unité à l'ensemble de données modifié et calculé les estimations EBLUP et pseudo-EBLUP. Pour la modélisation au niveau du domaine, on a d'abord calculé les estimations directes sur échantillon $\hat{\theta}_i^{\mathrm{EAS}}$ fondées sur l'EAS. On a ensuite appliqué le modèle de Fay-Herriot aux estimations directes au niveau du domaine et calculé les estimations FH-EAS au niveau du domaine. La figure 5.1 illustre la comparaison entre les estimations directes au niveau du domaine et les estimations fondées sur le modèle au niveau de l'unité et au niveau du domaine. En termes d'estimation ponctuelle, les estimations EBLUP et pseudo-EBLUP sont presque identiques, comme dans You et Rao (2002). Ce résultat s'explique par le fait que le modèle au niveau de l'unité est exact pour ces données (Battese et coll. 1988). Les estimations FH-EAS au niveau du domaine fondées sur le modèle et les estimations directes au niveau du domaine concordent assez bien dans cet exemple.

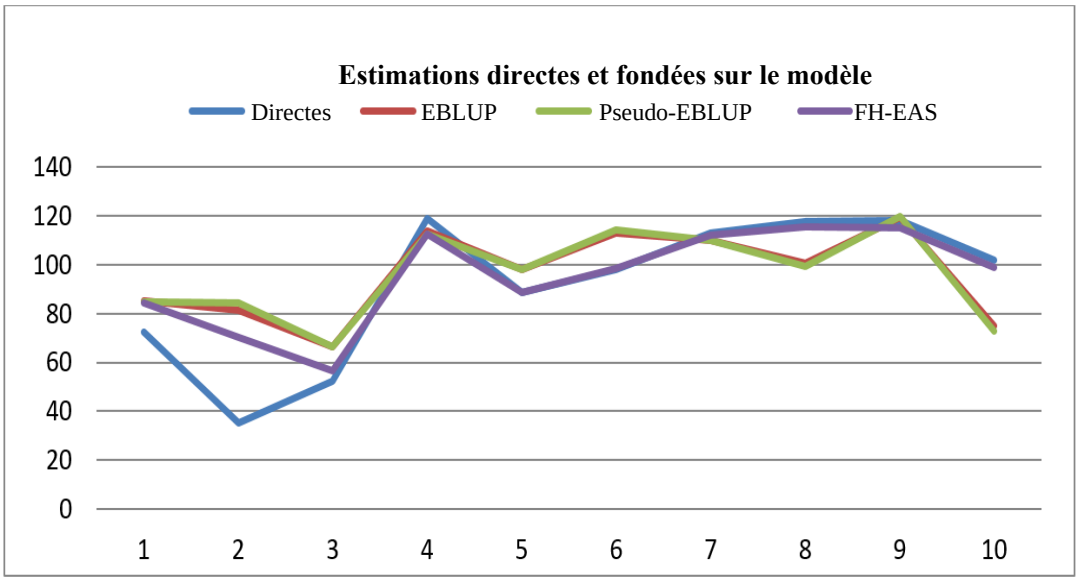

Figure 5.1 Comparaison des estimations directes et des estimations fondées sur le modèle.

La figure 5.2 illustre la comparaison entre les erreurs types des estimateurs directs et fondés sur le modèle. Les erreurs types des estimateurs fondés sur le modèle correspondent à la racine carrée de l'EQM estimée. Les deux estimateurs au niveau de l'unité, EBLUP et pseudo-EBLUP, ont des erreurs types faibles et stables. Comme prévu, l'estimateur pseudo-EBLUP est assorti d'erreurs types légèrement plus grandes que celles de l'estimateur EBLUP. Il est clair que les erreurs types des estimateurs directs et FH-EAS sont très variables et très instables. Cet exemple illustre l'efficacité des estimateurs au niveau de l'unité EBLUP et pseudo-EBLUP.

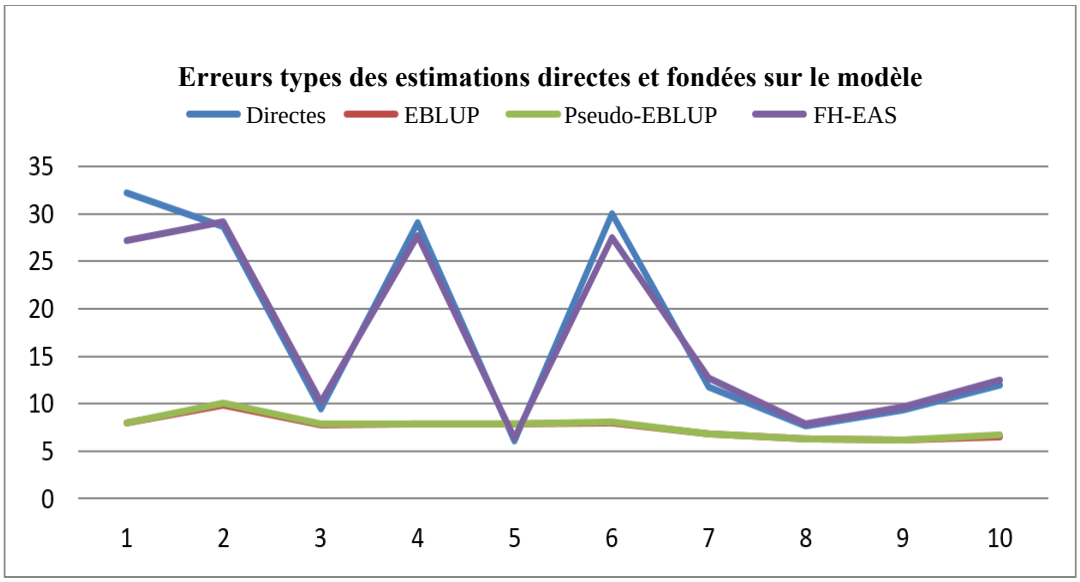

Figure 5.2 Comparaison des erreurs types des estimations directes et des estimations fondées sur le modèle.

6 Conclusions

Dans le présent article, les auteurs ont comparé l'efficacité des estimateurs fondés sur un modèle de régression à erreur emboîtée au niveau de l'unité et sur le modèle de Fay-Herriot au niveau du domaine à l'aide d'une étude par simulations fondée sur le plan. Ils ont comparé les estimations ponctuelles et les taux de couverture des intervalles de confiance des estimateurs au niveau de l'unité et au niveau du domaine. Dans l'ensemble, l'estimateur pseudo-EBLUP au niveau de l'unité est le plus efficace en termes de biais et de taux de couverture, que l'échantillonnage soit ou non informatif. L'estimateur EBLUP est efficace sous une modélisation exacte, puisque l'échantillonnage est non informatif en vertu du modèle au niveau de l'unité exact décrit en (2.2). L'estimateur pseudo-EBLUP est également assez robuste à une spécification inexacte du modèle. En pratique, les auteurs recommandent de construire les estimateurs pseudo-EBLUP à l'aide des poids d'enquête et des observations au niveau de l'unité dont il est question à la section 2.2. Dans le cas des modèles au niveau du domaine, l'estimateur FH-HA donne de meilleurs résultats que l'estimateur FH-HT; l'estimateur FH-EAS donne des résultats médiocres. On recommande donc de construire les estimateurs HA pondérés et d'appliquer ensuite le modèle de Fay-Herriot pour obtenir les estimateurs fondés sur le modèle correspondants si on utilise des estimateurs sur petits domaines au niveau du domaine.

Remerciements

Les auteurs remercient le rédacteur en chef adjoint et deux évaluateurs pour leurs suggestions et commentaires, qui ont permis d'améliorer considérablement la présentation des résultats. Ils tiennent à remercier plus particulièrement l'un des évaluateurs pour ses commentaires fort minutieux et constructifs.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Cressie, N. (1992). Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Baye. *Techniques d'enquête*, 18, 1, 83-103.

Datta, G.S., et Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistics Sinica*, 10, 613-627.

Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Pfeffermann, D., et Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 480, 1427-1439.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Rivest, L.-P., et Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, (Éd., J.N.K. Rao).

Torabi, M., et Rao, J.N.K. (2010). The mean squared error estimators of small area means using survey weights. *The Canadian Journal of Statistics*, 38, 598-608.

Verret, F., Rao, J.N.K. et Hidiroglou, M.A. (2015). Estimation sur petits domaines fondée sur un modèle sous échantillonnage informatif. *Techniques d'enquête*, 41, 2, 353-368.

Wang, J., et Fuller, W.A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2010). *Small Area Estimation under the Fay-Herriot Model Using Different Model Variance Estimation Methods and Different Input Sampling Variances*. Document de travail de la Direction de la méthodologie, SRID-2010-003E, Statistique Canada, Ottawa, Canada.

You, Y., et Rao, J.N.K. (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K. et Kovacevic, M. (2003). Estimation des effets fixes et des composantes de la variance par un modèle à valeur aléatoire à l'origine en utilisant des données d'enquête. Recueil : Symposium 2003, *Défis reliés à la réalisation d'enquêtes pour la prochaine décennie*, Statistique Canada.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 69-93
Statistique Canada, nº 12-001-X au catalogue

Comparaison de certains estimateurs de variance positifs pour le modèle d'estimation sur petits domaines Fay-Herriot

Susana Rubin-Bleuer et Yong You[1]

Résumé

La méthode du maximum de vraisemblance restreint (méthode REML pour *restricted maximum likelihood*) est généralement utilisée pour estimer la variance de l'effet aléatoire de domaine sous le modèle de Fay-Herriot (Fay et Herriot 1979) afin d'obtenir le meilleur estimateur linéaire sans biais empirique (estimateur EBLUP pour *empirical best linear unbiased predictor*) d'une moyenne de petit domaine. Lorsque l'estimation REML correspond à zéro, le poids de l'estimateur d'échantillon direct est zéro et l'EBLUP devient un estimateur synthétique, ce qui est rarement souhaitable. Pour résoudre le problème, Li et Lahiri (2011) et Yoshimori et Lahiri (2014) ont élaboré des estimateurs de variance constante par la méthode du maximum de vraisemblance ajusté (méthode ADM pour *adjusted maximum likelihood*), qui produisent toujours des estimations de variance positives. Certains des estimateurs ADM produisent toujours des estimations positives, mais génèrent un biais élevé, ce qui influe sur l'estimation de l'erreur quadratique moyenne (EQM) de l'estimateur EBLUP. Nous proposons d'utiliser un estimateur de variance MIX, défini comme étant une combinaison des méthodes REML et ADM. Nous montrons que cet estimateur est sans biais jusqu'à l'ordre deux et qu'il produit toujours une estimation de variance positive. Nous proposons également un estimateur de l'EQM sous la méthode MIX et montrons au moyen d'une simulation fondée sur un modèle que, dans de nombreuses situations, cet estimateur donne de meilleurs résultats que d'autres estimateurs de l'EQM par « linéarisation de Taylor » récemment proposés.

Mots-clés : Estimation de variance; maximum de vraisemblance ajusté; REML; ordre du biais; estimation de l'EQM.

1 Introduction

Le modèle de Fay-Herriot (Fay et Herriot 1979) est un modèle de base au niveau du domaine utilisé pour estimer les moyennes de petit domaine lorsque les estimations d'enquête directes disponibles sont imprécises en raison de la petite taille des échantillons. Dans ce modèle, la moyenne de petit domaine est représentée par un terme linéaire non aléatoire dans les covariables, avec un effet aléatoire de domaine. Dans le modèle de Fay-Herriot, on peut obtenir le meilleur estimateur linéaire sans biais (estimateur BLUP pour *best linear unbiased predictor*) d'une moyenne de petit domaine en minimisant l'erreur quadratique moyenne (EQM) dans la classe des estimateurs linéaires sans biais. L'estimateur BLUP correspond à une moyenne pondérée de l'estimateur d'enquête direct et de l'estimateur synthétique de type régression, les poids dépendant de la variance des effets aléatoires de domaine, σ_v^2. En règle générale, cette variance doit être estimée à partir des données sous le modèle de Fay-Herriot. On obtient le meilleur estimateur linéaire sans biais empirique (EBLUP) de la moyenne de petit domaine en remplaçant la variance dans la formule de l'estimateur BLUP par une estimation. De nombreuses méthodes bien connues d'estimation de variance sont utilisées dans ce contexte, mais la plus courante est la méthode du maximum de vraisemblance restreint (méthode REML pour *restricted maximum likelihood*), car elle tient compte de la perte des degrés de liberté attribuable à l'estimation du coefficient de régression. De plus, cette méthode est sans biais jusqu'à l'ordre deux et exige moins d'itérations, car elle converge plus rapidement. Malgré ces caractéristiques

1. Susana Rubin-Bleuer et Yong You, Division de la coopération internationale et des méthodes statistiques institutionnelles, Statistique Canada. Courriel : susana.rubin-bleuer@canada.ca; yong.you@canada.ca.

importantes, il arrive parfois, particulièrement lorsque le nombre de domaines, m, est petit ou modéré, que la méthode REML produise une estimation de variance nulle. Cela donnerait un poids nul à l'estimateur d'enquête direct dans la formule EBLUP, et l'estimateur EBLUP devient donc un estimateur synthétique de type régression. La plupart des praticiens hésitent toutefois à utiliser les estimateurs synthétiques pour les moyennes de petit domaine, car ceux-ci ne tiennent pas compte des données d'enquête et sont souvent très biaisés. Lorsqu'on a affaire à des ensembles de données réels, pour lesquels les modèles ne sont jamais parfaits, une estimation positive pour σ_v^2 réduit le biais de l'estimateur EBLUP par rapport au modèle synthétique. Certes, une estimation positive de la variance des effets aléatoires produit un estimateur EBLUP « prudent », en ce sens qu'elle attribue un poids positif à l'estimateur d'enquête direct. Elle peut également être considérée comme la somme de l'estimateur de régression et du terme non nul qui tient compte d'une partie du « biais de modèle ». Cette caractéristique donne lieu à une série de méthodes d'estimation de variance qui donnent des estimations positives.

Dans cet article, nous mettons l'accent sur les estimateurs de variance par maximum de vraisemblable ajusté mis au point par Lahiri et Li (2009), et nous proposons un estimateur de variance MIX. Notre estimateur de variance MIX combine un estimateur REML et une des méthodes d'estimation par maximum de vraisemblance ajusté. Nous proposons également un estimateur de l'EQM de l'EBLUP sous la méthode MIX et examinons les propriétés théoriques et en échantillon fini de l'estimateur de variance MIX et de l'estimateur de l'EQM.

Morris (2006) et Lahiri et Li (2009) ont proposé des estimateurs de variance par maximum de vraisemblable ajusté découlant de l'optimisation de la vraisemblance profilée et de la vraisemblance résiduelle ajustée par un facteur $h\left(\sigma_v^2\right), \sigma_v^2 > 0.$ Li et Lahiri (2011) ont proposé deux méthodes d'estimation de variance (méthodes AM.LL et AR.LL, associées respectivement à la vraisemblance profilée et à la vraisemblance résiduelle) qui garantissent des estimations positives avec un facteur d'ajustement $h_{\mathrm{LL}}\left(\sigma_v^2\right) = \sigma_v^2.$ Yoshimori et Lahiri (2014) ont proposé deux autres méthodes d'estimation de variance (méthodes AM.YL et AR.YL) par ajustement de la vraisemblance profilée et de la vraisemblance résiduelle avec le facteur

$$h_{\mathrm{YL}}\left(\sigma_v^2\right) = \left\{\arctan\left[\sum_{i=1}^{m} \sigma_v^2 \Big/ \left(\sigma_v^2 + \psi_i\right)\right]\right\}^{1/m}$$

où ψ_i est la variance d'échantillonnage pour le i^{e} domaine. Il est bien connu que les estimateurs LL sont biaisés, particulièrement lorsque le nombre de domaines est faible ou modéré (Lahiri et Pramanik 2011). La méthode YL, qui ajuste la vraisemblance profilée, produit aussi un estimateur biaisé de σ_v^2. Le biais de l'estimateur de variance n'affecte toutefois pas l'EQM de l'EBLUP. En effet, l'approximation asymptotique d'ordre deux de l'EQM montre que l'EQM dépend de la variance asymptotique et non du biais de l'estimateur de variance. Cependant, le biais des estimateurs de variance affecte les estimateurs de l'EQM par linéarisation de Taylor et peut produire des estimateurs de l'EQM présentant un biais négatif. Il est alors souhaitable d'examiner d'autres estimateurs de variance positifs.

Yuan (2009) a été le premier à mentionner la méthode consistant à combiner les estimateurs de variance AM.LL et REML pour le modèle Fay-Herriot, mais il n'a pas étudié ses propriétés, empiriques ou autres.

Rubin-Bleuer, Yung et Landry (2010, 2011 et 2012) ont effectué des comparaisons empiriques d'un estimateur de variance MIX dans un modèle chronologique et transversal au niveau du domaine, tandis que Rubin-Bleuer et You (2012) ont étudié les propriétés asymptotiques et en échantillon fini de l'estimateur de variance MIX pour le modèle de Fay-Herriot.

Dans cet article, nous formalisons la méthode MIX pour le modèle de Fay-Herriot et prouvons que l'estimateur de variance MIX est sans biais jusqu'à l'ordre deux. En outre, nous proposons un estimateur de l'EQM par linéarisation de Taylor. Nous examinons également les résultats empiriques de l'estimateur MIX pour un nombre faible ou modéré de domaines. En ce qui concerne l'estimation de l'EQM, Rubin-Bleuer et You (2012) et Molina, Rao et Datta (2015) ont tous proposé des estimateurs d'EQM « fractionnés » différents sous l'estimation de variance MIX. Nous montrons que les estimateurs de l'EQM de Rubin-Bleuer et You (2012) et de Molina et coll. (2015) sont sans biais jusqu'à l'ordre deux. Ces estimateurs d'EQM « fractionnés » ont été assujettis à une règle pour les populations qui a donné des estimations nulles sous l'estimation de variance REML et à une autre règle pour les populations qui a donné des estimations positives sous l'estimation de variance REML. Les deux articles susmentionnés ont montré que, pour un petit nombre de domaines, ces estimateurs « fractionnés » ont donné de bons résultats empiriques en termes de biais relatif moyen. Or, ce résultat pourrait être trompeur, car les estimateurs de l'EQM présentent généralement un biais négatif pour les populations où l'estimation de variance REML est nulle, et un biais positif pour les populations où l'estimation REML est positive, le biais s'annulant en moyenne. Étant donné ce qui précède, nous proposons un nouvel estimateur de l'EQM, et nous le comparons à d'autres pour les populations où l'estimation REML est nulle.

Dans la section 2, nous présentons le modèle de Fay-Herriot, l'estimateur EBLUP de la moyenne de petit domaine et une approximation d'ordre deux de l'EQM de l'EBLUP sous le modèle. Dans la section 3, nous décrivons l'estimateur REML et les estimateurs de variance *.LL et *.YL. Dans la section 4, nous présentons un estimateur de variance MIX général et prouvons que son biais est du même ordre que celui de l'estimateur REML. Nous proposons un estimateur sans biais (jusqu'à l'ordre deux) de l'EQM sous la méthode MIX. Dans la section 5, nous menons une étude empirique afin de comparer les différents estimateurs de variance. Il est à noter que nous avons défini l'estimateur de variance MIX comme étant une combinaison de l'estimateur REML et d'un des estimateurs de variance par maximum de vraisemblable ajusté, mais l'estimateur de variance MIX que nous avons choisi pour cette étude combine l'estimateur REML et l'estimateur de variance AM.LL. Nous avons sélectionné cette combinaison, parce que Li et Lahiri (2011) ont déclaré que la méthode de la vraisemblance profilée ajustée avait donné de meilleurs résultats que celle de la vraisemblance résiduelle ajustée (AR.LL) et que le facteur d'ajustement des estimateurs de variance de Yoshimori et Lahiri (2014) était trop proche de zéro (en termes logarithmiques) pour améliorer de façon significative la méthode REML. Enfin, dans la section 6, nous présentons et analysons les résultats de la simulation et tirons des conclusions.

2 EBLUP et EQM de l'EBLUP sous le modèle de Fay-Herriot

Soit $y_i, i = 1, \ldots, m,$ les estimateurs d'enquête directs des moyennes de petit domaine $\theta_i, i = 1, \ldots, m.$ Le modèle de Fay-Herriot se compose des modèles d'échantillonnage et de lien suivants :

$$\textbf{Modèle d'échantillonnage :}\ y_i = \theta_i + e_i, e_i|\theta_i \overset{\text{i.d.}}{\sim} (0, \psi_i),\ \ i = 1, \ldots, m, \tag{2.1}$$

$$\textbf{Modèle de lien :}\ \theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i,\ \ v_i \overset{\text{i.i.d.}}{\sim} \left(0, \sigma_v^2\right),\ \ \sigma_v^2 > 0,\ \ i = 1, \ldots, m, \tag{2.2}$$

où les e_i sont les erreurs d'échantillonnage, indépendamment distribuées de moyenne de zéro et de variances d'échantillonnage « connues » ψ_i, $\mathbf{z}_i\ (p \times 1)$ sont des vecteurs connus de valeurs de covariables; $\boldsymbol{\beta}$ est un vecteur $p \times 1$ de coefficients de régression fixes inconnus; et v_i sont des effets aléatoires indépendants et identiquement distribués avec une moyenne de zéro et une variance de modèle σ_v^2. La combinaison de (2.1) et (2.2) donne :

$$y_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i,\ \ i = 1, \ldots, m, \tag{2.3}$$

avec des erreurs de modèle et d'échantillonnage. Les $y_i, i = 1, \ldots, m,$ peuvent être considérés comme des résultats dans l'espace conjoint plan de sondage-modèle (Rubin-Bleuer et Schiopu-Kratina 2005).

Dans le modèle (2.3), l'estimateur EBLUP de la moyenne de petit domaine θ_i est donné par :

$$\hat{\theta}_i\left(\hat{\sigma}_v^2\right) = \mathbf{z}_i'\hat{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2\right) + \hat{\gamma}_i\left[y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2\right)\right] = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)\,\mathbf{z}_i'\hat{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2\right),\ \ i = 1, \ldots, m, \tag{2.4}$$

où $\hat{\sigma}_v^2$ est un estimateur convergent de σ_v^2,

$$\hat{\gamma}_i = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \psi_i\right),\ \ \text{et}\ \ \hat{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2\right) = \left[\sum_{i=1}^{m} \mathbf{z}_i\mathbf{z}_i' \Big/ \left(\hat{\sigma}_v^2 + \psi_i\right)\right]^{-1}\left[\sum_{i=1}^{m} \mathbf{z}_i y_i \Big/ \left(\hat{\sigma}_v^2 + \psi_i\right)\right]. \tag{2.5}$$

Pour calculer l'erreur quadratique moyenne (EQM) de l'EBLUP, nous posons les conditions de régularité suivantes :

1) Les $\boldsymbol{\psi}_i$ ont une borne supérieure et sont loin de zéro;
2) Les $\mathbf{z}_i, 1 \le i \le m$ sont bornés;
3) $\liminf \boldsymbol{\lambda}_{\min}\left(1/m \sum_i \mathbf{z}_i \cdot \mathbf{z}_i'\right) > 0$ où $\boldsymbol{\lambda}_{\min}(A)$ = valeur propre minimum de la matrice A.

Sous la normalité des erreurs d'échantillonnage e_i associées au modèle (2.3) et les conditions de régularité ci-dessus, une approximation d'ordre deux de l'EQM est donnée par :

$$\text{EQM}\left\{\hat{\theta}_i\left(\hat{\sigma}_v^2\right)\right\} = g_{1i}\left(\sigma_v^2\right) + g_{2i}\left(\sigma_v^2\right) + g_{3i}\left(\sigma_v^2\right) + o\left(\frac{1}{m}\right), \tag{2.6}$$

avec $g_{1i}\left(\sigma_v^2\right) = \gamma_i\psi_i$, $g_{2i}\left(\sigma_v^2\right) = (1 - \gamma_i)^2\,\mathbf{z}_i'\left[\sum_{i=1}^{m} \mathbf{z}_i\mathbf{z}_i' \Big/ \left(\sigma_v^2 + \psi_i\right)\right]^{-1}\mathbf{z}_i$ et

$$g_{3i}\left(\sigma_v^2\right) = (\psi_i)^2\,\bar{V}\left(\hat{\sigma}_v^2\right) \Big/ \left(\sigma_v^2 + \psi_i\right)^3, \tag{2.7}$$

où $\bar{V}\left(\hat{\sigma}_v^2\right)$ est la variance asymptotique de $\hat{\sigma}_v^2$ (Das, Jiang et Rao 2004).

3 Examen des méthodes REML et du maximum de vraisemblance ajusté

3.1 Méthode REML

Nous examinons le modèle combiné de Fay-Herriot (2.3) où $\sigma_v^2 > 0$. On obtient l'estimateur de variance REML de σ_v^2 en maximisant la fonction de vraisemblance résiduelle pour σ_v^2 :

$$L_{\text{REML}}\left(\sigma_v^2\right) \propto \left|\left[\sum_{i=1}^{m} \mathbf{z}_i \mathbf{z}_i' \Big/ \left(\sigma_v^2 + \psi_i\right)\right]\right|^{-1/2} \prod_{i=1}^{m} \left(\sigma_v^2 + \psi_i\right)^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}\right\}$$

où $\mathbf{y} = (y_1, \ldots, y_m)'$, $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}$, $\mathbf{V} = \text{Var}(\mathbf{y})$, et $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m)'$. (Cressie 1992; Datta et Lahiri 2000; Rao 2003, chapitre 6). L'estimateur de variance REML est donné par :

$$\hat{\sigma}_{v\text{REML}}^2 = \max\left(\tilde{\sigma}_{v\text{REML}}^2, 0\right), \tag{3.1}$$

où $\tilde{\sigma}_{v\text{REML}}^2$ est la valeur convergente de l'algorithme REML. Le biais asymptotique et la variance de l'estimateur REML jusqu'à l'ordre deux sont donnés respectivement par :

$$\text{Biais}\left(\hat{\sigma}_{v\text{REML}}^2\right) = o\left(\frac{1}{m}\right) \text{ et } V\left(\hat{\sigma}_{v\text{REML}}^2\right) = \frac{2}{\text{tr}\left(\mathbf{V}^{-2}\right)} + o\left(\frac{1}{m}\right). \tag{3.2}$$

Un estimateur sans biais d'ordre deux de l'EQM de l'EBLUP sous l'estimation de variance REML est donné par (Datta et Lahiri 2000; Chen et Lahiri 2008, 2011) :

$$\text{eqm}\left\{\hat{\theta}_i\left(\hat{\sigma}_{v\text{REML}}^2\right)\right\} = \begin{cases} g_{1i}\left(\hat{\sigma}_{v\text{REML}}^2\right) + g_{2i}\left(\hat{\sigma}_{v\text{REML}}^2\right) + 2g_{3i}\left(\hat{\sigma}_{v\text{REML}}^2\right) & \text{si } \hat{\sigma}_{v\text{REML}}^2 > 0 \\ g_{2i}(0) & \text{si } \hat{\sigma}_{v\text{REML}}^2 = 0. \end{cases} \tag{3.3}$$

Remarque 3.1. Quand $\hat{\sigma}_v^2 = 0$, l'EBLUP est réduit à l'estimateur synthétique. Cependant, lorsque

$$\hat{\sigma}_v^2 = 0, g_{1i}\left(\hat{\sigma}_v^2\right) = 0, \quad g_{2i}\left(\hat{\sigma}_v^2\right) = \mathbf{z}_i'\left[\sum_{i=1}^{m} \mathbf{z}_i \mathbf{z}_i' \Big/ \psi_i\right]^{-1} \mathbf{z}_i,$$

et $g_{3i}\left(\hat{\sigma}_v^2\right) = \bar{V}\left(\hat{\sigma}_v^2\right)\Big/\psi_i > 0$, c'est-à-dire que $\text{eqm}\left\{\hat{\theta}_i\left(\hat{\sigma}_v^2\right)\right\}$ n'est pas une fonction continue de $\hat{\sigma}_v^2$. Nous verrons dans l'étude empirique que, lorsque nous procédons au conditionnement sur $\left\{\hat{\sigma}_v^2 = 0\right\}$, l'estimateur de l'EQM en (3.3) présente un biais négatif significatif, à moins que le rapport signal/bruit sous-jacent σ_v^2/ψ_i ne soit négligeable.

3.2 Méthodes du maximum de vraisemblance ajusté

On obtient les estimateurs de variance par maximum de vraisemblance ajusté en optimisant soit la vraisemblance profilée ajustée (AM pour *adjusted maximum*), soit la vraisemblance résiduelle ajustée (AR pour *adjusted residual*) avec le facteur $h\left(\sigma_v^2\right)$. Comme il est mentionné dans l'introduction, les estimateurs

AM.LL et AR.LL utilisent le facteur d'ajustement $h_{\mathrm{LL}}\left(\sigma_v^2\right) = \sigma_v^2$, tandis que les estimateurs AM.YL et AR.YL utilisent le facteur d'ajustement

$$h_{\mathrm{YL}}\left(\sigma_v^2\right) = \left\{\arctan\left[\sum_{i=1}^{m} \sigma_v^2 \Big/ \left(\sigma_v^2 + \psi_i\right)\right]\right\}^{1/m}.$$

Nous désignons par $\hat{\sigma}^2_{v\mathrm{AM.LL}}$ et $\hat{\sigma}^2_{v\mathrm{AM.YL}}$ les estimateurs de variance obtenus par maximisation des fonctions de vraisemblance profilée ajustée, pour σ_v^2 :

$$L_{\mathrm{AM.}*}\left(\sigma_v^2\right) \propto h\left(\sigma_v^2\right) \cdot \prod_{i=1}^{m}\left(\sigma_v^2 + \psi_i\right)^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}\right\}, \tag{3.4}$$

où $h\left(\sigma_v^2\right) = h_{\mathrm{LL}}\left(\sigma_v^2\right)$ et $h\left(\sigma_v^2\right) = h_{\mathrm{YL}}\left(\sigma_v^2\right)$ pour AM.LL et AM.YL respectivement. La matrice $\mathbf{P}$ est comme en (3.1). Le biais des estimateurs AM jusqu'à l'ordre deux (désigné par $\approx$) correspond à :

$$B\left(\hat{\sigma}^2_{v\mathrm{AM.LL}}\right) \approx \frac{\mathrm{tr}\left\{\mathbf{P} - \mathbf{V}^{-1}\right\} + 2/\sigma_v^2}{\mathrm{tr}\left(\mathbf{V}^{-2}\right)} = O\left(\frac{1}{m}\right) \quad \text{et} \quad B\left(\hat{\sigma}^2_{v\mathrm{AM.YL}}\right) \approx \frac{\mathrm{tr}\left\{\mathbf{P} - \mathbf{V}^{-1}\right\}}{\mathrm{tr}\left(\mathbf{V}^{-2}\right)} = O\left(\frac{1}{m}\right), \tag{3.5}$$

(Li et Lahiri 2011; Yoshimori et Lahiri 2014). On obtient les estimateurs de variance AR.LL et AR.YL, désignés par $\hat{\sigma}^2_{v\mathrm{AR.LL}}$ et $\hat{\sigma}^2_{v\mathrm{AR.YL}}$, en maximisant les fonctions de vraisemblance résiduelle ajustée (AR) pour σ_v^2 :

$$L_{\mathrm{AR.}*}\left(\sigma_v^2\right) \propto h\left(\sigma_v^2\right) \cdot \left|\sum_{i=1}^{m} \mathbf{z}_i \mathbf{z}_i' \Big/ \left(\sigma_v^2 + \psi_i\right)\right|^{-1/2} \prod_{i=1}^{m}\left(\sigma_v^2 + \psi_i\right)^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}\right\} \tag{3.6}$$

où $h\left(\sigma_v^2\right) = h_{\mathrm{LL}}\left(\sigma_v^2\right)$ et $h\left(\sigma_v^2\right) = h_{\mathrm{YL}}\left(\sigma_v^2\right)$ pour AR.LL et AR.YL respectivement, et $\mathbf{P}$ est comme en (3.1). Les biais asymptotiques des estimateurs AR sont donnés respectivement par :

$$B\left(\hat{\sigma}^2_{v\mathrm{AR.LL}}\right) \approx \frac{2/\sigma_v^2}{\mathrm{tr}\left(\mathbf{V}^{-2}\right)} = O\left(\frac{1}{m}\right) \quad \text{et} \quad B\left(\hat{\sigma}^2_{v\mathrm{AR.YL}}\right) = o\left(\frac{1}{m}\right). \tag{3.7}$$

Sous les conditions de régularité données dans la section 2 et sous $\sigma_v^2 > 0$, les deux LL et les deux estimateurs de variance YL existent et sont $\sqrt{m}$ – convergents (Li et Lahiri 2011; Yoshimori et Lahiri 2014). Lahiri et ses coauteurs ont proposé les estimateurs de l'EQM suivants :

$$\mathrm{eqm}\left\{\hat{\theta}_i(\cdot)\right\} = g_{1i}(\cdot) + g_{2i}(\cdot) + 2g_{3i}(\cdot) - \psi_i^2 \cdot B(\cdot)\big/(\cdot + \psi_i)^2 \tag{3.8}$$

où l'argument en $(\cdot)$ ci-dessus est soit $\hat{\sigma}^2_{v\mathrm{AM.LL}}$, $\hat{\sigma}^2_{v\mathrm{AR.LL}}$ ou $\hat{\sigma}^2_{v\mathrm{AM.YL}}$ sous les estimateurs de variance AM.LL, AR.LL et AM.YL respectivement et sous $\hat{\sigma}^2_{v\mathrm{AR.YL}}$:

$$\mathrm{eqm}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\mathrm{AR.YL}}\right)\right\} = g_{1i}\left(\hat{\sigma}^2_{v\mathrm{AR.YL}}\right) + g_{2i}\left(\hat{\sigma}^2_{v\mathrm{AR.YL}}\right) + 2g_{3i}\left(\hat{\sigma}^2_{v\mathrm{AR.YL}}\right). \tag{3.9}$$

Les estimateurs (3.8) et (3.9) sont sans biais jusqu'à l'ordre deux.

Remarque 3.2. Il n'est pas nécessaire que les erreurs d'échantillonnage aient une distribution normale pour assurer la convergence et la normalité asymptotique des estimateurs LL et YL (voir, par exemple, Rubin-Bleuer et coll. 2011).

3.3 Algorithmes d'optimisation

Vu les données, la fonction de vraisemblance REML peut atteindre sa valeur maximale à $\sigma_v^2 = 0,$ même lorsque la valeur sous-jacente réelle de $\boldsymbol{\sigma}_v^2$ est positive. Par ailleurs, les vraisemblances LL et YL atteignent toujours leur valeur maximale à $\sigma_v^2 > 0.$ Pourtant, la vraisemblance résiduelle YL est très proche de la vraisemblance REML. Des études empiriques montrent que l'algorithme de score sous AR.YL donne $\hat{\sigma}_{v\text{AR.YL}}^2 = 0$ dans une proportion presque aussi importante que sous REML pour les ensembles de données suivant un modèle de Fay-Herriot avec une variance sous-jacente réelle faible mais non nulle. Cela se produit lorsque l'algorithme de score passe à côté de la valeur maximale positive de la vraisemblance AR.YL et produit une valeur nulle (pour plus de détails, voir l'annexe B). Pour éviter ce problème, nous utilisons une méthode de grille pour l'optimisation (Estevao 2014). Dans notre étude, nous établissons la limite supérieure de l'intervalle de recherche à $1\,000 \times \sigma_v^2,$ car nous connaissons σ_v^2 a priori. Pour les applications avec des données réelles, nous suggérons d'obtenir une estimation initiale $\hat{\sigma}_{v\text{AM.LL}}^2$ en utilisant la méthode de score et de fixer la limite supérieure à $1\,000 \times \hat{\sigma}_{v\text{AM.LL}}^2,$ puis d'augmenter graduellement la limite jusqu'à ce que l'estimation de variance se situe dans l'intervalle de recherche.

4 Estimateur de variance MIX

4.1 Estimation de variance

L'estimateur de variance MIX est une procédure qui commence par calculer l'estimation de variance REML et qui la remplace par l'estimation de variance par le maximum de vraisemblance ajusté seulement lorsque l'estimation REML est négative. L'estimateur de variance MIX est toujours positif et sans biais jusqu'à un terme d'ordre $o(1/m).$ L'estimateur de variance MIX de σ_v^2 est défini par :

$$\hat{\sigma}_{v\text{MIX}}^2 = \begin{cases} \hat{\sigma}_{v\text{REML}}^2 & \text{si } \hat{\sigma}_{v\text{REML}}^2 > 0 \\ \hat{\sigma}_{v\text{adj}}^2 & \text{si } \hat{\sigma}_{v\text{REML}}^2 = 0, \end{cases} \tag{4.1}$$

où $\hat{\sigma}_{v\text{adj}}^2$ est un des estimateurs de la vraisemblance ajustée définis dans la section 3.

Remarque 4.1. L'estimateur de variance MIX présente automatiquement certaines des propriétés communes partagées par l'estimateur REML et l'estimateur de variance par le maximum de vraisemblance ajusté. Par exemple, il est pair et invariant de translation. Ainsi, sous l'hypothèse de normalité des erreurs d'échantillonnage, l'approximation d'ordre deux (2.6) de l'EQM de l'EBLUP est également valide : le théorème 4.1 ci-après montre que l'EQM de l'EBLUP sous l'estimateur de variance MIX hérite des mêmes propriétés asymptotiques que l'EQM sous l'estimateur de variance REML.

Théorème 4.1. Sous les conditions de régularité 1 à 3 données dans la section 2 et l'hypothèse que $\sigma_v^2 > 0$, l'EQM de l'EBLUP sous l'estimateur de variance MIX est égale à l'EQM sous l'estimateur de variance REML jusqu'à l'ordre deux. Le théorème découle du fait que la variance asymptotique de $\hat{\sigma}^2_{v\text{MIX}}$ coïncide avec la variance asymptotique de $\hat{\sigma}^2_{v\text{REML}}$ (pour obtenir plus de détails, voir l'annexe A).

Théorème 4.2. Sous les conditions du théorème 4.1, le $\text{Bias}\left(\hat{\sigma}^2_{v\text{MIX}}\right) = o(1/m)$. La preuve est présentée à l'annexe A.

4.2 Estimation de l'EQM

Le fait que l'estimateur MIX $\hat{\sigma}^2_{v\text{MIX}}$, est sans biais jusqu'à l'ordre deux est essentiel pour démontrer que l'estimateur d'EQM que nous proposons est lui aussi sans biais jusqu'à l'ordre deux.

Corollaire 4.2. L'estimateur de l'EQM de l'EBLUP sous $\hat{\sigma}^2_{v\text{MIX}}$ donné par :

$$\text{eqm}\left[\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{MIX}}\right)\right] = g_{1i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) + g_{2i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) + 2g_{3i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) \tag{4.2}$$

est sans biais jusqu'à l'ordre deux. Étant donné que $\hat{\sigma}^2_{v\text{MIX}}$ est sans biais jusqu'à l'ordre deux, le résultat suit le modèle de Datta et Lahiri (2000).

4.3 Autres estimateurs de l'EQM

Dans l'équation qui suit, l'estimateur de variance MIX est la combinaison des estimateurs REML et AM.LL.

Rubin-Bleuer et You (2012) ont suggéré un autre estimateur de l'EQM, qui est lui aussi sans biais jusqu'à l'ordre deux. Il s'agit d'un estimateur de l'EQM « fractionné » qui prend la forme :

$$\text{eqm}*\left[\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{MIX}}\right)\right] = \begin{cases} g_{1i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) + g_{2i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) + 2g_{3i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) & \text{si } \hat{\sigma}^2_{v\text{MIX}} = \hat{\sigma}^2_{v\text{REML}}, \\ g_{1i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) + g_{2i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) & \\ \quad + 2g_{3i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) - \left(1-\hat{\gamma}_{i\text{MIX}}\right)^2 \cdot \text{Biais}\left(\hat{\sigma}^2_{v\text{MIX}}\right) & \text{si } \hat{\sigma}^2_{v\text{MIX}} = \hat{\sigma}^2_{v\text{AM.LL}}. \end{cases} \tag{4.3}$$

L'estimateur $\text{eqm}*$ présente un biais relatif moyen (BRM) plus faible que l'estimateur de l'EQM donné en (4.2). Le BRM est plus faible parce que l'EQM est surestimée lorsque le REML est positif et sous-estimée lorsque le REML est nul. L'estimateur $\text{eqm}*$ est généralement satisfaisant, mais il peut prendre des valeurs négatives pour un ensemble particulier de données.

Molina et coll. (2015) ont proposé deux estimateurs de l'EQM pour l'EBLUP sous la méthode MIX. Dans l'équation qui suit, le test préliminaire proposé de l'hypothèse de variance nulle est représenté par le sigle TP, et les estimateurs sont :

$$\text{eqm}_0\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{MIX}}\right)\right\} = \begin{cases} \text{eqm}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{REML}}\right)\right\} & \text{si } \hat{\sigma}^2_{v\text{REML}} > 0 \\ g_{2i}(0) & \text{si } \hat{\sigma}^2_{v\text{REML}} = 0 \end{cases} \tag{4.4}$$

et

$$\text{eqm}_{\text{TP}}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{MIX}}\right)\right\} = \begin{cases} \text{eqm}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{REML}}\right)\right\} & \text{si } \hat{\sigma}^2_{v\text{REML}} > 0 \text{ et le TP est rejeté} \\ g_{2i}(0) & \text{si } \hat{\sigma}^2_{v\text{REML}} = 0 \text{ ou le TP n'est pas rejeté.} \end{cases} \tag{4.5}$$

La justification de eqm_0 et eqm_{TP} se fonde sur l'EQM du BLUP où $\sigma_v^2 = 0$. Molina et coll. (2015) ont montré dans une étude empirique que les estimateurs de l'EQM proposés donnaient de bons résultats en moyenne lorsque σ_v^2 et le nombre de domaines m étaient faibles.

Remarque 4.2. Les estimateurs eqm_0 et eqm_{TP} sont eux aussi sans biais jusqu'à l'ordre deux (l'annexe contient une preuve abrégée de cette propriété). Notre argument contre $\text{eqm}\left\{\hat{\theta}_i\left(\hat{\sigma}_v^2\right)\right\}$ (en 3.3) s'applique aussi à eqm_0 et à eqm_{TP} : pour un nombre modéré de domaines, le pourcentage de populations où $\hat{\sigma}^2_{v\text{REML}} = 0$ peut être significatif, même si σ_v^2 / ψ_i n'est pas négligeable. Dans ce cas-ci, l'estimateur de l'EQM de l'EBLUP doit aussi tenir compte de la variation due à l'estimation de variance ou à la sous-estimation du risque.

5 Conditions de simulation et mesures de performance

5.1 Conditions de simulation

Nous avons réalisé une simulation Monte Carlo fondée sur un modèle, en suivant l'exemple de Rubin-Bleuer et You (2012), afin d'examiner la performance en échantillon fini des différentes méthodes. Les estimations « directes » $(y_1, \ldots, y_m)$ où $m = 15, m = 45$ et $m = 100,$ sont générées à partir du modèle de Fay-Herriot en (2.3) où $\boldsymbol{\beta}' = (5, 4, 3, 2, 1)$ et les covariables $\mathbf{z}_i' = \left(1, z_{i2}, \ldots, z_{ip}\right),$ générées une fois à partir de distributions normales $z_{ik} \sim k + N(1,1),$ $k = 2, \ldots, 5,$ $i = 1, \ldots, m,$ et maintenues fixes sur les populations répétées. Les effets aléatoires normaux indépendants de domaine v_i sont générés avec la variance $\sigma_v^2 = 1.$ Des erreurs d'échantillonnage indépendantes $e_i,$ sont générées avec des variances d'échantillonnage $\psi_i \triangleq 50/n_i,$ où n_i est la taille de l'échantillon pour le domaine $i, i = 1, \ldots, m.$ Il y a cinq groupes de variances d'échantillonnage déterminés par $n_i = 3,$ 5, 7, 10 ou 15, où les rapports signal/bruit $\sigma_v^2 / \psi_i = 0{,}06;$ 0,1; 0,14; 0,2 et 0,3, respectivement. Ainsi, lorsque $m = 100,$ il y a 20 domaines par rapport signal/bruit. Nous avons d'abord généré 50 000 ensembles d'estimateurs directs pour chaque cas, puis calculé l'EBLUP et l'EQM Monte Carlo réelle de l'EBLUP à l'aide des estimateurs de variance REML, AM.LL, MIX, AM.YL et AR.YL. Nous n'avons pas étudié l'estimateur AR.LL en raison de sa performance médiocre mentionnée par Li et Lahiri (2011). Nous avons ensuite généré 10 000 ensembles d'estimateurs directs indépendamment des 50 000 premiers. Pour chaque ensemble généré, nous avons calculé les cinq estimateurs de variance. Pour l'estimateur de variance MIX, nous avons examiné trois des quatre estimateurs de l'EQM par linéarisation qui font l'objet d'une discussion dans la section 4. Comme il arrive souvent que les estimateurs de l'EQM par linéarisation n'estiment pas le biais de façon exacte, nous avons également jeté un coup d'œil à l'estimateur bootstrap paramétrique de l'EQM

(BP EQM) corrigé pour le biais en utilisant la méthode de Pfeffermann et Glickman (2004) ainsi que l'estimateur BP naïf de l'EQM avec 500 répétitions chacune (voir l'annexe B pour la construction des poids bootstrap). Les mesures de performance Monte Carlo sont définies ci-après.

1. L'EQM de l'EBLUP, $\overline{\text{EQM}}_\ell\left(\hat{\theta}_i\right)$, par groupe de variances d'échantillonnage :

$$\text{EQM}\left(\hat{\theta}_i\right) = \frac{1}{50\ 000}\sum_{r=1}^{50\ 000}\left(\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right)^2, \quad \overline{\text{EQM}}_\ell\left(\hat{\theta}\right) = \frac{5}{m}\sum_{i\in\{j:\psi_j=50/n_\ell\}}\text{EQM}\left(\hat{\theta}_i\right), \quad \ell = 1,...,5.$$

2. $E\left(\hat{\sigma}_v^2\right) = \sum_{r=1}^{10\ 000}\hat{\sigma}_v^{2(r)}\big/10\ 000$, $V\left(\hat{\sigma}_v^2\right) = \sum_{r=1}^{10\ 000}\left(\hat{\sigma}_v^{2(r)} - E\left(\hat{\sigma}_v^2\right)\right)^2\big/10\ 000$, où $\hat{\sigma}_v^{2(r)}$ est la valeur de $\hat{\sigma}_v^2$ pour la r^{e} simulation $(r = 1,\ldots,10\ 000)$.

3. Le biais relatif moyen (BRM) de l'EQM par groupe de variances d'échantillonnage :

$$\text{BRM}_\ell\,(\text{eqm}) = \frac{5}{m}\sum_{i\in\{j:\psi_j=50/n_\ell\}}\text{BR}\left(\text{eqm}\left(\hat{\theta}_i\right)\right), \quad \ell = 1,\ldots 5,$$

où $\text{BR}\left(\text{eqm}\left(\hat{\theta}_i\right)\right) = \left[\sum_{r=1}^{10\ 000}\text{eqm}\left(\hat{\theta}_i^{(r)}\right)\big/10\ 000 - \text{EQM}\left(\hat{\theta}_i\right)\right]\Big/\text{EQM}\left(\hat{\theta}_i\right)$.

4. La racine de l'EQM relative des estimateurs de l'EQM par groupe de variances d'échantillonnage :

$$\text{REQMR}_\ell\,(\text{eqm}) = \left(\frac{5}{m}\sum_{i\in\{j:\psi_j=50/n_\ell\}}\frac{\sum_{r=1}^{10\ 000}\left(\text{eqm}\left(\hat{\theta}_i^{(r)}\right) - \text{EQM}\left(\hat{\theta}_i\right)\right)^2\big/10\ 000}{\text{EQM}\left(\hat{\theta}_i\right)}\right)^{1/2}.$$

Nous examinons également le biais des estimateurs conditionnels de l'EQM étant donné que $\left\{\hat{\sigma}^2_{v\text{REML}} = 0\right\}$, car ce sont les populations pour lesquelles les estimateurs positifs ont été élaborés.

5. Le biais relatif moyen des estimateurs conditionnels de l'EQM :

$$\text{BRM}_{\mathbb{C}} = \frac{5}{m}\sum_{i\in\ell}\text{E}\left[\text{eqm}\left(\hat{\theta}_i\right)\middle|\,\hat{\sigma}^2_{v\text{REML}} = 0\right]\Big/\text{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\middle|\,\hat{\sigma}^2_{v\text{REML}} = 0\right] - 1.$$

6 Résultats de la simulation et analyse

6.1 Distribution Monte Carlo des estimateurs de variance

Le tableau 6.1 montre que l'estimateur de variance REML présente le biais le plus faible $\left(\sigma_v^2 = 1\right)$ et la variance la plus élevée. L'efficacité plus faible du REML pourrait être attribuable au fait qu'il ne s'agit pas d'une fonction lisse des données causée par sa définition fractionnée (3.1). L'estimateur MIX hérite d'une partie de cette efficacité faible. Les autres estimateurs de variance ont une variabilité plus faible et un biais

positif plus élevé, mais l'espérance conditionnelle des estimateurs AM.YL et AR.YL étant donné que $\hat{\sigma}^2_{vREML} = 0$ est proche de zéro. Le biais inconditionnel de l'estimateur AM.LL est plus élevé que celui du MIX. Selon la définition de l'estimateur MIX, les biais conditionnels des estimateurs MIX et AM.LL coïncident. En outre, le MIX converge plus rapidement que les autres estimateurs. Par exemple, étant donné la distribution des probabilités sur les 10 000 estimations de la variance où $m = 45,$ nous avons calculé la probabilité que les estimations se situent dans un intervalle contenant $\sigma^2_v = 1.$ La probabilité que les estimations se situent entre 0,6 et 1,4 est de 0,47 pour le MIX et de 0,16 pour l'AM.YL. Par ailleurs, la probabilité que les estimations soient inférieures à 0,2 est de 0,05 pour le MIX et de 0,53 pour l'AM.YL.

Tableau 6.1
Espérance, variance et espérance et variance conditionnelle de $\hat{\sigma}^2_v$

Méthode	m	$E(\hat{\sigma}^2_v)$	$V(\hat{\sigma}^2_v)$	%REML = 0	$E(\hat{\sigma}^2_v / \text{REML} = 0)$	$V(\hat{\sigma}^2_v / \text{REML} = 0)$
REML	15	1,48	3,38	43 %	N/A	N/A
	45	1,21	1,67	29 %	N/A	N/A
	100	1,07	0,81	16 %	N/A	N/A
AM.LL	15	2,80	1,37	43 %	1,80	0,11
	45	1,88	1,01	29 %	0,94	0,03
	100	1,49	0,51	16 %	0,63	0,01
MIX	15	2,28	1,87	43 %	1,80	0,11
	45	1,48	1,31	29 %	0,94	0,03
	100	1,17	0,66	16 %	0,63	0,01
AR.YL	15	1,66	2,99	43 %	0,27	0,01
	45	1,24	1,72	29 %	0,06	0,00
	100	1,08	0,80	16 %	0,02	0,00
AM.YL	15	0,52	0,84	43 %	0,10	0,00
	45	0,65	0,85	29 %	0,03	0,00
	100	0,76	0,59	16 %	0,01	0,00

6.2 EQM réelle de l'EBLUP, biais relatif moyen et racine de l'EQM relative moyenne des estimateurs de l'EQM

Tous les estimateurs de variance sont convergents et asymptotiquement normaux, la variance convergeant au même rythme. Leurs biais sont différents : ceux des estimateurs REML, AR.YL et MIX sont de l'ordre $o(1/m),$ tandis que ceux des estimateurs AM.LL et AM.YL sont de l'ordre $O(1/m).$ Le biais inhérent aux trois dernières méthodes ont un impact sur l'estimation de l'EQM de l'EBLUP, même pour un nombre modéré de domaines.

Pour $m = 100,$ les tableaux 6.2a et 6.2b montrent que l'EQM de l'EBLUP diminue à mesure que σ^2_v / ψ_i augmente, et que cette relation se maintient quel que soit le nombre de domaines. Nous observons que l'EQM de $\hat{\theta}_i$ sous les estimateurs de variance REML et MIX est légèrement plus élevée que le reste des EQM en raison de la plus grande variabilité inhérente à ces estimateurs de variance. Le tableau 6.2a présente les résultats pour l'estimateur de l'EQM par linéarisation de Taylor et les deux estimateurs paramétriques de l'EQM sous les estimateurs de variance REML, AM.LL, AR.YL et AM.YL. Le tableau 6.2b présente les résultats pour les estimateurs suivants de l'EQM sous l'estimation de variance MIX : RB_Y1 défini en (4.3), RB_Y2 défini en (4.2), M_et_coll défini en (4.5), BP EQM et BP EQM naïf. Parmi les estimateurs de

l'EQM de Taylor, RB_Y1 et M_et_coll sous la méthode MIX présentent le biais le plus faible. Parmi les estimateurs bootstrap de l'EQM, BP sous MIX et BP naïf sous AR.YL présentent le biais le plus faible. Quant à la racine de l'erreur quadratique moyenne relative (REQMR) des estimateurs de l'EQM, elle diminue à mesure que σ_v^2 / ψ_i augmente. Les différences entre l'estimateur RB_Y2 de l'EQM sous le MIX et l'estimateur de l'EQM de Taylor sous l'AM.YL semblent faibles mais consistantes. Alors que le RB_Y1, le M_et_coll et les estimateurs naïfs de l'EQM sous la méthode MIX présentent un BRM plus faible que le RB_Y2 sous la même méthode, et que le BRM de l'estimateur de Taylor et de l'estimateur BP naïf sous la méthode AR.YL est plus faible que celui du RB_Y2 sous la méthode MIX, c'est le contraire pour la REQMR. Cela s'explique en partie par le biais conditionnel négatif extrême de ces estimateurs de l'EQM (c'est-à-dire les estimateurs RB_Y1 et M_et_coll sous la méthode MIX et les estimateurs de Taylor et BP naïf sous la méthode AR.YL), comme le montre le tableau 6.3. Même pour $m = 100$, une proportion relativement élevée (16 %) des populations donnent $\hat{\sigma}^2_{v\text{REML}} = 0$ et, dans ces populations, les estimations obtenues au moyen de la plupart des méthodes d'estimation de variance et la plupart des estimateurs de l'EQM sont les plus inférieures à la valeur réelle. C'est-à-dire que, pour ces estimateurs de l'EQM, les estimateurs conditionnels ne donnent pas de bons résultats. L'estimateur BP EQM semble corriger pour le biais de façon satisfaisante, mais il est plus variable que le BP EQM naïf. Lorsque nous incluons le BRM, la REQMR et le $\text{BRM}_{\mathbb{C}}$ dans l'évaluation, c'est le RB_Y2 sous la méthode MIX, suivi de près par l'estimateur BP naïf, qui l'emporte sous la méthode MIX qui semble donner les meilleurs résultats. Nous pourrions donc conclure à la supériorité des estimateurs RB_Y2 et naïf sous la méthode MIX pour $m = 100$, ce qui est un nombre modéré de domaines pour les données de ce genre.

