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

Techniques d'enquête 41-1

Date de diffusion : le 29 juin 2015

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Programme des services de dépôt

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

Normes de service à la clientèle

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

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)

Techniques d'enquête

Nº 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Juin 2015

•

Volume 41

•

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, (rte@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca/Techniquesdenquete).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 41, numéro 1, juin 2015

Table des matières

Articles réguliers

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 2015
Vol. 41, n° 1, pp. 1-20
Statistique Canada, n° 12-001-X au catalogue

Estimation sur petits domaines sous un modèle de Fay-Herriot avec test préliminaire pour la présence d'effets aléatoires de domaine

Isabel Molina, J.N.K. Rao et Gauri Sankar Datta[1]

Résumé

Le modèle de Fay-Herriot est un modèle au niveau du domaine d'usage très répandu pour l'estimation des moyennes de petit domaine. Ce modèle contient des effets aléatoires en dehors de la régression linéaire (fixe) basée sur les covariables au niveau du domaine. Les meilleurs prédicteurs linéaires sans biais empiriques des moyennes de petit domaine s'obtiennent en estimant les effets aléatoires de domaine, et ils peuvent être exprimés sous forme d'une moyenne pondérée des estimateurs directs propres aux domaines et d'estimateurs synthétiques de type régression. Dans certains cas, les données observées n'appuient pas l'inclusion des effets aléatoires de domaine dans le modèle. L'exclusion de ces effets de domaine aboutit à l'estimateur synthétique de type régression, autrement dit un poids nul est appliqué à l'estimateur direct. L'étude porte sur un estimateur à test préliminaire d'une moyenne de petit domaine obtenu après l'exécution d'un test pour déceler la présence d'effets aléatoires de domaine. Parallèlement, elle porte sur les meilleurs prédicteurs linéaires sans biais empiriques des moyennes de petit domaine qui donnent toujours des poids non nuls aux estimateurs directs dans tous les domaines, ainsi que certains estimateurs de rechange basés sur le test préliminaire. La procédure de test préliminaire est également utilisée pour définir de nouveaux estimateurs de l'erreur quadratique moyenne des estimateurs ponctuels des moyennes de petit domaine. Les résultats d'une étude par simulation limitée montrent que, si le nombre de domaines est petit, la procédure d'essai préliminaire mène à des estimateurs de l'erreur quadratique moyenne présentant un biais relatif absolu moyen considérablement plus faible que les estimateurs de l'erreur quadratique moyenne usuels, surtout quand la variance des effets aléatoires est faible comparativement aux variances d'échantillonnage.

Mots-clés : Modèle au niveau du domaine; meilleur prédicteur linéaire sans biais empirique; erreur quadratique moyenne; test préliminaire; estimation sur petits domaines.

1 Introduction

Un modèle de base au niveau du domaine, appelé modèle de Fay-Herriot (FH), est souvent utilisé pour obtenir des estimateurs efficaces des moyennes de domaine quand les tailles d'échantillon dans les domaines sont petites. Ce modèle comprend des effets aléatoires de domaine non observables, et le meilleur prédicteur linéaire sans biais empirique (EBLUP pour *empirical best linear unbiased predictor*) d'une moyenne de petit domaine s'obtient en estimant l'effet aléatoire associé. L'EBLUP est une combinaison pondérée d'un estimateur direct propre au domaine et d'un estimateur synthétique de type régression qui utilise toutes les données. Un estimateur de l'erreur quadratique moyenne (EQM) de l'EBLUP a été obtenu pour la première fois par Prasad et Rao (1990) en utilisant un estimateur par la méthode des moments de la variance des effets aléatoires, et plus tard par Datta et Lahiri (2000) pour l'estimateur du maximum de vraisemblance restreint (MVRE) de la variance. Rao (2003, Chapitre 7) donne un compte rendu détaillé des EBLUP et des estimateurs de leur EQM pour les modèles FH.

Parfois, les données observées n'appuient pas l'inclusion des effets de domaine dans le modèle. L'exclusion de ces effets mène à l'estimateur synthétique de type régression. Partant de cette idée, Datta,

1. Isabel Molina, Département de statistiques, Université Carlos III de Madrid, C/Madrid 126, 28903 Getafe (Madrid), Espagne et Instituto de Ciencias Matemáticas (ICMAT), (Madrid), Espagne. Courriel : isabel.molina@uc3m.es; J.N.K. Rao, École de mathématiques et de statistiques, Université Carleton, Ottawa, Canada; Gauri Sankar Datta, Département de statistiques, Université de Géorgie, Athens, États-Unis.

Hall et Mandal (2011) ont proposé d'effectuer un test préliminaire pour déterminer la présence d'effets aléatoires de domaine à un seuil de signification spécifié, et de définir l'estimateur pour petits domaines en fonction du résultat du test. Si l'hypothèse nulle de l'absence d'effets aléatoires de domaine n'est pas rejetée, le modèle sans effets de domaine est pris en considération pour estimer les moyennes de petit domaine, c'est-à-dire que l'estimateur synthétique de type régression est utilisé. Si l'hypothèse nulle est rejetée, l'EBLUP habituel sous le modèle FH avec effets de domaine est utilisé. Datta et coll. (2011) ont remarqué que l'estimateur à test préliminaire (ETP) pouvait aboutir à des gains d'efficacité importants comparativement à l'EBLUP, particulièrement quand le nombre de petits domaines est modéré. En guise de test préliminaire, ils ont considéré un test fondé sur la normalité, ainsi qu'un test de type bootstrap qui permet d'éviter l'hypothèse de normalité.

Quand la variance estimée des effets de domaine est nulle, l'EBLUP devient automatiquement l'estimateur synthétique de type régression. Cependant, l'EQM estimée obtenue par Prasad et Rao (1990) ou par Datta et Lahiri (2000) ne se réduit pas à l'EQM estimée de l'estimateur synthétique de type régression. Donc, les estimateurs habituels de l'EQM sont biaisés pour une faible variance des effets aléatoires. Par conséquent, nous proposons des estimateurs de l'EQM de l'EBLUP basés sur la procédure de test préliminaire (TP). Si le test indique que la variance des effets aléatoires n'est pas importante, nous considérons l'estimateur de l'EQM de l'estimateur synthétique. Sinon, nous considérons les estimateurs habituels de l'EQM de l'EBLUP.

L'EBLUP applique un poids nul aux estimations directes pour tous les domaines quand la variance estimée des effets de domaine est nulle. Par ailleurs, les praticiens des sondages préfèrent souvent appliquer un poids strictement positif aux estimations directes, parce que ces dernières utilisent les données au niveau de l'unité propres au domaine disponibles et intègrent aussi le plan de sondage. Li et Lahiri (2010) ont présenté un estimateur du maximum de vraisemblance ajusté (MVA) de la variance des effets aléatoires qui est toujours positif et, par conséquent, mène à des EBLUP donnant des poids strictement positifs aux estimateurs directs. Comme nous le verrons, un prix est payé sous forme de biais lorsqu'on utilise l'EBLUP basé sur l'estimateur MVA. Nous proposons ici d'autres options d'estimateurs pour petits domaines qui donnent toujours un poids positif aux estimateurs directs, mais avec un biais plus faible.

Dans le présent article, nous étudions empiriquement les propriétés des estimateurs ETP des moyennes de petit domaine comparativement aux EBLUP habituels et à d'autres estimateurs proposés. En particulier, nous étudions le choix du seuil de signification pour les estimations de domaine et pour les estimations de l'EQM basées sur le test préliminaire (TP). Les EBLUP basés sur l'estimateur MVA de la variance des effets aléatoires de Li et Lahiri (2010), qui donnent des poids non nuls aux estimateurs directs dans tous les domaines, sont également étudiés et comparés aux versions TP de l'estimateur MVA (TP-MVA). Différents estimateurs de l'EQM de ces estimateurs ETP-MVA sont également étudiés en ce qui a trait au biais relatif. En nous fondant sur les résultats de simulation, nous recommandons les EBLUP et les estimateurs de l'EQM qui ont de bonnes propriétés. Enfin, nous examinons la couverture et la longueur des intervalles de prédiction fondés sur l'hypothèse de normalité, obtenus en utilisant les EBLUP et les estimateurs de l'EQM associés.

La présentation de l'article est la suivante. À la section 2, nous décrivons le modèle FH et les EBLUP des moyennes de petit domaine. À la section 3, nous commentons l'estimation de l'EQM. À la section 4, nous présentons les estimateurs ETP des moyennes de petit domaine et les estimateurs de l'EQM basés sur

le TP. À la section 5, nous décrivons les estimateurs pour petits domaines et les estimateurs de l'EQM associés sous estimation MVA de la variance des effets de domaine. À la section 6, nous présentons des estimateurs de rechange qui appliquent également des poids positifs aux estimateurs directs, ainsi que les estimateurs de l'EQM proposés. À la section 7, nous présentons les résultats de l'étude par simulation. Enfin, à la section 8, nous tirons certaines conclusions.

2 Estimation des moyennes de petit domaine

Considérons une population partitionnée en m domaines et soit θ_i la moyenne de la variable d'intérêt pour le domaine $i, i = 1, \ldots, m$. Nous supposons qu'un échantillon est tiré indépendamment dans chaque domaine. Soit y_i un estimateur direct sans biais sous le plan de sondage de θ_i obtenu en utilisant des données d'enquête provenant du domaine échantillonné i. Les estimateurs directs sont très inefficaces pour les domaines dont l'échantillon est de petite taille. Nous étudions l'estimation sur petits domaines sous un modèle au niveau du domaine, dans lequel les valeurs des covariables au niveau du domaine sont disponibles pour tous les domaines. Le modèle fondamental de ce type est le modèle de Fay-Herriot, introduit par Fay et Herriot (1979) pour estimer le revenu par habitant dans de petites localités aux États-Unis. Ce modèle comprend deux parties. La première repose sur l'hypothèse que les estimateurs directs, y_i, des moyennes de petit domaine, θ_i, sont sans biais sous le plan de sondage, et satisfont

$$y_i = \theta_i + e_i, \quad e_i \overset{\text{ind}}{\sim} N(0, D_i), \quad i = 1, \ldots, m. \tag{2.1}$$

Ici, la variance d'échantillonnage $D_i = \text{Var}(y_i | \theta_i)$ est supposée connue pour tous les domaines $i = 1, \ldots, m$. En pratique, les valeurs de D_i sont déterminées à partir de sources externes ou en lissant les variances d'échantillonnage estimées en utilisant une méthode à fonction de variance généralisée (Fay et Herriot 1979).

Dans la deuxième partie du modèle de Fay-Herriot, θ_i est traité comme étant aléatoire et l'on suppose qu'un vecteur p de covariables au niveau du domaine, $\mathbf{x}_i$, relié linéairement à θ_i, est disponible pour chaque domaine i, c'est-à-dire

$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i, \quad v_i \overset{\text{iid}}{\sim} N(0, A), \quad i = 1, \ldots, m, \tag{2.2}$$

où v_i est l'effet aléatoire du domaine i, supposé indépendant de e_i, et $A \geq 0$ est la variance des effets aléatoires. Observons que, marginalement,

$$y_i \overset{\text{ind}}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta}, D_i + A), \quad i = 1, \ldots, m. \tag{2.3}$$

En posant que $\mathbf{y} = (y_1, \ldots, y_m)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)'$ et $\mathbf{D} = \text{diag}(D_1, \ldots, D_m)$, le modèle (2.3) peut être exprimé en notation matricielle sous la forme $\mathbf{y} \sim N\{\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(A)\}$ avec $\boldsymbol{\Sigma}(A) = \mathbf{D} + A\mathbf{I}_m$, où $\mathbf{I}_m$ désigne la matrice identité de dimensions $m \times m$. Si A est connue, le meilleur prédicteur linéaire sans biais (BLUP) par composante de $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ est donné par

$$\tilde{\boldsymbol{\theta}}(A) = \left(\tilde{\theta}_1(A), \ldots, \tilde{\theta}_m(A)\right)' = \mathbf{X}\tilde{\boldsymbol{\beta}}(A) + A\boldsymbol{\Sigma}^{-1}(A)\{\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}(A)\}, \tag{2.4}$$

où

$$\begin{aligned} \tilde{\boldsymbol{\beta}}(A) &= \{\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\,\mathbf{X}\}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\,\mathbf{y} \\ &= \left\{\sum_{i=1}^{m}(A + D_i)^{-1}\mathbf{x}_i\mathbf{x}_i'\right\}^{-1}\sum_{i=1}^{m}(A + D_i)^{-1}\mathbf{x}_i y_i \end{aligned} \tag{2.5}$$

est l'estimateur des moindres carrés pondérés (MCP) de $\boldsymbol{\beta}$. Toutefois, en pratique, A est inconnue. En substituant un estimateur convergent $\hat{A}$ à A dans le BLUP (2.4), nous obtenons l'EBLUP donné par

$$\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_m\right)' = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{A}\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \tag{2.6}$$

où $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{A})$ et $\hat{\boldsymbol{\Sigma}} = \mathbf{D} + \hat{A}\mathbf{I}_m$. Pour le i^{e} domaine, l'EBLUP de θ_i peut être exprimé comme une combinaison linéaire convexe de l'estimateur synthétique de type régression $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ et de l'estimateur direct y_i, sous la forme

$$\hat{\theta}_i = B_i(\hat{A})\,\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \{1 - B_i(\hat{A})\}\,y_i, \tag{2.7}$$

où le poids appliqué à l'estimateur synthétique de type régression $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ est donné par $B_i(\hat{A})$, où $B_i(A) = D_i/(A + D_i)$. Notons que le poids augmente avec la variance d'échantillonnage D_i. Donc, quand l'estimateur direct n'est pas fiable, c'est-à-dire que D_i est grande comparativement à la variance totale $\hat{A} + D_i$, un poids plus important est appliqué à l'estimateur synthétique de type régression $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$. Par ailleurs, si l'estimateur direct est efficace, D_i est petite comparativement à $\hat{A} + D_i$, et un plus grand poids est alors donné à l'estimateur direct y_i.

Plusieurs estimateurs de A ont été proposés dans la littérature, y compris des estimateurs des moments sans hypothèse de normalité, l'estimateur du maximum de vraisemblance (MV) et l'estimateur du maximum de vraisemblance restreint (ou résiduel) (MVRE). L'estimateur MV de A est $\hat{A}_{\text{MV}} = \max\left(0, \hat{A}^*_{\text{MV}}\right)$, où $\hat{A}^*_{\text{MV}}$ peut être obtenu en maximisant la fonction de vraisemblance profil donnée par

$$L_P(A) = c\,|\boldsymbol{\Sigma}(A)|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}(A)\,\mathbf{y}\right\},$$

où c désigne une constante générique et

$$\mathbf{P}(A) = \boldsymbol{\Sigma}^{-1}(A) - \boldsymbol{\Sigma}^{-1}(A)\,\mathbf{X}\{\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\,\mathbf{X}\}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A).$$

L'estimateur MVRE de A est $\hat{A}_{\text{RE}} = \max\left(0, \hat{A}^*_{\text{RE}}\right)$, où $\hat{A}^*_{\text{RE}}$ s'obtient en maximisant la vraisemblance restreinte/résiduelle, donnée par

$$L_{\mathrm{RE}}(A) = c\,|\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\,\mathbf{X}|^{-1/2}\,|\boldsymbol{\Sigma}(A)|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}(A)\,\mathbf{y}\right\}.$$

Dans le présent article, nous nous concentrons sur l'estimateur MVRE $\hat{A}_{\mathrm{RE}}$ qui est fréquemment utilisé en pratique, et nous désignons par $\hat{\boldsymbol{\theta}}_{\mathrm{RE}} = \left(\hat{\theta}_{\mathrm{RE},1},\ldots,\hat{\theta}_{\mathrm{RE},m}\right)'$ l'EBLUP donné en (2.6) obtenu avec $\hat{A} = \hat{A}_{\mathrm{RE}}$.

3 Erreur quadratique moyenne

Notons que le BLUP $\tilde{\theta}_i(A)$ de la moyenne de petit domaine θ_i est une fonction linéaire de $\mathbf{y}$. Donc, son EQM peut être calculée facilement et est donnée par la somme de deux termes :

$$\mathrm{EQM}\left\{\tilde{\theta}_i(A)\right\} = g_{1i}(A) + g_{2i}(A),$$

où $g_{1i}(A)$ est dû à l'estimation de l'effet aléatoire de domaine v_i et $g_{2i}(A)$ est dû à l'estimation du paramètre de régression $\boldsymbol{\beta}$, avec

$$\begin{aligned} g_{1i}(A) &= D_i\{1 - B_i(A)\}, \\ g_{2i}(A) &= B_i^2(A)\,\mathbf{x}_i'\{\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\,\mathbf{X}\}^{-1}\mathbf{x}_i. \end{aligned}$$

Cependant, l'EBLUP $\hat{\theta}_i$ donné en (2.7) n'est pas linéaire en $\mathbf{y}$ en raison de l'estimation de la variance des effets aléatoires A. En utilisant un estimateur des moments de A, Prasad et Rao (1990) ont obtenu une approximation d'ordre deux correcte de l'EQM de l'EBLUP. Plus tard, Datta et Lahiri (2000) et Das, Jiang et Rao (2004) ont obtenu une approximation d'ordre deux correcte de l'EQM sous estimation du MV et du MVRE de A. En utilisant l'estimateur MVRE de A, leur approximation de l'EQM, pour une grande valeur de m, est donnée par

$$\mathrm{EQM}\left(\hat{\theta}_{\mathrm{RE},i}\right) = g_{1i}(A) + g_{2i}(A) + g_{3i}(A) + o(m^{-1}), \tag{3.1}$$

où

$$g_{3i}(A) = B_i^2(A)\frac{V_{\mathrm{RE}}(A)}{A + D_i} \text{ et } V_{\mathrm{RE}}(A) = \frac{2}{\sum_{i=1}^{m}(A + D_i)^{-2}}.$$

Notons que, quand $m \to \infty$, $g_{1i}(A) = O(1)$, $g_{2i}(A) = O(m^{-1})$ et $g_{3i}(A) = O(m^{-1})$, de sorte que $g_{1i}(A)$ est le terme principal dans l'EQM quand m est grand. Cependant, si A est petite, le terme $g_{1i}(A)$ est approximativement nul et $g_{3i}(A)$ pourrait alors devenir le terme principal quand m est petit. Par exemple, en ne prenant qu'une seule covariable $(p = 1)$ avec des valeurs constantes $x_i = 1$ et des variances d'échantillonnage constantes $D_i = D, i = 1,\ldots,m$, et en posant que $A = 0$, nous obtenons $g_{1i}(0) = 0, g_{2i}(0) = D/m$ et $g_{3i}(0) = 2D/m$; autrement dit, $g_{3i}(0)$ est deux fois plus grand que $g_{2i}(0)$.

Datta et Lahiri (2000) ont obtenu un estimateur de l'EQM de l'EBLUP $\hat{\theta}_{\mathrm{RE},i}$ donné par

$$\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right) = g_{1i}\left(\hat{A}_{\text{RE}}\right) + g_{2i}\left(\hat{A}_{\text{RE}}\right) + 2g_{3i}\left(\hat{A}_{\text{RE}}\right). \tag{3.2}$$

L'estimateur de l'EQM (3.2) est sans biais d'ordre deux en ce sens que

$$E\left\{\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)\right\} = \text{EQM}\left(\hat{\theta}_{\text{RE},i}\right) + o\left(m^{-1}\right).$$

Dans le cas où $A = 0$, le BLUP $\tilde{\theta}_{\text{RE},i}$ de θ_i devient l'estimateur synthétique de type régression $\hat{\theta}_{\text{SYN},i} = \mathbf{x}_i'\tilde{\boldsymbol{\beta}}(0)$. Mais étonnamment, l'approximation de l'EQM de l'EBLUP donnée en (3.1) peut être très différente de l'EQM de l'estimateur synthétique. Notons que cette dernière est donnée par

$$\text{EQM}\left(\hat{\theta}_{\text{SYN},i}\right) = g_{2i}(0) < g_{2i}(0) + g_{3i}(0),$$

parce que le terme $g_{3i}(0)$ est strictement positif, même pour $A = 0$. En fait, dans l'exemple simple d'une seule covariable $(p = 1)$ avec valeurs constantes $x_i = 1$ et variances d'échantillonnage constantes $D_i = D, i = 1, \ldots, m$, nous avons $\text{EQM}\left(\hat{\theta}_{\text{SYN},i}\right) = g_{2i}(0) = D/m$, tandis que l'approximation de l'EQM de l'EBLUP donnée en (3.1) avec $A = 0$ donne $\text{EQM}\left(\hat{\theta}_{\text{RE},i}\right) \approx g_{2i}(0) + g_{3i}(0) = 3D/m$, qui est trois fois plus grande. Il se fait que (3.1) n'est pas une bonne approximation de l'EQM de l'EBLUP quand $A = 0$ et, nous devrions plutôt utiliser $\text{EQM}\left(\hat{\theta}_{\text{RE},i}\right) = g_{2i}(0)$. En outre, puisque pour $A = 0$, cette quantité ne dépend d'aucun paramètre inconnu, nous pouvons aussi la prendre comme estimateur de l'EQM, c'est-à-dire que nous pouvons prendre $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right) = g_{2i}(0)$.

En pratique, la vraie valeur de A est inconnue, mais nous avons un estimateur convergent $\hat{A}_{\text{RE}}$. Quand $\hat{A}_{\text{RE}} = 0$, l'EBLUP devient l'estimateur synthétique de type régression pour tous les domaines, c'est-à-dire

$$\hat{\theta}_{\text{RE},i} = \hat{\theta}_{\text{SYN},i} = \mathbf{x}_i'\tilde{\boldsymbol{\beta}}(0), i = 1, \ldots, m.$$

Dans ce cas, $g_{1i}\left(\hat{A}_{\text{RE}}\right) = 0$ pour tous les domaines et l'estimateur de l'EQM donné en (3.2) se réduit à

$$\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right) = g_{2i}(0) + 2g_{3i}(0) > g_{2i}(0) = \text{EQM}\left(\hat{\theta}_{\text{SYN},i}\right), i = 1, \ldots, m.$$

Donc, l'estimateur de l'EQM donné en (3.2) peut gravement surestimer l'EQM pour $\hat{A}_{\text{RE}} = 0$. Afin de réduire la surestimation, nous considérons un estimateur modifié de l'EQM de $\hat{\theta}_{\text{RE},i}$ donné par

$$\text{eqm}_0\left(\hat{\theta}_{\text{RE},i}\right) = \begin{cases} g_{2i} & \text{si } \hat{A}_{\text{RE}} = 0, \\ g_{1i}\left(\hat{A}_{\text{RE}}\right) + g_{2i}\left(\hat{A}_{\text{RE}}\right) + 2g_{3i}\left(\hat{A}_{\text{RE}}\right) & \text{si } \hat{A}_{\text{RE}} > 0, \end{cases} \tag{3.3}$$

où $g_{2i} = g_{2i}(0) = \mathbf{x}_i'\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{x}_i, i = 1, \ldots, m.$

En fait, pour une valeur de A proche de zéro, il se peut que g_{2i} soit plus proche de la vraie EQM que l'estimateur de l'EQM complet $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)$, mais la question qui se pose est celle de savoir quand A est suffisamment proche de zéro. Cette question motive le recours à une procédure de test préliminaire de $A = 0$ pour définir des estimateurs de rechange de l'EQM de l'EBLUP à la section 4.

4 Estimateurs à test préliminaire

L'estimateur de A utilisé dans l'EBLUP de θ_i introduit une incertitude qui pourrait ne pas être négligeable quand m est petit. En effet, dans l'estimateur (3.2) de l'EQM, le terme g_{3i} découle de l'estimation de A. Cependant, quand la valeur de A est suffisamment faible par rapport aux variances d'échantillonnage, cette incertitude pourrait être évitée en utilisant l'estimateur synthétique de type régression $\mathbf{x}_i'\tilde{\boldsymbol{\beta}}(0)$ au lieu de l'EBLUP. Datta et coll. (2011) ont proposé un estimateur pour petits domaines basé sur une procédure de test préliminaire de l'hypothèse $H_0 : A = 0$ contre $H_1 : A > 0$. Si H_0 n'est pas rejetée, l'estimateur synthétique de type régression est utilisé comme estimateur de θ_i; sinon, l'EBLUP habituel est utilisé. Ils ont proposé la statistique de test

$$T = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{TP}}\right)' \mathbf{D}^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{TP}}\right),$$

où $\hat{\boldsymbol{\beta}}_{\mathrm{TP}} = \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{D}^{-1}\mathbf{y}$ est l'estimateur MCP de $\boldsymbol{\beta}$ obtenu en supposant que $H_0 : A = 0$ est vérifiée. La statistique de test T suit une loi X^2_{m-p} avec $m - p$ degrés de liberté sous H_0. Alors, pour un seuil de signification spécifié α, l'estimateur ETP de $\boldsymbol{\theta}$ défini par Datta et coll. (2011) est donné par

$$\hat{\boldsymbol{\theta}}_{\mathrm{TP}} = \left(\hat{\theta}_{\mathrm{TP},1}, \ldots, \hat{\theta}_{\mathrm{TP},m}\right)' = \begin{cases} \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{TP}} & \text{si } T \leq X^2_{m-p,\alpha}; \\ \hat{\boldsymbol{\theta}}_{\mathrm{RE}} & \text{si } T > X^2_{m-p,\alpha}, \end{cases}$$

où $X^2_{m-p,\alpha}$ est la valeur critique supérieure au seuil α de X^2_{m-p}. L'estimateur ETP est conçu spécialement pour traiter les cas où le nombre de petits domaines est modeste, disons $m = 15$.

Ici, nous proposons d'utiliser la procédure TP pour l'estimation de l'EQM de l'EBLUP, en ne considérant que l'EQM de l'estimateur synthétique g_{2i} quand l'hypothèse nulle n'est pas rejetée, et l'estimation complète de l'EQM autrement. Mais soulignons que la statistique de test T dans la procédure TP ne dépend pas de l'estimateur de A. Cela signifie que, même quand H_0 est rejetée, il peut arriver que $\hat{A}_{\mathrm{RE}} = 0$. Donc, ici, nous définissons l'estimateur ETP de l'EQM de l'EBLUP $\hat{\theta}_{\mathrm{RE},i}$ sous la forme

$$\mathrm{eqm}_{\mathrm{TP}}\left(\hat{\theta}_{\mathrm{RE},i}\right) = \begin{cases} g_{2i} & \text{si } T \leq X^2_{m-p,\alpha} \quad \text{ou} \quad \hat{A}_{\mathrm{RE}} = 0, \\ g_{1i}\left(\hat{A}_{\mathrm{RE}}\right) + g_{2i}\left(\hat{A}_{\mathrm{RE}}\right) + 2g_{3i}\left(\hat{A}_{\mathrm{RE}}\right) & \text{si } T > X^2_{m-p,\alpha} \quad \text{et} \quad \hat{A}_{\mathrm{RE}} > 0. \end{cases} \tag{4.1}$$

5 Maximum de vraisemblance ajusté

Les méthodes d'estimation de A décrites à la section 2 pourraient produire des estimations nulles. Le cas échéant, les EBLUP attribueront un poids nul aux estimateurs directs dans tous les domaines, quelle que soit l'efficacité de l'estimateur direct dans chaque domaine. Par ailleurs, les praticiens des sondages préfèrent souvent attribuer systématiquement un poids strictement positif aux estimateurs directs, parce qu'ils sont fondés sur des données au niveau de l'unité propres au domaine pour la variable d'intérêt, sans l'hypothèse d'un modèle de régression. Pour cette situation, Li et Lahiri (2010) ont proposé l'estimateur du maximum de vraisemblance ajusté (MVA) qui donne un estimateur strictement positif de A. Cet estimateur, désigné ici $\hat{A}_{\mathrm{MVA}}$, s'obtient en maximisant la vraisemblance ajustée définie par

$$L_{\text{MVA}}(A) = A \times L_P(A).$$

L'EBLUP donné en (2.6) avec $\hat{A} = \hat{A}_{\text{MVA}}$ sera noté ci-après sous la forme $\hat{\boldsymbol{\theta}}_{\text{MVA}} = \left(\hat{\theta}_{\text{MVA},1}, \ldots, \hat{\theta}_{\text{MVA},m}\right)'$. Notons que $\hat{\boldsymbol{\theta}}_{\text{MVA}}$ attribue des poids strictement positifs aux estimateurs directs.

Li et Lahiri (2010) ont proposé un estimateur sans biais d'ordre deux de l'EQM de $\hat{\theta}_{\text{MVA},i}$ donné par

$$\begin{aligned} \text{eqm}\left(\hat{\theta}_{\text{MVA},i}\right) &= g_{1i}\left(\hat{A}_{\text{MVA}}\right) + g_{2i}\left(\hat{A}_{\text{MVA}}\right) + 2g_{3i}\left(\hat{A}_{\text{MVA}}\right) \\ &- B_i^2\left(\hat{A}_{\text{MVA}}\right) b_{\text{MVA}}\left(\hat{A}_{\text{MVA}}\right), \end{aligned} \tag{5.1}$$

où $b_{\text{MVA}}(A)$ est le biais de $\hat{A}_{\text{MVA}}$ qui est donné par

$$b_{\text{MVA}}(A) = \frac{\text{trace}\left\{\mathbf{P}(A) - \boldsymbol{\Sigma}^{-1}(A)\right\} + 2/A}{\text{trace}\left\{\boldsymbol{\Sigma}^{-2}(A)\right\}}.$$

6 Estimateurs combinés

L'estimateur MVA strictement positif de A présente habituellement un plus grand biais que les estimateurs MV ou MVRE quand A est relativement petite par rapport aux D_i. Donc, si nous voulons encore obtenir un estimateur pour petits domaines qui applique un poids strictement positif à l'estimateur direct, afin de réduire le biais susmentionné, il sera préférable de n'utiliser l'estimateur MVA que quand cela est strictement nécessaire; c'est-à-dire, quand les données ne fournissent pas suffisamment de preuves que l'égalité $A = 0$ n'est pas vraie ou que l'estimateur MVRE résultant de A est nul. Nous présentons ici deux estimateurs pour petits domaines de $\boldsymbol{\theta}$ donnant un poids strictement positif à l'estimateur direct, qui ont été obtenus sous forme d'une combinaison de l'EBLUP basé sur la méthode du MVA et de l'EBLUP basé sur l'estimation du MVRE.

Dans la première combinaison proposée, la méthode du MVA est utilisée pour estimer A quand le test préliminaire ne donne pas lieu au rejet de l'hypothèse nulle et dans la deuxième combinaison proposée, elle est utilisée quand l'estimation du MVRE n'est pas positive. Plus précisément, le premier estimateur combiné, appelé ci-après TP-MVA, est défini par

$$\hat{\boldsymbol{\theta}}_{\text{TPMVA}} = \begin{cases} \hat{\boldsymbol{\theta}}_{\text{MVA}} & \text{si } T \le X^2_{m-p,\alpha} \quad \text{ou} \quad \hat{A}_{\text{RE}} = 0, \\ \hat{\boldsymbol{\theta}}_{\text{RE}} & \text{si } T > X^2_{m-p,\alpha} \quad \text{et} \quad \hat{A}_{\text{RE}} > 0. \end{cases} \tag{6.1}$$

Le deuxième estimateur combiné, appelé MVRE-MVA, est donné par

$$\hat{\boldsymbol{\theta}}_{\text{REMVA}} = \begin{cases} \hat{\boldsymbol{\theta}}_{\text{MVA}} & \text{si } \hat{A}_{\text{RE}} = 0, \\ \hat{\boldsymbol{\theta}}_{\text{RE}} & \text{si } \hat{A}_{\text{RE}} > 0, \end{cases} \tag{6.2}$$

voir Rubin-Bleuer et Yu (2013). Pour l'estimation de l'EQM de $\hat{\boldsymbol{\theta}}_{\text{REMVA}}$, ces auteurs ont proposé

$$\text{eqm}\left(\hat{\theta}_{\text{REMVA},i}\right) = \begin{cases} \text{eqm}\left(\hat{\theta}_{\text{MVA},i}\right) & \text{si } \hat{A}_{\text{RE}} = 0, \\ \text{eqm}\left(\hat{\theta}_{\text{RE},i}\right) & \text{si } \hat{A}_{\text{RE}} > 0. \end{cases} \tag{6.3}$$

L'utilisation de $\text{eqm}\left(\hat{\theta}_{\text{MVA},i}\right)$ quand $\hat{A}_{\text{RE}} = 0$ donne lieu à une surestimation importante si la valeur vraie de A est faible, parce que $\hat{\theta}_{\text{MVA},i}$ sera plus proche de l'estimateur synthétique de type régression. Donc, nous proposons l'estimateur de l'EQM de rechange

$$\text{eqm}_0\left(\hat{\theta}_{\text{REMVA},i}\right) = \begin{cases} g_{2i} & \text{si } \hat{A}_{\text{RE}} = 0, \\ \text{eqm}\left(\hat{\theta}_{\text{RE},i}\right) & \text{si } \hat{A}_{\text{RE}} > 0. \end{cases} \tag{6.4}$$

De nouveau, puisque, quand la variance A est petite, $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)$ pourrait encore surestimer la vraie valeur de l'EQM de $\hat{\theta}_{\text{REMVA},i}$, nous considérons également l'estimateur ETP suivant

$$\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{REMVA},i}\right) = \begin{cases} g_{2i} & \text{si } T \leq X^2_{m-p,\alpha} \quad \text{ou} \quad \hat{A}_{\text{RE}} = 0, \\ \text{eqm}\left(\hat{\theta}_{\text{RE},i}\right) & \text{si } T > X^2_{m-p,\alpha} \quad \text{et} \quad \hat{A}_{\text{RE}} > 0. \end{cases} \tag{6.5}$$

7 Expériences de simulation

Une étude par simulation a été conçue en vue de répondre aux objectifs suivants :

a) Étudier les propriétés, en termes de biais et d'EQM, des estimateurs ETP quand α varie pour une valeur fixe de A, et quand A varie pour une valeur fixe de α. Nous souhaitons déterminer quelles valeurs de α sont adéquates pour une valeur donnée de A.

b) Comparer les estimateurs ETP aux EBLUP basés sur le MVRE et aux EBLUP basés sur le MVA.

c) Étudier les propriétés des estimateurs proposés de l'EQM en ce qui concerne le biais relatif, ainsi que la couverture et la longueur des intervalles de prédiction.

d) Comparer les trois estimateurs pour petits domaines présentés qui attribuent un poids strictement positif à l'estimateur direct pour tous les domaines, à savoir l'EBLUP fondé sur les estimateurs MVA, TP-MVA et MVRE-MVA.