Tableau 6.2a
EQM, BRM et REQMR (pourcentage) des estimateurs de l'EQM, $m = 100$

			Estimateur de l'EQM de Taylor		Estimateur BP		Estimateur BP naïf	
Méthode	σ_v^2 / ψ_i	$\overline{\text{EQM}}$	**BRM**	**REQMR**	**BRM**	**REQMR**	**BRM**	**REQMR**
REML	0,06	135,4	5,1	71,1	-4,4	80,7	1,6	69,9
	0,1	132,1	5,3	64,7	-4,7	74,0	-0,2	63,0
	0,14	119,5	6,0	61,9	-5,5	71,3	-1,8	59,9
	0,2	119,2	6,5	53,6	-5,8	62,4	-3,4	51,7
	0,3	106,6	8,2	46,7	-6,8	55,0	-5,6	44,8
AM.LL	0,06	134,9	6,1	75,4	8,2	66,9	31,3	63,8
	0,1	131,2	6,8	68,1	7,8	59,5	27,5	55,7
	0,14	118,3	8,1	64,6	7,8	55,6	26,5	51,2
	0,2	117,6	8,4	55,4	6,5	46,7	21,6	42,1
	0,3	104,5	10,2	46,7	5,5	38,8	18,2	34,0
AR.YL	0,06	135,4	6,6	69,3	-4,3	80,2	2,1	69,4
	0,1	132,0	7,4	61,9	-4,5	73,4	0,3	62,5
	0,14	119,4	9,0	58,0	-5,3	70,6	-1,2	59,3
	0,2	119,0	10,6	48,2	-5,6	61,8	-2,9	51,1
	0,3	106,4	14,7	38,5	-6,6	54,3	-5,1	44,1
AM.YL	0,06	134,7	10,0	63,2	-12,3	81,0	-19,6	65,9
	0,1	131,3	12,0	56,6	-12,5	75,2	-19,7	61,2
	0,14	118,8	15,0	53,1	-13,7	73,3	-21,4	59,8
	0,2	118,6	18,1	44,8	-13,4	65,2	-20,7	53,5
	0,3	106,4	25,2	38,4	-14,4	58,8	-21,7	48,6

Tableau 6.2b
EQM, BRM et REQMR (pourcentage) des estimateurs de l'EQM, $m = 100$

			RB_Y1		RB_Y2		M_et_coll		Estimateur BP		Estimateur BP naïf	
	σ_v^2/ψ_i	$\overline{\text{EQM}}$	BRM	REQMR	BRM	REQMR	BRM	REQMR	BRM	REQMR	BRM	REQMR
MIX	0,06	135,4	2,7	75,7	13,6	63,0	5,2	71,1	-3,0	75,3	8,8	62,4
	0,1	132,1	3,6	68,3	14,9	56,1	5,3	64,7	-3,2	68,3	6,6	55,4
	0,14	119,5	4,9	64,7	16,0	52,4	6,0	61,9	-3,9	65,1	5,3	51,8
	0,2	119,1	6,3	55,2	16,7	43,8	6,5	53,6	-4,4	56,3	2,9	43,7
	0,3	106,5	9,4	46,2	19,9	36,0	8,3	46,7	-5,4	48,6	0,6	36,7

Tableau 6.3
$\text{EQM}_{\mathbf{C}}\left(\text{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2 \mid \hat{\sigma}^2_{v\text{REML}} = 0\right]\right)$ et $\text{BRM}_{\mathbf{C}}$ (pourcentage), $m = 100$

Méthode	σ_v^2/ψ_i	$\overline{\text{EQM}}_{\mathbf{C}}$	Estimateur de l'EQM de Taylor	Estimateur BP	Estimateur BP naïf
REML	0,06	135,6	-76,5	-98,6	-74,8
	0,1	133,0	-74,5	-94,4	-71,8
	0,14	121,5	-78,6	-98,0	-74,9
	0,2	120,4	-73,1	-89,8	-68,6
	0,3	108,0	-73,6	-88,2	-67,3
AM.LL	0,06	135,0	-92,0	-67,6	-26,1
	0,1	132,2	-85,2	-62,0	-24,6
	0,14	120,2	-85,4	-62,3	-25,4
	0,2	118,8	-74,4	-54,1	-22,1
	0,3	105,9	-65,9	-49,7	-20,6
AR.YL	0,06	135,5	-68,6	-96,9	-73,0
	0,1	132,9	-62,4	-92,6	-70,0
	0,14	121,4	-61,1	-96,1	-73,0
	0,2	120,2	-48,9	-87,9	-66,7
	0,3	107,8	-34,5	-86,1	-65,4
AM.YL	0,06	134,9	-45,9	-88,6	-74,7
	0,1	132,1	-39,4	-85,4	-72,2
	0,14	120,4	-36,0	-89,3	-75,7
	0,2	119,6	-23,6	-82,3	-69,7
	0,3	107,6	-6,5	-81,7	-69,3

			RB_Y1	RB_Y2	M_et_coll	Estimateur BP	Estimateur BP naïf
MIX	0,06	135,0	-92,0	-22,0	-76,4	-46,0	-27,0
	0,1	132,2	-85,2	-17,7	-74,3	-42,7	-25,9
	0,14	120,2	-85,4	-15,0	-78,3	-43,3	-27,0
	0,2	118,8	-74,4	-7,6	-72,8	-37,6	-23,9
	0,3	105,9	-65,9	1,5	-73,1	-34,6	-22,6

Les tableaux 6.4a et 6.4b ci-dessous présentent les résultats pour $m = 45$ avec 9 domaines par σ_v^2/ψ_i. L'estimateur AM.YL donne des EQM plus petites que le MIX, les différences ne dépassant pas 2 %. Le biais des estimateurs de variance augmente à mesure que le nombre de domaines diminue, ce qui a un impact sur les estimateurs de l'EQM. En effet, les BRM de tous les estimateurs de l'EQM ont augmenté. En particulier, le BRM des estimateurs de l'EQM de Taylor sous l'estimation de variance YL et LL et le BRM de l'estimateur RB_Y2 ont augmenté de 100 % par rapport au BRM avec 100 domaines. En ce qui concerne la REQMR, l'estimateur de l'EQM de Taylor sous AM.YL a une REQMR légèrement inférieure à celle du RB_Y2 sous la méthode MIX pour une σ_v^2/ψ_i très faible. En général, la variabilité (en termes de REQMR) du RB_Y2 est plus faible que celle de l'estimateur de Taylor sous LL et YL et des estimateurs RB_Y1 et

M_et_coll. Cela pourrait être attribuable en partie à la sous-estimation des EQM pour les populations avec des estimations REML nulles, dont le pourcentage tourne autour de 30 % lorsque $m = 45$. Le tableau 6.5 est plus instructif à cet égard : étant donné $\hat{\sigma}^2_{vREML} = 0$, RB_Y1 et M_et_coll conduisent à une sous-estimation importante.

Tableau 6.4a
EQM, BRM et REQMR (pourcentage) des estimateurs de l'EQM, m = 45 domaines

			Estimateur de l'EQM de Taylor		Estimateur BP		Estimateur BP naïf	
Méthode	σ_v^2/ψ_i	$\overline{EQM}$	BRM	REQMR	BRM	REQMR	BRM	REQMR
REML	0,06	171,4	11,8	94,7	-4,7	107,0	6,2	89,2
	0,1	174,1	11,9	83,9	-5,3	93,8	3,0	76,2
	0,14	171,3	12,6	74,5	-5,4	81,9	1,1	65,3
	0,2	166,6	13,9	63,4	-5,8	66,7	-1,2	52,0
	0,3	128,9	20,1	63,0	-7,0	61,4	-3,1	46,7
AM.LL	0,06	171,1	15,5	100,0	16,0	84,9	43,5	83,3
	0,1	173,4	16,8	87,0	14,4	71,1	36,7	68,5
	0,14	170,4	17,7	75,7	12,6	59,7	30,7	56,7
	0,2	165,3	18,2	61,7	9,9	46,2	23,5	43,2
	0,3	127,5	25,6	55,0	10,0	39,7	22,6	36,6
AR.YL	0,06	171,1	17,2	89,9	-3,7	105,0	8,0	87,6
	0,1	173,6	19,6	76,9	-4,3	91,8	4,8	74,6
	0,14	170,8	22,6	65,8	-4,4	79,9	2,7	63,7
	0,2	166,0	27,3	53,7	-4,8	64,8	0,3	50,5
	0,3	128,3	43,8	54,8	-5,7	59,3	-1,3	45,0
AM.YL	0,06	167,5	30,2	78,4	-18,0	97,3	-23,8	73,3
	0,1	169,6	36,5	72,2	-18,0	87,7	-23,6	66,7
	0,14	167,0	42,7	69,3	-17,2	78,0	-22,3	59,7
	0,2	162,8	52,1	70,8	-15,8	65,4	-20,3	50,6
	0,3	126,0	81,3	91,1	-18,0	62,3	-22,9	48,4

Tableau 6.4b
EQM, BRM et REQMR (pourcentage) des estimateurs de l'EQM, m = 45 domaines

			RB_Y1		RB_Y2		M_et_coll		Estimateur BP		Estimateur BP naïf	
	σ_v^2/ψ_i	$\overline{EQM}$	BRM	REQMR	BRM	REQMR	BRM	REQMR	BRM	REQMR	BRM	REQMR
MIX	0,06	171,4	9,8	99,4	31,9	84,0	11,8	94,7	3,5	93,8	21,9	78,5
	0,1	174,0	12,1	86,2	33,2	73,1	11,9	83,9	2,6	80,4	17,5	65,1
	0,14	171,2	14,5	74,9	34,4	64,6	12,6	74,5	2,0	68,7	14,0	54,4
	0,2	166,5	17,7	61,7	36,0	55,8	13,9	63,4	0,7	54,5	9,8	41,8
	0,3	128,9	28,8	57,6	48,8	58,2	20,2	63,1	0,3	48,6	8,7	35,9

Compte tenu du BRM, de la REQMR et du $BRM_{\mathbb{C}}$ des estimateurs de l'EQM, c'est l'estimateur BP naïf de l'EQM sous la méthode MIX qui donne les meilleurs résultats pour les σ_v^2/ψ_i plus importantes. Le tableau 6.6 présente la moyenne des mesures de performance, calculée sur les cinq groupes de variances

d'échantillonnage, pour les trois estimateurs de l'EQM de Taylor sous la méthode MIX avec des données du modèle décrit en 5.1, mais avec trois valeurs différentes de σ_v^2. L'estimateur RB_Y2 donne de meilleurs résultats quand $\sigma_v^2 = 1$, mais lorsque σ_v^2 diminue, c'est l'estimateur de l'EQM de M_et_coll qui prend le dessus, précisément parce qu'il a été construit en partant du principe que σ_v^2 est d'environ zéro.

Tableau 6.5
EQM_C et BRM_C (pourcentage). m = 45 domaines

Méthode	σ_v^2/ψ_i	$\overline{EQM}_C$	Estimateur de l'EQM de Taylor			Estimateur BP	Estimateur BP naïf
REML	0,06	170,2		-64,3		-89,7	-60,7
	0,1	173,0		-62,4		-83,7	-57,1
	0,14	170,2		-58,1		-75,5	-51,8
	0,2	165,8		-51,9		-65,1	-44,8
	0,3	131,1		-59,0		-70,5	-49,2
AM.LL	0,06	170,0		-71,5		-49,0	-3,1
	0,1	172,3		-61,5		-42,1	-2,3
	0,14	169,1		-51,1		-35,7	-2,1
	0,2	164,7		-38,3		-28,3	-1,6
	0,3	129,9		-28,8		-29,1	-3,7
AR.YL	0,06	169,9		-48,3		-86,2	-56,7
	0,1	172,6		-38,0		-80,2	-53,2
	0,14	169,7		-25,9		-72,2	-48,2
	0,2	165,3		-7,4		-61,9	-41,5
	0,3	130,5		19,3		-66,8	-45,5
AM.YL	0,06	166,6		-8,2		-73,5	-60,7
	0,1	168,8		3,8		-70,1	-58,1
	0,14	166,1		16,1		-64,1	-53,3
	0,2	162,2		35,9		-56,1	-46,8
	0,3	128,1		72,8		-62,5	-52,5
			RB_Y1	**RB_Y2**	**M_et_coll**		
MIX	0,06	170,0	-71,5	6,2	-64,3	-28,1	-4,0
	0,1	172,3	-61,5	13,2	-62,3	-23,8	-3,5
	0,14	169,1	-51,1	18,9	-57,8	-20,0	-3,3
	0,2	164,7	-38,3	26,8	-51,6	-15,7	-2,9
	0,3	129,9	-28,8	40,4	-58,7	-16,7	-5,1

Tableau 6.6
EQM, BRM, BRM_C et REQMR (pourcentage), 45 domaines

			RB_Y1			RB_Y2			M_et_coll		
%REML = 0	σ_v^2	$\overline{EQM}$	**BRM**	BRM_C	**REQMR**	**BRM**	BRM_C	**REQMR**	**BRM**	BRM_C	**REQMR**
29	1	108	16	-50	75	36	21	66	14	-59	75
48	0,2	99	48	-36	101	113	88	114	47	-38	94
51	0,1	91	58	-33	108	137	107	127	58	-32	100

Les tableaux 6.7a et 6.7b ci-dessous montrent les résultats pour $m = 15$ domaines avec 3 domaines par σ_v^2/ψ_i. Les EQM ne varient pas de plus de 5 %, quelle que soit la méthode d'estimation de variance utilisée.

Il n'existe pas de relation monotone entre le BRM ou la REQMR et σ_v^2/ψ_i, ce qui pourrait indiquer que l'approximation d'ordre deux pour estimer l'EQM est médiocre, quelle que soit la méthode d'estimation de variance utilisée. Les BRM des estimateurs de l'EQM de Taylor sous les méthodes d'estimation de

variance de LL et YL sont trop élevés, et il en va de même pour la REQMR. L'estimateur RB_Y2 sous la méthode MIX ne s'en sort pas très bien non plus. La raison de ce résultat est claire : le pourcentage élevé d'estimations REML nulles (43 %) suggère que le MIX coïncide avec l'AM.LL pour les populations REML nulles. Ainsi, le MIX a un biais positif pour $m = 15,$ et le RB_Y2 ne tient pas compte de ce biais. Le RB_Y1 tient compte du biais dans le MIX, mais l'estimateur du biais n'est pas très précis pour $m = 15.$ L'estimateur de l'EQM de M_et_coll coïncide presque avec le BRM et la REQMR de l'estimateur de l'EQM de Taylor sous l'estimation de variance REML, car ils sont égaux par définition lorsque $\hat{\sigma}^2_{v\text{REML}} = 0.$ Le $\text{BRM}_{\mathbb{C}}$ des trois estimateurs de l'EQM de Taylor sous la méthode MIX laisse à désirer. Compte tenu de toutes les mesures de performance, les estimateurs bootstrap de l'EQM donnent de meilleurs résultats que les estimateurs de l'EQM de Taylor. Pour $m = 15$ domaines avec 3 domaines par σ_v^2 / ψ_i, l'estimateur BP sous la méthode MIX donne les meilleurs résultats, l'estimateur BP naïf sous AR.YL et AM.YL venant en deuxième place.

Tableau 6.7a
EQM, BRM et REQMR (pourcentage) des estimateurs de l'EQM, $m = 15$ domaines

			Estimateur de l'EQM de Taylor		Estimateur BP		Estimateur BP naïf	
Méthode	σ_v^2 / ψ_i	$\overline{\text{EQM}}$	BRM	REQMR	BRM	REQMR	BRM	REQMR
REML	0,06	584,8	12,6	87,9	1,2	85,9	6,9	64,5
	0,1	376,7	26,5	106,3	2,3	85,6	9,6	62,8
	0,14	352,5	25,2	90,1	0,7	54,1	4,3	39,3
	0,2	209,4	43,0	123,0	0,4	74,0	6,3	51,1
	0,3	198,7	50,6	124,7	-1,0	46,3	2,6	31,5
AM.LL	0,06	589,3	24,1	89,3	13,7	61,2	24,1	65,8
	0,1	380,7	48,3	107,1	19,4	58,6	32,5	62,9
	0,14	355,7	40,2	88,6	10,0	36,2	16,8	38,1
	0,2	212,5	76,3	117,9	17,8	45,1	28,7	47,3
	0,3	200,7	76,5	105,1	10,7	26,9	17,2	27,6
AR.YL	0,06	583,3	23,8	83,3	3,2	79,5	3,2	61,6
	0,1	375,1	53,3	106,7	5,4	78,6	5,4	59,7
	0,14	351,3	53,3	102,7	2,4	49,4	2,4	37,1
	0,2	207,7	107,3	153,1	4,1	66,2	4,1	47,2
	0,3	197,5	142,0	199,4	1,9	41,1	1,9	28,9
AM.YL	0,06	571,4	41,6	103,5	-8,0	61,2	-9,2	43,3
	0,1	363,3	95,0	161,4	-11,3	62,9	-13,2	44,1
	0,14	342,0	97,2	179,7	-6,7	40,4	-7,8	29,3
	0,2	197,0	198,4	274,6	-14,5	58,2	-16,7	41,7
	0,3	191,4	270,2	362,4	-11,5	38,4	-13,1	28,7

Tableau 6.7b
EQM, BRM et REQMR (pourcentage) des estimateurs de l'EQM, $m = 15$ domaines

			RB_Y1		RB_Y2		M_et_coll		Estimateur BP		Estimateur BP naïf	
	σ_v^2 / ψ_i	$\overline{\text{EQM}}$	BRM	REQMR	BRM	REQMR	BRM	REQMR	%BRM	%REQMR	%BRM	%REQMR
MIX	0,06	584,9	21,0	84,7	35,4	93,7	12,6	87,9	10,0	53,8	19,3	62,1
	0,1	377,1	46,0	103,9	68,4	122,6	26,4	106,1	14,8	52,7	26,6	59,9
	0,14	353,0	41,9	91,5	59,4	112,7	25,0	89,9	7,6	33,2	13,7	36,7
	0,2	209,7	83,2	127,8	108,9	155,8	42,8	122,8	14,0	42,5	23,7	46,0
	0,3	198,9	94,8	136,7	117,1	162,2	50,4	124,6	8,7	26,6	14,5	27,7

En résumé, sous le modèle de Fay-Herriot avec une σ_v^2 positive, MIX et AR.YL sont les seuls estimateurs de variance positifs à l'étude qui présentent un biais asymptotique négligeable. Le biais asymptotique des estimateurs de variance AM.YL et LL est plus grand. En revanche, notre simulation a démontré que, pour un nombre modéré de domaines et pour les populations qui donnent des estimations REML nulles, les deux estimateurs de variance de YL présentaient un biais négatif et produisaient des EBLUP proches de l'estimateur synthétique de la moyenne. Par contre, le MIX, qui combine les estimateurs AM.LL et REML, ne présentait qu'un biais légèrement négatif dans ces populations. De plus, la distribution inconditionnelle du MIX approchait de la normalité beaucoup plus rapidement que celle des autres estimateurs de variance.

Tableau 6.8
$\mathbf{EQM_C}$ et $\mathbf{BRM_C}$ m = 15 domaines

Méthode	σ_v^2/ψ_i	$\overline{\mathbf{EQM_C}}$	Estimateur de l'EQM de Taylor			Estimateur BP	Estimateur BP naïf
REML	0,06	594,2		-22,6		-31,7	-16,5
	0,1	381,2		-32,9		-43,2	-22,5
	0,14	345,1		-17,7		-22,7	-10,7
	0,2	212,7		-41,1		-47,3	-25,5
	0,3	197,9		-30,4		-32,7	-17,6
AM.LL	0,06	595,6		-4,1		-5,7	12,1
	0,1	385,7		8,6		-7,0	15,6
	0,14	351,2		18,9		-2,0	10,4
	0,2	216,0		46,4		-5,8	14,4
	0,3	199,5		67,0		-2,9	9,8
AR.YL	0,06	592,2		-0,8		-27,1	-11,0
	0,1	379,7		21,0		-36,5	-14,8
	0,14	344,5		44,0		-18,6	-6,3
	0,2	210,9		98,2		-38,6	-16,4
	0,3	196,6		177,3		-26,1	-11,0
AM.YL	0,06	581,7		30,7		-21,9	-18,0
	0,1	368,6		79,8		-31,5	-25,8
	0,14	333,9		98,3		-15,2	-11,9
	0,2	198,9		198,0		-36,4	-30,0
	0,3	190,0		296,3		-26,2	-21,5
			RB_Y1	**RB_Y2**	**M_et_coll**	**Estimateur BP**	**Estimateur BP naïf**
MIX	0,06	595,6	-4,1	27,9	-22,9	3,4	17,8
	0,1	385,7	8,6	57,1	-33,7	5,1	22,8
	0,14	351,2	18,9	58,5	-19,1	4,9	14,3
	0,2	216,0	46,4	102,4	-42,0	5,9	20,4
	0,3	199,5	67,0	116,3	-30,9	4,8	13,4

En ce qui concerne l'estimateur de l'EQM de l'EBLUP, il était beaucoup plus précis que l'estimateur direct, sous toutes les méthodes d'estimation de variance examinées ici, même pour un petit nombre de domaines. Les estimateurs de variance AM.LL, AM.YL et AR.YL étaient tous moins variables que le REML et le MIX. L'impact sur l'EQM de l'EBLUP était minime, car il y avait peu de différences entre les EQM pour le même rapport signal/bruit. Ces différences s'accentuaient à mesure que le nombre de domaines ou le rapport signal/bruit diminuait. Ainsi, pour un rapport signal/bruit extrêmement faible, l'EQM sous la méthode MIX pourrait être un peu plus grande que sous l'estimateur de variance AM.YL.

Sous la méthode MIX d'estimation de variance, nous avons comparé trois estimateurs de l'EQM de type Taylor et deux estimateurs de l'EQM bootstrap. Les trois estimateurs de l'EQM de Taylor sous la méthode MIX (RB_Y1, RB_Y2 et M_et_coll) sont sans biais jusqu'à l'ordre deux. Les estimateurs de l'EQM de type Taylor sous LL et YL sont eux aussi sans biais jusqu'à l'ordre deux. Les estimateurs RB_Y1, AM.LL et AM.YL peuvent produire des estimations négatives de l'EQM.

L'estimateur de l'EQM de Taylor sous la méthode REML d'estimation de variance et l'estimateur M_et_coll sous la méthode MIX coïncidant par définition, les différences entre leurs mesures de performance sont négligeables (leurs EQM réelles sont différentes; dans notre étude cependant, pour $m = 100,$ l'estimateur MIX coïncidait avec le REML 84 % du temps). Pour un nombre modéré de domaines, qui pourrait être $m = 45$ ou 100 pour ces données, et pour les populations qui donnent des estimations REML nulles, les deux estimateurs de l'EQM de Taylor sous le REML et les estimateurs de l'EQM de M_et_coll ne tiennent pas compte de la variation attribuable à l'estimation de $\sigma_v^2,$ ce qui se reflète dans leur $\mathrm{BRM}_{\mathbb{C}},$ très négatif, qui est inférieur à -60 % pour les rapports signal/bruit plus petits. Par ailleurs, l'estimateur RB_Y1 tient compte de la variation due à l'estimation de $\sigma_v^2,$ mais son $\mathrm{BRM}_{\mathbb{C}}$ est lui aussi très négatif. En effet, le RB_Y1 est un estimateur de l'EQM fractionné qui, pour les populations où $\hat{\sigma}^2_{v\mathrm{REML}} = 0,$ soustrait un facteur du biais inconditionnel de l'AM.LL, qui est toujours positif. Une meilleure formule pour un estimateur de l'EQM fractionné serait d'utiliser un estimateur du biais conditionnel $E\left(\hat{\sigma}_v^2 / \hat{\sigma}^2_{v\mathrm{REML}} = 0\right).$ En fait, même pour un nombre modéré de domaines $(m = 100),$ le tableau 6.1 montre que l'estimateur MIX a un biais inconditionnel de 49 %, mais un biais conditionnel de -37 %.

L'estimateur BP de l'EQM sous les méthodes AR.YL et MIX était bien corrigé pour le biais, mais la variance en souffrait. Le bootstrap naïf semble être l'estimateur de l'EQM qui donne les meilleurs résultats, et ces résultats sont encore meilleurs sous l'estimation de variance MIX, même lorsque les trois mesures (BRM, $\mathrm{BRM}_{\mathbb{C}}$ et REQMR) sont prises en compte. Nous avons constaté que, pour un nombre modéré de domaines, le RB_Y2 était l'estimateur de Taylor sous la méthode MIX qui présentait la REQMR la plus faible. Par ailleurs, l'estimateur de M_et_coll est le plus fiable lorsque la variance sous-jacente réelle σ_v^2 est très faible : dans ce cas-ci, M_et_coll est effectivement l'estimateur de l'EQM de l'estimateur synthétique de la moyenne de petit domaine. Nous ne recommandons pas de compter sur l'approximation d'ordre deux de l'EQM lorsque m est petit : l'approximation (2.6) de l'EQM ne tient pas nécessairement, les mesures de performance obtenues dans notre étude sont très instables, et elles peuvent varier d'un ensemble de données à l'autre.

En conclusion, sous l'hypothèse $\sigma_v^2 > 0,$ les performances relatives des estimateurs de variance positifs comparés dépendent de la taille de $\sigma_v^2,$ du rapport signal/bruit, du nombre de domaines et de la fonction objectif. Pour un nombre modéré de domaines, l'estimateur de variance MIX semblait donner de meilleurs résultats que les estimateurs LL et YL dans cette étude. Sous la méthode MIX, l'estimateur BP naïf de l'EQM présentait la combinaison $\mathrm{BRM}_{\mathbb{C}}$ et REQMR la plus faible. L'estimateur de l'EQM de M_et_coll sous l'estimateur de variance MIX donnait des résultats légèrement meilleurs que le RB_Y1 lorsque la variance σ_v^2 sous-jacente était très petite. Cependant, le pourcentage de REML nuls produits sous le modèle de simulation montre qu'un résultat de $\hat{\sigma}^2_{v\mathrm{REML}} = 0$ et/ou que des tests d'hypothèses négatifs ne signifient pas nécessairement que la variance σ_v^2 est assez faible pour que l'on puisse se fier à l'estimateur de M_et_coll. En l'absence d'autres renseignements, l'estimateur BP naïf sous la méthode MIX semble donner de meilleurs résultats.

Remerciements

Les auteurs désirent remercier le professeur J.N.K. Rao de l'Université Carleton pour ses commentaires utiles ainsi que Victor Estevao de Statistique Canada, qui a développé un algorithme de maximisation de grille spécialement pour ce projet. Ils tiennent également à remercier les membres du comité de revue pour leur évaluation consciencieuse de cet article et pour leurs suggestions d'amélioration.

Annexe A

Preuve du théorème 4.1

La variance asymptotique de $\hat{\sigma}^2_{v\text{MIX}}$ est donnée par : $\overline{V}\left(\hat{\sigma}^2_{v\text{MIX}}\right) = \lim_{m\to\infty} E\left(\hat{\sigma}^2_{v\text{MIX}} - \sigma^2_v\right)^2$

Nous montrons que $E\left(\hat{\sigma}^2_{v\text{MIX}} - \sigma^2_v\right)^2 \le E\left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right)^2 + o(1/m)$ à mesure que $m \to \infty$.

$$
\begin{aligned}
E\left(\hat{\sigma}^2_{v\text{MIX}} - \sigma^2_v\right)^2 &= \int_{\left\{\hat{\sigma}^2_{v\text{REML}} > 0\right\}} \left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right)^2 dP + \int_{\left\{\hat{\sigma}^2_{v\text{REML}} = 0\right\}} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2 dP \\
&\le \int_{\Omega} \left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right)^2 dP + \int_{\left\{\hat{\sigma}^2_{v\text{REML}} = 0\right\}} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2 dP = E\left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right)^2 \quad \text{(A.1)} \\
&\quad + o\left(\frac{1}{m}\right).
\end{aligned}
$$

En effet, par les inégalités de Holder et Minkowski, avec $1 < p < \infty, 1/p + 1/q = 1,$ et si $X \equiv \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2 = \text{O}_p(1/m)$ et l'indicateur $I\left(\hat{\sigma}^2_{v\text{REML}} = 0\right)$ des populations avec $\hat{\sigma}^2_{v\text{REML}} = 0,$ nous avons :

$$
\begin{aligned}
\int_{\left\{\hat{\sigma}^2_{v\text{REML}} < 0\right\}} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2 dP &\le \left(\int_{\Omega} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^{2p} dP\right)^{1/p} . \left(P\left\{\hat{\sigma}^2_{v\text{REML}} = 0\right\}\right)^{1/q} \\
&= \left(O\left(\frac{1}{m^p}\right)\right)^{1/p} . (o(1))^{1/q} = o\left(\frac{1}{m}\right),
\end{aligned}
\qquad \text{(A.2)}
$$

puisque $\left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2$ est uniformément borné et $\hat{\sigma}^2_{v\text{REML}} \xrightarrow{P} \sigma^2_v > 0.$ Il est à noter que les estimateurs AM.LL et REML de σ^2_v sont uniformément bornés en conséquence de leur convergence presque certaine vers σ^2_v (voir, par exemple, Yuan et Jennrich 1998).

Preuve du théorème 4.2

Nous désignons par $\hat{\boldsymbol{\sigma}}^2_{v\text{ML}}$ l'estimateur de variance par le maximum de vraisemblance.

Nous montrons d'abord que $\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} - \hat{\boldsymbol{\sigma}}^2_{v\text{ML}} = O_p(1/m).$ Soit $G_*\left(\boldsymbol{\sigma}^2_v\right) = \partial \log\left(L_*\right) \big/ \partial \boldsymbol{\sigma}^2_v = 0$ l'équation d'estimation qui donne l'estimateur de variance *. L'équation (3.4) implique que :

$$G_{\mathrm{AM.LL}}\left(\boldsymbol{\sigma}_v^2\right) - G_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right) = \partial \log \boldsymbol{\sigma}_v^2 \big/ \partial\, \boldsymbol{\sigma}_v^2 = \frac{1}{m\boldsymbol{\sigma}_v^2} = O\left(\frac{1}{m}\right). \tag{A.3}$$

Avec $G'_{\mathrm{ML}}(\cdot) \triangleq \left(\partial G_{\mathrm{ML}} \big/ \partial \boldsymbol{\sigma}_v^2\right)(\cdot)$ et $G''_{\mathrm{ML}}(\cdot) \triangleq \left(\partial G'_{\mathrm{ML}} \big/ \partial \boldsymbol{\sigma}_v^2\right)(\cdot)$, l'équation (A.3) implique que :

$$G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right) - G'_{\mathrm{AM.LL}}\left(\boldsymbol{\sigma}_v^2\right) = O\left(\frac{1}{m}\right). \tag{A.4}$$

Maintenant, en utilisant l'équation (A.4), la $\sqrt{m}$ – convergence des estimateurs ML et AM.LL de $\boldsymbol{\sigma}_v^2$, le développement en série de Taylor à deux termes de $G_{\mathrm{ML}}(\cdot)$ et $G_{\mathrm{AM.LL}}(\cdot)$ à $\boldsymbol{\sigma}_v^2$ et $G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right) = O(1)$ à mesure que $m \to \infty$, le côté gauche de l'équation en (A.3) est égal à :

$$= G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} - \boldsymbol{\sigma}_v^2\right) - G'_{\mathrm{AM.LL}}\left(\boldsymbol{\sigma}_v^2\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}} - \boldsymbol{\sigma}_v^2\right) + O_p\left(\frac{1}{m}\right)$$

$$= G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} - \hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}}\right) + \left(G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right) - G'_{\mathrm{AM.LL}}\left(\boldsymbol{\sigma}_v^2\right)\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}} - \boldsymbol{\sigma}_v^2\right) + O_p\left(\frac{1}{m}\right)$$

$$= G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} - \hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}}\right) + O_p\left(\frac{1}{m^{3/2}}\right) + O_p\left(\frac{1}{m}\right).$$

La dernière égalité ci-dessus implique que :

$$\hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}} - \hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} = O_p\left(\frac{1}{m}\right) \text{ à mesure que } m \to \infty. \tag{A.5}$$

De même, nous établissons une relation entre $G_{\mathrm{REML}}\left(\boldsymbol{\sigma}_v^2\right)$ et $G_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right)$: étant donné que $\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right) = O(1)$ découle des conditions 1 à 3 de la section 3 et de l'équation (3.1), nous avons :

$$G_{\mathrm{REML}}\left(\boldsymbol{\sigma}_v^2\right) - G_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right) = \frac{1}{m}\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right) = O\left(\frac{1}{m}\right) \text{ à mesure que } m \to \infty, \tag{A.6}$$

l'équation (A.6) et le même argument que pour l'estimateur AM.LL impliquent que :

$$\hat{\boldsymbol{\sigma}}^2_{v\mathrm{REML}} - \hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} = O_p\left(\frac{1}{m}\right) \text{ à mesure que } m \to \infty. \tag{A.7}$$

Ensemble, les équations (A.5) et (A.7) donnent :

$$\left(\hat{\sigma}^2_{v\mathrm{REML}} - \hat{\sigma}^2_{v\mathrm{AM.LL}}\right) = O_P\left(\frac{1}{m}\right). \tag{A.8}$$

Nous exprimons maintenant le biais de l'estimateur MIX comme suit :

$$B_{\mathrm{MIX}}\left(\hat{\sigma}^2_{v\mathrm{MIX}}\right) = \int_{\left\{\hat{\sigma}^2_{v\mathrm{REML}} > 0\right\}} \left(\hat{\sigma}^2_{v\mathrm{REML}} - \sigma_v^2\right) dP + \int_{\left\{\hat{\sigma}^2_{v\mathrm{REML}} = 0\right\}} \left(\hat{\sigma}^2_{v\mathrm{AM.LL}} - \sigma_v^2\right) dP.$$

Nous ajoutons et soustrayons $\int_{\{\hat{\sigma}^2_{v\text{REML}}=0\}} \left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right) dP$ du côté droit de l'équation ci-dessus pour obtenir :

$$\begin{aligned} B_{\text{MIX}}\left(\hat{\sigma}^2_{v\text{MIX}}\right) &= \int_{\Omega} \left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right) dP + \int_{\{\hat{\sigma}^2_{v\text{REML}}=0\}} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \hat{\sigma}^2_{v\text{REML}}\right) dP \\ &= \text{Biais}\left(\hat{\sigma}^2_{v\text{REML}}\right) + \int_{\{\hat{\sigma}^2_{v\text{REML}}=0\}} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \hat{\sigma}^2_{v\text{REML}}\right) dP. \end{aligned} \tag{A.9}$$

Puisque $\hat{\sigma}^2_{v\text{AM.LL}} - \hat{\sigma}^2_{v\text{REML}}$ est uniformément borné, nous appliquons l'inégalité de Holder et Minkowski avec $p = q = 2$ et l'équation (A.8) au dernier terme en (A.9) pour obtenir :

$$\begin{aligned} B_{\text{MIX}}\left(\hat{\sigma}^2_{v\text{MIX}}\right) &= \text{Biais}\left(\hat{\sigma}^2_{v\text{REML}}\right) + \left(\int_{\Omega} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \hat{\sigma}^2_{v\text{REML}}\right)^2 dP\right)^{1/2} \cdot P\left\{\hat{\sigma}^2_{v\text{REML}} = 0\right\}^{1/2} \\ &= \text{Biais}\left(\hat{\sigma}^2_{v\text{REML}}\right) + O\left(\frac{1}{m}\right) \cdot o(1) = \text{Biais}\left(\hat{\sigma}^2_{v\text{REML}}\right) + o\left(\frac{1}{m}\right). \end{aligned} \tag{A.10}$$

Preuve de la remarque 4.2 : eqm_0 est sans biais jusqu'à l'ordre deux

$$\begin{aligned} E(\text{eqm}_0) - \text{EQM}\left(\hat{\boldsymbol{\theta}}_i\right) &= \int_{\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}>0\}} (g_{1i} + g_{2i} + 2g_{3i})\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP + \int_{\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}=0\}} g_{2i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP - \text{EQM} \\ &= \left[\int_{\Omega} (g_{1i} + g_{2i} + 2g_{3i})\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP - \text{EQM}\right] \\ &\quad + \int_{\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}=0\}} g_{2i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP - \int_{\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}=0\}} (g_{1i} + g_{2i} + 2g_{3i})\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP \\ &= \left[\text{o}\left(\frac{1}{m}\right)\right] - \int_{\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}=0\}} 2g_{3i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP, \end{aligned} \tag{A.11}$$

puisque $g_{1i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) = g_{1i}(0) = 0$ en $\left\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} = 0\right\}$ et $g_{2i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right)$ s'annulent en (A.11). Mais

$$g_{3i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) = g_{3i}(0) = \frac{\bar{V}(0)}{\boldsymbol{\psi}_i} = O_p\left(\frac{1}{m}\right)$$

et est uniformément borné sous les conditions de régularité données dans la section 2. Le dernier terme en (A.11) est donc aussi un $\text{o}(1/m)$, ce qui rend eqm_0 sans biais jusqu'à l'ordre deux.

Annexe B

B.1 Comparaison entre les estimateurs REML et AR.YL au moyen de l'algorithme de score

L'algorithme de score donnait parfois des estimations nulles pour la vraisemblance de l'estimateur AR.YL. En effet, pour les ensembles de données simulés sous le modèle donné dans la section 5, où $m = 45$ et $\sigma_v^2 = 1$, les algorithmes de score REML et AR.YL produisaient 28 % et 26 % d'estimations nulles respectivement. On peut voir pourquoi dans les figures B.1 à B.3 : les vraisemblances correspondent à une seule population générée sous le modèle avec $\boldsymbol{\sigma}_v^2 = 1$ pour lequel $\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} = 0$.

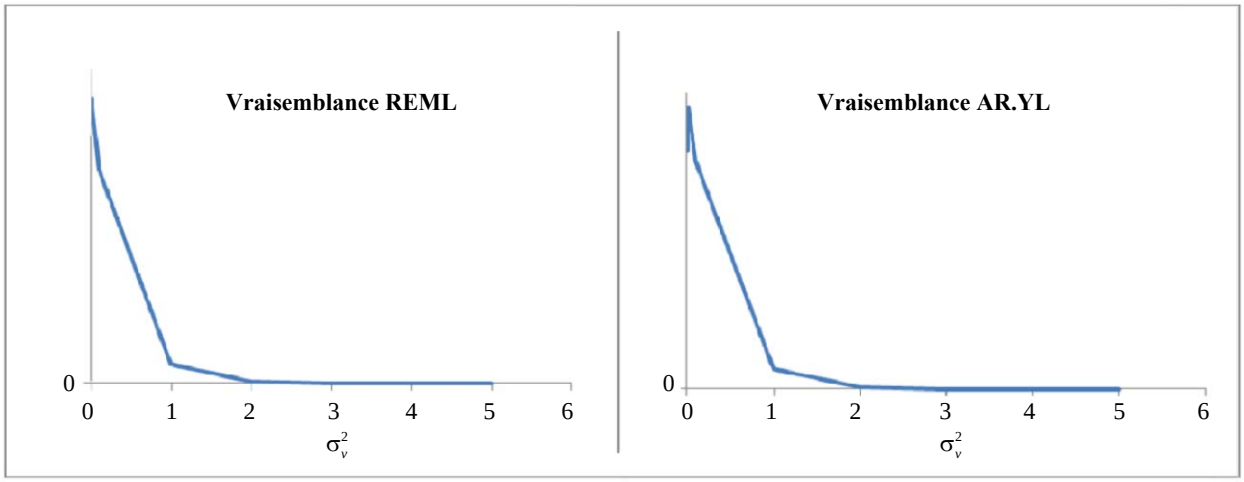

Figure B.1 $L = L_{\text{REML}}\left(\sigma_v^2 \middle| y_1, \ldots, y_{45}\right)$.

Figure B.2 $L = L_{\text{AR.YL}}\left(\sigma_v^2 \middle| y_1, \ldots, y_{45}\right)$.

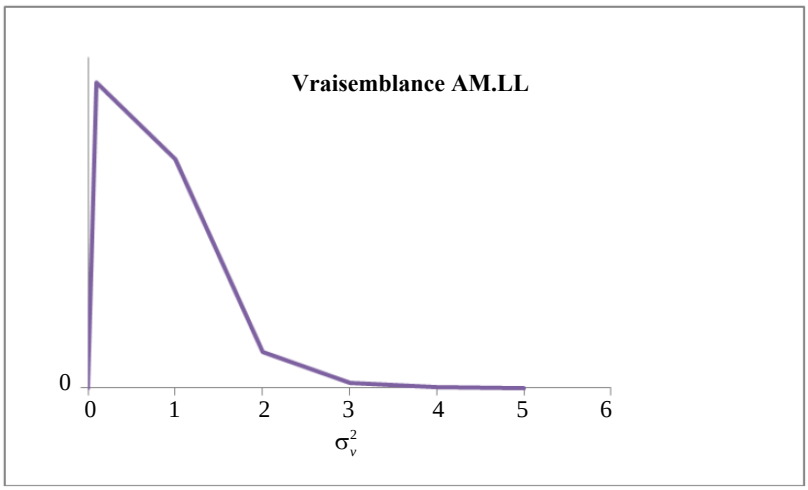

Figure B.3 $L = L_{\text{AM.LL}}\left(\sigma_v^2 \middle| y_1, \ldots, y_{45}\right)$.

La figure B.2 montre que la valeur maximale de la vraisemblance AR.YL est très proche de la limite. Il arrive parfois que l'algorithme de score passe à côté du maximum et donne une valeur nulle. La figure B.3 montre que la vraisemblance AM.LL a une valeur maximale qui se différencie mieux de la limite.

B.2 Traitement des zéros dans l'estimateur bootstrap paramétrique

Pour chaque estimation $\hat{\sigma}_v^2 = \hat{\sigma}_v^2\left(\mathbf{y}^{(r)}\right), r = 1,\ldots 10K,$ et chaque méthode d'estimation de variance :

i. Générer un grand nombre B d'effets aléatoires de domaine $v_i^{(b)} \overset{\text{i.i.d.}}{\sim} N\left(0, \hat{\sigma}_v^2\right), b = 1,\ldots,B,$ et générer, indépendamment de $v_i^{(b)},$ des erreurs d'échantillonnage $e_i^{(b)} \overset{\text{i.i.d.}}{\sim} N\left(0, \psi_i\right),\ i = 1,\ldots,m,$ $b = 1,\ldots,B.$ Générer des données bootstrap $y_i^{(b)} = \theta_i^{(b)} + e_i^{(b)},\ \theta_i^{(b)} = \mathbf{x}_i'\hat{\beta} + v_i^{(b)},\ i = 1,\ldots,m.$ Si $\hat{\sigma}^2_{v\text{REML}}\left(y^{(r)}\right) = 0,$ générer $\left(y_i^{(b)}, \theta_i^{(b)}\right), b = 1,\ldots,B,$ à partir du modèle synthétique (voir aussi Rao et Molina 2015).

ii. Adapter le modèle aux données bootstrap et obtenir $\hat{\sigma}_v^{2(b)}$; pour l'estimateur MIX, calculer $\hat{\sigma}^{2(b)}_{v\text{MIX}} = \hat{\sigma}^{2(b)}_{v\text{REML}}$ si $\hat{\sigma}_v^{2(b)}$ est positif et $\hat{\sigma}^{2(b)}_{v\text{MIX}} = \hat{\sigma}^{2(b)}_{v\text{AM}}$ autrement.

iii. Obtenir $\hat{\beta}^{(b)},$ l'EBLUP correspondant $\hat{\theta}_i^{(b)},$ les composantes bootstrap $g_{1i}^{(b)} = g_{1i}\left(\hat{\sigma}_v^2\left(y^{(b)}\right)\right),$ $g_{2i}^{(b)} = g_{2i}\left(\hat{\sigma}_v^2\left(y^{(b)}\right)\right)$ et $\overline{g}_{ji}^{\text{BP}} = B^{-1}\sum_b g_{ji}^{(b)}, j = 1,2.$

iv. L'estimateur bootstrap naïf de l'EQM est $\text{eqm}_{\text{naive}} = B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_i^{(b)} - \theta_i^{(b)}\right)^2.$

v. L'estimateur BP de l'EQM (qui est corrigé du biais (Pfeffermann et Glickman 2004)) est : $\text{eqm}_{\text{BP}}\left(\hat{\theta}_i\right) = g_{1i}\left(\hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_v^2\right) - \overline{g}_{1i}^{\text{BP}} - \overline{g}_{2i}^{\text{BP}} + \text{eqm}_{\text{naive}}.$

vi. Pour calculer le $\text{BRM}_{\mathbb{C}},$ faire la moyenne de $\left(\text{eqm}_{\text{BP}}^{(r)}\left(\hat{\theta}_i\right) - \text{EQM}\left(\hat{\theta}_i\right)\right)\Big/\text{EQM}\left(\hat{\theta}_i\right)$ pour les populations où $(r)\Big/\hat{\sigma}^2_{v\text{REML}}\left(\mathbf{y}^{(r)}\right) = 0$ et faire de même pour le $\text{BRM}_{\mathbb{C}}$ de $\text{eqm}_{\text{naive}}.$

Bibliographie

Chen, S., et Lahiri, P. (2008). On mean squared prediction error estimation in small area estimation problems. *Communications in Statistics-Theory and Methods*, 37, 1792-1798.

Chen, S., et Lahiri, P. (2011). On the estimation of Mean Squared Prediction Error in small area estimation. *Calcutta Statistical Association Bulletin*, 63, (Special 7th Triennial Proceedings Volume), Nos. 249-252.

Cressie, N. (1992). Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Baye. *Techniques d'enquête*, 18, 1, 83-103.

Das, K., Jiang, J. et Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 2, 818-840.

Datta, G., et Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

Estevao, V. (2014). Grid optimization algorithm for maximum likelihood. Rapport interne, Division de la recherche et de l'innovation en statistique (DRIS), Statistique Canada.

Fay, R.E., et Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein Procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Lahiri, P., et Li, H. (2009). Generalized maximum likelihood method in linear mixed models with an application in small area estimation. Dans *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, disponible au http//www.fcsm.gov/events/papers2009.html.

Lahiri, P., et Pramanik, S. (2011). Discussion of "Estimating random effects via adjustment for density maximization" par C. Morris et R. Tang. *Statistical Science*, 26, 2, 291-295.

Li, H., et Lahiri, P. (2011). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.

Molina, I., Rao, J.N.K. et Datta, G.S. (2015). Estimation sur petits domaines sous un modèle de Fay-Herriot avec test préliminaire pour la présence d'effets aléatoires de domaine. *Techniques d'enquête*, 41, 1, 1-20.

Morris, C.N. (2006). Mixed model prediction and small area estimation (with discussions). *Test*, 15, 72-76.

Pfeffermann, D., et Glickman, H. (2004). Mean squarred error approximation in small area estimation by use of parametric and non-parametric bootstrap. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, Alexandria, VA. 4167-78.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation, second edition*. New York : John Wiley & Sons, Inc.

Rubin-Bleuer, S., et Schiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33, 6, 2789-2810.

Rubin-Bleuer, S., et You, Y. (2012). A positive variance estimator for the Fay-Herriot small area model. SRID-2012-009E, Division de la recherche et de l'innovation en statistique (DRIS), Statistique Canada.

Rubin-Bleuer, S., Yung, W. et Landry, S. (2010). Adjusted maximum likelihood method for a small area model accounting for time and area effects. SRID-2010-006E, Division de la recherche et de l'innovation en statistique (DRIS), Statistique Canada.

Rubin-Bleuer, S., Yung, W. et Landry, S. (2011). Adjusted maximum likelihood method for a small area model accounting for time and area effects. Long abstract, *Small Area Estimation*, (SAE 20122) à Trèves, Allemagne, International Statistical Institute Satellite Conference.

Rubin-Bleuer, S., Yung, W. et Landry, S. (2012). Variance Component Estimation through the Adjusted Maximum Likelihood Approach. Exposé donné à la conférence en l'honneur du 75^{e} anniversaire de J.N.K. Rao, Université Carleton, mai 2012, Ottawa.

Yoshimori, M., et Lahiri, P. (2014). A new adjusted maximum likelihood method for the Fay-Herriot small area model. *Journal of Multivariate Analysis*, 124, 281-294.

Yuan, P. (2009). Comparison of SAE methods of variance estimation. Document interne, Division de la recherche et de l'innovation en statistique (DRIS), Statistique Canada.

Yuan, K.H., et Jennrich, R. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis,* 65, 2, 245-260.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 95-130
Statistique Canada, nº 12-001-X au catalogue

Une comparaison d'estimateurs non paramétriques pour les fonctions de répartition de populations finies

Leo Pasquazzi et Lucio de Capitani[1]

Résumé

Le présent travail a pour objet de comparer des estimateurs non paramétriques pour des fonctions de répartition de populations finies fondés sur deux types de valeurs prédites, à savoir celles données par l'estimateur bien connu de Kuo et une version modifiée de ces dernières, qui intègre une estimation non paramétrique de la fonction de régression à la moyenne. Pour chaque type de valeurs prédites, nous considérons l'estimateur fondé sur un modèle correspondant et, après incorporation des poids de sondage, l'estimateur par la différence généralisée. Nous montrons sous des conditions assez générales que le terme principal de l'erreur quadratique moyenne sous le modèle n'est pas affecté par la modification des valeurs prédites, même si cette modification réduit la vitesse de convergence pour le biais sous le modèle. Les termes d'ordre deux des erreurs quadratiques moyennes sous le modèle sont difficiles à obtenir et ne seront pas calculés dans le présent article. La question est de savoir si les valeurs prédites modifiées offrent un certain avantage du point de vue de l'approche fondée sur un modèle. Nous examinons aussi les propriétés des estimateurs sous le plan de sondage et proposons pour l'estimateur par la différence généralisée un estimateur de variance fondé sur les valeurs prédites modifiées. Enfin, nous effectuons une étude en simulation. Les résultats des simulations laissent entendre que les valeurs prédites modifiées entraînent une réduction importante de l'erreur quadratique moyenne si l'échantillon est de petite taille.

Mots-clés : Échantillonnage en population finie; estimateur de fonction de répartition; valeur prédite; estimateur de Kuo.

1 Introduction

Depuis la publication de l'article fondamental de Chambers et Dunstan (1986), plusieurs estimateurs ont été proposés pour les fonctions de répartition de populations finies. La plupart sont fondés sur différents types de valeurs prédites ou sur différents moyens de combiner ces valeurs en un estimateur. Ainsi, l'estimateur proposé par Chambers et Dunstan (1986) s'appuie sur des valeurs prédites tirées d'un modèle de superpopulation dans lequel le lien entre la variable étudiée et une variable auxiliaire est donné par un modèle de régression linéaire à composantes d'erreur indépendantes dont les variances sont supposées connues. En remplaçant les fonctions indicatrices non observées par les valeurs prédites dans la définition de la fonction de répartition de la population de la variable étudiée, on obtient l'estimateur de Chambers et Dunstan. Rao, Kovar et Mantel (1990) intègrent les poids de sondage dans les valeurs prédites de Chambers et Dunstan, puis utilisent celles-ci dans un estimateur par la différence généralisée. Kuo (1988) recourt à la régression non paramétrique pour estimer directement la relation de régression entre les fonctions indicatrices et la variable auxiliaire, et obtient des valeurs prédites qui admettent pratiquement n'importe quel modèle de superpopulation. Comme Chambers et Dunstan, elle remplace les fonctions indicatrices non observées par les valeurs prédites correspondantes et obtient un estimateur fondé sur un modèle. Chambers, Dorfman et Wehrly (1993) combinent les valeurs prédites de Chambers et Dunstan (1986) et de Kuo (1988) et proposent un autre estimateur fondé sur un modèle qui vise à être plus efficace que l'estimateur de Kuo si le modèle de superpopulation linéaire supposé par Chambers et Dunstan est vérifié, et qui ne souffre pas d'un biais de spécification incorrecte du modèle autrement. À la suite de ces premiers travaux, un assez grand nombre de propositions ont été faites en vue de réaliser un gain d'efficacité par rapport à l'estimateur

1. Leo Pasquazzi et Lucio de Capitani, Università degli Studi di Milano-Bicocca, Milan, Italie. Courriel : leo.pasquazzi@unimib.it, lucio.decapitani1@unimib.it.

de Horvitz-Thompson, tout en préservant la robustesse de ce dernier et parfois aussi l'une de ses propriétés souhaitables suivantes ou les deux, à savoir i) le fait qu'il s'agit d'une combinaison linéaire des fonctions indicatrices dans l'échantillon dont les coefficients ne dépendent pas de la variable étudiée et ii) le fait qu'il produit toujours des estimations non décroissantes pour la fonction de répartition.

Le présent travail part de l'idée d'améliorer les valeurs prédites proposées par Kuo (1988) en y incorporant une estimation de la fonction de régression à la moyenne (voir la section 2). Cette idée, avancée dans un ouvrage récent de Chambers et Clark (2012), repose sur l'hypothèse d'un modèle de superpopulation sous-jacent caractérisé par une relation de régression lisse entre la variable étudiée et une variable auxiliaire, ainsi qu'une variation lisse des distributions des composantes de l'erreur. Selon cette idée, les valeurs prédites sont le résultat d'une procédure en deux étapes : à la première étape, la fonction de régression à la moyenne est estimée par régression paramétrique ou non paramétrique, et à la deuxième étape, en utilisant les résidus de cette régression, les fonctions de répartition des composantes de l'erreur sont estimées par régression non paramétrique afin de tenir compte de la possibilité d'une variation lisse des distributions des composantes de l'erreur. En combinant les deux estimations, on peut calculer les valeurs prédites pour les fonctions indicatrices qui figurent dans la fonction de répartition de la population finie de la variable étudiée. Chambers et Clark (2012) analysent l'estimateur fondé sur un modèle obtenu en remplaçant les fonctions indicatrices non observées par les valeurs prédites correspondantes, et ils esquissent une preuve qui mène à une expression pour la variance sous le modèle de l'estimateur résultant. Dans cette preuve, ils supposent que la fonction de régression à la moyenne est estimée au moyen d'un estimateur convergent et que la contribution de son erreur d'estimation à la variance sous le modèle de l'estimateur de la fonction de répartition finale peut être négligée. Dans le présent travail, nous considérons la régression linéaire locale pour estimer à la fois la fonction de régression à la moyenne sous le modèle et les distributions des composantes de l'erreur. Nous donnons des développements asymptotiques pour le biais et pour la variance sous le modèle de l'estimateur résultant et les comparons à ceux correspondant à l'estimateur de Kuo fondé sur la régression linéaire locale. Il s'avère que les termes principaux dans les variances sous le modèle sont les mêmes et que, pour des suites de fenêtres de lissage choisies comme il convient, le carré du biais sous le modèle des deux estimateurs tend vers zéro plus rapidement que la variance sous le modèle. Pour établir quel estimateur est asymptotiquement plus efficace du point de vue de la modélisation, il est donc nécessaire de connaître les termes d'ordre deux des variances sous le modèle. Cependant, ces derniers dépendent d'hypothèses plus précises que celles considérées dans le présent travail et, du moins pour l'estimateur fondé sur les valeurs prédites modifiées, il semble que la détermination des termes d'ordre deux des variances sous le modèle ne soit pas une tâche facile. La question de savoir quel estimateur est le plus efficace du point de vue de la modélisation reste donc à résoudre.

En plus des estimateurs fondés sur un modèle susmentionnés, nous analysons les estimateurs par la différence généralisée fondés sur les deux types de valeurs prédites dans leurs versions pondérées selon le plan de sondage. Les résultats présentés à la section 3 montrent que les vitesses de convergence de leurs biais et de leurs variances sous le modèle sont les mêmes que pour leurs équivalents fondés sur un modèle. Les propriétés sous le plan de sondage sont discutées dans une certaine mesure à la section 4, de même que la question de l'estimation de la variance. Il serait évidemment intéressant d'établir et de comparer les développements asymptotiques pour les biais et les variances sous le plan de sondage. Breidt et Opsomer (2000) obtiennent sous des conditions faibles une expression générale pour le terme d'ordre un dans l'erreur quadratique moyenne sous le plan des estimateurs de régression par polynômes locaux, dont l'estimateur par la différence généralisée fondé sur les valeurs prédites de Kuo est un cas particulier. L'estimateur par la

différence généralisée fondé sur les valeurs prédites modifiées ne rentre toutefois pas dans cette classe. À l'instar de Särndal, Swensson et Wretman (1992), nous conjecturons que, sous des conditions générales, le terme d'ordre un de son erreur quadratique moyenne sous le plan est le même que celui de l'estimateur par la différence généralisée fondé sur les valeurs prédites de Kuo. Des preuves formelles pourraient peut-être être obtenues en adaptant et en étendant certains résultats présentés dans Wang et Opsomer (2011). Pour vérifier cette conjecture et comparer la performance de l'estimateur par la différence généralisée et de l'estimateur fondé sur un modèle dans diverses conditions, nous effectuons une étude en simulation dont les résultats sont présentés à la section 5.

2 Définition des estimateurs

Soit (y_i, x_i) les valeurs prises par une variable étudiée Y et une variable auxiliaire X sur l'unité i d'une population finie $U := \{1, 2, \ldots, N\}$. Supposons que

$$y_i = m(x_i) + \varepsilon_i, \qquad i \in U, \tag{2.1}$$

où $m(x)$ est une fonction lisse et où les ε_i sont des variables aléatoires indépendantes de moyenne nulle dont les fonctions de répartition $P(\varepsilon_i \le \varepsilon) = G(\varepsilon | x_i)$ varient continûment en fonction de x_i. Soit $s \subset U$ un échantillon tiré de la population U selon un certain plan de sondage. Comme d'habitude dans le contexte de l'information auxiliaire complète, nous supposons que les valeurs x_i sont connues pour toutes les unités de la population, tandis que les valeurs y_i sont observées uniquement pour les unités de la population qui appartiennent à l'échantillon s.

Pour estimer la fonction de répartition inconnue de la population

$$F_N(t) := \frac{1}{N} \sum_{i \in U} I(y_i \le t),$$

Kuo (1988) propose l'estimateur donné par

$$\hat{F}(t) := \frac{1}{N} \left(\sum_{j \in s} I\left(y_j \le t\right) + \sum_{i \notin s} \sum_{j \in s} w_{i,j} I\left(y_j \le t\right) \right), \tag{2.2}$$

où, à la place de $w_{i,j}$, elle propose d'utiliser soit les poids de régression constants locaux

$$w_{i,j} := \frac{K\left(\frac{x_i - x_j}{\lambda}\right)}{\sum_{k \in s} K\left(\frac{x_i - x_k}{\lambda}\right)}$$

avec une fonction noyau (intégrable) à la place de $K(u)$ et $\lambda > 0$, soit les poids des k plus proches voisins

$$w_{i,j} := \begin{cases} 1/k, & \text{si } x_j \text{ est l'un des } k \text{ plus proches voisins de } x_i \\ 0, & \text{sinon.} \end{cases}$$

Notons que, dans la définition $\hat{F}(t)$,

$$\hat{G}_i(t) := \sum_{j \in s} w_{i,j} I\left(y_j \le t\right) \tag{2.3}$$

est utilisé comme valeur prédite remplaçant la fonction indicatrice non observée $I(y_i \le t)$ pour $i \notin s$.

En nous inspirant d'une idée avancée dans l'ouvrage de Chambers et Clark (2012), nous allons analyser un estimateur de $F_N(t)$ basé sur des valeurs prédites de rechange qui intègrent une estimation non paramétrique de la fonction de régression à la moyenne $m(x)$. Les valeurs prédites en question sont données par

$$\hat{G}_i^*(t) := \sum_{j \in s} w_{i,j} I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right) \tag{2.4}$$

où

$$\hat{m}_i := \sum_{k \in s} w_{i,j} y_j$$

est un estimateur non paramétrique de $m(x)$ à $x = x_i$, et l'estimateur résultant de $F_N(t)$ est donné par

$$\hat{F}^*(t) := \frac{1}{N}\left(\sum_{j \in s} I\left(y_j \le t\right) + \sum_{i \notin s} \sum_{j \in s} w_{i,j} I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right)\right). \tag{2.5}$$

Les valeurs prédites en (2.3) et (2.4), ou leurs versions modifiées de manière appropriée comprenant l'intégration des probabilités d'inclusion dans l'échantillon dans les poids de régression $w_{i,j}$, peuvent de toute évidence être calculées également pour $i \in s$, et elles peuvent être utilisées, par exemple, dans les estimateurs par la différence généralisée (Särndal et coll. 1992, page 221) ou dans les estimateurs calés sur un modèle (voir par exemple Wu et Sitter 2001; Chen et Wu 2002; Wu 2003; Montanari et Ranalli 2005; Rueda, Martínez, Martínez et Arcos 2007; Rueda, Sànchez-Borrego, Arcos et Martínez 2010). En plus des estimateurs fondés sur un modèle donné en (2.2) et (2.5), nous examinerons les estimateurs par la différence généralisée donnés par

$$\tilde{F}(t) := \frac{1}{N}\left(\sum_{i \in U} \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j \le t\right)\right) + \sum_{i \in s} \pi_i^{-1}\left(I\left(y_i \le t\right) - \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j \le t\right)\right)$$

et par

$$\tilde{F}^*(t) := \frac{1}{N}\left(\sum_{i \in U} \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j - \tilde{m}_j \le t - \tilde{m}_i\right)\right) + \sum_{i \in s} \pi_i^{-1}\left(I\left(y_i \le t\right) - \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j - \tilde{m}_j \le t - \tilde{m}_i\right)\right)$$

où π_i désigne les probabilités d'inclusion d'ordre un dans l'échantillon, $\tilde{w}_{i,j}$ désigne les poids de régression pondérés selon le plan de sondage dont la définition est donnée plus bas, et $\tilde{m}_i := \sum_{k \in s} \tilde{w}_{i,k} y_k$. Notons que, $\tilde{F}(t)$ et $\tilde{F}^*(t)$ sont fondés sur les équivalents des valeurs prédites pondérées selon le plan de sondage $\hat{G}_i(t)$ et $\hat{G}_i^*(t)$ qui sont donnés par

$$\tilde{G}_i(t) := \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j \le t\right)$$

et

$$\tilde{G}_i^*(t) := \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j - \tilde{m}_j \le t - \tilde{m}_i\right),$$

respectivement.

Quant aux poids de régression $w_{i,j}$ et $\tilde{w}_{i,j}$, nous les remplaçons dans le présent travail par des poids de régression linéaires locaux. Dans la suite de l'exposé, $w_{i,j}$ et $\tilde{w}_{i,j}$, sont donc définis par

$$w_{i,j} := \frac{1}{n\lambda} K\left(\frac{x_i - x_j}{\lambda}\right) \frac{M_{2,s}(x_i) - \left(\frac{x_i - x_j}{\lambda}\right) M_{1,s}(x_i)}{M_{2,s}(x_i) M_{0,s}(x_i) - M_{1,s}^2(x_i)}$$

et

$$\tilde{w}_{i,j} := \frac{1}{\pi_j n\lambda} K\left(\frac{x_i - x_j}{\lambda}\right) \frac{\tilde{M}_{2,s}(x_i) - \left(\frac{x_i - x_j}{\lambda}\right) \tilde{M}_{1,s}(x_i)}{\tilde{M}_{2,s}(x_i) \tilde{M}_{0,s}(x_i) - \tilde{M}_{1,s}^2(x_i)},$$

où n est le nombre d'unités dans l'échantillon s,

$$M_{r,s}(x) := \sum_{k \in s} \frac{1}{n\lambda} K\left(\frac{x - x_k}{\lambda}\right)\left(\frac{x - x_k}{\lambda}\right)^r, \qquad r = 0,1,2,$$

et

$$\tilde{M}_{r,s}(x) := \sum_{k \in s} \frac{1}{\pi_k n\lambda} K\left(\frac{x - x_k}{\lambda}\right)\left(\frac{x - x_k}{\lambda}\right)^r, \qquad r = 0,1,2.$$

Il convient de souligner que les estimateurs non paramétriques de la présente section ne sont pas bien définis si les poids de régression $w_{i,j}$ et $\tilde{w}_{i,j}$, inclus dans leurs définitions ne sont pas bien définis. Ce problème se pose, par exemple, quand le support de la fonction noyau $K(u)$ est donné par l'intervalle [–1,1] (par exemple, noyau uniforme, noyau d'Epanechnikov), et quand il n'existe pas au moins deux $j \in s$ tels que $\left|x_i - x_j\right| < \lambda$. Pour contourner ce problème, on peut utiliser une fonction noyau dont le support correspond à la courbe réelle entière (par exemple, noyau gaussien) ou choisir la fenêtre de lissage de manière adaptative. La dernière solution peut aussi aboutir à des estimateurs plus efficaces (voir par exemple, Fan et Gijbels 1992). Pour ce qui est des estimateurs $\hat{F}^*(t)$ et $\tilde{F}^*(t)$ fondés sur les valeurs prédites modifiées, il convient en outre de noter que l'on pourrait en principe appliquer différentes fenêtres de lissage et (ou) différents poids de régression aux valeurs y_i et aux fonctions indicatrices. Par souci de simplicité, nous ne considérerons ici ni la sélection adaptative de la fenêtre de lissage ni la possibilité de différents poids de régression pour estimer la fonction de régression à la moyenne et les distributions des composantes de l'erreur.

Si l'on compare les définitions des estimateurs fondés sur les deux types de valeurs prédites, il saute aux yeux que $\hat{F}(t)$ et $\tilde{F}(t)$ sont plus faciles à calculer puisqu'il s'agit de combinaisons linéaires des fonctions indicatrices observées $I(y_j \le t)$. Les coefficients de ces combinaisons linéaires ne dépendent pas de la variable étudiée Y et peuvent par conséquent être utilisés pour estimer les moyennes d'autres fonctions que les fonctions indicatrices, ou de fonctions de plusieurs variables étudiées, en particulier quand il y a tout lieu de croire que ces dernières sont reliées à la variable auxiliaire X. Ce fait est particulièrement précieux pour les praticiens qui veulent que les estimations reliées à plusieurs variables étudiées soient cohérentes. Néanmoins, il existe aussi un argument puissant en faveur des estimateurs $\hat{F}^*(t)$ et $\tilde{F}^*(t)$ fondés sur les valeurs prédites modifiées : si $y_i = a + bx_i$ pour tout $i \in U$, il s'ensuit que $\hat{F}^*(t) = \tilde{F}^*(t) = F_N(t)$ pour chaque échantillon s tel que les estimateurs sont bien définis. On s'attendrait donc à ce que $\hat{F}^*(t)$ et $\tilde{F}^*(t)$ soient plus efficaces que $\hat{F}(t)$ et $\tilde{F}(t)$ quand il existe une forte relation de régression entre Y et X.

3 Propriétés sous le modèle

À la présente section, nous donnons des développements asymptotiques pour le biais et la variance sous le modèle des estimateurs présentés à la section précédente. Ces développements s'appuient sur les hypothèses suivantes :

(C1) $N \to \infty$ et les suites de valeurs x_i et de plans de sondage sont telles que

$$H_{N,s}(x) := \frac{1}{n}\sum_{i \in s} I(x_i \le x)$$

et

$$H_{N,\bar{s}}(x) := \frac{1}{N-n}\sum_{i \notin s} I(x_i \le x)$$

convergent vers des fonctions de répartition absolument continues $H_s(x) := \int_a^x h_s(z)\,dz$ et $H_{\bar{s}}(x) := \int_a^x h_{\bar{s}}(z)\,dz$ respectivement. Le support de $H_s(x)$ et $H_{\bar{s}}(x)$ est donné par un intervalle borné $[a,b]$ et les dérivées premières des fonctions de densité $h_s(x)$ et $h_{\bar{s}}(x)$ sont bornées pour $x \in (a,b)$. $h_s(x)$ possède une borne inférieure strictement positive.

(C2) La fonction noyau $K(u)$ est symétrique, a pour support $[-1,1]$ et possède une dérivée bornée pour $u \in (-1,1)$. La suite de fenêtres de lissage λ tend vers zéro suffisamment lentement pour que

$$\alpha := \max\left\{\sup_{x \in [a,b]} \left|H_{N,s}(x) - H_s(x)\right|, \sup_{x \in [a,b]} \left|H_{N,\bar{s}}(x) - H_{\bar{s}}(x)\right|\right\}$$

soit d'ordre $o(\lambda)$.

(C3) Les valeurs y_i de la population sont générées à partir du modèle (2.1). La fonction $m(x)$ est telle que

$$\left| m(x) - m(x_0) - m'(x_0)(x - x_0) - \frac{1}{2} m''(x_0)(x - x_0)^2 \right| \leq C \,|\, x - x_0 \,|^{2+\delta}$$

pour un certain $\delta > 0,$ et la famille des fonctions de répartition des composantes de l'erreur $G(\varepsilon \,|\, x)$ est telle que

$$\begin{aligned} &\left| G(\varepsilon \,|\, x) - G(\varepsilon_0 | x_0) - G^{(1,0)}(\varepsilon_0 | x_0)(\varepsilon - \varepsilon_0) - G^{(0,1)}(\varepsilon_0 | x_0)(x - x_0) \right. \\ &\left. - \frac{1}{2}\left(G^{(2,0)}(\varepsilon_0 | x_0)(\varepsilon - \varepsilon_0)^2 + 2G^{(1,1)}(\varepsilon_0 | x_0)(\varepsilon - \varepsilon_0)(x - x_0) + G^{(0,2)}(\varepsilon_0 | x_0)(x - x_0)^2 \right) \right| \\ &\quad \leq C\left(|\, \varepsilon - \varepsilon_0 \,|^{2+\delta} + |\, x - x_0 \,|^{2+\delta} \right) \end{aligned}$$

pour $C > 0$ et $\delta > 0,$ où

$$G^{(r,s)}(\varepsilon \,|\, x) := \partial^{r+s} G(\varepsilon \,|\, x) \big/ \left(\partial \varepsilon^r \partial x^s \right) \quad \text{pour} \quad r, s = 0, 1, 2.$$

L'hypothèse (C1) impose une contrainte sur la façon dont les valeurs x_i dans l'échantillon et hors de celui-ci sont générées. Conjuguée à l'hypothèse (C2), elle fait en sorte que les erreurs d'estimation des estimateurs à noyau de la densité pour $h_s(x)$ et $h_{\bar{s}}(x)$ tendent vers zéro uniformément pour $x \in [a + \lambda, b - \lambda]$ et qu'elles sont bornées uniformément pour $x \in [a, b].$ Le remplacement de (C1) par des hypothèses plus précises pourrait permettre de relâcher (C2) et d'accroître la vitesse de convergence uniforme pour l'erreur d'estimation des estimateurs à noyau de la densité (voir par exemple les résultats dans Hansen 2008). Enfin, l'hypothèse (C3) est nécessaire pour que les erreurs quadratiques moyennes des deux estimateurs sous le modèle convergent vers zéro. Elle peut être relâchée au prix d'une réduction des vitesses de convergence. En plus des hypothèses (C1) à (C3), nous aurons besoin de l'hypothèse (C4) qui suit pour nous assurer que les erreurs quadratiques moyennes des estimateurs par la différence généralisée sous le modèle tendent vers zéro :

(C4) Les probabilités d'inclusion d'ordre un dans l'échantillon sont données par

$$\pi_i := n^* \frac{\pi(x_i)}{\sum_{j \in U} \pi(x_j)}, \qquad i \in U,$$

où n^* est la taille d'échantillon espérée et $\pi(x)$ est une fonction dont la borne inférieure est strictement positive et qui possède une dérivée première bornée pour $x \in (a, b).$

Proposition 1. *Sous les hypothèses (C1) à (C3), il s'ensuit que :*

$$E\left(\hat{F}(t)-F_N(t)\right) = \lambda^2 \frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b \Big[G^{(2,0)}(t-m(x)|x)\left(m'(x)\right)^2 - G^{(1,0)}(t-m(x)|x)m''(x) - 2G^{(1,1)}(t-m(x)|x)m'(x) + G^{(0,2)}(t-m(x)|x)\Big]h_{\bar{s}}(x)\,dx + o\left(\lambda^2\right)$$

et

$$\operatorname{var}\left(\hat{F}(t)-F_N(t)\right) = \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \Big[G(t-m(x)|x)-G^2(t-m(x)|x)\Big]\Big[h_{\bar{s}}(x)/h_s(x)\Big]h_{\bar{s}}(x)\,dx + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2 \int_a^b \Big[G(t-m(x)|x)-G^2(t-m(x)|x)\Big]h_{\bar{s}}(x)\,dx + o\left(n^{-1}\right),$$

où $\mu_r := \int_{-1}^{-1} K(u)u^r du$ *pour* $r = 0,1,2.$

En ajoutant l'hypothèse (C4), on peut montrer que

$$E\left(\tilde{F}(t)-F_N(t)\right) = \lambda^2 \frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b \Big[G^{(2,0)}(t-m(x)|x)\left(m'(x)\right)^2 - G^{(1,0)}(t-m(x)|x)m''(x) - 2G^{(1,1)}(t-m(x)|x)m'(x) + G^{(0,2)}(t-m(x)|x)\Big]h(x)\,dx + o\left(\lambda^2\right),$$

où

$$h(x) := h_{\bar{s}}(x) + \left(1-\pi^{-1}(x)\right)h_s(x),$$

et l'on peut montrer que

$$\operatorname{var}\left(\tilde{F}(t)-F_N(t)\right) = \operatorname{var}\left(\hat{F}(t)-F_N(t)\right) + o\left(n^{-1}\right).$$

Proposition 2. *Sous les hypothèses (C1) à (C3) et en supposant que*

i) la fonction

$$\sigma^2(x) := \int_{-\infty}^{\infty} \varepsilon^2 dG(\varepsilon|x)$$

possède une dérivée première bornée pour $x \in (a,b)$,

ii)

$$\sup_{x\in[a,b]} \int_{-\infty}^{\infty} \varepsilon^4 dG(\varepsilon|x) < \infty,$$

on peut montrer que

$$E\left(\hat{F}^*(t)-F_N(t)\right) = \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)|x)h_{\bar{s}}(x)dx$$

$$+\frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)|x)(t-m(x))h_s^{-1}(x)h_{\bar{s}}(x)dx\right.$$

$$\left.+\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)|x)\sigma^2(x)h_s^{-1}(x)h_{\bar{s}}(x)dx\right]+o\left(\lambda^2+(n\lambda)^{-1}\right),$$

où $\kappa := \int_{-1}^{1}K^2(u)du$ *et* $\theta := \int_{-1}^{1}K(v)\int_{-1}^{1}K(u+v)K(u)dudv$, *et on peut montrer que*

$$\mathrm{var}\left(\hat{F}^*(t)-F_N(t)\right)=\mathrm{var}\left(\hat{F}(t)-F_N(t)\right)+o\left(n^{-1}+\lambda^5\right).$$

En ajoutant l'hypothèse (C4), on peut également montrer que

$$E\left(\tilde{F}^*(t)-F_N(t)\right) = \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)|x)h(x)dx$$

$$+\frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)|x)(t-m(x))h_s^{-1}(x)h(x)dx\right.$$

$$\left.+\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)|x)\sigma^2(x)h_s^{-1}(x)h(x)dx\right]$$

$$+o\left(\lambda^2+(n\lambda)^{-1}\right)$$

et que

$$\mathrm{var}\left(\tilde{F}^*(t)-F_N(t)\right)=\mathrm{var}\left(\hat{F}(t)-F_N(t)\right)+o\left(n^{-1}+\lambda^5\right).$$

Les preuves des propositions sont données en annexe. Dorfman et Hall (1993) ont dérivé des développements similaires pour l'estimateur de Kuo en utilisant des poids de régression locaux constants au lieu de linéaires.

Notons qu'étant donné les développements asymptotiques, il est possible de choisir des suites de fenêtres de lissage λ de manière à être certain que les carrés des biais de modèle soient d'un ordre de grandeur inférieur aux variances sous le modèle correspondantes. Pour les estimateurs fondés sur les valeurs prédites de Kuo, cette condition est réalisée quand $\lambda = o\left(n^{-1/4}\right)$, tandis que pour les estimateurs utilisant les valeurs prédites modifiées, cela exige que λ tende vers zéro plus rapidement que $O\left(n^{-1/4}\right)$ et plus lentement que $O\left(n^{-1/2}\right)$. Les vitesses de convergence pour les biais des derniers estimateurs sous le modèle sont optimisées quand $\lambda = O\left(n^{-1/3}\right)$ et, dans ce cas, les biais sous le modèle résultants sont tous deux d'ordre $O\left(n^{-2/3}\right)$. Pour les estimateurs fondés sur les valeurs prédites de Kuo, la convergence des biais sous le modèle peut être rendue plus rapide, en fonction des suites $H_{N,s}(x)$ et $H_{N,\bar{s}}(x)$ et de la suite de fenêtres de lissage λ.

Étant donné les considérations susmentionnées concernant les biais sous le modèle et vu que les termes principaux des variances sous le modèle sont les mêmes pour les deux types de valeurs prédites, il serait intéressant de connaître les termes d'ordre deux de ces variances afin d'établir quel estimateur est le plus efficace sous l'angle de l'approche fondée sur un modèle. Les preuves présentées en annexe font toutefois penser que les termes d'ordre deux dépendent d'hypothèses plus spécifiques que (C1) à (C3) et que, en particulier pour les estimateurs fondés sur les valeurs prédites modifiées, ils sont difficiles à déterminer.

4 Propriétés sous le plan de sondage

À la section précédente, nous avons montré que les estimateurs fondés sur le modèle $\hat{F}(t)$ et $\hat{F}^*(t)$ sont asymptotiquement sans biais sous le modèle et convergents en termes d'erreur quadratique moyenne sous le modèle. Cependant, ils ne sont pas sans biais sous le plan de sondage en général et ne devraient donc pas être utilisés quand les probabilités d'inclusion dans l'échantillon ne sont pas constantes. Dans ces cas, il convient de se servir des estimateurs par la différence généralisée $\tilde{F}(t)$ et $\tilde{F}^*(t)$. En fait, il découle des résultats présentés dans Breidt et Opsomer (2000) que, sous des conditions assez générales, $\tilde{F}(t)$ est asymptotiquement sans biais sous le plan de sondage et que son erreur quadratique moyenne sous le plan est donnée par

$$E_d\left(\left|\tilde{F}(t)-F_N(t)\right|^2\right)=\frac{1}{N^2}\sum_{i,j\in U}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}\left[I(y_i\le t)-\bar{G}_i(t)\right]\left[I\left(y_j\le t\right)-\bar{G}_j(t)\right]+o\left(n^{-1}\right),$$

où $E_d(\cdot)$ désigne l'espérance par rapport au plan de sondage, $\pi_{i,j}$ désigne la probabilité d'inclusion conjointe des unités i et j dans l'échantillon (il est entendu que $\pi_{i,i}=\pi_i\big)$, et où

$$\bar{G}_i(t):=\sum_{j\in U}\bar{w}_{i,j}I\left(y_j\le t\right).$$

Les poids de régression $\bar{w}_{i,j}$ qui figurent dans la définition de $\bar{G}_i(t)$ s'appliquent à la population finie entière U et sont donnés par

$$\bar{w}_{i,j}:=\frac{1}{N\lambda}K\left(\frac{x_i-x_j}{\lambda}\right)\frac{\bar{M}_{2,s}(x_i)-\left(\frac{x_i-x_j}{\lambda}\right)\bar{M}_{1,s}(x_i)}{\bar{M}_{2,s}(x_i)\bar{M}_{0,s}(x_i)-\bar{M}_{1,s}^2(x_i)},$$

où

$$\bar{M}_{r,s}(x):=\sum_{k\in U}\frac{1}{N\lambda}K\left(\frac{x-x_k}{\lambda}\right)\left(\frac{x-x_k}{\lambda}\right)^r,\qquad r=0,1,2.$$

En outre, selon Breidt et Opsomer (2000),

$$\tilde{V}\left(\tilde{F}(t)\right) := \frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j}\left[I\left(y_i \le t\right)-\tilde{G}_i(t)\right]\left[I\left(y_j \le t\right)-\tilde{G}_j(t)\right]$$

est un estimateur convergent pour l'erreur quadratique moyenne sous le plan de $\tilde{F}(t)$.

Malheureusement, on ne peut appliquer les résultats de Breidt et Opsomer (2000) à l'estimateur par la différence généralisée $\tilde{F}^*(t)$, puisque celui-ci ne rentre pas dans la classe des estimateurs de régression par polynômes locaux en raison de la présence des estimateurs des fonctions de régression $\tilde{m}_i$ et $\tilde{m}_j$ à l'intérieur des fonctions indicatrices dans les valeurs prédites $\tilde{G}_i^*(t)$. Cependant, les résultats pour $\tilde{F}(t)$ donnent à penser que, dans les grands échantillons, $\tilde{G}_i^*(t)$ et

$$\bar{G}_i^*(t) := \sum_{j\in U}\bar{w}_{i,j}I\left(y_j-\bar{m}_j \le t-\bar{m}_i\right),$$

où les $\bar{m}_i := \sum_{j\in U}\bar{w}_{i,j}y_j$, sont approximativement les mêmes et que

$$E_d\left(\left|\tilde{F}^*(t)-F_N(t)\right|^2\right) = \frac{1}{N^2}\sum_{i,j\in U}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}\left[I\left(y_i \le t\right)-\bar{G}_i^*(t)\right]\left[I\left(y_j \le t\right)-\bar{G}_j^*(t)\right]+o\left(n^{-1}\right).$$

Partant de cette conjecture, nous avons testé

$$\tilde{V}\left(\tilde{F}^*(t)\right) := \frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j}\left[I\left(y_i \le t\right)-\tilde{G}_i^*(t)\right]\left[I\left(y_j \le t\right)-\tilde{G}_j^*(t)\right]$$

comme estimateur pour l'erreur quadratique moyenne sous le plan de l'estimateur par la différence généralisée $\tilde{F}^*(t)$ dans l'étude en simulation décrite à la section suivante.

5 Étude en simulation

À la présente section, nous analysons certains résultats de simulation. Notre objectif est de comparer l'efficacité par rapport au plan de sondage des estimateurs des fonctions de répartition présentés à la section 2 et des estimateurs de la variance présentés à la section 4. Les résultats des simulations s'appliquent à l'échantillonnage aléatoire simple sans remise et à l'échantillonnage de Poisson avec probabilités d'inclusion inégales. À titre de référence, nous avons également inclus dans l'étude en simulation l'estimateur de la fonction de répartition de Horvitz-Thompson

$$\hat{F}_\pi(t) := \frac{1}{N}\sum_{j\in s}\pi_j^{-1}I\left(y_j \le t\right)$$

et l'estimateur de variance correspondant

$$\tilde{V}\left(\hat{F}_\pi(t)\right) := \frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j}I\left(y_i \le t\right)I\left(y_j \le t\right).$$

Nous avons considéré des populations artificielles ainsi que réelles. Les premières ont été obtenues en générant $N = 1000$ valeurs x_i à partir de variables aléatoires i.i.d. de loi uniforme avec support sur l'intervalle $(0,1)$ et en les combinant avec trois types de fonction de régression $m(x)$ et deux types de composantes de l'erreur ε_i. Les fonctions de régression sont i) $m(x) = 0$ (uniforme), ii) $m(x) = 10x$ (linéaire) et iii) $m(x) = 10x^{1/4}$ (concave), tandis que les composantes de l'erreur ε_i sont soit des réalisations indépendantes tirées d'une loi t de Student unique à $\nu = 5$ dl, ou des réalisations indépendantes tirées de N lois t de Student non centrales décalées à $\nu = 5$ dl et avec paramètres de non-centralité donnés par $\mu = 15x_i$. Les décalages appliqués aux composantes de l'erreur dans le dernier cas font en sorte que les moyennes des lois t de Student non centrales à partir desquelles elles sont générées soient nulles. Les populations artificielles sont présentées aux figures 5.1 à 5.3. En ce qui concerne les populations réelles, nous avons pris la population $MU284$ de municipalités suédoises de Särndal et coll. (1992) (taille de la population $N = 284$) et considéré le logarithme naturel de $RMT85 =$ Revenus de l'imposition municipale de 1985 (en millions de couronnes) comme variable étudiée Y, et le logarithme naturel de $P85 =$ population de 1985 (en milliers) ou de $REV84 =$ valeurs immobilières selon les évaluations de 1984 (en millions de couronnes) comme variable auxiliaire X. Les populations réelles sont présentées à la figure 5.4.

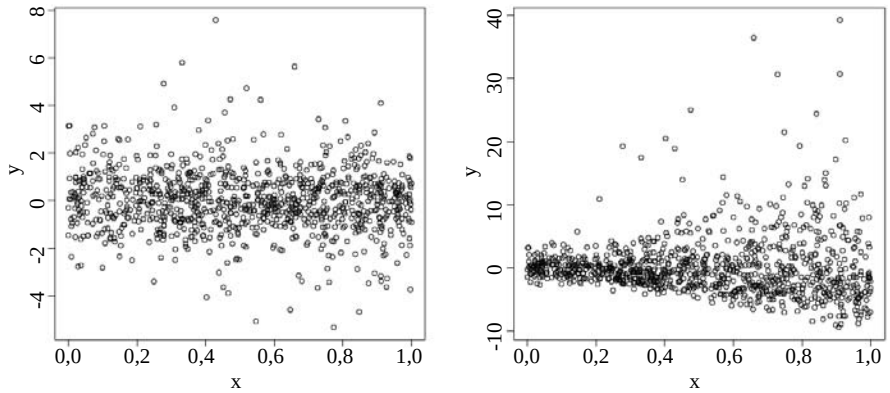

Figure 5.1 Populations générées à partir de $y_i = \varepsilon_i$, où $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$ (à gauche) et $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$ (à droite).

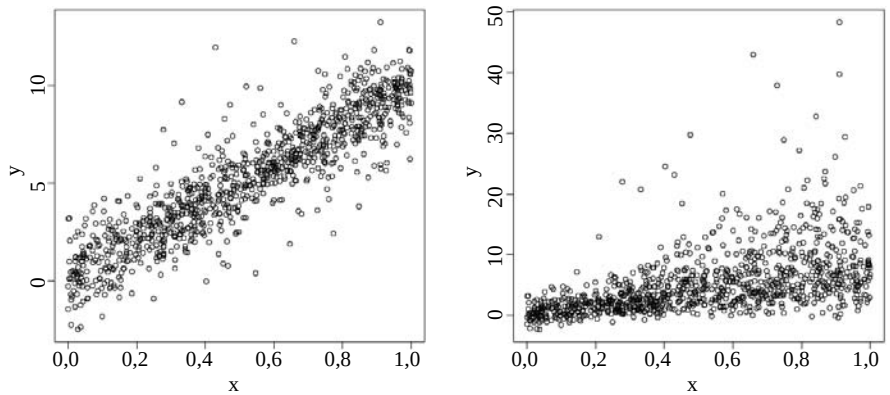

Figure 5.2 Populations générées à partir de $y_i = 10x_i + \varepsilon_i$, où $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$ (à gauche) et $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$ (à droite).

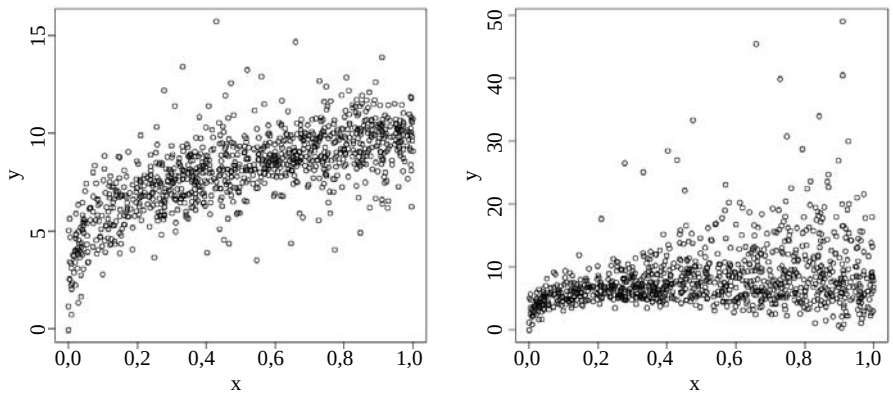

Figure 5.3 Populations générées à partir de $y_i = 10x_i^{1/4} + \varepsilon_i$, où $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$ (à gauche) et $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$ (à droite).

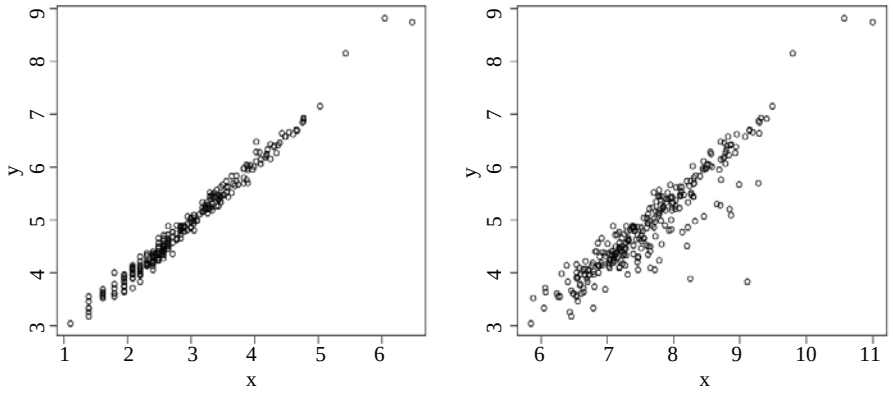

Figure 5.4 Population MU284 de municipalités suédoises de Särndal et coll. (1992). $y_i = \ln RMT85_i$ pour la i^e municipalité, et $x_i = \ln P85_i$ (à gauche) ou $x_i = \ln REV84_i$ (à droite).

Pour chaque population, nous avons sélectionné indépendamment $B = 1\,000$ échantillons. Pour le tirage d'échantillons à partir des populations artificielles, en cas d'échantillonnage aléatoire simple sans remise, nous avons fixé la taille d'échantillon à $n = 100,$ et en cas d'échantillonnage de Poisson, nous avons fixé la taille d'échantillon espérée à $n^* = 100$ et fait en sorte que les probabilités d'inclusion dans l'échantillon soient proportionnelles aux écarts-types des lois t de Student non centrales décalées susmentionnées. Pour le tirage d'échantillons dans les populations réelles, nous avons fixé la taille d'échantillon à $n = 30$ en cas d'échantillonnage aléatoire simple sans remise. Pour l'échantillonnage de Poisson, nous avons fixé la taille d'échantillon espérée à $n^* = 30$ et fait en sorte que les probabilités d'inclusion dans l'échantillon soient proportionnelles aux valeurs absolues des résidus des régressions linéaires par les moindres carrés des valeurs y_i de la population sur les valeurs x_i de la population.

Comme pour la définition des estimateurs non paramétriques, nous avons utilisé la fonction noyau d'Epanechnikov $K(u) := 0{,}75\left(1-u^2\right)$ avec $\lambda = 0{,}15$ ou $\lambda = 0{,}3$ pour les échantillons tirés des populations artificielles et la fonction noyau gaussienne $K(u) := 1/\sqrt{2\pi}e^{-(1/2)u^2}$ avec $\lambda = 1$ ou $\lambda = 2$ pour les échantillons tirés des populations réelles. Dans les tableaux présentant les résultats des simulations, les estimateurs non paramétriques correspondant aux petites et aux grandes valeurs de fenêtre de lissage sont désignés par un s (pour *small*) ou par un l (pour *large*), respectivement, dans l'indice inférieur. Nous avons recouru à la fonction noyau gaussienne pour les échantillons tirés des populations réelles afin d'éviter les problèmes de singularité qui se posent en cas de vides dans le jeu de valeurs x_i échantillonnées. De tels vides sont nettement plus susceptibles d'exister dans le cas des populations réelles que dans celui des populations artificielles, parce que les lois des variables auxiliaires sont asymétriques dans les premières. En fait, dans les populations artificielles, les estimateurs non paramétriques étaient bien définis pour chacun des $B = 1\,000$ échantillons sélectionnés selon le plan d'échantillonnage aléatoire simple sans remise. Pour le plan d'échantillonnage de Poisson, par contre, 47 des $B = 1\,000$ échantillons simulés étaient tels que les estimateurs non paramétriques avec la petite valeur de fenêtre de lissage n'ont pas pu être calculés et seulement un de ces échantillons était tel que les estimateurs non paramétriques avec la grande valeur de fenêtre de lissage étaient indéfinis. Les résultats des simulations s'appliquant aux estimateurs non paramétriques dans les tableaux 5.2 et 5.5 tiennent compte uniquement des échantillons pour lesquels les estimateurs étaient bien définis et sont donc fondés sur un peu moins que les $B = 1\,000$ réalisations.

Les tableaux 5.1 à 5.4 donnent le biais simulé (BIAIS) et la racine carrée de l'erreur quadratique moyenne simulée (REQM) pour chaque estimateur de la fonction de répartition à différents niveaux de t auxquels $F_N(t)$ a été estimée : en se basant, par exemple, sur les valeurs $\tilde{F}_b(t)$, $b = 1,2,\ldots,B,$ tirées de l'estimateur $\tilde{F}(t)$,

$$\text{BIAIS} := \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right) \times 10\,000$$

et

$$\text{REQM} := \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right)^2} \times 10\,000.$$

La REQM montre que les estimateurs fondés sur les valeurs prédites modifiées sont habituellement plus efficaces. Dans le cas de l'échantillonnage dans les populations réelles, l'augmentation des REQM est parfois assez grande. Comme prévu, les estimateurs fondés sur le modèle ont tendance à être plus efficaces que les estimateurs par la différence généralisée sous échantillonnage aléatoire simple sans remise quand les deux types d'estimateurs sont approximativement sans biais. Sous échantillonnage de Poisson, le BIAIS des estimateurs fondés sur le modèle augmente, mais demeure néanmoins concurrentiel. Une plus grande variabilité des probabilités d'inclusion dans l'échantillon modifierait certainement ce résultat, car elle augmenterait le BIAIS des estimateurs fondés sur le modèle. Les résultats des simulations ne doivent donc pas être considérés comme contredisant Johnson, Breidt et Opsomer (2008) qui se prononcent en faveur des estimateurs par la différence généralisée (appelés estimateurs assistés par modèle dans leur article), soutenant qu'il s'agit d'« un bon choix global pour les estimateurs de la fonction de répartition ».

Tableau 5.1
Populations artificielles (taille de population $N = 1\,000$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student centrale avec $\nu = 5$										
$\hat{F}_s(t)$	6	216	-3	433	31	512	23	434	12	207
$\hat{F}_l(t)$	15	219	10	430	0	502	-10	429	3	213
$\hat{F}_s^*(t)$	6	209	-30	411	22	484	22	414	3	200
$\hat{F}_l^*(t)$	15	214	-9	409	10	477	1	407	-10	207
$\tilde{F}_s(t)$	6	213	8	425	24	504	-4	430	8	207
$\tilde{F}_l(t)$	6	210	10	417	22	494	-8	422	6	206
$\tilde{F}_s^*(t)$	8	213	9	426	25	503	-5	432	5	206
$\tilde{F}_l^*(t)$	7	210	10	417	23	494	-6	424	4	206
$\tilde{F}_\pi(t)$	7	208	11	411	19	489	-5	417	6	200
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	26	225	33	376	8	477	26	419	33	209
$\hat{F}_l(t)$	52	236	23	374	-5	475	38	421	29	213
$\hat{F}_s^*(t)$	20	195	-29	351	-89	471	11	407	30	202
$\hat{F}_l^*(t)$	36	201	-11	357	-94	473	28	410	21	204
$\tilde{F}_s(t)$	8	211	11	370	-7	473	4	415	16	211
$\tilde{F}_l(t)$	5	208	8	367	-5	468	5	411	16	212
$\tilde{F}_s^*(t)$	11	210	11	372	-11	475	4	416	15	210
$\tilde{F}_l^*(t)$	7	208	11	368	-7	468	8	412	15	211
$\tilde{F}_\pi(t)$	1	211	1	391	-6	477	8	399	18	210
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\hat{F}_s(t)$	32	201	25	275	13	250	-14	264	-36	217
$\hat{F}_l(t)$	114	250	152	304	12	236	-180	312	-86	242
$\hat{F}_s^*(t)$	-50	165	12	226	51	216	26	230	13	172
$\hat{F}_l^*(t)$	-46	155	-14	199	69	195	23	211	17	156
$\tilde{F}_s(t)$	-5	186	4	275	15	248	11	269	-2	201
$\tilde{F}_l(t)$	-5	184	7	274	17	250	5	269	-2	196
$\tilde{F}_s^*(t)$	-10	180	5	275	16	245	14	266	-1	200
$\tilde{F}_l^*(t)$	-9	176	3	272	15	242	13	262	-1	194
$\tilde{F}_\pi(t)$	-7	203	14	413	37	472	17	405	1	206
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	24	204	23	351	27	403	26	382	29	208
$\hat{F}_l(t)$	94	242	135	372	51	392	13	380	15	212
$\hat{F}_s^*(t)$	55	182	-9	301	-18	368	-23	359	37	202
$\hat{F}_l^*(t)$	124	210	-31	278	-63	363	-8	356	48	200
$\tilde{F}_s(t)$	-2	194	-4	349	11	401	18	377	13	208
$\tilde{F}_l(t)$	-2	190	-5	345	12	398	17	374	11	209
$\tilde{F}_s^*(t)$	0	191	-5	352	14	401	20	376	13	207
$\tilde{F}_l^*(t)$	-1	189	-6	344	13	397	18	375	12	209
$\tilde{F}_\pi(t)$	-4	205	-5	401	21	470	24	401	14	207
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\hat{F}_s(t)$	81	207	44	316	17	384	-2	376	23	203
$\hat{F}_l(t)$	138	258	183	356	35	367	-50	374	8	208
$\hat{F}_s^*(t)$	7	146	-14	274	16	352	-8	358	15	197
$\hat{F}_l^*(t)$	9	144	10	246	-2	323	-18	339	24	186
$\tilde{F}_s(t)$	3	175	3	319	10	383	17	374	10	203
$\tilde{F}_l(t)$	0	178	5	316	11	380	17	370	8	202
$\tilde{F}_s^*(t)$	1	167	5	320	12	383	17	374	9	203
$\tilde{F}_l^*(t)$	-1	164	6	316	13	379	20	368	8	201
$\tilde{F}_\pi(t)$	4	209	11	412	25	477	27	422	10	200

Tableau 5.1 (suite)
Populations artificielles (taille de population $N = 1\,000$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	59	234	95	402	66	455	51	395	26	208
$\hat{F}_l(t)$	94	259	190	441	147	467	98	400	16	212
$\hat{F}_s^*(t)$	30	184	33	343	-123	435	-34	385	40	203
$\hat{F}_l^*(t)$	57	201	58	331	-148	437	2	382	34	203
$\tilde{F}_s(t)$	1	205	7	386	12	449	17	392	13	208
$\tilde{F}_l(t)$	-1	204	0	385	9	445	20	389	11	209
$\tilde{F}_s^*(t)$	3	201	8	389	7	449	13	392	14	207
$\tilde{F}_l^*(t)$	0	198	6	383	9	446	19	390	13	208
$\tilde{F}_\pi(t)$	0	205	-2	399	9	463	25	398	14	208

Tableau 5.2
Populations artificielles (taille de population $N = 1\,000$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage de Poisson avec probabilités d'inclusion dans l'échantillon π_i proportionnelles aux écarts-types des lois t de Student non centrales avec $\nu = 5$ dl et avec paramètres de non-centralité $\mu = 15x_i$. Taille espérée d'échantillon $n^* = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student centrale avec $\nu = 5$										
$\hat{F}_s(t)$	-10	252	-11	593	-22	738	-20	743	6	357
$\hat{F}_l(t)$	-1	237	9	543	-15	621	-5	590	11	302
$\hat{F}_s^*(t)$	22	244	-29	485	-3	555	9	515	-17	297
$\hat{F}_l^*(t)$	14	238	-10	492	-5	564	14	524	-1	283
$\tilde{F}_s(t)$	-6	247	0	579	-27	724	-40	736	3	349
$\tilde{F}_l(t)$	-2	231	11	526	-1	598	-10	566	7	285
$\tilde{F}_s^*(t)$	23	248	23	505	-4	562	-27	531	-20	304
$\tilde{F}_l^*(t)$	12	240	20	504	1	573	-13	538	-6	287
$\tilde{F}_\pi(t)$	-6	220	-7	543	-37	741	-44	929	-48	1 058
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	17	164	30	411	4	749	14	590	15	190
$\hat{F}_l(t)$	47	173	19	383	-1	602	57	498	15	187
$\hat{F}_s^*(t)$	21	175	-7	378	-89	554	-11	473	3	192
$\hat{F}_l^*(t)$	29	152	-3	367	-99	555	27	481	3	184
$\tilde{F}_s(t)$	1	159	10	406	-11	737	-5	579	-2	194
$\tilde{F}_l(t)$	1	158	9	388	-5	586	14	482	-1	192
$\tilde{F}_s^*(t)$	14	186	27	409	-3	562	-17	487	-10	200
$\tilde{F}_l^*(t)$	3	160	22	399	-11	566	-5	482	-2	193
$\tilde{F}_\pi(t)$	-3	162	-7	451	-31	738	-29	980	-55	1 067
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\hat{F}_s(t)$	8	461	21	561	-12	259	-18	218	-30	164
$\hat{F}_l(t)$	78	429	183	451	2	248	-161	261	-79	189
$\hat{F}_s^*(t)$	-69	306	12	340	10	267	15	199	6	143
$\hat{F}_l^*(t)$	-59	294	4	302	56	205	15	172	17	124
$\tilde{F}_s(t)$	-25	441	4	560	-10	257	9	219	5	153
$\tilde{F}_l(t)$	-14	372	35	410	-10	262	4	219	5	151
$\tilde{F}_s^*(t)$	-31	333	-2	386	-29	294	4	227	-1	161
$\tilde{F}_l^*(t)$	-20	339	15	372	-10	259	11	215	4	151
$\tilde{F}_\pi(t)$	-15	385	3	746	-37	917	-35	1 004	-48	1 070

Tableau 5.2 (suite)
Populations artificielles (taille de population $N = 1\,000$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage de Poisson avec probabilités d'inclusion dans l'échantillon π_i proportionnelles aux écarts-types des lois t de Student non centrales avec $\nu = 5$ dl et avec paramètres de non-centralité $\mu = 15x_i$. Taille espérée d'échantillon $n^* = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	-4	516	30	671	7	453	11	344	6	182
$\hat{F}_l(t)$	63	409	129	539	61	421	9	341	1	180
$\hat{F}_s^*(t)$	44	300	-29	433	-45	422	-47	345	12	180
$\hat{F}_l^*(t)$	107	314	-41	420	-60	397	-22	323	31	171
$\tilde{F}_s(t)$	-27	502	8	667	-8	450	0	344	-8	185
$\tilde{F}_l(t)$	-10	364	16	510	11	425	-2	345	-7	182
$\tilde{F}_s^*(t)$	-6	325	-9	479	-25	447	-14	356	-10	187
$\tilde{F}_l^*(t)$	-7	332	-9	489	-5	426	-3	344	-6	182
$\tilde{F}_\pi(t)$	-16	349	-2	705	-21	886	-42	1 013	-61	1 069
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\hat{F}_s(t)$	36	497	47	629	9	418	-11	320	15	191
$\hat{F}_l(t)$	56	393	186	490	43	383	-48	308	13	184
$\hat{F}_s^*(t)$	-29	276	-19	383	-18	380	-43	335	-1	204
$\hat{F}_l^*(t)$	-29	274	10	355	7	336	-29	290	23	179
$\tilde{F}_s(t)$	-30	475	12	630	4	421	7	317	6	191
$\tilde{F}_l(t)$	-42	336	31	452	11	390	8	312	8	186
$\tilde{F}_s^*(t)$	-31	306	5	429	-18	406	-14	344	-8	210
$\tilde{F}_l^*(t)$	-28	308	14	424	7	387	5	315	7	191
$\tilde{F}_\pi(t)$	-15	380	10	739	-23	891	-37	993	-47	1 064
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	24	308	69	687	53	690	38	406	2	188
$\hat{F}_l(t)$	47	301	131	553	139	561	91	393	-2	186
$\hat{F}_s^*(t)$	15	237	2	435	-135	513	-59	411	12	186
$\hat{F}_l^*(t)$	27	235	18	435	-149	506	-5	374	13	179
$\tilde{F}_s(t)$	-28	274	-8	673	4	688	3	403	-10	191
$\tilde{F}_l(t)$	-29	251	-12	512	17	541	7	395	-9	188
$\tilde{F}_s^*(t)$	-3	255	-12	481	-7	536	-20	422	-12	196
$\tilde{F}_l^*(t)$	-12	251	-16	489	2	538	-4	399	-9	189
$\tilde{F}_\pi(t)$	-10	267	-8	608	-4	860	-38	1 009	-63	1 066

Tableau 5.3
Populations réelles (taille de population $N = 284$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAISR**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln P85$										
$\hat{F}_s(t)$	133	421	339	625	180	529	-265	490	-187	439
$\hat{F}_l(t)$	52	380	67	588	45	555	-63	469	-87	370
$\hat{F}_s^*(t)$	8	81	-154	203	90	130	62	123	6	54
$\hat{F}_l^*(t)$	28	66	-170	212	69	112	57	109	2	50
$\tilde{F}_s(t)$	-28	300	-24	497	8	483	-48	421	-38	319
$\tilde{F}_l(t)$	-28	326	-96	569	-52	544	3	466	1	319
$\tilde{F}_s^*(t)$	26	177	-11	302	0	244	1	308	-18	102
$\tilde{F}_l^*(t)$	29	179	-10	302	-2	243	-1	308	-21	104
$\tilde{F}_\pi(t)$	22	388	-10	771	9	864	5	731	-43	394

Tableau 5.3 (suite)
Populations réelles (taille de population $N = 284$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAISR**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln REV84$										
$\hat{F}_s(t)$	143	449	303	643	138	554	-217	543	-166	446
$\hat{F}_l(t)$	62	395	62	611	36	582	-49	519	-71	376
$\hat{F}_s^*(t)$	-11	204	-32	300	-101	328	42	285	31	155
$\hat{F}_l^*(t)$	36	183	-40	288	-149	345	6	261	34	122
$\tilde{F}_s(t)$	5	340	-22	548	4	557	-30	498	-23	332
$\tilde{F}_l(t)$	-2	349	-78	599	-36	588	10	522	8	331
$\tilde{F}_s^*(t)$	24	303	7	446	-6	494	2	439	-13	209
$\tilde{F}_l^*(t)$	29	304	4	443	-6	495	-1	432	-18	192
$\tilde{F}_\pi(t)$	34	395	1	766	16	880	9	744	-37	398

Tableau 5.4
Populations réelles (taille de population $N = 284$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage de Poisson avec probabilités d'inclusion proportionnelles à la valeur absolue des résidus de la régression linéaire des valeurs y_i de la population sur les valeurs x_i de la population. Taille espérée $n^* = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAISR**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln P85$										
$\hat{F}_s(t)$	204	420	485	668	239	519	-412	626	-90	317
$\hat{F}_l(t)$	180	424	417	684	319	614	-239	548	-148	348
$\hat{F}_s^*(t)$	-41	97	-118	199	132	178	40	140	-71	104
$\hat{F}_l^*(t)$	11	70	-147	211	63	128	-25	122	-85	106
$\tilde{F}_s(t)$	24	360	30	649	0	675	-68	614	58	368
$\tilde{F}_l(t)$	9	390	-63	737	-64	774	-7	682	75	414
$\tilde{F}_s^*(t)$	16	184	-14	307	36	283	16	323	-11	103
$\tilde{F}_l^*(t)$	25	187	-15	312	30	286	14	328	-11	112
$\tilde{F}_\pi(t)$	40	445	73	1 983	12	2 498	-43	3 094	-49	3 341
Population MU284 avec $Y = \ln RMT85$ et $X = \ln REV84$										
$\hat{F}_s(t)$	349	660	1 185	1 373	890	1 059	458	654	-32	270
$\hat{F}_l(t)$	287	601	1 003	1 236	771	989	484	695	42	263
$\hat{F}_s^*(t)$	317	453	739	866	761	879	624	701	159	207
$\hat{F}_l^*(t)$	364	471	720	842	718	824	572	647	96	158
$\tilde{F}_s(t)$	35	488	82	818	-31	772	7	634	-8	326
$\tilde{F}_l(t)$	22	500	3	878	-98	852	40	704	27	354
$\tilde{F}_s^*(t)$	37	317	32	498	-13	513	32	412	7	157
$\tilde{F}_l^*(t)$	51	313	30	498	-30	518	12	411	-10	149
$\tilde{F}_\pi(t)$	32	671	19	1 658	-172	2 354	-173	2 787	-191	2 935

Considérons enfin les résultats des simulations concernant les estimateurs de variance de la section 4. Les tableaux 5.5 à 5.8 donnent le biais relatif (BIAISR) et la racine carrée de l'erreur quadratique moyenne relative (REQMR) pour chacun d'eux. Par exemple, selon les estimations de variance $\tilde{V}_b\left(\tilde{F}(t)\right)$, $b = 1, 2, \ldots, B$, obtenues au moyen de l'estimateur $\tilde{V}\left(\tilde{F}(t)\right)$,

$$\text{BIAISR} := \frac{1}{B}\sum_{b=1}^{B}\frac{\tilde{V}_b\left(\tilde{F}(t)\right) - V_B\left(\tilde{F}(t)\right)}{V_B\left(\tilde{F}(t)\right)} \times 10\,000$$

et

$$\text{REQMR} := \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{V}_b\left(\tilde{F}(t)\right) - V_B\left(\tilde{F}(t)\right)\right)^2}}{V_B\left(\tilde{F}(t)\right)} \times 10\,000$$

où

$$V_B\left(\tilde{F}(t)\right) := \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right)^2.$$

À titre de référence, nous donnons également les BIAISR et REQMR de l'estimateur

$$\tilde{V}\left(\tilde{F}_\pi(t)\right) := \frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j} - \pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j} I(y_i \le t) I\left(y_j \le t\right)$$

pour la variance de l'estimateur de Horvitz-Thompson.