Pour réaliser les objectifs susmentionnés, nous avons généré des données à partir du modèle de Fay-Herriot donné par les équations (2.1) et (2.2) avec une moyenne constante, c'est-à-dire avec $p = 1, \boldsymbol{\beta} = \mu$ et $\mathbf{x}_i = 1, i = 1, \ldots, m$. Nous posons que $\mu = 0$ sans perte de généralité, que le nombre de domaines est $m = 15$ et que $D_i = 1, i = 1, \ldots, m$. L'étude par simulation a été répétée pour des valeurs croissantes de la variance du modèle, $A \in \{0{,}01;\ 0{,}02;\ 0{,}05;\ 0{,}1;\ 0{,}2;\ 1\}$, ainsi que pour six seuils de signification du test de $H_0 : A = 0$ contre $H_0 : A > 0$, à savoir $\alpha = \{0{,}05;\ 0{,}1;\ 0{,}2;\ 0{,}3;\ 0{,}4;\ 0{,}5\}$. Pour chaque combinaison de A et α, nous avons procédé aux étapes qui suivent pour chaque exécution de la simulation $\ell = 1, \ldots, L$ avec $L = 10\,000$ exécutions :

1. Générer les données au moyen du modèle hypothétique de moyenne nulle constante; c'est-à-dire

$$\begin{aligned}\theta_i^{(\ell)} &= v_i^{(\ell)}, \quad v_i^{(\ell)} \overset{\text{ind}}{\sim} N(0, A),\\ y_i^{(\ell)} &= \theta_i^{(\ell)} + e_i^{(\ell)}, \quad e_i^{(\ell)} \overset{\text{ind}}{\sim} N(0, D_i), \quad i = 1, \ldots, m.\end{aligned}$$

2. Calculer les estimateurs suivants de $\boldsymbol{\theta}$: l'EBLUP basé sur l'estimation du MVRE de $A, \hat{\boldsymbol{\theta}}_{\text{RE}}^{(\ell)}$, l'estimation ETP, $\hat{\boldsymbol{\theta}}_{\text{TP}}^{(\ell)}$, l'EBLUP basé sur l'estimation du MVA de $A, \hat{\boldsymbol{\theta}}_{\text{MVA}}^{(\ell)}$, l'estimation combinée TP-MVA $\hat{\boldsymbol{\theta}}_{\text{TPMVA}}^{(\ell)}$ et l'estimation MVRE-MVA $\hat{\boldsymbol{\theta}}_{\text{REMVA}}^{(\ell)}$.
3. Pour chaque domaine $i = 1, \ldots, m,$ calculer : les trois estimations de l'EQM de l'EBLUP $\hat{\theta}_{\text{RE},i}$ données dans (3.2), (3.3) et (4.1), désignées respectivement par $\text{eqm}^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right), \text{eqm}_0^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)$ et $\text{eqm}_{\text{TP}}^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)$, et les trois estimations (6.3), (6.4) et (6.5) de l'EQM de l'estimateur combiné pour petits domaines $\hat{\theta}_{\text{REMVA},i}$, désignées $\text{eqm}^{(\ell)}\left(\hat{\theta}_{\text{REMVA},i}\right), \text{eqm}_0^{(\ell)}\left(\hat{\theta}_{\text{REMVA},i}\right)$ et $\text{eqm}_{\text{TP}}^{(\ell)}\left(\hat{\theta}_{\text{REMVA},i}\right)$, respectivement.
4. Pour chaque domaine $i = 1, \ldots, m,$ obtenir les intervalles de prédiction $1 - \alpha$ fondés sur l'hypothèse de normalité pour la moyenne de petit domaine θ_i basée sur les trois estimateurs considérés de l'EQM de l'EBLUP :

$$\begin{aligned}\text{IC}_i^{(\ell)} &= \hat{\theta}_{\text{RE},i}^{(\ell)} \mp Z_{\alpha/2}\sqrt{\text{eqm}^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)},\\ \text{IC}_{0,i}^{(\ell)} &= \hat{\theta}_{\text{RE},i}^{(\ell)} \mp Z_{\alpha/2}\sqrt{\text{eqm}_0^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)},\\ \text{IC}_{\text{TP},i}^{(\ell)} &= \hat{\theta}_{\text{RE},i}^{(\ell)} \mp Z_{\alpha/2}\sqrt{\text{eqm}_{\text{TP}}^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)},\end{aligned}$$

   où $Z_{\alpha/2}$ est la valeur critique supérieure au seuil $\alpha/2$ d'une loi normale centrée réduite.
5. Répéter les étapes 1 à 4 pour $\ell = 1, \ldots, L,$ pour $L = 10\,000.$ Puis, pour chaque estimateur pour petits domaines $\hat{\theta}_i \in \left\{\hat{\theta}_{\text{RE},i}, \hat{\theta}_{\text{TP},i}, \hat{\theta}_{\text{MVA},i}, \hat{\theta}_{\text{TPMVA},i}, \hat{\theta}_{\text{REMVA},i}\right\}, i = 1, \ldots, m,$ calculer le biais et l'EQM empiriques sous la forme

$$B\left(\hat{\theta}_i\right) = \frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{\theta}_i^{(\ell)} - \theta_i^{(\ell)}\right), \quad \text{EQM}\left(\hat{\theta}_i\right) = \frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{\theta}_i^{(\ell)} - \theta_i^{(\ell)}\right)^2.$$

   Obtenir ensuite la moyenne sur les domaines des biais et des EQM absolus sous la forme

$$\overline{\text{BA}}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{m}\sum_{i=1}^{m}\left|B\left(\hat{\theta}_i\right)\right|, \quad \overline{\text{EQMA}}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{m}\sum_{i=1}^{m}\text{EQM}\left(\hat{\theta}_i\right).$$

6. Calculer le biais relatif de chaque estimateur de l'EQM, $\text{eqm}\left(\hat{\theta}_i\right)$, comme il suit

$$\text{BR}\left\{\text{eqm}\left(\hat{\theta}_i\right)\right\} = \left\{\frac{1}{L}\sum_{\ell=1}^{L}\text{eqm}^{(\ell)}\left(\hat{\theta}_i\right) - \text{EQM}\left(\hat{\theta}_i\right)\right\} \Big/ \text{EQM}\left(\hat{\theta}_i\right).$$

Calculer la moyenne sur les domaines des biais relatifs absolus sous la forme

$$\overline{\text{BRA}}\left\{\text{eqm}\left(\hat{\boldsymbol{\theta}}\right)\right\} = \frac{1}{m}\sum_{i=1}^{m}\left|\text{BR}\left\{\text{eqm}\left(\hat{\theta}_i\right)\right\}\right|.$$

7. Pour chaque type d'intervalle de prédiction $\text{IC}_i^{(\ell)} = \left(L_i^{(\ell)}, U_i^{(\ell)}\right)$, pour $\text{IC}_i^{(\ell)} \in \left\{\text{IC}_i^{(\ell)}, \text{IC}_{0,i}^{(\ell)}, \text{IC}_{\text{TP},i}^{(\ell)}\right\}$ donné à l'étape 4, calculer le taux de couverture (TC) et la longueur moyenne (LM) empiriques comme il suit

$$\text{TC}(\text{IC}_i) = \frac{\#\left\{\theta_i^{(\ell)} \in \text{IC}_i^{(\ell)}\right\}}{L}, \quad \text{LM}\left(\text{IC}_i\right) = \frac{1}{L}\sum_{\ell=1}^{L}\left(U_i^{(\ell)} - L_i^{(\ell)}\right).$$

Enfin, calculer la moyenne sur les domaines des taux de couverture et des longueurs moyennes, comme il suit

$$\overline{\text{TC}}\left(\text{IC}\right) = \frac{1}{m}\sum_{i=1}^{m}\text{TC}\left(\text{IC}_i\right), \quad \overline{\text{LM}}\left(\text{IC}\right) = \frac{1}{m}\sum_{i=1}^{m}\text{LM}\left(\text{IC}_i\right).$$

Les figures 7.1 et 7.2 représentent graphiquement les EQM moyennes des estimateurs ETP pour chaque valeur de $A \in \{0{,}05; 0{,}1; 0{,}2\}$, ainsi que l'EQM moyenne des EBLUP basés sur le MVRE et le MVA en fonction du seuil de signification α. Notons que, quand la valeur de A est petite, pour une grande valeur de α, la procédure TP donne lieu plus souvent au rejet de H_0 et par conséquent l'estimateur ETP devient plus fréquemment l'EBLUP usuel, tandis que si la valeur de α est faible, la procédure TP donne lieu moins fréquemment au rejet de H_0 et l'estimateur synthétique de type régression est alors utilisé plus souvent. Par contre, pour une grande valeur de A, l'estimateur ETP devient plus fréquemment l'EBLUP quelle que soit la valeur de α. Les biais absolus des estimateurs ne sont pas présentés ici, parce qu'ils sont à peu près les mêmes pour tous les estimateurs ETP pour les différentes valeurs de α. Il en est ainsi parce que, quand le modèle est vérifié, les deux composantes de l'estimateur ETP, l'estimateur synthétique et l'EBLUP, sont sans biais pour le paramètre étudié. Notons que l'estimateur synthétique est sans biais même quand $A > 0$. La première conclusion qui se dégage des figures 7.1 et 7.2 est que l'EQM de l'estimateur ETP est pratiquement constante pour les diverses valeurs de $\alpha \geq 0{,}1$. Nous voyons aussi que l'EQM moyenne de l'estimateur ETP pour une valeur donnée de α augmente avec A, parce que l'estimateur ETP se réduit plus fréquemment à l'EBLUP quand A augmente et que l'EQM de l'EBLUP augmente avec A. Observons aussi que l'estimateur ETP et l'EBLUP basé sur le MVRE donnent des résultats très similaires pour $\alpha \geq 0{,}2$. Cependant, pour $\alpha < 0{,}2$, l'estimateur ETP devient plus efficace que l'EBLUP aussitôt que A s'approche de l'hypothèse nulle $(A < 0{,}1)$, ce qui concorde avec la remarque de Datta et coll. (2011).

Pour l'EBLUP basé sur le MVA, les figures 7.1 et 7.2 montrent que l'EQM moyenne est considérablement plus grande que celles des deux autres estimateurs, mais que les écarts par rapport aux autres diminuent à mesure que A augmente. Cette situation est attribuable au biais de l'estimateur MVA de A quand la valeur de A est petite. Nous étudierons plus loin les estimateurs pour petits domaines combinés TP-MVA et MVRE-MVA, qui n'utilisent l'EBLUP basé sur le MVA que si l'hypothèse nulle n'est pas rejetée ou que l'estimation réalisée de A est nulle.

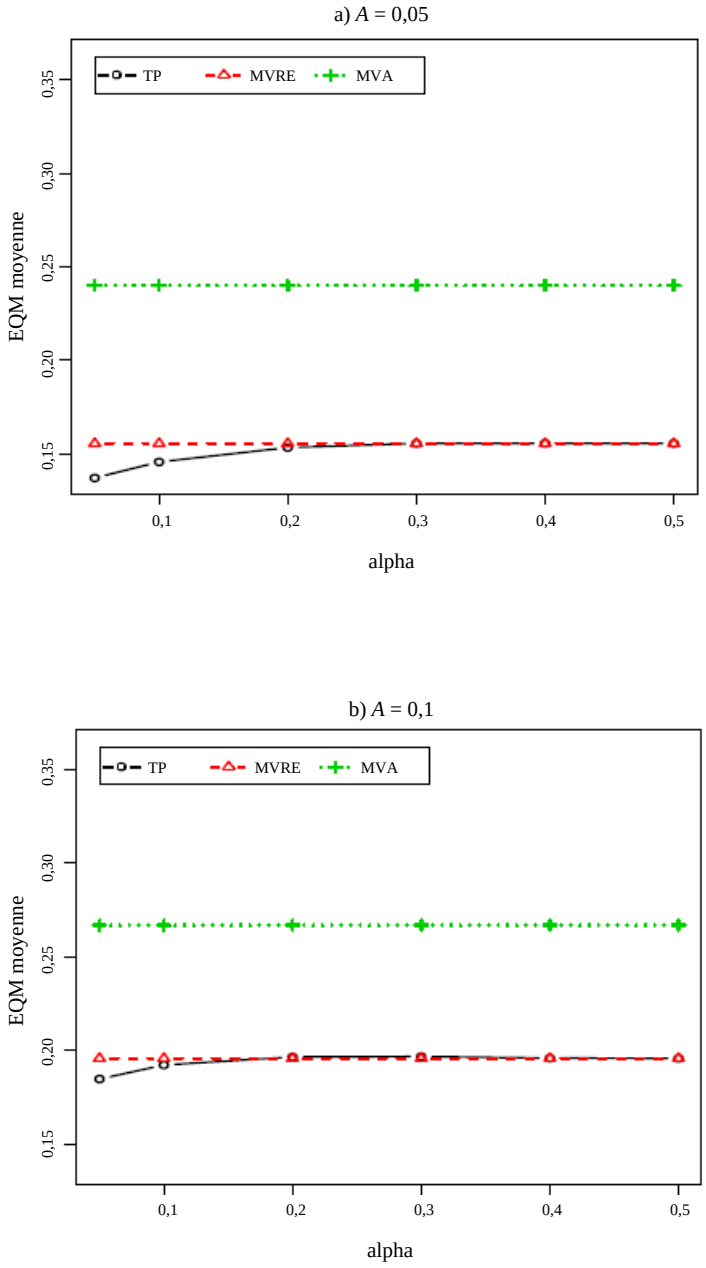


Figure 7.1 EQM moyennes de l'ETP, de l'EBLUP basé sur le MVRE et de l'EBLUP basé sur le MVA en fonction de α, pour a) $A = 0{,}05$ et b) $A = 0{,}1$.

Datta et coll. (2011, page 366) ont recommandé d'utiliser $\alpha \geq 0{,}2$ pour l'ETP. En outre, selon la littérature sur l'estimation TP pour les modèles à effets fixes, un bon choix de α en ce qui concerne le biais et l'EQM est $\alpha = 0{,}2$ (Bancroft 1944; Han et Bancroft 1968). Cependant, les résultats susmentionnés donnent à penser que, pour $\alpha \geq 0{,}2$, l'estimateur ETP est pratiquement le même que l'EBLUP et qu'on pourrait par conséquent choisir de toujours utiliser l'EBLUP.

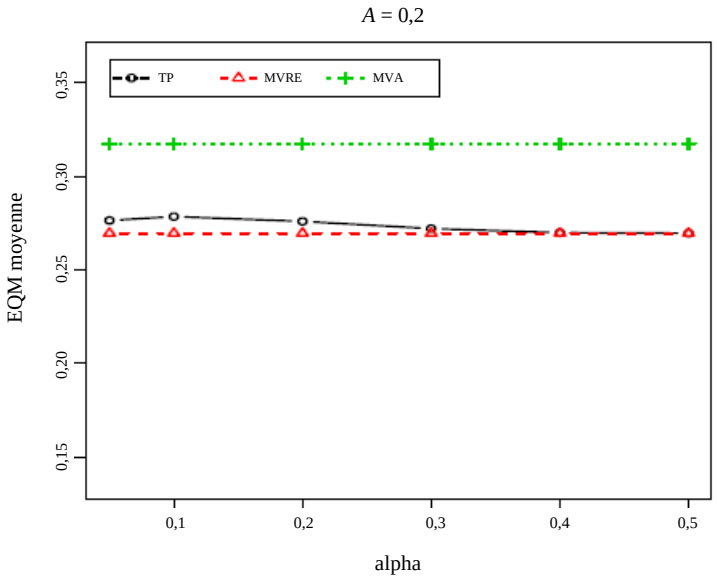


Figure 7.2 EQM moyennes de l'ETP, de l'EBLUP basé sur le MVRE et de l'EBLUP basé sur le MVA en fonction de α, pour $A = 0{,}2$.

Nous allons maintenant étudier les propriétés de l'estimateur ETP pour l'estimation de l'EQM en fonction de α. La figure 7.3 représente graphiquement le biais relatif absolu moyen des estimateurs de l'EQM $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ étiqueté TP en fonction du seuil de signification α pour chaque valeur $A \in \{0{,}05;\ 0{,}1;\ 0{,}2;\ 1\}$. Lorsque l'on choisit α très petit $\alpha < 0{,}1$, l'hypothèse nulle $H_0 : A = 0$ est rejetée moins fréquemment et $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ devient souvent égal à g_{2i}, ce qui entraîne une sous-estimation. Pour une grande valeur de α $(\alpha > 0{,}2)$, l'hypothèse nulle est rejetée plus fréquemment et $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ devient l'estimateur usuel de l'EQM de l'EBLUP, qui surestime fortement la valeur de l'EQM quand A est petite. La valeur $\alpha = 0{,}2$ semble être un bon compromis, avec un biais relatif absolu moyen de l'ordre de 10 % pour $A \geq 0{,}1$ et de 20 % pour $A = 0{,}05$.

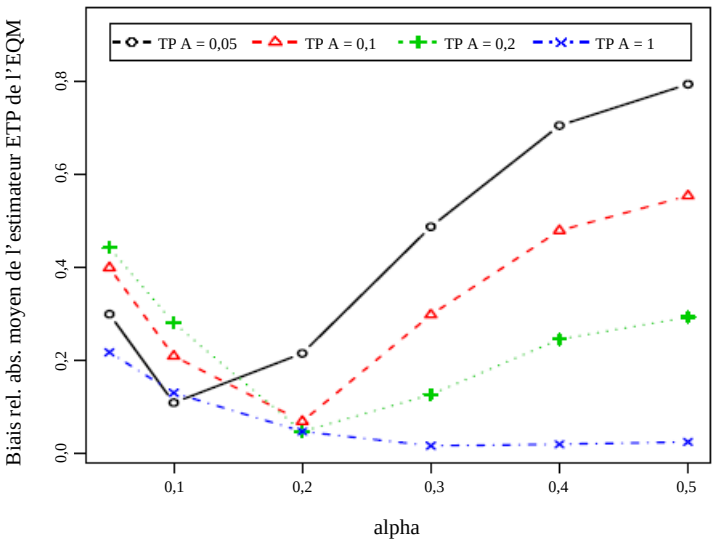


Figure 7.3 Moyenne sur les domaines des biais relatifs absolus de l'estimateur de l'EQM $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$, étiqueté TP, pour $A \in \{0{,}05;\ 0{,}1;\ 0{,}2;\ 1\}$ en fonction du seuil de signification α.

Les résultats susmentionnés donnent à penser que $\alpha = 0{,}2$ est un bon choix lorsqu'on utilise la procédure TP pour estimer l'EQM de l'EBLUP usuel. Cette constatation a été étudiée de manière plus approfondie en examinant les biais relatifs (affectés d'un signe) de $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ pour chaque domaine. Ces résultats sont représentés graphiquement aux figures 7.4 et 7.5, avec quatre graphiques, un pour chaque valeur de $A \in \{0{,}05;\ 0{,}1;\ 0{,}2;\ 1\}$. Les chiffres qui figurent dans les légendes de ces graphiques sont les seuils de signification α pour l'estimateur ETP de l'EQM $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$. Ces graphiques confirment nos observations antérieures, à savoir que l'estimateur de l'EQM fondé sur l'ETP, $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$, sous-estime $\text{EQM}\left(\hat{\theta}_{\text{RE},i}\right)$ pour les faibles valeurs de α et la surestime pour les grandes valeurs de α. Il s'avère que $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ avec $\alpha = 0{,}2$ convient bien pour toutes les valeurs de A.

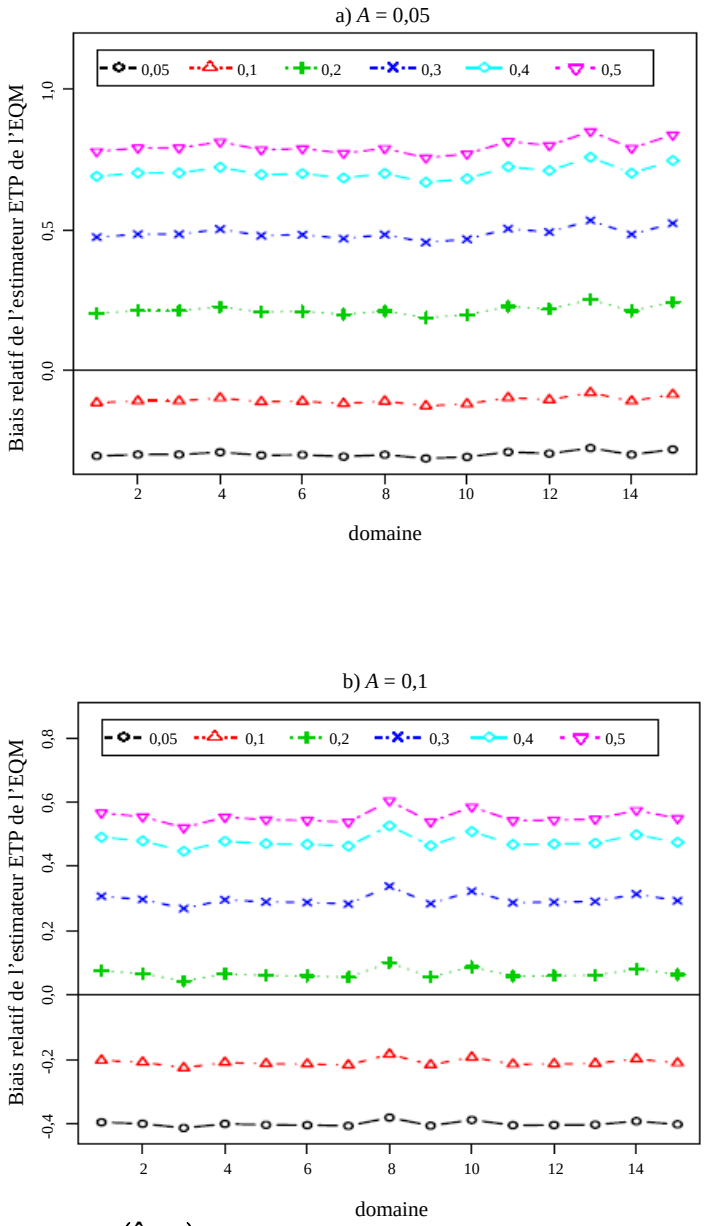


Figure 7.4 Biais relatif de $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ pour chaque seuil de signification $\alpha \in \{0{,}05;\ 0{,}1;\ 0{,}2;\ 0{,}3;\ 0{,}4;\ 0{,}5\}$ en fonction du domaine i, pour a) $A = 0{,}05$ et b) $A = 0{,}1$.

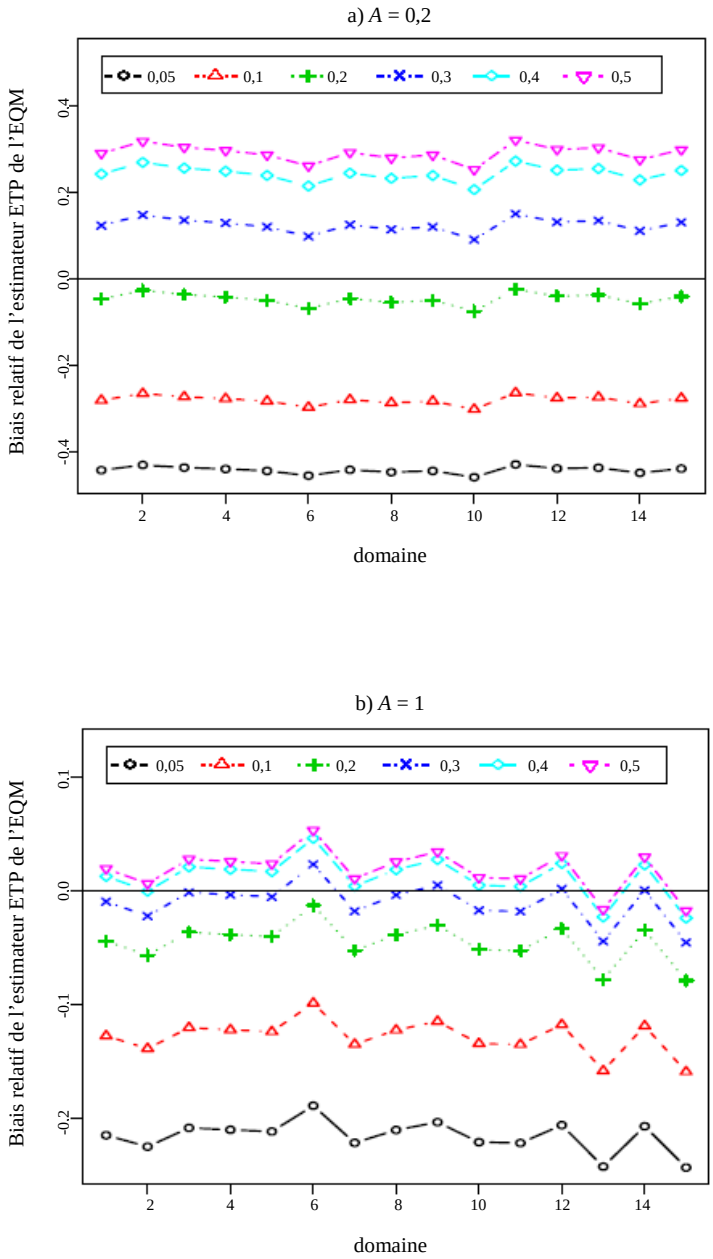


Figure 7.5 Biais relatif de $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ pour chaque seuil de signification $\alpha \in \{0{,}05;\ 0{,}1;\ 0{,}2;\ 0{,}3;\ 0{,}4;\ 0{,}5\}$ en fonction du domaine i, pour a) $A = 0{,}2$ et b) $A = 1$.

Comparons maintenant $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ pour le seuil de signification choisi de $\alpha = 0{,}2$ aux deux autres estimateurs de l'EQM, $\text{eqm}_0\left(\hat{\theta}_{\text{RE},i}\right)$ et $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)$, donnés par (3.3) et (3.2), respectivement. La figure 7.6 représente graphiquement les biais relatifs absolus moyens des trois estimateurs de l'EQM, étiquetés respectivement TP, MVRE0 et MVRE. Nous constatons que $\text{eqm}_0\left(\hat{\theta}_{\text{RE},i}\right)$ donne de meilleurs résultats que $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)$ pour tous les domaines, mais que $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ demeure meilleur que $\text{eqm}_0\left(\hat{\theta}_{\text{RE},i}\right)$ pour toutes les valeurs considérées de A sauf $A = 1$, valeur pour laquelle les différences entre les trois estimateurs sont négligeables. Les écarts diminuent à mesure que A augmente, mais

soulignons que l'estimateur de l'EQM usuel, $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)$, peut être sévèrement biaisé si la valeur de A est petite, avec un biais relatif absolu moyen supérieur à 50 % pour $A < 0{,}2$ et croissant exponentiellement quand A tend vers zéro. La conclusion est que, quand H_0 n'est pas rejetée, même si l'estimation réalisée de A est positive, il semble préférable d'omettre le terme g_{3i} dans l'estimateur de l'EQM et de ne considérer que g_{2i}.

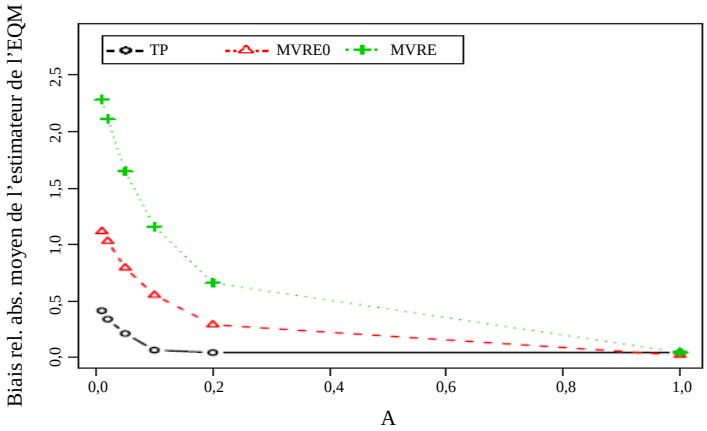


Figure 7.6 Moyenne sur les domaines des biais relatifs absolus des estimateurs de l'EQM $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ avec $\alpha = 0{,}2$, étiqueté TP, $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)$ étiqueté MVRE et $\text{eqm}_0\left(\hat{\theta}_{\text{RE},i}\right)$ étiqueté MVRE0, en fonction de A.

Examinons maintenant les estimateurs pour petits domaines qui appliquent un poids strictement positif à l'estimateur direct pour tous les domaines, à savoir l'EBLUP basé sur le MVA, $\hat{\boldsymbol{\theta}}_{\text{MVA}}$, et les deux estimateurs combinés, TP-MVA donné en (6.1) et MVRE-MVA donné en (6.2). Les EQM moyennes sont représentées graphiquement à la figure 7.7 pour ces trois estimateurs. Dans ce graphique, $\hat{\boldsymbol{\theta}}_{\text{MVA}}$ semble être un peu moins efficace, et est suivi par TP-MVA. L'estimateur combiné MVRE-MVA semble donner d'un peu meilleurs résultats que les deux autres pour une faible valeur de A, quoique pour $A \geq 0{,}2$, l'estimateur TP-MVA est très proche. Pour l'estimation de l'EQM, nous nous concentrons sur l'estimateur MVRE-MVA en raison de sa meilleure performance.

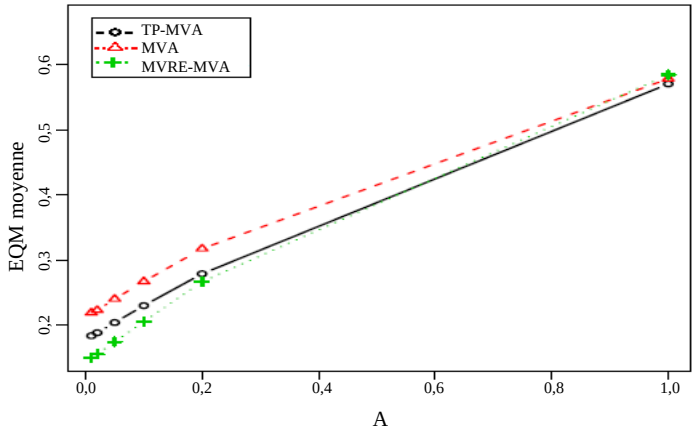


Figure 7.7 Moyenne sur les domaines des EQM pour l'estimateur TP-MVA avec $\alpha = 0{,}2$, l'EBLUP basé sur le MVA et l'estimateur MVRE-MVA en fonction de A.

Pour l'estimateur combiné MVRE-MVA, la figure 7.8 montre que l'estimateur de l'EQM basé sur le test préliminaire TP, $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{REMVA},i}\right)$ qui utilise seulement g_{2i} quand $\hat{A}_{\text{RE}} = 0$ ou que l'hypothèse nulle n'est pas rejetée, présente un biais relatif absolu moyen inférieur à 10 % pour $A \geq 0{,}1$ et est plus faible que les valeurs correspondantes pour $\text{eqm}\left(\hat{\theta}_{\text{REMVA},i}\right)$ et $\text{eqm}_0\left(\hat{\theta}_{\text{REMVA},i}\right)$, spécialement pour $A \leq 0{,}4$.

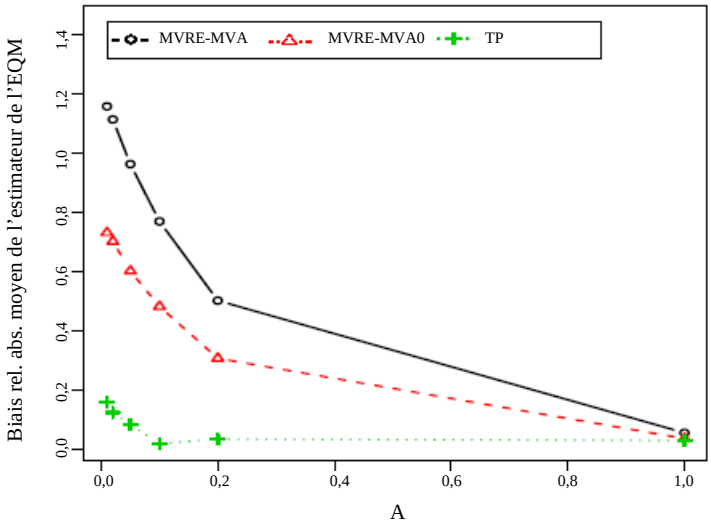


Figure 7.8 Moyenne sur les domaines des biais relatifs absolus des estimateurs de l'EQM $\text{eqm}\left(\hat{\theta}_{\text{REMVA},i}\right)$, $\text{eqm}_0\left(\hat{\theta}_{\text{REMVA},i}\right)$ et $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{REMVA},i}\right)$, étiquetés respectivement MVRE-MVA, MVRE-MVA0 et TP, en fonction de *A*.

Enfin, nous analysons la moyenne sur les domaines des taux de couverture et des longueurs moyennes des intervalles de prédiction fondés sur l'hypothèse de normalité pour la moyenne de petit domaine θ_i en utilisant l'EBLUP basé sur le MVRE comme estimation ponctuelle et les trois estimateurs différents de l'EQM de l'EBLUP, à savoir $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right), \text{eqm}_0\left(\hat{\theta}_{\text{RE},i}\right)$ et $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$. La figure 7.9 représente les taux de couverture des trois types d'intervalles, où les estimateurs de l'EQM basés sur la procédure TP ont été obtenus en prenant $\alpha = 0{,}2; 0{,}3$. Il semble que les bonnes propriétés de biais relatif de l'estimateur de l'EQM basé sur la procédure TP, $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$, pour une valeur faible de A ne peuvent pas être extrapolées à la couverture basée sur les intervalles de prédiction normaux, et présentent une sous-couverture surtout pour $A = 0{,}2$. Dans ce cas, choisir un seuil de signification plus élevé, $\alpha = 0{,}3$, réduit un peu la couverture insuffisante des intervalles de prédiction obtenus en utilisant $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$. Néanmoins, les taux de couverture de $\text{eqm}_0\left(\hat{\theta}_{\text{RE},i}\right)$ sont meilleurs pour toutes les valeurs de A. Comme prévu, l'estimateur usuel de l'EQM $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)$ donne une surcouverture pour les petites valeurs de A, laquelle résulte de la forte surestimation de l'EQM. Par ailleurs, les intervalles pour lesquels on observe une sous-couverture entraînent aussi des intervalles de prédiction plus courts, comme le montre la figure 7.10.

Il est utile de mentionner que la construction des intervalles de prédiction pour θ_i basés sur le modèle de Fay-Herriot avec des taux de couverture exacts n'est pas une tâche évidente. Plusieurs articles traitant de ce problème ont été publiés. Par exemple, Chatterjee, Lahiri et Li (2008) ont proposé des intervalles de

prédiction avec taux de couverture corrects jusqu'à l'ordre deux en utilisant uniquement le terme g_{1i} comme estimation de l'EQM et en appliquant une procédure bootstrap pour trouver les quantiles calés. Diao, Smith, Datta, Maiti et Opsomer (2014) ont obtenu récemment des intervalles de prédiction avec taux de couverture corrects jusqu'à l'ordre deux en évitant d'utiliser des procédures de rééchantillonnage et en utilisant l'estimateur complet de l'EQM. L'obtention d'intervalles de prédiction dont la couverture est exacte en utilisant d'autres estimations de l'EQM pose encore des difficultés et dépasse le cadre du présent article.

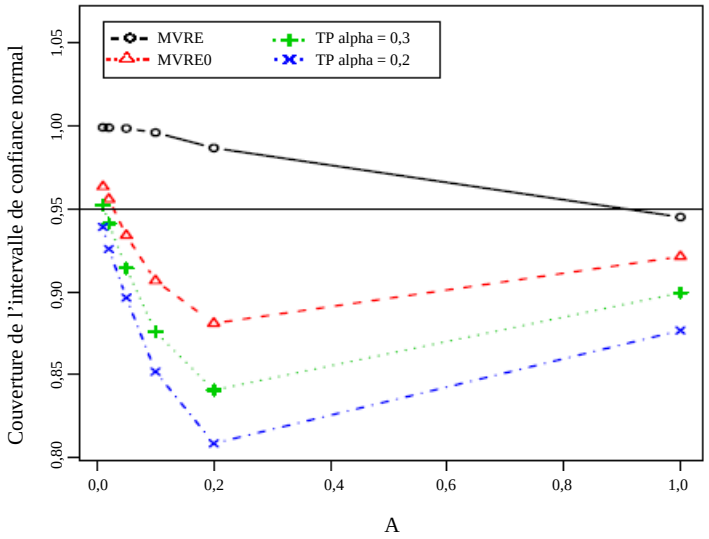


Figure 7.9 Moyenne sur les domaines des taux de couverture des intervalles de prédiction fondés sur la normalité pour θ_i en utilisant les estimateurs de l'EQM $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)$, $\text{eqm}_0\left(\hat{\theta}_{\text{RE},i}\right)$ et $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ avec α = 0,2; 0,3, étiquetés respectivement MVRE, MVRE0 et TP, en fonction de *A*.

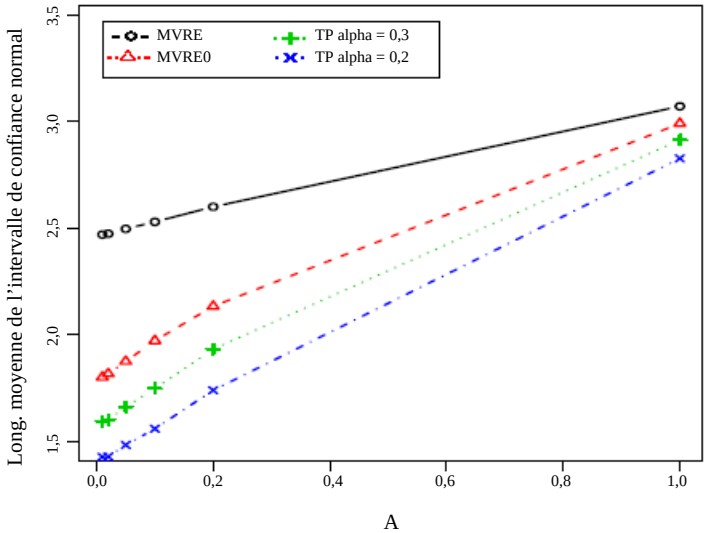


Figure 7.10 Moyenne sur les domaines des longueurs moyennes des intervalles basés sur l'hypothèse de normalité pour θ_i en utilisant les estimateurs de l'EQM $\text{eqm}\left(\hat{\theta}_{\text{RE},i}\right)$, $\text{eqm}_0\left(\hat{\theta}_{\text{RE},i}\right)$ et $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$ avec α = 0,2; 0,3, étiquetés respectivement MVRE, MVRE0 et TP, en fonction de *A*.

L'étude par simulation dont la description précède a été répétée pour plusieurs profils de variances d'échantillonnage inégales D_i. Les résultats ne sont pas présentés ici, mais les conclusions sont très semblables à condition que le profil de variance ne soit pas extrêmement irrégulier.

8 Conclusion

Les principales conclusions qui se dégagent des résultats de notre étude par simulation de l'estimation des moyennes de petit domaine basée sur le modèle de Fay-Herriot au niveau du domaine quand le nombre de domaines est modéré (disons $m = 15$) sont les suivantes : 1) Sous le modèle de Fay-Herriot avec une valeur de la variance des effets aléatoires, A, clairement différente de zéro, l'estimateur ETP ne semble pas améliorer appréciablement l'efficacité relative de l'EBLUP usuel, à moins que le seuil de signification soit faible ($\alpha \leq 0{,}1$ dans notre étude par simulation). 2) Nos résultats de simulation indiquent que l'utilisation de la procédure TP avec une valeur modérée de α, en particulier $\alpha = 0{,}2$, pour estimer l'EQM de l'EBLUP usuel donne lieu à une réduction du biais comparativement à l'estimateur de l'EQM usuel. D'où, nous recommandons d'utiliser $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{RE},i}\right)$, donné par (4.1), pour estimer l'EQM de l'EBLUP. 3) Parmi les estimateurs qui appliquent un poids strictement positif à l'estimateur direct pour tous les domaines, nous recommandons l'estimateur combiné MVRE-MVA donné par (6.2), parce que son efficacité est un peu plus élevée que celle de l'EBLUP basé sur le MVA et le TP-MVA donné par (6.1). 4) Pour estimer l'EQM de l'estimateur MVRE-MVA recommandé, l'estimateur $\text{eqm}_{\text{TP}}\left(\hat{\theta}_{\text{REMVA},i}\right)$ donné par (6.5) produit de meilleurs résultats que les estimateurs de rechange. 5) Nos résultats concernant les intervalles de prédiction, basés sur la théorie de la normalité, indiquent que la bonne performance des estimateurs proposés de l'EQM pourrait ne pas se traduire en bonnes propriétés de couverture de ces intervalles. La construction d'intervalles de prédiction donnant une couverture exacte en utilisant les estimations proposées de l'EQM semble être une tâche difficile.

Des options lisses des estimations avec test préliminaire dans le cas des paramètres de position ont été proposées dans la littérature en utilisant des moyennes pondérées des estimations obtenues sous les hypothèses nulle et alternative, avec des poids dépendant de la statistique de test, voir par exemple, Saleh (2006). Les estimations de l'erreur quadratique moyenne de ce type n'ont pas été étudiées et nous réservons ce sujet pour de futurs travaux de recherche.

Remerciements

Nous tenons à remercier le rédacteur de ses suggestions très constructives. Les travaux de recherche de Gauri S. Datta ont été financés en partie par la subvention H98230-11-1-0208 de la *National Security Agency*, les travaux de recherche d'Isabel Molina, par les subventions MTM2009-09473, MTM2012-37077-C02-01 et SEJ2007-64500 du *Ministerio de Educación y Ciencia* de l'Espagne, et les travaux de recherche de J.N.K. Rao, par le Conseil de recherches en sciences naturelles et en génie du Canada.

Bibliographie

Bancroft, T.A. (1944). On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics*, 15, 190-204.

Chatterjee, S., Lahiri, P. et Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics*, 36, 1221-1245.

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

Datta, G.S., 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.

Datta, S., Hall, P. et Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, 106, 362-374.

Diao, L., Smith, D.D., Datta, G.S., Maiti, T. et Opsomer, J.D. (2014). Accurate confidence interval estimation of small area parameters under the Fay-Herriot model. *Scandinavian Journal of Statistics*, à paraitre.

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.

Han, C.-P., et Bancroft, T.A. (1968). On pooling means when variance is unknown. *Journal of the American Statistical Association*, 63, 1333-1342.

Li, H., et Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.

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

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken, NJ : Wiley.

Rubin-Bleuer, S., et Yu, Y. (2013). A positive variance estimator for the Fay-Herriot small area model. SRID2-12-001E, Statistique Canada.