Tableau 5.5
Populations artificielles (taille de population $N = 1\,000$). BIAISR et REQMR des estimateurs de variance sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student centrale avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-1 092	32 442	-1 249	3 895	-1 714	3 077	-1 536	3 828	-824	34 601
$\tilde{V}(\tilde{F}_l(t))$	-576	31 726	-603	3 838	-1 122	3 374	-951	3 758	-441	33 055
$\tilde{V}(\tilde{F}_s^*(t))$	-1 091	32 579	-1 292	3 914	-1 708	3 085	-1 640	3 828	-802	34 809
$\tilde{V}(\tilde{F}_l^*(t))$	-556	31 881	-622	3 857	-1 148	3 361	-1 025	3 749	-425	33 184
$\tilde{V}(\tilde{F}_\pi(t))$	42	30 952	57	3 928	-592	3 776	-287	3 825	551	33 462
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1 900	29 622	50	4 707	-917	3 557	-998	3 695	-1 480	29 417
$\tilde{V}(\tilde{F}_l(t))$	-1 359	29 623	535	4 572	-395	3 881	-527	3 736	-1 277	28 267
$\tilde{V}(\tilde{F}_s^*(t))$	-1 832	30 119	-101	4 710	-991	3 530	-1 077	3 704	-1 398	29 927
$\tilde{V}(\tilde{F}_l^*(t))$	-1 362	29 713	465	4 559	-420	3 865	-591	3 718	-1 236	28 489
$\tilde{V}(\tilde{F}_\pi(t))$	-351	29 132	1 096	4 215	-78	4 074	574	4 067	-638	29 507
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-2 170	11 624	-1 027	2 480	-816	3 274	-1 424	2 583	-1 946	8 681
$\tilde{V}(\tilde{F}_l(t))$	-1 534	11 605	-529	2 632	-148	2 975	-859	2 590	-1 151	9 015
$\tilde{V}(\tilde{F}_s^*(t))$	-1 765	12 107	-1 108	2 529	-714	3 366	-1 318	2 660	-1 905	8 658
$\tilde{V}(\tilde{F}_l^*(t))$	-1 062	11 948	-671	2 735	-212	3 291	-762	2 785	-1 048	8 590
$\tilde{V}(\tilde{F}_\pi(t))$	254	31 545	-52	3 726	136	4 152	267	3 992	35	30 264
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1 642	25 809	-855	3 541	-1 076	3 038	-1 081	3 030	-1 361	21 157
$\tilde{V}(\tilde{F}_l(t))$	-950	25 692	-323	3 509	-597	3 312	-617	3 164	-1 124	20 231
$\tilde{V}(\tilde{F}_s^*(t))$	-1 385	26 406	-997	3 505	-1 089	3 045	-1 096	3 033	-1 310	21 393
$\tilde{V}(\tilde{F}_l^*(t))$	-832	26 212	-292	3 556	-614	3 317	-716	3 154	-1 135	20 286
$\tilde{V}(\tilde{F}_\pi(t))$	105	29 621	507	3 857	209	4 244	425	3 910	-337	29 082

Tableau 5.5 (suite)
Populations artificielles (taille de population $N = 1\,000$). BIAISR et REQMR des estimateurs de variance sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-2 465	30 612	-1 121	4 594	-1 512	3 183	-1 958	3 076	-863	19 720
$\tilde{V}(\tilde{F}_l(t))$	-1 780	28 103	-663	4 420	-1 092	3 319	-1 491	3 140	-439	18 985
$\tilde{V}(\tilde{F}_s^*(t))$	-2 052	33 980	-1 150	4 619	-1 537	3 217	-1 948	3 127	-954	19 637
$\tilde{V}(\tilde{F}_l^*(t))$	-1 194	33 573	-691	4 472	-1 124	3 368	-1 438	3 228	-357	19 245
$\tilde{V}(\tilde{F}_\pi(t))$	-81	30 001	9	3 756	-110	3 996	-598	3 661	440	32 455
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1 873	29 437	-758	3 759	-621	3 476	-709	3 599	-1 298	27 679
$\tilde{V}(\tilde{F}_l(t))$	-1 267	28 511	-284	3 661	-131	3 758	-321	3 552	-1 075	26 790
$\tilde{V}(\tilde{F}_s^*(t))$	-1 710	30 670	-928	3 741	-628	3 510	-777	3 603	-1 245	27 972
$\tilde{V}(\tilde{F}_l^*(t))$	-939	30 486	-270	3 764	-171	3 803	-375	3 581	-1 014	26 926
$\tilde{V}(\tilde{F}_\pi(t))$	178	29 640	599	3 816	533	4 324	590	3 874	-404	28 917

Tableau 5.6
Populations artificielles (taille de population $N = 1\,000$). BIAISR et REQMR des estimateurs de variance sous échantillonnage de Poisson avec probabilités d'inclusion dans l'échantillon π_i proportionnelles aux écarts-types des lois t de Student non centrale avec $\nu = 5$ dl et avec paramètre de non-centralité $\mu = 15x_i$. Taille espérée d'échantillon $n^* = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student centrale avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-3 306	65 777	-4 248	8 032	-5 093	4 242	-6 258	4 844	-5 652	32 037
$\tilde{V}(\tilde{F}_l(t))$	-2 048	47 035	-2 656	4 705	-2 434	3 116	-3 310	3 939	-3 092	29 380
$\tilde{V}(\tilde{F}_s^*(t))$	-3 362	36 855	-2 488	4 409	-1 910	3 147	-2 869	3 910	-4 329	23 247
$\tilde{V}(\tilde{F}_l^*(t))$	-2 696	39 509	-2 076	4 450	-1 768	3 163	-2 648	3 811	-3 244	26 343
$\tilde{V}(\tilde{F}_\pi(t))$	113	129 637	259	15 120	618	6 327	193	5 429	273	6 097
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-740	125 975	-2 522	14 864	-5 466	3 658	-4 896	6 691	-1 551	83 262
$\tilde{V}(\tilde{F}_l(t))$	-391	83 047	-1 503	8 946	-2 428	4 099	-2 228	5 526	-1 154	54 680
$\tilde{V}(\tilde{F}_s^*(t))$	-3 260	58 072	-2 649	7 661	-2 260	3 936	-2 795	5 011	-2 116	48 739
$\tilde{V}(\tilde{F}_l^*(t))$	-716	77 935	-2 000	7 979	-1 934	4 235	-2 279	5 243	-1 243	52 531
$\tilde{V}(\tilde{F}_\pi(t))$	666	251 134	-564	26 553	-87	7 344	-2	6 029	407	6 610
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-6 801	7 898	-6 470	4 281	-1 059	22 596	-398	32 401	-1 650	72 632
$\tilde{V}(\tilde{F}_l(t))$	-4 978	5 826	-2 898	4 473	-603	9 530	206	15 226	-1 157	40 466
$\tilde{V}(\tilde{F}_s^*(t))$	-4 520	6 691	-2 710	4 213	-3 245	6 723	-1 156	12 681	-2 458	32 907
$\tilde{V}(\tilde{F}_l^*(t))$	-4 226	6 206	-1 674	5 062	-978	7 874	55	12 781	-1 283	33 737
$\tilde{V}(\tilde{F}_\pi(t))$	-707	47 550	118	7 214	609	4 409	743	4 628	435	4 800

Tableau 5.6 (suite)
Populations artificielles (taille de population $N = 1\,000$). BIAISR et REQMR des estimateurs de variance sous échantillonnage de Poisson avec probabilités d'inclusion dans l'échantillon π_i proportionnelles aux écarts-types des lois t de Student non centrale avec $\nu = 5$ dl et avec paramètre de non-centralité $\mu = 15x_i$. Taille espérée d'échantillon $n^* = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-7 398	8 847	-6 235	3 667	-2 493	8 171	-1 051	16 299	-1 440	71 943
$\tilde{V}(\tilde{F}_l(t))$	-4 548	9 463	-3 136	3 282	-1 187	4 246	-832	7 638	-982	45 182
$\tilde{V}(\tilde{F}_s^*(t))$	-3 902	11 727	-2 808	3 409	-2 411	3 501	-1 721	6 737	-1 671	41 389
$\tilde{V}(\tilde{F}_l^*(t))$	-3 598	10 771	-2 610	3 462	-1 284	3 988	-852	7 008	-972	43 017
$\tilde{V}(\tilde{F}_\pi(t))$	146	57 044	-42	8 708	520	4 784	214	4 686	390	5 085
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-7 731	8 568	-6 597	3 484	-2 442	7 775	-903	16 067	-1 967	56 480
$\tilde{V}(\tilde{F}_l(t))$	-4 611	9 378	-2 990	3 252	-874	4 119	-347	7 420	-1 310	35 051
$\tilde{V}(\tilde{F}_s^*(t))$	-4 747	11 909	-2 679	3 298	-1 896	3 272	-2 248	5 747	-3 382	27 222
$\tilde{V}(\tilde{F}_l^*(t))$	-4 223	10 380	-2 100	3 494	-788	3 731	-550	5 975	-1 795	29 856
$\tilde{V}(\tilde{F}_\pi(t))$	-428	47 038	-206	7 350	641	4 504	738	4 708	487	4 943
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-4 936	40 696	-6 111	4 579	-5 549	4 035	-1 864	14 381	-1 509	84 892
$\tilde{V}(\tilde{F}_l(t))$	-3 004	29 404	-2 764	3 962	-2 436	3 606	-1 234	7 357	-1 103	53 875
$\tilde{V}(\tilde{F}_s^*(t))$	-4 328	27 704	-2 516	4 235	-2 671	3 332	-2 586	5 955	-1 939	47 601
$\tilde{V}(\tilde{F}_l^*(t))$	-3 454	28 267	-2 263	4 160	-2 329	3 574	-1 433	6 682	-1 171	50 985
$\tilde{V}(\tilde{F}_\pi(t))$	152	98 607	663	12 879	15	5 376	20	5 080	429	5 619

Tableau 5.7
Populations réelles (taille de population $N = 284$). BIAISR et REQMR des estimateurs de variance sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln P85$										
$\tilde{V}(\tilde{F}_s(t))$	-2 853	16 809	-1 700	3 037	-1 554	2 984	-1 100	4 633	-5 503	16 257
$\tilde{V}(\tilde{F}_l(t))$	-1 110	16 374	-1 827	2 760	-1 683	2 847	-927	4 387	-3 016	18 685
$\tilde{V}(\tilde{F}_s^*(t))$	-1 043	19 081	-91	7 728	-448	9 120	-484	7 715	-1 877	65 298
$\tilde{V}(\tilde{F}_l^*(t))$	-424	18 971	104	7 819	-382	9 110	-301	7 799	-1 058	62 968
$\tilde{V}(\tilde{F}_\pi(t))$	-186	29 720	-603	3 901	31	3 971	500	4 383	-74	28 418
Population MU284 avec $Y = \ln RMT85$ et $X = \ln REV84$										
$\tilde{V}(\tilde{F}_s(t))$	-2 283	16 303	-1 450	3 538	-945	3 526	-1 071	4 300	-4 832	19 401
$\tilde{V}(\tilde{F}_l(t))$	-1 095	16 755	-1 427	3 181	-938	3 390	-780	4 051	-2 753	20 551
$\tilde{V}(\tilde{F}_s^*(t))$	-1 737	14 642	-298	5 648	-546	5 282	-736	5 679	-3 564	38 344
$\tilde{V}(\tilde{F}_l^*(t))$	-1 174	14 111	-27	5 856	-422	5 452	-228	5 974	-1 433	43 923
$\tilde{V}(\tilde{F}_\pi(t))$	-307	28 421	-460	3 963	-344	3 850	112	4 235	-401	27 987

Tableau 5.8
Populations réelles (taille de population $N = 284$). BIAISR et REQMR des estimateurs de variance sous échantillonnage de Poisson avec probabilités d'inclusion proportionnelles à la valeur absolue des résidus de la régression linéaire des valeurs y_i de la population sur les valeurs x_i de la population. Taille espérée $n^* = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln P85$										
$\tilde{V}(\tilde{F}_s(t))$	-3 502	26 342	-1 841	14 037	-2 691	12 087	-3 415	9 674	-5 932	26 823
$\tilde{V}(\tilde{F}_l(t))$	-2 159	27 610	-1 782	14 010	-2 840	12 002	-3 186	10 177	-4 455	26 802
$\tilde{V}(\tilde{F}_s^*(t))$	-434	22 455	515	15 503	-506	31 296	-1 460	23 496	-2 649	78 527
$\tilde{V}(\tilde{F}_l^*(t))$	-80	22 921	677	15 575	-280	33 294	-1 283	26 612	-1 597	72 166
$\tilde{V}(\tilde{F}_\pi(t))$	-294	361 991	522	75 891	43	48 764	-241	36 354	90	32 354
Population MU284 avec $Y = \ln RMT85$ et $X = \ln REV84$										
$\tilde{V}(\tilde{F}_s(t))$	-5 220	18 699	-3 667	8 749	-3 222	7 537	-3 018	9 279	-4 955	44 597
$\tilde{V}(\tilde{F}_l(t))$	-4 254	20 765	-3 100	9 180	-3 435	7 231	-3 196	8 540	-3 461	43 206
$\tilde{V}(\tilde{F}_s^*(t))$	-2 938	18 922	-1 110	11 828	-1 265	8 726	-1 040	10 963	-3 682	89 262
$\tilde{V}(\tilde{F}_l^*(t))$	-1 938	19 997	-699	12 641	-1 003	9 305	-599	11 545	-1 558	98 798
$\tilde{V}(\tilde{F}_\pi(t))$	-143	128 401	493	33 934	-255	18 473	-91	17 904	327	16 463

Comme le montrent les résultats des simulations, les estimateurs de variance souffrent d'une grande variabilité. Ce problème touche aussi l'estimateur de variance pour l'estimateur de Horvitz-Thompson qui, à l'occasion, présente de très grandes REQMR. Il est en outre intéressant de noter que, si le BIAISR des estimateurs de variance pour les estimateurs par la différence généralisée est presque toujours négatif et parfois assez grand en valeur absolue, celui de l'estimateur de variance pour l'estimateur de Horvitz-Thompson est positif dans la plupart des cas considérés.

Remerciements

La présente étude a été financée en partie par la subvention FAR 2014-ATE-0200 octroyée par *University of Milano-Bicocca*.

Annexe

Soit β une suite de nombres réels. Tout au long de la présente annexe, nous désignerons par $O_{i_1,i_2,\ldots,i_k}(\beta)$ les termes de reste qui peuvent dépendre de $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ et qui sont de même ordre que la suite β uniformément pour $i_1, i_2, \ldots, i_k \in U$. Formellement, $R\left(x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right) = O_{i_1,i_2,\ldots,i_k}(\beta)$ si

$$\sup_{i_1,i_2,\ldots,i_k \in U} \left| R\left(x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right) \right| = O(\beta).$$

En outre, pour simplifier la notation, nous écrirons m_i à la place de $m(x_i)$ et σ_i^2 à la place de $\sigma^2(x_i)$.

Biais de l'estimateur fondé sur le modèle de Kuo

$$
\begin{aligned}
E\left(\hat{F}(t)-F_N(t)\right) &= E\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[I\left(\varepsilon_j \le t-m_j\right)-I\left(\varepsilon_i \le t-m_i\right)\right]\right)\\
&= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t-m_j|x_j\right)-G\left(t-m_i|x_i\right)\right]\\
&= \frac{1}{2N}\sum_{i\notin s}\Big[G^{(2,0)}\left(t-m_i|x_i\right)\left(m_i'\right)^2-G^{(1,0)}\left(t-m_i|x_i\right)m_i''\\
&\qquad -2G^{(1,1)}\left(t-m_i|x_i\right)m_i'+G^{(0,2)}\left(t-m_i|x_i\right)\Big]\sum_{j\in s} w_{i,j}\left(x_j-x_i\right)^2+o\left(\lambda^2\right)\\
&= \lambda^2\frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b\Big[G^{(2,0)}\left(t-m(x)|x\right)\left(m'(x)\right)^2-G^{(1,0)}\left(t-m(x)|x\right)m''(x)\\
&\qquad -2G^{(1,1)}\left(t-m(x)|x\right)m'(x)+G^{(0,2)}\left(t-m(x)|x\right)\Big]h_{\bar{s}}(x)\,dx+o\left(\lambda^2\right).
\end{aligned}
$$

Biais de l'estimateur par la différence généralisée de Kuo

Écrivons

$$
\begin{aligned}
\tilde{F}(t)-F_N(t) &= \frac{1}{N}\Bigg\{\sum_{i\notin s}\sum_{j\in s}\tilde{w}_{i,j}\left[I\left(\varepsilon_j \le t-m_j\right)-I\left(\varepsilon_i \le t-m_i\right)\right]\\
&\qquad +\sum_{i\in s}\left(1-\frac{1}{\pi_i}\right)\sum_{j\in s}\tilde{w}_{i,j}\left[I\left(\varepsilon_j \le t-m_j\right)-I\left(\varepsilon_i \le t-m_i\right)\right]\Bigg\}.
\end{aligned}
$$

Des étapes similaires à celles suivies pour $\hat{F}(t)$ montrent que

$$
\begin{aligned}
E\left(\tilde{F}(t)-F_N(t)\right) &= \lambda^2\frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b\Big[G^{(2,0)}\left(t-m(x)|x\right)\left(m'(x)\right)^2-G^{(1,0)}\left(t-m(x)|x\right)m''(x)\\
&\qquad -2G^{(1,1)}\left(t-m(x)|x\right)m'(x)+G^{(0,2)}\left(t-m(x)|x\right)\Big]h(x)\,dx+o\left(\lambda^2\right),
\end{aligned}
$$

où

$$
h(x) := h_{\bar{s}}(x)+\left(1-\pi^{-1}(x)\right)h_s(x).
$$

Variance de l'estimateur fondé sur le modèle de Kuo

$$
\begin{aligned}
\operatorname{var}\left(\hat{F}(t)-F_N(t)\right) &= \operatorname{var}\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}I\left(\varepsilon_j \le t-m_j\right)-\frac{1}{N}\sum_{i\notin s}I\left(y_i \le t\right)\right)\\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\\
&\quad +\frac{1}{N^2}\sum_{i\notin s}\left[G\left(t-m_i|x_i\right)-G^2\left(t-m_i|x_i\right)\right]\\
&= A_1+A_2,
\end{aligned}
$$

où

$$\begin{aligned} A_1 &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right] \\ &= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\left(\sum_{i\notin s} w_{i,j}\right)^2 \\ &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|x\right)-G^2\left(t-m(x)|x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx \\ &\quad + O\left((n\lambda)^{-1}\alpha\right) \end{aligned}$$

et

$$\begin{aligned} A_2 &:= \frac{1}{N^2}\sum_{i\notin s}\left[G\left(t-m_i|x_i\right)-G^2\left(t-m_i|x_i\right)\right] \\ &= \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|x\right)-G^2\left(t-m(x)|x\right)\right]h_{\bar{s}}(x)\,dx+O\left(n^{-1}\alpha\right). \end{aligned}$$

Donc,

$$\begin{aligned} \operatorname{var}\left(\hat{F}(t)-F_N(t)\right) &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|\,x\right)-G^2\left(t-m(x)|\,x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx \\ &\quad + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|\,x\right)-G^2\left(t-m(x)|\,x\right)\right]h_{\bar{s}}(x)\,dx+O\left((n\lambda)^{-1}\alpha\right). \end{aligned}$$

Variance de l'estimateur par la différence généralisée de Kuo

Notons que,

$$\tilde{F}(t)-F_N(t)=\frac{1}{N}\left\{\sum_{j\in s} I\left(y_j\le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j}-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)+\left(\pi_j^{-1}-1\right)\right]-\sum_{i\notin s} I\left(y_i\le t\right)\right\}$$

de sorte que

$$\begin{aligned} \operatorname{var}\left(\tilde{F}(t)-F_N(t)\right) &= \operatorname{var}\left(\frac{1}{N}\sum_{j\in s} I\left(y_j\le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]\right) \\ &\quad + \operatorname{var}\left(\frac{1}{N}\sum_{i\notin s} I\left(y_i\le t\right)\right) \\ &= B_1+A_2, \end{aligned}$$

où le terme A_2 est le même que dans la variance de $\hat{F}(t)$, et où

$$
\begin{aligned}
B_1 &:= \operatorname{var}\left(\frac{1}{N}\sum_{j\in s} I\left(y_j \le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j} + \left(\pi_j^{-1}-1\right) - \sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]\right) \\
&= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j \mid x_j\right) - G^2\left(t-m_j \mid x_j\right)\right]\left[\sum_{i\notin s}\tilde{w}_{i,j} + \left(\pi_j^{-1}-1\right) - \sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]^2 \\
&= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j \mid x_j\right) - G^2\left(t-m_j \mid x_j\right)\right]\left[\sum_{i\notin s}\tilde{w}_{i,j} + \left(\pi_j^{-1}-1\right)\left(1-\sum_{i\in s}\tilde{w}_{i,j}\right)\right]^2 + O\left(\lambda n^{-1}\right) \\
&= \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \left[G\left(t-m(x) \mid x\right) - G^2\left(t-m(x) \mid x\right)\right]\left[h_{\tilde{s}}(x)/h_s(x)\right]h_{\tilde{s}}(x)\,dx \\
&\quad + O\left((n\lambda)^{-1}\alpha + \lambda n^{-1}\right) \\
&= A_1 + O\left((n\lambda)^{-1}\alpha + \lambda n^{-1}\right).
\end{aligned}
$$

Donc,

$$
\operatorname{var}\left(\tilde{F}(t) - F_N(t)\right) = \operatorname{var}\left(\hat{F}(t) - F_N(t)\right) + O\left((n\lambda)^{-1}\alpha + \lambda n^{-1}\right).
$$

Biais de l'estimateur fondé sur le modèle avec valeurs prédites modifiées

Soit $\hat{\tilde{m}}_i := \sum_{k\in s} w_{i,k} m_k$, $c_{i,j} := 1 - w_{j,j} + w_{i,j}$ et

$$
d_{i,j} := \frac{1}{c_{i,j}}\left[\left(1-c_{i,j}\right)\left(t-m_i\right) + \left(\hat{\tilde{m}}_j - m_j\right) - \left(\hat{\tilde{m}}_i - m_i\right) + \sum_{k\in s, k\ne j}\left(w_{j,k} - w_{i,k}\right)\varepsilon_k\right].
$$

Observons que $w_{i,j} = O_{i,j}\left((n\lambda)^{-1}\right)$ d'où

$$
y_j - \hat{\tilde{m}}_j \le t - \hat{\tilde{m}}_i
$$

est (asymptotiquement, aussitôt que $c_{i,j} > 0$) équivalent à

$$
\varepsilon_j \le t - m_i + d_{i,j}.
$$

Comme $d_{i,j}$ ne dépend pas de ε_j, il s'ensuit que

$$
\begin{aligned}
E\left(I\left(y_j - \hat{\tilde{m}}_j \le t - \hat{\tilde{m}}_i\right)\right) &= E\left(I\left(\varepsilon_j \le t - m_i + d_{i,j}\right)\right) \\
&= E\left(E\left(I\left(\varepsilon_j \le t - m_i + d_{i,j}\right) \mid \varepsilon_k, k \ne j\right)\right) \qquad \text{(A.1)} \\
&= E\left(G\left(t - m_i + d_{i,j} \mid x_j\right)\right).
\end{aligned}
$$

Or, en utilisant le fait que

$$d_{i,j} = (1-c_{i,j})(t-m_i) + (\hat{\hat{m}}_j - m_j) - (\hat{\hat{m}}_i - m_i) + \sum_{k\in s, k\neq j} (w_{j,k} - w_{i,k})\varepsilon_k + R(d_{i,j}), \tag{A.2}$$

où

$$E^{1/4}\left(|R(d_{i,j})|^4\right) = O_{i,j}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right), \tag{A.3}$$

on voit en examinant (A.1) que

$$\begin{aligned} E\left(I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right)\right) &= E\left(G\left(t - m_i + d_{i,j}\right)\middle| x_j\right) \\ &= G\left(t - m_i \middle| x_j\right) + G^{(1,0)}\left(t - m_i \middle| x_j\right) E\left(d_{i,j}\right) \\ &\quad + \frac{1}{2} G^{(2,0)}\left(t - m_i \middle| x_j\right) E\left(d_{i,j}^2\right) + o_{i,j}\left(\lambda^4 + (n\lambda)^{-1}\right). \end{aligned} \tag{A.4}$$

Donc,

$$\begin{aligned} E\left(\hat{F}^*(t) - F_N(t)\right) &= E\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left(I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right) - I\left(y_i \le t\right)\right)\right) \\ &= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t - m_i \middle| x_j\right) - G\left(t - m_i \middle| x_i\right)\right] \\ &\quad + \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t - m_i \middle| x_j\right) E\left(d_{i,j}\right) \\ &\quad + \frac{1}{2N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(2,0)}\left(t - m_i \middle| x_j\right) E\left(d_{i,j}^2\right) + o\left(\lambda^4 + (n\lambda)^{-1}\right) \\ &:= C_1 + C_2 + C_3 + o\left(\lambda^4 + (n\lambda)^{-1}\right). \end{aligned} \tag{A.5}$$

Considérons d'abord C_1 et notons que

$$\begin{aligned} C_1 &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t - m_i \middle| x_j\right) - G\left(t - m_i \middle| x_i\right)\right] \\ &= \frac{1}{2N}\sum_{i\notin s} G^{(0,2)}\left(t - m_i \middle| x_i\right)\sum_{j\in s} w_{i,j}\left(x_j - x_i\right)^2 + o\left(\lambda^2\right) \\ &= \lambda^2 \frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}\left(t - m(x) \middle| x\right) h_{\bar{s}}(x)\,dx + o\left(\lambda^2\right). \end{aligned}$$

Considérons ensuite C_2. (A.2) et (A.3) impliquent que

$$\begin{aligned} E\left(d_{i,j}\right) &= (1-c_{i,j})(t-m_i) + (\hat{\hat{m}}_j - m_j) - (\hat{\hat{m}}_i - m_i) + O_{i,j}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right) \\ &= (w_{j,j} - w_{i,j})(t-m_i) + m_j'' \sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 - m_i'' \sum_{k\in s} w_{i,k}\left(x_k - x_i\right)^2 \\ &\quad + o_{i,j}\left(\lambda^2\right) + O_{i,j}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right) \\ &= (w_{j,j} - w_{i,j})(t-m_i) + (m_j'' - m_i'')\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 \\ &\quad + m_i''\left(\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 - \sum_{k\in s} w_{i,k}\left(x_k - x_i\right)^2\right) \\ &\quad + o_{i,j}\left(\lambda^2\right) + O_{i,j}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right) \end{aligned}$$

de sorte que

$$C_2 = C_{2,a} + C_{2,b} + C_{2,c} + o\left(\lambda^2\right) + O\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right),$$

où

$$\begin{aligned} C_{2,a} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t - m_i \mid x_j\right)\left(w_{j,j} - w_{i,j}\right)(t - m_i) \\ &= \frac{1}{N}\sum_{i\notin s} G^{(1,0)}\left(t - m_i \mid x_i\right)(t - m_i)\sum_{j\in s} w_{i,j}\left(w_{j,j} - w_{i,j}\right) + O\left(n^{-1}\right) \\ &= \frac{1}{n\lambda}\frac{N-n}{N}\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}\left(t - m(x) \mid x\right)(t - m(x))\left[h_{\bar{s}}(x)/h_s(x)\right]dx \\ &\quad + O\left((n\lambda)^{-1}\lambda^{-1}\alpha + n^{-1}\right) \end{aligned}$$

avec $\kappa := \int_{-1}^{1} K^2(u)\,du$,

$$\begin{aligned} C_{2,b} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t - m_i \mid x_j\right)\left(m_j'' - m_i''\right)\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 \\ &= o\left(\lambda^2\right) \end{aligned}$$

et

$$\begin{aligned} C_{2,c} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t - m_i \mid x_j\right) m_i''\left(\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 - \sum_{k\in s} w_{i,k}\left(x_k - x_i\right)^2\right) \\ &= \frac{1}{N}\sum_{i\notin s} G^{(1,0)}\left(t - m_i \mid x_i\right) m_i''\left(\sum_{j\in s} w_{i,j}\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 - \sum_{k\in s} w_{i,k}\left(x_k - x_i\right)^2\right) + o\left(\lambda^2\right) \\ &= o\left(\lambda^2\right). \end{aligned}$$

Considérons enfin C_3. Notons que, d'après (A.2) et (A.3),

$$E\left(d_{i,j}^2\right) = \sum_{k\in s}\left(w_{j,k} - w_{i,k}\right)^2 \sigma_k^2 + O_{i,j}\left(\lambda^4 + (n\lambda)^{-2}\right) \tag{A.6}$$

d'où

$$\begin{aligned} C_3 &= \frac{1}{2N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(2,0)}\left(t - m_i \mid x_j\right)\sum_{k\in s}\left(w_{j,k} - w_{i,k}\right)^2 \sigma_k^2 + O\left(\lambda^4 + (n\lambda)^{-2}\right) \\ &= \frac{1}{2N}\sum_{i\notin s} G^{(2,0)}\left(t - m_i \mid x_i\right)\sigma_i^2 \sum_{j\in s} w_{i,j}\sum_{k\in s}\left(w_{j,k} - w_{i,k}\right)^2 + o\left((n\lambda)^{-1}\right) + O\left(\lambda^4\right) \\ &= \frac{1}{n\lambda}\frac{N-n}{N}\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}\left(t - m(x) \mid x\right)\sigma^2(x)\left[h_{\bar{s}}(x)/h_s(x)\right]dx + o\left((n\lambda)^{-1}\right) + O\left(\lambda^4\right) \end{aligned}$$

avec $\theta := \int_{-1}^{1} K(v) \int_{-1}^{1} K(u+v) K(u)\, du dv$.

En substituant les développements susmentionnés à C_1, C_2 et C_3 dans (A.5), on obtient finalement

$$\begin{aligned} E\left(\hat{F}^*(t) - F_N(t)\right) &= \lambda^2 \frac{N-n}{N} \frac{\mu_2}{\mu_0} \int_a^b G^{(0,2)}(t - m(x) \mid x) h_{\bar{s}}(x)\, dx \\ &\quad + \frac{1}{n\lambda} \frac{N-n}{N} \left[\frac{K(0) - \kappa}{\mu_0} \int_a^b G^{(1,0)}(t - m(x) \mid x)(t - m(x)) h_s^{-1}(x) h_{\bar{s}}(x)\, dx \right. \\ &\quad \left. + \frac{\kappa - \theta}{\mu_0^2} \int_a^b G^{(2,0)}(t - m(x) \mid x) \sigma^2(x) h_s^{-1}(x) h_{\bar{s}}(x)\, dx \right] \\ &\quad + o\left(\lambda^2 + (n\lambda)^{-1}\right). \end{aligned}$$

Biais de l'estimateur par la différence généralisée avec valeurs prédites modifiées

Soit $\tilde{d}_{i,j}$ l'équivalent pondéré selon le plan de sondage de $d_{i,j}$ et observons que

$$\begin{aligned} \tilde{F}^*(t) - F_N(t) &= \frac{1}{N} \left[\sum_{i \notin s} \sum_{j \in s} \tilde{w}_{i,j} \left(I\left(\varepsilon_j \le t - m_i + \tilde{d}_{i,j}\right) - I(y_i \le t) \right) \right. \\ &\quad \left. + \sum_{i \in s} \left(1 - \pi_i^{-1}\right) \sum_{j \in s} \tilde{w}_{i,j} \left(I\left(\varepsilon_j \le t - m_i + \tilde{d}_{i,j}\right) - I(y_i \le t) \right) \right]. \end{aligned} \tag{A.7}$$

En adaptant la preuve qui mène à (A.4), on voit que le développement asymptotique en (A.4) est également vérifié en prenant $\tilde{d}_{i,j}$ à la place de $d_{i,j}$. L'adaptation de la partie restante de la preuve mène en bout de ligne à

$$\begin{aligned} E\left(\tilde{F}^*(t) - F_N(t)\right) &= \lambda^2 \frac{N-n}{N} \frac{\mu_2}{\mu_0} \int_a^b G^{(0,2)}(t - m(x) \mid x) h(x)\, dx \\ &\quad + \frac{1}{n\lambda} \frac{N-n}{N} \left[\frac{K(0) - \kappa}{\mu_0} \int_a^b G^{(1,0)}(t - m(x) \mid x)(t - m(x)) h_s^{-1}(x) h(x)\, dx \right. \\ &\quad \left. + \frac{\kappa - \theta}{\mu_0^2} \int_a^b G^{(2,0)}(t - m(x) \mid x) \sigma^2(x) h_s^{-1}(x) h(x)\, dx \right] \\ &\quad + o\left(\lambda^2 + (n\lambda)^{-1}\right), \end{aligned}$$

où

$$h(x) := h_{\bar{s}}(x) + \left(1 - \pi^{-1}(x)\right) h_s(x).$$

Variance de l'estimateur fondé sur le modèle avec valeurs prédites modifiées

Écrivons

$$\hat{F}^*(t) - F_N(t) = \frac{1}{N}\left(\sum_{i \notin s} \sum_{j \in s} w_{i,j} I\left(\varepsilon_j \le t - m_i + d_{i,j}\right) - \sum_{i \notin s} I\left(\varepsilon_i \le t - m_i\right) \right)$$

et observons que

$$\text{var}\left(\hat{F}^*(t) - F_N(t)\right) = D_1 + D_2 + D_3,$$

où

$$D_1 := \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j \in s} w_{i_1,j} w_{i_2,j} \,\text{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + d_{i_1,j}\right), I\left(\varepsilon_j \le t - m_{i_2} + d_{i_2,j}\right)\right),$$

$$D_2 := \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s, j_2 \ne j_1} w_{i_1,j_1} w_{i_2,j_2} \times \text{cov}\left(I\left(\varepsilon_{j_1} \le t - m_{i_1} + d_{i_1,j_1}\right), I\left(\varepsilon_{j_2} \le t - m_{i_2} + d_{i_2,j_2}\right)\right)$$

et où $D_3 := A_2$ provenant de la variance de l'estimateur fondé sur le modèle de Kuo.

Considérons D_1. Observons que

$$\begin{aligned} \text{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + d_{i_1,j}\right), I\left(\varepsilon_j \le t - m_{i_2} + d_{i_2,j}\right)\right) &= E\left(G\left(t - m_{i_1} + d_{i_1,j} \wedge t - m_{i_2} + d_{i_2,j} \middle| x_j\right)\right) \\ &\quad - E\left(G\left(t - m_{i_1} + d_{i_1,j} \middle| x_j\right)\right) E\left(G\left(t - m_{i_2} + d_{i_2,j} \middle| x_j\right)\right). \end{aligned} \tag{A.8}$$

Puisque

$$\left|\left(t - m_{i_1} + d_{i_1,j} \wedge t - m_{i_2} + d_{i_2,j}\right) - \left(t - m_{i_1} \wedge t - m_{i_2}\right)\right| \le \left|d_{i_1,j}\right| + \left|d_{i_2,j}\right|,$$

il découle de (A.6) que

$$E\left(G\left(t - m_{i_1} + d_{i_1,j} \wedge t - m_{i_2} + d_{i_2,j} \middle| x_j\right)\right) = G\left(t - m_{i_1} \wedge t - m_{i_2} \middle| x_j\right) + O_{i_1,i_2,j}\left(\lambda^2 + (n\lambda)^{-1/2}\right). \tag{A.9}$$

En outre, de (A.1), (A.4) et (A.6), il découle que

$$E\left(G\left(t - m_i + d_{i,j} \middle| x_j\right)\right) = G\left(t - m_i \middle| x_j\right) + O_{i,j}\left(\lambda^2 + (n\lambda)^{-1/2}\right). \tag{A.10}$$

En utilisant (A.9) et (A.10) pour obtenir un développement asymptotique pour la covariance en (A.8) et en introduisant par substitution le résultat dans la définition de D_1, on obtient

$$
\begin{aligned}
D_1 &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+d_{i_1,j}\right), I\left(\varepsilon_j \le t-m_{i_2}+d_{i_2,j}\right)\right)\\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[E\left(G\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\,\middle|\, x_j\right)\right)\right.\\
&\qquad\qquad \left. -E\left(G\left(t-m_{i_1}+d_{i_1,j}\,\middle|\, x_j\right)\right)E\left(G\left(t-m_{i_2}+d_{i_2,j}\,\middle|\, x_j\right)\right)\right]\\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[G\left(t-m_{i_1}\wedge t-m_{i_2}\,\middle|\, x_j\right)-G\left(t-m_{i_1}\,\middle|\, x_j\right)G\left(t-m_{i_2}\,\middle|\, x_j\right)\right]\\
&\quad + O\left(\lambda^2 n^{-1}+(n\lambda)^{-1/2}\,n^{-1}\right)\\
&= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j\,\middle|\, x_j\right)-G^2\left(t-m_j\,\middle|\, x_j\right)\right]\left(\sum_{i\notin s} w_{i,j}\right)^2 + O\left(\lambda n^{-1}+(n\lambda)^{-1/2}\,n^{-1}\right)\\
&= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)\,|\, x\right)-G^2\left(t-m(x)\,|\, x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx\\
&\quad + O\left((n\lambda)^{-1}\alpha+n^{-1}\lambda+n^{-1}(n\lambda)^{-1/2}\right).
\end{aligned}
\tag{A.11}
$$

Considérons ensuite

$$
D_2 := \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s, j_2\neq j_1} w_{i_1,j_1}w_{i_2,j_2}\times\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right), I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right).
$$

Puisque

$$
\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right), I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right)=0
$$

si $\left|x_{i_1}-x_{i_2}\right|>2\lambda,$ il s'ensuit que les termes de reste R_{i_1,j_1,i_2,j_2}, dont la contribution à la covariance susmentionnée est d'ordre $O_{i_1,j_1,i_2,j_2}(\beta)$ pour une suite β qui tend vers zéro, apportent à D_2 un terme d'ordre $O(\lambda\beta)$. Or, soit

$$
\begin{aligned}
b_{i,j_1,j_2} &:= c_{i,j_1}^{-1}\left(w_{j_1,j_2}-w_{i,j_2}\right),\\
a_{i,j_1,j_2} &:= t-m_i+d_{i,j_1}-b_{i,j_1,j_2}\varepsilon_{j_2}
\end{aligned}
$$

et notons que

$$
t-m_i+d_{i,j_1}=a_{i,j_1,j_2}+b_{i,j_1,j_2}\varepsilon_{j_2}.
$$

Puisque a_{i,j_1,j_2} ne dépend pas de ε_{j_1} ni de ε_{j_2}, il s'ensuit que

$$
\begin{aligned}
&E\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right)I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right)\\
&\quad = E\left(E\left(I\left(\varepsilon_{j_1}\le a_{i_1,j_1,j_2}+b_{i_1,j_1,j_2}\varepsilon_{j_2}\right)I\left(\varepsilon_{j_2}\le a_{i_2,j_2,j_1}+b_{i_2,j_2,j_1}\varepsilon_{j_1}\right)\,\middle|\,\varepsilon_k, k\neq j_1, j_2\right)\right)\\
&\quad = E\left(\int_{-\infty}^{\varepsilon^*_{i_1,i_2,j_1,j_2}} G\left(a_{i_2,j_2,j_1}+b_{i_2,j_2,j_1}\varepsilon\,\middle|\, x_{j_2}\right)dG\left(\varepsilon\,|\, x_{j_1}\right)\right)\\
&\quad + E\left(\int_{-\infty}^{\varepsilon^*_{i_2,i_1,j_2,j_1}} G\left(a_{i_1,j_1,j_2}+b_{i_1,j_1,j_2}\varepsilon\,\middle|\, x_{j_1}\right)dG\left(\varepsilon\,|\, x_{j_2}\right)\right)\\
&\quad - E\left(G\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\,\middle|\, x_{j_1}\right)G\left(\varepsilon^*_{i_2,i_1,j_2,j_1}\,\middle|\, x_{j_2}\right)\right),
\end{aligned}
\tag{A.12}
$$

où

$$\varepsilon^*_{i_1,i_2,j_1,j_2} := \frac{a_{i_1,j_1,j_2} + a_{i_2,j_2,j_1} b_{i_1,j_1,j_2}}{1 - b_{i_1,j_1,j_2} b_{i_2,j_2,j_1}}.$$

Notons que les deux espérances aux troisième et quatrième lignes de (A.12) sont les mêmes si i_1 et j_1 sont remplacés par i_2 et j_2, respectivement. Donc, il suffit d'analyser la première espérance. Étant donné que

$$\varepsilon^*_{i_1,i_2,j_1,j_2} = t - m_{i_1} + d_{i_1,j_1} + b_{i_1,j_1,j_2}\left(t - m_{i_2} - \varepsilon_{j_2}\right) + R\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\right),$$

où

$$E^{1/4}\left(\left|R\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\right)\right|^4\right) = O_{i_1,i_2,j_1,j_2}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right),$$

on voit que

$$\begin{aligned}
&E\left(\int_{-\infty}^{\varepsilon^*_{i_1,i_2,j_1,j_2}} G\left(a_{i_2,j_2,j_1} + b_{i_2,j_2,j_1}\varepsilon \,\middle|\, x_{j_2}\right) dG\left(\varepsilon \,\middle|\, x_{j_1}\right)\right) \\
&= G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right)\left[E\left(d_{i_1,j_1}\right) + b_{i_1,j_1,j_2}\left(t - m_{i_2}\right)\right] \\
&\quad + G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) E\left(d_{i_2,j_2}\right) + G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) b_{i_2,j_2,j_1} \int_{-\infty}^{t - m_{i_1}} \varepsilon dG\left(\varepsilon \,\middle|\, x_{j_1}\right) \\
&\quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d^2_{i_1,j_1}\right) + \frac{1}{2} G^{(2,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) E\left(d^2_{i_2,j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d_{i_1,j_1} d_{i_2,j_2}\right) \\
&\quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right),
\end{aligned} \tag{A.13}$$

et que

$$\begin{aligned}
&E\left(G\left(\varepsilon^*_{i_1,i_2,j_1,j_2} \,\middle|\, x_{j_1}\right) G\left(\varepsilon^*_{i_2,i_1,j_2,j_1} \,\middle|\, x_{j_2}\right)\right) \\
&= G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right)\left[E\left(d_{i_1,j_1}\right) + b_{i_1,j_1,j_2}\left(t - m_{i_2}\right)\right] \\
&\quad + G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right)\left[E\left(d_{i_2,j_2}\right) + b_{i_2,j_2,j_1}\left(t - m_{i_1}\right)\right] \\
&\quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d^2_{i_1,j_1}\right) \\
&\quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) E\left(d^2_{i_2,j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d_{i_1,j_1} d_{i_2,j_2}\right) \\
&\quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right).
\end{aligned} \tag{A.14}$$

En utilisant les développements asymptotiques en (A.4), (A.13) et (A.14), on obtient

$$\begin{aligned}
&\operatorname{cov}\left(I\left(\varepsilon_{j_1} \le t - m_{i_1} + d_{i_1,j_1}\right), I\left(\varepsilon_{j_2} \le t - m_{i_2} + d_{i_2,j_2}\right)\right) \\
&= G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) b_{i_2,j_2,j_1} \gamma_{i_1,j_1} + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) b_{i_1,j_1,j_2} \gamma_{i_2,j_2} \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) \operatorname{cov}\left(d_{i_1,j_1}, d_{i_2,j_2}\right) \\
&\quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right),
\end{aligned} \tag{A.15}$$

où

$$\gamma_{i,j} := \int_{-\infty}^{t-m_i} \varepsilon dG\left(\varepsilon \,|\, x_j\right).$$

Observons maintenant que

$$b_{i,j_1,j_2} = w_{j_1,j_2} - w_{i,j_2} + O_{i,j_1,j_2}\left((n\lambda)^{-2}\right)$$

et que

$$\begin{aligned}
\operatorname{cov}\left(d_{i_1,j_1}, d_{i_2,j_2}\right) &= \frac{1}{c_{i_1,j_1} c_{i_2,j_2}} \sum_{k \in s; k \ne j_1, j_2} \left(w_{j_1,k} - w_{i_1,k}\right)\left(w_{j_2,k} - w_{i_2,k}\right)\sigma_k^2 \\
&= \sum_{k \in s} \left(w_{j_1,k} - w_{i_1,k}\right)\left(w_{j_2,k} - w_{i_2,k}\right)\sigma_k^2 + O_{i_1,i_2,j_1,j_2}\left((n\lambda)^{-2}\right)
\end{aligned}$$

de sorte que

$$D_2 = 2D_{2a} + D_{2b} + o\left(\lambda^5 + n^{-1}\right), \tag{A.16}$$

où

$$\begin{aligned}
D_{2a} &:= \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s, j_2 \ne j_1} w_{i_1,j_1} w_{i_2,j_2} G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right)\left(w_{j_1,j_2} - w_{i_1,j_2}\right)\gamma_{i_2,j_2} \\
&= \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s} w_{i_1,j_1} w_{i_2,j_2} G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right)\left(w_{j_1,j_2} - w_{i_1,j_2}\right)\gamma_{i_2,j_2} + O\left(n^{-1}(n\lambda)^{-1}\right) \\
&= \frac{1}{N^2} \sum_{j_2 \in s} G^{(1,0)}\left(t - m_{j_2} \,\middle|\, x_{j_2}\right)\gamma_{j_2,j_2} \left[\sum_{j_1 \in s} w_{j_1,j_2} \sum_{i_1 \notin s} w_{i_1,j_1} \sum_{i_2 \notin s} w_{i_2,j_2} - \left(\sum_{i \notin s} w_{i,j_2}\right)^2\right] \\
&\quad + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\
&= O\left((n\lambda)^{-1}\alpha + n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right)
\end{aligned} \tag{A.17}$$

et

$$
\begin{aligned}
D_{2b} \quad &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s, j_2\neq j_1} w_{i_1,j_1}w_{i_2,j_2}G^{(1,0)}\left(t-m_{i_1}\,|\,x_{j_1}\right)G^{(1,0)}\left(t-m_{i_2}\,|\,x_{j_2}\right) \\
&\quad \times\sum_{k\in s}\left(w_{j_1,k}-w_{i_1,k}\right)\left(w_{j_2,k}-w_{i_2,k}\right)\sigma_k^2 \\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s} w_{i_1,j_1}w_{i_2,j_2}G^{(1,0)}\left(t-m_{i_1}\,|\,x_{j_1}\right)G^{(1,0)}\left(t-m_{i_2}\,|\,x_{j_2}\right) \\
&\quad \times\sum_{k\in s}\left(w_{j_1,k}-w_{i_1,k}\right)\left(w_{j_2,k}-w_{i_2,k}\right)\sigma_k^2 + O\left(n^{-1}(n\lambda)^{-1}\right) \qquad \text{(A.18)} \\
&= \frac{1}{N^2}\sum_{k\in s}\sigma_k^2\left[G^{(1,0)}\left(t-m_k\,|\,x_k\right)\right]^2\left(\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left(w_{j,k}-w_{i,k}\right)\right)^2 + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\
&= \frac{1}{N^2}\sum_{k\in s}\sigma_k^2\left[G^{(1,0)}\left(t-m_k\,|\,x_k\right)\right]^2\left(\sum_{j\in s} w_{j,k}\sum_{i\notin s} w_{i,j} - \sum_{i\notin s} w_{i,k}\right)^2 + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\
&= O\left((n\lambda)^{-1}\alpha + n^{-1}\lambda\right).
\end{aligned}
$$

En regroupant tout, on obtient finalement

$$
\begin{aligned}
\operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right) \quad &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)\,|\,x\right)-G^2\left(t-m(x)\,|\,x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx \\
&\quad + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)\,|\,x\right)-G^2\left(t-m(x)\,|\,x\right)\right]h_{\bar{s}}(x)\,dx + o\left(\lambda^5+n^{-1}\right).
\end{aligned}
$$

Variance de l'estimateur par la différence généralisée avec valeurs prédites modifiées

Étant donné (A.7), nous allons montrer que

$$
\operatorname{var}\left(\tilde{F}^*(t)-F_N(t)\right) = \operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right) + o\left(n^{-1}\right) \qquad \text{(A.19)}
$$

en démontrant que

$$
\operatorname{var}\left(\frac{1}{N}\sum_{i\in s}\left(1-\pi_i^{-1}\right)\sum_{j\in s}\tilde{w}_{i,j}\left(I\left(\varepsilon_j\le t-m_i+\tilde{d}_{i,j}\right)-I\left(y_i\le t\right)\right)\right) = o\left(n^{-1}\right). \qquad \text{(A.20)}
$$

Pour prouver (A.20), observons que la variance dans le premier membre peut s'écrire

$$
E_1+E_2+E_3-2E_4-2E_5,
$$

où

$$E_1 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j\in s}\tilde{w}_{i_1,j}\tilde{w}_{i_2,j}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_j \le t-m_{i_2}+\tilde{d}_{i_2,j}\right)\right),$$

$$E_2 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j_1\in s}\sum_{j_2\in s, j_2\neq j_1}\tilde{w}_{i_1,j}\tilde{w}_{i_2,j_2}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_{j_1} \le t-m_{i_1}+\tilde{d}_{i_1,j_1}\right), I\left(\varepsilon_{j_2} \le t-m_{i_2}+\tilde{d}_{i_2,j_2}\right)\right),$$

$$E_3 := \frac{1}{N^2}\sum_{i\in s}\left(1-\pi_i^{-1}\right)^2 \operatorname{var}\left(I\left(\varepsilon_i \le t-m_i\right)\right),$$

$$E_4 := \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\tilde{w}_{i_1,j}\left(1-\pi_i^{-1}\right)\left(1-\pi_j^{-1}\right)\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_i+\tilde{d}_{i,j}\right), I\left(\varepsilon_j \le t-m_j\right)\right),$$

et finalement

$$E_5 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j\in s, j\neq i_2}\tilde{w}_{i_1,j}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \le t-m_{i_2}\right)\right).$$

Pour commencer, considérons E_1 et E_2. Notons que, à part i) le fait que les indices de sommation i_1 et i_2 s'étendent sur s au lieu du complément de s dans U, ii) la présence des facteurs $\left(1-\pi_i^{-1}\right)$ et iii) le fait que les $w_{i,j}$ et les $d_{i,j}$ sont remplacés par leurs équivalents pondérés selon le plan de sondage $\tilde{w}_{i,j}$ et $\tilde{d}_{i,j}$, E_1 et E_2 sont semblables à D_1 et D_2 provenant de $\operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right)$, respectivement. L'adaptation des preuves qui mènent aux développements asymptotiques pour D_1 et D_2 montre donc que

$$E_1 = \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G(t-m(x)|x)-G^2(t-m(x)|x)\right]\left[1-\pi^{-1}(x)\right]^2 h_s(x)\,dx + o\left(n^{-1}\right)$$

et que

$$E_2 = o\left(\lambda^5 + n^{-1}\right).$$

Comme pour E_3, on constate immédiatement que

$$E_3 = E_1 + o\left(n^{-1}\right),$$

tandis que, pour traiter E_4 et E_5, on a besoin des développements asymptotiques pour

$$\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \le t-m_{i_2}\right)\right) \tag{A.21}$$

pour le cas où $j = i_2$ et celui où $j \neq i_2$. Dans le premier cas, nous pouvons faire appel à des arguments similaires à ceux utilisés pour prouver (A.9) et (A.10), ce qui donne

$$\begin{aligned}&\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_j \le t-m_j\right)\right)\\ &\quad = G\left(t-m_{i_1}\wedge t-m_j \,\middle|\, x_j\right) - G\left(t-m_{i_1}\,\middle|\, x_j\right)G\left(t-m_j\,\middle|\, x_j\right) + O\left(\lambda^2+(n\lambda)^{-1/2}\right).\end{aligned}$$

Par contre, quand $j \neq i_2$, la covariance dans (A.21) diffère de zéro uniquement si $\left|x_j - x_{i_2}\right| \leq \lambda$ ou $\left|x_{i_1} - x_{i_2}\right| \leq \lambda$, et en adaptant (A.12), on peut montrer que

$$\begin{aligned}
&E\left(I\left(\varepsilon_j \leq t - m_{i_1} + \tilde{d}_{i_1,j}\right) I\left(\varepsilon_{i_2} \leq t - m_{i_2}\right)\right) \\
&\quad = E\left(E\left(I\left(\varepsilon_j \leq \tilde{a}_{i_1,j,i_2} + \tilde{b}_{i_1,j,i_2}\varepsilon_{i_2}\right) I\left(\varepsilon_{i_2} \leq t - m_{i_2}\right) \middle| \varepsilon_k, k \neq i, j\right)\right) \\
&\quad = E\left(\int_{-\infty}^{t-m_{i_2}} G\left(\tilde{a}_{i_1,j,i_2} + \tilde{b}_{i_1,j,i_2}\varepsilon \middle| x_j\right) dG\left(\varepsilon \middle| x_{i_2}\right)\right) \\
&\quad = G\left(t - m_{i_1} \middle| x_j\right) G\left(t - m_{i_2} \middle| x_{i_2}\right) + G\left(t - m_{i_2} \middle| x_{i_2}\right) G^{(1,0)}\left(t - m_{i_1} \middle| x_j\right) E\left(d_{i_1,j}\right) \\
&\qquad + G^{(1,0)}\left(t - m_{i_1} \middle| x_j\right) \tilde{b}_{i_1,j,i_2}\gamma_{i_2,i_2} + \frac{1}{2} G\left(t - m_{i_2} \middle| x_{i_2}\right) G^{(2,0)}\left(t - m_{i_1} \middle| x_j\right) E\left(d_{i_1,j}^2\right) \\
&\qquad + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right),
\end{aligned}$$

où $\tilde{a}_{i,j,k}$ et $\tilde{b}_{i,j,k}$ sont les équivalents pondérés selon le plan de sondage de $a_{i,j,k}$ et $b_{i,j,k}$, respectivement. En adaptant également (A.4) pour tenir compte des poids de sondage, on constate que

$$\begin{aligned}
\operatorname{cov}\left(I\left(\varepsilon_j \leq t - m_{i_1} + \tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \leq t - m_{i_2}\right)\right) &= G^{(1,0)}\left(t - m_{i_1} \middle| x_j\right) \tilde{b}_{i_1,j,i_2}\gamma_{i_2,i_2} + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right) \\
&= G^{(1,0)}\left(t - m_i \middle| x_j\right)\left(\tilde{w}_{j,i_2} - \tilde{w}_{i_1,i_2}\right)\gamma_{i_2,i_2} + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right)
\end{aligned}$$

de sorte que (voir les étapes qui mènent aux développements asymptotiques des termes D_1 et D_2 dans la variance de l'estimateur en deux étapes fondé sur le modèle)

$$E_4 = E_1 + o\left(n^{-1}\right)$$

et

$$E_5 = o\left(\lambda^5 + n^{-1}\right).$$

Cela achève la preuve de (A.20) et donc (A.19) s'ensuit.

Bibliographie

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals Statistics*, 28(4), 1026-1053.

Chambers, R.L., et Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*, Oxford Statistical Science Series 37.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597-604.

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993). Bias robust estimation in finite populations using non-parametric calibration. *Journal of the American Statistical Association*, 88(421), 268-277.

Chen, J., et Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

Dorfman, A.H., et Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21(3), 1452-1475.

Fan, J., et Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20(4), 2008-2036.

Hansen, B.E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24, 726-748.

Johnson, A.A., Breidt, F.J. et Opsomer, J.D. (2008). Estimating distribution functions from survey data using nonparametric regression. *Journal of Statistical Theory and Practice*, 2(3), 419-431.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. Dans les *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, 280-285.

Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472), 1429-1442.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77(2), 365-375.

Rueda, M., Martínez, S., Martínez, H. et Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137(2), 435-448.

Rueda, M., Sànchez-Borrego, I., Arcos, A. et Martínez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika*, 71(1), 33-44.

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

Wang, J.C., et Opsomer, J.D. (2011). On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika*, 98(1), 91-106.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90(4), 937-951.

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(453), 185-193.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 131-145
Statistique Canada, nº 12-001-X au catalogue

Remarque concernant l'estimation par régression lorsque la taille de la population est inconnue

Michael A. Hidiroglou, Jae Kwang Kim et Christian Olivier Nambeu[1]

Résumé

L'estimateur par régression est utilisé de façon intensive en pratique, car il peut améliorer la fiabilité de l'estimation des paramètres d'intérêt tels que les moyennes ou les totaux. Il utilise les totaux de contrôle des variables connues au niveau de la population qui sont incluses dans le modèle de régression. Dans cet article, nous examinons les propriétés de l'estimateur par régression qui utilise les totaux de contrôle estimés à partir de l'échantillon, ainsi que ceux connus au niveau de la population. Cet estimateur est comparé aux estimateurs par régression qui utilisent uniquement les totaux connus du point de vue théorique et par simulation.

Mots-clés : Estimateur optimal; échantillonnage; pondération.

1 Introduction

Les grands organismes statistiques utilisent de plus en plus l'estimation par régression afin d'améliorer la fiabilité des estimateurs des paramètres d'intérêt (comme les totaux et les moyennes) lorsque des variables auxiliaires sont disponibles à l'échelle de la population. Cassel, Särndal et Wretman (1976) ainsi que Fuller (2009), entre autres, donnent un aperçu détaillé de l'estimateur par régression dans le contexte de l'échantillonnage. Nous montrons comment utiliser l'estimateur par régression pour estimer le total, $Y = \sum_{i \in U} y_i$ où $U = \{1, \ldots, N\}$ désigne la population cible. Un échantillon s de taille attendue n est sélectionné selon un plan de sondage $p(s)$ de U, où π_i est la probabilité d'inclusion du premier ordre. En l'absence de variables auxiliaires, nous utilisons l'estimateur d'Horvitz-Thompson donné par $\hat{Y}_\pi = \sum_{i \in s} d_i y_i$ (Horvitz et Thompson 1952), où $d_i = 1/\pi_i$ est le poids de sondage associé à l'unité i. L'estimateur par régression est donné par

$$\hat{Y}_{\text{REG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}, \tag{1.1}$$

où $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$, $\hat{\mathbf{X}}_\pi = \sum_{i \in s} d_i \mathbf{x}_i$, $\mathbf{x}_i = \left(1, x_{2i}, \ldots, x_{pi}\right)^{\text{T}}$, et $\hat{\mathbf{B}}$ est un vecteur de coefficients de régression estimés de dimension p, s'exprimant comme une fonction des variables observées $\left(y_i, \mathbf{x}_i^{\text{T}}\right)^{\text{T}}$ dans l'échantillon s.

Il est à noter que les composantes du vecteur du total de population $\mathbf{X}$ sont connues pour chacune des variables correspondantes du vecteur $\mathbf{x}_i = \left(1, x_{2i}, \ldots, x_{pi}\right)^{\text{T}}$ utilisé pour calculer $\hat{\mathbf{B}}$. Cependant, il arrive parfois qu'il y ait plus de variables auxiliaires observées dans l'échantillon que dans la population. Supposons que l'échantillon comprend q variables observées $(q > p)$, et que les p variables de la

1. Michael A. Hidiroglou, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, (Ontario), Canada K1A 0T6. Courriel : hidirog@yahoo.ca; Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011. Courriel : jkim@iastate.edu; Christian Olivier Nambeu, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, (Ontario), Canada K1A 0T6. Courriel : christianolivier.nambeu@canada.ca.

population sont un sous-ensemble des q variables observées dans l'échantillon. Supposons par ailleurs que certaines des $q - p$ variables supplémentaires de l'échantillon sont bien corrélées avec la variable d'intérêt y. Ces variables supplémentaires peuvent-elles être incorporées dans l'estimateur par régression afin d'en accroître l'efficacité ? Singh et Raghunath (2011) ont essayé de répondre à cette question lorsque $q = p + 1$. La variable supplémentaire de l'échantillon était l'ordonnée à l'origine, qu'ils ont utilisée pour estimer la taille de population inconnue N au moyen de l'équation $\hat{N} = \sum_{i \in s} d_i$.

Dans cet article, nous comparons l'estimateur proposé par Singh et Raghunath (2011) à d'autres estimateurs par régression lorsque N est connu et lorsqu'il ne l'est pas. Dans la section 2, nous décrivons les estimateurs par régression standard pour l'estimation des totaux lorsque N est connu, ainsi que par la régression proposée par Singh et Raghunath (2011) lorsque N est inconnu. Dans la section 3, nous proposons un autre estimateur lorsque N est inconnu. Dans la section 4, nous procédons à une étude par simulations afin d'illustrer la performance des différents estimateurs examinés en termes de biais et d'erreur quadratique moyenne. Enfin, dans la section 5, nous présentons nos conclusions et recommandations générales.

2 Estimateurs par régression

Sous des conditions générales de régularité (Isaki et Fuller 1982; Montanari 1987), une approximation de l'estimateur par régression (1.1) est

$$\tilde{Y}_{\text{REG}} = \hat{Y}_{\pi} + \left(\mathbf{X} - \hat{\mathbf{X}}_{\pi}\right)^{\text{T}} \mathbf{B}, \tag{2.1}$$

où $\mathbf{B}$ est la limite en probabilité de $\hat{\mathbf{B}}$ lorsque la taille de l'échantillon et celle de la population tendent vers l'infini. Pour de grands échantillons, la variance de l'estimateur par régression (1.1) peut être étudiée avec (2.1). Notons que $\tilde{Y}_{\text{REG}}$ est sans biais sous le plan de sondage $p(s)$ et peut être réexprimé sous la forme :

$$\tilde{Y}_{\text{REG}} = \mathbf{X}^{\text{T}}\mathbf{B} + \sum_{i \in s} d_i E_i, \tag{2.2}$$

où $E_i = y_i - \mathbf{x}_i^{\text{T}}\mathbf{B}$.

Une approximation de la variance par rapport au plan de $\hat{Y}_{\text{REG}}$ peut être donnée par

$$\text{AV}_p\left(\hat{Y}_{\text{REG}}\right) = \sum_{i \in U}\sum_{j \in U} \Delta_{ij} \frac{E_i}{\pi_i} \frac{E_j}{\pi_j}, \tag{2.3}$$

où $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$ et π_{ij} est la probabilité d'inclusion du second ordre pour les unités i et j. Notons que $\mathbf{B}$ peut être estimée selon l'approche assistée par modèle (Särndal, Swensson et Wretman 1992) et l'approche de la variance optimale (Montanari 1987). Les deux méthodes permettent d'obtenir des estimateurs approximativement sans biais. Dans le cas de l'approche assistée par modèle, les propriétés de base (biais et variance) sont valides même lorsque le modèle n'est pas spécifié correctement. Sous l'approche de la variance optimale, aucune hypothèse n'est formulée au sujet de la variable d'intérêt.