Saleh, A.K. Md. E. (2006). *Theory of Preliminary Test and Stein-type Estimation with Applications*. New York : John Wiley & Sons, Inc.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 21-37
Statistique Canada, n° 12-001-X au catalogue

Estimation sur petits domaines en combinant des données provenant de plusieurs sources

Jae-kwang Kim, Seunghwan Park et Seo-young Kim[1]

Résumé

Une approche basée sur un modèle au niveau du domaine pour combiner des données provenant de plusieurs sources est examinée dans le contexte de l'estimation sur petits domaines. Pour chaque petit domaine, plusieurs estimations sont calculées et reliées au moyen d'un système de modèles d'erreur structurels. Le meilleur prédicteur linéaire sans biais du paramètre de petit domaine peut être calculé par la méthode des moindres carrés généralisés. Les paramètres des modèles d'erreur structurels sont estimés en s'appuyant sur la théorie des modèles d'erreur de mesure. L'estimation des erreurs quadratiques moyennes est également discutée. La méthode proposée est appliquée au problème réel des enquêtes sur la population active en Corée.

Mots-clés : Modèle au niveau du domaine; information auxiliaire; modèles d'erreur de mesure; modèle d'erreur structurel; intégration des enquêtes.

1 Introduction

Combiner des données provenant de diverses sources est un problème important en statistique. Dans le contexte des sondages, combiner les données de plusieurs enquêtes peut améliorer la qualité des estimations sur petits domaines. Les données peuvent provenir d'un échantillon probabiliste sur lequel sont faites des mesures directes, d'un autre échantillon probabiliste sur lequel sont faites des mesures indirectes (comme l'état de santé autodéclaré), ou d'information auxiliaire au niveau du domaine. Bon nombre d'approches de combinaison de données, telles que les méthodes à bases de sondage multiples et les méthodes d'appariement statistique, requièrent l'accès à des données au niveau individuel, ce qui n'est pas toujours possible en pratique.

Nous considérons une approche de l'estimation sur petits domaines basée sur un modèle au niveau du domaine lorsqu'il existe plusieurs sources d'information auxiliaire. Pfeffermann (2002) et Rao (2003) ont procédé à une recension détaillée des méthodes utilisées en estimation sur petits domaines. Lohr et Prasad (2003) ont utilisé des modèles multivariés pour combiner l'information provenant de plusieurs enquêtes. Ybarra et Lohr (2008) ont considéré le problème de l'estimation sur petits domaines quand les données auxiliaires au niveau du domaine contiennent des erreurs de mesure. Merkouris (2010) a discuté de l'estimation sur petits domaines lorsque l'on combine des données provenant de plusieurs enquêtes. Raghunathan, Xie, Schenker, Parsons, Davis, Dodd et Feuer (2007), ainsi que Manzi, Spiegelhalter, Turner, Flowers et Thompson (2011) se sont servi de modèles hiérarchiques bayésiens pour combiner les données provenant de plusieurs enquêtes pour l'estimation sur petits domaines. Kim et Rao (2012) ont examiné une approche fondée sur le plan de sondage pour combiner les données provenant de deux enquêtes indépendantes.

Afin de décrire la situation, supposons que la population finie est constituée de H sous-populations, désignées par $U_1, \ldots, U_H$, et que nous souhaitons estimer les totaux de sous-population $X_h = \sum_{i \in U_h} x_i$

1. Jae-kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa, 50011, É.-U.; Seunghwan Park, Department of Statistics, Seoul National University, Seoul, 151-747, Corée. Courriel : kkampsh@gmail.com; Seo-young Kim, Statistical Research Institute, Statistics Korea, Daejon, 302-847, Corée.

d'une variable x pour chaque domaine h. Nous supposons qu'il existe une enquête conçue pour mesurer x_i à partir de l'échantillon, mais que la taille de cet échantillon n'est pas suffisamment grande pour obtenir des estimations de X_h d'une précision raisonnable. Considérons l'une des enquêtes, appelée enquête A, comme étant l'enquête principale, et soit $\hat{X}_h$ un estimateur convergent sous le plan de X_h obtenu à partir de l'enquête A. Souvent, nous calculons $\hat{X}_h = \sum_{i \in A_h} w_{ia} x_i$, où A_h est le jeu d'unités de l'échantillon A pour la sous-population h et w_{ia} est le poids de l'unité i dans l'échantillon A.

En plus de l'enquête principale, supposons qu'il en existe une autre, appelée enquête B, donnant une mesure qui est une estimation grossière de x_i. Soit y_{1i} la mesure prise au moyen de l'enquête B. Nous pouvons supposer que y_{1i} est une mesure grossière de x_i présentant un certain niveau d'erreur de mesure. Donc, nous pouvons émettre l'hypothèse que

$$y_{1i} = \beta_0 + \beta_1 x_i + e_{1i} \tag{1.1}$$

pour certains paramètres (β_0, β_1), où $e_{1i} \sim (0, \sigma_{e1}^2)$. Le modèle (1.1) étant propre à la variable, l'hypothèse de régression linéaire ou les hypothèses de variance égale peuvent être relâchées plus tard. Si $(\beta_0, \beta_1) = (0,1)$, alors le modèle (1.1) signifie qu'il n'y a pas de biais de mesure. Notons que, dans (1.1), les paramètres du modèle (β_0, β_1) ne sont pas propres au domaine, mais peuvent différer pour des groupes de domaines, comme il est démontré dans l'application à l'enquête coréenne sur la population active présentée à la section 5. La spécification de modèles de régression distincts pour différents groupes peut donner lieu à de plus petites erreurs de modélisation et donc accroître l'efficacité statistique de la méthode proposée. Partant de l'enquête B, nous pouvons obtenir un autre estimateur $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$ de X_h, où w_{ib} est le poids de l'unité i dans l'échantillon de l'enquête B, et B_h est l'échantillon B pour la sous-population h. Notons que l'on peut obtenir $\hat{Y}_{1h}$ pour chaque domaine, si les mêmes domaines sont définis dans les deux enquêtes A et B. Le modèle (1.1) peut être utilisé pour combiner l'information provenant des deux enquêtes.

Enfin, les données de recensement peuvent représenter une autre source d'information. Les données de recensement ne souffrent pas d'une erreur de couverture ni d'une erreur d'échantillonnage. Toutefois, elles peuvent présenter des erreurs de mesure et ne fournissent pas d'information mise à jour pour chaque mois ou chaque année. Soit y_{2i} la mesure de l'unité i d'après le recensement. Le total de sous-population $Y_{2h} = \sum_{i \in C_h} y_{2i}$ est disponible quand C_h est le jeu d'unités du recensement C pour la sous-population h.

Le tableau 1.1 résume les principales sources d'information que nous pouvons prendre en considération dans l'estimation sur petits domaines.

Tableau 1.1
Information disponible pour l'estimation sur petits domaines

Données	Observation	Estimation au niveau du domaine	Propriétés
Enquête A	Observation directe (x_i)	$\hat{X}_h, \hat{V}(\hat{X}_h)$	Erreur d'échantillonnage (grande)
Enquête B	Observation auxiliaire (y_{1i})	$\hat{Y}_{1h}, \hat{V}(\hat{Y}_{1h})$	Biais Erreur de mesure Erreur d'échantillonnage
Recensement	Observation auxiliaire (y_{2i})	Y_{2h}	Erreur de mesure Pas d'information mise à jour

Dans le présent article, nous considérons une approche d'estimation sur petits domaines au moyen d'un modèle au niveau du domaine combinant toute l'information disponible. L'approche proposée est basée sur les modèles d'erreur de mesure, dans lesquels les erreurs d'échantillonnage des estimateurs directs sont traitées comme des erreurs de mesure, et toutes les autres données auxiliaires sont combinées au moyen d'un ensemble de modèles de lien. L'approche proposée est appliquée au problème de l'estimation sur petits domaines dans le cas des enquêtes sur la population active en Corée, où trois estimations sont combinées pour produire des estimations sur petits domaines des taux de chômage.

La présentation de l'article est la suivante. À la section 2, nous exposons la théorie de base et nous envisageons le problème d'estimation sur petits domaines comme un problème de prédiction d'un modèle d'erreur de mesure. À la section 3, nous discutons de l'estimation des paramètres du modèle d'estimation sur petits domaines au niveau du domaine. À la section 4, nous décrivons brièvement l'estimation de l'erreur quadratique moyenne. À la section 5, nous appliquons la méthode proposée aux données de l'enquête sur la population active en Corée. Enfin, à la section 6, nous présentons nos conclusions.

2 Théorie de base

À la présente section, nous commençons par présenter la théorie de base qui sous-tend la combinaison de l'information pour l'estimation sur petits domaines. Nous examinons d'abord le cas simple de la combinaison de deux enquêtes. Supposons qu'il existe deux enquêtes, A et B, réalisées selon deux plans d'échantillonnage probabiliste distincts. Les deux enquêtes ne sont pas forcément indépendantes. À partir de l'enquête A, nous obtenons un estimateur sans biais sous le plan $\hat{X}_{h,a} = \sum_{i \in A_h} w_{ia} x_i$ et l'estimateur de sa variance $\hat{V}\left(\hat{X}_h\right)$. À partir de l'enquête B, nous obtenons un estimateur sans biais sous le plan $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$ de $Y_{1h} = \sum_{i \in U_h} y_{1i}$. L'erreur d'échantillonnage de $\left(\hat{X}_h, \hat{Y}_{1h}\right)$ peut être exprimée par le *modèle d'erreur d'échantillonnage*

$$\begin{pmatrix} \hat{X}_h \\ \hat{Y}_{1h} \end{pmatrix} = \begin{pmatrix} X_h \\ Y_{1h} \end{pmatrix} + \begin{pmatrix} N_h a_h \\ N_h b_h \end{pmatrix} \tag{2.1}$$

et a_h et b_h représentent les erreurs d'échantillonnage associées à $\hat{X}_h / N_h$ et à $\hat{Y}_{1h} / N_h$ telles que

$$\begin{pmatrix} a_h \\ b_h \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V(a_h) & \text{Cov}(a_h, b_h) \\ \text{Cov}(a_h, b_h) & V(b_h) \end{pmatrix} \right].$$

Le paramètre d'intérêt est le total de population X_h de x dans le domaine h.

Partant de (1.1), nous obtenons le modèle au niveau du domaine qui suit :

$$Y_{1h} = N_h \beta_0 + \beta_1 X_h + \tilde{e}_{1h}, \tag{2.2}$$

où $(N_h, X_h, Y_{1h}, \tilde{e}_{1h}) = \sum_{i \in U_h} (1, x_i, y_{1i}, e_{1i})$. Nous pouvons exprimer (2.2) en fonction de la moyenne de population

$$\bar{Y}_{1h} = \beta_0 + \bar{X}_h \beta_1 + \bar{e}_{1h}, \tag{2.3}$$

où $(\bar{X}_h, \bar{Y}_{1h}, \bar{e}_{1h}) = N_h^{-1} \sum_{i \in U_h} (x_i, y_{1i}, e_{1i})$. Si nous utilisons un modèle d'erreurs emboîtées

$$e_{1hi} = \varepsilon_h + u_{hi} \tag{2.4}$$

où $\varepsilon_h \sim (0, \sigma_e^2)$ et $u_{hi} \sim (0, \sigma_u^2)$, alors $\bar{e}_{1h} \sim (0, \sigma_{e,h}^2), \sigma_{e,h}^2 = \sigma_e^2 + \sigma_u^2 / N_h$. Le modèle d'erreurs emboîtées, dont l'usage est assez fréquent en estimation sur petits domaines (par exemple, Battese, Harter et Fuller 1988), repose sur l'hypothèse que $\text{Cov}(e_{1hi}, e_{1hj}) = \sigma_e^2$ pour $i \neq j$. Comme N_h est souvent assez grand, nous pouvons supposer sans risque que $\bar{e}_{1h} \sim (0, \sigma_{e,h}^2 = \sigma_e^2)$. Le modèle (2.2) est appelé *modèle d'erreur structurel* parce qu'il décrit la relation structurelle entre les deux variables latentes Y_{1h} et X_h. Les deux modèles, (2.1) et (2.2), sont souvent mentionnés dans la littérature traitant des modèles d'erreur de mesure (Fuller 1987). Donc, le modèle pour l'estimation sur petits domaines peut être considéré comme un modèle d'erreur de mesure, comme l'a suggéré Fuller (1991) qui a été le premier à utiliser l'approche du modèle d'erreur de mesure dans la modélisation au niveau de l'unité pour l'estimation sur petits domaines.

Maintenant, si nous définissons $(\bar{y}_{1h}, \bar{x}_h) = N_h^{-1} (\hat{Y}_{1h}, \hat{X}_h)$, en combinant (2.1) et (2.3), nous obtenons

$$\begin{pmatrix} \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \bar{X}_h \end{pmatrix} + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

qui peut également s'écrire sous la forme

$$\begin{pmatrix} \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}. \tag{2.5}$$

Donc, quand tous les paramètres du modèle (2.5) sont connus, le meilleur estimateur de $\bar{X}_h$ peut être calculé par

$$\hat{\bar{X}}_h = \left\{ (\beta_1, 1) V_h^{-1} (\beta_1, 1)' \right\}^{-1} (\beta_1, 1) V_h^{-1} (\bar{y}_{1h} - \beta_0, \bar{x}_h)' \tag{2.6}$$

où V_h est la matrice de variance-covariance de $(b_h + \bar{e}_{1h}, a_h)'$. La variance de $\hat{\bar{X}}_h$ est donnée par $\left\{ (\beta_1, 1) V_h^{-1} (\beta_1, 1)' \right\}^{-1}$. L'estimateur en (2.6) peut être appelé estimateur par les moindres carrés généralisés (MCG), parce qu'il s'appuie sur la méthode des moindres carrés généralisés de la théorie des modèles linéaires. La méthode MCG est utile parce qu'elle est optimale et qu'elle permet d'incorporer naturellement des sources d'information supplémentaires. Par exemple, si un autre estimateur $\bar{y}_{2h}$ de $\bar{Y}_{2h}$ est également disponible et satisfait

$$\bar{Y}_{2h} = \gamma_0 + \gamma_1 \bar{X}_h + \bar{e}_{2h}$$

et

$$\bar{y}_{2h} = \bar{Y}_{2h} + c_h,$$

alors le modèle MCG étendu s'écrit

$$\begin{pmatrix} \bar{y}_{2h} - \gamma_0 \\ \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} c_h + \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \tag{2.7}$$

et l'estimateur MCG peut être obtenu par

$$\hat{\bar{X}}_{h2} = \left\{(\gamma_1, \beta_1, 1) V_{h2}^{-1} (\gamma_1, \beta_1, 1)'\right\}^{-1} (\gamma_1, \beta_1, 1) V_{h2}^{-1} (\bar{y}_{2h} - \gamma_0, \bar{y}_{1h} - \beta_0, \bar{x}_h)'$$

où V_{h2} est la matrice de variance-covariance de $(c_h + \bar{e}_{2h}, b_h + \bar{e}_{1h}, a_h)'$. La variance de l'estimateur MCG est $\left\{(\gamma_1, \beta_1, 1) V_{h2}^{-1} (\gamma_1, \beta_1, 1)'\right\}^{-1}$. Si $\bar{y}_{2h}$ est indépendant de $(\bar{x}_h, \bar{y}_{1h})$, le gain d'efficacité, en termes de variance relative, qui découle de l'incorporation de $\bar{y}_{2h}$ dans l'estimateur MCG peut s'exprimer sous la forme

$$\frac{V\left(\hat{\bar{X}}_{h2}\right) - V\left(\hat{\bar{X}}_h\right)}{V\left(\hat{\bar{X}}_h\right)} = -\frac{\{V(\bar{y}_{2h}/\gamma_1)\}^{-1}}{\left\{V\left(\hat{\bar{X}}_h\right)\right\}^{-1} + \{V(\bar{y}_{2h}/\gamma_1)\}^{-1}},$$

où $V(\bar{y}_{2h}/\gamma_1) = V(c_h + \bar{e}_{2h})/\gamma_1^2$. Le gain est important si la variance d'échantillonnage de $\bar{y}_{2h}$ ainsi que la variance du modèle $V(\bar{e}_{2h})$ sont faibles. Si $\gamma_1 = 0$, alors le gain est nul.

Remarque 1 *Notons que le modèle (2.5) peut également s'écrire*

$$\begin{pmatrix} \beta_1^{-1}(\bar{y}_{1h} - \beta_0) \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} (b_h + \bar{e}_{1h})/\beta_1 \\ a_h \end{pmatrix}. \tag{2.8}$$

L'estimateur MCG obtenu à partir de (2.8), qui est le même que l'estimateur MCG obtenu à partir de (2.5), peut être exprimé sous la forme

$$\hat{\bar{X}}_h = \alpha_h \bar{x}_h + (1 - \alpha_h)\tilde{x}_h \tag{2.9}$$

où $\tilde{x}_h = \beta_1^{-1}(\bar{y}_{1h} - \beta_0)$ *et*

$$\begin{aligned} \alpha_h &= \frac{V(\tilde{x}_h) - \mathrm{Cov}(\bar{x}_h, \tilde{x}_h)}{V(\bar{x}_h) + V(\tilde{x}_h) - 2\mathrm{Cov}(\bar{x}_h, \tilde{x}_h)} \\ &= \frac{\sigma_{e,h}^2 + V(b_h) - \beta_1 \mathrm{Cov}(a_h, b_h)}{\sigma_{e,h}^2 + V(b_h) + \beta_1^2 V(a_h) - 2\beta_1 \mathrm{Cov}(a_h, b_h)}, \end{aligned}$$

L'estimateur $\tilde{x}_h$, *lorsqu'il est calculé en utilisant le paramètre estimé* $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$, *est appelé estimateur synthétique, et l'estimateur optimal en (2.9) est souvent appelé estimateur composite. On peut montrer qu'en ignorant l'effet de l'estimation de* β, *la variance de l'estimateur composite est égale à*

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \alpha_h V(\bar{x}_h) + (1 - \alpha_h)\mathrm{Cov}(\bar{x}_h, \tilde{x}_h) \tag{2.10}$$

et, comme $\alpha_h < 1$, *l'estimateur composite est plus efficace que l'estimateur direct.*

3 Estimation des paramètres

Maintenant, nous discutons de l'estimation des paramètres du modèle (2.3). L'estimateur MCG de $\beta = (\beta_0, \beta_1)$ peut être obtenu par minimisation de

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} \frac{(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2}{V(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)}. \tag{3.1}$$

Puisque

$$V(\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1) = \sigma_{e,h}^2 + (-\beta_1, 1)\Sigma_h(-\beta_1, 1)', \tag{3.2}$$

où $\sigma_{e,h}^2 = V(\bar{e}_{1h})$ et $\Sigma_h = V\{(a_h, b_h)'\}$, nous pouvons écrire

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} w_h(\beta_1)(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2, \tag{3.3}$$

où $w_h(\beta_1) = \left\{\sigma_{e,h}^2 + (-\beta_1, 1)\Sigma_h(-\beta_1, 1)'\right\}^{-1}$. Maintenant, en résolvant $\partial Q^* / \partial \beta = 0$, nous obtenons

$$\hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w \tag{3.4}$$

et

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{(\bar{x}_h - \bar{x}_w)(\bar{y}_{1h} - \bar{y}_{1w}) - C(a_h, b_h)\}}{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{(\bar{x}_h - \bar{x}_w)^2 - V(a_h)\}}, \tag{3.5}$$

où

$$(\bar{x}_w, \bar{y}_w) = \left\{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\right\}^{-1} \sum_{h=1}^{H} w_h(\hat{\beta}_1)(\bar{x}_h, \bar{y}_h).$$

Notons que le poids $w_h(\beta_1)$ dépend de β_1. Donc, la solution (3.5) peut être obtenue à l'aide d'un algorithme itératif. Après avoir calculé $\hat{\beta}_1$ en utilisant (3.5), on obtient $\hat{\beta}_0$ en utilisant (3.4).

Passons maintenant à l'estimation de la variance du modèle $\sigma_{e,h}^2$. La méthode la plus simple est la méthode des moments (MOM). Autrement dit, nous pouvons utiliser

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_{e,h}^2 \tag{3.6}$$

pour obtenir un estimateur sans biais de $\sigma_{e,h}^2$. Sous le modèle des erreurs emboîtées donné par (2.4), nous avons $\sigma_{e,h}^2 = \sigma_e^2$ et

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_e^2. \tag{3.7}$$

Donc, comme dans Fuller (2009), l'estimateur MOM de σ_e^2 peut être exprimé par

$$\hat{\sigma}_e^2 = \sum_{h=1}^{H} \kappa_h \left\{ \left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h \hat{\beta}_1 \right)^2 - \left(-\hat{\beta}_1, 1 \right) \Sigma_h \left(-\hat{\beta}_1, 1 \right) \right\} \tag{3.8}$$

où

$$\kappa_h \propto \left\{ \hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1 \right) \Sigma_h \left(-\hat{\beta}_1, 1 \right) \right\}^{-1}$$

et $\sum_{h=1}^{H} \kappa_h = 1$. Comme κ_h dépend de $\hat{\sigma}_e^2$, la solution (3.8) peut être obtenue itérativement, en utilisant $\hat{\sigma}_e^2 = 0$ comme valeur initiale. Fay et Herriot (1979) ont utilisé une autre méthode qui est fondée sur la solution itérative de l'équation non linéaire :

$$\sum_{h=1}^{H} \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_h \right)^2}{\sigma_e^2 + \left(-\hat{\beta}_1, 1 \right) \Sigma_h \left(-\hat{\beta}_1, 1 \right)'} = H - 2.$$

En écrivant l'équation susmentionnée sous la forme $g\left(\sigma_e^2\right) = H - 2$, une méthode de type Newton pour $g(\theta) = 0$ avec $\theta = \sigma_e^2$ peut être obtenue par

$$\theta^{(t+1)} = \theta^{(t)} + \frac{1}{g'\left(\theta^{(t)}\right)} \left(H - 2 - g\left(\theta^{(t)}\right) \right) \tag{3.9}$$

où

$$g'(\theta) = -\sum_{h=1}^{H} \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_h \right)^2}{\left\{ \theta + \left(-\hat{\beta}_1, 1 \right) \Sigma_h \left(-\hat{\beta}_1, 1 \right)' \right\}^2}.$$

En supposant que $\sigma_{e,h}^2 \equiv \sigma_e^2$, nous décrivons maintenant la procédure complète d'estimation des paramètres comme il suit :

Étape 1 Calculer l'estimateur initial de (β_0, β_1) en posant que $\hat{\sigma}_e^2 = 0$ dans (3.4) et (3.5).

Étape 2 En se basant sur la valeur courante de $\left(\hat{\beta}_0, \hat{\beta}_1\right)$, calculer $\hat{\sigma}_e^2$ en utilisant l'algorithme itératif en (3.9).

Étape 3 Utiliser la valeur courante de $\hat{\sigma}_e^2$, calculer l'estimateur mis à jour de (β_0, β_1) au moyen de (3.4) et (3.5).

Étape 4 Répéter [Étape 2]-[Étape 3] jusqu'à la convergence.

La méthode d'estimation des paramètres proposée comprend l'estimation de $\beta = (\beta_0, \beta_1)$ par les MCG et l'estimation de σ_e^2 par les MOM itérativement. Notons que l'estimation de β est fondée sur des données provenant de tous les domaines. Si des modèles de régression distincts sont utilisés, la méthode d'estimation des paramètres proposée peut être appliquée à des groupes de domaines. Au lieu de cette

méthode d'estimation itérative distincte, nous pouvons également considérer une autre méthode fondée sur l'estimation du maximum de vraisemblance (EMV) sous des hypothèses distributionnelles paramétriques. Voir Carroll, Rupert et Stefanski (1995) et Schafer (2001) pour une discussion de l'EMV pour les paramètres des modèles d'erreur de mesure.

Remarque 2 *Si l'égalité* $\sigma_{e,h}^2 = \sigma_e^2$ *n'est pas vérifiée, nous pouvons considérer un modèle de rechange tel que*

$$\bar{e}_h \sim \left(0, \bar{X}_h \sigma_e^2\right). \tag{3.10}$$

Pour vérifier si le modèle (3.10) tient, on peut calculer

$$\nu_h = \left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h \hat{\beta}_1\right)^2 - \hat{\beta}_1^2 V(a_h) + 2\hat{\beta}_1 \hat{C}(a_h, b_h) - V(b_h) \tag{3.11}$$

et représenter graphiquement ν_h *en fonction de* $\bar{x}_h$. *Si le graphique montre une relation linéaire, alors (3.10) peut être traité comme un modèle raisonnable. Sous le modèle (3.10), nous pouvons obtenir* σ_e^2 *par une méthode du ratio :*

$$\hat{\sigma}_e^2 = \frac{\sum_{h=1}^{H} \kappa_h \nu_h}{\sum_{h=1}^{H} \kappa_h \hat{\bar{X}}_h} \tag{3.12}$$

où

$$\kappa_h \propto \left\{\hat{\bar{X}}_h \hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1\right) \Sigma_h \left(-\hat{\beta}_1, 1\right)\right\}^{-1}$$

avec $\sum_{h=1}^{H} \kappa_h = 1$, $\hat{\bar{X}}_h$ *défini en (2.9), et* ν_h *défini en (3.11). Comme* κ_h *dépend aussi de* σ_e^2, *la solution (3.12) peut être obtenue par itération.*

Remarque 3 *Nous pouvons également considérer une transformation* $\bar{x}_h^* = T(\bar{x}_h)$ *et* $\bar{y}_{1h}^* = T(\bar{y}_{1h})$ *afin d'améliorer l'approximation par une loi normale asymptotique. Pour vérifier l'écart par rapport à la normalité, nous représentons graphiquement* $n_{ha} \bar{V}(\bar{x}_h)$ *en fonction de* $\bar{x}_h$. *Si le graphique révèle une relation structurelle de* $\bar{x}_h$, *l'hypothèse de normalité peut être mise en doute. Maintenant, considérons la transformation suivante*

$$T(x) = \log(x). \tag{3.13}$$

Notons que la variance asymptotique de $\bar{x}_h^* = T(\bar{x}_h)$ *est égale à*

$$V\left(\bar{x}_h^*\right) \doteq \frac{1}{\left(\bar{x}_h\right)^2} V(\bar{x}_h).$$

Il s'agit d'une transformation stabilisant la variable qui est utile lorsque nous voulons améliorer l'approximation par la loi normale.

Après avoir obtenu l'estimateur MCG $\hat{\bar{X}}_h^*$ *de* $\bar{X}_h^*$, *nous devons appliquer la transformation inverse pour obtenir le meilleur estimateur de* $\bar{X}_h = T^{-1}\left(\bar{X}_h^*\right) := Q\left(\bar{X}_h^*\right)$. *La simple application de la transformation inverse donnera une estimation biaisée. Afin de corriger le biais, nous pouvons utiliser une linéarisation de Taylor d'ordre deux. En effectuant un développement en série de Taylor, nous obtenons*

$$Q\left(\hat{\bar{X}}_h^*\right) \doteq Q\left(\bar{X}_h^*\right) + Q'\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right) + \frac{1}{2}Q''\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right)^2$$

et donc, si nous utilisons $Q\left(\hat{\bar{X}}_h^*\right)$ *comme estimateur de* $\bar{X}_h = Q\left(\bar{X}_h^*\right)$, *nous obtenons, en laissant tomber les termes d'ordre plus faible,*

$$E\left\{Q\left(\hat{\bar{X}}_h^*\right)\right\} = \bar{X}_h + \frac{1}{2}Q''\left(\bar{X}_h^*\right)V\left(\hat{\bar{X}}_h^*\right).$$

Pour la transformation donnée par (3.13), nous avons $Q\left(\bar{X}_h^*\right) = \exp\left(\bar{X}_h^*\right)$ *et donc* $Q''\left(\bar{X}_h^*\right) = \bar{X}_h$. *Donc,* $\hat{\bar{X}}_h = Q\left(\hat{\bar{X}}_h^*\right)$, *et nous obtenons*

$$E\left(\hat{\bar{X}}_h\right) \cong \bar{X}_h + \frac{1}{2}\bar{X}_h V\left(\hat{\bar{X}}_h^*\right)$$

et l'estimateur de $\bar{X}_h$ *corrigé pour le biais est*

$$\hat{\bar{X}}_{h,bc} = \frac{\hat{\bar{X}}_h}{1 + 0,5V\left(\hat{\bar{X}}_h^*\right)}, \tag{3.14}$$

où $V\left(\hat{\bar{X}}_h^*\right)$ *est calculée par la méthode d'estimation de l'EQM dont nous discuterons à la section 4.*

4 Estimation de l'EQM

Passons maintenant à l'estimation de l'erreur quadratique moyenne (EQM) de l'estimateur MCG $\hat{\bar{X}}_h$ qui est donné par (2.9). Notons que l'estimateur MCG est une fonction de (β_0, β_1) et de σ_e^2. Si les paramètres du modèle sont connus, alors l'EQM de $\hat{\bar{X}}_h$ est égale à $M_{h1} = \alpha_h V(\bar{x}_h) + (1 - \alpha_h)\mathrm{Cov}(\bar{x}_h, \tilde{x}_h)$, comme il est discuté dans la remarque 1. Autrement dit, en écrivant $\theta = \left(\beta_0, \beta_1, \sigma_e^2\right)$ et $\hat{\bar{X}}_h = \hat{\bar{X}}_h(\theta)$, la prédiction réelle de $\bar{X}_h$ est calculée par $\hat{\bar{X}}_{eh} = \hat{\bar{X}}_h\left(\hat{\theta}\right)$. Afin de tenir compte de l'effet de l'estimation des paramètres du modèle, nous notons d'abord la décomposition qui suit de $\mathrm{EQM}\left(\hat{\bar{X}}_h^*\right)$:

$$\begin{aligned}\mathrm{EQM}\left(\hat{\bar{X}}_{eh}\right) &= \mathrm{EQM}\left(\hat{\bar{X}}_h\right) + E\left\{\left(\hat{\bar{X}}_{eh} - \hat{\bar{X}}_h\right)^2\right\} \\ &=: M_{h1} + M_{h2},\end{aligned}$$

qui a été prouvée pour la première fois par Kackar et Harville (1984) sous des hypothèses de normalité. Le premier terme, M_{h1}, est d'ordre $1/n_h$, où n_h est la taille de A_h, et le deuxième terme, M_{h2}, est d'ordre $1/n$ avec $n = \sum_{h=1}^{H} n_h$. Le deuxième terme est souvent beaucoup plus petit que le premier.

Nous considérons une approche jackknife pour estimer l'EQM. L'utilisation du jackknife pour obtenir une estimation corrigée pour le biais a été proposée au départ par Quenouille (1956). Jiang, Lahiri et Wan (2002) ont produit une justification rigoureuse de la méthode du jackknife pour l'estimation de l'EQM en estimation sur petits domaines. Les étapes qui suivent peuvent être utilisées pour le calcul du jackknife.

Étape 1 Calculer la k^{e} réplique $\hat{\theta}^{(-k)}$ de $\hat{\theta}$ en supprimant le k^{e} jeu de données de domaine $(\bar{x}_k, \bar{y}_{1k})$ du jeu de données complet $\{(\bar{x}_h, \bar{y}_{1h}); h = 1, 2, \ldots, H\}$. Ce calcul est effectué pour chaque k pour obtenir H répliques de θ : $\{\hat{\theta}^{(-k)}; k = 1, \ldots, H\}$ qui, à leur tour, fournissent H répliques de $\hat{\bar{X}}_h$: $\{\hat{\bar{X}}_h^{(-k)}; k = 1, 2, \ldots, H\}$, où $\hat{\bar{X}}_h^{(-k)} = \hat{\bar{X}}_h(\hat{\theta}^{(-k)})$.

Étape 2 Calculer l'estimateur de M_{h2} sous la forme

$$\hat{M}_{2h} = \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\bar{X}}_h^{(-k)} - \hat{\bar{X}}_h\right)^2. \tag{4.1}$$

Étape 3 Calculer l'estimateur de M_{h1} sous la forme

$$\hat{M}_{1h} = \hat{\alpha}_h^{(\text{JK})} V(\bar{x}_h) + \left(1 - \hat{\alpha}_h^{(\text{JK})}\right)\text{Cov}(\bar{x}_h, \tilde{x}_h) \tag{4.2}$$

où $\hat{\alpha}_h^{(\text{JK})}$ est un estimateur de α_h corrigé pour le biais donné par

$$\hat{\alpha}_h^{(\text{JK})} = \hat{\alpha}_h - \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\alpha}_h^{(-k)} - \hat{\alpha}_h\right),$$

$$\hat{\alpha}_h = \frac{\hat{\sigma}_e^2 + V(b_h) - \hat{\beta}_1\text{Cov}(a_h, b_h)}{\hat{\sigma}_e^2 + V(b_h) + \hat{\beta}_1^2 V(a_h) - 2\hat{\beta}_1\text{Cov}(a_h, b_h)},$$

et

$$\hat{\alpha}_h^{(-k)} = \frac{\hat{\sigma}_e^{(-k)2} + V(b_h) - \hat{\beta}_1^{(-k)}\text{Cov}(a_h, b_h)}{\hat{\sigma}_e^{(-k)2} + V(b_h) + \left(\hat{\beta}_1^{(-k)}\right)^2 V(a_h) - 2\hat{\beta}_1^{(-k)}\text{Cov}(a_h, b_h)}.$$

Remarque 4 *Pour la transformation donnée par (3.13), nous utilisons l'estimateur corrigé pour le biais (3.14) et la méthode d'estimation de son EQM doit être modifiée. En utilisant* $\hat{\bar{X}}_{eh,bc}$ *pour désigner l'estimateur corrigé pour le biais (3.14) évalué à* $\hat{\theta}$, *nous pouvons obtenir*

$$\begin{aligned}
\text{EQM}\left(\hat{\bar{X}}_{eh,bc}\right) &= \text{EQM}\left(\hat{\bar{X}}_{eh}\right) \\
&= \text{EQM}\left\{Q\left(\hat{\bar{X}}_{eh}^*\right)\right\} \\
&\cong \left\{Q'\left(\bar{X}_h^*\right)\right\}^2 \cdot \text{EQM}\left(\hat{\bar{X}}_{eh}^*\right) \\
&= \bar{X}_h^2 \cdot \text{EQM}\left(\hat{\bar{X}}_{eh}^*\right),
\end{aligned}$$

où la première égalité découle du fait que $\hat{\bar{X}}_{h,bc} - \hat{\bar{X}}_h$ *est d'ordre* $O_p\left(n_h^{-1}\right)$. *L'EQM de* $\hat{\bar{X}}_h^*$, *l'estimateur MCGE de* $\bar{X}_h^*$ *après transformation, est calculée au moyen de (4.1) et (4.2). Lorsque*

$\mathrm{EQM}\left(\hat{\bar{X}}_{eh}^{*}\right)$ *est estimée, nous devons la multiplier par* $\hat{\bar{X}}_h^2$ *pour obtenir l'estimateur de l'EQM de l'estimateur MCGE* $\hat{\bar{X}}_{eh,bc}$ *rétrotransformé.*

5 Application à l'Enquête sur la population active de la Corée

Nous examinons maintenant une application de la méthode proposée aux enquêtes sur la population active en Corée. Dans ce pays, deux enquêtes distinctes sur la population active sont utilisées pour obtenir des renseignements au sujet de l'emploi. L'une d'elles est l'Enquête sur la population active coréenne (PAC) et l'autre est l'Enquête sur la population active locale (PAL). L'enquête PAC est réalisée auprès d'un échantillon d'environ 7 000 ménages, tandis que l'enquête PAL est réalisée auprès d'un échantillon d'environ 200 000 ménages. Comme la PAL est une enquête à grande échelle faisant appel à un grand nombre d'intervieweurs à temps partiel, les données comportent un certain niveau d'erreurs de mesure. Nous supposons que l'enquête PAC est exempte d'erreur de mesure, quoiqu'elle présente d'importantes erreurs d'échantillonnage au niveau des petits domaines. L'échantillon de l'enquête PAC est un échantillon de deuxième phase tiré de l'échantillon de l'enquête PAL. Donc, les erreurs d'échantillonnage des estimations d'après les deux enquêtes sont corrélées. Soit $\bar{X}_h$ le taux de chômage (réel) dans le domaine h. Le niveau de petit domaine que nous considérons est appelé « Gu ». La Corée compte 229 « Gu ».

Nous observons $\bar{x}_h$ au moyen de l'enquête PAC et $\bar{y}_{1h}$ au moyen de l'enquête PAL. Pour construire des modèles de lien, nous commençons par diviser la population en deux régions, une région urbaine et une région rurale, en nous basant sur la proportion de ménages travaillant en agriculture. Nous spécifions des modèles distincts pour chaque région (même modèle mais en permettant des paramètres différents) et estimons les paramètres du modèle séparément. Le modèle structurel est

$$\bar{Y}_h = \beta_1 \bar{X}_h + e_h \tag{5.1}$$

avec $e_h \sim \left(0, \sigma_e^2\right)$. Ici, nous posons que $\beta_0 = 0$ pour garantir que l'estimateur MCG de $\bar{X}_h$ n'est pas négatif. Le modèle d'erreur d'échantillonnage reste le même. Dans ce cas, nous pouvons estimer β_1 comme il suit

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\left\{\bar{x}_h \bar{y}_{1h} - C\left(a_h, b_h\right)\right\}}{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\left\{\bar{x}_h^2 - V\left(a_h\right)\right\}}. \tag{5.2}$$

La variance d'échantillonnage de $\left(a_h, b_h\right)$ est calculée en utilisant la méthode d'échantillonnage à deux phases inverse décrite à l'annexe. La variance sous le modèle est estimée par la méthode des moments dans (3.8) avec $\hat{\beta}_0 = 0$. L'estimateur MCG peut être calculé en utilisant (2.9) avec $\tilde{x}_h = \hat{\beta}_1^{-1} \bar{y}_{1h}$.

En plus des deux enquêtes, nous pouvons aussi utiliser l'information provenant du recensement. Le modèle MCG intégrant les trois sources d'information peut être exprimé sous la forme

$$\begin{pmatrix} \bar{Y}_{2h} \\ \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

où $\bar{Y}_{2h}$ est le résultat du recensement pour le domaine h. Comme l'estimation d'après le recensement ne présente pas d'erreur d'échantillonnage, nous avons une seule erreur de modélisation e_{2h} qui représente l'erreur commise quand nous modélisons $E(\bar{Y}_{h2}) = \gamma_1 \bar{X}_h$. Les paramètres du modèle peuvent être obtenus en utilisant la méthode décrite à la section 3 avec $\Sigma_h = \text{diag}(0, V(a_h, b_h))$. L'estimateur MCG de $\bar{X}_h$ s'obtient facilement. L'EQM peut être calculée en utilisant le fait que

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix}' \left\{ V \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \right\}^{-1} \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} := M_{h1}$$

et en appliquant la méthode du jackknife pour corriger le biais.

La figure 5.1 donne le graphique du taux de chômage selon l'enquête PAC en fonction du taux de chômage selon l'enquête PAL pour les domaines urbains. La figure 5.1 montre qu'il existe une relation structurelle linéaire entre les estimations PAC et PAL. Au lieu du résidu habituel $\hat{\bar{e}}_h$ dans le modèle d'erreur structurel, nous utilisons $\hat{v}_h$ en tant que résidu dans le modèle de régression avec erreurs de mesure, où $\hat{v}_h = \bar{y}_{1h} - \hat{\beta}_1 \bar{x}_h$. La figure 5.2 donne le graphique de $\hat{v}_h$ en fonction de $\hat{\bar{X}}_h$ pour les domaines urbains. Le graphique montre que l'hypothèse de variance σ_e^2 égale est légèrement violée. Nous avons également considéré le modèle de variance hétéroscédastique décrit dans la remarque 2, mais les résultats n'ont pas varié de manière significative.

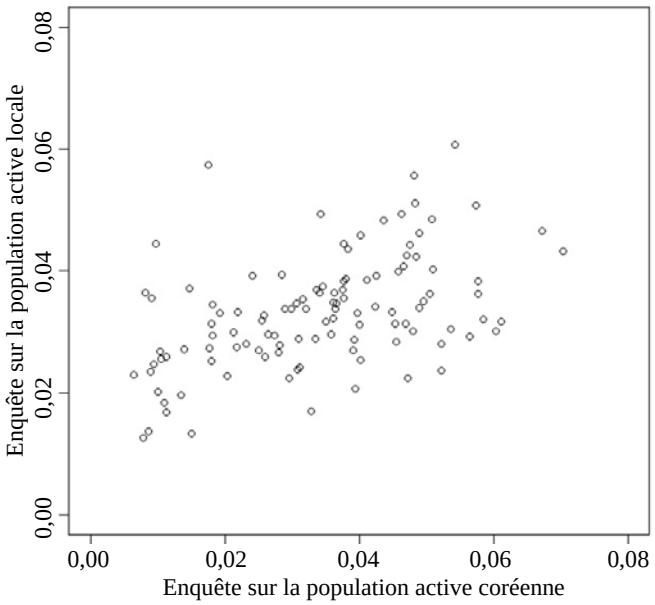


Figure 5.1 Graphique du taux de chômage selon les enquêtes PAC et PAL pour les domaines urbains.

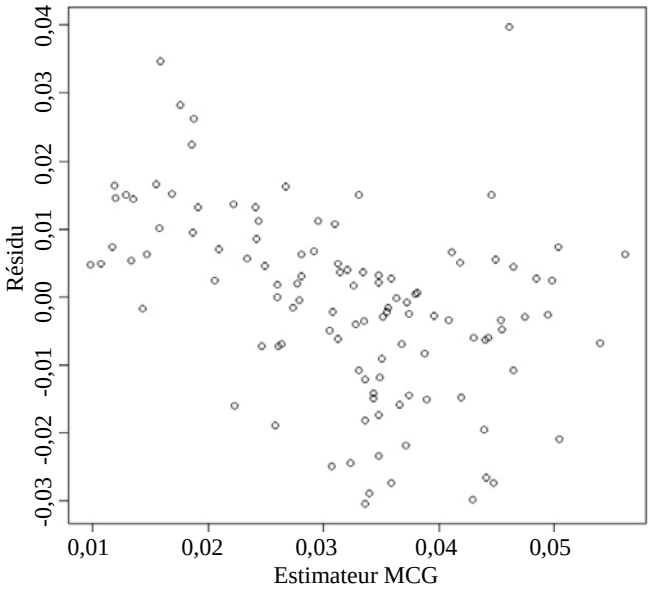


Figure 5.2 Graphique des résidus en fonction des valeurs estimées pour les domaines urbains.

Le tableau 5.1 donne les propriétés des estimations sur petits domaines en ce qui concerne l'EQM estimée. Nous avons examiné quatre estimateurs distincts de $\bar{X}_h$. PAC représente le résultat obtenu en utilisant les données de l'enquête sur la population active coréenne uniquement, PAL représente le résultat obtenu en utilisant les données de l'enquête sur la population active locale uniquement, MCG 1 représente le résultat obtenu en combinant les données des deux enquêtes PAC et PAL, et MCG 2 représente le résultat obtenu en combinant les données des enquêtes PAC et PAL et du recensement. Le tableau 5.1 montre que l'estimateur MCG 2 est celui qui donne les erreurs quadratiques moyennes les plus petites.

Tableau 5.1
Quartile de la performance des estimations sur petits domaines selon l'EQM pour les 229 domaines

EQM	1er Q	Médiane	3e Q	Moyenne
PAC	0,0000630	0,0001210	0,0002395	0,0002476
PAL	0,0001123	0,0001330	0,0001695	0,0001482
MCG 1	0,0000444	0,0000738	0,0001210	0,0000893
MCG 2	0,0000405	0,0000543	0,0000721	0,0000575

6 Conclusion

Le présent article décrit le traitement d'un problème d'estimation sur petits domaines comme un problème de prédiction d'un modèle d'erreur de mesure où les covariables, qui sont les estimations directes pour les petits domaines, sont sujettes à des erreurs d'échantillonnage. Dans notre approche du modèle d'erreur de mesure, les erreurs d'échantillonnage des estimateurs directs sont traitées comme des erreurs de mesure et le modèle d'erreur structurel peut être utilisé pour relier les autres estimations auxiliaires aux estimateurs directs. Le modèle proposé est en fait l'opposé du modèle d'Ybarra et Lohr

(2008), qui traitent l'estimateur direct comme une variable dépendante dans le modèle de régression et les estimations auxiliaires des erreurs non dues à l'échantillonnage comme des erreurs de mesure.

Dans notre approche, chaque estimation auxiliaire est traitée comme une variable dépendante dans le modèle de régression en utilisant l'estimation directe en tant que covariable et l'erreur d'échantillonnage de l'estimateur direct en tant qu'erreur de mesure. La variance de l'erreur de mesure est facile à estimer, parce qu'elle est essentiellement la variance d'échantillonnage de l'estimation directe. L'approche du modèle d'erreur de mesure est également très utile quand il existe plusieurs sources d'information auxiliaire au niveau des domaines. Contrairement à l'approche bayésienne, l'estimateur résultant ne s'appuie pas sur des hypothèses de modélisation paramétrique au sujet du modèle d'erreur structurel et reste optimal au sens de la minimisation des erreurs quadratiques moyennes parmi la classe d'estimateurs sans biais qui sont linéaires dans les données disponibles.

Dans l'exemple de l'application à l'enquête sur la population active de la Corée, deux estimations sur échantillon et l'information provenant du recensement sont utilisées pour calculer les estimations MCG des paramètres de petit domaine et les deux estimations sur échantillon sont corrélées en raison du plan d'échantillonnage à deux phases. Nous avons utilisé simplement des modèles de régression linéaire comme modèles de lien, principalement par souci de simplicité des calculs. Au lieu du modèle linéaire, on pourrait envisager un modèle linéaire généralisé afin d'améliorer le pouvoir de prédiction du modèle. Une telle extension ferait intervenir la théorie des modèles d'erreur de mesure non linéaires. Une étude plus approfondie de cette extension sera le sujet de futurs travaux de recherche.

Remerciements

Nous remercions un examinateur anonyme et le rédacteur associé de leurs commentaires constructifs. Les travaux de recherche du premier auteur ont été financés partiellement par l'entente de coopération NSF (MMS-121339).

Annexe

Échantillonnage à deux phases inverse

En échantillonnage à deux phases classique, l'échantillon de deuxième phase (A_2) est un sous-ensemble de l'échantillon de première phase (A_1). Nous considérons un autre type de plan d'échantillonnage possédant la structure inverse du plan d'échantillonnage à deux phases. Dans le plan d'échantillonnage à deux phases inverse, les étapes d'échantillonnage sont les suivantes :

Étape 1 À partir de la population finie, nous sélectionnons l'échantillon de première phase A_1 de taille n_1.

Étape 2 Dans l'échantillon de deuxième phase, nous sélectionnons A_2 à partir de $U - A_1$ de taille n_2. L'échantillon final A est constitué de A_1 et A_2. C'est-à-dire que $A = A_1 \cup A_2$ et $|A| = n = n_1 + n_2$.

L'échantillonnage à deux phases inverse est utilisé lorsqu'on augmente l'échantillon par une procédure d'échantillonnage additionnelle.

Pour discuter de l'estimation des paramètres sous échantillonnage à deux phases inverse, posons que $\pi_{1i} = \Pr(i \in A_1)$ est la probabilité d'inclusion d'ordre un pour A_1. Soit $\pi_{2i|1} = \Pr\left(i \in A_2 | A_1^c\right)$ la probabilité d'inclusion d'ordre un conditionnelle pour A_2 sachant $A_1^c = U - A_1$. Pour calculer la probabilité d'inclusion pour A, nous avons

$$\Pr(i \in A) = \Pr(i \in A_1) + \Pr\left(i \in A_2 | A_1^c\right) \Pr\left(i \in A_1^c\right).$$

Donc, nous pouvons utiliser $\pi_i = \pi_{1i} + (1 - \pi_{1i})\,\pi_{2i|1}$ pour calculer l'estimateur d'Horvitz-Thompson de la forme

$$\hat{Y}_{r,\mathrm{HT}} = \sum_{i \in A} \frac{1}{\pi_i} y_i. \tag{A.1}$$

Notons que, au lieu de (A.1), nous pouvons considérer la classe d'estimateurs suivante :

$$\hat{Y}_w = W \sum_{i \in A_1} \frac{1}{\pi_{1i}} y_i + (1 - W) \sum_{i \in A_2} \frac{1}{\pi_{2i|1}(1 - \pi_{1i})} y_i := W\hat{Y}_1 + (1 - W)\hat{Y}_2. \tag{A.2}$$

Puisque $\hat{Y}_1$ et $\hat{Y}_2$ sont tous deux sans biais pour Y, $\hat{Y}_w$ est également sans biais quel que soit le choix de W. Un choix raisonnable de W est $W = n_1/n$.

Sous échantillonnage aléatoire simple dans les deux plans, les deux estimateurs sont égaux à $\hat{Y} = N\bar{y}_n$, où $\bar{y}_n$ est la moyenne d'échantillon de y dans A. En écrivant $\bar{y}_1 = n_1^{-1} \sum_{i \in A_1} y_i$ et $\bar{y}_2 = \sum_{i \in A_2} y_i / n_2$, nous obtenons

$$\bar{y}_n = W\bar{y}_1 + (1 - W)\,\bar{y}_2 \tag{A.3}$$

où $W = n_1/n$. En utilisant

$$V(\bar{y}_1) = \left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 \tag{A.4}$$

$$V(\bar{y}_2) = \left(\frac{1}{n_2} - \frac{1}{N}\right) S_y^2$$

$$\mathrm{Cov}(\bar{y}_1, \bar{y}_2) = \mathrm{Cov}\left(\bar{y}_1, \bar{y}_1^c\right) = -\frac{n_1}{N - n_1}\left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 = -\frac{1}{N} S_y^2,$$

où $\bar{y}_1^c = \sum_{i \in A_1^c} y_i / (N - n_1)$, nous obtenons, pour $W = n_1/n$,

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2. \tag{A.5}$$

En outre,

$$\mathrm{Cov}(\bar{y}_1, \bar{y}_n) = \mathrm{Cov}\left[\bar{y}_1, W\bar{y}_1 + (1 - W)\,\bar{y}_2\right] = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2. \tag{A.6}$$

Si l'égalité $W = n_1/n$ n'est pas vérifiée, alors (A.5) et (A.6) ne sont pas vérifiées.

Dans l'application à l'enquête sur la population active de la Corée à la section 5, puisque x et y mesurent le même item, nous pouvons supposer que $S_x^2 = S_y^2 = S_{xy}$ et la matrice de variance-covariance des erreurs d'échantillonnage peut être lissée sous la forme

$$V(a_h, b_h) = \begin{pmatrix} n_1^{-1} & n^{-1} \\ n^{-1} & n^{-1} \end{pmatrix} S_y^2.$$

Bibliographie

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

Carroll, R.J., Rupert, D. et Stefanski, L.A. (1995). *Measurement error in nonlinear models*. New York : Chapman & Hall.

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.

Fuller, W.A. (1987). *Measurement error models*. New York : John Wiley & Sons, Inc.

Fuller, W.A. (1991). Small area estimation as a measurement error problem. Dans *Economic Models, Estimation, and Socioeconomic Systems: Essays in Honor of Karl A. Fox*, (Éds., Tij K. Kaul et Jati K. Sengupta), Elsevier Science Publishers, 333-352.

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

Jiang, J., Lahiri, P. et Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Annals of Statistics,* 30, 1782-1810.

Kackar, R.N., et Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association,* 79, 853-862.

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

Lohr, S.L., et Prasad, N.G.N. (2003). Small area estimation with auxiliary survey data. *The Canadian Journal of Statistics,* 31, 383-396.

Manzi, G., Spiegelhalter, D.J., Turner, R.M., Flowers, J. et Thompson, S.G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society A,* 174, 31-50.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society B,* 68, 509-521.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Revue Internationale de Statistique,* 70, 125-144.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika,* 43, 353-360.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.I., Davis, W.W., Dodd, K.W. et Feuer, E.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association,* 102, 474-486.

Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, NJ.

Schafer, D.W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics,* 57, 53-61.

Ybarra, L.M.R., et Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika,* 95, 919-931.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 39-58
Statistique Canada, n° 12-001-X au catalogue

Meilleure prédiction observée par régression à erreurs emboîtées sous spécification éventuellement inexacte de la moyenne et de la variance

Jiming Jiang, Thuan Nguyen et J. Sunil Rao[1]

Résumé

Nous considérons la méthode de la meilleure prédiction observée (MPO; Jiang, Nguyen et Rao 2011) pour l'estimation sur petits domaines sous le modèle de régression à erreurs emboîtées, où les fonctions moyenne et variance peuvent toutes deux être spécifiées inexactement. Nous montrons au moyen d'une étude par simulation que la MPO peut donner de nettement meilleurs résultats que la méthode du meilleur prédicteur linéaire sans biais empirique (MPLSBE) non seulement en ce qui concerne l'erreur quadratique moyenne de prédiction (EQMP) globale, mais aussi l'EQMP au niveau du domaine pour chacun des petits domaines. Nous proposons, pour estimer l'EQMP au niveau du domaine basée sur le plan de sondage, une méthode du bootstrap simple qui produit toujours des estimations positives de l'EQMP. Nous évaluons les propriétés de l'estimateur de l'EQMP proposé au moyen d'une étude par simulation. Nous examinons une application à la Television School and Family Smoking Prevention and Cessation study.

Mots-clés : EQMP basée sur le plan; hétéroscédasticité; spécification inexacte du modèle; MPO; estimation sur petits domaines; TVSFP.

1 Introduction

La meilleure prédiction observée (MPO; Jiang, Nguyen et Rao 2011) est une nouvelle méthode d'estimation sur petits domaines (EPD; par exemple, Rao 2003). Elle est motivée par le fait que le meilleur prédicteur linéaire sans biais (MPLSB) est un hybride de la meilleure prédiction et de l'estimation du maximum de vraisemblance (MV), alors qu'habituellement en EPD, on s'intéresse surtout à un problème de prédiction. Dans le cas de la méthode MPO, l'estimation du paramètre est basée sur des considérations purement prédictives, menant à ce que l'on appelle le meilleur estimateur prédictif (MEP) des paramètres du modèle. Le développement de la méthode MPO dans Jiang et coll. (2011) est axé principalement sur le modèle de Fay-Herriot (Fay et Herriot 1979). Une autre classe importante de modèles d'EPD est le modèle de régression à erreurs emboîtées (REE) introduit par Battese, Harter et Fuller (1988). Le modèle REE peut être exprimé sous la forme

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij}, \tag{1.1}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, où les v_i sont les effets aléatoires au niveau du domaine et les e_{ij} sont les erreurs qui sont supposés être indépendants et suivre une loi normale de moyenne nulle et de variance $\text{var}(v_i) = \sigma_v^2$ et $\text{var}(e_{ij}) = \sigma_e^2$, où σ_v^2 et σ_e^2 sont inconnues. Sous le modèle REE, la moyenne de petit domaine, en supposant que la population est infinie, est $\theta_i = \bar{X}_i'\beta + v_i$ pour le i^{e} petit domaine, où $\bar{X}_i$ est la moyenne de population des x_{ij} (supposée connue; par exemple, Rao 2003). On voit que θ_i est un

1. Jiming Jiang, Thuan Nguyen et J. Sunil Rao, University of California, Davis, Oregon Health and Science University et University of Miami. Courriel : jimjiang@ucdavis.edu.

effet mixte (linéaire). Soit $\gamma = \sigma_v^2 / \sigma_e^2$. Dès lors, le meilleur prédicteur (MP) de θ_i s'obtient en minimisant l'erreur quadratique moyenne de prédiction (EQMP) basée sur le modèle

$$\mathrm{E}_M \left(\breve{\theta}_i - \theta_i \right)^2, \tag{1.2}$$

où E_M désigne l'espérance sous le modèle REE supposé, et $\breve{\theta}_i$ désigne un prédicteur de θ_i. En vertu de la théorie gaussienne (par exemple, Jiang 2007, page 237), le MP est donné par

$$\tilde{\theta}_i = \mathrm{E}_M (\theta_i | y_i) = \bar{X}_i' \beta + \frac{n_i \gamma}{1 + n_i \gamma} (\bar{y}_{i\cdot} - \bar{x}_{i\cdot}' \beta), \tag{1.3}$$

où $y_i = \left(y_{ij} \right)_{1 \le j \le n_i}$, β et γ sont les paramètres réels, $\bar{y}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ et $\bar{x}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$. La méthode classique du meilleur prédicteur linéaire sans biais (MPLSB) est basée sur l'équation (1.3) dans laquelle β est remplacé par son estimateur du MV, en supposant que γ est connu; et le MPLSB empirique (MPLSBE) est dérivé du MPLSB en remplaçant γ par un estimateur convergent.

Dans la méthode MPO (Jiang et coll. 2011), des estimateurs de β et γ, nommément le MEP, sont calculés en minimisant l'EQMP basée sur le plan observée, ce qui diffère entièrement des méthodes conventionnelles, dont celles du maximum de vraisemblance (MV) et du maximum de vraisemblance restreint (MVR ou REML en anglais; par exemple, Jiang 2007). Tout au long du présent exposé, nous supposons que les échantillons sont tirés de chaque petit domaine par échantillonnage aléatoire simple sans remise, ce qui est le fondement de l'approche basée sur le plan de sondage. Écrivons $\psi = (\beta', \gamma)'$. Notons qu'en pratique, les populations des petits domaines sont finies. À l'instar de Jiang et coll. (2011), nous considérons un modèle REE de superpopulation. Supposons que les sous-populations de réponses $\{Y_{ik}, k = 1, \ldots, N_i\}$ et les données auxiliaires $\{X_{ikl}, k = 1, \ldots, N_i\}, l = 1, \ldots, p$ sont des réalisations provenant des superpopulations correspondantes qui sont supposées satisfaire le modèle REE. Il s'ensuit que

$$Y_{ik} = X_{ik}' \beta + v_i + e_{ik}, \quad i = 1, \ldots, m, \quad k = 1, \ldots, N_i, \tag{1.4}$$

où β, v_i et e_{ik} satisfont les mêmes hypothèses que dans (1.1). Sous les conditions de population finie, la moyenne de petit domaine réelle est $\theta_i = \bar{Y}_i = N_i^{-1} \sum_{k=1}^{N_i} Y_{ik}$ (par opposition à $\theta_i = \bar{X}_i' \beta + v_i$ sous les conditions de population infinie) pour $1 \le i \le m$. En outre, écrivons $r_i = n_i / N_i$. Alors, la version en population finie du MP (1.3) a pour expression (par exemple, Rao 2003, section 7.2.5)

$$\tilde{\theta}_i = \mathrm{E}_M (\theta_i | y_i) = \bar{X}_i' \beta + \left\{ r_i + (1 - r_i) \frac{n_i \gamma}{1 + n_i \gamma} \right\} (\bar{y}_{i\cdot} - \bar{x}_{i\cdot}' \beta), \tag{1.5}$$

où E_M désigne l'espérance (conditionnelle) sous le modèle REE de superpopulation supposé, et β et γ sont les paramètres réels. Notons que le MP est dépendant du modèle.

En pratique, tout modèle supposé est sujet à l'erreur de spécification. Jiang et coll. (2011) considèrent la spécification inexacte de la fonction moyenne, tout en supposant que la structure de variance-covariance

des données est spécifiée correctement. Cependant, en pratique, cette dernière peut elle aussi être mal spécifiée. Dans le présent article, nous étendons la spécification éventuellement inexacte du modèle à la fonction moyenne ainsi qu'à la structure de variance-covariance. Une spécification inexacte possible de la structure de variance-covariance est l'hétéroscédasticité, définie en termes de $\text{var}\left(e_{ij}\right) = \sigma_i^2$ pour le domaine $i, 1 \le i \le m,$ où les σ_i^2 sont inconnues et éventuellement différentes. Cependant, en dépit de la spécification éventuellement inexacte du modèle, il existe des raisons de ne pas pouvoir « abandonner » le modèle supposé, et le MP basé sur le modèle. Premièrement, le modèle supposé et le MP sont relativement simples à utiliser, et par conséquent, attrayants pour les praticiens; en particulier, ils s'appuient sur une relation simple (linéaire) entre la réponse et les autres variables. Par exemple, contrairement à (1.4), qui peut être sujet à une spécification inexacte de la fonction moyenne, $X_{ik}'\beta,$ on peut supposer que $Y_{ik} = \mu_{ik} + v_i + e_{ik},$ où les μ_{ik} sont des constantes inconnues, entièrement non spécifiées. Le dernier modèle est presque toujours exact, mais est inutile, parce qu'il n'utilise aucune relation entre Y et $X.$ En fait, en pratique, si des données auxiliaires sont disponibles, il est souvent considéré « politiquement incorrect » de ne pas les utiliser. Deuxièmement, même si l'on s'inquiète de la spécification inexacte du modèle, on manque souvent de preuves (statistiques) des raisons pour lesquelles une autre spécification est plus raisonnable ou qu'une complication est nécessaire. Par exemple, on émet parfois des réserves quant à l'hypothèse de normalité, alors que rien n'indique pourquoi une autre loi, disons, $t_5,$ est plus raisonnable. En guise d'autre exemple, supposons que l'on ajuste un modèle quadratique et que le coefficient du terme quadratique soit non significatif. Dans ces conditions, il n'est pas certain que la complication de la modélisation quadratique comparativement à la modélisation linéaire soit nécessaire. Par conséquent, dans le présent article, nous ne tentons pas de modifier le modèle supposé, ni le MP, (1.5), basé sur le modèle supposé. En particulier, nous supposons que nous avons un seul paramètre, $\gamma,$ dans (1.5) pour le ratio $\sigma_v^2/\sigma_e^2,$ au lieu de considérer un modèle REE hétéroscédastique semblable à ceux de Jiang et Nguyen (2012) et Nandram et Sun (2012). Notre objectif est de trouver un meilleur moyen d'estimer les paramètres, $\psi,$ sous le modèle supposé qui interviennent dans (1.5), de sorte que le MP résultant, (1.5), soit plus robuste aux spécifications inexactes du modèle. Nous le faisons en considérant une EQMP objective qui ne dépend pas du modèle, définie comme il suit. Soit $\theta = (\theta_i)_{1\le i\le m}$ le vecteur des moyennes de petit domaine, et $\tilde{\theta} = \left[\tilde{\theta}_i\right]_{1\le i\le m}$ le vecteur des MP. Notons que $\tilde{\theta}_i$ dépend de $\psi,$ c'est-à-dire $\tilde{\theta}_i = \tilde{\theta}_i(\psi).$ L'EQMP basée sur le plan est

$$\text{EQMP}\left(\tilde{\theta}\right) = \text{E}\left(\left|\tilde{\theta} - \theta\right|^2\right) = \sum_{i=1}^{m} \text{E}\left\{\tilde{\theta}_i(\psi) - \theta_i\right\}^2. \tag{1.6}$$

Notons que l'espérance E dans (1.6) est différente de E_M dans (1.2), (1.3) ou (1.5) en ce sens que E est entièrement exempte d'un modèle; autrement dit, dans (1.6), l'espérance est calculée par rapport à l'échantillonnage aléatoire simple dans les domaines, ce qui n'a rien à voir avec le modèle supposé. Jiang et coll. (2011) ont montré que l'EQMP donnée en (1.6) possède une autre expression, qui est une idée clé de la MPO. Nommément, nous avons $\text{EQMP}\left(\tilde{\theta}\right) = \text{E}\{Q(\psi) + \cdots\},$ où $\cdots$ ne dépend pas de $\psi,$ et

$$Q(\psi) = \sum_{i=1}^{m}\left\{\tilde{\theta}_i^2(\psi) - 2\frac{1-r_i}{1+n_i\gamma}\bar{y}_{i\cdot}\bar{X}_i'\beta + b_i(\gamma)\hat{\mu}_i^2\right\} = \sum_{i=1}^{m} Q_i. \tag{1.7}$$

Dans (1.7), ψ est considéré comme un vecteur de paramètres, plutôt que le vecteur des paramètres réels, $b_i(\gamma) = 1 - 2a_i(\gamma)$ avec $a_i(\gamma) = r_i + (1 - r_i)n_i\gamma(1 + n_i\gamma)^{-1}$. En outre, $\hat{\mu}_i^2$ est un estimateur sans biais sous le plan de $\bar{Y}_i^2$ dont l'expression est :

$$\hat{\mu}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{N_i - 1}{N_i(n_i - 1)}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2. \tag{1.8}$$

Le MEP de $\psi, \hat{\psi},$ est le minimiseur de $Q(\psi)$ par rapport à ψ. Pour faciliter la lecture, les calculs en vue d'établir (1.7) et (1.8) sont présentés en annexe. Notons aussi que le MP est fondé sur l'EQMP (basée sur le modèle) au niveau du domaine (de sorte qu'elle est optimale pour chaque petit domaine, si le modèle supposé est exact), tandis que le MEP est fondé sur l'EQMP globale (basée sur le plan de sondage). Il en est ainsi parce que nous ne voulons pas que l'estimateur de ψ dépende du domaine. L'une des raisons est que les estimateurs dépendants du domaine sont souvent instables en raison de la petite taille de l'échantillon du domaine, tandis qu'un estimateur obtenu en utilisant tous les domaines, tel que le MEP défini dans le présent article, a tendance à être beaucoup plus stable.

La prise en considération de l'EQMP basée sur le plan de sondage, comme nous le faisons dans le présent article, est due au fait qu'elle est entièrement exempte de modélisation. Notons que, dans Jiang et coll. (2011), où les auteurs ont considéré le modèle de Fay-Herriot, il était impossible d'évaluer l'EQMP basée sur le plan de sondage, parce que les échantillons réels provenant des domaines n'étaient pas disponibles (seuls des résumés des données étaient disponibles au niveau du domaine). Donc, les auteurs ont plutôt considéré l'EQMP basée sur un modèle sous le modèle le plus général, ou le moins contraignant, qui repose simplement sur l'hypothèse que la fonction moyenne est μ_i, où μ_i est complètement inconnue, pour le i^{e} petit domaine. En général, il existe une « règle empirique » pour déterminer le type d'EQMP que l'on doit prendre en considération. Essentiellement, la règle est que l'EQMP doit être exempte de modélisation dans la mesure du possible, afin qu'elle soit objective et (relativement) robuste aux erreurs de spécification du modèle.

À la section 2, nous considérons un exemple simulé dans lequel nous comparons les propriétés prédictives basées sur le plan de sondage de la MPO à celles du MPLSBE. Des comparaisons de ce genre ont été faites dans Jiang et coll. (2011) sous le modèle de Fay-Herriot, mais n'ont jamais été effectuées sous le modèle REE. En outre, les conditions de simulation comprennent la spécification inexacte à la fois de la fonction moyenne et de la fonction variance, ce qui, de nouveau, n'avait pas été considéré auparavant. Les résultats des simulations montrent que la MPO peut donner de meilleurs résultats que le MPLSBE non seulement en ce qui concerne l'EQMP globale basée sur le plan, mais aussi l'EQMP au niveau du domaine (basée sur le plan) pour chacun d'un grand nombre de petits domaines. Il s'agit clairement d'une propriété inédite. Par exemple, Jiang et coll. (2011) ont montré que la MPO donnait de meilleurs résultats que le MPLSBE pour l'EQMP globale, mais pas nécessairement pour chaque petit domaine.

L'estimation des EQMP au niveau des domaines, ici les EQMP basées sur le plan de sondage, représente un important problème d'intérêt pratique. À la section 3, nous proposons un estimateur bootstrap de l'EQMP au niveau du domaine qui a l'avantage d'être simple et toujours positif. Nous décrivons une autre étude par simulation exécutée pour évaluer la performance de l'estimateur de l'EQMP

proposé. Une application au *Television School and Family Smoking Prevention and Cessation Project* (TVSFP) est discutée à la section 4.

2 Études par simulation : MPO c. MPLSBE

2.1 Une démonstration

Nous présentons d'abord un exemple simulé simple pour montrer l'effet que peut avoir la spécification inexacte du modèle sur les propriétés prédictives, basées sur le plan de sondage, de la MPO et du MPLSBE. Soit le cas d'une covariable unique, x_{ij}, considérée comme linéairement associée à la réponse y_{ij} conformément au modèle REE suivant :

$$y_{ij} = \beta x_{ij} + v_i + e_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, 5 \tag{2.1}$$

(donc, nous avons $n_i = 5, 1 \leq i \leq m$ dans ce cas), où β est un coefficient inconnu, et les termes v_i, e_{ij} sont les mêmes que dans (1.1). Donc, nous croyons en particulier que la réponse moyenne doit être nulle quand la valeur de la covariable est nulle.

Nous considérons trois tailles d'échantillon différentes : $m = 50, 100$ ou $400,$ ainsi que deux valeurs réelles différentes de b : $b = 0{,}5$ ou $1{,}0,$ où b est défini ci-après. Dès lors, il existe six cas, chacun étant une combinaison de taille d'échantillon et de valeur de b. Dans chaque cas, une sous-population x est générée à partir de la loi normale de moyenne égale à 1 et d'écart-type égal à $\sqrt{0{,}1} \approx 0{,}32$. La sous-population y est alors générée à partir du modèle de superpopulation REE hétéroscédastique suivant :

$$Y_{ik} = b + v_i + e_{ik}, \quad i = 1, \ldots, m, \quad k = 1, \ldots, 1\,000 \tag{2.2}$$

(donc la taille de la sous-population est $N_i = 1\,000, 1 \leq i \leq m$), où v_i est tiré de la loi normale de moyenne 0 et d'écart-type $\sqrt{0{,}1} \approx 0{,}32; e_{ij}$ est tiré de la loi normale de moyenne 0 et d'écart-type σ_i, où les σ_i^2 sont générés indépendamment à partir de la loi uniforme $[0{,}05;\ 0{,}15]$ (de sorte que l'intervalle pour σ_i est environ de 0,22 à 0,39); et les v_i et les e_{ik} sont générés indépendamment. On voit que le modèle REE supposé est spécifié incorrectement en ce qui concerne les fonctions moyenne ainsi que variance. Une fois que les sous-populations x et y sont générées, elles demeurent fixes dans toutes les simulations.

Dans chaque simulation, nous tirons un échantillon aléatoire simple de taille 5 de $\{1, \ldots, 1\,000\}$ qui détermine les échantillons x_{ij} et $y_{ij}, j = 1, \ldots, 5,$ pour chaque i. L'exercice est répété pour $K = 1\,000$ simulations. Nous effectuons des comparaisons des mêmes données pour la MPO et le MPLSBE, en utilisant l'estimateur du MV de γ pour le second, en ce qui concerne à la fois l'EQMP globale et l'EQMP au niveau du domaine. L'EQMP globale est définie comme étant $\text{EQMP}(\hat{\theta}) = \text{E}\left(\|\hat{\theta} - \theta\|^2\right) = \sum_{i=1}^{m} \text{E}\left(\hat{\theta}_i - \theta_i\right)^2$, où $\theta = (\theta_i)_{1 \leq i \leq m}$ est le vecteur des moyennes réelles de petit domaine avec $\theta_i = \bar{Y}_i$, et $\hat{\theta} = \left(\hat{\theta}_i\right)_{1 \leq i \leq m}$ est le vecteur des valeurs prédites (par la MPO ou par le MPLSBE). Notons que la même

mesure a été utilisée dans Jiang et coll. (2011). Le tableau 2.1 donne les résultats pour l'EQMP globale, où l'EQMP est évaluée empiriquement par $K^{-1}\sum_{k=1}^{K}\left|\hat{\theta}^{(k)} - \theta^{(k)}\right|^2 = K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{m}\left\{\hat{\theta}_i^{(k)} - \theta_i^{(k)}\right\}^2$, et $\theta^{(k)} = \left[\theta_i^{(k)}\right]_{1\le i\le m}$ et $\hat{\theta}^{(k)} = \left[\hat{\theta}_i^{(k)}\right]_{1\le i\le m}$ sont θ et $\hat{\theta}$ dans la k^{e} simulation, respectivement. On voit que l'augmentation en pourcentage de l'EQMP globale du MPLSBE comparativement à celle de la MPO varie d'environ 20 % à presque 1 000 %, selon la taille de l'échantillon et la valeur de b. Les tendances qui se dégagent ici concordent avec celles décrites dans Jiang et coll. (2011) sous le modèle de Fay-Herriot, où les propriétés prédictives basées sur un modèle sont évaluées. Cependant, l'amélioration apportée par la MPO est nettement plus importante, pour $m = 100$ et $m = 400$, que celle mentionnée dans Jiang et coll. (2011).

Tableau 2.1
EQMP globale empirique (augmentation en % pour le MPLSBE par rapport à la MPO)

m	b	MPO	MPLSBE	Augmentation en %
50	0,5	0,130	0,161	24
50	1,0	0,503	0,598	19
100	0,5	0,076	0,277	264
100	1,0	0,396	1,077	172
400	0,5	0,096	0,965	905
400	1,0	0,393	4,046	930

Dans le cas des EQMP au niveau du domaine, à l'instar de Jiang et coll. (2011), nous utilisons des boîtes à moustache pour représenter les distributions des EQMP au niveau du domaine associées aux deux méthodes. Voir la figure 2.1. Les graphiques montrent des détails non révélés par les EQMP globales. Ainsi, on pourrait se demander si l'augmentation en pourcentage de l'EQMP globale dans le cas du MPLSBE est simplement due au nombre accru de domaines additionnés. Un simple calcul donne à penser que cela pourrait ne pas être le cas, par exemple, $(400/50) \times 19$ % vaut seulement 152 % (et non 930 %). Une raison plus explicite est donnée à la figure 2.1. Par exemple, si l'on compare le cas où $m = 50, b = 1$ au cas $m = 400, b = 1,$ on constate que, tandis que le chevauchement entre les boîtes à moustache pour la MPO et le MPLSBE est important dans le premier cas, les boîtes à moustache sont entièrement séparées dans le deuxième; autrement dit, la plus grande EQMP de la MPO au niveau du domaine est plus petite que la plus petite EQMP du MPLSBE au niveau du domaine. Cette constatation ne peut pas être attribuée simplement à l'addition ou à la duplication des domaines. En fait, dans le dernier cas, la MPO donne de nettement meilleurs résultats que le MPLSBE, non seulement globalement, mais aussi pour chacun des 400 petits domaines. Il s'agit clairement d'un résultat inédit. Par exemple, dans le premier exemple simulé de Jiang et coll. (2011), les auteurs ont constaté que l'EQMP de la MPO était plus petite que celle du MPLSBE pour la moitié des petits domaines, tandis que celle du MPLSBE était plus petite que celle de la MPO pour l'autre moitié; des tendances comparables ont été observées dans le deuxième exemple simulé dans Jiang et coll. (2011).

L'estimation des EQMP de la MPO au niveau du domaine est examinée à la section 3.

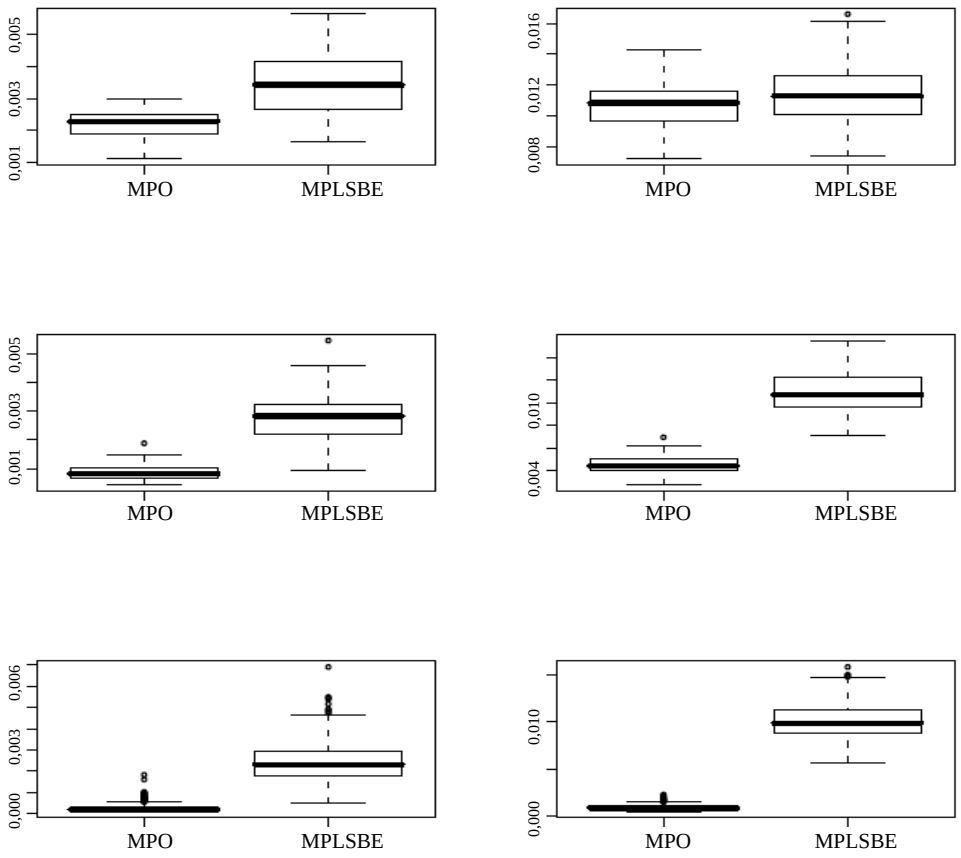


Figure 2.1 EQMP empiriques au niveau du domaine (boîtes à moustache). En haut à gauche : $m = 50, b = 0{,}5$; en haut à droite : $m = 50, b = 1{,}0$; au milieu à gauche : $m = 100, b = 0{,}5$; au milieu à droite : $m = 100, b = 1{,}0$; en bas à gauche : $m = 400, b = 0{,}5$; en bas à droite : $m = 400, b = 1{,}0$.