L'estimateur assisté par modèle de Särndal et coll. (1992) suppose un modèle de travail entre la variable d'intérêt (y) et les variables auxiliaires $(\mathbf{x})$. Le modèle de travail est désigné par m : $y_i = \mathbf{x}_i^{\text{T}}\boldsymbol{\beta} + \varepsilon_i$ où

$\boldsymbol{\beta}$ est un vecteur de p paramètres inconnus, $E_m(\varepsilon_i \mid \mathbf{x}_i) = 0$, $V_m(\varepsilon_i \mid \mathbf{x}_i) = \sigma_i^2$, et $\mathrm{Cov}_m(\varepsilon_i, \varepsilon_j \mid \mathbf{x}_i, \mathbf{x}_j) = 0, i \neq j$. Sous cette approche, $\mathbf{B}$ dans l'équation (2.1) est l'estimateur des moindres carrés ordinaires de $\boldsymbol{\beta}$ dans la population et est donné par

$$\mathbf{B}_{\mathrm{GREG}} = \left(\sum_{i \in U} c_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \left(\sum_{i \in U} c_i \mathbf{x}_i y_i\right), \tag{2.4}$$

où $c_i = \sigma_i^{-2}$. Cela donne l'estimateur suivant pour le total Y

$$\hat{Y}_{\mathrm{GREG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{GREG}}, \tag{2.5}$$

où

$$\hat{\mathbf{B}}_{\mathrm{GREG}} = \left(\sum_{i \in s} c_i d_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \left(\sum_{i \in s} c_i d_i \mathbf{x}_i y_i\right). \tag{2.6}$$

L'estimateur optimal de Montanari (1987), obtenu en minimisant la variance par rapport au plan de

$$\tilde{Y}_{\mathrm{REG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \mathbf{B},$$

est

$$\tilde{Y}_{\mathrm{OPT}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \mathbf{B}_{\mathrm{OPT}}, \tag{2.7}$$

où

$$\begin{aligned} \mathbf{B}_{\mathrm{OPT}} &= \left\{V\left(\hat{\mathbf{X}}_\pi\right)\right\}^{-1} \mathrm{Cov}\left(\hat{\mathbf{X}}_\pi, \hat{Y}_\pi\right) \\ &= \left(\sum_{i \in U}\sum_{j \in U} \Delta_{ij} \frac{\mathbf{x}_i}{\pi_i} \frac{\mathbf{x}_j^{\mathrm{T}}}{\pi_j}\right)^{-1} \left(\sum_{i \in U}\sum_{j \in U} \Delta_{ij} \frac{\mathbf{x}_i}{\pi_i} \frac{y_j}{\pi_j}\right). \end{aligned} \tag{2.8}$$

L'estimateur optimal pour le total Y est estimé par

$$\hat{Y}_{\mathrm{OPT}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{OPT}}, \tag{2.9}$$

où

$$\hat{\mathbf{B}}_{\mathrm{OPT}} = \left(\sum_{i \in s}\sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\mathbf{x}_i}{\pi_i} \frac{\mathbf{x}_j^{\mathrm{T}}}{\pi_j}\right)^{-1} \left(\sum_{i \in s}\sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\mathbf{x}_i}{\pi_i} \frac{y_j}{\pi_j}\right). \tag{2.10}$$

Il est à noter que, pour que nous puissions calculer les vecteurs de régression, la première composante qui les définit doit être inversible. Nous pouvons nous assurer qu'elle l'est en réduisant le nombre de variables auxiliaires qui entrent dans la régression si l'efficience de l'estimateur par régression qui en découle n'en souffre pas trop. Par contre, si la perte d'efficience est importante, nous pouvons inverser ces matrices singulières en utilisant des inverses généralisés.

Comme il est mentionné dans l'introduction, les totaux de population ne sont pas nécessairement connus pour toutes les composantes du vecteur auxiliaire $\mathbf{x}$. La régression utilise normalement les variables auxiliaires pour lesquelles un total de population correspondant est connu. En décomposant $\mathbf{x}_i$ en $\left(1, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ où $\mathbf{x}_i^* = \left(x_{2i}, \ldots, x_{pi}\right)^{\mathrm{T}}$, Singh et Raghunath (2011) ont proposé un estimateur semblable au GREG qui suppose une régression fondée sur une ordonnée à l'origine et la variable $\mathbf{x}^*$, même si seul le total de population de $\mathbf{x}^*$ est connu.

Si N est inconnu et que le total de population de $\mathbf{x}^*$ est connu, leur estimateur est

$$\hat{Y}_{\mathrm{SREG}} = \hat{Y}_\pi + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \hat{\mathbf{B}}_{2,\mathrm{GREG}}, \tag{2.11}$$

où $\mathbf{X}^* = \sum_{i \in U} \mathbf{x}_i^*$ et $\hat{\mathbf{X}}_\pi^* = \sum_{i \in s} d_i \mathbf{x}_i^*$. Le vecteur de régression des coefficients estimés $\hat{\mathbf{B}}_{2,\mathrm{GREG}}$ est obtenu à partir de $\hat{\mathbf{B}}_{\mathrm{GREG}} = \left(\hat{B}_{1,\mathrm{GREG}}, \hat{\mathbf{B}}_{2,\mathrm{GREG}}^{\mathrm{T}}\right)^{\mathrm{T}}$ donné par (2.6). La variance approximative par rapport au plan de $\hat{Y}_{\mathrm{SREG}}$ prend la même forme que l'équation (2.3), où $E_i = y_i - \mathbf{x}_i^{*\mathrm{T}} \mathbf{B}_{2,\mathrm{GREG}}$, et

$$\mathbf{B}_{2,\mathrm{GREG}} = \left\{\sum_{i \in U} c_i \left(\mathbf{x}_i^* - \bar{\mathbf{X}}_N^*\right)\left(\mathbf{x}_i^* - \bar{\mathbf{X}}_N^*\right)^{\mathrm{T}}\right\}^{-1} \sum_{i \in U} c_i \left(\mathbf{x}_i^* - \bar{\mathbf{X}}_N^*\right) y_i$$

et $\bar{\mathbf{X}}_N^* = \sum_{i \in U} \mathbf{x}_i^* \big/ N$.

Nous pouvons obtenir les propriétés de (2.11) en notant que

$$\begin{aligned} \hat{Y}_{\mathrm{SREG}} - Y &= \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \hat{\mathbf{B}}_{2,\mathrm{GREG}} \\ &= \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \mathbf{B}_{2,\mathrm{GREG}} + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \left(\hat{\mathbf{B}}_{2,\mathrm{GREG}} - \mathbf{B}_{2,\mathrm{GREG}}\right). \end{aligned}$$

Étant donné que $\hat{\mathbf{B}}_{2,\mathrm{GREG}} - \mathbf{B}_{2,\mathrm{GREG}} = O_p\left(n^{-1/2}\right)$ sous certaines conditions de régularité examinées dans Fuller (2009, chapitre 2), le dernier terme est d'ordre plus faible. Ainsi, en ignorant les termes d'ordre plus faible, nous obtenons l'approximation

$$\hat{Y}_{\mathrm{SREG}} - Y \cong \sum_{i \in s} d_i E_i - \sum_{i \in U} E_i, \tag{2.12}$$

où $E_i = y_i - \mathbf{x}_i^{*\mathrm{T}} \mathbf{B}_{2,\mathrm{GREG}}$. Par conséquent, $\hat{Y}_{\mathrm{SREG}}$ est approximativement sans biais sous le plan. Nous pouvons calculer la variance asymptotique en utilisant

$$V\left\{\sum_{i \in s} d_i E_i - \sum_{i \in U} E_i\right\} = E\left\{\left(\sum_{i \in s} d_i E_i - \sum_{i \in U} E_i\right)^2\right\}.$$

Comme nous pouvons le voir, la variance asymptotique peut être assez importante à moins que $\sum_{i \in U} E_i = 0$.

Remarque 2.1 *Si* $y_i = a + bx_i$, *nous avons* $\hat{Y}_{\mathrm{SREG}} - Y = \left(\hat{N}_\pi - N\right)a$, *ce qui implique que* $V\left(\hat{Y}_{\mathrm{SREG}}\right) = a^2 V\left(\hat{N}_\pi\right)$. *Cela signifie que si* $V\left(\hat{N}_\pi\right) > 0$, *nous pouvons accroître artificiellement* $a^2 V\left(\hat{N}_\pi\right)$, *la variance de* $\hat{Y}_{\mathrm{SREG}}$, *en choisissant des valeurs élevées de* a.

Il est à noter que l'estimateur par régression optimal obtenu en utilisant $\mathbf{x}^* = \left(x_2, \ldots, x_p\right)^{\mathrm{T}}$ est lui aussi approximativement sans biais sous le plan, car

$$\begin{aligned}\hat{Y}_{\mathrm{OPT}}^* - Y &= \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{OPT}}^* \\ &= \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \mathbf{B}_{\mathrm{OPT}}^* + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \left(\hat{\mathbf{B}}_{\mathrm{OPT}}^* - \mathbf{B}_{\mathrm{OPT}}^*\right),\end{aligned}$$

où $\mathbf{B}_{\mathrm{OPT}}^*$ est obtenu en remplaçant $\mathbf{x}_i$ par $\mathbf{x}_i^*$ dans l'équation (2.8). Étant donné que $\hat{\mathbf{B}}_{\mathrm{OPT}}^* - \mathbf{B}_{\mathrm{OPT}}^* = O_p\left(n^{-1/2}\right)$ sous certaines conditions de régularité examinées dans Fuller (2009, chapitre 2), en ignorant les termes d'ordre plus faible, nous obtenons

$$\hat{Y}_{\mathrm{OPT}}^* - Y \cong \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \mathbf{B}_{\mathrm{OPT}}^*.$$

La variance asymptotique de $\hat{Y}_{\mathrm{OPT}}^*$ est plus faible que celle de $\hat{Y}_{\mathrm{SREG}}$, car l'estimateur optimal minimise la variance asymptotique dans la classe d'estimateurs de la forme

$$\hat{Y}_B = \hat{Y}_\pi + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \hat{\mathbf{B}} \tag{2.13}$$

indexée par $\hat{\mathbf{B}}$.

3 Estimateur par régression alternatif

Nous examinons maintenant un estimateur alternatif qui n'utilise pas l'information sur la taille de population (N). Il utilise plutôt les probabilités d'inclusion connues π_i, à condition qu'elles soient connues pour chaque unité de la population. Étant donné que $\sum_{i \in U} \pi_i = n$, nous pouvons utiliser $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ comme données auxiliaires dans le modèle

$$y_i = \mathbf{z}_i^{\mathrm{T}} \beta + e_i,$$

où $e_i \overset{\text{ind}}{\sim} \left(0, \sigma^2 \pi_i\right)$. Cela signifie que l'introduction de la structure de variance c_i de l'erreur dans le vecteur de régression est donnée par $c_i = d_i / \sigma^2$. L'estimateur qui en découle est donné par

$$\hat{Y}_{\mathrm{KREG}} = \hat{Y}_\pi + \left(\mathbf{Z} - \hat{\mathbf{Z}}_\pi\right)^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{KREG}}, \tag{3.1}$$

où $\mathbf{Z} = \sum_{i \in U} \mathbf{z}_i$, $\hat{\mathbf{Z}} = \sum_{i \in s} d_i \mathbf{z}_i$ et

$$\hat{\mathbf{B}}_{\mathrm{KREG}} = \left(\sum_{i \in s} c_i d_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}\right)^{-1} \sum_{i \in s} c_i d_i \mathbf{z}_i y_i. \tag{3.2}$$

Cet estimateur correspond exactement à celui fourni par Isaki et Fuller (1982).

Remarque 3.1 *Par construction,*

$$\sum_{i \in s} d_i^2 \left(y_i - \mathbf{z}_i^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{KREG}}\right) \mathbf{z}_i = \mathbf{0}$$

et, comme π_i *est une composante de* $\mathbf{z}_i$*, nous avons* $\sum_{i\in s} d_i \left(y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{KREG}}\right) = 0$*, ce qui aboutit à*

$$\hat{Y}_{\mathrm{KREG}} = \mathbf{Z}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{KREG}}.$$

Ainsi, $\hat{Y}_{\mathrm{KREG}}$ *est le meilleur prédicteur linéaire sans biais de* $Y = \sum_{i=1}^{N} y_i$ *sous le modèle*

$$y_i = \pi_i\beta_1 + \mathbf{x}_i^{*\mathrm{T}}\boldsymbol{\beta}_2 + e_i,$$

où $e_i \sim \left(0, \sigma^2\pi_i\right)$.

Il est à noter que nous pouvons exprimer $\hat{\mathbf{B}}_{\mathrm{KREG}}$ sous la forme $\hat{\mathbf{B}}_{\mathrm{GREG}}$ en posant que $c_i = d_i/\sigma^2$ et $\mathbf{x}_i = \mathbf{z}_i$. L'estimateur par régression proposé peut donc être considéré comme un cas spécial de l'estimateur GREG. En utilisant un argument semblable à (2.12), nous obtenons

$$\hat{Y}_{\mathrm{KREG}} - Y \cong \sum_{i\in s} d_i E_i^* - \sum_{i\in U} E_i^*, \tag{3.3}$$

où $E_i^* = y_i - \mathbf{z}_i^{\mathrm{T}}\mathbf{B}_{\mathrm{KREG}}$ et

$$\mathbf{B}_{\mathrm{KREG}} = \left(\sum_{i\in U} c_i\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right)^{-1}\sum_{i\in U} c_i\mathbf{z}_i y_i.$$

L'estimateur proposé est approximativement sans biais, et sa variance asymptotique

$$V\left\{\sum_{i\in s} d_i\left(y_i - \mathbf{z}_i^{\mathrm{T}}\mathbf{B}_{\mathrm{KREG}}\right)\right\} = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{E_i^*}{\pi_i}\frac{E_j^*}{\pi_j}$$

est souvent plus faible que celle de l'estimateur de Singh et Raghunath (2011).

La version optimale de $\hat{Y}_{\mathrm{KREG}}$ utilise $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ en tant que données auxiliaires. Elle est donnée par

$$\hat{Y}_{\mathrm{KOPT}} = \hat{Y}_\pi + \left(\mathbf{Z} - \hat{\mathbf{Z}}_\pi\right)^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{KOPT}}, \tag{3.4}$$

où $\hat{\mathbf{B}}_{\mathrm{KOPT}}$ est obtenu en remplaçant $\mathbf{x}_i$ par $\mathbf{z}_i$ dans l'équation (2.10).

Remarque 3.2 *Pour les plans de sondage de taille fixe, nous avons* $V_p\left(\sum_{i\in s} d_i\pi_i\right) = 0$*. Dans ce cas, le vecteur du coefficient de régression optimal* $\mathbf{B}_{\mathrm{KOPT}} = V_p\left(\hat{\mathbf{Z}}_\pi\right)^{-1}\mathrm{Cov}_p\left(\hat{\mathbf{Z}}_\pi, \hat{Y}_\pi\right)$ *ne peut pas être calculé, car la matrice de variances-covariances* $V_p\left(\hat{\mathbf{Z}}_\pi\right)$ *n'est pas inversible. En conséquence, l'estimateur optimal où* $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ *se réduit à l'estimateur optimal (2.9) seulement si nous utilisons* $\mathbf{x}_i^*$.

Remarque 3.3 *Pour les plans de sondage de taille aléatoire,* $V_p\left(\sum_{i\in s} d_i\pi_i\right) \geq 0$*. Dans ce cas, toutes les composantes de* $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ *peuvent être utilisées dans l'estimateur par régression optimal sous le plan (2.9).*

Une difficulté liée à l'utilisation de l'estimateur optimal $\hat{Y}_{\mathrm{KOPT}}$ est qu'il faut calculer les probabilités d'inclusion conjointe π_{ij}, ce qui peut s'avérer difficile sous certains plans de sondage. Nous pouvons obtenir

un estimateur qui ne nous oblige pas à calculer les probabilités d'inclusion conjointe en supposant que $\pi_{ij} = \pi_i \pi_j$. Nous donnons à cet estimateur le nom d'estimateur pseudo-optimal $\hat{Y}_{\text{POPT}}$. Il est donné par

$$\hat{Y}_{\text{POPT}} = \hat{Y}_{\pi} + \left(\mathbf{Z} - \hat{\mathbf{Z}}_{\pi}\right)^{\text{T}} \hat{\mathbf{B}}_{\text{POPT}}, \tag{3.5}$$

où

$$\hat{\mathbf{B}}_{\text{POPT}} = \left(\sum_{i \in s} c_i d_i \mathbf{z}_i \mathbf{z}_i^{\text{T}}\right)^{-1} \sum_{i \in s} c_i d_i \mathbf{z}_i y_i$$

et

$$c_i = d_i - 1.$$

En général, l'estimateur pseudo-optimal $\hat{Y}_{\text{POPT}}$ devrait produire des estimations proches de celles produites par $\hat{Y}_{\text{KREG}}$ lorsque la fraction de sondage est faible. Il est à noter que $\hat{Y}_{\text{POPT}}$ est exactement égal à l'estimateur optimal $\hat{Y}_{\text{KOPT}}$ dans le cas d'un plan de sondage de Poisson. Sous ce plan, les probabilités d'inclusion des unités de l'échantillon sont indépendantes. La variance approximative par rapport au plan pour $\hat{Y}_{\text{KREG}}$, $\hat{Y}_{\text{KOPT}}$ et $\hat{Y}_{\text{POPT}}$ a la même forme que celle donnée par l'équation (2.3) où les E_i sont donnés respectivement par $y_i - \mathbf{z}_i^{\text{T}} \hat{\mathbf{B}}_{\text{KREG}}$, $y_i - \mathbf{z}_i^{\text{T}} \hat{\mathbf{B}}_{\text{KOPT}}$ et $y_i - \mathbf{z}_i^{\text{T}} \hat{\mathbf{B}}_{\text{POPT}}$.

4 Simulations

Nous avons réalisé deux études par simulations. La première utilisait un ensemble de données fourni dans l'ouvrage de Rosner (2006), tandis que la deuxième se fondait sur une population artificielle créée selon un modèle de régression linéaire simple. La première simulation évaluait la performance de tous les estimateurs sous les différents plans de sondage, alors que la deuxième mettait l'accent sur l'impact de la modification de la valeur de l'ordonnée à l'origine dans le modèle.

Le paramètre d'intérêt pour ces deux simulations est le total de la variable d'intérêt y: $Y = \sum_{i \in U} y_i$. Tous les estimateurs $\left(\hat{Y}_{\text{GREG}}, \hat{Y}_{\text{OPT}}, \hat{Y}_{\text{POPT}}, \hat{Y}_{\text{SREG}}, \hat{Y}_{\text{KREG}} \text{ et } \hat{Y}_{\text{KOPT}}\right)$ ont été utilisés avec les données auxiliaires disponibles. Le tableau 4.1 résume les données auxiliaires et la structure de variance des erreurs (s'il y a lieu) qui sont associées aux estimateurs utilisés dans les deux études.

Tableau 4.1
Estimateurs utilisés dans l'étude de simulation

***N* connu**	***N* inconnu**
$\hat{Y}_{\text{GREG2}}$ défini par (2.5) où $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$ et $c_i = c$	$\hat{Y}_{\text{SREG1}}$ défini comme étant un cas spécial de (2.11) où $\mathbf{x}_i^* = (x_{2i})$
$\hat{Y}_{\text{OPT2}}$ défini par (2.9) où $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$	$\hat{Y}_{\text{OPT1}}$ défini par (2.9) où $\mathbf{x}_i = (x_{2i})$
$\hat{Y}_{\text{OPT3}}$ défini par (2.9) où $\mathbf{x}_i = (1, \pi_i, x_{2i})^{\text{T}}$	$\hat{Y}_{\text{KREG2}}$ défini par (3.1) où $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$ et $c_i = d_i / \sigma^2$
$\hat{Y}_{\text{POPT3}}$ défini par (3.5) où $\mathbf{z}_i = (1, \pi_i, x_{2i})^{\text{T}}$ et $c_i = d_i - 1$	$\hat{Y}_{\text{KOPT2}}$ défini par (3.4) où $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$
	$\hat{Y}_{\text{POPT2}}$ défini par (3.5) où $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$ et $c_i = d_i - 1$

La performance de tous les estimateurs a été évaluée en fonction du biais relatif, de l'efficacité relative de Monte Carlo et de l'efficacité relative approximative. Des expressions de ces quantités sont présentées ci-dessous.

1. *Biais relatif :*

$$\mathrm{RB}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{100}{R}\sum_{i=1}^{R}\frac{\left(\hat{Y}_{\mathrm{EST}(r)} - Y\right)}{Y}, \tag{4.1}$$

où $\hat{Y}_{\mathrm{EST}(r)}$ représente un des estimateurs présentés au tableau 4.1 tel que calculé dans le r^{e} échantillon de Monte Carlo.

2. *Efficacité relative Monte Carlo*

$$\mathrm{RE}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{EST}}\right)}{\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{GREG2}}\right)}, \tag{4.2}$$

où

$$\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_{\mathrm{EST}(r)} - Y\right)^2.$$

RE mesure l'efficacité relative de l'estimateur $\hat{Y}_{\mathrm{EST}}$ en ce qui concerne $\hat{Y}_{\mathrm{GREG2}}$.

3. *Efficacité relative approximative*

$$\mathrm{AR}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{\mathrm{AV}_p\left(\hat{Y}_{\mathrm{EST}}\right)}{\mathrm{AV}_p\left(\hat{Y}_{\mathrm{GREG2}}\right)}, \tag{4.3}$$

où

$$\mathrm{AV}_p\left(\hat{Y}_{\mathrm{EST}}\right) = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{E_i}{\pi_i}\frac{E_j}{\pi_i},$$

est la variance approximative de $\hat{Y}_{\mathrm{EST}}$ où $E_i = y_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{B}_{\mathrm{EST}}$. L'efficacité relative approximative (AR) mesure le gain d'efficacité relatif de $\hat{Y}_{\mathrm{EST}}$ par rapport à $\hat{Y}_{\mathrm{GREG2}}$ en utilisant le résidu de population obtenu au moyen de la technique de linéarisation de Taylor. On s'attend à ce que RE et AR donnent des résultats comparables. Cependant, comme nous allons le voir, ce n'est pas nécessairement le cas.

4.1 Simulation 1

La population était l'ensemble de données (FEV.DAT) disponible sur le CD qui accompagne l'ouvrage de Rosner (2006). Le fichier de données contient 654 enregistrements tirés d'une étude réalisée à Boston sur les maladies respiratoires des enfants. Les variables du fichier étaient l'âge, la taille, le sexe (masculin ou féminin), le tabagisme (c'est-à-dire si la personne fume ou non) et le volume expiratoire maximal (VEM). Singh et Raghunath (2011) ont utilisé le même ensemble de données. Le paramètre d'intérêt est la taille totale (y) de la population. La variable âge (x_1) a été utilisée comme variable auxiliaire dans la régression. La variable VEM (x_2) a été choisie comme variable de taille pour calculer les probabilités de sélection sous les plans de sondage examinés dans cette simulation. Les variables sexe et tabagisme ont été écartées. Le tableau 4.2 résume les mesures de la tendance centrale des trois variables dans la population. La moyenne et la médiane étaient similaires pour chaque variable, ce qui indique une répartition symétrique des trois variables.

Tableau 4.2
Statistiques descriptives de y, x_1 et x_2

	Minimum	Q1	Médiane	Moyenne	Q3	Maximum
y	46	57	61,5	61,14	65,5	74
x_1	3	8	10	9,931	12	19
x_2	0,79	1,98	2,55	2,64	3,12	5,79

La figure 4.1 illustre la relation entre la variable d'intérêt y et la variable auxiliaire x_1. La relation entre la taille (y) et l'âge (x_1) semble linéaire, mais ne passe pas par l'origine. Le coefficient de corrélation de Pearson entre y et x_1 était de 0,79.

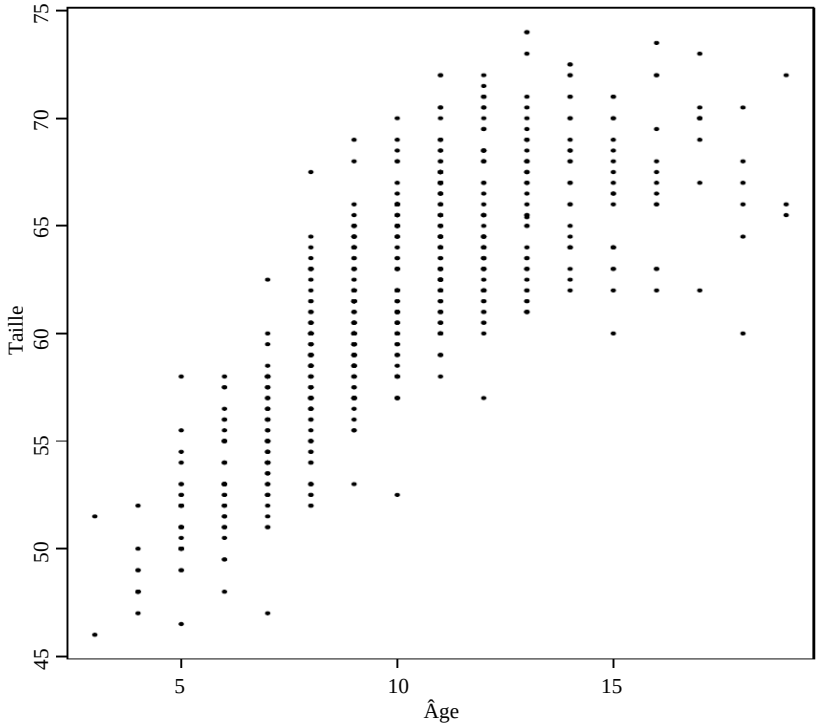

Figure 4.1 Relation entre la variable d'intérêt *Taille* et la variable auxiliaire *Âge*.

L'objectif de cette étude par simulations était d'évaluer la performance des estimateurs présentés au tableau 4.1 en utilisant différents plans de sondage. Nous avons examiné les plans de sondage de Midzuno, de Sampford et de Poisson. La variable x_2 a été utilisée comme mesure de taille sous les trois plans de sondage pour calculer les probabilités d'inclusion. Ces plans de sondage se présentent comme suit :

1. *Plan de sondage de Midzuno* (voir Midzuno 1952) : La première unité est échantillonnée avec la probabilité p_i et les $n-1$ unités restantes sont sélectionnées par échantillonnage aléatoire simple sans remise parmi les $N-1$ unités restantes de la population. Les probabilités de

sélection p_i pour l'unité i sont données par $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$. La probabilité d'inclusion de premier ordre pour l'unité i est donnée par $\pi_i = (N-1)^{-1}\left[(N-n)\, p_i + (n-1)\right]$.

2. *Plan de sondage de Sampford* (voir Sampford 1967) : Dans l'algorithme de sélection de l'échantillon, la première unité est sélectionnée avec la probabilité $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$ tandis que les $n-1$ unités restantes sont sélectionnées avec remise et avec la probabilité $\lambda_i = (1 - np_i)^{-1}\, p_i$. S'il y a des unités qui ont été sélectionnées plus d'une fois, la procédure est répétée jusqu'à ce que tous les éléments de l'échantillon soient différents. La probabilité d'inclusion de premier ordre est donnée par $\pi_i = np_i$.
3. *Plan de sondage de Poisson* : Chaque unité est sélectionnée indépendamment, ce qui donne une taille d'échantillon aléatoire. La probabilité de sélection de l'unité i est $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$. La probabilité d'inclusion associée à l'unité i est $\pi_i = np_i$. Une bonne description de cette procédure figure dans l'ouvrage de Särndal et coll. (1992).

Le paramètre d'intérêt était le total de $Y = \sum_{i \in U} y_i$. En nous basant sur chacun de ces plans de sondage, nous avons sélectionné $R = 2\,000$ échantillons Monte Carlo de taille $n = 50$. Nous avons ensuite calculé les estimateurs du tableau 4.1 pour chaque échantillon, puis avons évalué leur performance en utilisant le biais relatif, l'efficacité relative Monte Carlo et l'efficacité relative approximative tels que décrits dans les équations (4.1), (4.2) et (4.3) respectivement.

4.2 Résultats de la simulation 1

Les résultats de la simulation sont présentés au tableau 4.3. Tous les estimateurs étudiés sont approximativement sans biais, et leur biais relatif est inférieur à 1 %. Nous aborderons séparément l'efficacité relative approximative (AR) et l'efficacité relative (RE) des estimateurs lorsque la taille de la population N est connue et lorsqu'elle est inconnue.

Cas 1 : *La taille de la population N est connue*

Nous comparons les efficacités AR et RE des estimateurs $\hat{Y}_{\text{GREG2}}$, $\hat{Y}_{\text{OPT2}}$, $\hat{Y}_{\text{OPT3}}$ et $\hat{Y}_{\text{POPT3}}$ pour chacun des trois plans de sondage. Nous pouvons le faire pour presque tous ces estimateurs sauf $\hat{Y}_{\text{OPT3}}$ sous les plans de sondage de Midzuno et de Sampford. En l'occurrence, nous ne pouvons pas calculer $\mathbf{B}_{\text{OPT3}}$ pour une raison semblable à celle décrite dans la remarque 3.2.

Selon les efficacités AR et RE, l'estimateur pseudo-optimal $\hat{Y}_{\text{OPT3}}$ est l'estimateur le plus fiable, quel que soit le plan de sondage. Il est proche de l'estimateur optimal $\hat{Y}_{\text{OPT2}}$ seulement pour AR. Les efficacités RE et AR de l'estimateur optimal $\hat{Y}_{\text{OPT2}}$ n'étaient pas aussi proches que prévu dans le plan de sondage de Midzuno. Montanari (1998) a lui aussi observé la faible efficacité relative de l'estimateur optimal $\hat{Y}_{\text{OPT2}}$. La figure 4.2 montre ce qui se passe. Nous pouvons observer que la plupart des estimations obtenues au moyen de l'estimateur optimal $\hat{Y}_{\text{OPT2}}$ pour les 2 000 échantillons Monte Carlo sont proches de la moyenne. Cependant, dans certains échantillons, les estimations sont très éloignées de la moyenne. Cela contraste avec

$\hat{Y}_{\text{POPT3}}$, où les valeurs sont concentrées autour de la moyenne. Il est à noter que les efficacités RE et AR associées sont très proches l'une de l'autre.

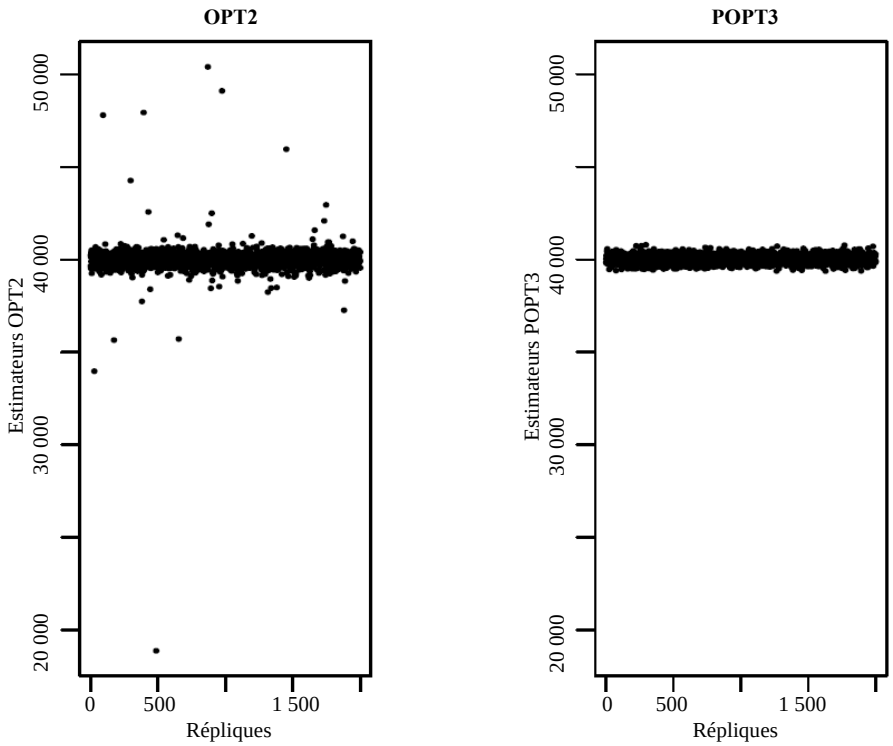

Figure 4.2 Nuages de points des estimateurs Monte Carlo sous le plan de sondage de Midzuno.

L'estimateur optimal $\hat{Y}_{\text{OPT3}}$ est équivalent à l'estimateur pseudo-optimal $\hat{Y}_{\text{POPT3}}$ sous le plan de sondage de Poisson. Il faut se rappeler que l'estimateur optimal $\hat{Y}_{\text{OPT2}}$ utilisait $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$ comme données auxiliaires, tandis que l'estimateur optimal $\hat{Y}_{\text{OPT3}}$ utilisait $\mathbf{x}_i = (1, \pi_i, x_{2i})^{\text{T}}$. L'ajout de π_i a beaucoup amélioré l'efficacité de l'estimateur optimal sous le plan de sondage de Poisson.

Singh et Raghunath (2011) utilisaient $\hat{Y}_{\text{SREG1}}$ lorsque N était connu, mais ne l'incluaient pas comme total de contrôle. Ils ont néanmoins observé que $\hat{Y}_{\text{SREG1}}$ était assez comparable à $\hat{Y}_{\text{GREG2}}$ en ce qui a trait aux efficacités AR et RB sous le plan de sondage de Midzuno. Pourquoi ? Parce que ce plan de sondage ressemble beaucoup à l'échantillonnage aléatoire simple sans remise. Cependant, si nous utilisons ces deux mesures, $\hat{Y}_{\text{SREG1}}$ est de loin le pire estimateur sous les deux autres plans de sondage.

Cas 2 : *La taille de la population N est inconnue*

Cinq estimateurs sont présentés au tableau 4.3 pour ce cas. Toutefois, comme $\hat{Y}_{\text{KREG2}}$ est très proche de $\hat{Y}_{\text{KOPT2}}$ et $\hat{Y}_{\text{POPT2}}$, nous commentons les résultats obtenus pour $\hat{Y}_{\text{SREG1}}$, $\hat{Y}_{\text{OPT1}}$ et $\hat{Y}_{\text{KREG2}}$. Les estimateurs $\hat{Y}_{\text{SREG1}}$, $\hat{Y}_{\text{OPT1}}$ et $\hat{Y}_{\text{KREG2}}$ sont très semblables pour ce qui est de l'efficacité relative et de l'efficacité relative approximative sous le plan de sondage de Midzuno. Sous le plan de sondage de Sampford, $\hat{Y}_{\text{OPT1}}$, $\hat{Y}_{\text{KREG2}}$ et $\hat{Y}_{\text{POPT2}}$ étaient comparables et donnaient des résultats légèrement meilleurs que ceux de l'estimateur $\hat{Y}_{\text{SREG1}}$. Sous le plan de sondage de Poisson, $\hat{Y}_{\text{OPT1}}$ et $\hat{Y}_{\text{KREG2}}$ étaient plus efficaces que $\hat{Y}_{\text{SREG1}}$. Nous constatons également que $\hat{Y}_{\text{SREG1}}$ était très inefficace, son efficacité relative étant au moins 10 fois plus

élevée que celles associées à $\hat{Y}_{\text{KREG2}}$ et $\hat{Y}_{\text{POPT2}}$. Notons que $\hat{Y}_{\text{KREG2}}$ donnait de meilleurs résultats que $\hat{Y}_{\text{OPT1}}$, ce qui est raisonnable puisque $\hat{Y}_{\text{KREG2}}$ utilise deux variables auxiliaires, tandis que $\hat{Y}_{\text{OPT1}}$ utilise seulement la variable auxiliaire x_{2i}.

Tableau 4.3
Comparaison des estimateurs en ce qui concerne le biais relatif et les efficacités relatives

		Taille de population connue				Taille de population inconnue				
		$\hat{Y}_{\text{GREG2}}$	$\hat{Y}_{\text{OPT2}}$	$\hat{Y}_{\text{OPT3}}$	$\hat{Y}_{\text{POPT3}}$	$\hat{Y}_{\text{SREG1}}$	$\hat{Y}_{\text{OPT1}}$	$\hat{Y}_{\text{KREG2}}$	$\hat{Y}_{\text{KOPT2}}$	$\hat{Y}_{\text{POPT2}}$
Midzuno	RB (en %)	0,08	0,04		0,07	0,07	0,07	0,07		0,07
	RE	1,00	5,84		0,54	0,94	0,93	0,93		0,93
	AR	1,00	0,55		0,55	0,94	0,93	0,93		0,93
Sampford	RB (en %)	0,11	0,11		0,07	-0,01	0,07	0,02		0,02
	RE	1,00	0,59		0,58	14,72	13,69	13,55		13,56
	AR	1,00	0,55		0,56	15,77	14,39	14,39		14,40
Poisson	RB (en %)	0,11	0,11	0,08	0,08	0,09	0,14	0,16	0,16	0,16
	RE	1,00	0,96	0,57	0,57	160,47	15,49	13,85	13,85	13,85
	AR	1,00	0,96	0,55	0,56	180,36	16,73	14,40	14,39	15,73

Note : Nous n'avons pas produits des résultats pour les estimateurs $\hat{Y}_{\text{OPT3}}$ et $\hat{Y}_{\text{KOPT2}}$ sous les plans de Midzuno et de Sampford car la matrice de variance-covariance n'était pas inversible.

4.3 Simulation 2

La performance des estimateurs a été évaluée pour différentes valeurs de l'ordonnée à l'origine dans le modèle. Nous nous sommes limités au plan de sondage de Poisson afin d'illustrer la remarque 2.1 de la section 2, à savoir que l'efficacité de $\hat{Y}_{\text{SREG}}$ se détériore au fur et à mesure que l'ordonnée à l'origine augmente. La population a été générée selon le modèle suivant :

$$y_i = a + x_i + e_i. \tag{4.4}$$

Les valeurs e_i ont été générées à partir de la loi normale de moyenne 0 et de variance $\sigma_i^2 = 1$. Les valeurs x ont été générées suivant une loi du chi-carré à un degré de liberté. Trois populations de taille $N = 5\,000$ ont été générées à l'aide de l'équation (4.4) avec différentes valeurs de l'ordonnée à l'origine a. Il est à noter que les valeurs x ont été générées à nouveau pour chaque population. Les trois populations étaient désignées A, B et C selon l'ordonnée à l'origine utilisée. Les valeurs de l'ordonnée à l'origine ont été fixées à 3, 5 et 10 respectivement pour les populations A, B et C. Dans chacune de ces populations, nous avons prélevé $R = 2\,000$ échantillons Monte Carlo d'une taille prévue $n = 50$ en utilisant le plan de sondage de Poisson. La première probabilité d'inclusion était égale à $\pi_i = nz_i \Big/ \sum_{i \in U} z_i$ pour chaque unité i. Les valeurs z ont été générées suivant le modèle

$$z_i = 0{,}5y_i + u_i,$$

où u_i est une erreur aléatoire générée selon la loi exponentielle de moyenne k égale à 0,5 ou 1.

4.4 Résultats de la simulation 2

Les résultats numériques sont présentés au tableau 4.4 pour $k = 1$ et au tableau 4.5 pour $k = 0{,}5$. Tous les estimateurs sont approximativement sans biais, les biais relatifs étant inférieurs à 1 %.

Cas 1 : *La taille de la population N est connue*

Comme prévu, les estimateurs optimaux $\hat{Y}_{\text{OPT2}}$ et $\hat{Y}_{\text{OPT3}}$ sont plus efficaces que $\hat{Y}_{\text{GREG2}}$. L'estimateur optimal $\hat{Y}_{\text{OPT2}}$ fondé sur $(1, x_{2i})^{\text{T}}$ donne des résultats légèrement meilleurs que ceux de $\hat{Y}_{\text{GREG2}}$. L'inclusion de la variable supplémentaire π_i engendrant $\hat{Y}_{\text{OPT3}}$ permet d'améliorer considérablement les efficacités RE et AR : ces gains diminuent au fur et à mesure que l'ordonnée à l'origine augmente. Là encore, $\hat{Y}_{\text{SREG1}}$ est très inefficace et, comme nous le soulignons dans la remarque 2.1, cette inefficacité augmente avec l'ordonnée à l'origine. Les observations qui précèdent restent valables, quelle que soit k. L'efficacité des estimateurs optimaux $\hat{Y}_{\text{OPT2}}$ et $\hat{Y}_{\text{OPT3}}$, quant à elle, diminue avec k.

Cas 2 : *La taille de la population N est inconnue*

L'estimateur le plus efficace est $\hat{Y}_{\text{KREG2}}$. Il surpasse $\hat{Y}_{\text{OPT1}}$, car il utilise plus de variables auxiliaires. L'estimateur $\hat{Y}_{\text{SREG1}}$ est de loin le plus inefficace. Lorsque l'ordonnée à l'origine dans le modèle de population augmente, les efficacités relatives RE et AR restent assez stables pour $\hat{Y}_{\text{KREG2}}$. Par contre, les efficacités relatives associées à $\hat{Y}_{\text{SREG1}}$ et $\hat{Y}_{\text{OPT1}}$ se détériorent rapidement à mesure que l'ordonnée à l'origine dans le modèle de population augmente. L'effet de k sur les efficacités des estimateurs est tel que décrit lorsque la taille de la population est connue.

Tableau 4.4
Biais relatif et efficacités relatives des estimateurs pour $k = 1$ sous le plan de sondage de Poisson

Ordonnée à l'origine		Taille de la population connue				Taille de la population inconnue				
		$\hat{Y}_{\text{GREG2}}$	$\hat{Y}_{\text{OPT2}}$	$\hat{Y}_{\text{OPT3}}$	$\hat{Y}_{\text{POPT3}}$	$\hat{Y}_{\text{SREG1}}$	$\hat{Y}_{\text{OPT1}}$	$\hat{Y}_{\text{KREG2}}$	$\hat{Y}_{\text{KOPT2}}$	$\hat{Y}_{\text{POPT2}}$
3	RB (en %)	0,23	0,38	0,56	0,56	0,18	0,77	0,22	0,22	0,22
	RE	1,00	0,95	0,67	0,67	7,72	5,42	0,94	0,94	0,94
	AR	1,00	0,94	0,60	0,98	7,08	5,01	0,85	0,85	0,91
5	RB (en %)	0,04	0,07	0,18	0,18	-0,01	0,67	-0,07	-0,07	-0,07
	RE	1,00	0,99	0,76	0,76	23,91	16,63	1,50	1,50	1,50
	AR	1,00	0,98	0,70	0,73	23,48	16,20	1,45	1,45	1,52
10	RB (en %)	-0,01	-0,02	0,06	0,06	-0,57	0,79	-0,02	-0,02	-0,02
	RE	1,00	1,00	0,80	0,80	88,30	67,47	2,20	2,20	2,20
	AR	1,00	0,99	0,73	0,74	97,92	66,13	2,15	2,15	2,20

Tableau 4.5
Biais relatif et efficacités relatives des estimateurs pour $k = 0,5$ sous le plan de sondage de Poisson

Ordonnée à l'origine		Taille de la population connue				Taille de la population inconnue				
		$\hat{Y}_{\text{GREG2}}$	$\hat{Y}_{\text{OPT2}}$	$\hat{Y}_{\text{OPT3}}$	$\hat{Y}_{\text{POPT3}}$	$\hat{Y}_{\text{SREG1}}$	$\hat{Y}_{\text{OPT1}}$	$\hat{Y}_{\text{KREG2}}$	$\hat{Y}_{\text{KOPT2}}$	$\hat{Y}_{\text{POPT2}}$
3	RB (en %)	0,13	0,25	0,42	0,42	-0,18	0,54	-0,02	-0,02	-0,02
	RE	1,00	0,99	0,89	0,89	8,42	5,93	1,78	1,78	1,78
	AR	1,00	0,96	0,83	0,95	8,30	5,83	1,79	1,79	2,10
5	RB (en %)	0,03	0,09	0,22	0,22	0,72	1,49	0,18	0,18	0,18
	RE	1,00	1,00	0,91	0,91	24,35	17,39	3,26	3,26	3,26
	AR	1,00	0,98	0,88	0,94	23,83	16,41	3,15	3,15	3,54
10	RB (en %)	0,06	0,07	0,12	0,12	0,33	1,42	0,13	0,13	0,13
	RE	1,00	1,00	0,96	0,96	98,69	73,93	6,26	6,26	6,26
	AR	1,00	0,99	0,91	0,92	98,65	66,20	5,89	5,89	6,24

5 Conclusions

L'estimateur par régression peut être très efficace lorsque les données auxiliaires qu'il utilise sont bien corrélées avec la variable d'intérêt. Il faut aussi que les totaux de population correspondant aux variables auxiliaires soient disponibles. Dans cet article, nous avons examiné le comportement de l'estimateur par régression $\left(\hat{Y}_{\mathrm{SREG}}\right)$ proposé par Singh et Raghunath (2011). Cet estimateur utilise le total estimé de la population comme total de contrôle et les totaux de population connus des variables auxiliaires. Nous l'avons comparé à l'estimateur par régression généralisée $\left(\hat{Y}_{\mathrm{GREG}}\right)$, son analogue optimal $\left(\hat{Y}_{\mathrm{OPT}}\right)$, et à un estimateur de rechange $\left(\hat{Y}_{\mathrm{KREG}}\right)$ qui utilise les probabilités d'inclusion de premier ordre et les données auxiliaires pour lesquelles les totaux de population sont connus. Comme l'estimateur par régression optimale nécessite le calcul des probabilités d'inclusion de second ordre, nous avons aussi inclus un estimateur pseudo-optimal $\left(\hat{Y}_{\mathrm{POPT}}\right)$ qui n'utilise pas ces probabilités. Nous avons examiné les propriétés de ces estimateurs en termes de biais et d'efficacité au moyen d'une simulation incluant différents plans de sondage et différentes valeurs de l'ordonnée à l'origine dans le modèle pour une population artificielle générée. Nous avons comparé les résultats obtenus lorsque la taille de la population était connue et inconnue.

Lorsque la taille de population est connue, l'estimateur optimal $\hat{Y}_{\mathrm{OPT}}$ est le plus efficace. Cependant, comme cet estimateur peut être instable, l'estimateur pseudo-optimal $\hat{Y}_{\mathrm{POPT}}$ est un bon substitut. Notre conclusion concorde avec celle de Rao (1994), qui préférait l'estimateur optimal $\hat{Y}_{\mathrm{POPT}}$ à l'estimateur par régression généralisée $\hat{Y}_{\mathrm{GREG}}$. La proposition de Singh et Raghunath (2011), qui recommandaient d'utiliser $\hat{Y}_{\mathrm{SREG}}$, n'est pas viable, car cet estimateur peut être très inefficace. Lorsque la taille de la population est inconnue, l'estimateur de rechange par régression $\hat{Y}_{\mathrm{KREG}}$ donne les meilleurs résultats.

Remerciements

Les auteurs remercient le rédacteur associé et les arbitres pour leurs suggestions qui ont considérablement améliorées la qualité de cet article.

Bibliographie

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimators and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

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

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

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

Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics*, 3, 99-107.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.

Montanari, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 1, 71-79.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary data information at the estimation stage. *Journal of Official Statistics*, 10(2), 153-165.

Rosner, B. (2006). *Fundamentals of Biostatistics*. Sixième édition, Duxbury Press.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of section. *Biometrika*, 54, 499-513.

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

Singh, S., et Raghunath, A. (2011). On calibration of design weights. *METRON International Journal of Statistics*, vol. LXIX, 2, 185-205.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 147-172
Statistique Canada, nº 12-001-X au catalogue

Échantillonnage fondé sur des registres pour les panels auprès des ménages

Jan A. van den Brakel[1]

Résumé

Aux Pays-Bas, les données statistiques sur le revenu et le patrimoine reposent sur deux grands panels auprès des ménages qui sont entièrement dérivés de données administratives. L'utilisation de ménages comme unités d'échantillonnage dans les plans de sondage des panels pose problème en raison de l'instabilité de ces unités au fil du temps. Les changements dans la composition des ménages influent sur les probabilités d'inclusion nécessaires aux méthodes d'inférence fondées sur le plan et assistées par modèle. Dans les deux panels auprès des ménages susmentionnés, ces problèmes sont surmontés par la sélection de personnes que l'on suit au fil du temps. À chaque période, les membres des ménages auxquels appartiennent les personnes choisies sont inclus dans l'échantillon. Il s'agit d'une méthode équivalente à un échantillonnage selon des probabilités proportionnelles à la taille du ménage, selon laquelle les ménages peuvent être sélectionnés plus d'une fois jusqu'à concurrence du nombre de membres du ménage. Dans le présent article, nous décrivons les propriétés de ce plan d'échantillonnage et les comparons avec la méthode généralisée du partage des poids pour l'échantillonnage indirect (Lavallée 1995, 2007). Les méthodes sont illustrées au moyen d'une application à la *Dutch Regional Income Survey*.

Mots-clés : Probabilités proportionnelles à la taille; échantillonnage indirect; pondération cohérente des personnes et des ménages; *Regional Income Survey*; méthode généralisée du partage des poids.

1 Introduction

Statistics Netherlands réalise deux grandes enquêtes par sondage pour recueillir des données sur le revenu et le patrimoine de la population néerlandaise. D'abord, la *Regional Income Survey* (RIS) brosse un portrait de la situation du revenu et du patrimoine à un niveau régional très détaillé. Des données exactes sur les répartitions des revenus par personne et par ménage au niveau des quartiers sont publiées annuellement, à partir d'un grand échantillon fondé sur un petit ensemble des principales composantes du revenu dérivées de manière assez directe à partir des données fiscales. Ensuite, des données sur le revenu annuel et les caractéristiques du patrimoine de la population néerlandaise sont publiées chaque année dans le cadre l'*Income Panel Survey* (IPS), à un niveau régional plus agrégé. Cette enquête est fondée sur un grand ensemble de variables faisant appel à toutes les composantes du revenu des ménages qu'il est possible de dériver à partir des données administratives disponibles aux Pays-Bas. Comme la dérivation des variables aux fins de cette enquête exige plus de temps, son échantillon est beaucoup plus petit que celui de la RIS. Les deux enquêtes sont conçues sous forme de panels auprès des ménages et permettent d'observer les variables sur le revenu et le patrimoine des personnes et des ménages.

Les ménages sont souvent considérés comme les unités d'échantillonnage dans les panels mis sur pied pour recueillir de l'information au niveau des ménages et des personnes (Lynn 2009; Smith, Lynn et Elliot 2009). De tels panels servent à la réalisation d'analyses longitudinales ainsi qu'à la production d'estimations transversales. L'utilisation des ménages comme unités d'échantillonnage dans le cadre d'une enquête par panel présente toutefois des inconvénients majeurs, en raison de l'instabilité des ménages au fil

1. Jan A. van den Brakel, Département des méthodes statistiques, Statistics Netherlands, C.P. 4481, 6401 CZ Heerlen, Pays-Bas et Département d'économie quantitative, Maastricht University School of Business and Economics, C.P. 616, 6200 MD, Maastricht, Pays-Bas. Courriel : ja.vandenbrakel@cbs.nl.

du temps. De fait, les ménages peuvent se défaire, se fusionner ou se séparer, de nouveaux membres peuvent s'y ajouter et d'autres les quitter pour différentes raisons. Kalton et Brick (1995) expliquent que ces changements peuvent influer sur les probabilités de sélection des ménages dans l'échantillon. La reconstruction des bonnes probabilités d'inclusion des unités d'échantillonnage est essentielle pour déterminer les poids à utiliser aux fins d'analyse, particulièrement si le panel sert à produire des estimations transversales.

Supposons un panel où les ménages sont sélectionnés par échantillonnage aléatoire simple, par exemple au moment $t = 0$. Dans beaucoup de panels, les personnes qui se joignent à un ménage échantillonné à une date ultérieure sont aussi incluses dans le panel. Lavallée (1995) appelle ces personnes des cohabitants. Au fil du temps, de plus en plus de cohabitants sont inclus dans l'échantillon et perturbent le plan d'échantillonnage à probabilités égales utilisé pour sélectionner l'échantillon initial (Kalton et Brick 1995). Prenons l'exemple du ménage A, sélectionné dans l'échantillon au moment de l'établissement du panel au temps $t = 0$. Si après un certain temps ce ménage fusionne avec le ménage B, qui n'a pas été sélectionné initialement pour le panel au temps $t = 0$, la probabilité de sélection de ce nouveau ménage correspond maintenant à la somme des probabilités de sélection des ménages A et B au temps $t = 0$. Le fait de ne pas corriger pour les différences dans les probabilités de sélection à cause de la croissance graduelle de la part des cohabitants dans l'échantillon donne lieu à une inférence biaisée. Ernst (1989) propose la méthode du partage des poids pour résoudre ce problème. Lavallée (1995) élargit cette solution à la méthode généralisée du partage des poids afin de faire des inférences à propos des populations cibles échantillonnées au moyen d'une base de sondage se rapportant à une population différente.

La RIS et l'IPS sont toutes deux des enquêtes par panel et sont réalisées afin de recueillir des données sur les ménages et les personnes. Afin d'éviter les problèmes associés au fait que les panels utilisent des ménages comme unités d'échantillonnage, un plan de sondage différent est utilisé. Au lieu de sélectionner des ménages, on sélectionne plutôt selon un plan d'échantillonnage à probabilités égales des « personnes principales », que l'on suit au fil du temps. Tous les membres du ménage auquel appartient une personne principale à chaque période particulière sont inclus dans l'échantillon. On obtient ainsi un plan d'échantillonnage en vertu duquel les ménages sont tirés proportionnellement à la taille du ménage et peuvent être sélectionnés plus d'une fois, jusqu'à concurrence du nombre de personnes dans le ménage. Ce plan est une application de l'échantillonnage indirect (Lavallée 1995, 2007; Deville et Lavallée 2006).

Le présent article décrit un plan d'échantillonnage assorti d'une technique d'estimation utile pour les panels recueillant des données au niveau de la personne et du ménage. La méthodologie employée est particulièrement utile dans le cas de l'échantillonnage fondé sur des registres, puisque les personnes principales sont incluses dans l'échantillon pendant une période indéterminée. Ce plan d'échantillonnage peut aussi servir aux panels en ligne, moyennant une forme quelconque de renouvellement afin de remédier au problème de l'attrition des participants. Cela signifie que les unités d'échantillonnage peuvent s'ajouter au panel, être observées plusieurs fois puis quitter le panel selon un schéma prédéterminé (Smith et coll. 2009). La principale contribution du présent article concerne la dérivation d'expressions explicites pour la variance des paramètres cibles à l'aide d'espérances d'inclusion plutôt que de probabilités d'inclusion en vertu du plan d'échantillonnage susmentionné. Une mesure de l'exactitude minimale pour une répartition

estimée du revenu est proposée et des expressions explicites pour déterminer la taille minimale d'échantillon sont dérivées. On utilise la RIS pour illustrer les techniques d'échantillonnage décrites.

L'article se présente comme suit. Le plan d'échantillonnage de la RIS est présenté à la section 2. À la section 3, on introduit le concept d'espérances d'inclusion comme solution de rechange pratique aux probabilités d'inclusion. Ensuite, les espérances d'inclusion de premier et de deuxième ordre sont dérivées pour le plan d'échantillonnage proposé. Ces espérances d'inclusion sont nécessaires pour construire l'estimateur π ou estimateur de Horvitz-Thompson (HT) (Narain 1951; Horvitz et Thompson 1952). On montre aussi que les mêmes poids peuvent être dérivés comme cas particulier de la méthode généralisée du partage des poids pour l'échantillonnage indirect (Lavallée 1995, 2007). Les variables cibles clés pour la RIS sont les répartitions estimées du revenu. À la section 4, les formules relatives à la taille minimale d'échantillon requise sont dérivées d'après une mesure de précision pour les répartitions estimées du revenu. Comme les ménages peuvent être sélectionnés plus d'une fois, une expression pour le nombre prévu de ménages uniques est dérivée à la section 4. La méthode d'estimation utilisée pour la RIS, fondée sur une pondération linéaire à l'aide de l'estimateur de régression généralisée (GREG) (Särndal, Swensson et Wretman 1992), est décrite à la section 5. La méthode de pondération intégrée de Lemaître et Dufour (1987), Nieuwenbroek (1993) et Steel et Clark (2007) est appliquée pour obtenir des poids égaux pour les personnes appartenant au même ménage. À la section 6, on dérive les approximations des variances pour l'estimateur GREG en vertu du plan d'échantillonnage proposé. Une application à la RIS est présentée à la section 7. L'article se termine à la section 8 par une discussion.

2 Plan d'échantillonnage

La population de la RIS comprend toutes les personnes physiques résidant aux Pays-Bas. La base de sondage est un registre de toutes les personnes physiques de 15 ans ou plus résidant aux Pays-Bas selon les dossiers du Bureau de l'impôt. À partir de ce registre, on tire un échantillon aléatoire simple stratifié de personnes dites « principales », selon une fraction de sondage de 0,16. Les quartiers servent de variable de stratification. Bien que l'on utilise un plan d'échantillonnage à probabilités égales, l'échantillonnage stratifié est utile pour éliminer la variation entre les strates et pour satisfaire aux exigences minimales en matière de précision pour chaque strate. Les Pays-Bas sont divisés en quelque 2 830 quartiers d'en moyenne 5 000 personnes de 15 ans ou plus.

La RIS est réalisée sous forme de panel depuis 1994. Pour garantir l'exactitude de l'inférence transversale à partir de ce panel, il faut d'abord établir avec justesse les espérances d'inclusion de premier et de deuxième ordre pour les unités d'échantillonnage; ces espérances sont dérivées à la section 3. Il faut ensuite s'assurer que le panel demeure représentatif de la population cible. Pour ce faire, on détermine annuellement quelle partie de la population est entrée dans la population cible de la RIS soit par naissance, soit par immigration. À partir de cette sous-population, on sélectionne un échantillon aléatoire simple stratifié de personnes principales selon une fraction de sondage de 0,16. Ces personnes principales sont ajoutées au panel de la RIS dans le but de maintenir un échantillon représentatif.

Les quartiers sont le niveau de publication le plus détaillé pour la RIS et sont donc utilisés comme strates. À la section 4, on dérive les expressions pour les tailles minimales des échantillons en fonction des

exigences de précision. Les personnes principales font partie du panel pendant une période indéterminée. À chaque période d'enquête, tous les membres des ménages des personnes principales sont aussi inclus dans l'échantillon. Les personnes qui quittent le ménage d'une personne principale sont éliminées du panel. Les nouvelles personnes qui se joignent au ménage d'une personne principale sont suivies dans le panel tant qu'elles font partie du ménage de la personne principale. Les données sur la composition des ménages des personnes principales sont obtenues auprès de la *Municipal Basis Administration* (MBA), le registre du gouvernement néerlandais recensant tous les résidents du pays. Les citoyens néerlandais sont tenus par la loi de déclarer aux municipalités tout changement d'ordre démographique. La MBA est utilisée conjointement avec les données provenant des autorités fiscales afin d'identifier les membres du ménage des personnes principales de l'échantillon.