2.2 Autres considérations

La situation considérée à la sous-section 2.1 pourrait être un peu extrême (raison pour laquelle nous la qualifions de « démonstration théorique »). En pratique, le modèle supposé peut ne pas être entièrement faux, ou être presque exact. À la présente sous-section, nous examinons d'abord un cas où le modèle supposé est « partiellement exact ». Plus précisément, la pente dans (2.1) n'est pas nulle (de sorte que le modèle supposé est correct à cet égard); l'ordonnée à l'origine n'est pas nulle, mais sa valeur est nettement plus faible que celles prises en considération à la sous-section 2.1 (de sorte que le modèle supposé est inexact, mais n'est pas « terriblement inexact »). Plus précisément, le modèle sous-jacent réel est

$$Y_{ij} = b_0 + b_1 X_{ik} + v_i + e_{ik}, \quad i = 1, \ldots, m, \quad k = 1, \ldots, 1\,000, \tag{2.3}$$

par opposition à (2.2), où $b_0 = 0{,}2, b_1 = 0{,}1$; les v_i sont générés indépendamment à partir de la loi normale de moyenne 0 et d'écart-type 0,1; et les e_{ik} sont générées à partir de la loi normale hétéroscédastique comme à la sous-section 2.1. En plus de l'EQMP globale, nous présentons la

contribution à l'EQMP résultant du « biais » et de la « variance ». Posons que $d_i = \hat{\theta}_i - \theta_i$, et que $d_i^{(k)}$ est d_i basé sur le k^{e} ensemble de données simulé, $1 \le k \le K$. Nous définissons le biais et la variance empiriques pour le i^{e} petit domaine comme étant $\bar{d}_i = K^{-1}\sum_{k=1}^{K} d_i^{(k)}$ et $v_i^2 = (K-1)^{-1}\sum_{k=1}^{K}\{d_i^{(k)} - \bar{d}_i\}^2$, respectivement. Notons EQMP_i l'EQMP empirique pour le i^{e} petit domaine. Il est facile de montrer que l'EQMP empirique globale est

$$\sum_{i=1}^{m} \text{EQMP}_i = \frac{K-1}{K}\sum_{i=1}^{m} v_i^2 + \sum_{i=1}^{m}(\bar{d}_i)^2 .$$

Donc, les contributions du biais et de la variance à l'EQMP globale sont définies par $\sum_{i=1}^{m}(\bar{d}_i)^2$ et $\sum_{i=1}^{m} v_i^2$, respectivement. Les résultats basés sur $K = 1\,000$ simulations sont présentés au tableau 2.2. On peut voir que, pour la plus petite valeur de $m, m = 50$, la MPO donne des résultats (légèrement) moins bons que le MPLSBE, mais que pour les plus grandes valeurs de $m, m = 100$ et $m = 400$, la MPO donne des résultats (légèrement) meilleurs, et que l'avantage augmente avec la valeur de m. En ce qui concerne la contribution du biais et de la variance, la MPO semble posséder un biais plus faible, et une variance plus faible pour les valeurs de m plus élevées $(m = 100, 400)$.

Tableau 2.2
EQMP globale empirique (contribution du biais, de la variance) : Le modèle supposé est partiellement exact; augmentation en % donnée pour l'EQMP du MPLSBE par rapport à l'EQMP de la MPO (une valeur négative indique une diminution)

m	MPO	MPLSBE	Augmentation en %
50	0,421 (0,224; 0,197)	0,405 (0,238; 0,167)	-4,0
100	0,733 (0,448; 0,285)	0,748 (0,457; 0,291)	2,1
400	2,745 (1,847; 0,899)	2,848 (1,878; 0,971)	3,8

Ensuite, nous considérons le cas où le modèle supposé est effectivement exact. À savoir, le vrai modèle sous-jacent donné en (2.3) avec $b_0 = 0$; les erreurs e_{ik} sont homoscédastiques de variance égale à 0,1, et tous les autres éléments restent les mêmes que pour le cas susmentionné. Les résultats basés sur $K = 1\,000$ simulations sont présentés au tableau 2.3. Cette fois-ci, nous voyons que le MPLSBE donne des résultats légèrement meilleurs que la MPO sous différentes valeurs de m, mais que l'écart diminue à mesure que la taille d'échantillon augmente. En ce qui concerne la contribution du biais et de la variance, le MPLSBE semble posséder une plus petite variance, et un plus petit biais pour les valeurs de m plus grandes $(m = 100, 400)$, mais les avantages en ce qui concerne tant le biais que la variance se réduisent à mesure que m augmente.

Tableau 2.3
EQMP globale empirique (contribution du biais, de la variance) : Le modèle supposé est exact; augmentation en % donnée pour l'EQMP du MPLSBE par rapport à l'EQMP de la MPO (une valeur négative indique une diminution)

m	MPO	MPLSBE	Augmentation en %
50	0,335 (0,204; 0,131)	0,330 (0,205; 0,125)	-1,4
100	0,749 (0,457; 0,292)	0,746 (0,456; 0,290)	-0,4
400	2,796 (1,800; 0,997)	2,794 (1,799; 0,996)	-0,1

Brièvement, selon les résultats de la simulation, quand la spécification du modèle supposé est légèrement inexacte, la MPO ne donne pas nécessairement de meilleurs résultats que le MPLSBE quand m, le nombre de petits domaines, est relativement faible. Par contre, la MPO devrait surpasser le MPLSBE quand m est relativement grand, et l'avantage de la MPO par rapport au MPLSBE augmente avec m (souvenons-nous de la définition de l'EQMP globale). Par ailleurs, si la spécification du modèle supposé est exacte, le MPLSBE devrait donner de meilleurs résultats que la MPO, quoique l'écart pourrait être ignorable; et l'avantage du MPLSBE par rapport à la MPO s'estompe à mesure que m augmente. Ces résultats, ainsi que ceux de la sous-section 2.1, concordent bien avec ceux de Jiang et coll. (2011; section 4) sous le modèle de Fay-Herriot.

3 Estimation de l'EQMP au niveau du domaine

L'EQMP au niveau du domaine basée sur le plan est définie comme étant

$$\mathrm{EQMP}\left(\hat{\theta}_i\right) = \mathrm{E}\left(\hat{\theta}_i - \theta_i\right)^2, \tag{3.1}$$

où, ici et dans la suite de l'exposé, E désigne l'espérance basée sur le plan, et $\hat{\theta}_i$ est la MPO de θ_i, donnée par (1.5) avec $\psi = (\beta', \gamma)'$ remplacé par son MEP, $\hat{\psi} = \left(\hat{\beta}', \hat{\gamma}\right)'$. Comme l'ont souligné Jiang et coll. (2011), il est difficile d'obtenir un estimateur de l'EQMP au niveau du domaine sans biais d'ordre deux sous une spécification éventuellement inexacte du modèle. En effet, les techniques asymptotiques conventionnelles, telles que la méthode de linéarisation de Prasad-Rao (Prasad et Rao 1990) et la méthode du jackknife (Jiang, Lahiri et Wan 2002), ne s'appliquent plus quand le modèle sous-jacent est spécifié incorrectement. Jiang et coll. (2011) ont utilisé une technique différente pour obtenir un estimateur de l'EQMP par linéarisation qui est sans biais d'ordre deux. Cependant, il n'est pas garanti que cet estimateur soit non négatif. En outre, son terme principal est une fonction $O(1)$ des données au niveau du domaine plutôt que de toutes les données. Plus précisément, le terme principal de l'EQMP de $\hat{\theta}_i$, où $\hat{\theta}_i$ est la MPO de la i^{e} moyenne de petit domaine, θ_i, est $\left(\hat{\theta}_i - y_i\right)^2 + D_i\left(2\hat{B}_i - 1\right)$ sous le modèle de Fay-Herriot, où y_i est l'observation provenant du i^{e} domaine (l'estimateur direct), D_i est la variance d'échantillonnage (connue), $\hat{B}_i = \hat{A}/\left(\hat{A} + D_i\right)$, et $\hat{A}$ est le MEP de la variance de l'effet aléatoire au niveau du domaine. Il s'agit du terme principal parce qu'il est d'ordre $O(1)$, tandis que les autres termes de l'expression de l'EQMP estimée sont d'ordre $O\left(m^{-1}\right)$ ou inférieur. Comme y_i est une observation provenant d'un seul petit domaine, sa variance est assez grande, c'est-à-dire d'ordre $O(1)$, si n_i est borné. Par ailleurs, le MEP $\hat{A}$ est obtenu en utilisant les données provenant de tous les petits domaines et, si bien que sa variance est relativement parlant (beaucoup) plus petite; et $\hat{\theta}_i$ est un mélange de y_i et des MEP. Dès lors, $\left(\hat{\theta}_i - y_i\right)^2$ est le terme qui contribue le plus à la variance, qui peut être assez grande en raison de la variation de y_i. D'autre part, le terme $D_i\left(2\hat{B}_i - 1\right)$ peut être négatif. Par conséquent, en raison de la forte variation de $\left(\hat{\theta}_i - y_i\right)^2$, il existe une probabilité, qui ne disparaît pas (à mesure que m augmente), que le terme principal, donc l'EQMP estimée, soit négatif. Si nous adoptons une approche de linéarisation similaire sous le modèle REE, nous pouvons calculer un estimateur de l'EQMP sans biais

d'ordre deux faisant intervenir $\bar{y}_{i\cdot}$ dans le terme principal, qui est basé sur des données provenant d'un seul petit domaine. Alors, de nouveau, nous nous heurtons au problème d'une forte variation et d'une probabilité qui ne disparaît pas d'une valeur négative de l'estimateur de l'EQMP.

Jiang et coll. (2011) ont aussi utilisé une méthode du bootstrap paramétrique pour obtenir un autre estimateur de l'EQMP; cependant, la justification de l'utilisation de cette méthode est douteuse vu la possibilité que la spécification du modèle soit inexacte. Ici, nous proposons d'utiliser le bootstrap non paramétrique conformément à l'idée originale d'Efron (Efron 1979). La méthode ne s'appuie pas sur le modèle REE et n'est donc pas affectée par la spécification inexacte du modèle. Par conséquent, la méthode courante est mieux justifiée. En outre, il est garanti que l'estimateur de l'EQMP proposé est non négatif, et positif avec une probabilité de 1, ce qui représente un avantage considérable par rapport à l'estimateur de l'EQMP par linéarisation de Jiang et coll. (2011).

Supposons que les sous-populations de petit domaine, ou les N_i, sont suffisamment grandes pour que l'échantillonnage à partir de ces sous-populations puisse être traité approximativement comme étant effectué avec remise. Soit $z_{ij} = \left(x_{ij}', y_{ij}\right)'$, $j = 1, \ldots, n_i$ les échantillons (originaux) provenant du i^{e} petit domaine, $1 \le i \le m$. Nous tirons alors des échantillons, $z_{ij}^{(a)} = \left[\left\{x_{ij}^{(a)}\right\}', y_{ij}^{(a)}\right]'$, $j = 1, \ldots, n_i$, avec remise, de $\left\{z_{ij}, j = 1, \ldots, n_i\right\}$, indépendamment pour $1 \le i \le m$. Supposons que B échantillons bootstrap sont tirés, donnant les échantillons $z^{(a)} = \left\{z_{ij}^{(a)}, 1 \le j \le n_i, 1 \le i \le m\right\}, 1 \le a \le B$. La version sous bootstrap du MP (1.5) est

$$\tilde{\theta}_i^{(a)} = \bar{X}_i'\beta + \left\{r_i + (1 - r_i)\frac{n_i\gamma}{1 + n_i\gamma}\right\}\left[\bar{y}_{i\cdot}^{(a)} - \left\{\bar{x}_{i\cdot}^{(a)}\right\}'\beta\right], \tag{3.2}$$

où β et γ sont les mêmes paramètres de population β et γ, respectivement, que pour la population originale. Notons que les échantillons originaux de z_{ij} sont supposés satisfaire le même modèle REE (1.4), avec X_{ik} (Y_{ik}) remplacé par x_{ij} $\left(y_{ij}\right)$. Puisque les échantillons originaux sont traités comme étant la population bootstrap, suivant l'idée originale d'Efron, les paramètres de population, β, γ, sont les mêmes pour les échantillons bootstrap que pour les échantillons originaux. Néanmoins, comme nous l'avons mentionné, la procédure bootstrap proposée est non paramétrique en ce sens que le modèle supposé, (1.4), ne joue aucun rôle dans le tirage des échantillons bootstrap. En particulier, les MEP de β et γ, basés sur les échantillons originaux, ne sont utilisés nulle part dans la procédure bootstrap; et les quantités d'intérêt dans la population sont $\bar{Y}_i, 1 \le i \le m$, dont les analogues bootstrap sont $\bar{y}_{i\cdot}, 1 \le i \le m$. Cela diffère du bootstrap paramétrique de Jiang et coll. (2011), où les MEP des paramètres du modèle, basés sur les échantillons originaux, sont utilisés pour tirer les échantillons bootstrap sous le modèle supposé. Notons aussi que, parce que les $\bar{X}_i$ sont connus, ils sont traités comme des constantes connues, et par conséquent ne changent pas durant la procédure bootstrap (cela n'a aucun sens d'« estimer » quelque chose que l'on connaît déjà). À part cela, la procédure suit de près l'idée du bootstrap classique (par exemple, Efron et Tibshirani (1993); voir aussi Chatterjee, Lahiri et Li (2008) pour une application à l'estimation sur petits domaines). L'estimateur bootstrap de $\text{EQMP}\left(\hat{\theta}_i\right) = \text{E}\left(\hat{\theta}_i - \bar{Y}_i\right)^2$ est

$$\widehat{\text{EQMP}}\left(\hat{\theta}_i\right) = \frac{1}{B}\sum_{a=1}^{B}\left\{\hat{\theta}_i^{(a)} - \bar{y}_{i\cdot}\right\}^2, \tag{3.3}$$

où $\hat{\theta}_i^{(a)}$ est (3.2) avec β, γ remplacés par leurs MEP basés sur les échantillons bootstrap.

Nota. On pourrait s'inquiéter du fait que, comme les n_i peuvent être petits dans les problèmes d'EPD types, il puisse ne pas exister de nombreux échantillons bootstrap distincts pour chaque petit domaine. Cependant, les données se rapportent non pas à un seul, mais à un grand nombre de petits domaines. Quand tous les petits domaines sont combinés, il reste encore un grand nombre d'échantillons bootstrap distincts, même si les n_i sont petits.

Nous évaluons les propriétés de l'estimateur de l'EQMP proposé en considérant l'échantillon simulé de la sous-section 2.1 avec $b = 0{,}5,$ mais sous des tailles d'échantillon plus petites. C'est-à-dire que nous prenons pour point de départ la taille d'échantillon de base $m = 10$ et $n_i = 5,$ puis nous augmentons $n_i,$ pour passer de 5 à 10, ou nous augmentons $m,$ pour passer de 10 à 20. Nous considérons d'abord le biais sous le plan de sondage de l'estimateur $\widehat{\text{EQMP}}\left(\hat{\theta}_i\right).$ Nous générons deux populations finies que nous fixons ensuite de manière que la population finie pour $m = 10$ soit une sous-population de la population finie pour $m = 20.$ Le tableau 3.1 donne, pour les dix premiers petits domaines (il s'agit de tous les petits domaines qui sont communs sous différentes valeurs de m), l'EQMP réelle simulée (EQMP), obtenue de la même façon qu'à la section 2, la moyenne simulée de $\widehat{\text{EQMP}}\left(\hat{\theta}_i\right)\left(\widehat{\text{EQMP}}\right),$ et le biais relatif en pourcentage (BR %) défini comme étant

$$100 \times \left\{\frac{\text{E}\left(\widehat{\text{EQMP}}\right) - \text{EQMP réelle}}{\text{EQMP réelle}}\right\},$$

où l'espérance est basée sur les simulations. Une autre mesure de performance est la racine carrée de l'erreur quadratique moyenne (REQM) sur l'ensemble des simulations, définie par

$$\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\text{EQMP}}_{i,k} - \text{EQMP}_i\right)^2}$$

pour le i^{e} petit domaine, où EQMP_i est l'EQMP réelle pour le i^{e} petit domaine (qui ne dépend pas de k), évaluée sur l'ensemble des simulations, et $\widehat{\text{EQMP}}_{i,k}$ est l'estimation de l'EQMP basée sur le k^{e} ensemble de données simulé. Nous considérons $B = 100$ comme étant le nombre d'échantillons bootstrap utilisés pour évaluer l'estimateur de l'EQMP, (3.3). Tous les résultats sont basés sur 1 000 simulations. On voit que, globalement, les résultats s'améliorent quand n_i ou m augmente, mais en ce qui concerne le biais relatif en pourcentage (BR %), l'amélioration est plus universelle, ou efficace, quand n_i augmente. Cela tient principalement au fait que, quand n_i augmente, l'échantillon est une meilleure approximation de la population; d'où, la distribution bootstrap est une meilleure approximation de la distribution de la population. En outre, notons que, selon le domaine, le signe du BR peut être positif ou négatif. Cela tient principalement aux différences d'un domaine à l'autre (rappelons que les populations sont fixes) ainsi qu'aux erreurs bootstrap. Pour obtenir certaines mesures globales, nous donnons la moyenne et l'écart-type (é.-t.) des biais relatifs en pourcentage (BR %) sur les dix petits domaines définis comme il suit : $m = 10, n_i = 5$: moyenne = 4,2 %, é.-t. = 14,8 %; $m = 10,$ $n_i = 10$: moyenne = 1,5 %, é.-t. = 4,2 %; $m = 20, n_i = 5$: moyenne = –0,6 %, é.-t. = 8,1 %. Les boîtes à moustache pour BR % sont présentées à la figure 3.1. Les graphiques illustrent aussi le schéma d'amélioration. Par ailleurs, en ce qui concerne la REQM, l'amélioration est beaucoup plus importante

quand m augmente que quand n_i augmente. Il en est ainsi parce qu'une plus grande valeur de m réduit les EQMP en général; donc, naturellement, les estimations correspondantes de l'EQMP diminuent également. Autrement dit, l'estimateur ainsi que le paramètre (l'EQMP) diminuent, ce qui se traduit habituellement par une réduction de la REQM. Le sommaire et les boîtes à moustache pour la REQM sont omis.

En outre, au tableau 3.1, le BR en % et la REQM fluctuent d'un domaine à l'autre, ce qui s'explique surtout par les différences de domaine en domaine. Rappelons que les populations des petits domaines sont générées chacune à partir d'une population de taille $N_i = 1\,000,$ puis fixées tout au long de la simulation. Bien que les superpopulations utilisées pour générer les populations des petits domaines, y compris X et $Y,$ soient les mêmes, il persiste certaines différences entre les populations finies générées, en particulier parce que la taille de population, $N_i,$ n'est pas très grande.

Tableau 3.1
Propriétés empiriques de $\widehat{\text{EQMP}}$

m	n_i	i	EQMP	$\widehat{\text{EQMP}}$	BR %	REQM	i	EQMP	$\widehat{\text{EQMP}}$	BR %	REQM
10	5	1	0,041	0,042	4,5	0,103	6	0,034	0,043	26,3	0,070
10	10	1	0,036	0,036	-0,4	0,068	6	0,034	0,036	6,4	0,070
20	5	1	0,031	0,032	4,1	0,051	6	0,028	0,031	12,5	0,046
10	5	2	0,046	0,038	-16,1	0,078	7	0,032	0,040	25,4	0,078
10	10	2	0,035	0,033	-4,1	0,078	7	0,033	0,034	2,7	0,068
20	5	2	0,031	0,029	-7,2	0,050	7	0,030	0,031	3,6	0,055
10	5	3	0,038	0,042	10,2	0,121	8	0,042	0,042	-0,4	0,150
10	10	3	0,037	0,036	-1,7	0,091	8	0,033	0,035	7,5	0,067
20	5	3	0,031	0,032	4,4	0,052	8	0,030	0,031	4,1	0,058
10	5	4	0,056	0,052	-7,6	0,121	9	0,050	0,042	-15,0	0,074
10	10	4	0,037	0,040	6,3	0,072	9	0,034	0,034	-1,0	0,063
20	5	4	0,040	0,035	-11,3	0,068	9	0,034	0,030	-11,1	0,049
10	5	5	0,033	0,037	11,8	0,066	10	0,041	0,043	3,1	0,082
10	10	5	0,032	0,033	2,5	0,066	10	0,034	0,033	-2,9	0,073
20	5	5	0,024	0,025	2,9	0,052	10	0,035	0,033	-7,9	0,062

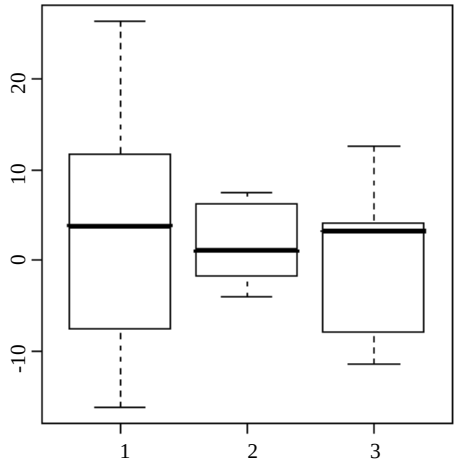


Figure 3.1 Boîtes à moustache de BR %. 1 : $m = 10, n_i = 5$; 2 : $m = 10, n_i = 10$; 3 : $m = 20, n_i = 5$.

Nous concluons la présente section par certains commentaires de nature théorique. Bien que des études approfondies de l'estimation de l'EQMP dans le contexte de l'EPD aient été effectuées depuis la publication de l'article fondateur de Prasad et Rao (Prasad et Rao 1990), la grande majorité de ces travaux était axés sur l'EQMP basée sur un modèle. Voir, par exemple, Datta, Kubokawa, Molina et Rao (2011),

Lahiri (2012), et Torabi et Rao (2012) pour certains travaux récents sur l'estimation de l'EQMP basée sur le plan de sondage en EPD. Comme il est mentionné dans Jiang et coll. (2011), sous une spécification éventuellement inexacte du modèle, l'EQMP au niveau du domaine basée sur un modèle n'est pas estimable de manière convergente, et cela vaut également pour l'EQMP au niveau du domaine basée sur le plan de sondage. En effet, quand le modèle est mal spécifié en ce qui concerne la fonction moyenne, l'EQMP n'est pas une fonction d'un nombre fini de paramètres (tel que β, γ et σ_e^2). En fait, comme nous travaillons sous spécification éventuellement inexacte du modèle, les quantités telles que $\bar{Y}_i^2, 1 \le i \le m$ interviennent dans les expressions des EQMP au niveau du domaine, qui devraient toutes être traitées comme des paramètres inconnus. En outre, la taille effective de l'échantillon pour l'estimation de $\bar{Y}_i^2$ est n_i, si le modèle supposé est défaillant. Il s'ensuit que $\bar{Y}_i^2$ ne peut pas être estimé de manière convergente en utilisant les données provenant du domaine seulement si n_i est borné. Généralement parlant, si l'EQMP peut être estimée de manière convergente, la différence entre l'estimateur de l'EQMP et l'EQMP est d'ordre $O_{\mathrm{P}}\left(m^{-1/2}\right)$; par conséquent, le biais est habituellement d'ordre $O\left(m^{-1}\right)$ sans correction du biais. Par ailleurs, si l'EQMP (au niveau du domaine) ne peut pas être estimée de manière convergente, la différence entre l'estimateur de l'EQMP et l'EQMP est habituellement d'ordre $O_{\mathrm{P}}\left\{(m \wedge n_i)^{-1/2}\right\}$, où $m \wedge n = \min(m, n)$, d'où le biais est habituellement $O\left\{(m \wedge n_i)^{-1}\right\}$, sans la correction du biais. L'estimateur bootstrap de l'EQMP, $\widehat{\mathrm{EQMP}}$, possède la dernière propriété, en plus du fait qu'il est toujours non négatif. Bien qu'il soit possible de corriger le biais de $\widehat{\mathrm{EQMP}}$ afin de réduire l'ordre du biais à $o\left\{(m \wedge n_i)^{-1}\right\}$ (par exemple, Hall et Maiti 2006), la propriété de non-négativité peut disparaître après la correction du biais. Compte tenu de la discussion qui précède, il semble que, sous spécification éventuellement inexacte du modèle, il est raisonnable de définir l'absence de biais d'ordre un et d'ordre deux d'un estimateur de l'EQMP au niveau du domaine en termes de $O\left\{(m \wedge n_i)^{-1}\right\}$ et $o\left\{(m \wedge n_i)^{-1}\right\}$, au lieu des $O\left(m^{-1}\right)$ et $o\left(m^{-1}\right)$ classiques (par exemple, Rao 2003).

4 Une application

Nous considérons une application des méthodes développées aux sections précédentes aux données du TVSFP. Pour une description complète de l'étude du TVSFP, voir Hedeker, Gibbons et Flay (1994). L'étude originale a été conçue pour tester les effets indépendants ainsi que combinés d'un programme de résistance sociale en milieu scolaire, d'une part, et télévisé, d'autre part, concernant la prévention et l'arrêt du tabagisme. Les sujets étaient des élèves de septième année de Los Angeles (LA) et de San Diego, dans l'État de Californie, aux États-Unis. Les élèves ont été prétestés en janvier 1986 dans le cadre d'une première étude. Les mêmes élèves ont rempli un questionnaire directement après l'intervention en avril 1986, un questionnaire de suivi un an plus tard (en avril 1987), et un questionnaire de suivi deux ans plus tard (en avril 1988). Dans la présente analyse, nous considérons un sous-ensemble des données du TVSFP portant sur les élèves de 28 écoles de Los Angeles, où les écoles ont été affectées aléatoirement à l'une de quatre conditions d'étude : a) un programme scolaire de résistance sociale (PS); b) une intervention médiatique (télévision) (TV); c) une combinaison des conditions PS et TV; et d) un groupe de contrôle sans traitement. L'une des principales variables de résultat de l'étude était la cote obtenue sur une échelle

des connaissances concernant le tabac et la santé (THKS pour *tobacco and health knowledge scale*), et est celle utilisée dans la présente analyse. La THKS consistait en un questionnaire à sept items utilisé pour évaluer les connaissances des élèves concernant le tabac et la santé. La cote THKS de l'élève a été définie comme la somme des items auxquels l'élève avait répondu correctement. Seules les données du prétest et de l'évaluation directement après l'intervention sont disponibles pour la présente analyse. Plus précisément, les données portent uniquement sur les sujets qui avaient rempli le questionnaire THKS à ces deux points dans le temps. D'une part, les données des enregistrements complets représentent une situation « avant-après » idéale; d'autre part, les données manquantes, c'est-à-dire celles fournies par les sujets qui ont rempli le questionnaire à un seul point dans le temps, auraient pu fournir des renseignements supplémentaires utiles. Par exemple, il se peut qu'un sujet n'ait pas rempli le questionnaire de suivi parce qu'il n'avait pas trouvé le programme utile. Malheureusement, les données incomplètes n'étaient pas disponibles. Par conséquent, l'analyse des enregistrements complets seulement comporte un risque de biais de sélection. Dans l'ensemble, l'échantillon comprenait 1 600 élèves répartis entre 28 écoles, le nombre d'élèves provenant de chaque école variant de 18 à 137.

Hedeker et coll. (1994) ont procédé à une analyse avec modèles mixtes basée sur un certain nombre de modèles REE pour illustrer l'estimation du maximum de vraisemblance pour l'analyse des données groupées. Ici, nous considérons le problème d'estimation des moyennes de petit domaine de l'écart entre les cotes THKS (la réponse) obtenues directement après l'intervention et au prétest. Ici, le « petit domaine » s'entend d'un certain nombre de caractéristiques importantes (par exemple, région de résidence, ratio enseignant/élèves) qui affectent la réponse, mais dont ne rendent pas compte les covariables du modèle (c'est-à-dire combinaison linéaire des indicateurs PS, TV et PSTV). Notons qu'habituellement, le terme « petit domaine » fait référence à de petites régions géographiques ou sous-populations pour lesquelles un échantillon adéquat n'est pas disponible (par exemple, Rao 2003), et des renseignements tels que les caractéristiques résidentielles ou les ratios enseignant/élèves seraient utilisés comme covariables supplémentaires. Cependant, les données sur ce genre de caractéristiques ne sont pas disponibles. C'est pourquoi nous définissons cette information non disponible comme étant « au niveau du domaine », afin qu'elle puisse être traitée comme les effets aléatoires (de petit domaine). Cette approche est en harmonie avec les caractéristiques fondamentales des effets aléatoires qui sont souvent utilisées pour traduire les effets ou l'information inobservable (par exemple, Jiang 2007), et étend la notion classique d'estimation sur petits domaines. Donc, un petit domaine correspond aux élèves de septième année dans toutes les écoles des États-Unis dont les caractéristiques principales sont similaires à celles d'une école de Los Angeles comprise dans les données durant une période raisonnable (par exemple, cinq ans) afin que ni ces caractéristiques, ni la pertinence sociale/éducative des programmes PS et TV n'aient beaucoup évolué au fil du temps. Les données du TVSFP englobent 28 écoles de Los Angeles qui correspondent à 28 ensembles de caractéristiques, de sorte que les données sont considérées comme des échantillons aléatoires provenant de 28 petits domaines définis comme il est indiqué plus haut. Ainsi, chaque population de petit domaine est suffisamment grande pour que $n_i/N_i \approx 0{,}1 \le i \le 28$. Rappelons que les n_i dans l'échantillon TVSFP varient de 18 à 137, tandis que les N_i devraient être au moins de l'ordre de dizaines de milliers. Notons que, dans le calcul de la MPO, le seul endroit où il est nécessaire de connaître N_i est dans le ratio n_i/N_i. Le modèle REE proposé peut être exprimé comme en (1.1) avec $x_{ij}'\beta = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1}x_{i,2}$, où $x_{i,1} = 1$ dans le cas PS, et 0 autrement; $x_{i,2} = 1$ dans le cas TV, et 0 autrement. Il s'ensuit que les données auxiliaires x_i sont des données au niveau du domaine; par conséquent, la valeur de $\bar{X}_i$ est connue pour chaque i.

Comme nous l'avons mentionné, les tailles d'échantillon de certains petits domaines sont assez grandes, mais il existe aussi des domaines dont les tailles d'échantillon sont relativement parlant (beaucoup) plus petites, ce qui est assez fréquent dans les situations réelles. Comme les données auxiliaires sont des données au niveau du domaine, nous avons $\bar{X}_i'\beta = \bar{x}_i'\beta$; donc, il est facile de montrer que le MP (1.5) peut être exprimé sous la forme

$$\tilde{\theta}_i = \left\{ r_i + (1 - r_i)\frac{n_i\gamma}{1 + n_i\gamma} \right\} \bar{y}_i + \frac{1 - r_i}{1 + n_i\gamma}\bar{x}_i'\beta.$$

Nous voyons que, quand n_i est grand, la MP est approximativement égal à $\bar{y}_i$, l'estimateur sous le plan de sondage, qui n'a rien à voir avec l'estimation du paramètre. Par conséquent, quand n_i est grand, la différence entre la MPO et le MPLSBE est faible. Par contre, si n_i est petit ou moyen, nous nous attendons à observer une certaine différence entre la MPO et le MPLSBE en ce qui concerne l'EQMP. Cependant, il est difficile de dire quelle est la grandeur de cette différence dans le présent exemple sur données réelles. Nos résultats de simulation de la section 2 montrent que la différence entre la MPO et le MPLSBE concernant l'EQMP dépend de la mesure dans laquelle la spécification du modèle supposé est inexacte. Il convient de souligner que la réponse, y_{ij}, est la différence entre les cotes THKS, et que les valeurs possibles de la cote THKS sont des nombres entiers compris entre 0 et 7. Manifestement, de telles données ne suivent pas une loi normale. L'effet possible de la non-normalité est double. D'une part, il est probable que le modèle REE, tel qu'il est proposé par Hedeker et coll. (1994), est spécifié incorrectement, auquel cas l'expression (1.5) n'est plus le MP, et les estimateurs du MV (MVR) gaussien ne sont plus les vrais estimateurs du MV (MVR). D'autre part, même si les données ne suivent pas une loi normale, il reste possible de justifier que (1.5) est le meilleur prédicteur linéaire (MPL; par exemple, Searle, Casella et McCulloch 1992, section 7.3). En outre, les estimateurs du MV (MVR) gaussiens sont convergents et asymptotiquement normaux, même sans l'hypothèse de normalité (Jiang 1996; voir aussi Jiang 2007, chapitre 1). D'autres aspects du modèle REE comprennent l'homoscédasticité de la variance de l'erreur sur l'ensemble des petits domaines. La figure 4.1 montre l'histogramme des variances d'échantillon des 28 petits domaines. La forme bimodale de l'histogramme donne à penser que la variance de l'erreur pourrait être hétéroscédastique, soit encore un autre type possible de spécification inexacte du modèle. Par conséquent, la méthode de la MPO est un choix naturel.

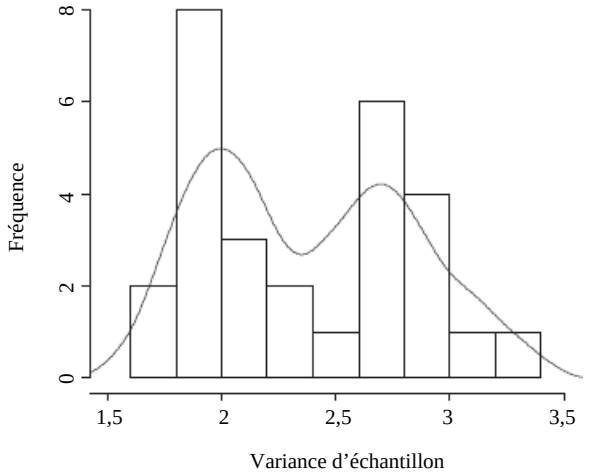


Figure 4.1 Histogramme des variances d'échantillon; un lisseur à noyau de la densité est ajusté.

Nous effectuons l'analyse de la MPO pour les 28 petits domaines et les résultats sont présentés au tableau 4.1. Les MEP des paramètres sont $\hat{\beta}_0 = 0{,}206$, $\hat{\beta}_1 = 0{,}687$, $\hat{\beta}_2 = 0{,}213$, $\hat{\beta}_3 = -0{,}288$ et $\hat{\gamma} = 0{,}003$. Bien que l'on puisse donner une interprétation des estimations des paramètres, il se pourrait que la spécification du modèle soit inexacte (auquel cas l'interprétation pourrait ne pas avoir de sens), comme nous l'avons mentionné plus haut. Quoi qu'il en soit, nous nous intéressons principalement à la prédiction et non à l'estimation; donc, nous nous concentrons sur la MPO. En plus des MPO, nous calculons aussi les estimateurs $\widehat{\text{EQMP}}$ correspondants, et leurs racines carrées comme mesures de l'incertitude. Aux fins de comparaison, nous incluons aussi dans le tableau les MPLSBE pour les petits domaines, ainsi que les racines carrées des estimations de l'EQMP correspondantes, $\widetilde{\text{EQMP}}$, en utilisant la méthode de Prasad-Rao (P-R; Prasad et Rao 1990). Nous voyons que les MPO sont toutes positives, même pour les petits domaines dans le groupe de contrôle. En ce qui concerne la signification statistique (ici, la « signification » est définie comme le fait que la MPO est plus grande en valeur absolue que 2 fois la racine carrée de l'estimation de l'EQMP correspondante), les moyennes de petit domaine sont significativement positives pour tous les petits domaines du groupe (1,1). Par contre, aucune des moyennes de petit domaine n'est significativement positive pour les petits domaines du groupe (0,0). Pour les deux autres groupes, les moyennes de petit domaine sont significativement positives pour tous les petits domaines du groupe (1,0), tandis qu'elles sont significativement positives pour tous les petits domaines sauf deux du groupe (0,1). Les groupes (0,0), (0,1), (1,0) et (1,1) contiennent 7, 8, 7 et 7 petits domaines, respectivement.

Tableau 4.1
MPO, MPLSBE, mesures de l'incertitude pour les données du TVSFP (Partie 1)

ID	PS	TV	MPO	$\sqrt{\widehat{\text{EQMP}}}$	MPLSBE	$\sqrt{\widetilde{\text{EQMP}}}$
403	1	0	0,886	0,171	0,913	0,121
404	1	1	0,844	0,296	0,856	0,121
193	0	0	0,215	0,207	0,217	0,120
194	0	0	0,221	0,137	0,221	0,134
196	1	0	0,878	0,171	0,907	0,124
197	0	0	0,225	0,158	0,223	0,126
198	1	1	0,771	0,220	0,807	0,131
199	0	1	0,426	0,142	0,453	0,130
401	1	1	0,826	0,133	0,844	0,127
402	0	0	0,188	0,171	0,199	0,123
405	0	1	0,394	0,147	0,432	0,129
407	0	1	0,508	0,300	0,508	0,133
408	1	0	0,871	0,240	0,903	0,123
409	0	0	0,230	0,125	0,227	0,136

Tableau 4.2
MPO, MPLSBE, mesures de l'incertitude pour les données du TVSFP (Partie 2)

ID	PS	TV	MPO	$\sqrt{\widehat{\text{EQMP}}}$	MPLSBE	$\sqrt{\widetilde{\text{EQMP}}}$
410	1	1	0,778	0,304	0,813	0,124
411	0	1	0,409	0,195	0,444	0,115
412	1	0	0,913	0,219	0,930	0,126
414	1	0	0,929	0,257	0,941	0,127
415	1	1	0,869	0,199	0,872	0,135
505	1	1	0,790	0,154	0,818	0,136
506	0	1	0,389	0,169	0,428	0,134
507	0	1	0,426	0,148	0,452	0,135
508	0	1	0,411	0,108	0,442	0,136
509	1	0	0,915	0,097	0,929	0,143
510	1	0	0,880	0,119	0,905	0,143
513	0	0	0,185	0,215	0,197	0,123
514	1	1	0,866	0,144	0,870	0,140
515	0	0	0,180	0,102	0,192	0,143

Si l'on compare la MPO au MPLSBE, les valeurs du second sont généralement plus élevées, et les estimations de l'EQMP correspondantes sont en majeure partie plus faibles. Du point de vue de la signification statistique, les résultats du MPLSBE sont significatifs pour les groupes (1,1), (1,0) et (0,1), et non significatifs pour le groupe (0,0). Il convient de souligner que l'estimateur P-R de l'EQMP du MPLSBE est calculé sous l'hypothèse de normalité, alors qu'ici, les données ont clairement une distribution non normale, comme il est mentionné plus haut. Donc, il se peut que la mesure de l'incertitude pour le MPLSBE ne soit pas exacte. En particulier, le fait que les (racines carrées des) EQMP pour les MPLSBE sont plus faibles, comparativement à celles des MPO ne signifie pas nécessairement que les EQMP réelles correspondantes des MPLSBE sont plus faibles que celles des MPO. En fait, nos résultats de simulation (voir la section 2) ont montré l'opposé. Nous constatons aussi que les estimations de l'EQMP des MPLSBE sont plus homogènes dans les divers petits domaines. Cela pourrait tenir au fait que l'estimateur P-R de l'EQMP du MPLSBE est obtenu en supposant que le modèle REE est correct, alors que l'estimateur proposé de l'EQMP pour la MPO ne s'appuie pas sur une telle hypothèse.