Le plan d'échantillonnage permet d'établir un échantillon de ménages sélectionnés selon des probabilités proportionnelles au nombre de personnes de 15 ans ou plus appartenant au ménage à ce moment-là. Les ménages peuvent être sélectionnés plus d'une fois, jusqu'à concurrence du nombre de membres du ménage de 15 ans ou plus. Dans le présent article, l'expression « personne principale » désigne toute personne qui faisait partie de l'échantillon initial et fait l'objet d'un suivi au fil du temps dans le panel. Le mot « personnes » désigne l'échantillon obtenu si tous les membres du ménage à une période particulière sont inclus dans l'échantillon.

L'IPS repose sur un plan d'échantillonnage similaire, mais assorti d'une fraction de sondage beaucoup plus faible. La RIS et l'IPS sont toutes deux fondées sur des échantillons constitués à partir de registres, ce qui signifie que pour chaque personne incluse dans l'échantillon, les données nécessaires pour les variables de la RIS sont obtenues à partir des registres du Bureau de l'impôt. Les personnes principales et les membres de leurs ménages ne savent donc pas qu'elles font partie des échantillons. Cette méthode a pour avantage de ne poser aucun problème de non-réponse sélective ou d'attrition des participants au panel. Elle permet aussi d'inclure les personnes principales sur une période indéterminée. Dans le cas d'un panel où les unités d'échantillonnage doivent répondre à un questionnaire, il faut mettre en place un plan quelconque de renouvellement afin d'éliminer le biais de sélection attribuable à l'attrition. Par ailleurs, la méthode employée élimine les problèmes de biais de mesure associés à la collecte de données par questionnaire. Bien sûr, d'autres types d'erreurs de mesure se produisent lorsque l'enquête repose sur des registres (Wallgren et Wallgren 2007). Cela suppose notamment que tous les renseignements requis à propos du revenu pour estimer les paramètres cibles de la RIS et de l'IPS figurent dans les registres. Comme tous les renseignements requis se trouvent dans un registre, un dénombrement complet de la population est possible. Toutefois, par le passé, l'infrastructure de TI ne permettait pas de produire rapidement des données statistiques régionales sur le revenu pour toute la population des Pays-Bas. En conséquence, la RIS est réalisée par tradition sur un grand échantillon de personnes principales selon une fraction de sondage de 0,16. Pour la même raison, l'IPS repose par tradition sur un échantillon d'environ 80 000 personnes principales. Avec la capacité de calcul actuelle, un dénombrement complet serait possible mais tout de même très exigeant. La principale raison justifiant la réalisation de l'enquête à partir d'un échantillon est le maintien du panel aux fins des analyses longitudinales couvrant des périodes antérieures où un recensement n'était pas possible.

3 Poids d'inclusion

3.1 Pondération selon les espérances d'inclusion

Pour l'inférence fondée sur le plan de sondage, on a besoin des probabilités d'inclusion de premier et de deuxième ordre pour les ménages et les personnes. Soit M le nombre de ménages de la population, N le nombre de personnes de 15 ans et plus dans la population et g_k le nombre de personnes de 15 ans et plus appartenant au k^{e} ménage. Selon le plan d'échantillonnage décrit à la section 2, le ménage k peut être inclus plus d'une fois, jusqu'à concurrence de g_k fois. Cela complique la dérivation des probabilités d'inclusion, car la probabilité de sélectionner le ménage k est égale à la probabilité de sélection de l'union des membres du ménage (k, j) de 15 ans ou plus. Cette probabilité est définie comme suit :

$$\begin{aligned} P(k \in s) = P\left(\bigcup_{j=1}^{g_k}[(k, j) \in s]\right) &= \sum_{j=1}^{g_k} P((k, j) \in s) \\ &\quad - \sum_{j=1}^{g_k}\sum_{j'=j+1}^{g_k} P([(k, j) \cap (k, j')] \in s) \\ &\quad + \sum_{j=1}^{g_k}\sum_{j'=j+1}^{g_k}\sum_{j''=j+j'+1}^{g_k} P([(k, j) \cap (k, j') \cap (k, j'')] \in s) - \ldots \end{aligned}$$

On peut éviter ce type de calcul en utilisant le concept d'espérances d'inclusion, plutôt que les probabilités d'inclusion. Bethlehem (2009, chapitre 2) généralise l'estimateur HT au concept d'espérance d'inclusion pour un échantillonnage avec remise. Soit a_k le nombre de fois que le ménage k est sélectionné dans l'échantillon. Selon le plan d'échantillonnage proposé, $a_k \in [0, 1, \ldots, g_k]$. Soit $\text{E}(.)$ l'espérance relative au plan d'échantillonnage. Maintenant, $\pi_k = \text{E}(a_k)$ désigne l'espérance d'inclusion de l'unité d'échantillonnage k. Puisque a_k peut être plus grand que un, π_k peut aussi prendre une valeur supérieure à un et ne peut donc plus être interprétée comme une probabilité d'inclusion. Elle peut toutefois être interprétée comme une espérance.

Le paramètre d'intérêt est le total de population, défini par

$$t_y = \sum_{k=1}^{M}\sum_{j=1}^{N_k} y_{kj} \equiv \sum_{k=1}^{M} y_k. \tag{3.1}$$

L'estimateur HT du total de population en (3.1) peut être défini par

$$\hat{t}_y = \sum_{k=1}^{M} \frac{a_k y_k}{\pi_k}. \tag{3.2}$$

Puisque $\text{E}(a_k) = \pi_k$, il s'ensuit que cet estimateur HT est sans biais par rapport au plan. Soit $\pi_{kk'}$ l'espérance d'inclusion des unités k et k', c'est-à-dire $\pi_{kk'} = \text{E}(a_k a_{k'})$. Par définition, la variance de l'estimateur HT est égale à

$$\begin{aligned}
\mathrm{V}\left(\hat{t}_y\right) &= \sum_{k=1}^{M}\sum_{k'=1}^{M}\mathrm{Cov}\left(a_k a_{k'}\right)\frac{y_k}{\pi_k}\frac{y_{k'}}{\pi_{k'}} \\
&= \sum_{k=1}^{M}\sum_{k'=1}^{M}\left[\mathrm{E}\left(a_k a_{k'}\right) - \mathrm{E}\left(a_k\right)\mathrm{E}\left(a_{k'}\right)\right]\frac{y_k}{\pi_k}\frac{y_{k'}}{\pi_{k'}} \\
&= \sum_{k=1}^{M}\sum_{k'=1}^{M}\left(\pi_{kk'} - \pi_k\pi_{k'}\right)\frac{y_k}{\pi_k}\frac{y_{k'}}{\pi_{k'}}.
\end{aligned}$$

Soulignons que dans le cas d'un échantillonnage sans remise, a_k est une variable factice qui prend la valeur zéro ou un pour indiquer si l'unité k est sélectionnée dans l'échantillon. Dans ce cas, π_k et $\pi_{kk'}$ sont les probabilités d'inclusion de premier et de deuxième ordre habituelles. Cela montre que l'estimateur HT standard, fondé sur les probabilités d'inclusion, peut facilement être élargi aux espérances d'inclusion. Dans le cas des plans d'échantillonnage en vertu desquels les unités peuvent être sélectionnées plus d'une fois, il est plus commode de travailler avec des espérances d'inclusion, puisqu'elles peuvent être dérivées relativement facilement. Dans le reste de la présente sous-section, on dérive les espérances d'inclusion de premier et de deuxième ordre pour le plan d'échantillonnage décrit à la section 2.

Les personnes principales sont tirées à l'aide d'un échantillonnage aléatoire simple stratifié. Comme la stratification repose sur des régions géographiques, tous les membres d'un ménage k appartiennent à la même strate h au moment du tirage des personnes principales. Soit N_h le nombre de personnes de 15 ans ou plus dans la population de la strate h, n_h le nombre de personnes principales sélectionnées dans l'échantillon de la strate h et g_k le nombre de personnes de 15 ans ou plus appartenant au ménage k. Enfin, a_{jk} correspond à un indicateur égal à un si la personne j du ménage k est sélectionnée dans l'échantillon, et à zéro dans le cas contraire. L'espérance d'inclusion de premier ordre du k^{e} ménage égale

$$\pi_{kh} = \mathrm{E}\left(a_k\right) = E\left(\sum_{j=1}^{g_k} a_{jk}\right) = \sum_{j=1}^{g_k} E\left(a_{jk}\right) = g_k\,\frac{n_h}{N_h}. \tag{3.3}$$

Les espérances d'inclusion de deuxième ordre pour les ménages k et k' pour $k \neq k'$ appartenant à la même strate h égalent

$$\pi_{kk'} = \mathrm{E}\left(a_k a_{k'}\right) = E\left(\sum_{j=1}^{g_k} a_{jk}\sum_{j'=1}^{g_{k'}} a_{j'k'}\right) = \sum_{j=1}^{g_k}\sum_{j'=1}^{g_{k'}} E\left(a_{jk}a_{j'k'}\right) = g_k g_{k'}\,\frac{n_h\left(n_h-1\right)}{N_h\left(N_h-1\right)}. \tag{3.4}$$

L'espérance d'inclusion de deuxième ordre pour le ménage $k = k'$ pour la même strate h est donnée par

$$\begin{aligned}
\pi_{kk} &= \mathrm{E}\left(a_k a_k\right) = E\left(\sum_{j=1}^{g_k} a_{jk}\sum_{j'=1}^{g_k} a_{j'k}\right) = E\left(\sum_{j=1}^{g_k} a_{jk} + \sum_{j=1}^{g_k}\sum_{j'\neq j=1}^{g_k} a_{jk}a_{j'k}\right) \\
&= \sum_{j=1}^{g_k} E\left(a_{jk}\right) + \sum_{j=1}^{g_k}\sum_{j'\neq j=1}^{g_k} E\left(a_{jk}a_{j'k}\right) = g_k\,\frac{n_h}{N_h} + g_k\left(g_k-1\right)\frac{n_h\left(n_h-1\right)}{N_h\left(N_h-1\right)}.
\end{aligned} \tag{3.5}$$

Les espérances d'inclusion de deuxième ordre pour les ménages k et k' pour $k \neq k'$ appartenant à deux strates différentes h et h' égalent

$$\pi_{kk'} = \mathrm{E}\left(a_k a_{k'}\right) = \mathrm{E}\left(\sum_{j=1}^{g_k} a_{jk} \sum_{j'=1}^{g_{k'}} a_{j'k'}\right) = \sum_{j=1}^{g_k} \sum_{j'=1}^{g_{k'}} E\left(a_{jk} a_{j'k'}\right) = g_{kh} g_{k'h'} \frac{n_h n_{h'}}{N_h N_{h'}}. \tag{3.6}$$

Une autre preuve reposant sur la définition d'une espérance, qui ne fait pas appel à la règle voulant que l'espérance de la somme de variables mutuellement dépendantes soit égale à la somme des espérances de ces variables, est donnée par van den Brakel (2013).

Au fil du temps, la composition des ménages des personnes principales change, ce qui influe sur les espérances d'inclusion des ménages dans l'échantillon. Si les fractions de sondage diffèrent d'une strate à l'autre, les espérances d'inclusion (3.3) à (3.6) se complexifient et exigent de l'information sur l'appartenance à la strate pour toutes les personnes des ménages des personnes principales. On élimine cet inconvénient en choisissant un plan d'échantillonnage autopondéré. Dans ce cas, chaque membre du ménage d'une personne principale a la même probabilité d'inclusion et le seul renseignement propre au ménage nécessaire pour dériver les espérances d'inclusion du ménage est le nombre de personnes de 15 ans ou plus dans le ménage de la personne principale.

Comme tous les membres d'un ménage sélectionné sont inclus dans l'échantillon, il s'ensuit que les espérances d'inclusion de premier ordre pour les personnes appartenant au ménage k sont égales aux espérances d'inclusion de premier ordre du ménage k définies en (3.3). Les espérances d'inclusion de deuxième ordre pour les personnes appartenant à deux ménages différents k et k' sont égales à (3.4) si les deux ménages appartiennent à la même strate ou à (3.6) s'ils appartiennent à des strates différentes. Les espérances d'inclusion de deuxième ordre pour les personnes d'un même ménage sont établies par (3.5).

Les examinateurs du comité de lecture ont soulevé la question à savoir si les espérances d'inclusion elles-mêmes avaient une variance dont il faut tenir compte dans la variance des estimateurs HT ou GREG lorsque ceux-ci sont fondés sur des espérances d'inclusion plutôt que sur des probabilités d'inclusion. Dans la population finie, chaque personne et chaque ménage a une espérance d'inclusion prédéterminée. Dans le cas des ménages observés dans l'échantillon, ces espérances peuvent être calculées avec exactitude et sans incertitude, puisqu'on dispose de tous les renseignements nécessaires pour évaluer leur vraie valeur. La substitution des probabilités d'inclusion par des espérances n'introduit donc pas une variance supplémentaire.

3.2 Méthode généralisée du partage des poids

Le plan d'échantillonnage décrit à la section 2 peut être considéré comme un cas particulier d'échantillonnage indirect (Lavallée 2007). L'échantillonnage indirect s'entend d'une situation où la population d'intérêt est échantillonnée à partir d'une base de sondage se rapportant à une population différente. Lavallée (1995) a mis au point la méthode généralisée du partage des poids pour établir les poids

à utiliser dans une telle situation; cette méthode peut servir à dériver les poids de sondage pour les ménages et les personnes du plan d'échantillonnage décrit à la section 2.

Selon la notation de Lavallée (1995) pour les cas d'échantillonnage indirect, il y a une population U^A de taille N^A à partir de laquelle un échantillon s^A de taille n est tiré selon les probabilités de sélection π_i^A. De plus, il y a une population cible U^B de taille N^B. Cette population peut être divisée en M^B grappes. Chaque grappe k contient N_k^B unités, de sorte que $N^B = \sum_{k=1}^{M^B} N_k^B$. La situation du plan d'échantillonnage décrit à la section 2 est illustrée à la figure 3.1. Les grappes sont les ménages, U^A est la population des personnes de 15 ans ou plus et U^B est la population de toutes les personnes résidant aux Pays-Bas. Les personnes appartenant à U^A et à U^B sont représentées par des cercles et les ménages appartenant à U^B, par des carrés gris; les cercles se trouvant dans un carré gris représentent les personnes appartenant à un même ménage. La figure 3.1 montre respectivement un ménage d'une seule personne, un ménage de deux personnes formé par exemple d'un parent divorcé et d'un enfant de moins de 15 ans, un ménage de deux personnes formé de deux adultes sans enfant et un ménage de quatre personnes formé de deux parents et de deux enfants, l'un ayant moins de 15 ans, et l'autre, plus de 15 ans. Les flèches représentent les liens entre les unités de U^A et U^B. Dans le plan d'échantillonnage présenté à la section 2, chaque unité de U^A a exactement un seul lien avec une unité de U^B. Les grappes de U^B ont au moins un lien avec les unités de U^A. Les liens sont définis par une variable indicatrice

$$l_{ij} = \begin{cases} 1 & \text{s'il y a un lien entre } i \in U^A \text{ et } j \in U^B \\ 0 & \text{s'il n'y a pas de lien entre } i \in U^A \text{ et } j \in U^B. \end{cases}$$

Si une unité i de U^A est sélectionnée dans l'échantillon, toute la grappe k à laquelle cette unité appartient est incluse dans l'échantillon. Le paramètre d'intérêt est le total de population de U^B; il se compare à (3.1) et est défini par $t_y = \sum_{k=1}^{M^B} \sum_{j=1}^{N_k^B} y_{kj}$. Un estimateur de t_y est défini par

$$\hat{t}_y = \sum_{k=1}^{m} \sum_{j=1}^{N_k^B} w_{kj} y_{kj}, \tag{3.7}$$

avec m le nombre de grappes uniques (ménages) incluses dans l'échantillon et w_{kj} le poids associé à chaque unité j de la grappe k. En règle générale, on se sert de l'inverse des probabilités de sélection des unités (k, j) observées dans l'échantillon pour établir les poids dans l'estimateur HT. Dans ce cas, les unités de l'échantillon n'ont pas toutes une probabilité d'inclusion connue. D'abord, les unités de U^B n'ont pas toutes un lien avec une unité de U^A. Ensuite, la composition des ménages change au fil du temps en raison des mariages, des divorces, du départ des enfants et de la cohabitation. En conséquence, au fil du temps, les unités liées à U^A sont intégrées dans les grappes de l'échantillon même si elles ne faisaient pas initialement partie de l'échantillon tiré à partir de U^A. Même si leurs probabilités d'inclusion ne sont pas nécessairement connues, ces unités influent sur les espérances d'inclusion des grappes dans l'échantillon. Pour reconstruire les probabilités d'inclusion, il faut disposer de données sur les probabilités de sélection de toutes les unités de la population au moment du tirage de l'échantillon. Dans la pratique, on ne dispose généralement pas de cette information.

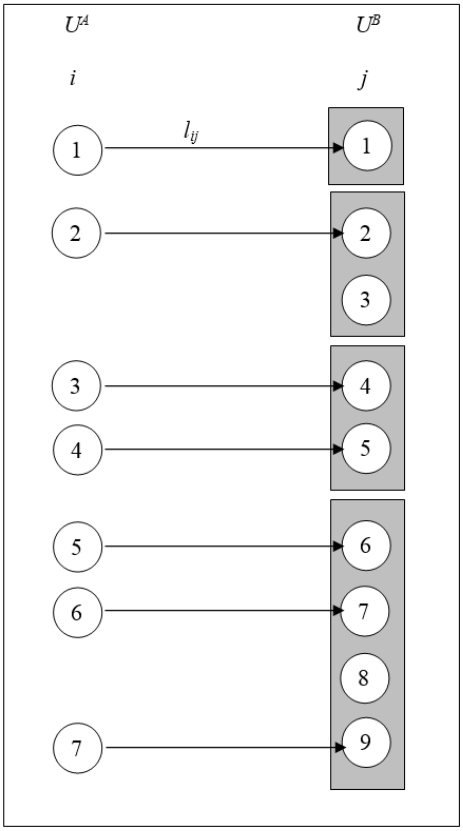

Figure 3.1 Liens entre les unités de la base de sondage et les unités de la population cible.

La méthode généralisée du partage des poids peut servir à dériver des poids non nuls pour toutes les unités de l'échantillon. Pour ce faire, il faut d'abord dériver les poids initiaux, définis par

$$w_{kj}^{*} = \begin{cases} \dfrac{\delta_i^A}{\pi_i^A} & \text{si } (k, j) \text{ a un lien avec } i \in U^A \\ 0 & \text{autrement} \end{cases},$$

avec δ_i^A une variable indicatrice égale à un si i est inclus dans l'échantillon s^A et à zéro autrement. Cette expression découle directement de Lavallée (1995), équation (2), ainsi que du fait que dans la présente application, chaque unité de U^A a exactement un seul lien avec une unité de U^B (voir la figure 3.1). Ensuite, un poids dit « de base » pour chaque grappe k est dérivé et correspond à la moyenne de tous les poids initiaux de chaque grappe :

$$w_k = \frac{\sum_{j=1}^{N_k^B} w_{kj}^{*}}{\sum_{j=1}^{N_k^B} l_{kj}},$$

qui découle de Lavallée (1995), équation (7). Enfin, toutes les personnes j appartenant au même ménage k reçoivent le même poids affecté à leur ménage, c'est-à-dire $w_{kj} = w_k$ pour tout $j \in k$. Une preuve que le recours à des poids de base en (3.7) constitue un estimateur non biaisé du total de population est aussi donnée par Lavallée (1995).

Soit $\sum_{j=1}^{N_k^B} l_{kj} = g_k$ le nombre de personnes de 15 ans ou plus du ménage k et a_k le nombre de personnes principales du ménage k, c'est-à-dire le nombre de personnes du ménage k faisant partie de l'échantillon s^A. Puisque s^A est tiré au moyen d'un échantillonnage aléatoire simple stratifié, il s'ensuit que $\pi_i^A = n_h^A / N_h^A$ avec N_h^A le nombre de personnes de 15 ans ou plus dans la population de la strate h, et n_h^A le nombre de personnes principales sélectionnées dans l'échantillon à partir de la strate h. Il s'ensuit que

$$w_k = \frac{a_k}{g_k} \frac{N_h^A}{n_h^A}. \tag{3.8}$$

Si l'on intègre l'espérance d'inclusion de premier ordre (3.3) dans (3.2), on obtient le même estimateur HT que celui qui est dérivé à partir de la méthode généralisée du partage des poids, c'est-à-dire en intégrant (3.8) dans (3.7).

Le calcul des espérances d'inclusion à la sous-section 3.1 s'applique à l'échantillonnage stratifié des ménages avec espérances d'inclusion proportionnelles à la taille du ménage et constitue un cas particulier de la méthode généralisée du partage des poids. Le fait qu'un échantillonnage des ménages proportionnel à la taille du ménage est efficient pour les variables cibles corrélées positivement avec la taille du ménage constitue un argument en faveur de l'utilisation d'un plan comme celui qui est décrit à la section 2.

Lavallée (1995) fournit aussi des expressions de la variance pour (3.7) fondées sur la méthode généralisée du partage des poids. Ces expressions reposent sur les probabilités d'inclusion de premier et de deuxième ordre des unités de l'échantillon tirées de U^A et sur une transformation de la variable cible. Par conséquent, le fait que la probabilité de sélection des grappes soit proportionnelle à leur taille n'est pas explicite, pas plus que le fait qu'elles soient tirées partiellement avec remise. On souligne à la section 6 que les expressions de la variance de Lavallée (1995) pour cette application sont égales aux expressions de la variance fondées sur les espérances d'inclusion calculées aux expressions (3.3) à (3.6).

4 Détermination de la taille de l'échantillon

La RIS a pour objet de publier les répartitions du revenu des ménages et des personnes à différentes échelles géographiques. Les répartitions du revenu des ménages pour une région ou une zone r sont définies par

$$P_{lr} = \frac{M_{lr}}{M_{+r}}, \qquad l = 1, \ldots, L, \tag{4.1}$$

où M_{lr} correspond au nombre de ménages dans la région r appartenant à la l^{e} catégorie de revenus et $M_{+r} = \sum_l M_{lr}$, le nombre total de ménages dans la région r. Cette répartition du revenu est estimée par

$$\hat{P}_{lr} = \frac{\hat{M}_{lr}}{M_{+r}}, \qquad l = 1, \ldots, L, \tag{4.2}$$

où $\hat{M}_{lr}$ correspond à un estimateur direct approprié du nombre total de ménages dans la région r appartenant à la l^{e} catégorie de revenus. Pour le moment, l'estimateur HT est présumé être un estimateur approprié de M_{lr}, c'est-à-dire

$$\hat{M}_{lr} = \sum_{h \in r} \sum_{k=1}^{m_h} \frac{y_{khl}}{\pi_k},$$

où $y_{khl} = 1$ si le ménage k de la strate h appartient à la l^{e} catégorie de revenus et $y_{khl} = 0$ autrement, et où m_h correspond au nombre total de ménages sélectionnés dans la strate h. Pour la RIS, $L = 10$. Les répartitions du revenu des personnes sont définies et estimées comme en (4.1) et (4.2), où M_{lr} correspond au nombre de personnes de la région r appartenant à la l^{e} catégorie de revenus. L'estimateur HT pour M_{lr} correspond maintenant à

$$\hat{M}_{lr} = \sum_{h \in r} \sum_{k=1}^{m_h} \frac{1}{\pi_k} \sum_{j=1}^{N_k} y_{kjhl},$$

où $y_{kjhl} = 1$ si la personne j du ménage k et de la strate h appartient à la l^{e} catégorie de revenus et $y_{kjhl} = 0$ autrement.

Pour déterminer la taille de l'échantillon, des spécifications précises pour les répartitions estimées du revenu sont nécessaires. Dans le cas des plans d'échantillonnage stratifié, les répartitions de Neyman sont souvent considérées pour déterminer les tailles minimales des échantillons et les répartitions optimales pour satisfaire aux exigences en matière de précision aux niveaux agrégés (Cochran 1977). Les répartitions exponentielles sont utiles pour trouver le juste équilibre entre les exigences de précision pour les agrégats et les strates (Bankier 1988). Dans la présente application, la taille minimale d'échantillon est fondée sur les exigences de précision pour les strates individuelles, c'est-à-dire les quartiers, qui constituent le niveau de publication le plus détaillé.

Si les exigences de précision sont spécifiées pour les catégories distinctes des répartitions du revenu, alors la catégorie de revenu ayant la plus grande variance de population détermine la taille minimale de l'échantillon requis, ce qui donne des tailles d'échantillon inutilement grandes. Comme solution de rechange, on propose d'utiliser plutôt la racine carrée de la moyenne des variances des catégories estimées des revenus d'une répartition du revenu comme mesure de précision pour les répartitions estimées des revenus. Avec cette mesure, l'influence de la catégorie de revenus la moins précise sur la taille minimale de l'échantillon est réduite. La racine carrée de la moyenne des variances des catégories estimées des revenus d'une répartition du revenu s'appelle la mesure de l'erreur type moyenne et est définie par

$$s = \sqrt{\frac{1}{L}\sum_{l=1}^{L} V\left(\hat{P}_{lr}\right)}. \tag{4.3}$$

Dans la présente section, on calcule une expression exacte pour s et on établit une approximation qui peut servir à estimer la taille minimale d'échantillon requise qui n'exige pas que l'on dispose de données à propos des répartitions du revenu ou des variances.

Comme les quartiers sont les régions les plus détaillées pour lesquelles les répartitions du revenu sont publiées, les exigences de précision pour la détermination de la taille d'échantillon sont spécifiées à ce niveau. Puisque les quartiers sont utilisés comme variable de stratification dans le plan d'échantillonnage, les expressions pour s peuvent être dérivées en vertu d'un échantillonnage aléatoire simple sans remise des personnes principales dans chaque quartier. On trouve en annexe la preuve qu'une expression de la mesure de l'erreur type moyenne s_h en (4.3) pour une répartition du revenu est donnée par

$$s_h = \sqrt{\frac{1}{L}\frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{l=1}^{L}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} - \sum_{l=1}^{L}\left(\frac{M_{lh}}{M_h}\right)^2\right)}, \tag{4.4}$$

avec M_h le nombre de ménages dans la strate h et M_{lh} le nombre de ménages de la strate h appartenant à la l^{e} catégorie de revenu. Soulignons que si $g_{kh} = 1$ pour tous les ménages de la population de la strate h, il s'ensuit que $M_h = N_h$ et la formule (4.1) est simplifiée comme suit :

$$V\left(\hat{P}_{lh}\right) = \frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(P_{lh}\left(1 - P_{lh}\right)\right),$$

ce qui correspond à la variance d'une fraction estimée en vertu d'un échantillonnage aléatoire simple sans remise (Cochran 1977, chapitre 3).

Le calcul des exigences quant à la taille minimale d'échantillon selon (4.4) exige des données sur la répartition du revenu et ses variances des périodes antérieures. Puisque ces données ne sont généralement pas disponibles à l'étape de la conception d'un panel, il est utile de fixer une borne supérieure pour la mesure de l'erreur type moyenne pour la répartition du revenu dans l'équation (4.4). Cela revient à considérer la variance comme un paramètre défini sous forme de proportion, qui atteint un maximum lorsque la proportion est de 0,5, pour calculer la taille minimale de l'échantillon requise pour une enquête. On présente en annexe la preuve qu'une borne supérieure pour la mesure de l'erreur type moyenne s_h relative à une répartition du revenu précisée en (4.4) est donnée par

$$s_h \leq \sqrt{\frac{1}{L}\frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{t=1}^{T}\frac{M_{th}}{t} - \frac{1}{L}\right)}, \tag{4.5}$$

avec M_{th} le nombre de ménages de taille t dans la strate h.

Si $g_{kh} = 1$ pour tous les ménages de la population de la strate h et si le nombre de catégories de la répartition du revenu $L = 2$, il s'ensuit que l'approximation de la mesure de l'erreur type moyenne s_h en (4.5) peut être simplifiée comme suit :

$$s_h \leq \sqrt{\frac{N_h - n_h}{n_h}\frac{1}{(N_h - 1)}\frac{1}{4}},$$

ce qui est égal à la racine carrée de la variance maximale d'une fraction estimée à $\hat{P} = 0{,}5$ sous un échantillonnage aléatoire simple. Ainsi, l'approximation de la mesure de l'erreur type moyenne de l'équation (4.5) peut être interprétée comme une généralisation de l'approximation de la variance maximale d'une fraction estimée à $\hat{P} = 0{,}5,$ souvent utilisée pour la détermination de la taille d'échantillon. La mesure de l'erreur type moyenne atteint sa valeur maximale dans le cas d'une répartition égale des ménages dans les catégories de revenus, c'est-à-dire $\hat{P}_{lh} = 1/L$ pour $l = 1, \ldots, L.$ Dans ce cas, l'approximation de s_h est exacte, ce qui découle directement de l'équation (4.3).

En fixant l'expression de s_h en (4.5) à une valeur maximale prédéterminée, par exemple $\Delta_h,$ on obtient l'expression suivante pour la taille minimale d'échantillon des personnes principales :

$$n_h \geq \frac{\left(\frac{N_h}{M_h}\right)^2 \sum_{t=1}^{T} \frac{M_{th}}{t} - \frac{N_h}{L}}{(N_h - 1)L\Delta_h^2 + \frac{N_h}{M_h^2}\sum_{t=1}^{T}\frac{M_{th}}{t} - \frac{1}{L}}. \tag{4.6}$$

Les données nécessaires pour estimer la taille minimale d'échantillon sont le nombre total de personnes et le nombre total de ménages de même taille dans les quartiers. On n'a besoin d'aucun renseignement sur la répartition prévue du revenu ou sur sa variance. Des estimations plus précises de la taille minimale d'échantillon peuvent être obtenues au moyen de l'expression (4.4), mais il faut pour cela disposer de données sur les répartitions du revenu, par exemple celles des périodes antérieures.

L'expression (4.6) donne la taille minimale d'échantillon pour les personnes principales. Par la suite, tous les membres du ménage de chaque personne principale sont inclus dans l'échantillon. Un même ménage peut donc être inclus plus d'une fois dans l'échantillon et la taille d'échantillon en termes de ménages uniques et de personnes uniques est aléatoire. Pour planifier une enquête et en gérer les coûts, il faut connaître le nombre espéré de ménages et de personnes uniques que l'on obtient si on tire un échantillon de personnes principales de taille $n_h.$ On trouve en annexe la preuve montrant que le nombre espéré de ménages uniques dans un échantillon de n_h personnes principales, tiré par échantillonnage aléatoire simple sans remise à partir d'une population finie de taille $N_h,$ est donné par

$$D_h = \sum_{t=1}^{T} M_{th}\left(1 - \frac{\prod_{i=0}^{t-1}(N_h - n_h - i)}{\prod_{i=0}^{t-1}(N_h - i)}\right). \tag{4.7}$$

Le nombre espéré de personnes uniques dans un échantillon de n_h personnes principales, tiré par échantillonnage aléatoire simple sans remise à partir d'une population finie de taille $N_h,$ découle directement de l'équation (4.7) et est donné par

$$D_h^{[p]} = \sum_{t=1}^{T} tM_{th} \left(1 - \frac{\prod_{i=0}^{t-1} (N_h - n_h - i)}{\prod_{i=0}^{t-1} (N_h - i)} \right). \tag{4.8}$$

Comme les nombres espérés de ménages et de personnes uniques sont des variables aléatoires, il est utile d'avoir une mesure d'incertitude pour ces valeurs espérées. Les expressions de la variance pour (4.7) et (4.8) ne peuvent toutefois pas être déterminées simplement et doivent donc faire l'objet d'autres recherches.

Les calculs relatifs à la taille d'échantillon sont effectués au niveau des quartiers. Il a été décidé de sélectionner les personnes principales selon une fraction de sondage de 0,16. Avec un échantillon de cette taille, la valeur maximale de la mesure de l'erreur type moyenne s_h au niveau des quartiers s'établit à environ 0,01 pour les répartitions estimées du revenu des ménages. Comme la population totale est d'environ 12 millions de personnes, on obtient un échantillon d'environ 2,1 millions de personnes principales, et un échantillon espéré d'environ 4,6 millions de personnes uniques. Cet échantillon a été tiré en 1994, l'année où le panel de la RIS néerlandaise a été mis sur pied.

5 Pondération linéaire

Pour des enquêtes auprès des ménages comme la RIS, il faut établir des estimations pour les caractéristiques des personnes et pour celles des ménages. Soit t_y la valeur totale de la variable cible y. Selon une pondération linéaire, un estimateur pour une variable cible fondée sur les personnes se définit comme suit :

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k \in 1}^{m_h} \sum_{j \in k} w_{kj} y_{kjh}, \tag{5.1}$$

avec y_{kjh} la valeur de la variable cible pour les personnes (k, j, h) et w_{kj} un poids pour la personne j appartenant au ménage k. Un estimateur de la variable cible fondée sur les ménages est donné par

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k=1}^{m_h} w_k y_{kh}, \tag{5.2}$$

avec y_{kh} la valeur de la variable cible pour le ménage k de la strate h et w_k un poids pour le ménage correspondant.

Les poids sont obtenus au moyen de l'estimateur GREG afin d'utiliser les variables auxiliaires observées dans l'échantillon et pour lesquelles on connaît les totaux de population grâce à d'autres sources (Särndal et coll. 1992). En conséquence, les poids reflètent les espérances (inégales) d'inclusion des unités d'échantillonnage et un ajustement faisant en sorte que pour les variables auxiliaires, la somme des observations pondérées corresponde aux totaux de population connus. Des variables catégoriques comme le sexe, l'âge, l'état matrimonial ou la région sont souvent utilisées comme variables auxiliaires. Comme les valeurs des variables auxiliaires varient d'une personne à l'autre dans le même ménage, différents poids

peuvent être calculés pour les personnes d'un même ménage. Pour s'assurer que la relation entre les variables du ménage et les variables des personnes est prise en compte dans les totaux estimés, il convient d'appliquer une méthode de pondération qui attribue un poids de ménage unique à tous les membres du ménage. Si les poids pour les personnes d'un ménage sont les mêmes, alors les estimations des mêmes variables cibles fondées sur le ménage et sur les personnes sont cohérentes entre elles (par exemple le revenu total estimé du ménage et celui des personnes). Pour ce faire, on peut utiliser des méthodes de pondération dites « intégrées ».

Lemaître et Dufour (1987) appliquent une méthode de pondération intégrée au niveau de la personne et remplacent les variables auxiliaires originales définies à ce niveau par la moyenne du ménage correspondant. Ainsi, les membres du même ménage ont la même espérance d'inclusion et partagent les mêmes données auxiliaires, ce qui fait que les poids de régression qui en découlent sont aussi forcément les mêmes. Nieuwenbroek (1993) propose une approche légèrement plus générale et applique une méthode de pondération linéaire au niveau du ménage, en vertu de laquelle les données auxiliaires sur les caractéristiques des personnes sont agrégées au niveau du ménage. Nieuwenbroek (1993) souligne que la méthode de pondération linéaire au niveau du ménage est équivalente à la méthode de pondération linéaire de Lemaître et Dufour (1987) au niveau de la personne, si la variance résiduelle du modèle de régression au niveau du ménage est choisie proportionnelle au nombre de personnes dans le ménage. Steel et Clark (2007) et Estevao et Särndal (2006) généralisent encore davantage la pondération intégrée dans les enquêtes auprès des personnes et des ménages. Steel et Clark (2007) ont étudié la question de savoir si les avantages cosmétiques de la pondération intégrée entraînent une variance accrue par rapport au plan dans les estimations GREG. Ils montrent que pour de grands échantillons, les variances par rapport au plan obtenues lorsqu'on applique une pondération linéaire au niveau du ménage sont inférieures ou égales à la variance par rapport au plan obtenue lorsqu'on applique une pondération linéaire au niveau de la personne. Pour de petits échantillons, il arrive que la pondération intégrée entraîne une légère augmentation de la variance par rapport au plan. On perd donc peu ou pas en efficience lorsqu'on applique une méthode de pondération intégrée.

Dans la présente étude, on applique la méthode de pondération intégrée au niveau du ménage. Soit $\mathbf{x}_{kh}$ un q-vecteur comprenant q variables auxiliaires pour le ménage k de la strate h. Les caractéristiques fondées sur la personne sont agrégées aux totaux pour le ménage. L'estimateur GREG est dérivé à partir d'un modèle de régression linéaire qui précise la relation entre la variable cible et les variables auxiliaires disponibles pour lesquelles les totaux de population sont connus et est défini par

$$y_{kh} = \mathbf{x}_{kh}^t \boldsymbol{\beta} + e_{kh}, \quad \text{avec} \quad \mathrm{E}_m(e_{kh}) = 0, \quad \mathrm{V}_m(e_{kh}) = \sigma_{kh}^2. \tag{5.3}$$

Dans (5.3), $\boldsymbol{\beta}$ désigne un vecteur comprenant les q coefficients de régression de la régression de y_{kh} sur $\mathbf{x}_{kh}$, e_{kh} désigne les résidus et E_m et V_m désignent l'espérance et la variance à l'égard du modèle de régression. Dans cette application, la structure de variance est considérée proportionnelle à la taille du ménage, c'est-à-dire $\sigma_{hk}^2 = g_k \sigma^2$. Nieuwenbroek (1993) montre que dans un tel cas, la pondération appliquée au niveau du ménage est égale à celle de la méthode de Lemaître et Dufour (1987).

Les poids de régression pour les ménages sont finalement obtenus par

$$w_k = \frac{1}{\pi_k}\left(1 + \left(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\boldsymbol{\pi}}\right)^t \left(\sum_{k=1}^{m} \frac{\mathbf{x}_{kh}\mathbf{x}_{kh}^t}{\pi_k g_k}\right)^{-1} \frac{\mathbf{x}_{kh}}{g_k}\right),$$

avec $\mathbf{t}_{\mathbf{x}}$ un vecteur q comprenant les totaux de population connus des variables auxiliaires $\mathbf{x}$, $\hat{\mathbf{t}}_{\mathbf{x}\boldsymbol{\pi}}$ l'estimateur HT pour $\mathbf{t}_{\mathbf{x}}$. Les poids calculés au niveau du ménage peuvent servir à pondérer les caractéristiques fondées sur la personne des membres du ménage correspondant, à l'aide de la formule énoncée en (5.1) puisque $w_{kj} = w_k$ pour toutes les personnes appartenant au même ménage k.

6 Estimation de la variance

Les paramètres de la RIS sont estimés sous forme de ratio de deux totaux de population :

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_z}, \tag{6.1}$$

où $\hat{t}_y$ et $\hat{t}_z$ sont des estimateurs GREG définis par (5.1) ou (5.2) selon que les variables cibles sont fondées sur la personne ou sur le ménage, respectivement. Une approximation de la variance de (6.1) par rapport à un plan d'échantillonnage où les personnes principales sont tirées par échantillonnage aléatoire simple stratifié et où tous les membres des ménages de ces personnes principales sont inclus dans l'échantillon peut être donnée par

$$V\left(\hat{R}\right) = \frac{1}{t_z^2}\sum_{h=1}^{H} \frac{N_h^2\left(1 - f_h\right)}{n_h} \frac{1}{N_h - 1}\sum_{k=1}^{N_h}\left(\frac{e_{kh}}{g_k} - \frac{1}{N_h}\sum_{k'=1}^{N_h}\frac{e_{k'h}}{g_{k'}}\right)^2, \tag{6.2}$$

où $f_h = n_h / N_h$, $e_{kh} = \left(y_{kh} - \mathbf{x}_{kh}^t \mathbf{b}_y\right) - R\left(z_{kh} - \mathbf{x}_{kh}^t \mathbf{b}_z\right)$, et $\mathbf{b}_y$ et $\mathbf{b}_z$ sont les coefficients de régression dans la population finie de la régression de y_{kh} et z_{kh}, respectivement, sur $\mathbf{x}_{kh}$. Un estimateur de la variance calculée par (6.2) est donné par

$$\hat{V}\left(\hat{R}\right) = \frac{1}{\hat{t}_z^2}\sum_{h=1}^{H}\left(1 - f_h\right)\frac{n_h}{n_h - 1}\sum_{k=1}^{n_h}\left(w_k \hat{e}_k - \frac{1}{n_h}\sum_{k'=1}^{n_h} w_{k'}\hat{e}_{k'h}\right)^2, \tag{6.3}$$

où $\hat{e}_{kh} = \left(y_{kh} - \mathbf{x}_{kh}^t \hat{\mathbf{b}}_y\right) - \hat{R}\left(z_{kh} - \mathbf{x}_{kh}^t \hat{\mathbf{b}}_z\right)$ et $\hat{\mathbf{b}}_y$ et $\hat{\mathbf{b}}_z$ sont les estimateurs de type HT pour $\mathbf{b}_y$ et $\mathbf{b}_z$. Ces résultats découlent directement de l'intégration des espérances d'inclusion de premier et de deuxième ordre calculées par les équations (3.3) à (3.6) dans l'approximation générale de la variance du ratio de deux estimateurs GREG et de son estimateur (Särndal et coll. 1992, section 7.13).

Les mêmes expressions pour la variance peuvent être dérivées des expressions de la variance proposées pour la méthode généralisée du partage des poids dans le cas d'un échantillonnage indirect. Lavallée (1995) a établi pour l'estimateur HT des expressions de la variance fondées sur le plan d'échantillonnage utilisé pour sélectionner l'échantillon s^A de n unités de la population U^A avec des variables cibles transformées, par exemple z_i. Dans cette application, chaque unité de U^A a exactement un lien avec une unité de U^B. Par conséquent, la variable z_i de Lavallée (1995) est dans ce cas définie comme la somme des variables

cibles de tous les éléments de la grappe k, divisée par le nombre d'unités de la grappe k ayant un lien avec la population U^A, c'est-à-dire $z_i = y_k / g_k$ pour tout $i \in U^A$ ayant un lien avec la grappe $k \in U^B$. En insérant les probabilités d'inclusion de premier et de deuxième ordre pour l'échantillonnage aléatoire simple stratifié sans remise et les variables transformées z_i (où la variable cible y_k est remplacée par le résidu de la régression sur les totaux de grappe e_k) dans la formule de variance pour un ratio, on obtient (6.2). L'expression (6.3) est obtenue de la même manière.

7 Application

Dans le cadre de la RIS, les personnes principales sont sélectionnées dans la population de 15 ans ou plus au moyen d'un échantillonnage aléatoire simple stratifié sans remise selon une fraction de sondage de 0,16. Dans la présente application, les résultats sont présentés pour une grande municipalité (Rotterdam), une municipalité de taille moyenne (Enschede) et une petite municipalité (Sevenum), pour trois années consécutives (2006, 2007 et 2008). Les tailles des populations et des échantillons pour ces trois municipalités sont indiquées dans le tableau 7.1.

Tableau 7.1
Taille de population et d'échantillon de la RIS pour trois municipalités néerlandaises

Municipalité	Population		Échantillon		
	Ménages	Personnes de 15 ans ou plus	Personnes principales	Ménages uniques	Personnes uniques
Rotterdam	293 400	484 000	73 000	67 600	171 400
Enschede	74 200	128 000	19 300	17 600	46 300
Sevenum	2 950	6 100	870	750	2 500

Les variables cibles d'intérêt pour la RIS sont les suivantes :

- Répartition du revenu des ménages dans dix catégories fondées sur des quantiles de dix points de pourcentage (déciles) de la répartition nationale, selon le revenu du ménage normalisé (abrégée RépRevMén);
- Revenu moyen normalisé du ménage (abrégé RevMén);
- Revenu disponible moyen des personnes ayant un revenu durant les 52 semaines de l'année (abrégé RevP).

Le revenu disponible d'une personne s'entend du revenu total d'une personne après impôt. Le revenu total comprend la rémunération, les profits, le revenu tiré du capital et de l'épargne, ainsi que les avantages sociaux et autres avantages. Le revenu normalisé du ménage s'entend du revenu disponible total d'un ménage, corrigé pour tenir compte des différences dans la taille et la composition du ménage. Dans les ouvrages publiés, on parle aussi de revenu disponible équivalent (OECD 2013).

Les estimations destinées aux publications officielles relatives à la RIS sont obtenues à l'aide de l'estimateur GREG selon la méthode de Lemaître et Dufour (1987). Comme l'enquête n'est pas touchée par la non-réponse, les données auxiliaires sont utilisées dans l'estimation pour réduire la variance et pour la

cohérence entre les cellules marginales des différents tableaux publiés. Les espérances d'inclusion sont fondées sur les formules dérivées à la sous-section 3.1. Pour chaque municipalité, on applique le schéma de pondération suivant dans l'estimateur GREG :

$$\text{Âge}\,(7) \times \text{Sexe} + \text{Âge}\,(4) \times \text{Sexe} \times \text{État matrimonial}\,(2) + \text{Adresse}\,(2) \times \text{Taille du ménage}\,(5).$$

Toutes les variables auxiliaires sont catégoriques. Les chiffres entre parenthèses correspondent au nombre de catégories. L'état matrimonial fait la distinction entre les personnes mariées et les autres états matrimoniaux. L'adresse fait la distinction entre les adresses où réside une famille et les autres types d'adresse. La taille du ménage fait la distinction entre les ménages comptant une, deux, trois, quatre et cinq personnes ou plus. Les estimations pour RevMén et RevP incluant les erreurs types fondées sur l'estimateur HT, l'estimateur GREG et l'estimateur GREG selon la méthode de Lemaître et Dufour (1987) sont données au tableau 7.2. À la figure 7.1, les répartitions du revenu RépRevMén estimées à l'aide de l'estimateur HT, de l'estimateur GREG et de l'estimateur GREG selon la méthode de Lemaître et Dufour (1987) sont représentées selon un intervalle de confiance à 95 % pour Rotterdam et Sevenum en 2008. Les erreurs types pour ces estimations sont comparées dans un histogramme distinct. À la figure 7.2, la RépRevMén pour Rotterdam et Sevenum estimée selon la méthode de Lemaître et Dufour (1987) est donnée pour 2006, 2007 et 2008. Voir van den Brakel (2013) pour en savoir davantage sur les répartitions du revenu.

Tableau 7.2
Résultats des estimations dans le cadre de la RIS pour Rotterdam (grande ville), Enschede (ville de taille moyenne) et Sevenum (petit village); erreurs types entre parenthèses

	Variable	Année	HT		GREG		GREG convergent (L et D)	
Rotterdam	RevMén	2006	19 790	(83)	20 134	(80)	20 161	(76)
		2007	22 306	(73)	22 950	(64)	22 866	(64)
		2008	23 750	(78)	24 511	(69)	24 410	(68)
	RevP	2006	22 074	(94)	22 219	(84)	22 233	(93)
		2007	24 094	(82)	24 362	(75)	24 432	(78)
		2008	25 325	(84)	25 625	(75)	25 705	(78)
Enschede	RevMén	2006	19 810	(128)	20 353	(111)	20 300	(107)
		2007	20 878	(128)	21 716	(107)	21 753	(105)
		2008	22 254	(148)	23 235	(125)	23 237	(123)
	RevP	2006	20 402	(102)	20 608	(92)	20 590	(92)
		2007	21 387	(115)	21 751	(103)	21 852	(106)
		2008	22 235	(123)	22 659	(110)	22 724	(114)
Sevenum	RevMén	2006	25 696	(799)	25 698	(734)	25 968	(711)
		2007	28 207	(618)	28 901	(520)	29 026	(490)
		2008	31 466	(795)	32 372	(715)	32 536	(694)
	RevP	2006	21 328	(466)	21 680	(428)	21 712	(428)
		2007	24 056	(456)	24 219	(396)	24 459	(393)
		2008	24 980	(468)	25 482	(426)	25 644	(455)

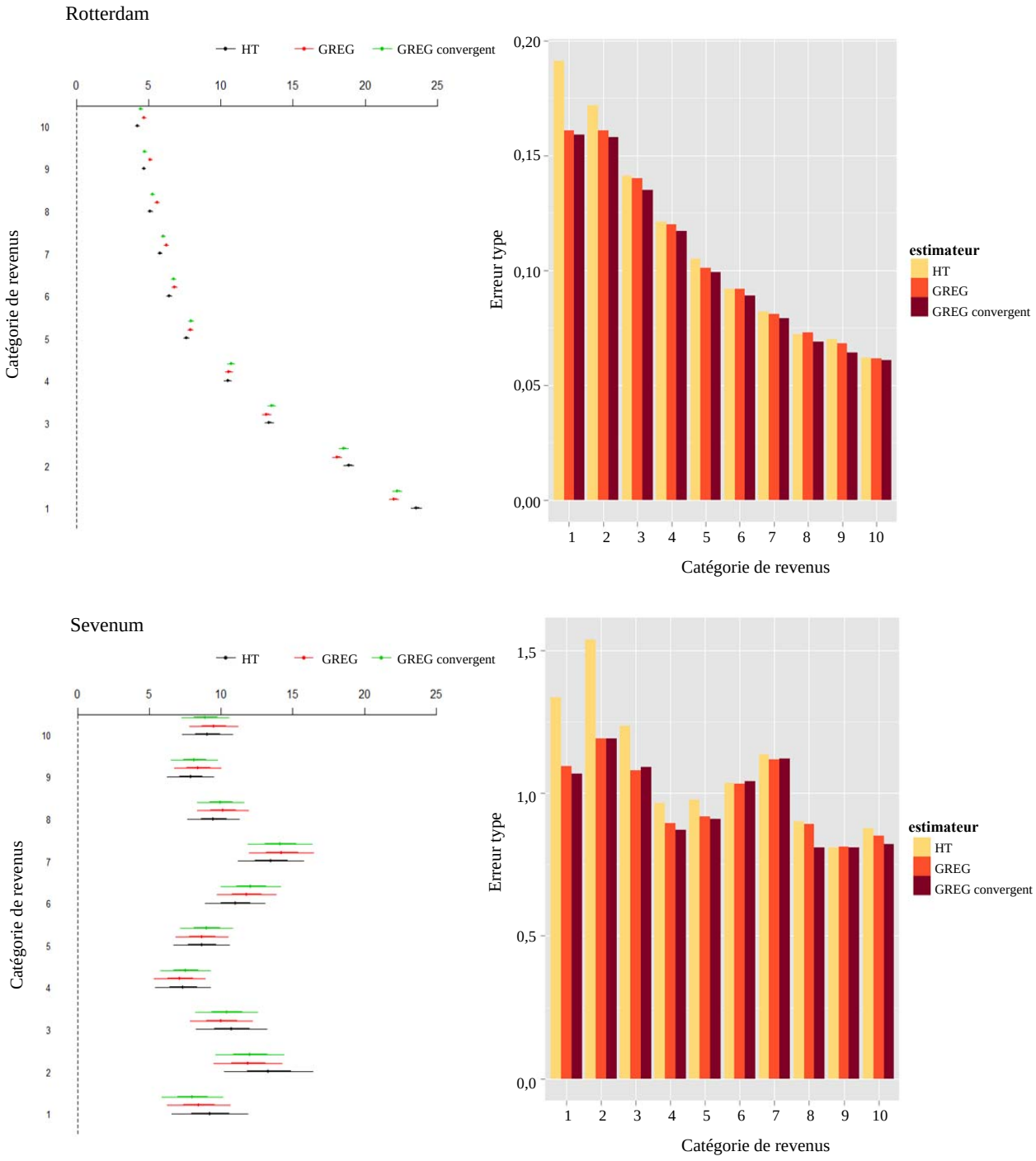

Figure 7.1 RépRevMén en pourcentage pour Rotterdam et Sevenum (à gauche) selon l'estimateur de Horvitz-Thompson, l'estimateur GREG et l'estimateur GREG intégré (GREG convergent), avec intervalles de confiance à 95 %; les erreurs types des estimateurs correspondants sont illustrées à droite.

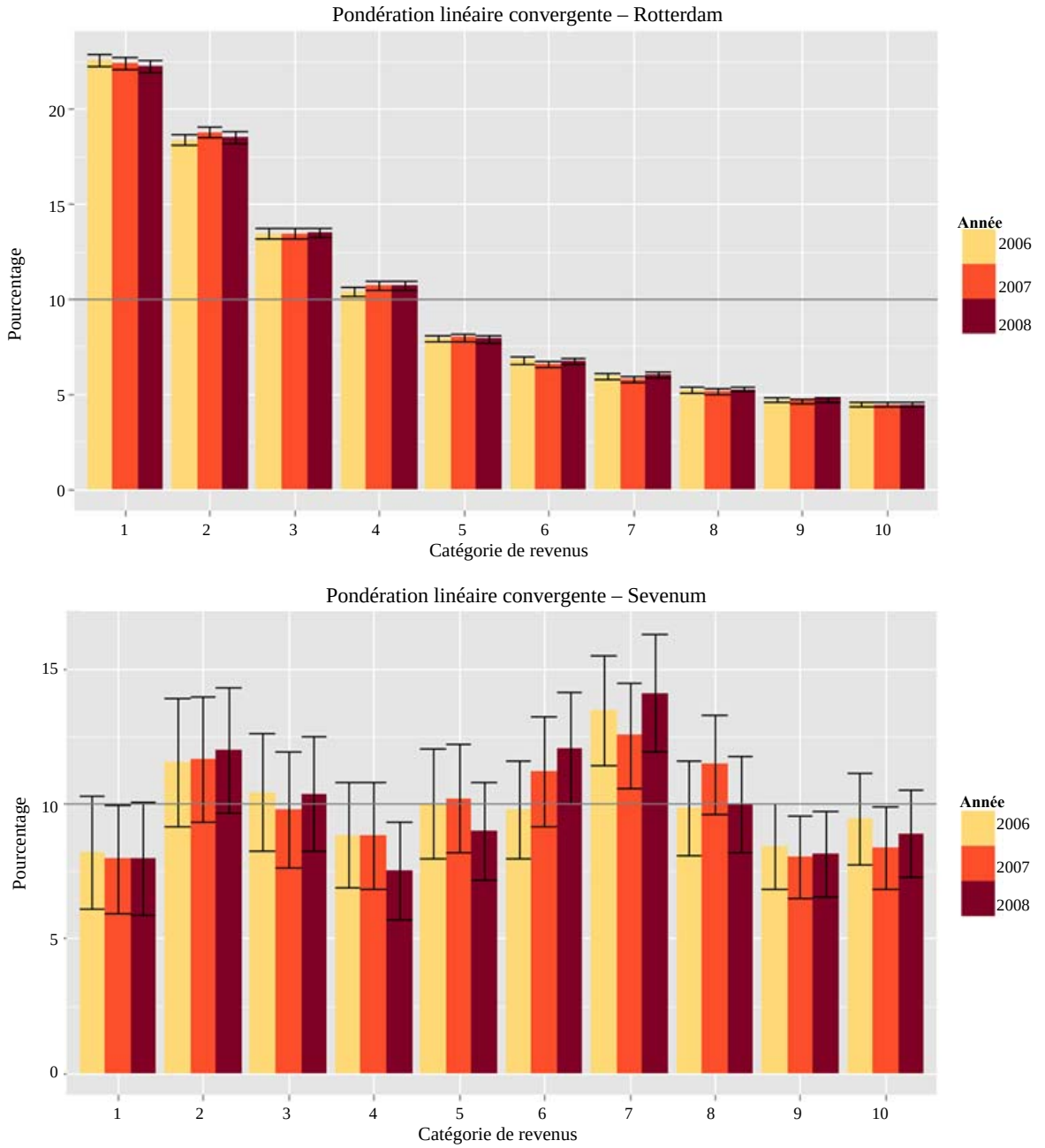

Figure 7.2 RépRevMén en pourcentage pour Rotterdam (en haut) et Sevenum (en bas) estimées selon une pondération intégrée pour 2006, 2007 et 2008 avec intervalles de confiance à 95 %; la ligne grise représente la répartition nationale du revenu.

Les répartitions du revenu observées illustrées aux figures 7.1 et 7.2 résultent de la composition démographique de chacune des deux municipalités. Rotterdam est une ville où la fraction des ménages se trouvant dans les catégories inférieures de revenus est supérieure à la moyenne nationale, les fractions des

trois premières catégories étant supérieures à 10 %. Le nombre de ménages se trouvant dans les catégories supérieures de revenus, en revanche, est inférieur à la moyenne nationale, les fractions de ces catégories étant inférieures à 10 %. Il s'agit d'une répartition type pour une grande ville universitaire où on trouve une fraction élevée d'immigrants non occidentaux. En revanche, Sevenum est un petit village situé près d'une grande ville industrielle. De tels villages ont généralement de petites fractions d'immigrants, pas d'étudiants et des fractions importantes de ménages comptant une ou deux personnes recevant un revenu 52 semaines par année. Cela explique pourquoi la fraction des ménages se trouvant dans la catégorie inférieure de revenus est sous la moyenne nationale, alors que la fraction des ménages se trouvant dans les catégories supérieures de revenus (6, 7 et 8) est supérieure à la moyenne nationale. Sevenum est un village qui n'attire pas les ménages extrêmement riches.