En conclusion, malgré les différences possibles entre les caractéristiques des petits domaines, les programmes PS et TV semblaient améliorer les cotes THKS des élèves (savoir si les meilleures cotes THKS signifient que la prévention et l'arrêt du tabagisme sont améliorés est toutefois une autre question). Il semble aussi que le programme PS était relativement plus efficace que le programme TV. Sans l'intervention d'un de ces programmes, la cote THKS ne semblait pas s'améliorer en ce qui concerne les moyennes de petit domaine. Pour ce qui est de la signification statistique des résultats, pour PS = 0 et TV = 0, la cote THKS ne semblait pas être améliorée; pour PS = 1, la cote THKS paraissait être améliorée; et pour PS = 0 et TV = 1, l'amélioration de la cote THKS n'était pas convaincante.

Remerciements

Les travaux de Jiming Jiang sont financés partiellement par les subventions DMS-0809127 et SES-1121794 de la NSF. Les travaux de Thuan Nguyen sont financés partiellement par la subvention SES-1118469 de la NSF. Les travaux de J. Sunil Rao sont financés partiellement par les subventions DMS-0806076 et SES-1122399 de la NSF. La recherche des trois auteurs est financée partiellement par la subvention R01-GM085205A1 du NIH. Les auteurs remercient le professeur Donald Hedeker d'avoir eu l'amabilité de fournir les données du TVSFP pour l'analyse. Enfin, les auteurs remercient le rédacteur associé et deux examinateurs de leurs commentaires.

Annexe

A.1. MPO sous régression à erreurs emboîtées. L'EQMP basée sur le plan de sondage est donnée par (1.6). Notons que toutes les espérances E, et plus tard les prédictions P, sont basées sur le plan de sondage, en supposant un échantillonnage aléatoire simple. Notons que $\mathrm{E}\left\{\tilde{\theta}_i(\psi) - \theta_i\right\}^2 = \mathrm{E}\left\{\tilde{\theta}_i^2(\psi)\right\} - 2\theta_i \mathrm{E}\left\{\tilde{\theta}_i(\psi)\right\} + \theta_i^2$. En outre, notons que $\mathrm{E}(\bar{y}_{i\cdot}) = \theta_i$ et $\mathrm{E}(\bar{x}_{i\cdot}) = \bar{X}_i$ ($\bar{y}_{i\cdot}$ et $\bar{x}_{i\cdot}$ sont les estimateurs sans biais sous le plan des moyennes de sous-population correspondantes). Donc, nous avons

$$\begin{aligned}\mathrm{E}\left\{\tilde{\theta}_i(\psi)\right\} &= \bar{X}_i'\beta + \left\{\frac{n_i}{N_i} + \left(1 - \frac{n_i}{N_i}\right)\frac{n_i\sigma_v^2}{\sigma_e^2 + n_i\sigma_v^2}\right\}(\theta_i - \bar{X}_i'\beta) \\ &= \left(1 - \frac{n_i}{N_i}\right)\frac{\sigma_e^2}{\sigma_e^2 + n_i\sigma_v^2}\bar{X}_i'\beta + \left\{\frac{n_i}{N_i} + \left(1 - \frac{n_i}{N_i}\right)\frac{n_i\sigma_v^2}{\sigma_e^2 + n_i\sigma_v^2}\right\}\theta_i.\end{aligned}$$

Par conséquent, en utilisant la notation présentée sous (1.7), nous avons

$$\mathrm{E}\left\{\tilde{\theta}_i(\psi) - \theta_i\right\}^2 = \mathrm{E}\left\{\tilde{\theta}_i^2(\psi)\right\} - 2\frac{1 - r_i}{1 + n_i\gamma}\bar{X}_i'\beta\theta_i + b_i(\gamma)\theta_i^2. \tag{A.1}$$

Nous pouvons exprimer le paramètre θ_i inconnu dans (A.1) par $\mathrm{E}(\bar{y}_{i\cdot})$. Nous avons également besoin d'un estimateur sans biais sous le plan de θ_i^2, qui est donné par (1.8). Autrement dit, nous avons $\theta_i^2 = \mathrm{E}\left(\hat{\mu}_i^2\right)$. Pour montrer l'absence de biais sous le plan de (1.8), notons que

$$\begin{aligned}\mathrm{E}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}^2\right) &= \frac{1}{n_i}\mathrm{E}\left\{\sum_{k=1}^{N_i} Y_{ik}^2 1_{(k\in I_i)}\right\} \\ &= \frac{1}{n_i}\sum_{k=1}^{N_i} Y_{ik}^2 \mathrm{P}(k \in I_i) = \frac{1}{N_i}\sum_{k=1}^{N_i} Y_{ik}^2,\end{aligned}$$

où I_i est l'ensemble d'indices échantillonnés correspondant au i^{e} petit domaine. En outre, nous avons

$$\begin{aligned}\mathrm{E}\left\{\frac{N_i - 1}{N_i(n_i - 1)}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2\right\} &= \frac{N_i - 1}{N_i(n_i - 1)}\mathrm{E}\left(\sum_{j=1}^{n_i} y_{ij}^2 - n_i\bar{y}_{i\cdot}^2\right) \\ &= \frac{N_i - 1}{N_i(n_i - 1)}\mathrm{E}\left(\sum_{j=1}^{n_i} y_{ij}^2\right) - \frac{(N_i - 1)n_i}{N_i(n_i - 1)}\mathrm{E}\left(\bar{y}_{i\cdot}^2\right) \\ &= \frac{(N_i - 1)n_i}{N_i(n_i - 1)}\left\{\frac{1}{N_i}\sum_{k=1}^{N_i} Y_{ik}^2 - \mathrm{E}\left(\bar{y}_{i\cdot}^2\right)\right\},\end{aligned}$$

et

$$\begin{aligned}\mathrm{E}\left(\bar{y}_{i\cdot}^2\right) &= \frac{1}{n_i^2}\mathrm{E}\left\{\sum_{k=1}^{N_i} Y_{ik}1_{(k\in I_i)}\right\}^2 \\ &= \frac{1}{n_i^2}\sum_{k,l=1}^{N_i} Y_{ik}Y_{il}\mathrm{P}(k \in I_i, l \in I_i) \\ &= \frac{1}{n_i^2}\left\{\sum_{k=1}^{N_i} Y_{ik}^2\frac{n_i}{N_i} + \sum_{k\neq l} Y_{ik}Y_{il}\frac{n_i(n_i - 1)}{N_i(N_i - 1)}\right\} \\ &= \frac{1}{n_i^2}\left[\frac{n_i}{N_i}\sum_{k=1}^{N_i} Y_{ik}^2 + \frac{n_i(n_i - 1)}{N_i(N_i - 1)}\left\{\left(\sum_{k=1}^{N_i} Y_{ik}\right)^2 - \sum_{k=1}^{N_i} Y_{ik}^2\right\}\right] \\ &= \frac{1}{n_i^2}\left\{\frac{n_i(N_i - n_i)}{N_i(N_i - 1)}\sum_{k=1}^{N_i} Y_{ik}^2 + \frac{N_i n_i(n_i - 1)}{N_i - 1}\theta_i^2\right\} \\ &= \frac{N_i - n_i}{N_i(N_i - 1)n_i}\sum_{k=1}^{N_i} Y_{ik}^2 + \frac{N_i(n_i - 1)}{(N_i - 1)n_i}\theta_i^2.\end{aligned}$$

Donc, après combinaison des éléments, nous obtenons

$$\mathrm{E}\left(\hat{\mu}_i^2\right) = \left[1 - \frac{(N_i - 1)\, n_i}{N_i\,(n_i - 1)}\left\{1 - \frac{N_i - n_i}{(N_i - 1)\, n_i}\right\}\right]\left(\frac{1}{N_i}\sum_{k=1}^{N_i} Y_{ik}^2\right) + \theta_i^2 \;=\; \theta_i^2.$$

Il s'ensuit que le deuxième membre de (A.1) peut être exprimé sous la forme

$$\mathrm{E}\left[\sum_{i=1}^{m}\left\{\tilde{\theta}_i^2\,(\psi) - 2\frac{1 - r_i}{1 + n_i\gamma}\,\bar{\mathbf{X}}_i'\beta\bar{y}_{i\cdot} + b_i\,(\gamma)\,\hat{\mu}_i^2\right\}\right].$$

Le MEP s'obtient en minimisant l'expression à l'intérieur de l'espérance, qui correspond à (1.7).

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association,* 83, 401, 28-36.

Chatterjee, S., Lahiri, P. et Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics,* 36, 3, 1221-1245.

Datta, G.S., Kubokawa, T., Molina, I. et Rao, J.N.K. (2011). Estimation of mean squared error of model-based small area estimators. *Test,* 20, 367-388.

Efron, B. (1979). Bootstrap method: Another look at the jackknife. *The Annals of Statistics,* 7, 1, 1-26.

Efron, B., et Tibshirani, R.J. (1993). *An Introduction to the Bootstrap,* Chapman & Hall/CRC.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 366a, 269-277.

Hall, P., et Maiti, T. (2006). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics,* 34, 4, 1733-1750.

Hedeker, D., Gibbons, R.D. et Flay, B.R. (1994). Random-effects regression models for clustered data with an example from smoking prevention research. *Journal of Consulting and Clinical Psychology,* 62, 4, 757-765.

Jiang, J. (1996). REML estimation: Asymptotic behavior and related topics. *The Annals of Statistics,* 24, 1, 255-286.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications,* New York : Springer.

Jiang, J., et Nguyen, T. (2012). Small area estimation via heteroscedastic nested-error regression. *The Canadian Journal of Statistics/La revue canadienne de statistique,* 40, 3, 588-603.

Jiang, J., Lahiri, P. et Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with *M* - estimation. *The Annals of Statistics,* 30, 6, 1782-1810.

Jiang, J., Nguyen, T. et Rao, J.S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association,* 106, 494, 732-745.

Lahiri, P. (2012). Estimation of average design-based mean squared error of synthetic small area estimators. Présenté au *40th Annual Meeting of the Statistical Society of Canada,* Guelph, ON.

Nandram, B., et Sun, Y. (2012). A Bayesian model for small area under heterogeneous sampling variances. Rapport technique.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association,* 85, 409, 163-171.

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

Searle, S.R., Casella, G. et McCulloch, C.E. (1992). *Variance Components,* New York : John Wiley & Sons, Inc.

Torabi, M., et Rao, J.N.K. (2012). Estimation of mean squared error of model-based estimators of small area means under a nested error linear regression model. Rapport technique.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 59-79
Statistique Canada, n° 12-001-X au catalogue

Une méthode de détermination du seuil pour la winsorisation avec application à l'estimation pour des domaines

Cyril Favre Martinoz, David Haziza et Jean-François Beaumont[1]

Résumé

Dans les enquêtes auprès des entreprises, il est courant de collecter des variables économiques dont la distribution est fortement asymétrique. Dans ce contexte, la winsorisation est fréquemment utilisée afin de traiter le problème des valeurs influentes. Cette technique requiert la détermination d'une constante qui correspond au seuil à partir duquel les grandes valeurs sont réduites. Dans cet article, nous considérons une méthode de détermination de la constante qui consiste à minimiser le plus grand biais conditionnel estimé de l'échantillon. Dans le contexte de l'estimation pour des domaines, nous proposons également une méthode permettant d'assurer la cohérence entre les estimations winsorisées calculées au niveau des domaines et l'estimation winsorisée calculée au niveau de la population. Les résultats de deux études par simulation suggèrent que les méthodes proposées conduisent à des estimateurs winsorisés ayant de bonnes propriétés en termes de biais et d'efficacité relative.

Mots-clés : Biais conditionnel; estimation robuste; estimateur winsorisé; valeurs influentes.

1 Introduction

Dans les enquêtes auprès des entreprises, il est courant de collecter des variables économiques dont la distribution est fortement asymétrique. Dans ce contexte, on est souvent confronté à la présence de valeurs influentes dans l'échantillon tiré. Ces dernières sont habituellement de très grandes valeurs dont la présence dans l'échantillon tend à rendre les estimateur classiques très instables.

Il est possible de se prémunir contre l'impact des valeurs influentes à l'étape du plan de sondage en sélectionnant d'office les unités potentiellement influentes. Par exemple, dans les enquêtes auprès des entreprises, il est de coutume d'utiliser un plan stratifié aléatoire simple sans remise comportant une ou plusieurs strates exhaustives composées habituellement des grandes unités. Malheureusement, il est rarement possible d'éliminer complètement le problème des valeurs influentes à l'étape du plan de sondage. En effet, les strates dans les enquêtes auprès des entreprises sont habituellement formées au moyen d'une variable géographique, d'une variable de taille (par exemple, le nombre d'employés) et d'une variable de classification (par exemple, le code SCIAN : Système de Classification des Industries de l'Amérique du Nord). Dans une enquête recueillant des dizaines de variables d'intérêt, il n'est pas improbable que certaines d'entre elles soient peu ou pas liées aux variables de stratification, pouvant alors conduire à la présence de valeurs influentes. C'est le cas notamment dans les enquêtes environnementales menées par Statistique Canada telles que l'Enquête sur l'eau dans l'agriculture dont l'un des objectifs consiste à quantifier la quantité d'eau utilisée par les fermes canadiennes pour l'irrigation. Il s'avère que la consommation d'eau utilisée une année donnée est peu liée aux variables de stratification car la consommation est en partie expliquée par les conditions météorologiques subies par les fermes échantillonnées. Un deuxième exemple est l'Enquête sur l'eau dans les industries dont l'un des objectifs

1. Cyril Favre Martinoz, Laboratoire de Statistique d'Enquête, CREST/ENSAI & IRMAR (UMR 6625), Campus de Ker Lann, 35170 Bruz, France; David Haziza, Département de mathématiques et statistique, Université de Montréal, Montréal, Canada, H3C 3J7 et Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France. Courriel : haziza@dms.umontreal.ca; Jean-François Beaumont, Division de la recherche et de l'innovation en statistique, Statistique Canada, Ottawa, Canada, K1A 0T6.

est de quantifier la quantité d'eau consommée. Dans le cas d'entreprises minières, la consommation d'eau utilisée pour extraire le minerai est fortement liée aux caractéristiques géophysiques du terrain qui ne sont pas prises en compte par les variables de stratification.

Un autre problème conduisant à la présence de valeurs influentes dans l'échantillon est celui des migrants interstrates plus connus sous le nom de "stratum jumpers" en anglais qui survient lorsque l'information de stratification recueillie sur le terrain est différente de celle disponible sur la base de sondage. Ces différences sont habituellement dues à des erreurs dans la base de sondage (par exemple, dans le cas d'une base obsolète). Un *stratum jumper* est une unité qui n'appartient pas à la strate à laquelle elle aurait dû appartenir si l'information sur la base de sondage avait été correcte. Si une unité avec une grande valeur est assignée à une strate non-exhaustive, elle combinera alors une grande valeur de la variable d'intérêt et éventuellement un grand poids de sondage, ce qui la rendra potentiellement très influente. En pratique, il n'est pas rare d'observer entre 5 % et 10 % de migrants interstrates.

Les estimateurs classiques (par exemple, l'estimateur par dilatation) ne présentent (pratiquement) pas de biais mais ils peuvent être très instables en présence de valeurs influentes. Les estimateurs robustes sont construits de manière à limiter l'impact des valeurs influentes, ce qui conduit à des estimateurs plus stables mais potentiellement biaisés. L'objectif consiste à développer des procédures d'estimation robustes dont l'erreur quadratique moyenne est significativement inférieure à celle des estimateurs classiques en présence de valeurs influentes dans la population mais qui ne souffrent pas d'une perte d'efficacité importante en leur absence. Le traitement des valeurs influentes permet donc habituellement d'obtenir un compromis entre le biais et la variance.

La winsorisation est une méthode souvent employée dans les enquêtes auprès des entreprises pour traiter les valeurs influentes. Cette méthode consiste à réduire la valeur et/ou le poids d'une ou plusieurs unités influentes afin de réduire leur impact. On considère deux formes de winsorisation : la winsorisation standard ainsi que la winsorisation décrite par Dalén (1987) et Tambay (1988). Ces méthodes sont décrites dans la section 4. Quel que soit le type utilisé, la winsorisation requiert la détermination d'une constante qui correspond au seuil à partir duquel les grandes valeurs sont réduites. Le choix de cette constante est crucial, un mauvais choix pouvant conduire à des estimateurs winsorisés ayant une erreur quadratique moyenne supérieure à celle des estimateurs classiques. Le choix de la constante a été étudié, entre autres, par Kokic et Bell (1994), Rivest et Hurtubise (1995). Dans le cas d'un plan stratifié aléatoire simple sans remise, ces auteurs ont déterminé la constante qui minimise l'erreur quadratique moyenne estimée des estimateurs winsorisés. Dans le cas d'enquêtes répétées, ils suggèrent d'utiliser les données historiques collectées à des occasions précédentes. Kokic et Bell (1994) ont déterminé la valeur optimale de la constante en postulant un modèle de moyenne commune dans chaque strate et en minimisant l'erreur quadratique moyenne de l'estimateur winsorisé calculée par rapport au modèle et au plan de sondage. Clark (1995) a généralisé les résultats de Kokic et Bell (1994) au cas d'un estimateur par le ratio et en calculant l'erreur quadratique moyenne par rapport au modèle seulement.

Dans un premier temps, nous considérons un critère différent qui consiste à trouver la constante qui minimise le plus grand biais conditionnel estimé de l'échantillon. Comme nous l'expliquons à la section 2, le biais conditionnel associé à une unité est une mesure d'influence qui tient compte du plan de sondage utilisé. La méthode proposée a l'avantage d'être simple à implémenter en pratique. De plus, contrairement aux méthodes proposées dans la littérature, elle ne requiert pas d'information historique ou un modèle

décrivant la distribution de la variable d'intérêt dans chaque strate. L'estimation robuste basée sur le biais conditionnel est présentée à la section 3.

Dans la section 5, nous traitons du problème de l'estimation pour des domaines, qui est un problème important en pratique. Il s'agira d'appliquer une méthode robuste séparément dans chaque domaine d'intérêt. Un estimateur au niveau de la population peut être simplement obtenu en agrégeant les estimateurs robustes obtenus dans chacun des domaines. Cependant, étant défini comme la somme d'estimateurs tous biaisés, l'estimateur agrégé risque de présenter un biais important. Ce point a été soulevé par Rivest et Hidiroglou (2004). Nous proposons une approche en trois étapes : d'abord, on appliquera une méthode robuste séparément dans chaque domaine d'intérêt afin d'obtenir des estimations initiales. Indépendamment, on obtiendra une estimation robuste initiale au niveau de la population. Finalement, en utilisant une méthode s'apparentant au calage (e.g., Deville et Särndal 1992), on modifiera les estimations initiales de manière à garantir la cohérence entre les estimations robustes obtenues au niveau des domaines et l'estimation robuste obtenue au niveau de la population. Le problème de la cohérence pour des domaines a été étudié dans un contexte d'estimation pour des petits domaines; voir par exemple, You, Rao et Dick (2004) et Datta, Gosh, Steorts et Maple (2011).

Nous terminons cette section par une discussion du concept de robustesse rencontré en statistique classique et celui rencontré en population finie. En statistique classique, on est en présence de populations infinies dont on cherche, par exemple, à estimer la moyenne. Dans ce contexte, une valeur aberrante est une valeur qui a été générée selon un modèle différent de celui qui a généré la majorité des observations. La présence de valeurs aberrantes dans l'échantillon peut s'expliquer par le fait que la population dont est générée l'échantillon est un mélange de distributions ou encore que certaines observations sont sujettes à des erreurs de mesure. En statistique classique, on cherche habituellement à mener des inférences sur la population des valeurs non aberrantes. Le but est donc de construire des estimateurs robustes au sens où ces derniers sont peu affectés par la présence de données aberrantes dans l'échantillon. Dans ce contexte, il est désirable de construire des estimateurs robustes ayant un point de rupture élevé et/ou une fonction d'influence bornée. En population finie, les erreurs de mesure sont corrigées à l'étape de la vérification si bien que l'on suppose qu'il n'en reste plus à l'étape de l'estimation. Le but est de mener une inférence sur la population "totale" qui comprend les valeurs aberrantes ainsi que les valeurs non aberrantes. Autrement dit, contrairement à la statistique classique, on ne s'intéresse pas qu'à la population de valeurs non aberrantes. Dans ce contexte, des estimateurs affichant un point de rupture élevé et/ou une fonction d'influence bornée ne sont généralement pas appropriés car ils peuvent conduire à des biais importants. On privilégiera des estimateurs robustes au sens où (i) ils sont plus stables que les estimateurs usuels en présence de valeurs influentes et, en l'absence de ces dernières, s'avèrent presque aussi efficaces que les estimateurs classiques et (ii) ils convergent vers les estimateurs classiques à mesure que la taille de l'échantillon et celle de la population augmentent. Des études par simulation sont présentées à la section 6. La section 7 conclut avec une discussion.

2 Mesure d'influence : le biais conditionnel

On se place dans le cadre d'une population finie d'individus, notée U, de taille N. On souhaite estimer le total de la variable d'intérêt y, noté $t = \sum_{i \in U} y_i$. De la population, on tire un échantillon S,

de taille (espérée) n selon un plan de sondage $p(S)$. Un estimateur classique de t est l'estimateur par dilatation, aussi appelé estimateur de Horvitz-Thompson, $\hat{t} = \sum_{i \in S} d_i y_i$, où $d_i = 1/\pi_i$ désigne le poids de sondage de l'unité i et π_i désigne sa probabilité d'inclusion dans l'échantillon. Bien que l'estimateur par dilatation, $\hat{t}$, soit sans biais pour t par rapport au plan de sondage, il peut présenter une grande instabilité en présence de valeurs influentes.

Afin de quantifier l'impact (ou l'influence) d'une unité échantillonnée sur l'estimateur par dilatation, nous utilisons le concept de biais conditionnel d'une unité ; voir Moreno-Rebollo, Muñoz-Reyez et Muñoz-Pichardo (1999), Moreno-Rebollo, Muñoz-Reyez, Jimenez-Gamero et Muñoz-Pichardo (2002) et Beaumont, Haziza et Ruiz-Gazen (2013). Soit I_i la variable indicatrice de sélection dans l'échantillon pour l'unité i telle que $I_i = 1$ si $i \in S$ et $I_i = 0$, sinon. Le biais conditionnel de l'estimateur $\hat{t}$ associé à une unité échantillonnée est défini selon

$$B_{1i}^{\mathrm{HT}} = E_p\left(\hat{t} \,|I_i = 1\right) - t = \sum_{j \in U}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) y_j, \tag{2.1}$$

où π_{ij} désigne la probabilité d'inclusion conjointe des unités i et j dans l'échantillon. Le biais conditionnel (2.1) est, en général, inconnu puisque les valeurs de la variable d'intérêt ne sont observées que pour les unités dans l'échantillon. En pratique, il s'agira de l'estimer. Nous considérons l'estimateur conditionnellement sans biais (voir, par exemple, Beaumont et coll. 2013) :

$$\begin{aligned} \hat{B}_{1i}^{\mathrm{HT}} &= \sum_{j \in S}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}}\right) y_j \\ &= (d_i - 1) y_i + \sum_{j \in S, j \neq i}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}}\right) y_j. \end{aligned} \tag{2.2}$$

Cet estimateur est conditionnellement sans biais au sens où $E_p\left(\hat{B}_{1i}^{\mathrm{HT}} \,|I_i = 1\right) = B_{1i}^{\mathrm{HT}}$. Nous faisons les remarques suivantes à propos du biais conditionnel et de son estimateur : (i) Le biais conditionnel (2.1) et son estimateur (2.2) dépendent des probabilités d'inclusion π_i et des probabilités d'inclusion conjointes π_{ij}. Autrement dit, le biais conditionnel est une mesure qui tient compte du plan de sondage. (ii) Si $\pi_i = 1$ alors $B_{1i}^{\mathrm{HT}} = 0$ de même que $\hat{B}_{1i}^{\mathrm{HT}} = 0$. En effet, lorsque $\pi_i = 1$, l'unité i est selectionnée dans tous les échantillons possibles et, par conséquent, $E_p\left(\hat{t} \,|I_i = 1\right) - t = E_p\left(\hat{t}\right) - t = 0$, l'estimateur $\hat{t}$ étant un estimateur sans biais de t par rapport au plan. Une unité sélectionnée d'office dans l'échantillon n'a donc aucune influence et ne contribue pas à la variance de l'estimateur $\hat{t}$. (iii) Le biais conditionnel estimé (2.2) dépend des probabilités d'inclusion d'ordre deux, π_{ij}. Pour certains plans de sondage, il peut s'avérer ardu d'obtenir ces probabilités, auquel cas on aura recours à des approximations. Pour des plans de sondage appartenant à la classe des plans à grande entropie (e.g., Berger 1998), plusieurs approximations des probabilités d'inclusion d'ordre deux ont été proposées dans la littérature; voir, par exemple, Haziza, Mecatti et Rao (2008). Une solution alternative consiste à obtenir des approximations des π_{ij} au moyen de méthodes Monte Carlo; voir Fattorini (2006) et Thompson et Wu (2008).

Pour un plan stratifié aléatoire simple, le biais conditionnel (2.1) associé à l'unité échantillonnée i dans la strate h est donné par

$$B_{1i}^{\mathrm{HT}} = \frac{N_h}{N_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_i - \bar{y}_{Uh}), \tag{2.3}$$

où n_h désigne la taille de l'échantillon tiré dans la strate $h, \bar{y}_{Uh} = N_h^{-1}\sum_{i\in U_h} y_i$, et U_h désigne la population des unités dans la strate h de taille $N_h, h = 1, \ldots, H$. L'estimateur du biais conditionnel (2.2) se réduit à

$$\hat{B}_{1i}^{\mathrm{HT}} = \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_i - \bar{y}_{Sh}),$$

où $\bar{y}_{Sh} = n_h^{-1}\sum_{i\in S_h} y_i$ et S_h est l'échantillon de la strate h.

Pour un plan de Poisson, le biais conditionnel de l'unité échantillonnée i est donné par

$$B_i^{\mathrm{HT}}(I_i = 1) = (d_i - 1)\, y_i. \tag{2.4}$$

Contrairement au plan aléatoire simple sans remise, le biais conditionnel (2.4) est connu pour toutes les unités de l'échantillon car il ne dépend pas de paramètres de la population finie.

3 Estimation robuste basée sur le biais conditionnel

Afin de se prémunir contre l'influence indue de certaines unités, il convient de construire des estimateurs robustes du total t ; c'est-à-dire des estimateurs qui réduisent l'impact des unités les plus influentes. Nous considérons une classe d'estimateurs de la forme

$$\hat{t}_R = \hat{t} + \Delta, \tag{3.1}$$

où Δ est une certaine variable aléatoire. Comme nous le verrons à la section 4, les estimateurs winsorisés considérés peuvent s'écrire sous la forme (3.1). Comme dans Beaumont et coll. (2013), on désire déterminer la valeur de Δ qui minimise le plus grand biais conditionnel estimé dans l'échantillon de l'estimateur $\hat{t}_R$. Formellement, on cherche la valeur de Δ qui minimise

$$\max_{i\in S}\left\{\left|\hat{B}_{1i}^{R}\right|\right\}, \tag{3.2}$$

où $\hat{B}_{1i}^{R}$ désigne le biais conditionnel estimé de l'estimateur $\hat{t}_R$ associé à l'unité échantillonnée i. Ce biais conditionnel est donné par

$$\begin{aligned} B_{1i}^{R} &= E_p\left(\hat{t}_R \,|\, I_i = 1\right) - t \\ &= B_{1i}^{\mathrm{HT}} + E_p\left(\Delta \,|\, I_i = 1\right) \end{aligned} \tag{3.3}$$

que l'on estimera par

$$\hat{B}_{1i}^{R} = \hat{B}_{1i}^{\mathrm{HT}} + \Delta, \tag{3.4}$$

où $\hat{B}_{1i}^{\mathrm{HT}}$ est un estimateur conditionnellement sans biais de B_{1i}^{HT}. En notant que Δ est un estimateur conditionnellement sans biais de $E_p(\Delta|I_i = 1)$, il découle que l'estimateur du biais conditionnel (3.4) est conditionnellement sans biais pour B_{1i}^R. Autrement dit, on a $E_p\left\{\hat{B}_{1i}^R|I_i = 1\right\} = B_{1i}^R$.

Beaumont et coll. (2013) ont montré que la valeur de Δ qui minimise (3.2) est donnée par

$$\Delta_{\mathrm{opt}} = -\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right),$$

où $\hat{B}_{\min} = \min_{i\in S}\left(\hat{B}_{1i}^{\mathrm{HT}}\right)$ et $\hat{B}_{\max} = \max_{i\in S}\left(\hat{B}_{1i}^{\mathrm{HT}}\right)$. L'estimateur (3.1) devient alors :

$$\hat{t}_R = \hat{t} - \frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right). \tag{3.5}$$

Sous certaines conditions de régularité, Beaumont et coll. (2013) ont montré que l'estimateur (3.5) est convergent par rapport au plan de sondage ; i.e., $\hat{t}_R - t = O_p\left(N/\sqrt{n}\right)$.

4 Application aux estimateurs winsorisés

L'estimateur (3.5) peut être écrit sous d'autres formes, ce qui peut parfois faciliter sa mise en oeuvre. Nous considérons la forme winsorisée. Cette forme a été souvent étudiée dans la littérature. Tel que mentionné à la section 1, on distingue la winsorisation standard de la winsorisation de Dalén-Tambay.

La winsorisation standard consiste à réduire la valeur des unités dépassant un certain seuil en tenant compte de leur poids. Soit $\tilde{y}_i$ la valeur de la variable y pour l'unité i après winsorisation. On a

$$\tilde{y}_i = \begin{cases} y_i & \text{si } d_i y_i \le K \\ \dfrac{K}{d_i} & \text{si } d_i y_i > K \end{cases} \tag{4.1}$$

où $K > 0$ est le seuil de winsorisation. L'estimateur winsorisé standard du total t est donné par

$$\begin{aligned} \hat{t}_s &= \sum_{i\in S} d_i \tilde{y}_i \\ &= \hat{t} + \Delta(K), \end{aligned} \tag{4.2}$$

où

$$\Delta(K) = -\sum_{i\in S} \max\left(0, d_i y_i - K\right).$$

L'estimateur (4.2) peut donc s'écrire sous la forme (3.1). Une écriture alternative consiste à exprimer $\hat{t}_s$ comme une somme pondérée des valeurs initiales au moyen de poids modifiés :

$$\hat{t}_s = \sum_{i\in S} \tilde{d}_i y_i,$$

où

$$\tilde{d}_i = d_i \frac{\min\left(y_i, \frac{K}{d_i}\right)}{y_i}. \tag{4.3}$$

Si $\min(y_i, K/d_i) = y_i$ (c'est-à-dire que l'unité i n'est pas influente), alors $\tilde{d}_i = d_i$. Le poids d'une unité non influente n'est donc pas modifié. Par contre, le poids modifié d'une unité influente est inférieur à d_i et peut même être inférieur à 1. Il convient de noter qu'une unité affichant une valeur $y_i = 0$ ne pose pas de problème particulier puisque sa contribution au total estimé, $\hat{t}_s$, est nulle. Dans ce cas, on peut assigner une valeur arbitraire au poids modifié $\tilde{d}_i$.

Dans le cas de la winsorisation de Dalén-Tambay, on définit les valeurs de la variable d'intérêt après winsorisation par

$$\tilde{y}_i = \begin{cases} y_i & \text{si } d_i y_i \leq K \\ \frac{K}{d_i} + \frac{1}{d_i}\left(y_i - \frac{K}{d_i}\right) & \text{si } d_i y_i > K \end{cases}. \tag{4.4}$$

Cela conduit à l'estimateur winsorisé du total t_y :

$$\begin{aligned} \hat{t}_{\mathrm{DT}} &= \sum_{i \in S} d_i \tilde{y}_i. \\ &= \hat{t} + \Delta(K), \end{aligned} \tag{4.5}$$

où

$$\Delta(K) = -\sum_{i \in S} \frac{(d_i - 1)}{d_i} \max(0, d_i y_i - K).$$

L'estimateur (4.5) peut également s'écrire sous la forme (3.1). Comme pour $\hat{t}_s$, une écriture alternative consiste à exprimer $\hat{t}_{\mathrm{DT}}$ comme une somme pondérée des valeurs initiales au moyen de poids modifiés :

$$\hat{t}_{\mathrm{DT}} = \sum_{i \in S} \tilde{d}_i y_i,$$

où

$$\tilde{d}_i = 1 + (d_i - 1) \frac{\min\left(y_i, \frac{K}{d_i}\right)}{y_i}. \tag{4.6}$$

Comme pour l'estimateur winsorisé standard, le poids d'une unité non-influente n'est pas modifié. Contrairement à la winsorisation standard, la winsorisation de Dalén-Tambay garantit que les poids modifiés ne peuvent être inférieurs à 1. Encore une fois, une unité affichant une valeur $y_i = 0$ ne pose pas de problème particulier puisque sa contribution au total estimé, $\hat{t}_{\mathrm{DT}}$, est nulle. Dans ce cas, on peut assigner une valeur arbitraire au poids modifié $\tilde{d}_i$.

Les estimateur winsorisés standard et de Dalén-Tambay étant de la forme (3.1), la constante optimale K_{opt} qui minimise (3.2) est obtenue en résolvant

$$\Delta(K) = -\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right)$$

ou encore

$$\sum_{j \in S} a_j \max\left(0, d_j y_j - K\right) = \frac{\hat{B}_{\min} + \hat{B}_{\max}}{2}, \tag{4.7}$$

où $a_j = 1$ dans le cas de l'estimateur $\hat{t}_s$ et $a_j = \left(d_j - 1\right)/d_j$ dans le cas de l'estimateur $\hat{t}_{\text{DT}}$. On montre dans l'annexe qu'une solution à l'équation (4.7) existe sous les conditions suivantes :

1. $\pi_{ij} - \pi_i \pi_j \leq 0$; et
2. $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \geq 0.$

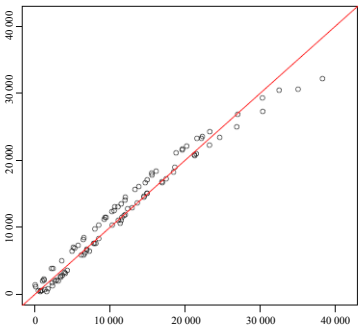
La condition 1 est satisfaite pour la plupart des plans de sondage à un degré utilisés en pratique tels que l'échantillonnage aléatoire simple stratifié et l'échantillonnage de Poisson. La condition 2 implique que $\hat{t}_R$ doit être plus petit ou égal à $\hat{t}$ puisqu'un estimateur winsorisé ne peut pas être plus grand que l'estimateur de Horvitz-Thompson par construction. On s'attend en général à ce que la condition 2 soit satisfaite dans la plupart des populations asymétriques que l'on retrouve dans les enquêtes auprès des entreprises et dans les enquêtes sociales. On montre aussi dans l'annexe que la solution à l'équation (4.7) est unique si les conditions précédentes tiennent et si $y_i \geq 0$ pour $i \in S$. On y décrit brièvement un algorithme pour trouver la solution à l'équation (4.7).

Il est à noter que bien que la valeur K_{opt} diffère selon l'estimateur winsorisé utilisé, les estimateurs robustes résultants sont identiques. Autrement dit, on a

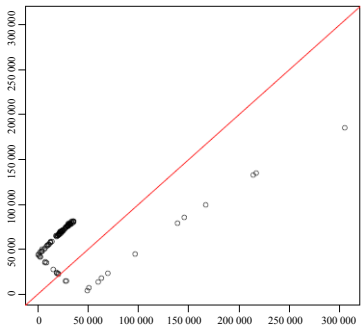
$$\hat{t}_s\left(K_{\text{opt}}\right) = \hat{t}_{\text{DT}}\left(K_{\text{opt}}\right) = \hat{t}_R = \hat{t} - \frac{\hat{B}_{\min} + \hat{B}_{\max}}{2}. \tag{4.8}$$

Afin de comparer l'influence de chacune des unités de la population par rapport à l'estimateur (non-robuste) par dilatation, $\hat{t}$, et sa version robuste (4.8), nous avons effectué une étude par simulation. Pour cela, nous avons généré deux populations, chacune de taille $N = 100$. La première population a été générée selon une loi normale de moyenne 4 108 et d'écart type 1 500 alors que la deuxième a été générée selon une loi log-normale de moyenne 4 108 et d'écart type 7 373. De chaque population, nous avons tiré $M = 500\ 000$ échantillons selon deux plans de sondage : (i) le plan aléatoire simple sans remise de taille $n = 10$ et (ii) le plan de Bernoulli de taille espérée $n = 10$. Dans un premier temps, nous avons calculé le biais conditionnel de l'estimateur de Horvitz-Thompson pour un sondage aléatoire simple sans remise donné en (2.3) ainsi que celui dans le cas d'un plan de Bernoulli donné en (2.4). Notons que le biais conditionnel de l'estimateur Horvitz-Thompson ne requiert pas d'être approximé par simulation puisque toutes les quantités de la population sont connues. Le biais conditionnel de l'estimateur robuste donné en (3.3) associé à l'unité i a été approximé de la manière suivante : parmi les 500 000 échantillons tirés, nous avons identifié les échantillons contenant l'unité i. Dans chacun de ces échantillons, nous avons calculé l'erreur, $\hat{t}_R - t$. Finalement, nous avons calculé la moyenne des valeurs de $\hat{t}_R - t$ sur tous les échantillons contenant l'unité i.