Comme RevMén et RevP sont fondées sur des définitions différentes du revenu et que RevP représente la moyenne des domaines des personnes qui reçoivent un revenu 52 semaines par année, les différences entre les deux moyennes varient d'une municipalité à l'autre. Dans une grande ville universitaire comme Rotterdam, le revenu moyen normalisé des ménages est généralement plus faible que la moyenne du revenu personnel disponible pour l'ensemble des personnes qui reçoivent un revenu 52 semaines par année. D'autres villes où se trouvent de grandes universités présentent un profil semblable. Dans un village petit mais riche comme Sevenum, la situation est inversée.

On remarque en outre qu'à Rotterdam et à Enschede, l'écart entre l'estimateur HT et l'estimateur GREG est relativement grand par rapport aux erreurs types. Compte tenu du grand échantillon et du fait qu'il n'y a pas de non-réponse, ces différences devraient être plus petites. Cela pourrait s'expliquer par le fait que Rotterdam et Enschede sont de grandes villes universitaires. Les étudiants sont souvent inscrits dans les registres de l'impôt (qui sont utilisés comme base de sondage) d'une manière différente que dans les registres de population (qui sont utilisés pour dériver les répartitions des variables auxiliaires dans la population), particulièrement en ce qui concerne la situation de leur ménage.

Pour chaque municipalité, on constate au fil du temps une augmentation constante de la moyenne du revenu des ménages et des personnes. De plus, les répartitions du revenu dans chaque municipalité affichent des tendances stables au fil des ans. Il s'agit là de résultats prévisibles si un panel est réalisé sur de grands échantillons pour estimer des phénomènes qui ne sont pas très volatils dans le temps.

La comparaison des estimations GREG avec et sans recours à la méthode de Lemaître et Dufour (1987) montre que les erreurs types des paramètres estimés des ménages sont plus faibles si on utilise la méthode de Lemaître et Dufour (1987). La différence est particulièrement visible lorsqu'on observe le revenu moyen des ménages dans le petit échantillon de Sevenum. En revanche, pour les paramètres estimés fondés sur la personne, la méthode de Lemaître et Dufour (1987) entraîne une erreur type légèrement plus élevée que celle de l'estimateur GREG ordinaire. Ce résultat donne à penser que la structure de variance présumée pour les résidus du modèle de régression sous-jacent dans le cas de la pondération intégrée convient mieux aux variables fondées sur le ménage qu'aux variables fondées sur la personne.

8 Discussion

En raison de leur instabilité au fil du temps, les ménages ne constituent pas des unités d'échantillonnage appropriées dans les panels visant à recueillir des données au niveau des ménages ou des personnes. Dans

le présent article, on propose un plan d'échantillonnage où les personnes sont tirées à l'aide d'un plan d'échantillonnage autopondéré. À chaque point dans le temps, les membres du ménage de ces personnes dites « principales » sont inclus dans l'échantillon. On obtient ainsi un échantillon où les ménages peuvent être tirés plus d'une fois, jusqu'à concurrence du nombre de personnes dans le ménage. Les ménages sont inclus selon une espérance proportionnelle à la taille du ménage. Les espérances d'inclusion de premier et de deuxième ordre pour les ménages sont calculées en vertu d'un plan d'échantillonnage à probabilités égales pour la sélection des personnes principales. Ces espérances d'inclusion peuvent être utilisées d'une façon comparable aux probabilités d'inclusion plus couramment employées pour l'inférence fondée sur le plan et assistée par modèle.

Le plan d'échantillonnage proposé dans la présente étude constitue un cas particulier de la méthode d'échantillonnage indirect (Lavallée 1995, 2007). Dans le cas d'un plan d'échantillonnage autopondéré, il est montré qu'il est possible de dériver d'une manière relativement simple les espérances d'inclusion de premier et de deuxième ordre pour ce plan d'échantillonnage à partir de la composition des ménages des personnes principales à chaque point dans le temps. Dans le cas des plans d'échantillonnage plus complexes, il faut employer la méthode généralisée du partage des poids (Lavallée 1995, 2007) pour établir les poids d'inclusion à chaque point dans le temps.

L'avantage du plan d'échantillonnage proposé est que la méthode d'estimation est plus simple que la méthode généralisée du partage des poids. Le plan est particulièrement utile si les personnes principales sont sélectionnées à l'aide d'un plan d'échantillonnage autopondéré. Si toutefois on a besoin, par exemple à cause des exigences minimales en matière de précision et maximales en matière de coûts, d'un plan à probabilités inégales pour la sélection des personnes principales, il faut alors utiliser la méthode généralisée du partage des poids. Comme les personnes principales demeurent dans le panel pour une période indéterminée, ce plan d'échantillonnage convient particulièrement bien aux panels auprès des ménages fondés sur des registres, en vertu desquels toute l'information requise est tirée de données administratives. Dans le cas des panels auprès des ménages fondés sur des interviews, il faut mettre en place un plan de renouvellement afin de remédier à certains problèmes comme l'attrition des participants.

La présente étude propose la mesure dite de l'erreur type moyenne, soit la racine carrée de la moyenne des variances des catégories des revenus estimés d'une répartition des revenus, comme mesure de précision pour déterminer la taille minimale d'échantillon. On montre que la valeur maximale de cette mesure de précision correspond à une distribution où les proportions dans les catégories sont égales. On montre aussi que ce résultat peut être vu comme une généralisation de la variance d'une fraction qui atteint sa valeur maximale à 0,5. On dérive ensuite une expression de la taille minimale d'échantillon requise pour satisfaire une précision prédéterminée pour les répartitions estimées. Comme un même ménage peut être inclus plus d'une fois dans l'échantillon, on dérive aussi une expression pour déterminer le nombre prévu de ménages uniques dans l'échantillon.

D'autres recherches sont nécessaires pour étudier la combinaison de cette mesure de l'erreur type moyenne avec une répartition de Neyman ou avec des répartitions exponentielles afin d'établir des expressions pour la taille minimale d'échantillon fondées sur les exigences en matière de précision pour les répartitions estimées au niveau des strates agrégées. On obtient actuellement un plan à probabilités inégales d'inclusion pour les personnes principales, pour lequel il faut employer la méthode généralisée du partage des poids afin de calculer les poids appropriés.

Dans le contexte des enquêtes et des panels auprès des ménages, il convient d'employer des méthodes de pondération qui appliquent des poids de régression égaux aux personnes d'un même ménage afin d'assurer la cohérence des estimations fondées sur les personnes et sur les ménages. Dans le cadre de la présente étude, on utilise pour la RIS une approche de pondération intégrée fondée sur les travaux de Lemaître et Dufour (1987). Les erreurs types obtenues en vertu de la méthode de Lemaître et Dufour (1987) sont plus faibles que celles qu'on obtient avec une méthode de pondération non intégrée pour les estimations fondées sur les ménages. Dans le cas des estimations fondées sur les personnes, les erreurs types peuvent être légèrement plus grandes. Ces résultats sont conformes à ceux de Steel et Clark (2007), qui montrent que la variance par rapport à un plan de sondage en grand échantillon avec pondération intégrée au niveau du ménage est plus faible ou égale à la variance par rapport au plan obtenue avec une pondération non intégrée au niveau de la personne. Ils rapportent aussi que leur simulation a donné de faibles augmentations des variances par rapport au plan à cause de la pondération intégrée utilisée sur des échantillons de petite taille.

La pondération intégrée de Lemaître et Dufour (1987) au niveau du ménage est obtenue grâce à une structure de variance pour les résidus proportionnelle à la taille du ménage (Nieuwenbroek 1993). Si les caractéristiques du ménage sont proportionnelles à la taille du ménage, on peut s'attendre à ce qu'une telle structure de variance explique mieux la variation des variables des ménages dans la population, comparativement à une structure de variance qui suppose une variance résiduelle égale pour les ménages. Dans le cas des variables fondées sur la personne, une telle structure de variance pourrait être moins efficiente, mais la pondération intégrée présente l'avantage supplémentaire que les totaux pour le revenu fondés sur les ménages et sur les personnes, qui peuvent être dérivés directement à partir des moyennes, sont cohérents.

Remerciements

Les opinions exprimées dans l'article sont celles de l'auteur et ne reflètent pas les politiques de *Statistics Netherlands*. L'auteur remercie le rédacteur en chef adjoint et les examinateurs anonymes de leurs commentaires constructifs au sujet de deux versions antérieures du présent article, ainsi que Drs. M. van den Brakel-Hofmans pour avoir mis à sa disposition les données de la RIS.

Annexe technique

Preuve de l'équation (4.4)

Une expression pour la variance de la fraction estimée des ménages de la catégorie de revenus l peut être dérivée de l'expression générale pour la variance de l'estimateur HT (Särndal et coll. 1992, section 2.8) :

$$V\left(\hat{P}_{lh}\right) = \frac{1}{M_h^2}\sum_{k=1}^{M_h}\sum_{k'=1}^{M_h}\left(\pi_{kk'h} - \pi_{kh}\pi_{k'h}\right)\frac{y_{khl}}{\pi_{kh}}\frac{y_{k'hl}}{\pi_{k'h}}. \qquad \text{(A.1)}$$

Si l'on insère les espérances d'inclusion de premier et de deuxième ordre précisées aux expressions (3.3) à (3.6) et si l'on exploite l'égalité $y_{khl} = y_{khl}^2$, puisque les valeurs de la variable cible sont limitées à zéro ou un, il s'ensuit, après quelques manipulations algébriques, que l'expression (A.1) peut être simplifiée comme suit :

$$V\left(\hat{P}_{lh}\right) = \frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} - \left(\frac{M_{lh}}{M_h}\right)^2\right). \tag{A.2}$$

On obtient l'équation (4.4) en insérant (A.2) dans (4.3).

Preuve de l'équation (4.5)

La *population* des ménages de la strate h peut être divisée en T sous-populations de ménages de taille égale. Soit M_{th} le nombre de ménages de taille t dans la strate h. Il s'ensuit maintenant pour la double sommation entre parenthèses pour l'expression de s en (4.4) que

$$\sum_{l=1}^{L}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} = \sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{k=1}^{M_{th}}\frac{y_{khl}}{t} = \sum_{t=1}^{T}\frac{M_{th}}{t}. \tag{A.3}$$

En vertu de l'inégalité de Cauchy-Schwartz (Cochran 1977, section 5.5), il s'ensuit pour la sommation entre parenthèses pour l'expression de s_h en (4.4) que

$$\sum_{l=1}^{L}\left(\frac{M_{lh}}{M_h}\right)^2 = \sum_{l=1}^{L}P_{lh}^2 \geq \frac{1}{L}. \tag{A.4}$$

On obtient l'équation (4.5) en insérant (A.3) et (A.4) dans l'expression de s en (4.4).

Preuve de l'équation (4.7)

Soit $\tilde{\pi}_{tkh}$ la probabilité d'inclusion pour le ménage k de la strate h de taille t. Comme les ménages de même taille ont les mêmes probabilités de premier ordre, il s'ensuit que $\tilde{\pi}_{tkh} = \tilde{\pi}_{tk'h} \equiv \tilde{\pi}_{th}$. Soit I_{tkh} une variable indicatrice, qui prend la valeur 1 si le ménage k de la strate h de taille t est inclus dans l'échantillon, et la valeur 0 autrement. Le nombre prévu de ménages uniques peut être calculé comme suit :

$$\begin{aligned} D_h &= \mathrm{E}\left(\sum_{t=1}^{T}\sum_{k=1}^{M_{th}} I_{tkh}\right) = \sum_{t=1}^{T} M_{th}\tilde{\pi}_{th} \\ &= \sum_{t=1}^{T} M_{th}\left(1 - \frac{\binom{N_h - t}{n_h}}{\binom{N_h}{n_h}}\right) = \sum_{t=1}^{T} M_{th}\left(1 - \frac{(N_h - n_h)(N_h - n_h - 1)\ldots(N_h - n_h - t + 1)}{N_h(N_h - 1)\ldots(N_h - t + 1)}\right). \end{aligned}$$

Bibliographie

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Bethlehem, J.G. (2009). *Applied Survey Methods*, New Jersey: John Wiley & Sons, Inc.

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

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, 32, 2, 185-196.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley & Sons, Inc., 135-159.

Estevao, V.M., et Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *Revue Internationale de Statistique*, 74, 127-147.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Kalton, G., et Brick, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 1, 37-49.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 1, 27-35.

Lavallée, P. (2007). *Indirect Sampling*, New York: Springer Verlag.

Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 2, 211-220.

Lynn, P. (2009). Methods for longitudinal surveys. Dans *Methodology of Longitudinal Surveys,* (Éd., P. Lynn), Wiley, Chichester, 1-19.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

Nieuwenbroek, N.J. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. Document de recherche, BPA nr: 8555-93-M1-1, Statistics Netherlands, Heerlen.

OECD (2013). *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*. OECD publishing, http://dx.doi.org/10.1787/9789264194830-en.

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

Smith, P., Lynn, P. et Elliot, D. (2009). Sample design for longitudinal surveys. Dans *Methodology of Longitudinal Surveys*, (Éd., P. Lynn), Wiley, Chichester, 21-33.

Steel, D.G., et Clark, R.G. (2007). Estimation par la régression au niveau de la personne et au niveau du ménage dans les enquêtes-ménages. *Techniques d'enquête*, 33, 1, 59-69.

van den Brakel, J.A. (2013). Sampling and estimation techniques for household panels. Document de discussion 2013-15, Statistics Netherlands, Heerlen. http://www.cbs.nl/NR/rdonlyres/B4F85FB9-52F2-4B8A-94C4-56DA43F2250D/0/201315x10pub.pdf.

Wallgren, A., et Wallgren, B. (2007). *Register-Based Statistics: Administrative Data for Statistical Purposes*. New York: John Wiley & Sons, Inc.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 173-190
Statistique Canada, nº 12-001-X au catalogue

Ajustements pour la non-réponse dans les plans stratifiés assortis de modèles aux spécifications erronées

Ismael Flores Cervantes et J. Michael Brick[1]

Résumé

L'ajustement des poids de base au moyen de classes de pondération est une méthode communément employée pour composer avec la non-réponse totale. Une approche courante consiste en l'application d'ajustements pour la non-réponse pondérés selon l'inverse de la propension à répondre supposée des répondants dans les classes de pondération en vertu d'une méthode de quasi-randomisation. Little et Vartivarian (2003) ont remis en question l'utilité de la pondération du facteur d'ajustement. Dans la pratique, les modèles utilisés sont mal spécifiés; il est donc essentiel de comprendre l'incidence que peut avoir la pondération dans un tel cas. Le présent article décrit les effets, sur les estimations corrigées pour la non-réponse de moyennes et de totaux pour l'ensemble de la population et pour certains domaines qui ont été calculés selon l'inverse pondéré et non pondéré de la propension à répondre en vertu de plans d'échantillonnage aléatoires simples stratifiés. Le rendement de ces estimateurs est évalué dans différentes conditions, par exemple selon des répartitions différentes de l'échantillon, le mécanisme de réponse et la structure de population. Les résultats montrent que pour les scénarios étudiés, l'ajustement pondéré présente des avantages considérables pour l'estimation des totaux, et que le recours à un ajustement non pondéré peut donner lieu à des biais importants, sauf dans des cas très limités. En outre, contrairement aux estimations non pondérées, les estimations pondérées ne sont pas sensibles à la façon dont la répartition de l'échantillon est faite.

Mots-clés : Non-réponse; stratification; poids d'échantillonnage; repondération des classes de pondération.

1 Introduction

L'ajustement des poids de base au moyen de classes de pondération pour tenir compte de la non-réponse totale est une méthode couramment employée pour pondérer les données d'enquête, mais les chercheurs et les organismes d'enquête ne font pas tous ces ajustements de la même manière. Little et Vartivarian (2003), ci-après désignés « L et V », constatent que le recours à un facteur d'ajustement pour la non-réponse pondéré en fonction de l'inverse de la probabilité de sélection semble être l'approche la plus courante. Ils soulignent aussi que le fait d'utiliser des poids de sondage pour calculer un ajustement pondéré pour la non-réponse n'élimine pas le biais de non-réponse dans les estimations de la moyenne de population lorsque le mécanisme de réponse n'est pas précisé correctement dans le modèle d'ajustement de la pondération. L et V ont donc réalisé une étude par simulation à l'aide d'un plan d'échantillonnage simple stratifié afin d'examiner l'effet de la pondération des facteurs d'ajustement pour la non-réponse. Ils ont conclu que la pondération de l'ajustement pour la non-réponse est peu utile, voire inutile.

Afin d'éliminer le biais de non-réponse, les justifications théoriques pour l'ajustement pour la non-réponse exigent une modélisation exacte soit du mécanisme de réponse, soit de la variable cible; nous ne connaissons aucune théorie stipulant que la pondération selon l'inverse de la probabilité de sélection élimine complètement le biais lorsque les spécifications du modèle sont erronées (par exemple Kalton 1983; Little 1986; Little et Rubin 2002; Särndal et Lundström 2005). C'est pourquoi l'intégration dans la modélisation de l'ajustement pour la non-réponse que préconisent L et V est essentielle à une bonne pratique

1. Ismael Flores Cervantes et J. Michael Brick, Westat, 1600 Research Blvd, Rockville, Maryland, États-Unis, 20850. Courriel : ismaelflorescervantes@westat.com.

statistique. Toutefois, la spécification exacte d'un modèle hautement prédictif est un objectif qu'il n'est pas possible d'atteindre dans la plupart des enquêtes à cause de la complexité du phénomène et du fait qu'il existe rarement des variables auxiliaires suffisamment puissantes. Les recherches visant à trouver de meilleures données auxiliaires pour cette modélisation ont mené à l'exploration des paradonnées, mais les modèles qui font appel à ces données sont toujours associés à de faibles corrélations avec la propension à répondre (Kreuter, Olson, Wagner, Yan, Ezzati-Rice, Casas-Cordero, Lemay, Peytchev, Groves et Raghunathan 2010). Dans la pratique, on a recours à des modèles imparfaits et le biais de non-réponse n'est jamais complètement éliminé.

En conséquence, il importe de comprendre les effets des méthodes d'ajustement pour la non-réponse et de déterminer s'il est utile de pondérer l'ajustement pour la non-réponse lorsque les spécifications du modèle de réponse sont erronées. Bien que L et V insistent entre autres sur la nécessité d'inclure les variables de plan dans la modélisation de la non-réponse, certains chercheurs semblent avoir conclu que la pondération de l'ajustement est inutile (par exemple Chadborn, Baster, Delpech, Sabin, Sinka, Rice et Evans 2005; Haukoos et Newgard 2007). Cependant, la conclusion de L et V, selon laquelle la pondération du facteur d'ajustement pour la non-réponse est incorrecte ou inefficace, est fondée sur des comparaisons avec des modèles correctement spécifiés qui produisent toujours des estimations non biaisées. Leur suggestion de conditionner le modèle sur les variables de plan (dans le scénario de L et V, la variable de plan correspondait à la strate) a donné lieu à des estimateurs avec et sans pondération identiques. Leurs simulations étaient aussi axées sur un plan d'échantillonnage stratifié spécifique et ils n'ont tenu compte que de l'estimation des moyennes. Comme il est expliqué plus loin, ces limitations sont considérables et il convient de revoir les conclusions de certains quant à l'inutilité de pondérer l'ajustement.

Après L et V, des chercheurs ont examiné les effets de la pondération dans d'autres cas. Sukasih, Jang, Vartivarian, Cohen et Zhang (2009) ont comparé les ajustements pour la non-réponse avec et sans pondération à l'aide de simulations dans le contexte d'une enquête particulière. West (2009) a utilisé une simulation pour étudier les estimations des moyennes de population en vertu de plans d'échantillonnage plus complexes comprenant des grappes et des taux d'échantillonnage différentiels. Ces deux études ont conclu que la pondération des ajustements pour la non-réponse à l'aide des poids de sondage était utile comparativement à une approche de non-pondération, même si les différences obtenues après pondération n'étaient pas importantes. Après avoir évalué la robustesse des ajustements sur le plan théorique et décrit les conditions en vertu desquelles les divers estimateurs des moyennes de population étaient le moins influencés par le biais de non-réponse, Kott (2012) recommande une approche de pondération. D'autres recherches ont été menées sur la nécessité de pondérer pour estimer les coefficients des modèles de la propension à répondre (Wun, Ezzati-Rice, Diaz-Tena et Greenblatt 2007; Grau, Potter, Williams et Diaz-Tena 2006), mais cette piste de recherche est assez éloignée de la nôtre et nous ne l'abordons pas ici.

Dans le présent article, nous explorons l'effet de la pondération des ajustements pour la non-réponse lorsque le modèle de non-réponse est imparfait. Dans la section 2, nous prenons les résultats de L et V comme point de départ, pour aller plus loin et examiner les estimateurs pour les totaux et pour les moyennes et totaux de domaine; L et V n'ont tenu compte que des moyennes globales. À l'aide de la même population et du même scénario de simulation de base que L et V, nous examinons aussi l'effet de différentes répartitions de l'échantillon dans les strates, tandis que L et V n'ont utilisé qu'une seule répartition de

l'échantillon. Les résultats des simulations présentés à la section 3 révèlent des différences importantes des propriétés des estimateurs avec et sans pondération, qui varient selon la répartition de l'échantillon. Nous expliquons les comportements des estimateurs à l'aide d'approximations simples afin d'illustrer pourquoi ils sont différents. Bien que la pondération des facteurs d'ajustement ne donne pas toujours des estimations assorties d'un biais et d'une racine de l'erreur quadratique moyenne (reqm) plus faibles que ceux des estimations obtenues sans pondération, elle présente des avantages substantiels pour les estimations des totaux et fournit une protection contre les erreurs importantes qui pourraient découler d'une approche sans pondération. En conséquence, nous recommandons de pondérer lorsque le véritable mécanisme de réponse n'est pas entièrement connu. La section 4 donne les conclusions.

2 Scénario

Les poids de sondage compensent pour différents types de données manquantes : les poids d'échantillonnage ou de base compensent les unités non échantillonnées; les poids d'ajustement de non-couverture tiennent compte des unités qui ne font pas partie de la base de sondage; et les poids d'ajustement pour la non-réponse compensent les unités qui font partie de l'échantillon, mais qui ne répondent pas. Nous nous concentrons ici sur les poids d'ajustement pour la non-réponse et sur l'effet du recours aux poids de base pour établir les ajustements pour la non-réponse.

Commençons par l'estimateur de Horvitz-Thompson non ajusté pour le total :

$$\hat{y}_{na} = \sum_s R_i d_i y_i, \tag{2.1}$$

où d_i est l'inverse de la probabilité de sélection de l'unité $i, R_i = 1$ si l'unité i répond et $= 0$ autrement; la somme est calculée pour l'ensemble des unités de l'échantillon s. La moyenne du ratio est $\hat{\bar{y}}_{na} = \hat{y}_{na} \big/ \sum_s R_i d_i$. Si toutes les données de l'échantillon sont observées et que la base de sondage est complète, alors $E(\hat{y}_{na}) = Y$, et la moyenne du ratio est convergente pour $\bar{Y}$.

Lorsqu'il y a non-réponse totale, on présume que la réponse est une variable aléatoire et que la probabilité de réponse ou la propension à répondre $(\phi_i = \Pr(R_i = 1))$ correspond à la probabilité pour une phase supplémentaire d'échantillonnage (Särndal, Swensson et Wretman 1992). Si on suppose que $\phi_i > 0$ pour toutes les valeurs de i, alors le biais de non-réponse d'une moyenne de ratio estimée en vertu du modèle stochastique correspond à :

$$\text{biais}(\hat{\bar{y}}_{na}) \approx \bar{\phi}^{-1}\sigma_\phi\sigma_y\rho_{\phi,y}, \tag{2.2}$$

où $\bar{\phi}$ correspond à la moyenne de population des propensions à répondre, σ_ϕ est l'écart type de ϕ, σ_y est l'écart type de y et $\rho_{\phi,y}$ est la corrélation entre ϕ et y (Bethlehem 1988). La moyenne estimée pour les répondants est non biaisée si ϕ et y ne sont pas corrélés. Brick et Jones (2008) élargissent ces résultats à d'autres types de statistiques et d'estimateurs.

Pour réduire le biais de non-réponse, on peut utiliser les variables auxiliaires associées à l'échantillon pour étayer les ajustements pour la non-réponse en fonction des poids de base. Les ajustements peuvent être mis en œuvre par modélisation de la répartition de ϕ ou de y, ou encore des deux à l'aide des variables auxiliaires. Nous nous intéressons particulièrement à la modélisation du mécanisme de réponse.

Les propensions à répondre estimées sont appliquées comme si elles correspondaient aux probabilités réelles de réponse. En d'autres termes, le facteur d'ajustement pour la non-réponse correspond à l'inverse de la propension à répondre estimée pour l'unité échantillonnée $i\left(\hat{\phi}_i\right)$. La propension à répondre peut être estimée de différentes manières, par exemple par régression logistique, mais pour la plupart des enquêtes, on établit des groupes mutuellement exclusifs appelés classes de pondération ou groupes de réponse homogènes qui renferment des unités ayant des propensions estimées similaires et on ajuste les poids dans chaque groupe ou classe en fonction d'un facteur commun, par exemple $\hat{f}_c = \hat{\phi}_c^{-1}$ pour toutes les valeurs de $i \in c$ (Särndal et coll. 1992; Little 1986). En vertu de cette approche, l'estimateur ajusté est appelé un estimateur de classe de pondération et s'écrit comme suit :

$$\hat{y}_{cp} = \sum_c \sum_{i \in s_c} R_{ci} d_{ci} \hat{f}_c y_{ci}, \tag{2.3}$$

où $c = 1, 2, \ldots, C$ correspond aux classes d'ajustement pour la non-réponse et $i \in s_c$ est une unité échantillonnée de la classe c.

La question spécifique à laquelle nous nous intéressons ici est l'effet de la pondération du facteur d'ajustement. Le facteur non pondéré s'écrit

$$\hat{f}_c^{np} = \frac{\sum_{i \in s_c} \delta_{ci}}{\sum_{i \in s_c} R_{ci} \delta_{ci}} = \frac{n_{c+}}{r_{c+}}$$

où $\delta_{ci} = 1$ si $i \in c$ et $\delta_{ci} = 0$ si $i \notin c$, et n_{c+} et r_{c+} correspondent au nombre d'unités échantillonnées et répondantes de la classe c. Le facteur d'ajustement pondéré s'écrit

$$\hat{f}_c^{p} = \frac{\sum_{i \in s_c} d_{ci}}{\sum_{i \in s_c} R_{ci} d_{ci}} = \frac{\hat{N}_c}{\hat{N}'_c},$$

où $\hat{N}_c = \sum_{i \in s_c} d_{ci}$ et $\hat{N}'_c = \sum_{i \in s_c} R_{ci} d_{ci}$. Les facteurs correspondent aux taux de réponse non pondéré et pondéré, respectivement. En substituant les facteurs dans l'estimateur (2.3), on obtient deux nouveaux estimateurs (2.4) et (2.5) de la population totale. Il s'agit de deux estimateurs de classes de pondération, dont la notation a été modifiée pour mettre en évidence le taux de réponse utilisé (pondéré ou non pondéré).

$$\hat{y}_{trnp} = \sum_c \hat{f}_c^{np} \sum_{i \in r_c} d_{ci} y_{ci} = \sum_c \frac{n_{c+}}{r_{c+}} \sum_{i \in r_c} d_{ci} y_{ci}, \tag{2.4}$$

$$\hat{y}_{trp} = \sum_c \hat{f}_c^{p} \sum_{i \in r_c} d_{ci} y_{ci} = \sum_c \frac{\hat{N}_c}{\hat{N}'_c} \sum_{i \in r_c} d_{ci} y_{ci}. \tag{2.5}$$

Ces deux estimateurs constituent les éléments de base pour tous les types de statistiques que nous examinons dans l'étude par simulation. Par exemple, les estimateurs des moyennes, des moyennes de domaine et des ratios sont de simples fonctions des estimateurs (2.4) et (2.5).

Pour respecter la structure, la notation et les simulations de L et V, la présente étude est restreinte à la même population et à un échantillon aléatoire simple stratifié où deux strates sont définies par la variable de plan binaire Z, et où deux classes d'ajustement pour la non-réponse sont définies par une variable

auxiliaire binaire C, qui recoupe les strates comme indiqué dans le tableau 2.1. Nous avons remplacé la lettre X utilisée par L et V par la lettre C dans la cellule de pondération introduite ci-dessus afin de faciliter l'identification de la cellule d'ajustement pour la non-réponse. Comme dans l'étude de L et V, la taille de la population est fixée à $N = 10\,000$.

Tableau 2.1
Chiffres de population par strate *Z* et par cellule d'ajustement pour la non-réponse *C*

Strate d'échantillonnage	Cellule d'ajustement pour la non-réponse	
	***C* = 0**	***C* = 1**
Z = 0	3 064	3 931
Z = 1	2 079	926

Source : Little et Vartivarian (2003), qui ont utilisé *X* au lieu de *C*.

La variable d'intérêt, Y, est une variable binaire pour laquelle la probabilité que $Y = 1$ est définie par un modèle logistique où $\text{logit}(Y = 1|C, Z) = 0{,}5 + \gamma_C (C - \bar{C}) + \gamma_Z (Z - \bar{Z}) + \gamma_{CZ} (C - \bar{C})(Z - \bar{Z})$. La variable de réponse R est aussi binaire, et la probabilité que $R = 1$ est générée à partir d'un modèle logistique où $\text{logit}(R|C, Z) = 0{,}5 + \beta_C (C - \bar{C}) + \beta_Z (Z - \bar{Z}) + \beta_{CZ} (C - \bar{C})(Z - \bar{Z})$. Différentes populations et propensions à répondre sont générées en fonction des valeurs de $\gamma_C, \gamma_Z, \gamma_{CZ}, \beta_C, \beta_Z$ et β_{CZ} indiquées dans le tableau 2.2. Nous avons adopté la notation de L et V pour les modèles linéaires généralisés afin de faciliter la comparaison avec leurs travaux. Les valeurs indiquées dans le tableau sont les mêmes variables de population et de réponse que celles que L et V ont produites en affectant des valeurs à $(\gamma_C, \gamma_Z, \gamma_{CZ}, \beta_C, \beta_Z, \beta_{CZ})$. Dans la notation $[A]^B$ présentée au tableau 2.2, la population (Y) ou la propension à répondre (R) sont indiquées par l'exposant B, alors que les paramètres et les interactions du modèle pour la répartition de la population ou de la réponse sont indiqués par la lettre A entre crochets. Par exemple, le modèle logistique additif qui génère la répartition de Y dans la strate d'échantillonnage Z et la cellule de non-réponse C est indiqué comme suit : $[C + Z]^Y$. De même, les modèles où R dépend de C seulement, de Z seulement ou ni de C ni de Z sont indiqués respectivement par $[C]^R$, $[Z]^R$ et $[C + Z]^R$. L et V donnent plus de détails sur les motifs justifiant le choix de ces populations et modèles de réponse en particulier.

Tableau 2.2
Modèles pour la variable de résultat *Y* et la probabilité de réponse *R*

Modèle pour *Y* (variable d'intérêt)	Modèle pour *R* (propension à répondre)	Paramètres		
		γ_C, β_C	γ_Z, β_Z	γ_{CZ}, β_{CZ}
$[CZ]^Y$	$[CZ]^R$	2	2	2
$[C + Z]^Y$	$[C + Z]^R$	2	2	0
$[C]^Y$	$[C]^R$	2	0	0
$[Z]^Y$	$[Z]^R$	0	2	0
$[\phi]^Y$	$[\phi]^R$	0	0	0

Source : Little et Vartivarian (2003).

L et V ont calculé des estimations des moyennes comme suit, selon notre notation :

$$\hat{\bar{y}}_{trnp} = \frac{\hat{y}_{trnp}}{\sum_c \hat{f}_c^{np} \sum_{i \in s_c} R_{ci} d_{ci}} = \frac{\hat{y}_{trnp}}{\sum_c \hat{f}_c^{np} \hat{N}'_c}, \tag{2.6}$$

et

$$\hat{\bar{y}}_{trp} = \frac{\hat{y}_{trp}}{\sum_c \hat{f}_c^{p} \sum_{i \in s_c} R_{ci} d_{ci}} = \frac{\hat{y}_{trp}}{\sum_c \hat{N}_c}. \tag{2.7}$$

Les dénominateurs des moyennes sont des estimations de la taille de population N. Dans l'estimateur (2.7), le dénominateur est une constante égale à N, mais dans l'estimateur (2.6), le dénominateur est une variable aléatoire. Dans le scénario de simulation comprenant le plan d'échantillonnage aléatoire simple stratifié décrit ci-dessous, ou tout plan où $\sum_{i \in s} d_i = N$ pour chaque valeur de s, l'estimateur (2.7) se réduit à l'estimateur linéaire $\hat{\bar{y}}_{trp} = N^{-1} \hat{y}_{trp}$, tandis que l'estimateur (2.6) est un estimateur par le ratio. Il s'agit là d'un point important sur lequel nous reviendrons.

Les moyennes de domaine peuvent avoir des propriétés différentes des moyennes globales parce que les dénominateurs des moyennes de domaine pondérées et non pondérées sont des variables aléatoires, sauf quand les domaines concordent avec la strate d'échantillonnage et que les tailles des domaines et les tailles des strates sont connues. L et V n'abordent pas la question des domaines et n'ont donc pas examiné ces estimations dans le cadre de leur simulation. Nous avons établi des domaines en générant au hasard une variable aléatoire ν_i à partir d'une distribution uniforme (0, 1) et en définissant la fonction d'appartenance $\tau(a) = 1$ si $a < 0$ et $\tau(a) = 0$ si $a \geq 0$. Des moyennes de domaine de 50 % ont été créées en substituant $d_{ci}^* = \tau(\nu_i - 0{,}5)\, d_{ci}$ dans les expressions (2.6) et (2.7) afin de produire les estimateurs $\hat{\bar{y}}_{trnp;0,5}$ et $\hat{\bar{y}}_{trp;0,5}$, respectivement. Les estimateurs pondérés et non pondérés des totaux de domaine $\hat{y}_{trnp;0,5}$ et $\hat{y}_{trp;0,5}$ ont été établis de la même manière. Nous avons utilisé la même méthode pour créer des moyennes de domaine de 25 % et des totaux de domaine de 25 %. Comme nous nous intéressons à l'effet des ajustements pour la non-réponse sur les moyennes calculées sous forme d'estimateurs par ratio, d'autres domaines comme ceux qui correspondent à près de 100 % de la population ont été exclus de l'analyse parce que le dénominateur des moyennes de domaine est proche de la population totale constante N et qu'alors la moyenne devient un estimateur linéaire. Les domaines plus proches de 0 % ont été exclus à cause de la petite taille des échantillons.

3 Résultats

La simulation a été effectuée dans le logiciel R (R Development Core Team 2011) à partir de 10 000 tirages (L et V en ont utilisé 1 000). Nous avons évalué les estimateurs en calculant la racine de l'erreur quadratique moyenne (reqm) et le biais des estimations, le biais et la reqm étant mesurés par les écarts par rapport aux quantités de population comme l'ont fait L et V. Nous avons utilisé la même taille d'échantillon total (312) que dans la simulation, mais avec différentes répartitions de l'échantillon ou

différents taux d'échantillonnage relatifs entre les strates. Nous avons reproduit l'ensemble des 25 configurations de L et V; les résultats sont présentés dans le tableau S-1 des documents supplémentaires. Le tableau S-2 des documents supplémentaires comprend aussi les 25 configurations, mais présente le biais relatif des moyennes et des totaux avec et sans pondération, ainsi que les ratios des variances et des reqm des estimations non pondérées à ceux des estimations pondérées. Le biais relatif et les ratios des variances et des reqm facilitent les comparaisons entre les estimations. Les documents supplémentaires comprennent les erreurs de simulation estimées, qui sont toutes relativement petites. Pour les estimateurs et les taux d'échantillonnage donnés par L et V, nos résultats correspondent aux valeurs publiées, compte tenu des erreurs de simulation. Commençons par examiner le biais des estimateurs.

3.1 Biais

Il y a deux situations pour lesquelles il existe des résultats théoriques bien connus (Little et Rubin 2002). La première est lorsque la propension à répondre est la même dans toutes les cellules – les données manquent complètement au hasard (MCAR, de l'anglais *missing completely at random*); ces données de type MCAR correspondent au modèle $[\phi]^R = (\beta_c = 0, \beta_z = 0, \beta_{cz} = 0)$ de la dernière ligne du tableau 2.2. Lorsqu'on a des données de type MCAR, les facteurs d'ajustement non pondéré et pondéré ont la même espérance mathématique, et tous deux produisent des estimations non biaisées. Les résultats de la simulation présentés dans le tableau V de l'article de L et V (lignes 5, 10, 15, 20 et 25) confirment cette observation. La deuxième situation est lorsque la propension à répondre est indépendante de la strate, ce qui correspond à des données qui manquent au hasard (MAR, de l'anglais *missing at random*) selon le modèle de réponse $[\phi]^C = (\beta_c = 2, \beta_z = 0, \beta_{cz} = 0)$ de la troisième ligne du tableau 2.2. Nous considérons ces situations comme étant de type MAR parce que le biais de l'estimateur ne dépend pas de l'utilisation de données à propos de Z dans le modèle. Encore une fois, les estimations avec et sans pondération sont toutes deux sans biais, et les ajustements ont la même espérance mathématique. Les résultats de la simulation présentés dans le tableau V de L et V (lignes 3, 8, 13, 18 et 23) confirment cette observation de façon empirique.

Afin de nous concentrer sur la situation dans laquelle les spécifications du modèle sont erronées, nous ne présentons pas les résultats des simulations pour les situations de type MCAR et MAR dans le présent article; ces résultats sont toutefois présentés dans les documents supplémentaires. Il importe de souligner que même si les ajustements avec et sans pondération pour les modèles de type MCAR et MAR ont la même espérance mathématique, ils ne sont pas identiques. Après avoir simulé les deux approches en vertu de modèles de type MAR, Sukasih et coll. (2009) se sont prononcés en faveur d'une approche de pondération, principalement en raison de la variabilité moindre des estimations des totaux pour l'ensemble des simulations, même si les deux approches donnent des résultats non biaisés.

Comme il est précisé plus haut, les taux d'échantillonnage varient dans le cadre de nos simulations, tandis que la taille globale de l'échantillon est fixée à 312; L et V ont utilisé un taux d'échantillonnage unique. Quand les taux d'échantillonnage sont les mêmes dans toutes les strates (c'est-à-dire que l'échantillon est réparti proportionnellement dans toutes les strates), les poids d'échantillonnage sont les mêmes pour chaque strate et, en conséquence, les estimateurs avec et sans pondération sont identiques. Le taux d'échantillonnage selon une répartition proportionnelle joue un rôle important dans notre présentation, parce que les deux estimations doivent converger à cette étape.

Le graphique présenté à la figure 3.1 (à gauche) illustre les résultats de la simulation pour le biais des estimateurs avec et sans pondération du total pour $[CZ]^Y$ et $[C+Z]^R$. Nous avons choisi cette configuration (ligne 2 dans les tableaux de L et V) parce que les simulations de L et V montrent que la moyenne non pondérée est assortie d'un biais et d'une reqm plus faibles que la moyenne pondérée dans ce cas particulier. L'axe horizontal indique le taux d'échantillonnage relatif calculé comme étant le ratio du taux d'échantillonnage de $Z = 0$ à $Z = 1$ ou $N_0 n_0^{-1} / \left(N_1 n_1^{-1}\right)$. Le taux d'échantillonnage relatif employé par L et V était d'environ 2,25. On voit tout de suite que le biais de l'estimateur pondéré est pratiquement constant pour les différents taux d'échantillonnage, alors que le biais de l'estimateur non pondéré varie considérablement selon le taux d'échantillonnage relatif. Pour certains taux d'échantillonnage, le biais des estimateurs non pondérés du total peut être plus de deux fois celui de l'estimateur pondéré. Les deux types d'estimateur sont biaisés pour presque tous les taux d'échantillonnage relatifs, et l'estimateur qui a le biais le plus faible dépend du taux d'échantillonnage relatif. Lorsque les taux d'échantillonnage relatifs sont égaux (répartition proportionnelle), les estimateurs sans pondération et avec pondération ont le même biais, comme prévu. Cependant, dans la pratique, il n'est généralement pas possible de reconnaître l'effet du taux d'échantillonnage sur le biais et de choisir à l'avance la méthode d'ajustement qui permet de réduire le biais pour un échantillon particulier.

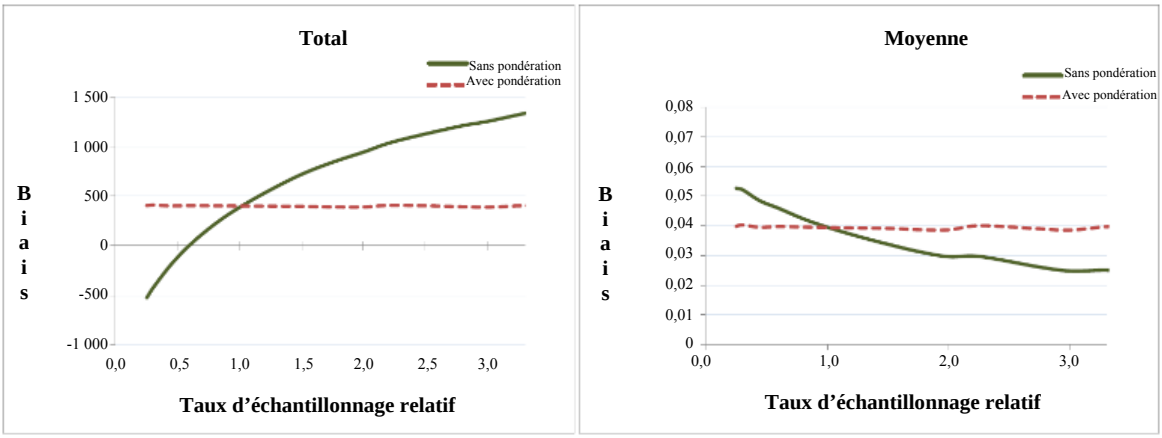

Figure 3.1 Biais des estimateurs avec et sans pondération pour le modèle de population $[CZ]^Y$ et le modèle de propension à répondre $[C+Z]^R$; le graphique de gauche correspond au total, et celui de droite, à la moyenne.

Pour comprendre ces résultats, nous avons appliqué des approximations standard qui se vérifient raisonnablement bien dans ce cas (c'est-à-dire $E\left(\eta^{-1}\right) \approx E^{-1}(\eta)\big)$. La valeur prévue approximative pour l'estimateur pondéré est

$$E\hat{y}_{trp} \approx \sum_z \sum_c \frac{N_c}{\left(\sum_z \phi_{cz} N_{cz}\right)} \phi_{cz} Y_{cz}, \tag{3.1}$$

où Y_{cz} est le total de population de la cellule cz. De même, la valeur prévue approximative pour l'estimateur non pondéré est

$$E\hat{y}_{trnp} \approx \sum_z \sum_c \frac{\left(\sum_z N_z n_z^{-1} N_{cz}\right)}{\left(\sum_z \phi_{cz} N_z n_z^{-1} N_{cz}\right)} \phi_{cz} Y_{cz}. \quad (3.2)$$

Si ϕ_{cz} est une constante (MCAR) ou ϕ_{cz} est une constante dans les cellules de pondération (MAR), alors les deux estimateurs ne sont pas biaisés à cet ordre d'approximation et concordent avec la théorie connue. Lorsque les taux d'échantillonnage sont les mêmes dans toutes les strates, les deux estimateurs ont la même valeur prévue (comme il est précisé plus haut, ils sont identiques dans ce cas). Surtout, ces approximations montrent que l'espérance mathématique de l'estimateur pondéré ne dépend pas du taux d'échantillonnage, mais que celle de l'estimateur non pondéré, elle, en dépend. Cela explique les courbes illustrées à la figure 3.1.

Quelques détails des estimations de la simulation pour cette configuration sont présentés dans le tableau 3.1 pour certains taux d'échantillonnage. Comme il est indiqué ci-dessus, les résultats complets de la simulation pour toutes les configurations et tous les taux d'échantillonnage utilisés pour dessiner les graphiques se trouvent dans les documents supplémentaires. Ces documents comprennent les biais relatifs, les ratios des variances et les ratios des reqm, qui constituent de meilleurs indicateurs pour évaluer l'incidence des ajustements sur les estimations. Nous avons constaté que pour toutes les configurations dont les estimations des totaux sont biaisées, les biais pour l'estimateur pondéré sont inférieurs d'un côté du taux d'échantillonnage relatif de 1, et supérieurs de l'autre côté. Toutes les configurations sont assorties d'un biais à peu près constant pour l'estimateur pondéré du total pour tous les taux d'échantillonnage relatifs, mais le biais de l'estimateur non pondéré varie en fonction du taux d'échantillonnage relatif.

Examinons maintenant les moyennes estimées – les seuls estimateurs examinés par L et V. Le graphique de droite de la figure 3.1 montre que le biais pour l'estimateur pondéré est encore une fois indépendant du taux d'échantillonnage relatif, alors que le biais de l'estimateur non pondéré varie en fonction du taux d'échantillonnage. L et V ont utilisé un taux d'échantillonnage de 2,25, ce qui explique pourquoi ils ont trouvé que l'estimateur non pondéré était associé à un biais inférieur pour la moyenne dans le cadre de leur exercice de simulation. Il importe de souligner deux choses à cet égard. D'une part, les biais pour les moyennes pour les deux méthodes d'ajustement sont tous relativement faibles, particulièrement par rapport aux biais relatifs potentiels des totaux obtenus à l'aide de l'estimateur non pondéré (graphique de gauche). D'autre part, il n'y a aucun moyen de déterminer si une estimation particulière tomberait du côté gauche ou du côté droit du taux d'échantillonnage relatif de 1. Le tableau 3.1 montre les biais estimés pour cette configuration.

Les graphiques illustrent aussi une relation quelque peu étonnante : les taux d'échantillonnage relatifs pour lesquels l'estimateur non pondéré du total est assorti d'un biais inférieur sont ceux pour lesquels l'estimateur non pondéré de la moyenne est assorti d'un biais supérieur. En d'autres termes, les moyennes se comportent différemment des totaux parce que la moyenne non pondérée est un ratio alors que la moyenne pondérée n'en est pas un. En conséquence, le biais relatif (br = biais/estimation) de l'estimateur non pondéré de la moyenne n'est pas égal au biais relatif de l'estimateur non pondéré du total (la relation est vérifiée pour l'estimateur pondéré). On peut approximer le biais relatif comme suit :

$$br\left(\hat{\bar{y}}_{trnp}\right) \approx \frac{1 + br\left(\hat{y}_{trnp}\right)}{1 + br\left(\hat{N}_{trnp}\right)},$$

où $\hat{N}_{trnp}$ est l'estimateur non pondéré du total (où $y_i = 1$ pour toutes les valeurs de i). Cette approximation se vérifie raisonnablement bien dans cette situation, puisque $\text{cov}\left(\hat{\bar{y}}_{trnp}, \hat{N}_{trnp}\right) / E\left(\hat{N}_{trnp}\right) \approx 0$. Le biais relatif de la moyenne non pondérée diminue donc quand les biais du numérateur et du dénominateur sont positivement corrélés.

Examinons maintenant les estimations de domaine – que L et V n'ont pas étudiées. Les biais pour les estimateurs du total de domaine avec et sans pondération et la relation avec les biais des estimateurs non pondérés qui varient en fonction du taux d'échantillonnage relatif sont les mêmes que ceux qui ont été observés pour les totaux globaux (voir le tableau 3.1), parce que les totaux de domaine demeurent des totaux et que les approximations (3.1) et (3.2) continuent de s'appliquer. Les moyennes de domaine sont aussi présentées dans le tableau, et elles aussi suivent la tendance des biais illustrée à la figure 3.1 pour la moyenne de l'échantillon complet. Il importe de souligner que les biais relatifs pour les estimations de la moyenne (globale et pour chaque domaine) ne varient pas beaucoup, la plupart d'entre eux se trouvant entre 5 % et 7 %.

Tableau 3.1
Biais (facteur 10 000), racine de l'erreur quadratique moyenne (facteur 10 000) et variance des estimateurs avec et sans pondération des moyennes et du total de l'échantillon complet et des domaines, configuration $[CZ]^Y$, $[C+Z]^R$ selon divers taux d'échantillonnage

				Taux d'échantillonnage relatif				
	Caractéristique	**Domaine**	**Ajustement**	**0,30**	**0,44**	**1,00**	**2,25**	**3,30**
Biais	Moyenne	Complet	trnp	515	491	404	301	248
			trp	398	403	404	404	394
		50 %	trnp	513	501	411	307	257
			trp	397	414	410	410	401
		25 %	trnp	523	498	407	298	252
			trp	408	411	407	400	395
	Total	Complet	trnp	-419	-184	401	1 058	1 335
			trp	398	403	404	404	394
		50 %	trnp	-214	-89	205	535	673
			trp	194	205	206	207	200
		25 %	trnp	-107	-48	101	264	335
			trp	97	98	102	101	100
reqm	Moyenne	Complet	trnp	643	614	546	536	566
			trp	553	547	545	587	616
		50 %	trnp	758	726	669	699	778
			trp	687	671	669	728	794
		25 %	trnp	949	898	863	952	1 062
			trp	895	859	863	955	1 041
	Total	Complet	trnp	537	376	543	1 183	1 485
			trp	553	547	545	587	616
		50 %	trnp	371	311	393	714	888
			trp	399	392	394	449	494
		25 %	trnp	255	233	282	451	553
			trp	285	273	283	328	365
Variance	Moyenne	Complet	trnp	15	14	14	20	26
			trp	15	14	14	18	22
		50 %	trnp	32	28	28	40	54
			trp	32	28	28	37	47
		25 %	trnp	64	57	59	83	107
			trp	64	58	59	76	93
	Total	Complet	trnp	11	11	14	28	43
			trp	15	14	14	18	22
		50 %	trnp	9	9	11	23	34
			trp	12	11	11	16	21
		25 %	trnp	5	5	7	14	20
			trp	7	7	7	10	12

3.2 Racine de l'erreur quadratique moyenne (reqm)

Malgré la petite taille de l'échantillon utilisé pour les simulations (312 avant la non-réponse) et le biais relatif plutôt modeste des estimations pour les moyennes, le biais demeure une composante importante de la reqm. Par exemple, le biais représente 56 % (sans pondération) à 69 % (avec pondération) de la reqm pour l'estimation de la moyenne selon la configuration $[CZ]^Y$ et $[C+Z]^R$ et le même taux d'échantillonnage que L et V. Lorsque l'échantillon est plus important, comme c'est généralement le cas pour les grandes enquêtes par sondage, le biais est souvent la composante dominante de la reqm (Brick 2013).

La figure 3.2 montre la reqm pour le total estimé (graphique de gauche) et pour la moyenne (graphique de droite) selon la même configuration que pour la figure précédente. La reqm pour le total pour l'estimateur pondéré est approximativement constante et inférieure à la reqm pour l'estimateur non pondéré, sauf lorsque le taux d'échantillonnage relatif est d'environ 0,5, ce qui correspond à la région où le biais est très faible pour l'estimateur non pondéré (voir la figure 3.1). Toutefois, lorsque le taux d'échantillonnage relatif est supérieur à un, la reqm pour l'estimateur non pondéré du total est beaucoup plus grande que la reqm pour l'estimateur pondéré (jusqu'à deux fois plus élevée pour certains taux d'échantillonnage). En revanche, pour les estimations de la moyenne illustrées à la figure 3.2 (graphique de droite), les reqm des estimateurs avec et sans pondération sont du même ordre de grandeur, et la symétrie autour du taux de répartition proportionnelle demeure. Même si L et V soulignent que l'estimateur non pondéré a une reqm inférieure (au taux d'échantillonnage relatif de 2,25), nous considérons les reqm des deux estimateurs comme étant approximativement égales pour tous les taux d'échantillonnage relatifs.

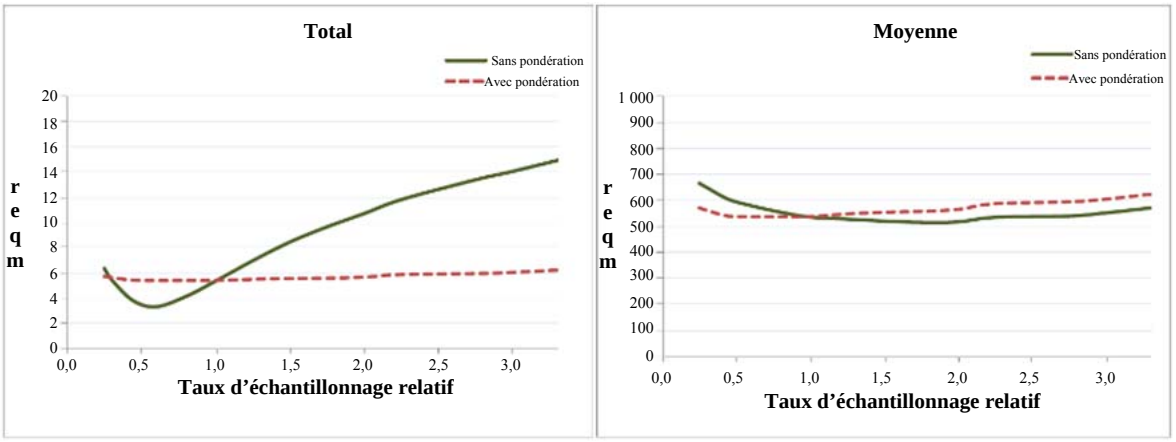

Figure 3.2 Racine de l'erreur quadratique moyenne pour les estimateurs avec et sans pondération pour $[CZ]^Y$ et $[C+Z]^R$; le graphique de gauche correspond au total (reqm en millions) et celui de droite, à la moyenne.

La figure 3.3 indique la reqm pour la moyenne estimée pour un domaine de 50 % (graphique de gauche) et un domaine de 25 % (graphique de droite), encore une fois pour $[CZ]^Y$ et $[C+Z]^R$. L'examen des trois graphiques de la reqm (pour la moyenne globale, la moyenne pour un domaine de 50 % et la moyenne pour un domaine de 25 %) révèle l'effet de l'estimateur par ratio. À mesure que la taille du domaine passe de

100 % à 25 %, l'estimateur pondéré ressemble de plus en plus à un estimateur par ratio inconditionnel et la corrélation entre le numérateur et le dénominateur réduit la reqm de l'estimation. En conséquence, les reqm des estimateurs de domaine avec et sans pondération sont très semblables. Même si l'estimateur pondéré est assorti d'une reqm inférieure à chacun des taux d'échantillonnage relatifs comparativement à l'estimateur non pondéré pour la moyenne pour un domaine de 25 %, les deux estimateurs sont essentiellement équivalents en termes de reqm. Le léger avantage de l'estimateur non pondéré qu'ont souligné L et V pour la moyenne pour l'ensemble de la population selon cette configuration disparaît pour les moyennes de domaine où l'estimateur pondéré est aussi un estimateur par ratio.

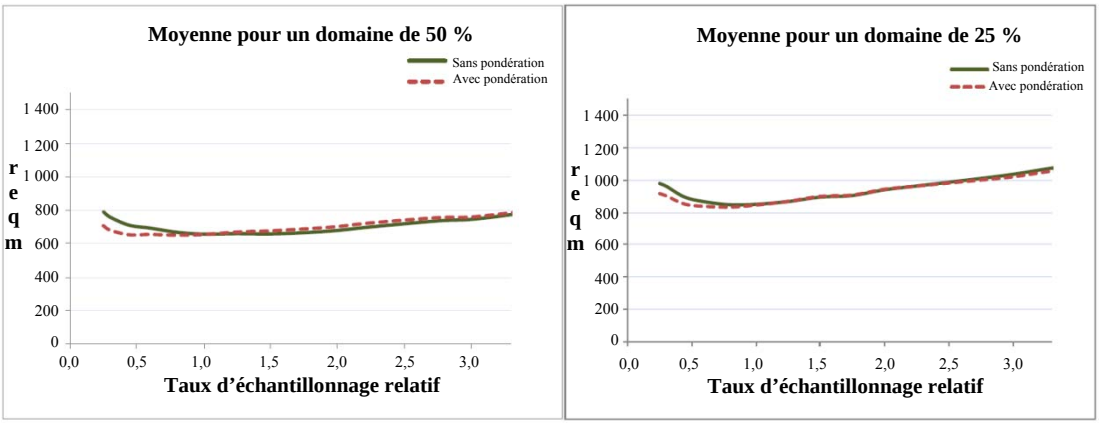

Figure 3.3 Racine de l'erreur quadratique moyenne pour les estimateurs avec et sans pondération pour $[CZ]^Y$ et $[C+Z]^R$; le graphique de gauche correspond à la moyenne pour un domaine de 50 % et celui de droite, à la moyenne pour un domaine de 25 %.