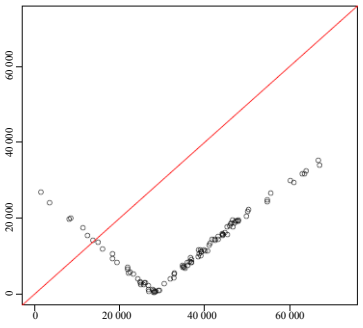
Les figures 4.1 (a) et 4.1 (b) présentent les résultats dans le cas de l'échantillonnage aléatoire simple sans remise pour les distributions normale et log-normale, respectivement. Les figures 4.1 (c) et 4.1 (d) présentent les résultats dans le cas de l'échantillonnage de Bernoulli pour les distributions normale et log-normale, respectivement. Dans chacune des figures, la valeur absolue du biais conditionnel de $\hat{t}_R$ est représentée en fonction de la valeur absolue du biais conditionnel de $\hat{t}$ pour chaque unité de la population. Les unités situées au dessus de la première bissectrice possèdent un biais conditionnel associé à $\hat{t}_R$ en valeur absolue supérieur au biais conditionnel associé à l'estimateur $\hat{t}$ en valeur absolue. Dans un premier temps, nous discutons des résultats ayant trait à l'échantillonnage aléatoire simple sans remise : le biais conditionnel de $\hat{t}_R$ en valeur absolue présente un comportement similaire au biais conditionnel en valeur absolue de $\hat{t}$, ce qui indique que l'influence des unités n'est pas modifiée de manière significative après avoir robustifié l'estimateur par dilatation. Ce résultat n'est pas surprenant puisque la population ne comprend pas d'unités fortement influentes. Dans le cas de la loi log-normale, on constate que l'influence des valeurs affichant un biais conditionnel associé à $\hat{t}$ élevé a été réduite de manière significative. En revanche, on constate que, pour la majorité des données, le biais conditionnel de $\hat{t}_R$ est légèrement plus élevé que celui de $\hat{t}$. Nous discutons maintenant les résultats ayant trait à l'échantillonnage de Bernoulli : dans le cas de la population normale, on constate que l'influence de la grande majorité des unités a été réduite puisque le biais conditionnel en valeur absolue de $\hat{t}_R$ est significativement moins élevé que le biais conditionnel en valeur absolue de $\hat{t}$. Dans le cas de la loi log-normale, les résultats obtenus sont similaires à ceux obtenus dans le cas de l'échantillonnage aléatoire simple sans remise pour la même distribution.



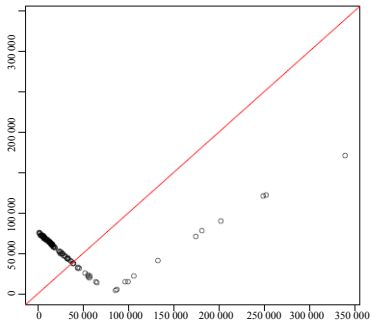
(a) Loi normale avec sondage aléatoire simple



(b) Loi Lognormale avec sondage aléatoire simple



(c) Loi normale avec tirage bernouillien



(d) Loi Lognormale avec tirage bernouillien

Figure 4.1 Représentation de la valeur absolue des biais conditionnels des estimateurs robuste et non robuste

5 Estimation robuste de totaux par domaine

En pratique, on cherche la plupart du temps à obtenir des estimations au niveau de domaines de la population ainsi qu'une estimation au niveau global. Soit $t_g = \sum_{i \in U_g} y_i$ le total de la variable y dans le domaine g. On va supposer que les domaines forment une partition de la population telle que $t = \sum_{i \in U} y_i = \sum_{g=1}^{G} t_g$, où G est le nombre de domaines. Soit S_g l'ensemble des unités échantillonnées dans le domaine g. L'estimateur par dilatation de t_g est donné par $\hat{t}_g = \sum_{i \in S_g} d_i y_i$. On a la relation de cohérence suivante : $\sum_{g=1}^{G} \hat{t}_g = \hat{t}$.

En présence de valeurs influentes, on peut appliquer une procédure robuste séparément pour chacun des domaines à l'aide de la méthode décrite à la section 3, ce qui conduit à G estimateurs robustes, $\hat{t}_{R,g}$. Un estimateur robuste, $\hat{t}_{R(\text{agr})}$, du total au niveau de la population est simplement obtenu en agrégeant les estimateurs robustes $\hat{t}_{R,g}$. On a alors $\hat{t}_{R(\text{agr})} = \sum_{g=1}^{G} \hat{t}_{R,g}$. La relation de cohérence entre les estimations calculées au niveau des domaines et l'estimation calculée au niveau de la population est donc satisfaite. Cependant, agréger G estimateurs robustes, chacun souffrant d'un biais potentiel, peut engendrer un estimateur robuste agrégé, $\hat{t}_{R(\text{agr})}$, fortement biaisé. Dans la grande majorité des cas, le biais de $\hat{t}_{R(\text{agr})}$ sera négatif, chacun des estimateurs $\hat{t}_{R,g}$ présentant un biais négatif.

Une solution permettant d'éviter un estimateur avec un biais inacceptable consiste d'abord à calculer l'estimateur robuste (4.8), $\hat{t}_{R,g}$, pour chacun des domaines. Ensuite, on obtient indépendamment un estimateur robuste du total t dans la population, $\hat{t}_{R,0}$, donné par (4.8). Cependant, dans ce cas, la relation de cohérence n'est plus nécessairement satisfaite. Autrement dit, on aura, $\hat{t}_{R,0} \neq \sum_{g=1}^{G} \hat{t}_{R,g}$, en général. Il s'agira alors de forcer la cohérence entre les estimations robustes dans les domaines et l'estimation robuste agrégée au moyen d'une méthode qui s'apparente au calage. Pour cela, on déterminera des estimations robustes finales $\hat{t}^*_{R,g}, g = 0,1,..,G,$ qui soient aussi proches que possible des estimations robustes initiales $\hat{t}_{R,g}$, au sens d'une certaine fonction de distance, et qui vérifient l'équation de calage

$$\sum_{g=1}^{G} \hat{t}^*_{R,g} = \hat{t}^*_{R,0}. \tag{5.1}$$

Dans le cas d'une fonction de distance de type khi-deux généralisé, on cherche des estimations robustes finales, $\hat{t}^*_{R,g}$, telles que

$$\sum_{g=0}^{G} \frac{\{\hat{t}^*_{R,g} - \hat{t}_{R,g}\}^2}{2q_g \hat{t}_{R,g}} \tag{5.2}$$

est minimum sous la contrainte (5.1). Le coefficient q_g dans l'expression précédente est un poids que l'on assigne à l'estimation initiale dans le domaine g, $\hat{t}_{R,g}$, et s'interprète comme l'importance de celui-ci dans le problème de minimisation. En utilisant la méthode des multiplicateurs de Lagrange, on peut facilement obtenir une solution au problème de minimisation ci-dessus. Cette solution est donnée par :

$$\hat{t}_{R,g}^{*} = \hat{t}_{R,g} - \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \delta_g q_g \hat{t}_{R,g}, \tag{5.3}$$

où $\delta_0 = -1$ et $\delta_g = 1$, pour $g = 1, \ldots, G$.

Nous faisons les remarques suivantes : (i) Si $q_g = 0$, alors l'estimation robuste finale $\hat{t}_{R,g}^{*}$ est identique à l'estimation robuste initiale $\hat{t}_{R,g}$. Ainsi, si l'on souhaite que l'estimation initiale dans le domaine g, ne soit pas trop modifiée, il suffit de lui associer une petite valeur de q_g. Cet aspect sera également illustré empiriquement à la section 6.2. (ii) Notons qu'en plus des estimations robustes initiales au niveau des domaines, $\hat{t}_{R,g}$, pour $g = 1, \ldots, G$, l'estimation robuste initiale au niveau de la population, $\hat{t}_{R,0}$, peut être également être modifiée. (iii) Si $q_0 = 0$ (autrement dit, l'estimation initiale robuste au niveau de la population n'est pas modifiée) et $q_g = q$ pour $g = 1, \ldots, G$, où q est une constante strictement positive, l'expression (5.3) se simplifie pour donner

$$\hat{t}_{R,g}^{*} = \hat{t}_{R,g} \left(\frac{\hat{t}_{R,0}}{\hat{t}_{R(\text{agr})}} \right). \tag{5.4}$$

Dans ce cas, les estimations initiales $\hat{t}_{R,g}$ sont toutes modifiées par le même facteur $\hat{t}_{R,0} / \hat{t}_{R(\text{agr})}$. (iv) Comment fixer les valeurs de q_g en pratique ? Il semble naturel de privilégier le choix suivant :

$$q_g = \widehat{\text{CV}}\left(\hat{t}_g\right) \Big/ \sum_{g=1}^{G} \widehat{\text{CV}}\left(\hat{t}_g\right),$$

où $\widehat{\text{CV}}\left(\hat{t}_g\right)$ désigne le coefficient de variation (CV) estimé associé au domaine g. Par exemple, dans une enquête répétée, il sera possible d'utiliser le CV estimé observé à une occasion précédente. Ce choix de q_g est motivé par le fait qu'on ne cherchera pas à modifier de manière importante l'estimation initiale associée à un domaine caractérisé par un petit CV estimé. Dans un tel domaine, il est clair que le problème des valeurs influentes est moins criant et l'on s'attend à ce que l'estimation robuste initiale $\hat{t}_{R,g}$ soit relativement proche du vrai total t_g. Autrement dit, l'estimateur robuste $\hat{t}_{R,g}$ devrait être peu biaisé et relativement stable. Il est donc naturel de ne pas chercher à modifier l'estimation robuste initiale de manière importante. (v) En (5.2), nous avons utilisé la distance du khi-deux généralisée conduisant à la méthode linéaire. Dans la littérature portant sur le calage (e.g., Deville et Särndal 1992), il existe plusieurs autres méthodes de calage. Mentionnons la distance de Kullback-Leibler conduisant à la méthode exponentielle et les méthodes logit et linéaire tronquée. Les deux dernières méthodes permettent de spécifier des bornes positives C_1 et C_2 telles que $C_1 \leq \hat{t}_{R,g}^{*} / \hat{t}_{R,g} \leq C_2$. Autrement dit, on s'assurera que le rapport $\hat{t}_{R,g}^{*} / \hat{t}_{R,g}$ se situe à l'intérieur des deux limites C_1 et C_2. Notons qu'il est possible que la procédure de calage conduise à $\hat{t}_{R,g}^{*} - \hat{t}_g \geq 0$, pour un certain g, ce qui est contre-intuitif. Dans ce cas, il suffit de rajouter la contrainte $\hat{t}_{R,g}^{*} \leq \hat{t}_g$ pour $g = 1, \ldots, G$, dans la procédure de calage. (vi) Une écriture alternative consiste à exprimer $\hat{t}_{R,g}^{*}$ comme une somme pondérée des valeurs initiales au moyen de poids modifiés :

$$\hat{t}_{R,g}^{*} = \sum_{i \in S_g} \tilde{d}_i^{\,*} y_i,$$

où

$$\tilde{d}_i^{\,*} = \tilde{d}_i \left(1 - \delta_g q_g \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \right)$$

et $\tilde{d}_i$ est donné soit par (4.3) ou par (4.6). On peut également écrire l'estimateur $\hat{t}_{R,g}^{*}$ comme une somme pondérée avec les poids initiaux au moyen de valeurs modifiées :

$$\hat{t}_{R,g}^{*} = \sum_{i \in S_g} d_i \tilde{y}_i^{*},$$

où

$$\tilde{y}_i^{*} = \tilde{y}_i \left(1 - \delta_g q_g \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \right), \quad i \in g$$

et $\tilde{y}_i$ est donné soit par (4.1) ou par (4.4). (vii) On peut vouloir trouver les seuils de winsorisation, $K_g, g = 1, \ldots, G,$ tels que l'estimateur winsorisé standard ou celui de Dalén-Tambay est égal à $\hat{t}_{R,g}^{*}$. On peut procéder de façon similaire à la section 4 et on peut utiliser un algorithme semblable à celui donné dans l'annexe. Une condition nécessaire pour l'existence d'une solution est que $\hat{t}_g - \hat{t}_{R,g}^{*} \geq 0$. (viii) La procédure de calage proposée permet de traiter conjointement plusieurs partitions de la population. Par exemple, on peut, à la fois, être intéressé à publier des estimations par province et des estimations par industrie. Dans ce cas, il suffit de poser les équations de calage suivantes dans la procédure de calage :

$$\sum_{g=1}^{G} \hat{t}_{R,g}^{*} = \hat{t}_{R,0}^{*},$$

$$\sum_{l=1}^{L} \hat{t}_{R,l}^{*} = \hat{t}_{R,0}^{*},$$

où G et L désigne le nombre de provinces et le nombre d'industries, respectivement. De même, la méthode est applicable au cas de plusieurs découpages de la population.

6 Études par simulation

6.1 Winsorisation dans un plan aléatoire simple sans remise

Nous avons effectué une étude par simulation afin d'étudier les propriétés de plusieurs estimateurs robustes au moyen de onze populations. Les dix premières de taille $N = 5\ 000$ consistent en une variable

d'intérêt y. Dans chaque population, les valeurs de la variable y ont été générées selon le modèle suivant :

$$Y_i = U_i + \delta_i V_i,$$

où U_i, δ_i et V_i sont des variables aléatoires dont les distributions sont décrites dans le tableau 6.1. La population 1 a été générée selon une loi normale. Les populations $2-5$ ont été générées au moyen d'un mélange de lois normales avec des taux de contamination variant de 0,5 % à 5 %. Les populations $6-8$ ont été générées selon des lois asymétriques. Les populations 9 et 10 ont été générées au moyen d'un mélange de lois log-normales avec des taux de contamination égaux à 0,5 % et 5 %. Enfin, la onzième population de taille $N = 5\ 000$ est issue de l'enquête sur les technologies de l'information produite par l'Institut National de la Statistique et des Études Économiques (INSEE) en 2011. Un des objectifs de cette enquête est d'estimer le chiffre d'affaires réalisé par commerce électronique pour les entreprises françaises. La variable "chiffre d'affaires" sera donc utilisée dans notre simulation. La distribution de la variable y dans chacune des populations est représentée graphiquement à la figure 6.1. De plus, le tableau 6.2 montre plusieurs statistiques descriptives pour chacune des populations utilisées. Pour des raisons de confidentialité, nous avons omis d'afficher les unités dans le graphique correspondant à la population 11. De même, le tableau 6.2 ne comporte aucune statistique descriptive pour la population 11.

Dans chaque population, nous avons tiré $M = 5\ 000$ échantillons selon un plan aléatoire simple sans remise de taille $n = 100, 300$ et 500. Dans chaque échantillon, nous avons calculé l'estimateur par dilatation $\hat{t}$ et l'estimateur robuste (4.8). Soient $y_{(1)}, \ldots, y_{(n)}$ les valeurs de la variable y rangées par ordre croissant. Nous avons également calculé les estimateurs winsorisés d'ordres 1, 2 et 3, où l'estimateur winsorisé d'ordre p est obtenu en remplaçant les p plus grandes valeurs de l'échantillon par la valeur $y_{(n-p)}, p = 1, 2, 3$. Dans un contexte de statistique classique, Rivest (1994) a montré que l'estimateur winsorisé d'ordre 1 possède de bonnes propriétés en termes d'erreur quadratique moyenne pour une grande classe de distributions asymétriques.

Comme mesure du biais d'un estimateur $\hat{\theta}$, nous avons calculé le biais relatif Monte Carlo (en %) :

$$\mathrm{BR}_{\mathrm{MC}}\left(\hat{\theta}\right) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)}{t} \times 100,$$

où $\hat{\theta}_{(m)}$ désigne l'estimateur $\hat{\theta}$ dans l'échantillon $m, m = 1, \ldots, 5\ 000$. Nous avons également calculé l'efficacité relative des estimateurs robustes relativement à l'estimateur par dilatation, $\hat{t}$:

$$\mathrm{RE}_{\mathrm{MC}}\left(\hat{\theta}\right) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)^2}{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{t}_{(m)} - t\right)^2} \times 100.$$

Les résultats sont présentés dans le tableau 6.3.

Les résultats présentés au tableau 6.3 montrent que l'estimateur winsorisé d'ordre 1 est moins biaisé et généralement plus efficace que les estimateurs winsorisés d'ordres 2 et 3, ce qui est cohérent avec les résultats obtenus par Rivest (1994). Il est intéressant de comparer l'estimateur robuste $\hat{t}_R$ et l'estimateur

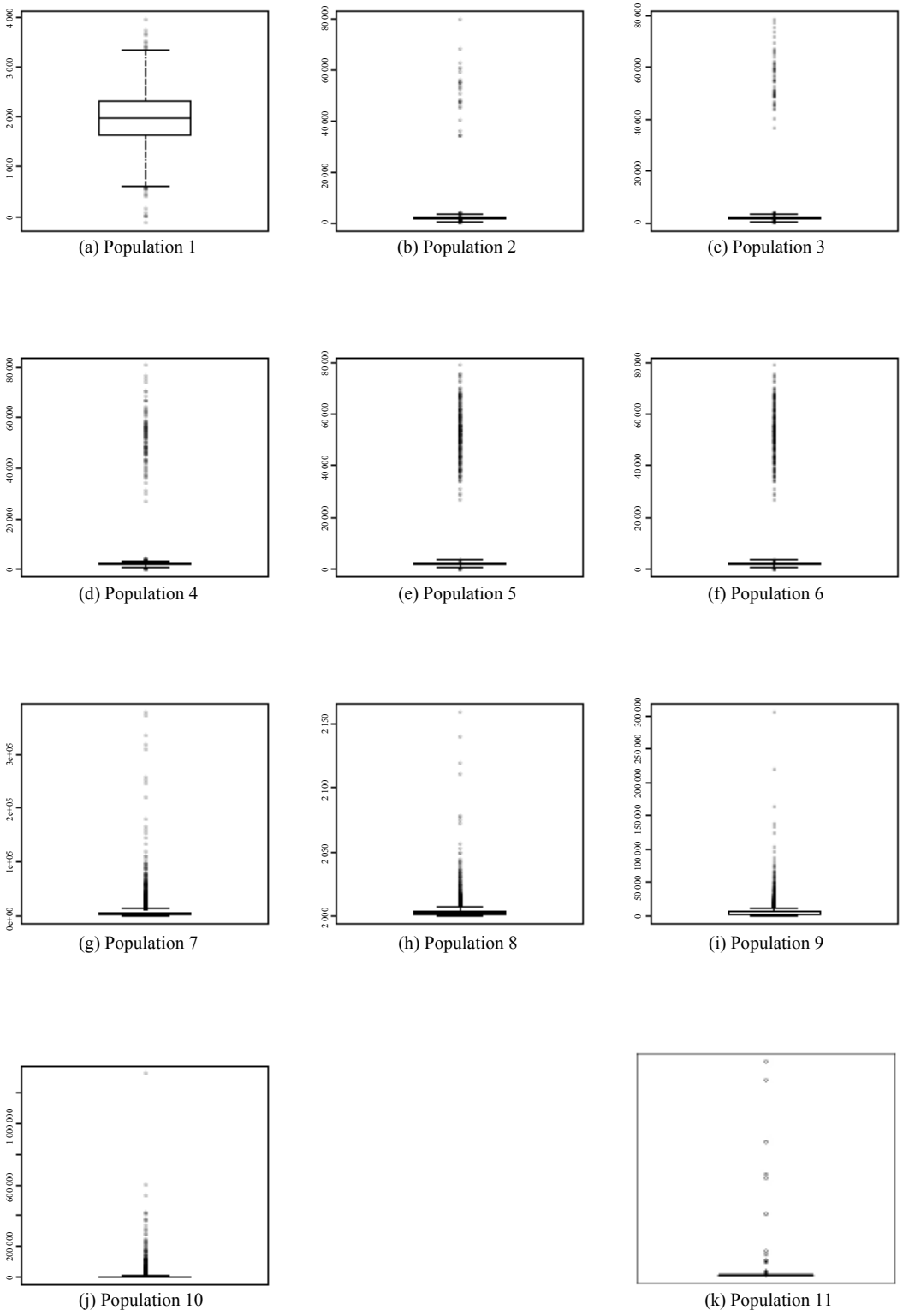
winsorisé d'ordre 1. Dans le cas de la population 1 ne contenant pas de valeurs influentes, on remarque que les deux estimateurs sont peu biaisés et sont aussi efficaces que l'estimateur par dilatation. Dans le cas des populations de mélange de lois normales (populations 2 à 5), nous constatons que l'estimateur winsorisé d'ordre 1 est moins efficace que l'estimateur robuste dans tous les scénarios à l'exception de la population 5 avec $n = 300$. En fait, l'estimateur winsorisé d'ordre 1 est moins efficace que l'estimateur par dilatation dans tous les scénarios à l'exception de la population 2 avec $n = 100$. L'estimateur robuste, quant à lui, affiche un gain en efficacité par rapport à l'estimateur par dilatation sauf pour les populations 4 et 5 pour lesquelles nous observons des valeurs d'efficacité relative variant de 91 % à 102 %. Dans le cas des populations de mélange de lois log-normales (populations 9 et 10), nous constatons que l'estimateur winsorisé d'ordre 1 et l'estimateur robuste présentent des performances très similaires en termes de biais et d'efficacité dans tous les scénarios. Il en va de même pour les populations asymétriques (populations 6 à 8), pour lesquelles les deux estimateurs présentent des résultats similaires. Dans le cas de la population 11, l'estimateur robuste affiche un biais moins élevé que l'estimateur winsorisé d'ordre 1 pour $n = 100$ bien qu'il soit moins efficace (41 % vs. 47 %). Pour $n = 300$ et $n = 500$, l'estimateur robuste est moins biaisé et significativement plus efficace que l'estimateur winsorisé d'ordre 1.

Tableau 6.1
Modèles utilisés afin de générer les populations

Population	Loi de U_i	Mélange	Loi de δ_i	Loi de V_i
1	$\mathcal{N}(2\,000;500)$	Non		
2	$\mathcal{N}(2\,000;500)$	Oui	$\mathcal{B}(0{,}005)$	$\mathcal{N}(50\,000;10\,000)$
3	$\mathcal{N}(2\,000;500)$	Oui	$\mathcal{B}(0{,}01)$	$\mathcal{N}(50\,000;10\,000)$
4	$\mathcal{N}(2\,000;500)$	Oui	$\mathcal{B}(0{,}02)$	$\mathcal{N}(50\,000;10\,000)$
5	$\mathcal{N}(2\,000;500)$	Oui	$\mathcal{B}(0{,}05)$	$\mathcal{N}(50\,000;10\,000)$
6	$\mathcal{L}\text{og} - \mathcal{N}(\log(2\,000);1{,}2)$	Non		
7	$\mathcal{L}\text{og} - \mathcal{N}(\log(2\,000);1{,}5)$	Non		
8	$\mathcal{F}\text{rechet}(2\,000;2{,}5;2{,}1)$	Non		
9	$\mathcal{L}\text{og} - \mathcal{N}(\log(2\,000);1{,}2)$	Oui	$\mathcal{B}(0{,}05)$	$\mathcal{L}\text{og} - \mathcal{N}(\log(5\,000);1{,}2)$
10	$\mathcal{L}\text{og} - \mathcal{N}(\log(2\,000);1{,}2)$	Oui	$\mathcal{B}(0{,}05)$	$\mathcal{L}\text{og} - \mathcal{N}(\log(5\,000);1{,}2)$

Tableau 6.2
Statistiques descriptives des dix populations simulées

Statistique descriptive	Population									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
min	132,3	314,9	105,3	275,9	187,4	23,6	7,6	2 000,9	20,5	26,6
max	3 968	79 506	78 526	80 540	78 690	252 612	379 751	2 159	305 612	$1{,}3\times10^{6}$
$Q1$	1 639	1 667	1 664	1 666	1 685	883	743	200	920	913
Médiane	1 986	1 993	1 997	2 015	2 053	1 996	1 981	2 002	2 167	2 041
$Q3$	2 330	2 337	2 339	2 349	2 421	4 505	5 337	2 004	5 018	4 927
Moyenne	1 985	2 267	2 536	2 976	4 661	4 005	6 118	2 004	4 738	7 883
Écart-type	503	3 709	5 506	7 119	11 470	7 353	17 190	5,89	9 796	33 111
Asymétrie	0,0	14,0	10,2	7,3	4,3	4,2	11,6	11,8	12,1	18,4
Aplatissement	3	209	109	56	20	19	196	228	267	570
CV	0,25	1,6	2,2	2,4	2,5	1,8	2,8	$2{,}9\times10^{-3}$	2,0	4,2



(a) Population 1 (b) Population 2 (c) Population 3

(d) Population 4 (e) Population 5 (f) Population 6

(g) Population 7 (h) Population 8 (i) Population 9

(j) Population 10 (k) Population 11

Figure 6.1 Distribution de la variable d'intérêt dans chacune des onze populations

Tableau 6.3
Biais relatif Monte Carlo (en %) et efficacité relative (entre parenthèses) de plusieurs estimateurs winsorisés

Population	n	$\hat{t}_R$	Winsorisation		
			ordre 1	ordre 2	ordre 3
1	100	-0,1(100)	-0,1(100)	-0,2(101)	-0,3(102)
	300	0,0(100)	-0,0(100)	-0,0(100)	-0,1(100)
	500	0,0(100)	-0,0(100)	-0,0(100)	-0,0(100)
2	100	-4,9(59)	-7,5(87)	-10,7(65)	-11,9(55)
	300	-2,9(87)	-3,0(129)	-6,8(158)	-9,5(169)
	500	-1,9(96)	-1,2(122)	-3,6(175)	-6,5(226)
3	100	-6,9(74)	-8,9(122)	-16,5(119)	-20,0(107)
	300	-3,5(99)	-1,9(122)	-5,6(171)	-10,6(232)
	500	-2,4(102)	-0,9(107)	-2,2(130)	-4,5(186)
4	100	-7,6(91)	-6,2(131)	-15,5(169)	-24,4(194)
	300	-2,9(101)	-0,6(103)	-2,1(118)	-4,4(154)
	500	-2,0(102)	-0,6(102)	-1,1(101)	-1,8(108)
5	100	-5,7(102)	-1,1(104)	-4,1(126)	-9,7(173)
	300	-2,2(102)	-0,4(100)	-0,8(101)	-1,4(102)
	500	-1,2(100)	-0,1(100)	-0,3(100)	-0,5(101)
6	100	-5,7(79)	-5,4(75)	-8,2(80)	-10,6(89)
	300	-2,6(84)	-2,6(79)	-3,9(81)	-5,1(88)
	500	-2,0(86)	-2,0(81)	-3,0(82)	-3,8(88)
7	100	-8,4(72)	-9,3(73)	-14,7(72)	-18,7(79)
	300	-4,5(86)	-4,4(95)	-7,8(91)	-10,2(95)
	500	-3,5(94)	-3,1(105)	-6,0(106)	-8,1(109)
8	100	-0,0(69)	-0,0(75)	-0,0(77)	-0,0(85)
	300	-0,0(82)	-0,0(88)	-0,0(87)	-0,0(95)
	500	-0,0(88)	-0,0(96)	-0,0(94)	-0,0(100)
9	100	-5,7(73)	-5,8(71)	-9,5(72)	-12,4(80)
	300	-3,5(87)	-3,5(85)	-5,4(88)	-6,8(98)
	500	-2,4(88)	-2,4(88)	-3,8(90)	-4,9(97)
10	100	-13,5(68)	-15,0(70)	-24,6(76)	-31,7(89)
	300	-7,5(80)	-7,2(79)	-12,1(85)	-16,3(97)
	500	-5,3(85)	-5,1(83)	-8,4(91)	-11,4(103)
11	100	-22,8(47)	-32,6(41)	-42,0(42)	-47,7(47)
	300	-14,7(65)	-20,0(77)	-29,6(68)	-34,3(75)
	500	-11,3(76)	-14,6(96)	-24,3(90)	-29,3(97)

6.2 Winsorisation dans un plan stratifié aléatoire simple sans remise

Nous avons également testé la méthode de calage décrite dans la section 5. Nous avons généré une population de taille $N = 5\ 000$ que nous avons divisée en cinq strates $U_1, \ldots, U_5$ de taille $N_1, \ldots, N_5$, respectivement; voir tableau 6.4 pour les valeurs de N_h. Dans chacune des strates, nous avons généré une variable y selon une loi log-normale de paramètres $\log(2\ 000)$ et 1,5.

De la population, nous avons tiré $M = 5\ 000$ échantillons selon un plan stratifié aléatoire simple sans remise. Dans la strate U_h, nous avons tiré un échantillon S_h de taille n_h selon un plan aléatoire simple sans remise ; voir le tableau 6.4 pour les tailles n_h et les fractions de sondage, $f_h = n_h / N_h$, correspondantes.

Ici, l'objectif est d'estimer le total dans la population, $t = \sum_{i \in U} y_i$, ainsi que les totaux au niveau des strates $t_h = \sum_{i \in U_h} y_i, h = 1, \ldots, H$. Autrement dit, dans notre exemple, les strates correspondent à des domaines d'intérêt. Les strates formant une partition de la population, on a bien la relation de cohérence, $t = \sum_{h=1}^{H} t_h$. De même, les estimateurs par dilatation satisfont la relation de cohérence : $\hat{t} = \sum_{h=1}^{H} \hat{t}_h$, où $\hat{t} = \sum_{i \in S} d_i y_i$ et $\hat{t}_h = \sum_{i \in S_h} d_i y_i$ avec $d_i = N_h / n_h$ si $i \in U_h$.

Dans chaque échantillon, nous avons, dans un premier temps, calculé l'estimateur robuste (4.8) dans chacune des strates et avons agrégé les estimations robustes obtenues afin d'obtenir une estimation robuste agrégée, $\hat{t}_{R(\text{agr})} = \sum_{h=1}^{H} \hat{t}_{R,h}$. Indépendamment, nous avons calculé l'estimateur robuste (4.8), noté $\hat{t}_{R,0}$, au niveau de la population. Afin de garantir la relation de cohérence (5.1), nous avons effectué un calage tel que décrit dans la section 5 de manière à obtenir les estimations robustes finales $\hat{t}^*_{R,h}, h = 0, \ldots, 5$. Nous avons utilisé quatre jeux de coefficients q_h : (1) $q_0 = 0$ et $q_1 = \ldots = q_5 = 1$; (2) $q_0 = 0$ et $q_h = n_h^{-1}(1 - f_h)$, $h = 1, \ldots, 5$; (3) $q_0 = 0$ et $q_h = \text{CV}(\hat{t}_h) = \sqrt{N_h^2 (1 - f_h) n_h^{-1} S_h^2} / t_h$, où $S_h^2 = (N_h - 1)^{-1} \sum_{i \in U_h} (y_i - \bar{y}_{Uh})^2$, $h = 1, \ldots, 5$; (4) $q_0 = 0$ et $q_h = \widehat{\text{CV}}(\hat{t}_h) = \sqrt{N_h^2 (1 - f_h) n_h^{-1} s_h^2} / \hat{t}_h$, où $s_h^2 = (n_h - 1)^{-1} \sum_{i \in S_h} (y_i - \bar{y}_{Sh})^2$, $h = 1, \ldots, 5$; Nous faisons les remarques suivantes à propos du choix des coefficients q_h : (i) Dans le cas des quatre jeux, nous avons assigné un poids $q_0 = 0$ à l'estimation $\hat{t}_{R,0}$, ce qui revient à ne pas modifier l'estimation robuste au niveau global. Autrement dit, on a $\hat{t}^*_{R,0} = \hat{t}_{R,0}$. (ii) Le premier jeu de poids assigne un poids égal à toutes les strates indépendamment de la taille de l'échantillon ou de la fraction de sondage. (iii) Dans le cas du deuxième jeu, le coefficient q_h est fonction de la taille de l'échantillon n_h et de la fraction de sondage f_h mais il est indépendant de la variabilité intra-strate S_h^2. (iv) Dans les troisième et quatrième jeux, le choix de q_h dépend du vrai CV et du CV estimé, respectivement, pour les raisons mentionnées dans la section 5.

Tableau 6.4
Caractéristiques des strates

Strate	1	2	3	4	5
N_h	2 000	1 500	1 000	400	100
n_h	20	75	100	80	80
f_h	0,01	0,05	0,1	0,2	0,8

Pour chacun des estimateurs robustes, nous avons calculé le biais relatif Monte Carlo (en %) ainsi que l'efficacité relative (relativement à l'estimateur par dilatation) ; voir section 6.1. Les résultats sont présentés dans le tableau 6.5.

Les résultats montrent que les estimateurs robustes initiaux $\hat{t}_{R,h}$ sont biaisés comme nous pouvions nous y attendre. Le biais est plus important dans les strates ayant une petite fraction de sondage. Par exemple, dans la strate 1 pour laquelle $f_1 = 1\ \%$, le biais relatif de $\hat{t}_{1,h}$ est égal à $-11{,}9\ \%$, alors qu'il

n'est que de $-1{,}5\ \%$ dans la strate 5 pour laquelle $f_5 = 80\ \%$. On note également que les estimateurs robustes initiaux sont tous plus efficaces que l'estimateur par dilatation correspondant avec des valeurs de l'efficacité relative variant de 57 % à 97 %. L'estimateur agrégé $\hat{t}_{R(\text{agr})}$ obtenu en sommant les estimateurs initiaux $\hat{t}_{R,h}, h = 1,\ldots,5$ présente un biais modeste égal à $-5{,}7\ \%$ mais se montre plus efficace que l'estimateur par dilatation calculé au niveau de la population $\hat{t}$ avec une efficacité relative égale à 87 %.

L'estimateur winsorisé calculé au niveau de la population, $\hat{t}_{R,0}$ présente un léger biais de $-2{,}8\ \%$ et se montre significativement plus efficace que l'estimateur par dilatation avec une efficacité relative égale à 81 %. Les estimateurs finaux $\hat{t}^*_{R,h}$ obtenus au moyen du jeu de coefficients $q_h = 1$ pour $h = 1,\ldots,5$, sont tous moins biaisés que l'estimateur initial $\hat{t}_{R,h}$, à l'exception de la strate 5. Ceci s'explique par le fait que l'on force la somme des estimations finales $\hat{t}^*_{R,h}$ à se caler sur un estimateur peu biaisé. Par contre, la diminution du biais s'accompagne d'une légère diminution de l'efficacité. Par exemple, pour la strate 4, on passe d'une efficacité relative de 63 % pour l'estimateur robuste $\hat{t}_{R,4}$ à une efficacité relative de 66 % pour l'estimateur final $\hat{t}^*_{R,4}$. Dans le cas de la strate 5, il est clair que le premier jeu de coefficients est inapproprié puisqu'il conduit à modifier l'estimation pour cette strate au même titre que toutes les autres strates, alors que cette dernière possède une fraction de sondage élevée égale à 80 %. En fait, pour ce jeu de coefficients, l'estimateur $\hat{t}^*_{R,5}$ est moins efficace que l'estimateur par dilatation avec une efficacité relative égale à 104. Le deuxième choix des coefficients q_h, qui prend en compte la fraction de sondage, f_h, et la taille de l'échantillon n_h conduit à des résultats intéressants. En effet, l'estimateur robuste final dans la première strate, $\hat{t}^*_{R,1}$, est considérablement moins biaisé que l'estimateur initial $\hat{t}_{R,1}$ et que l'estimateur final basé sur le premier jeu de coefficients et ce, au prix d'une légère perte d'efficacité. Quant à la strate 5, l'estimateur $\hat{t}^*_{R,5}$ est peu biaisé (avec un biais relatif égal à $-0{,}8\ \%$) et possède la même efficacité que l'estimateur initial $\hat{t}_{R,5}$ avec une valeur égale à 97 %. Les troisième et quatrième jeux de poids q_h conduisent à des résultats similaires en termes de biais relatif et d'efficacité relative. Pour la strate 1, ces deux jeux conduisent à des biais relatifs moins élevés que ceux obtenus avec le premier jeu de poids au prix d'une légère perte d'efficacité. Pour les strates 2 à 4, tous les jeux de coefficients sont similaires en termes de biais relatif et d'efficacité relative. Finalement, pour la strate 5, les estimateurs finaux sont quasiment sans biais et pas moins efficaces que l'estimateur par dilatation.

Tableau 6.5
Biais relatif Monte Carlo (en %) et efficacité relative (entre parenthèses) des estimateurs robustes au niveau global et au niveau des strates

Estimateur global		$\hat{t}_{R(\text{agr})}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$
		-5,7(87)	-2,8(81)	-2,8(81)	-2,8(81)	-2,8(81)
			$\hat{t}^*_{R,h}$			
		$\hat{t}_{R,h}$	$q_h = 1$	$q_h = n_h^{-1}(1 - f_h)$	$q_h = \text{CV}(\hat{t}_h)$	$q_h = \widehat{\text{CV}}(\hat{t}_h)$
Strate	1	-11,9(57)	-9,1(60)	-0,9(67)	-5,7(62)	-6,7(64)
	2	-6,3(74)	-3,4(76)	-3,3(76)	-3,3(76)	-3,1(78)
	3	-6,0(69)	-3,1(70)	-3,8(69)	-3,2(70)	-3,2(70)
	4	-6,6(63)	-3,7(66)	-4,2(65)	-3,3(66)	-3,4(70)
	5	-1,5(97)	1,5(104)	-0,8(97)	-0,2(98)	0,1(99)

7 Discussion

Dans cet article, nous avons proposé une méthode de détermination du seuil pour des estimateurs winsorisés. Cette méthode a l'avantage d'être simple à mettre en oeuvre en pratique et peut être utilisée pour des plans de sondage à probabilités inégales. Nous avons également proposé une méthode de calage permettant de satisfaire une relation de cohérence entre les estimations winsorisées obtenues au niveau des domaines et une estimation winsorisée au niveau de la population. Bien que nous n'ayons appliqué cette méthode que dans le cas d'estimateurs winsorisés, cette dernière peut être utilisée pour n'importe quel type d'estimateur robuste.

Remerciements

Les auteurs remercient un éditeur associé ainsi que deux arbitres pour leurs commentaires et suggestions qui ont grandement contribué à améliorer la qualité de l'article. Les travaux de recherche de David Haziza ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

Annexe

On veut montrer qu'il existe une solution à l'équation

$$-\Delta(K) = \sum_{j \in S} a_j \max\left(0, d_j y_j - K\right) = \frac{\hat{B}_{\min} + \hat{B}_{\max}}{2} = \hat{t} - \hat{t}_R$$

sous les conditions $\pi_{ij} - \pi_i \pi_j \leq 0$ et $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \geq 0$.

Ordonnons tout d'abord les unités de la plus petite à la plus grande selon la valeur de $b_i = d_i y_i, i \in S,$ de telle sorte que l'unité 1 devient celle qui a la plus petite valeur de b_i et l'unité n devient celle qui a la plus grande valeur. Considérons en premier le cas : $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) = 0$. Il faut résoudre l'équation $-\Delta(K) = 0$ et on peut facilement observer que cette équation est satisfaite pour tout $K \geq b_n$.