3.3 Variance

Quand les facteurs d'ajustement pour la non-réponse sont fondés sur un petit nombre de répondants, il est possible qu'ils accroissent la variance des estimations (Kalton 1983; Tremblay 1986). L et V sont d'avis que la pondération des facteurs d'ajustement pour la non-réponse pourrait entraîner une inflation de la variance supérieure à celle que l'on obtient lorsqu'on utilise des facteurs non pondérés. Les figures ci-dessus montrent que cela ne s'est pas produit dans le cadre de notre exercice de simulation. La figure 3.4 illustre le ratio de la variance de l'estimateur non pondéré à la variance de l'estimation pondérée pour la moyenne et le total pour l'ensemble de la population et pour le total du domaine de 50 % selon la configuration $[CZ]^Y$ et $[C+Z]^R$. Pour la moyenne, le ratio des variances est presque égal à un pour tous les taux d'échantillonnage relatifs; il n'y a pas d'inflation de la variance pour l'estimateur pondéré comparativement à l'estimateur non pondéré. En ce qui concerne les totaux, le ratio est inférieur à un pour les taux d'échantillonnage relatifs de moins de 1, et supérieur à un pour les taux d'échantillonnage relatifs de plus de 1. Cette relation se vérifie aussi pour le total du domaine de 50 %. Ces résultats semblent indiquer que la pondération de l'ajustement n'est pas une source de facteurs importants susceptibles de faire augmenter la variance des estimations. Par mesure de prudence, il convient d'examiner l'importance des facteurs de non-réponse, qu'ils soient ou non pondérés.

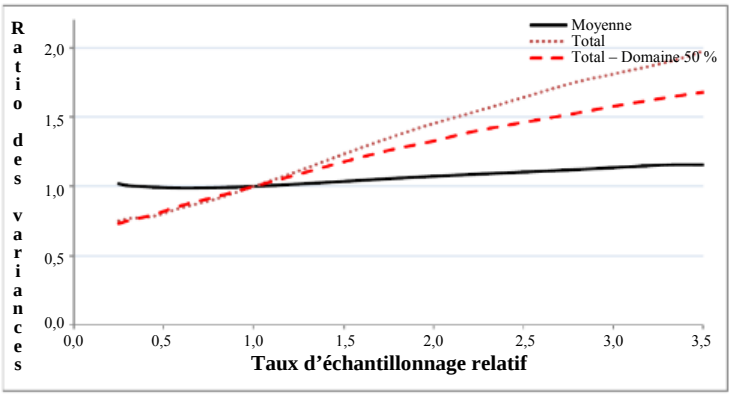

Figure 3.4 Ratio des variances des estimations non pondérées aux estimations pondérées de la moyenne, du total global et du total pour un domaine de 50 % selon $[CZ]^Y$ et $[C+Z]^R$.

Le tableau 3.2 présente les résultats de simulation pour une autre configuration, $[CZ]^Y$ et $[CZ]^R$, qui était favorable à l'ajustement non pondéré dans le cadre de l'étude de L et V (première ligne de leurs tableaux), alors que le tableau 3.3 présente les résultats de simulation pour la configuration $[C + Z]^Y$ et $[C + Z]^R$, qui était favorable à l'ajustement pondéré. Les résultats pour ces deux configurations montrent les mêmes tendances générales présentées ci-dessus pour $[CZ]^Y$ et $[C + Z]^R$.

Tableau 3.2
Biais (facteur 10 000), racine de l'erreur quadratique moyenne (facteur 10 000) et variance des estimateurs pondérés et non pondérés des moyennes et du total pour l'échantillon complet et pour les domaines, configuration $[CZ]^Y$, $[CZ]^R$ selon divers taux d'échantillonnage

				Taux d'échantillonnage relatif				
	Caractéristique	Domaine	Ajustement	0,30	0,44	1,00	2,25	3,30
Biais	Moyenne	Complet	trnp	329	329	289	255	237
			trp	294	299	289	298	298
		50 %	trnp	334	341	293	251	238
			trp	299	311	293	294	298
		25 %	trnp	336	344	306	257	247
			trp	302	314	306	299	307
	Total	Complet	trnp	-412	-187	287	732	901
			trp	294	299	289	298	298
		50 %	trnp	-209	-91	145	367	455
			trp	143	152	146	149	154
		25 %	trnp	-103	-46	72	184	230
			trp	74	76	73	75	79
reqm	Moyenne	Complet	trnp	530	507	476	501	533
			trp	505	487	476	520	554
		50 %	trnp	684	653	616	664	732
			trp	666	638	616	674	740
		25 %	trnp	911	859	832	920	1 016
			trp	900	849	832	920	1 011
	Total	Complet	trnp	550	395	474	886	1 078
			trp	505	487	476	520	554
		50 %	trnp	385	326	373	575	696
			trp	394	375	373	425	475
		25 %	trnp	263	244	278	390	464
			trp	285	274	278	321	361
Variance	Moyenne	Complet	trnp	17	15	14	19	23
			trp	17	15	14	18	22
		50 %	trnp	36	31	30	38	48
			trp	36	31	30	37	46
		25 %	trnp	73	63	61	79	98
			trp	73	63	61	76	94
	Total	Complet	trnp	14	12	14	25	35
			trp	17	15	14	18	22
		50 %	trnp	11	10	12	20	28
			trp	14	12	12	16	20
		25 %	trnp	6	6	7	12	16
			trp	8	7	7	10	13

Tableau 3.3
Biais (facteur 10 000), racine de l'erreur quadratique moyenne (facteur 10 000) et variance des estimateurs pondérés et non pondérés des moyennes et du total pour l'échantillon complet et pour les domaines, configuration $[C+Z]^Y$, $[C+Z]^R$ selon divers taux d'échantillonnage

				Taux d'échantillonnage relatif				
	Caractéristique	**Domaine**	**Ajustement**	**0,30**	**0,44**	**1,00**	**2,25**	**3,30**
Biais	Moyenne	Complet	trnp	763	735	654	566	529
			trp	665	661	654	654	652
		50 %	trnp	773	737	653	564	532
			trp	677	664	653	651	656
		25 %	trnp	773	739	659	574	513
			trp	679	668	659	660	636
	Total	Complet	trnp	-272	-8	651	1 411	1 744
			trp	665	661	654	654	652
		50 %	trnp	-133	-6	326	711	875
			trp	336	328	328	332	328
		25 %	trnp	-69	-2	157	359	438
			trp	165	166	158	168	165
reqm	Moyenne	Complet	trnp	854	818	745	699	711
			trp	767	753	745	764	790
		50 %	trnp	951	901	827	816	863
			trp	877	845	826	863	912
		25 %	trnp	1 101	1 046	981	1 023	1 098
			trp	1 044	1 004	981	1 045	1 107
	Total	Complet	trnp	426	313	741	1 503	1 868
			trp	767	753	745	764	790
		50 %	trnp	334	300	475	867	1 071
			trp	489	470	476	529	575
		25 %	trnp	246	240	314	530	649
			trp	320	316	314	372	409
Variance	Moyenne	Complet	trnp	15	13	13	17	23
			trp	15	13	13	16	20
		50 %	trnp	31	27	26	35	46
			trp	31	28	26	32	40
		25 %	trnp	62	56	54	73	95
			trp	63	57	54	67	83
	Total	Complet	trnp	11	10	13	27	45
			trp	15	13	13	16	20
		50 %	trnp	10	9	12	25	39
			trp	13	12	12	17	22
		25 %	trnp	6	6	7	15	23
			trp	8	7	8	11	14

3.4 Estimation de la taille de population

Sukasih et coll. (2009) ont étudié un type particulier d'estimation, soit l'estimation du nombre d'unités d'une population. On parle alors d'une estimation de la taille de population où la taille de population n'est qu'une estimation d'un total où $y_i = 1$ pour toutes les valeurs de i. Elle peut être estimée pour un domaine en affectant à toutes les unités en dehors du domaine la valeur $y_i = 0$. Dans le plan d'échantillonnage simple stratifié étudié ici, l'estimateur pondéré reproduit toujours la taille de population totale, $N = 10\,000$, mais pas l'estimateur non pondéré. Comme cette situation favorise clairement l'estimateur pondéré, nous examinons plutôt l'estimation de la taille de population d'un domaine.

Supposons que nous voulions estimer le nombre d'unités d'un domaine ou d'un sous-groupe qui ont une valeur en dessous d'un centile défini par une caractéristique pour la population totale (par exemple le revenu médian national). Ce type de statistique est extrêmement important dans les enquêtes, parce que les estimations de la taille de population pour les domaines sont souvent des statistiques clés. Ce type d'estimation peut être, par exemple, le nombre total de personnes ayant un revenu sous le seuil de pauvreté ou de faible revenu (Kovačević et Yung 1997).

Comme l'analyse de L et V ne tenait pas compte des estimations pour les tailles ou les moyennes de domaine, il n'existe pas de variable explicite qui pourrait servir à définir une sous-population. Pour ne pas

compliquer l'analyse, nous illustrons le rendement des deux estimateurs à l'aide d'un domaine artificiel créé par la sélection aléatoire de la moitié de la population (c'est-à-dire un domaine de 50 %). Selon une analyse semblable à celle dont il est question dans les sections précédentes, nous avons calculé les totaux et les moyennes pondérés et non pondérés pour le domaine de 50 %. Même si nous connaissons déjà la taille du domaine de l'exemple (c'est-à-dire 50 % de la population totale), l'analyse demeure valide. Dans la pratique, la taille du domaine n'est pas connue.

Quand on estime une statistique comme la taille de population d'un domaine, les deux estimateurs, pondéré et non pondéré, de la taille de population du domaine ne sont pas biaisés lorsque les données sont de type MCAR ou MAR, comme le soulignent Sukasih et coll. (2009). En outre, les reqm des estimateurs avec et sans pondération sont approximativement égales dans ce cas, comme le confirment les simulations.

Si les données ne sont pas de type MAR, la situation peut être très différente. L'estimateur pondéré d'une taille de population de domaine est à peu près non biaisé pour tous les taux d'échantillonnage relatifs et toutes les configurations, alors que l'estimateur non pondéré est toujours biaisé, sauf lorsqu'il est identique à l'estimateur pondéré (à un taux d'échantillonnage relatif de 1). En conséquence, la reqm de l'estimateur non pondéré pour la taille de domaine est souvent considérablement plus élevée que celle de l'estimateur pondéré. La figure 3.5 montre que la reqm de l'estimateur non pondéré de la taille de domaine de 50 % pour $[CZ]^Y$ et $[C + Z]^R$ est beaucoup plus grande que celle de l'estimateur pondéré pour la plupart des taux d'échantillonnage relatifs (jusqu'à deux fois la reqm de l'estimateur pondéré). La seule exception, c'est lorsque deux estimateurs sont à peu près égaux (répartition presque proportionnelle).

L'estimateur pondéré des tailles de domaine présente donc un avantage considérable par rapport à l'estimateur non pondéré pour tous les mécanismes de données manquantes présentés par L et V qui ne sont pas de type MCAR ou MAR.

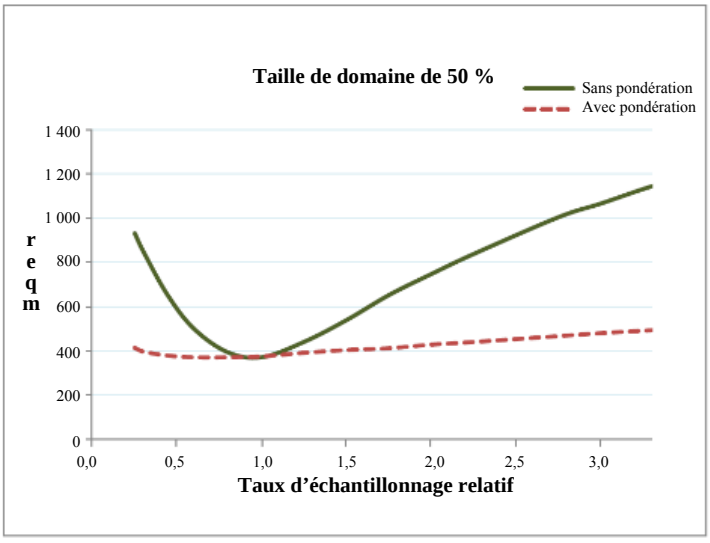

Figure 3.5 Racine de l'erreur quadratique moyenne (reqm) pour les estimateurs avec et sans pondération de la taille de domaine de 50 % pour $[CZ]^Y$ et $[C+Z]^R$.

4 Conclusions

Presque toutes les enquêtes sont touchées par la non-réponse; la méthode utilisée pour ajuster les poids de base pour la non-réponse totale est donc une question importante. L et V ont souligné à juste titre que l'utilisation de poids de sondage pour calculer un ajustement pondéré pour la non-réponse n'élimine pas le biais de non-réponse lorsque le mécanisme de réponse n'est pas spécifié correctement dans le modèle d'ajustement de la pondération. Toutefois, leur étude par simulation a porté au moins certains chercheurs à penser qu'un ajustement non pondéré pourrait mieux convenir qu'un ajustement pondéré dans la plupart des cas. Les résultats de notre évaluation, fondée sur le même scénario que celui de L et V, contredit cette perception. Nous avons examiné de façon plus approfondie les différences entre les estimateurs avec et sans pondération lorsque le modèle d'ajustement est inexact en utilisant le même scénario que L et V et en incluant différents taux d'échantillonnage et estimations des totaux et des domaines, en plus des moyennes étudiées par L et V.

Ces simulations élargies montrent que les ajustements avec et sans pondération ont effectivement des propriétés différentes. Le biais de l'estimateur pondéré des moyennes des totaux de plans d'échantillonnage aléatoires simples stratifiés est à peu près constant, quel que soit le taux d'échantillonnage, tandis que le biais de l'estimateur non pondéré dépend du taux d'échantillonnage. En revanche, le biais de l'estimateur non pondéré du total est considérablement plus important que celui de l'estimateur pondéré pour certains taux d'échantillonnage. Pour les moyennes, le biais et la reqm des deux estimateurs ne sont pas très différents, y compris pour les configurations que L et V ont décrites comme étant favorables à l'estimateur non pondéré. Les mêmes conclusions générales se vérifient pour les estimations des moyennes et des totaux de domaines à mesure que la moyenne pondérée se rapproche de plus en plus d'une estimation par ratio pour les domaines, ce qui finit par influencer quelque peu son comportement.

Nous avons aussi examiné l'estimation des tailles de domaine. Pour ce type de statistique, la reqm de l'estimateur pondéré est presque systématiquement inférieure à celle de l'estimateur non pondéré lorsque les données du scénario de simulation ne sont pas de type MAR. Les différences sont attribuables au biais de l'estimateur non pondéré de la taille de domaine; à cause de ce biais, l'estimateur non pondéré est assorti d'une reqm beaucoup plus grande que celle de l'estimateur pondéré pour certains taux d'échantillonnage.

Les modèles utilisés dans la plupart des enquêtes sont imparfaits; il est donc important de choisir la bonne méthode d'ajustement pour la non-réponse. Les résultats de la simulation élargie que nous présentons montrent que l'ajustement pondéré présente des avantages considérables pour certaines estimations et certains taux d'échantillonnage, comparativement à l'ajustement non pondéré. Plus particulièrement, une enquête selon le même plan qui produit des estimations des totaux et des statistiques autres que de simples moyennes semble bénéficier d'une pondération de l'ajustement. Bien sûr, la pondération de l'ajustement n'élimine pas le biais; elle diminue toutefois l'ampleur du biais dans bon nombre de situations et pour beaucoup des estimateurs que nous avons examinés. En outre, le biais de l'estimateur pondéré n'est pas sensible au taux d'échantillonnage relatif, alors que le biais de l'estimateur non pondéré l'est. L'inconvénient possible d'une augmentation de la variance de l'estimation lorsqu'on utilise un ajustement pondéré ne s'est pas concrétisé durant les simulations, et on peut l'éviter en examinant les facteurs d'ajustement, ce qu'il convient également de faire lorsqu'on utilise un ajustement non pondéré. Enfin, les

résultats de l'étude font ressortir le problème potentiel de la généralisation à partir de simulations. Même si les simulations constituent un outil précieux pour vérifier un point précis, une généralisation à plus grande échelle des résultats d'une simulation peut être trompeuse, particulièrement lorsque les résultats dépendent fortement des conditions du modèle utilisé pour la simulation.

Bibliographie

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics,* 29(2), 329-353.

Brick, J., et Jones, M. (2008). Propensity to respond and nonresponse bias. *Metron-International Journal of Statistics*, LXVI, 51-73.

Chadborn, T.R., Baster, K., Delpech, V., Sabin, C.A., Sinka, K., Rice, B.D. et Evans, B. (2005). No time to wait: How many HIV-infected homosexual men are diagnosed late and consequently die? (England and Wales, 1993-2002). *Aids*, 19(5), 513-520.

Grau, E., Potter, F., Williams, S. et Diaz-Tena, N. (2006). Nonresponse adjustment using logistic regression: To weight or not to weight? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3073-3080.

Haukoos, J.S., et Newgard, C.D. (2007). Advanced statistics: Missing data in clinical research - part 1: An introduction and conceptual framework. *Academic Emergency Medicine*, 14(7), 662-668.

Kalton, G. (1983). *Introduction to Survey Sampling, SAGE University Paper 35.* Thousand Oaks, CA: SAGE Publications.

Kott, P. (2012). Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes. *Techniques d'enquête*, 38, 1, 103-107.

Kovačević, M., et Yung, W. (1997). Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu – Étude empirique. *Techniques d'enquête*, 23, 1, 47-59.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M. et Raghunathan, T.E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. *Journal of the Royal Statistical Society Series A*, Royal Statistical Society, 173(2), 389-407.

Little, R.J. (1986). Survey nonresponse adjustments. *Revue Internationale de Statistique*, 54, 139-157.

Little, R.J., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data (2nd Ed.)*. New York : John Wiley & Sons, Inc.

Little, R., et Vartivarian, S. (2003). On weighting the rates in nonresponse weights. *Statistics in Medicine*, 22, 1589-1599.

R Development Core Team (2011). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. doi: http://www.R-project.org.

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

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

Sukasih, A., Jang, D., Vartivarian, S., Cohen, S. et Zhang, F. (2009). A simulation study to compare weighting methods for nonresponses in the National Survey of Recent College Graduates. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Récupéré le 21 octobre 2013, à partir de www.amstat.org/sections/srms/proceedings/y2009/Files/304345.pdf.

Tremblay, V. (1986). Critères pratiques pour la définition des classes de pondération. *Techniques d'enquête,* 12, 1, 91-103.

West, B.T. (2009). A simulation study of alternative weighting class adjustments for nonresponse when estimating a population mean from complex sample survey data. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Récupéré le 21 octobre 2013, à partir de www.amstat.org/sections/srms/proceedings/y2009/Files/305394.pdf.

Wun, L.-M., Ezzati-Rice, T.M., Diaz-Tena, N. et Greenblatt, J. (2007). On modelling response propensity for dwelling unit (DU) level non-response adjustment in the Medical Expenditure Panel Survey (MEPS). *Statistics in Medicine*, 26(8), 1875-1884.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 191-199
Statistique Canada, nº 12-001-X au catalogue

Note brève sur l'estimation fondée sur les quantiles et les expectiles dans les échantillons à probabilités inégales

Linda Schulze Waltrup et Göran Kauermann[1]

Résumé

L'estimation des quantiles est une question d'intérêt dans le contexte non seulement de la régression, mais aussi de la théorie de l'échantillonnage. Les expectiles constituent une solution de rechange naturelle ou un complément aux quantiles. En tant que généralisation de la moyenne, les expectiles ont gagné en popularité ces dernières années parce qu'en plus d'offrir un portrait plus détaillé des données que la moyenne ordinaire, ils peuvent servir à calculer les quantiles grâce aux liens étroits qui les associent à ceux-ci. Nous expliquons comment estimer les expectiles en vertu d'un échantillonnage à probabilités inégales et comment les utiliser pour estimer la fonction de répartition. L'estimateur ajusté de la fonction de répartition obtenu peut être inversé pour établir les estimations des quantiles. Nous réalisons une étude par simulations pour examiner et comparer l'efficacité de l'estimateur fondé sur des expectiles.

Mots-clés : Quantiles; expectiles; probabilité proportionnelle à la taille; approche fondée sur le plan de sondage; variable auxiliaire; fonction de répartition.

1 Introduction

Ces dernières années, l'estimation des quantiles et la régression quantile ont connu de nouveaux développements découlant des travaux de Koenker (2005). L'idée principale est d'estimer une fonction de répartition cumulative inversée, qu'on appelle généralement la fonction quantile $Q(\alpha) = F^{-1}(\alpha)$ pour $\alpha \in (0,1)$, où le quantile 0,5, $Q(0{,}5)$, la médiane, joue un rôle central. Pour le dépistage de données d'enquête à partir d'un échantillon à probabilités inégales et à probabilités connues d'inclusion, Kuk (1988) montre comment estimer les quantiles en tenant compte des probabilités d'inclusion. L'idée principale est d'estimer une fonction de répartition de la variable d'intérêt et de l'inverser pour obtenir la fonction quantile. Chambers et Dunstan (1986) proposent un estimateur fondé sur un modèle pour la fonction de répartition. Rao, Kovar et Mantel (1990) proposent un estimateur fondé sur le plan de sondage et faisant appel à des données auxiliaires pour la fonction de répartition cumulative. Chen, Elliott et Little (2010) et Chen, Elliott et Little (2012) ont également récemment proposé des approches bayésiennes allant dans le même sens.

L'estimation des quantiles résulte de la minimisation d'une fonction de perte L_1, comme l'a montré Koenker (2005). Si la perte L_1 est remplacée par la fonction de perte L_2, on obtient ce qu'on appelle des « expectiles », une notion présentée par Aigner, Amemiya et Poirier (1976) et par Newey et Powell (1987). Pour $\alpha \in (0,1)$, on obtient la fonction expectile $M(\alpha)$ qui, comme la fonction quantile $Q(\alpha)$, définit de façon unique la fonction de répartition cumulative $F(y)$. Les expectiles sont relativement faciles à estimer et suscitent un certain intérêt depuis quelques temps; voir par exemple Schnabel et Eilers (2009), Pratesi, Ranalli et Salvati (2009), Sobotka et Kneib (2012) et Guo et Härdle (2013). Ils ne sont toutefois pas faciles à interpréter, et sont donc moins acceptés et utilisés en statistique que les quantiles; voir Kneib (2013). Les quantiles et les expectiles sont reliés, c'est-à-dire qu'il existe une fonction de transformation unique et

1. Linda Schulze Waltrup, Administration des affaires et sciences sociales, Université Louis-et-Maximilien de Munich, Ludwigstraße 33, 80539 Munich, Allemagne. Courriel : lschulze_waltrup@stat.uni-muenchen.de; Göran Kauermann, Administration des affaires et sciences sociales, Université Louis-et-Maximilien de Munich, Ludwigstraße 33, 80539 Munich, Allemagne. Courriel : goeran.kauermann@stat.uni-muenchen.de.

inversible $h_y : [0,1] \to [0,1]$ de sorte que $M(h(\alpha)) = Q(\alpha)$; voir Yao et Tong (1996) et De Rossi et Harvey (2009). Cette relation peut être exploitée pour estimer les quantiles à partir d'un ensemble d'expectiles ajustés. Schulze Waltrup, Sobotka, Kneib et Kauermann (2014) ont utilisé ce principe et montré de façon empirique que les quantiles ainsi obtenus peuvent être plus efficients que les quantiles empiriques, même lorsque ces derniers font l'objet d'un lissage (voir Jones 1992). Ce résultat pourrait s'expliquer intuitivement par le fait que les expectiles tiennent compte de toutes les données, alors que les quantiles fondés sur la fonction de répartition empirique ne tiennent compte que des données de gauche (ou de droite). Autrement dit, la médiane est définie par la moitié gauche (ou droite) des données, alors que la moyenne (expectile de 50 %) est une fonction tenant compte de tous les points de données. Dans la présente note, nous utilisons ces constatations comme point de départ pour montrer comment les expectiles peuvent être estimés pour des échantillons à probabilités inégales et comment obtenir une fonction de répartition ajustée à partir d'expectiles ajustés.

La présentation de l'article est la suivante. À la section 2, on présente les éléments de notation utiles et on discute de la régression quantile dans un échantillonnage à probabilités inégales. Ce sujet est approfondi à la section 3, où l'on présente l'estimation des expectiles. À la section 4, on exploite la relation entre les expectiles et les quantiles pour montrer comment dériver les quantiles à partir d'expectiles ajustés. La section 5 présente des simulations pour illustrer le gain d'efficacité découlant de l'utilisation des quantiles dérivés d'expectiles; l'article se termine par une discussion à la section 6.

2 Estimation des quantiles

Considérons une population finie de N éléments et une variable d'enquête continue Y. On s'intéresse aux quantiles de la fonction de répartition cumulative $F(y) = \sum_{i=1}^{N} 1\{Y_i \leq y\}/N$, et on définit comme

$$Q(\alpha) = \inf\left\{\arg\min_q \sum_{i=1}^{N} w_\alpha(Y_i - q)\,|\,Y_i - q\,|\right\} \tag{2.1}$$

la fonction quantile de Y (voir Koenker 2005), où

$$w_\alpha(\varepsilon) = \begin{cases} \alpha & \text{pour } \varepsilon > 0 \\ 1-\alpha & \text{pour } \varepsilon \leq 0. \end{cases}$$

L'argument « inf » de l'expression (2.1) est nécessaire pour une population finie puisque « arg min » n'est pas unique. On tire un échantillon de la population selon des probabilités d'inclusion connues π_i, $i = 1, \ldots, N$. En notant $y_1, \ldots, y_n$ l'échantillon obtenu, on estime la fonction quantile en remplaçant (2.1) par la version avec échantillon pondéré

$$\hat{Q}_N(\alpha) = \inf\left\{\arg\min_q \sum_{j=1}^{n} \frac{1}{\pi_j} w_{\alpha,j}\,|\,y_j - q\,|\right\} \tag{2.2}$$

avec $w_{\alpha,j} = w_\alpha(y_j - q)$, selon la définition ci-dessus. Il est facile de voir que la somme en (2.2) est une estimation sans biais par rapport au plan de la somme dans $Q(\alpha)$ donnée en (2.1). Néanmoins, parce qu'on

admet « arg min », il s'ensuit que $\hat{Q}_N(\alpha)$ n'est pas sans biais pour $Q(\alpha)$. Examinons donc les énoncés de cohérence pour $\hat{Q}_N(\alpha)$ comme suit. Soit $R_i(q) = w_\alpha(y_i - q)|y_i - q|$ et

$$\bar{R}_N(q) := \frac{1}{N}\sum_i R_i(q).$$

On tire un échantillon à partir de $R_i(q), i = 1, \ldots, N$ en appliquant un plan de sondage cohérent de sorte que

$$\bar{r}_n(q) := \frac{1}{N}\sum_{j=1}^{n}\frac{1}{\pi_j} r_j(q)$$

converge par rapport au plan pour $\bar{R}_N(q)$, où $r_j(q)$ désigne l'échantillon de $R_i(q)$. Soulignons que $r_j(q)$ et donc $\bar{r}_n(q)$, $R_i(q)$ et $\bar{R}_N(q)$ dépendent aussi de α, qui a été supprimé de la notation par souci de lisibilité. Soit q_0 la valeur minimale de $\bar{R}_N(q)$, qui n'est pas nécessairement unique en raison de la structure finie de la population. On peut admettre l'argument « inf », c'est-à-dire $q_0 = \inf\{\arg\min \bar{R}_N(q)\}$, mais par souci de simplicité, on suppose un modèle de superpopulation (voir Isaki et Fuller 1982) en considérant la population finie comme un échantillon d'une superpopulation infinie. Pour cette dernière, on présume que la variable d'enquête Y a une fonction de répartition cumulative continue, de sorte que q_0 donne un quantile α unique. Pour $\delta > 0$, on obtient

$$P(\bar{r}_n(q_0) < \bar{r}_n(q_0 - \delta)) \Leftrightarrow P\left(\frac{1}{N}\sum_{j=1}^{n}\frac{1}{\pi_j}\{r_j(q_0) - r_j(q_0 - \delta)\} < 0\right).$$

Soulignons que l'argument dans l'énoncé de probabilité est une estimation convergente par rapport au plan de sondage pour $\bar{R}_N(q_0) - \bar{R}_N(q_0 - \delta)$, dont la valeur est inférieure à zéro puisque q_0 correspond à la valeur minimale de $\bar{R}_N(\cdot)$. En conséquence, la probabilité tend vers un au sens de la convergence par rapport au plan de sondage définie par Isaki et Fuller (1982). Il en va bien sûr de même pour $\delta < 0$. En vertu de cet énoncé, on peut conclure que la valeur estimée minimale $\hat{q}_0 = \arg\min \sum_{j=1}^{n} 1/\pi_j \; r_j(q)$ est une estimation convergente par rapport au plan de sondage pour q_0 de sorte que $\hat{Q}_N(\alpha)$ en (2.2) converge aussi par rapport au plan de sondage pour $Q_N(\alpha)$. Il est facile de montrer que $\hat{Q}_N(\alpha)$ est l'inverse de la fonction de répartition cumulative pondérée normalisée

$$\hat{F}_N(y) := \frac{\sum_{j=1}^{n} 1\{y_j \le y\}/\pi_j}{\sum_{j=1}^{n} 1/\pi_j}$$

selon la notation utilisée par Kuk (1988). Soulignons que $\hat{F}_N(y)$ correspond à l'estimation de Hajek (1971) pour la fonction de répartition cumulative (voir aussi Rao et Wu 2009) et n'est donc pas une estimation de Horvitz-Thompson. Par conséquent, $\hat{Q}_N(\alpha)$ n'est pas sans biais par rapport au plan de sondage. Néanmoins, $\hat{F}_N(y)$ est une fonction de répartition valide et peut donc être considérée comme une version normalisée de l'estimateur de Lahiri ou de Horvitz-Thompson de la fonction de répartition (voir Lahiri 1951), désignée par

$$\hat{F}_L(y) := \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j 1\{y_j \le y\}.$$

Kuk (1988) propose de remplacer $\hat{F}_L(\cdot)$ par d'autres estimations de la fonction de répartition : au lieu d'estimer la fonction de répartition elle-même, il suggère d'estimer la proportion complémentaire $\hat{S}_R(y)$ qui mène ensuite à l'estimation $\hat{F}_R(y)$ définie par

$$\hat{F}_R(y) = 1 - \hat{S}_R(y) = 1 - \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j 1\{y_j > y\}.$$

Directement à partir de ces définitions, on peut exprimer $\hat{F}_R(\cdot)$ en termes de $\hat{F}_N(\cdot)$ par

$$\hat{F}_R = 1 - \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j + \hat{F}_L \quad \text{et} \quad \hat{F}_L = \frac{\sum_{j=1}^{n} 1/\pi_j}{N}\hat{F}_N. \tag{2.3}$$

Kuk (1988) montre que, en vertu d'un échantillonnage à probabilités inégales, l'estimation de la médiane dérivée de $\hat{F}_R$ est plus efficace que celles qui sont dérivées de $\hat{F}_N$ et $\hat{F}_L$ en termes d'estimation de l'erreur quadratique moyenne. Soulignons que les estimateurs $\hat{F}_N$, $\hat{F}_L$ et $\hat{F}_R$ coïncident dans le cas d'un échantillonnage aléatoire simple sans remise où $\pi_j = \pi = n/N$.

3 Estimation des expectiles

Les expectiles sont une solution de rechange aux quantiles. La fonction expectile $M(\alpha)$ est donc définie par remplacement de la perte L_1 dans l'expression (2.1) par la perte L_2 pour donner

$$M(\alpha) = \arg\min_m \left\{\sum_{i=1}^{N} w_\alpha (Y_i - m)(Y_i - m)^2\right\}. \tag{3.1}$$

Soulignons que $M(\alpha)$ est continue en α même pour des populations finies. En outre, $M(0{,}5)$ est égale à la valeur moyenne $\bar{Y} = \sum_{i=1}^{N} Y_i/N$. À partir de l'échantillon $y_1, \ldots, y_n$ avec probabilités d'inclusion $\pi_1, \ldots, \pi_n$, on peut estimer $M(\alpha)$ en remplaçant la somme de l'expression (2.2) par sa version d'échantillon, c'est-à-dire

$$\hat{M}(\alpha) = \arg\min_m \left\{\sum_{j=1}^{n} \frac{1}{\pi_j} w_{\alpha,j} (y_j - m)^2\right\}$$

avec $w_{\alpha,j}$ correspondant à la définition ci-dessus. Il est facile de voir que la somme dans $\hat{M}(\alpha)$ est une estimation sans biais par rapport au plan de la somme dans $M(\alpha)$. L'estimation elle-même n'est toutefois pas sans biais par rapport au plan comme pour la fonction quantile susmentionnée. Toutefois, on peut utiliser les mêmes arguments que pour $Q_N(\alpha)$ en (2.2) pour établir la convergence par rapport au plan de sondage.

4 Des expectiles à la fonction de répartition

La fonction quantile $Q(\alpha)$ et la fonction expectile $M(\alpha)$ définissent toutes deux de façon unique une fonction de répartition $F(.)$. Tandis que $Q(\alpha)$ est une simple inversion de $F(.)$, la relation entre $M(\alpha)$ et $F(.)$ est plus complexe. Selon Schnabel et Eilers (2009) et Yao et Tong (1996), on peut établir la relation

$$M(\alpha) = \frac{(1-\alpha)\,G(M(\alpha)) + \alpha\{M(0{,}5) - G(M(\alpha))\}}{(1-\alpha)\,F(M(\alpha)) + \alpha\{1 - F(M(\alpha))\}}, \tag{4.1}$$

où $G(m)$ est la fonction génératrice des moments définie par $G(m) = \sum_{i=1}^{N} Y_i\, 1\{Y_i \le m\}/N$. L'expression (4.1) donne la relation unique de la fonction $M(\alpha)$ à la fonction de répartition $F(.)$. Il faut maintenant résoudre (4.1) pour $F(.)$, c'est-à-dire exprimer la répartition $F(.)$ en termes de la fonction expectile $M(.)$. Cela n'est apparemment pas possible sous une forme analytique, mais on peut effectuer le calcul numériquement. Pour ce faire, on évalue la fonction ajustée $\hat{M}(\alpha)$ selon un ensemble dense de valeurs $0 < \alpha_1 < \alpha_2 \ldots < \alpha_L < 1$, en désignant les valeurs ajustées par $\hat{m}_l = \hat{M}(\alpha_l)$. On définit aussi des bornes à gauche et à droite par $\hat{m}_o = \hat{m}_1 - c_0$ et $\hat{m}_{L+1} = \hat{m}_L + c_{L+1}$, où c_0 et c_L sont des constantes définies par l'utilisateur. Par exemple, on peut définir $c_0 = \hat{m}_2 - \hat{m}_1$ et $c_{L+1} = \hat{m}_L - \hat{m}_{L-1}$. Ce faisant, on dérive les valeurs ajustées pour la fonction de répartition cumulative $F(.)$ à $\hat{m}_l$, que l'on écrit $\hat{F}_l := \hat{F}(\hat{m}_l) = \sum_{j=1}^{l} \hat{\delta}_j$ pour les échelons non négatifs $\hat{\delta}_j \ge 0, j = 1, \ldots, L$ avec $\sum_{j=1}^{L} \hat{\delta}_j \le 1$. On définit $\hat{\delta}_{L+1} = 1 - \sum_{l=1}^{L} \hat{\delta}_l$ pour faire de $\hat{F}(.)$ une fonction de répartition. En supposant une répartition uniforme entre les points de support $\hat{m}_l$ de l'ensemble dense, on peut exprimer la fonction de génération des moments $G(.)$ par simple intégration séquentielle comme

$$\hat{G}_l := \hat{G}(\hat{m}_l) = \int_{-\infty}^{m_l} x\, d\hat{F}(x) = \sum_{j=1}^{l} \hat{d}_j \hat{\delta}_l,$$

où $\hat{d}_j = (\hat{m}_j - \hat{m}_{j-1})/2$ sous la contrainte que $\hat{G}_{L+1} = \hat{M}(0{,}5)$ et $\hat{M}(0{,}5) = \sum_{j=1}^{n}(y_j/\pi_j)\big/\sum_{j=1}^{n}(1/\pi_j)$. Avec les échelons $\hat{\delta}_l, l = 1, \ldots, L$, on peut maintenant réécrire l'expression (4.1) comme

$$\hat{m}_l = \frac{(1-\alpha)\sum_{j=1}^{l} \hat{d}_j \hat{\delta}_j + \alpha\left(\hat{M}(0{,}5) - \sum_{j=1}^{l} \hat{d}_j \hat{\delta}_j\right)}{(1-\alpha)\sum_{j=1}^{l} \hat{\delta}_j + \alpha\left(1 - \sum_{j=1}^{l} \hat{\delta}_j\right)}, \quad l = 1, \ldots, L,$$

que l'on résout ensuite pour $\hat{\delta}_1, \ldots, \hat{\delta}_L$. Il s'agit d'un exercice numérique relativement direct sur le plan conceptuel. On peut consulter Schulze Waltrup et coll. (2014) pour les détails. Une fois qu'on a calculé $\hat{\delta}_1, \ldots, \hat{\delta}_L$, on obtient une estimation pour la fonction de répartition cumulative, qu'on écrit $\hat{F}_N^M(y) = \sum_{l:\hat{m}_l < y} \hat{\delta}_l$. On peut aussi inverser $\hat{F}_N^M(.)$, ce qui donne une fonction quantile ajustée que l'on désigne $\hat{Q}_N^M(\alpha)$.

Comme le montre Kuk (1988), à la fois théoriquement et empiriquement, $\hat{F}_R(.)$ est plus efficiente que $\hat{F}_N(.)$. On exploite cette relation en l'appliquant à $\hat{F}_N^M(.)$ pour obtenir l'estimateur

$$\hat{F}_R^M := 1 - \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j + \frac{\sum_{j=1}^{n} 1/\pi_j}{N}\hat{F}_N^M .$$

Dans la section qui suit, on compare les quantiles calculés à partir de l'estimateur fondé sur les expectiles $\hat{F}_R^M$ avec les quantiles calculés à partir de $\hat{F}_R$. Soulignons que $\hat{F}_R^M$ et $\hat{F}_R$ ne sont pas des fonctions de répartition appropriées puisqu'elles ne sont pas normalisées pour prendre des valeurs situées entre 0 et 1.

5 Simulations

On a réalisé une petite étude par simulations pour illustrer l'efficacité des estimations fondées sur les expectiles. On utilise ci-dessous la méthode d'échantillonnage de Midzuno (voir Midzuno 1952); les probabilités d'inclusion π_j sont définies comme étant proportionnelles à une mesure de la taille x, selon le module « *sampling* » de Tillé et Matei (2015) dans le logiciel R. On examine deux ensembles de données, que Kuk (1988) a aussi utilisés. Le premier ensemble de données (Logements) comprend deux variables fortement corrélées (corrélation de 0,97), soit le nombre d'unités de logement (X) et le nombre d'unités louées (Y); voir aussi Kish (1965). Le deuxième ensemble de données (Villages) comprend de l'information sur la population (X) et sur le nombre de personnes travaillant dans des entreprises familiales (Y) dans 128 villages de l'Inde; voir Murthy (1967). Dans le deuxième ensemble de données, la corrélation entre Y et X s'établit à 0,54. Afin de comparer les résultats de la simulation à ceux de Kuk (1988), on a choisi un échantillon de la même taille, soit $n = 30$ (pour une population totale de $N = 270$ en ce qui concerne les données sur les logements et de $N = 128$ pour les données sur les villages).

On compare les quantiles définis par l'inversion de $\hat{F}_R$ avec les quantiles définis par l'inversion de $\hat{F}_R^M$. Le tableau 5.1 présente la racine de l'erreur quadratique moyenne (REQM) et l'efficience relative pour certains quantiles. On constate que la médiane pour les données relatives aux villages et, pour les données relatives aux logements, les quantiles supérieurs dérivés des expectiles ont une efficience accrue. En outre, le gain d'efficience n'est pas uniforme; on constate en effet une perte d'efficience dans les quantiles inférieurs.

Tableau 5.1
Comparaison de l'erreur quadratique moyenne fondée sur 500 répliques

	α	quantiles $\sqrt{\text{EQM}\left(\hat{Q}_R(\alpha)\right)}$	quantiles dérivés des expectiles $\sqrt{\text{EQM}\left(\hat{Q}_R^M(\alpha)\right)}$	efficience relative $\frac{\sqrt{\text{EQM}\left(\hat{Q}_R^M(\alpha)\right)}}{\sqrt{\text{EQM}\left(\hat{Q}_R(\alpha)\right)}}$
Logements	0,1	2,57	2,76	1,07
	0,25	1,77	1,97	1,11
	0,5	2,45	2,35	0,96
	0,75	3,15	2,91	0,92
	0,9	4,20	3,43	0,82
Villages	0,1	5,52	6,65	1,21
	0,25	11,41	10,31	0,90
	0,5	12,29	11,69	0,95
	0,75	16,24	15,41	0,95
	0,9	13,31	18,34	1,38

Pour mieux comprendre, on a réalisé une simulation à l'aide d'un échantillon plus grand de taille $n = 100$ sélectionné à partir de populations de tailles $N = 1\,000$ et $N = 10\,000$. On a tiré Y et X d'une loi log-normale standard bivariée avec $\mu = 0$ et $\sigma = 1$. Les variables Y et X sont tirées de façon que la corrélation entre les variables soit égale à 0,9. On a encore une fois calculé la racine de l'erreur quadratique moyenne pour une gamme de valeurs de α; l'efficience relative de l'approche fondée sur les expectiles est illustrée à la figure 5.1. Pour une meilleure présentation visuelle, les graphiques donnent une version lissée de l'efficience relative. On constate une diminution de la racine de l'erreur quadratique moyenne dans les deux cas, soient $N = 1\,000$ et $N = 10\,000$. On peut conclure que les expectiles peuvent être facilement ajustés dans un échantillonnage à probabilités inégales et que la relation entre les expectiles et la fonction de répartition peut être exploitée numériquement pour calculer les quantiles avec une efficience accrue. Ce gain d'efficience ne se vérifie que pour les quantiles supérieurs, c'est-à-dire pour les valeurs de α dont la borne inférieure est strictement positive. Soulignons toutefois que le plan de sondage est tel que les grandes valeurs de Y sont échantillonnées selon une probabilité supérieure, puisque le plan de sondage vise à obtenir des estimations plus fiables pour le côté droit de la fonction de répartition, c'est-à-dire pour les quantiles supérieurs. Si on s'intéresse aux quantiles inférieurs, il faut utiliser un plan de sondage différent en attribuant une probabilité d'inclusion accrue aux personnes ayant une valeur Y faible. Dans ce cas, on observerait un comportement correspondant au reflet de celui qui est illustré à la figure 5.1 en ce qui concerne α.

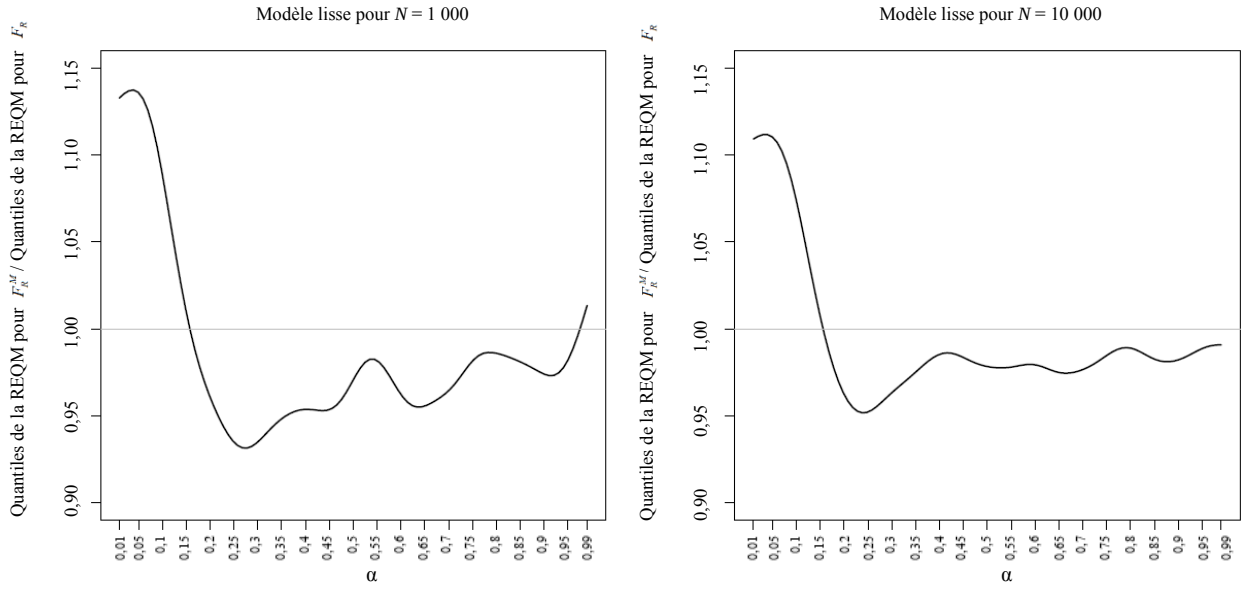

Figure 5.1 Racine de l'erreur quadratique moyenne (REQM) relative des quantiles et des quantiles dérivés des expectiles pour le plan de sondage avec probabilité proportionnelle à la taille (PPT), calculée à partir de 500 répliques (à gauche : $N = 1\,000$; à droite : $N = 10\,000$).

6 Discussion

À la section 4, on a augmenté la boîte à outils des expectiles à l'estimation des fonctions de répartition dans le cadre d'un échantillonnage à probabilités inégales. On a défini les expectiles pour des échantillons

à probabilités inégales. Quand on compare les quantiles fondés sur $\hat{F}_R$ avec les quantiles reposant sur l'estimateur fondé sur les expectiles $\hat{F}_R^M$, on constate que l'estimateur proposé offre un bon rendement en comparaison des méthodes existantes. Le calcul des expectiles empiriques est mis en œuvre dans le logiciel libre R (voir R Core Team 2014) et se trouve dans le module expectreg du logiciel R mis au point par Sobotka, Schnabel, et Schulze Waltrup (2013). Le calcul de l'estimateur de la fonction de répartition fondé sur les expectiles $\hat{F}_N^M$ fait aussi partie du module expectreg du logiciel R. Le calcul de $\hat{F}_R^M$ est toutefois plus exigeant que celui de $\hat{F}_R$ parce qu'il comporte trois étapes : il faut d'abord calculer les expectiles pondérés selon la méthode exposée à la section 3, puis estimer $\hat{F}_R^N$ et enfin, dériver $\hat{F}_R^M$ à partir de $\hat{F}_R^N$ (voir la section 4). Dans la simulation log-normale, il faut environ 2-3 secondes pour calculer $\hat{F}_R^M$ pour $N = 1\ 000$, tandis que l'effort pour calculer $\hat{F}_R$ est négligeable.

Remerciements

Les deux auteurs remercient la *Deutsche Forschungsgemeinschaft* DFG (KA 1188/7-1) pour le soutien financier.

Bibliographie

Aigner, D.J., Amemiya, T. et Poirier, D.J. (1976). On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review*, 17(2), 377-396.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597-604.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2010). Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 36, 1, 25-37.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2012). Inférence bayésienne pour les quantiles de population finie sous échantillonnage avec probabilités inégales. *Techniques d'enquête*, 38, 2, 221-233.

De Rossi, G., et Harvey, A. (2009). Quantiles, expectiles and splines. Nonparametric and robust methods in econometrics. *Journal of Econometrics*, 152(2), 179-185.

Guo, M., et Härdle, W. (2013). Simultaneous confidence bands for expectile functions. *AStA - Advances in Statistical Analysis*, 96(4), 517-541.

Hajek, J. (1971). Comment on "An essay on the logical foundations of survey sampling, part one". *The Foundations of Survey Sampling*, 236.

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

Jones, M. (1992). Estimating densities, quantiles, quantile densities and density quantiles. *Annals of the Institute of Statistical Mathematics*, 44(4), 721-727.

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

Kneib, T. (2013). Beyond mean regression (with discussion and rejoinder). *Statistical Modelling*, 13(4), 275-385.

Koenker, R. (2005). *Quantile Regression, Econometric Society Monographs*. Cambridge: Cambridge University Press.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75(1), 97-103.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute*, (33), 133-140.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics*, 3, 99-107.

Murthy, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

Newey, W.K., et Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55(4), 819-847.

Pratesi, M., Ranalli, M. et Salvati, N. (2009). Nonparametric M-quantile regression using penalised splines. *Journal of Nonparametric Statistics*, 21(3), 287-304.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienne, Autriche : R Foundation for Statistical Computing.

Rao, J., et Wu, C. (2009). Empirical likelihood methods. *Handbook of Statistics*, 29B, 189-207.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77(2), 365-375.

Schnabel, S.K., et Eilers, P.H. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, 53(12), 4168-4177.

Schulze Waltrup, L., Sobotka, F., Kneib, T. et Kauermann, G. (2014). Expectile and quantile regression - David and Goliath? *Statistical Modelling*, 15, 433-456.

Sobotka, F., et Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56(4), 755-767.

Sobotka, F., Schnabel, S. et Schulze Waltrup, L. (2013). *Expectreg: Expectile and Quantile Regression*. Avec la contribution de P. Eilers, T. Kneib et G. Kauermann, R package version 0.38.

Tillé, Y., et Matei, A. (2015). *Sampling: Survey Sampling. R package, version 2.7*. https://cran.r-project.org/web/packages/sampling/index.html.

Yao, Q., et Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics*, 6(2-3), 273-292.

ADDENDUM

Répartition optimale assistée par modèle pour des domaines planifiés en utilisant l'estimation composite

Wilford B. Molefe et Robert Graham Clark
Volume 41, numéro 2, (décembre 2015), 399-410

Dans le deuxième paragraphe de la page 378 de notre document, nous examinons le document de Choudhry, Rao et Hidiroglou, publié en 2012. Tel que formulé, le paragraphe sous-entend une critique de ce document, ce qui ne correspond pas à nos intentions et nous souhaitons par la présente corriger et clarifier notre examen. Les coefficients de variation (c.v.) auxquels nous avons fait référence se trouvent au tableau 5 de Choudhry et coll. (2012), et l'en-tête du tableau indique clairement que les c.v. sont des estimateurs composites, plutôt que de nature non spécifiée comme nous l'avons incorrectement mentionné. Nous avons également suggéré que certains c.v. de ce tableau sont étonnamment élevés. Ce serait bel et bien le cas si les c.v. (appelés en fait des racines carrées relatives de l'erreur quadratique moyenne, selon les règles courantes) avaient été calculés au moyen de l'approximation de Longford (2006) ou des erreurs quadratiques moyennes prévues, qui y sont très semblables. Cependant, Choudhry et coll. (2012) ont utilisé un estimateur de l'erreur quadratique moyenne différent (et plus conventionnel), et, compte tenu de ces renseignements, les valeurs élevées ne sont pas surprenantes.

Nous avons également affirmé que Choudhry et coll. (2012) n'ont pas étudié la question de savoir si d'autres plans, comme la répartition exponentielle, pouvaient donner des valeurs plus faibles du critère de Longford. Cela est exact, et cette situation nous a incités à effectuer la recherche sur le sujet dans notre document. Néanmoins, nous aurions dû préciser que Choudhry et coll. (2012) avaient pris en compte la répartition par la racine carrée, un cas particulier de la répartition exponentielle, selon d'autres critères comme l'établissement de niveaux de tolérance pour les c.v. sur de petits domaines.

Bibliographie

Choudhry, G.H., Rao, J.N.K. et Hidiroglou, M.A. (2012). À propos de la répartition de l'échantillon pour une estimation sur domaine efficace. *Techniques d'enquête,* 38, 1, 25-32.

Longford, N.T. (2006). Calcul de la taille de l'échantillon pour l'estimation pour petits domaines. *Techniques d'enquête*, 32, 1, 97-106.

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 31, No. 4, 2015

Letter to the Editor
Bijak, Jakub/Alberts, Isabel/Alho, Juha/Bryant, John/Buettner, Thomas/Falkingham, Jane/ Forster, Jonathan J./ Gerland, Patrick/King, Thomas/Onorante, Luca/Keilman, Nico/O'Hagan, Anthony/Owens, Darragh/ Raftery, Adrian/Ševčíková, Hana/Smith, Peter W.F. 537

Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations
Barron, Martin/Davern, Michael/Montgomery, Robert/Tao, Xian/Wolter, Kirk M./ Zeng, Wei/Dorell, Christina/Black, Carla 545

Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes
Bavdaž, Mojca/Giesen, Deirdre/Černe, Simona Korenjak/Löfgren, Tora/Raymond-Blaess, Virginie 559

Statistical Estimators Using Jointly Administrative and Survey Data to Produce French Structural Business Statistics
Brion, Philippe/Gros, Emmanuel 589

First Impressions of Telephone Survey Interviewers
Broome, Jessica 611

Quarterly Regional GDP Flash Estimates by Means of Benchmarking and Chain Linking
Cuevas, Ángel/Quilis, Enrique M./Espasa, Antoni 627

Coordination of Conditional Poisson Samples
Grafström, Anton/Matei, Alina 649

Cultural Variations in the Effect of Interview Privacy and the Need for Social Conformity on Reporting Sensitive Information
Mneimneh, Zeina M./Tourangeau, Roger/Pennell, Beth-Ellen/Heeringa, Steven G./Elliott, Michael R. 673

Frameworks for Guiding the Development and Improvement of Population Statistics in the United Kingdom
Raymer, James/Rees, Phil/Blake, Ann 699

B-Graph Sampling to Estimate the Size of a Hidden Population
Spreen, Marinus/Bogaerts, Stefan 723

Quality Indicators for Statistical Disclosure Methods: A Case Study on the Structure of Earnings Survey
Templ, Matthias 737

Effects of Cluster Sizes on Variance Components in Two-Stage Sampling
Valliant, Richard/Dever, Jill A./Kreuter, Frauke 763

On Proxy Variables and Categorical Data Fusion
Zhang, Li-Chun 783

Book Review: Online Panel Research: A Data Quality Perspective
Cornesse, Carina/Blom, Annelies G. 809

Book Review: Practical Tools for Designing and Weighting Survey Samples
Espejo, Mariano Ruiz 813

Book Review: Managing and Sharing Research Data: A Guide to Good Practice
Mulcahy, Timothy Michael 817

Editorial Collaborators 821
Index to Volume 31, 2015 827

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 32, No. 1, 2016

Micro- and Macrodata: a Comparison of the Household Finance and Consumption Survey with Financial Accounts in Austria
Andreasch, Michael/Lindner, Peter 1

Respondent-Driven Sampling – Testing Assumptions: Sampling with Replacement
Barash, Vladimir D./Cameron, Christopher J./Spiller, Michael W./Heckathorn, Douglas D. 29

Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes
Conrad, Frederick G. / Couper, Mick P. / Sakshaug, Joseph W. 75

Census Model Transition: Contributions to its Implementation in Portugal
Dias, Carlos A./Wallgren, Anders/Wallgren, Britt/Coelho, Pedro S. 93

Constructing Synthetic Samples
Dong, Hua/Meeden, Glen 113

A Discussion of Weighting Procedures for Unit Nonresponse
Haziza, David/Lesage, Éric 129

A Note on the Effect of Data Clustering on the Multiple-Imputation Variance Estimator:
A Theoretical Addendum to the Lewis et al. article in JOS 2014
He, Yulei/Shimizu, Iris/Schappert, Susan/Xu, Jianmin/Beresovsky, Vladislav/Khan, Diba/Valverde, Roberto/
Schenker, Nathaniel 147

Sample Representation and Substantive Outcomes Using Web With and Without Incentives Compared to
Telephone in an Election Survey
Lipps, Oliver/Pekari, Nicolas 165

Bayesian Predictive Inference of a Proportion under a Twofold Small-Area Model
Nandram, Balgobin 187

SELEKT – A Generic Tool for Selective Editing
Norberg, Anders 209

Synthetic Multiple-Imputation Procedure for Multistage Complex Samples
Zhou, Hanzhi/Elliott, Michael R./Raghunathan, Trivellore E. 231

Book Review
House, Carol 257

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 44, No. 1, March/mars 2016

Issue Information-Editorial Board....1

Issue Information-Masthead....2

Mary E. Thompson, Lilia L. Ramirez Ramirez, Vyacheslav Lyubchich and Yulia R. Gel
Using the bootstrap for statistical inference on random graphs....3

Yiwei Jiang and Zehua Chen
A sequential scaled pairwise selection approach to edge detection in nonparanormal graphical models....25

Esra Kürüm, John Hughes and Runze Li
A semivarying joint model for longitudinal binary and continuous outcomes....44

Wenhua Wei and Yong Zhou
Semiparametric maximum likelihood estimation for a two-sample density ratio model with right-censored data....58

Jiahua Chen, Pengfei Li and Yukun Liu
Sample-size calculation for tests of homogeneity....82

Dongliang Wang and Yichuan Zhao
Jackknife empirical likelihood for comparing two Gini indices....102

Acknowledgement of referees' services: Remerciements aux lecteurs critiques....120

CONTENTS TABLE DES MATIÈRES

Volume 44, No. 2, June/juin 2016

Issue Information - Ed board and Masthead....125

Qing Liu, Gong Tang, Joseph P. Costantino and Chung-Chou H. Chang
Robust prediction of the cumulative incidence function under non-proportional subdistribution hazards....127

James P. Long, Noureddine El Karoui and John A. Rice
Kernel density estimation with Berkson error....142

Baisuo Jin, Yuehua Wu and Xiaoping Shi
Consistent two-stage multiple change-point detection in linear models....161

Aixin Tan and Jian Huang
Bayesian inference for high-dimensional linear regression under mnet priors....180

Michelle Xia and Paul Gustafson
Bayesian regression models adjusting for unidirectional covariate misclassification....198

Marina M. De Queiroz, Roger W. C. Silva and Rosangela H. Loschi
Shannon entropy and Kullback–Leibler divergence in multivariate log fundamental skew-normal and related distributions....219

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.