Considérons maintenant le cas : $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) > 0$. Notons d'abord que la fonction $-\Delta(K)$ est continue et linéaire par morceaux pour $0 \leq K \leq b_n$. Les morceaux sont définis par les intervalles $\left[b_{j-1}, b_j\right[, j = 1, ..., n,$ où $b_0 = 0$. Notons aussi que $-\Delta(0) = \sum_{j=m}^{n} a_j b_j > 0,$ où m est le plus petit indice tel que $b_m \geq 0$. Par le théorème de la valeur intermédiaire, il existe une solution à l'équation (4.7) si on peut montrer que

$$-\Delta(b_n) = 0 < \frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \leq -\Delta(0) = \sum_{j=m}^{n} a_j b_j. \qquad \text{(A.1)}$$

La première inégalité découle directement de la condition $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) > 0$. Pour montrer la deuxième inégalité, on note d'abord que $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \leq \hat{B}_{\max}$. Si on utilise l'estimateur (2.2) du biais

conditionnel et la condition $\pi_{ij} - \pi_i \pi_j \leq 0$ alors on observe que $\hat{B}_{\max} \leq (d_k - 1)\, y_k$, l'indice k étant associé à l'unité qui a le plus grand biais conditionnel estimé. Pour l'estimateur winsorisé de Dalén-Tambay, cette dernière inégalité peut être réécrite comme suit : $\hat{B}_{\max} \leq a_k b_k$. Il en résulte que $a_k b_k \leq -\Delta(0) = \sum_{j=m}^{n} a_j b_j$, ce qui complète la preuve d'existence d'une solution à l'équation (4.7). Pour l'estimateur winsorisé standard, on peut aussi facilement montrer que $\hat{B}_{\max} \leq a_k b_k$ et donc qu'une solution existe. De plus, si les $y_i, i \in S$, sont tous positifs alors la fonction $-\Delta(K)$ est monotone décroissante pour $0 \leq K \leq b_n$ et la solution est unique.

Pour trouver la solution K_{opt}, on trouve le plus grand indice l tel que $-\Delta(b_l) \geq \frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right)$, pour $l \leq n$. La solution peut ensuite être obtenue par interpolation linéaire entre les points b_l et b_{l+1}; c'est-à-dire

$$K_{\text{opt}} = b_l \frac{\Delta(b_{l+1}) - \Delta\left(K_{\text{opt}}\right)}{\Delta(b_{l+1}) - \Delta(b_l)} + b_{l+1} \frac{\Delta\left(K_{\text{opt}}\right) - \Delta(b_l)}{\Delta(b_{l+1}) - \Delta(b_l)},$$

où $\Delta\left(K_{\text{opt}}\right) = -\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right)$.

Bibliographie

Beaumont, J.-F., Haziza, D. et Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika,* 100, 555-569.

Berger, Y.G. (1998). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference,* 74, 149-168.

Clark, R.G. (1995). Winsorization methods in sample surveys. Thèse de maîtrise, Department of Statistics, Australian National University.

Dalén, J. (1987). Practical estimators of a population total which reduce the impact of large observations. R and D Report. Statistics Sweden.

Datta, G.S., Gosh, M., Steorts, R. et Maple, J. (2011). Bayesian benchmarking with applications to small area estimation. *Test,* 20, 574-588.

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

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika,* 93, 269-278.

Haziza, D., Mecatti, F. et Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron,* 66, 91-108.

Kokic, P.N., et Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics,* 10, 419-435.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. et Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: Estimating the conditional bias. *Metrika,* 55, 209-214.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. et Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: Conditional bias. *Biometrika,* 86, 923-928.

Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika,* 81, 373-383.

Rivest, L.-P., et Hidiroglou, M. (2004). Outlier treatment for disaggregated estimates. *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, Virginie, 4248-4256.

Rivest, L.-P., et Hurtubise, D. (1995). Moyenne winsorisée de Searls pour populations asymétriques. *Techniques d'enquête,* 21, 2, 119-129.

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

Thompson, M.E., et Wu, C. (2008). Échantillonnage PPT systématique randomisé basé sur la simulation en cas de substitution d'unités. *Techniques d'enquête,* 34, 1, 3-11.

You, Y., Rao, J.N.K. et Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition,* 6, 631-640.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 81-100
Statistique Canada, n° 12-001-X au catalogue

Estimateur par la régression modifiée pour les enquêtes-entreprises répétées avec bases de sondage évolutives

John Preston[1]

Résumé

L'estimation composite est une technique applicable aux enquêtes répétées avec chevauchement contrôlé entre les enquêtes successives. Le présent article examine les estimateurs par la régression modifiée qui permettent d'intégrer l'information provenant de périodes antérieures dans les estimations pour la période courante. La gamme d'estimateurs par la régression modifiée est étendue au cas des enquêtes-entreprises dont la base de sondage évolue avec le temps en raison de l'ajout des « nouvelles entreprises » et de la suppression des « entreprises disparues ». Puisque les estimateurs par la régression modifiée peuvent s'écarter de l'estimateur par la régression généralisée au cours du temps, il est proposé d'utiliser un estimateur par la régression modifiée de compromis correspondant à la moyenne pondérée de l'estimateur par la régression modifiée et de l'estimateur par la régression généralisée. Une étude par simulation Monte Carlo montre que l'estimateur par la régression modifiée de compromis proposé donne lieu à d'importants gains d'efficacité en ce qui concerne les estimations ponctuelles ainsi que les estimations des variations.

Mots-clés : Bases de sondage évolutives; estimation composite; régression modifiée; enquêtes répétées; échantillons rotatifs.

1 Introduction

La méthode d'estimation composite a été utilisée amplement dans les enquêtes-ménages à panel rotatif afin d'améliorer l'efficacité des estimations des variations, en accordant plus de poids aux groupes de renouvellement « communs ». La plupart des estimateurs composites existants, dont l'estimateur composite AK (Gurney et Daly 1965), le meilleur estimateur linéaire sans biais (BLUE) (Yansaneh et Fuller 1998) et l'estimateur B1 (Bell 2001), nécessitent que toutes les unités primaires d'échantillonnage dans la population puissent être affectées aléatoirement à des groupes de renouvellement. Ces estimateurs composites n'ont pas été adoptés largement pour les enquêtes-entreprises, car le concept des groupes de renouvellement ne s'adapte pas bien aux enquêtes-entreprises répétées. Les plans de sondage à panel rotatif conviennent mal aux enquêtes-entreprises répétées en raison de la nature très dynamique des bases de sondage, dans lesquelles des changements sont causés par l'ajout des « nouvelles unités » de population et par la suppression des « unités disparues », ainsi que par les modifications des données de classification au cours du temps.

Un exemple type de ce genre d'enquête-entreprise répétée est la *Quarterly Business Indicators Survey* (Australian Bureau of Statistics (ABS) 2012b), dont la base de sondage est mise à jour trimestriellement pour tenir compte des nouvelles entreprises et des changements des caractéristiques des entreprises. De surcroît, environ le douzième des unités sectorielles échantillonnées sont supprimées de l'échantillon de l'enquête et remplacées par d'autres, afin de répartir le fardeau de déclaration équitablement.

L'estimateur par la régression modifiée qui a été introduit pour la première fois par Singh (1994) semble être le type le plus approprié d'estimateur composite pour tenir compte des bases de sondage évolutives. Les estimateurs par la régression modifiée les plus anciens sont l'estimateur RM1 (Singh et

1. John Preston, Australian Bureau of Statistics, 639 Wickham Street, Fortitude Valley QLD 4006, Australie. Courriel : john.preston@abs.gov.au.

Merkouris 1995; Singh 1996) et l'estimateur RM2 (Singh, Kennedy, Wu et Brisebois 1997). Le premier s'est avéré donner de meilleurs résultats pour les estimations ponctuelles et le second, pour les estimations des variations. Un compromis entre ces deux estimateurs par la régression modifiée, appelé estimateur par la régression modifiée composite, a été proposé par Fuller et Rao (2001). Cet estimateur par la régression modifiée composite a été étudié par Singh, Kennedy et Wu (2001), Gambino, Kennedy et Singh (2001), Bell (2001), et Beaumont et Bocci (2005).

Pour tous ces estimateurs par la régression modifiée, les meilleurs résultats s'obtiennent quand les unités de la population ne changent pas de la période précédente à la période courante. Cela ne pose pas vraiment de problèmes pour une enquête-ménage mensuelle type où les taux de natalité, de mortalité et de migration nette sont relativement faibles. Par exemple, en Australie durant la période 2011-2012, le taux mensuel moyen de natalité était de 0,11 %, le taux mensuel moyen de mortalité était de 0,05 %, et le taux mensuel moyen de migration nette était de 0,08 % (ABS 2012a). En revanche, cela pose davantage de problèmes pour une enquête-entreprise trimestrielle type où les taux de création et de disparition d'entreprises sont nettement plus élevés. Par exemple, en Australie durant la période 2011-2012, le taux trimestriel moyen de création (ou d'entrée) d'entreprises était de 3,38 % et le taux trimestriel moyen de disparition (ou de sortie) d'entreprises était de 3,28 % (ABS 2012c).

Si la population subit des changements importants au cours du temps, les estimateurs par la régression modifiée ne conviennent pas sous leur forme actuelle, car un biais important risque de s'accumuler. Ces estimateurs par la régression modifiée peuvent être étendus à la situation où les bases de sondage évoluent en apportant des ajustements aux variables auxiliaires composites, premièrement en ajoutant les « nouvelles unités » à la population de la période précédente, puis en ajoutant les « unités disparues » à la population de la période courante pour créer une « pseudo-population ». Ces « pseudo-populations » satisferont à l'exigence que les unités dans la population ne changent pas entre la période précédente et la période courante.

La section 2 décrit l'estimateur par la régression généralisée et les estimateurs par la régression modifiée, ainsi qu'une moyenne pondérée de ces deux estimateurs qui donne lieu à d'importants gains d'efficacité en ce qui concerne les estimations ponctuelles ainsi que les estimations des variations. Une extension de l'estimateur par la régression modifiée pour tenir compte des bases de sondage évolutives est également décrite à la section 3. Les résultats d'une étude par simulation sont présentés à la section 4. Certaines conclusions sont exposées à la section 5.

2 Estimation par la régression modifiée

Considérons une population finie $U^{(t)}$ à la période t partitionnée en H strates non chevauchantes $U_1^{(t)},\ldots,U_h^{(t)},\ldots,U_H^{(t)}$, où $U_h^{(t)}$ est constituée de $N_h^{(t)}$ unités. Un échantillon aléatoire simple sans remise $s_h^{(t)}$ de $n_h^{(t)}$ unités est sélectionné avec les probabilités d'inclusion $\pi_i^{(t)} = n_h^{(t)} / N_h^{(t)} \left(i \in U_h^{(t)}\right)$ dans chaque strate h à la période t, ce qui donne un échantillon total $s^{(t)} = \bigcup_{h=1}^{H} s_h^{(t)}$ de taille $n^{(t)} = \sum_{h=1}^{H} n_h^{(t)}$. Une estimation sans biais du total de population $Y^{(t)} = \sum_{h=1}^{H} \sum_{i \in U_h^{(t)}} y_i^{(t)}$ est donnée par l'estimateur de Horvitz-Thompson (HT) $\hat{Y}_{\mathrm{HT}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} w_i^{(t)} y_i^{(t)}$, où $w_i^{(t)} = 1 / \pi_i^{(t)}$ est le poids de sondage de l'unité i à la période t et $y_i^{(t)}$ est la valeur de la variable d'intérêt y pour l'unité i à la période t. Supposons

qu'il existe un ensemble de variables auxiliaires $\mathbf{x}^{(t)}$ à la période t pour lequel les totaux de population $\mathbf{X}^{(t)} = \sum_{i \in U^{(t)}} \mathbf{x}_i^{(t)}$ sont connus et les variables $\mathbf{x}_i^{(t)}$ sont connues pour chaque $i \in s^{(t)}$.

L'estimateur par la régression généralisée (RG) (Särndal, Swensson et Wretman 1992) est un estimateur assisté par modèle, conçu en vue d'améliorer l'exactitude des estimations en utilisant des variables auxiliaires qui sont corrélées à la variable d'intérêt. L'estimateur RG est donné par :

$$\hat{Y}_{\mathrm{RG}}^{(t)} = \hat{Y}_{\mathrm{HT}}^{(t)} + \left(\mathbf{X}^{(t)} - \hat{\mathbf{X}}_{\mathrm{HT}}^{(t)}\right)^T \hat{\boldsymbol{\beta}}_{\mathrm{RG}}^{(t)} \tag{2.1}$$

où $\hat{\boldsymbol{\beta}}_{\mathrm{RG}}^{(t)}$ est le vecteur des paramètres du modèle de régression linéaire donné par :

$$\hat{\boldsymbol{\beta}}_{\mathrm{RG}}^{(t)} = \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \mathbf{x}_i^{(t)} \mathbf{x}_i^{(t)T}}{c_i^{(t)}}\right)^{-1} \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \mathbf{x}_i^{(t)} y_i^{(t)}}{c_i^{(t)}}\right) \tag{2.2}$$

et les $c_i^{(t)}$ sont les facteurs spécifiés qui se rapportent à la structure de variance du modèle de régression linéaire associé à l'estimateur RG $y_i^{(t)} = \mathbf{x}_i^{(t)T} \hat{\boldsymbol{\beta}}_{\mathrm{RG}}^{(t)} + \varepsilon_i^{(t)}$, avec $E\left(\varepsilon_i^{(t)}\right) = 0$, $\mathrm{Var}\left(\varepsilon_i^{(t)}\right) = c_i^{(t)} \sigma^2$ et $\mathrm{Cov}\left(\varepsilon_i^{(t)}, \varepsilon_j^{(t)}\right) = 0$ pour tout $i \neq j$. L'estimateur RG peut aussi s'écrire sous la forme :

$$\hat{Y}_{\mathrm{RG}}^{(t)} = \sum_{i \in s^{(t)}} \tilde{w}_i^{(t)} y_i^{(t)} \tag{2.3}$$

où $\tilde{w}_i^{(t)} = w_i^{(t)} \tilde{g}_i^{(t)}$ et $\tilde{g}_i^{(t)}$ est le poids g pour l'unité i à la période t donné par :

$$\tilde{g}_i^{(t)} = 1 + \left(\mathbf{X}^{(t)} - \hat{\mathbf{X}}_{\mathrm{HT}}^{(t)}\right)^T \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \mathbf{x}_i^{(t)} \mathbf{x}_i^{(t)T}}{c_i^{(t)}}\right)^{-1} \frac{\mathbf{x}_i^{(t)}}{c_i^{(t)}}. \tag{2.4}$$

À la période $t > 1$, définissons un ensemble de variables auxiliaires composites $\mathbf{z}^{(t)}$ pour lequel les « pseudo-totaux de référence » $\tilde{\mathbf{Z}}^{(t)}$ (basés sur les estimations des variables clés de l'enquête à la période $t-1$) sont connus et $\mathbf{z}_i^{(t)}$ peut être calculé pour chaque $i \in s^{(t)}$. L' estimateur par la régression modifiée (RM) est l'estimateur RG dans lequel les variables du modèle de régression sont les variables auxiliaires $\mathbf{x}^{(t)}$ et les variables auxiliaires composites $\mathbf{z}^{(t)}$. L'estimateur RM est donné par :

$$\hat{Y}_{\mathrm{RM}}^{(t)} = \hat{Y}_{\mathrm{HT}}^{(t)} + \left(\left(\mathbf{X}^{(t)}, \tilde{\mathbf{Z}}^{(t)}\right) - \left(\hat{\mathbf{X}}_{\mathrm{HT}}^{(t)}, \hat{\mathbf{Z}}_{\mathrm{HT}}^{(t)}\right)\right)^T \hat{\boldsymbol{\beta}}_{\mathrm{RM}}^{(t)} \tag{2.5}$$

où $\hat{\boldsymbol{\beta}}_{\mathrm{RM}}^{(t)}$ est le vecteur des paramètres du modèle de régression linéaire donné par :

$$\hat{\boldsymbol{\beta}}_{\mathrm{RM}}^{(t)} = \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)}\right)\left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)}\right)^T}{c_i^{(t)}}\right)^{-1} \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)}\right) y_i^{(t)}}{c_i^{(t)}}\right). \tag{2.6}$$

L'estimateur RM peut aussi s'écrire sous la forme :

$$\hat{Y}_{\mathrm{RM}}^{(t)} = \sum_{i \in s^{(t)}} \breve{w}_i^{(t)} y_i^{(t)} \tag{2.7}$$

où $\breve{w}_i^{(t)} = w_i^{(t)} \breve{g}_i^{(t)}$ et $\breve{g}_i^{(t)}$ est le poids g pour l'unité i à la période t donné par :

$$\begin{aligned} \breve{g}_i^{(t)} &= 1 + \left(\left(\mathbf{X}^{(t)}, \tilde{\mathbf{Z}}^{(t)}\right) - \left(\hat{\mathbf{X}}_{\mathrm{HT}}^{(t)}, \hat{\mathbf{Z}}_{\mathrm{HT}}^{(t)}\right)\right)^T \\ &\times \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)}\right)\left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)}\right)^T}{c_i^{(t)}}\right)^{-1} \frac{\left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)}\right)}{c_i^{(t)}}. \end{aligned} \tag{2.8}$$

La clé de l'efficacité de l'estimateur RM tient à la définition des variables auxiliaires composites. Idéalement, les valeurs des variables auxiliaires composites à la période t seraient égales aux valeurs des variables clés de l'enquête à la période $t-1$. Cependant, en raison du roulement dû aux unités qui entrent

dans l'échantillon et aux unités qui en sortent d'une période à la suivante, les valeurs des variables clés de l'enquête à la période $t-1$ manqueront, par conception, pour les unités présentes dans l'échantillon à la période t, mais non à la période $t-1$.

Plusieurs méthodes existent pour définir les variables auxiliaires composites. Les estimateurs par la régression modifiée les plus anciens étaient l'estimateur RM1 (Singh et Merkouris 1995; Singh 1996) et l'estimateur RM2 (Singh, Kennedy, Wu et Brisebois 1997) dans lesquels les valeurs utilisées pour les variables auxiliaires composites étaient données, respectivement, par :

$$\mathbf{z}_{(\mathrm{RM}1)i}^{(t)} = \begin{cases} \mathbf{y}_i^{(t-1)}, & \text{si } i \in s_h^{(t)} \cap s_h^{(t-1)} \\ \bar{\mathbf{Y}}_{(\mathrm{RM})h}^{(t-1)}, & \text{si } i \in s_h^{(t)} \setminus s_h^{(t-1)} \end{cases} \tag{2.9}$$

$$\mathbf{z}_{(\mathrm{RM}2)i}^{(t)} = \begin{cases} \mathbf{y}_i^{(t)} + \left(\sum_{i \in s_h^{(t)}} w_i^{(t)} \Big/ \sum_{i \in s_h^{(t)} \cap s_h^{(t)-1}} w_i^{(t)}\right)\left(\mathbf{y}_i^{(t-1)} - \mathbf{y}_i^{(t)}\right), & \text{si } i \in s_h^{(t)} \cap s_h^{(t-1)} \\ \mathbf{y}_i^{(t)}, & \text{si } i \in s_h^{(t)} \setminus s_h^{(t)-1} \end{cases} \tag{2.10}$$

et $\bar{\mathbf{Y}}_{(\mathrm{RM})h}^{(t-1)}$ représente les estimateurs par la régression composites de la moyenne de population dans la strate h pour les variables clés de l'enquête à la période $t-1$.

Pour les valeurs RM1 des variables auxiliaires composites, on applique une méthode d'imputation par la moyenne pour imputer les valeurs manquantes, tandis que pour les valeurs RM2, on utilise une méthode d'imputation historique inverse pour imputer les valeurs manquantes, puis on modifie les valeurs qui n'ont pas été imputées de manière que l'estimateur HT des variables auxiliaires composites $\hat{\mathbf{Z}}_{\mathrm{HT}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} w_i^{(t)} \mathbf{z}_{(\mathrm{RM}2)i}^{(t)}$ à la période t soit sans biais pour les variables d'enquête clés correspondantes $\mathbf{Y}^{(t-1)}$ à la période $t-1$.

L'estimateur RM1 s'est avéré donner de meilleurs résultats pour les estimations ponctuelles, tandis que l'estimateur RM2 s'est avéré donner de meilleurs résultats pour les estimations des variations. Fuller et Rao (2001) ont proposé un estimateur de rechange qui offre un compromis entre l'amélioration des estimations ponctuelles et l'amélioration des estimations des variations grâce à l'utilisation de valeurs des variables auxiliaires composites données par :

$$\mathbf{z}_{(\mathrm{RM})i}^{(t)} = (1-\alpha)\,\mathbf{z}_{(\mathrm{RM}1)i}^{(t)} + \alpha \mathbf{z}_{(\mathrm{RM}2)i}^{(t)}. \tag{2.11}$$

L'expression (2.11) pour les variables auxiliaires composites requiert une décision quant au choix de α, qui dépendra des corrélations des variables d'enquête clés dans le temps et de l'importance relative des estimations ponctuelles et des estimations des variations.

Beaumont et Bocci (2005) ont proposé un perfectionnement des variables auxiliaires composites qui, selon eux, ne nécessite pas de choix arbitraire de α :

$$\mathbf{z}_{(\mathrm{RMP})i}^{(t)} = \begin{cases} \mathbf{y}_i^{(t-1)}, & \text{si } i \in s_h^{(t)} \cap s_h^{(t-1)} \\ \mathbf{y}_i^{(t)} + \left(\sum_{i \in s_h^{(t)} \cap s_h^{(t-1)}} w_i^{(t)} \left(\mathbf{y}_i^{(t-1)} - \mathbf{y}_i^{(t)}\right) \Big/ \sum_{i \in s_h^{(t)} \cap s_h^{(t-1)}} w_i^{(t)}\right), & \text{si } i \in s_h^{(t)} \setminus s_h^{(t-1)}. \end{cases} \tag{2.12}$$

Dans l'approche perfectionnée RMP, une méthode d'imputation historique inverse est utilisée pour imputer les valeurs manquantes des variables auxiliaires composites, puis les valeurs imputées sont modifiées afin que l'estimateur HT des variables auxiliaires composites $\hat{\mathbf{Z}}_{\mathrm{HT}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} w_i^{(t)} \mathbf{z}_{(\mathrm{RMP})i}^{(t)}$ à la période t soit sans biais pour les variables d'enquête clés $\mathbf{Y}^{(t-1)}$ à la période $t-1$.

Les estimateurs RM peuvent s'écarter de l'estimateur RG au cours du temps (Fuller et Rao 2001). Dans une enquête répétée, ce problème de « dérive » sera caractérisé par un écart important qui s'agrandit au cours du temps entre l'estimateur RM et l'estimateur RG, tandis qu'une étude par simulation sera caractérisée par une réduction au cours du temps de l'efficacité relative de l'estimateur RM comparativement à l'estimateur RG. Une solution éventuelle au problème de « dérive » consisterait à utiliser une moyenne pondérée de l'estimateur RM et de l'estimateur RG (Bell 1999) donnée par :

$$\hat{Y}_{\mathrm{RMC}}^{(t)} = \alpha \hat{Y}_{\mathrm{RG}}^{(t)} + (1-\alpha)\hat{Y}_{\mathrm{RM}}^{(t)}. \tag{2.13}$$

L'estimateur par la régression modifiée de compromis (RMC) doit aussi offrir un compromis entre les gains d'efficacité pour les estimations ponctuelles et les estimations des variations, parce que les estimateurs RM donnent généralement de meilleurs résultats que l'estimateur RG pour les estimations des variations, mais ne donnent pas toujours de meilleurs résultats pour les estimations ponctuelles; en particulier les estimateurs RM2 et RMP.

L'estimateur RMC requiert une décision quant au choix de α. En utilisant des méthodes de linéarisation (ou de développement en série de Taylor) pour approximer la variance de (2.13), il est possible de trouver une expression relativement simple pour α qui minimise la variance sur les estimations des variations tout en maintenant la variance sur les estimations ponctuelles produites en utilisant l'estimateur RG.

Les estimateurs RM courants donnent leurs meilleurs résultats lorsque les unités de la population ne changent pas entre la période précédente et la période courante. En cas de changements importants dans la population au cours du temps, ces estimateurs par la régression modifiée ne conviennent pas sous leur forme actuelle, car ils peuvent accumuler un biais important au cours du temps. Bien qu'un facteur simple $\left(\sum_{i\in s_h^{(t-1)}} w_i^{(t-1)} \Big/ \sum_{i\in s_h^{(t)}} w_i^{(t)}\right)$ puisse être appliqué aux valeurs RM1, RM2 et RMP pour tenir compte des changements de la taille de la population dans la strate h entre les périodes $t-1$ et t, ces estimateurs par la régression modifiée peuvent encore accumuler un biais considérable au cours du temps.

3 Estimation par la régression modifiée pour bases de sondage évolutives

Les estimateurs RM peuvent être étendus au cas des bases de sondage évolutives par ajout des « nouvelles unités » à la population de la période précédente et par ajout des « unités disparues » à la population de la période courante pour créer une « pseudo-population » (diagramme 3.1). Ces « pseudo-populations » satisferont à l'exigence que les unités de la population ne changent pas entre la période précédente et la période courante. L'extension de l'estimateur RM pour tenir compte des bases de sondage évolutives est décrite en détail ci-après.

Considérons une population dynamique qui évolue au cours du temps en raison de l'ajout des « nouvelles unités » et de la suppression des « unités disparues ». À la période t, l'union de $U_h^{(t)}$ et $U_h^{(t-1)}$ peut être subdivisée en trois composantes. La première composante comprend les unités de la population présentes dans la strate h à la période $t-1$ mais non à la période t, c'est-à-dire la population d'« unités disparues » $U_{dh}^{(t-1)}$ de la strate h, constituée de $N_{dh}^{(t-1)}$ unités. La deuxième composante comprend les unités présentes dans la population de la strate h à la période $t-1$ et à la période t, c'est-à-dire la population « commune » $U_{ch}^{(t-1)} = U_{ch}^{(t)}$ de la strate h, constituée de $N_{ch}^{(t-1)} = N_{ch}^{(t)}$ unités. La troisième

composante comprend les unités présentes dans la population de la strate h à la période t mais non à la période $t-1$, c'est-à-dire la population de « nouvelles unités » $U_{bh}^{(t)}$ de la strate h, constituée de $N_{bh}^{(t)}$ unités. Les unités de la population qui changent de strate entre les périodes $t-1$ et t sont incluses dans la population d'« unités disparues » $U_{dh}^{(t-1)}$ sous leur strate à la période $t-1$ et sont également incluses dans la population de « nouvelles unités » $U_{bh}^{(t)}$ sous leur strate à la période t.

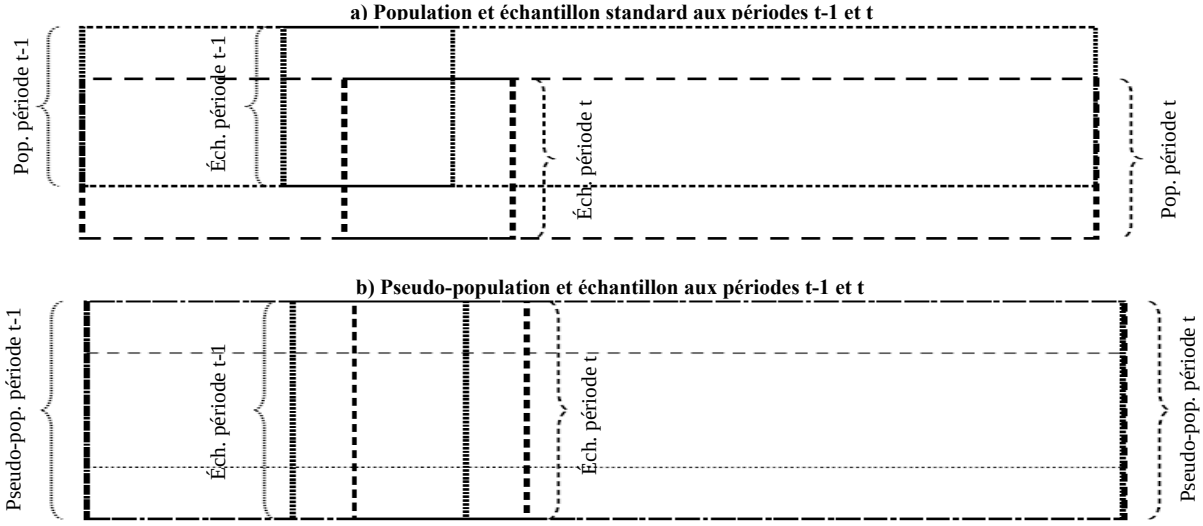


Diagramme 3.1 Populations et échantillons standard et pseudo-populations et échantillons

À la période $t > 1$, définissons la « pseudo-population » $U_h^{*(t-1)} = U_h^{*(t)}$ dans la strate h comme étant l'union de $U_h^{(t)}$ et $U_h^{(t-1)}$, constituée de $N_h^{*(t-1)} = N_h^{*(t)} = N_{dh}^{(t-1)} + N_{ch}^{(t-1)} + N_{bh}^{(t)}$ unités. Il est important de noter que la « pseudo-population » $U_h^{*(t-1)}$ à la période $t-1$ diffère de la « pseudo-population » $U_h^{*(t-1)}$ à la période t, car la « pseudo-population » $U_h^{*(t-1)}$ à la période $t-1$ est fondée sur l'union de $U_h^{(t-2)}$ et $U_h^{(t-1)}$, tandis que la « pseudo-population » $U_h^{*(t-1)}$ à la période t est fondée sur l'union de $U_h^{(t-1)}$ et $U_h^{(t)}$. Donc, les « pseudo-populations » pour les périodes courante et précédente doivent être calculées à chaque période. Définissons les « pseudo-valeurs » de la variable d'intérêt y pour l'unité i à la période $t-1$ et à la période t comme étant :

$$y_i^{*(t-1)} = \begin{cases} y_i^{(t-1)}, & \text{si } i \in U_{ch}^{(t-1)} \\ 0, & \text{si } i \in U_{bh}^{(t)} \end{cases}$$

$$y_i^{*(t)} = \begin{cases} y_i^{(t)}, & \text{si } i \in U_{ch}^{(t)} \\ 0, & \text{si } i \in U_{dh}^{(t-1)} \end{cases}$$

et définissons les « pseudo-valeurs » des variables auxiliaires $\mathbf{x}$ pour l'unité i à la période $t-1$ et à la période t comme étant :

$$\mathbf{x}_i^{*(t-1)} = \begin{cases} \mathbf{x}_i^{(t-1)}, & \text{si } i \in U_{ch}^{(t-1)} \\ 0, & \text{si } i \in U_{bh}^{(t)} \end{cases}$$

$$\mathbf{x}_i^{*(t)} = \begin{cases} \mathbf{x}_i^{(t)}, & \text{si } i \in U_{ch}^{(t)} \\ 0, & \text{si } i \in U_{dh}^{(t-1)}. \end{cases}$$

À la période $t > 1$, notons que $s_h^{*(t-1)}$ et $s_h^{*(t)}$ sont les « pseudo-échantillons » dans la strate h, où $s_h^{*(t-1)}$ est constitué de toutes les unités sélectionnées dans l'échantillon original $s_h^{(t-1)}$ dans la strate h à la période $t-1$ plus un échantillon aléatoire d'unités $s_{bh}^{(t)}$ provenant de la population de « nouvelles unités » $U_{bh}^{(t)}$ dans la strate h à la période t sélectionnées avec les probabilités d'inclusion $\pi_i^{(t-1)} = n_h^{(t-1)} / N_h^{(t-1)} \left(i \in U_h^{(t-1)}\right)$, et $s_h^{*(t)}$ est constitué de toutes les unités sélectionnées dans l'échantillon original $s_h^{(t)}$ dans la strate h à la période t plus un échantillon aléatoire d'unités $s_{dh}^{(t-1)}$ provenant de la population d'« unités disparues » $U_{dh}^{(t-1)}$ dans la strate h à la période $t-1$ sélectionnées avec les probabilités d'inclusion $\pi_i^{(t)} = n_h^{(t)} / N_h^{(t)} \left(i \in U_h^{(t)}\right)$. Soient $n_h^{*(t-1)}$ et $n_h^{*(t)}$ les tailles des « pseudo-échantillons » $s_h^{*(t-1)}$ et $s_h^{*(t)}$, respectivement. De nouveau, il est important de noter que le « pseudo-échantillon » $s_h^{*(t-1)}$ à la période $t-1$ diffère du « pseudo-échantillon » $s_h^{*(t-1)}$ à la période t, car le « pseudo-échantillon » $s_h^{*(t-1)}$ à la période $t-1$ comprend un échantillon aléatoire d'unités provenant de la population de « nouvelles unités » à la période $t-1$, tandis que le « pseudo-échantillon » $s_h^{*(t-1)}$ à la période t comprend un échantillon aléatoire d'unités provenant de la population d'« unités disparues » à la période $t-1$. Donc, les « pseudo-échantillons » pour les périodes courante et précédente doivent être calculés à chaque période.

Le choix d'une méthode appropriée de sélection de l'échantillon, pour la sélection des échantillons aléatoires supplémentaires d'unités tirées des populations de « nouvelles unités » et d'« unités disparues », dépendra de la méthode de sélection de l'échantillon utilisée pour sélectionner les échantillons originaux. Dans le cas de nombreuses enquêtes-entreprises répétées, les échantillons sont sélectionnés en utilisant une méthode de sélection par attribution de nombres aléatoires permanents (NAP) afin de pouvoir exercer un certain contrôle sur le roulement des unités qui entrent dans l'échantillon et qui en sortent d'une période à la suivante. Considérons le cas le plus simple où les échantillons originaux $s_h^{(t-1)}$ et $s_h^{(t)}$ dans la strate h décrits par $\left\{i \in U_h^{(t-1)} \text{ et } R_i \in \left[S_h^{(t-1)}, E_h^{(t-1)}\right)\right\}$ et $\left\{i \in U_h^{(t)} \text{ et } R_i \in \left[S_h^{(t)}, E_h^{(t)}\right)\right\}$, où $S_h^{(t)}$ et $E_h^{(t)}$ sont les points de début et de fin de l'intervalle de sélection dans la strate h à la période t, et R_i est le nombre aléatoire permanent attribué à l'unité i. Dans ce cas, les « pseudo-échantillons » $s_h^{*(t-1)}$ et $s_h^{*(t)}$ dans la strate h sont décrits par $\left\{i \in U_h^{*(t-1)} \text{ et } R_i \in \left[S_h^{(t-1)}, E_h^{(t-1)}\right)\right\}$ et $\left\{i \in U_h^{*(t)} \text{ et } R_i \in \left[S_h^{(t)}, E_h^{(t)}\right)\right\}$. Cette méthode de sélection donnera une même quantité de chevauchement entre les échantillons provenant de la population d'« unités disparues » aux périodes $t-1$ et t, et entre les échantillons provenant de la population de « nouvelles unités » aux périodes $t-1$ et t qu'entre les échantillons provenant de la population « commune » aux périodes $t-1$ et t. Manifestement, la quantité de chevauchements des échantillons provenant des populations d'« unités disparues » et de « nouvelles unités » aura une incidence sur le comportement des estimations, et l'optimisation de la quantité de chevauchements pourrait être étudiée.

Soit les « pseudo-poids de sondage » $w_i^{*(t-1)} = 1/\pi_i^{(t-1)}$ pour toutes les unités du « pseudo-échantillon » $s_h^{*(t-1)}$, et $w_i^{*(t)} = 1/\pi_i^{(t)}$ pour toutes les unités du « pseudo-échantillon » $s_h^{*(t)}$. Puisque les « pseudo-poids de sondage » pour les unités échantillonnées originales sont égaux aux poids de sondage originaux et les « pseudo-valeurs » de la variable d'intérêt sont égales à zéro pour les unités échantillonnées additionnelles provenant des populations de « nouvelles unités » et d'« unités disparues », l'estimateur HT $\hat{Y}_{\mathrm{HT}}^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} y_i^{*(t)}$ basé sur le « pseudo-échantillon », les « pseudo-valeurs » et les « pseudo-poids de sondage » est équivalent à l'estimateur HT $\hat{Y}_{\mathrm{HT}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} w_i^{(t)} y_i^{(t)}$ basé sur l'échantillon original, les valeurs originales et les poids de sondage originaux. D'où, l'inclusion de ces unités échantillonnées additionnelles dans le « pseudo-échantillon » provenant des populations de « nouvelles unités » et d'« unités disparues » n'introduira aucune variabilité supplémentaire dans les estimations ponctuelles.

L'estimateur RM proposé pour le cas particulier des bases de sondage évolutives peut s'écrire sous la forme :

$$\hat{Y}_{\mathrm{RM}}^{*(t)} = \sum_{i \in s^{*(t)}} \breve{w}_i^{*(t)} y_i^{*(t)} \tag{3.1}$$

où $\breve{w}_i^{*(t)} = w_i^{*(t)} \breve{g}_i^{*(t)}$ et $\breve{g}_i^{*(t)}$ est le « pseudo-poids g » pour l'unité i à la période t donné par :

$$\begin{aligned} \breve{g}_i^{*(t)} &= 1 + \left(\left(\mathbf{X}^{(t)}, \tilde{\mathbf{Z}}^{*(t)} \right) - \left(\hat{\mathbf{X}}_{\mathrm{HT}}^{(t)}, \hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)} \right) \right)^T \\ &\times \left(\sum_{i \in s^{*(t)}} \frac{w_i^{*(t)} \left(\mathbf{x}_i^{*(t)}, \mathbf{z}_i^{*(t)} \right) \left(\mathbf{x}_i^{*(t)}, \mathbf{z}_i^{*(t)} \right)^T}{c_i^{(t)}} \right)^{-1} \frac{\left(\mathbf{x}_i^{*(t)}, \mathbf{z}_i^{*(t)} \right)}{c_i^{(t)}} \end{aligned} \tag{3.2}$$

et les valeurs RM1, RM2 et RMP pour les « pseudo-variables auxiliaires composites » sont données par :

$$\mathbf{z}_{(\mathrm{RM1})i}^{*(t)} = \begin{cases} R_h^{(t-1,t)} \mathbf{y}_i^{*(t-1)}, & \text{si } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \text{ et } s_h^{*(t)} \setminus s_h^{*(t-1)} \neq \varnothing \\ R_h^{(t-1,t)} \left(\dfrac{\sum_{i \in s_h^{(t)}} w_i^{(t)}}{\sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}} \right) \mathbf{y}_i^{*(t-1)}, & \text{si } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \text{ et } s_h^{*(t)} \setminus s_h^{*(t-1)} = \varnothing \\ R_h^{(t-1,t)} \left(\dfrac{\left(\sum_{i \in s_h^{(t)}} w_i^{(t)} - \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)} \right)}{\sum_{i \in s_h^{*(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)}} \right) \overline{\mathbf{Y}}_{(\mathrm{RM})h}^{(t-1)}, & \text{si } i \in s_h^{*(t)} \setminus s_h^{*(t-1)}. \end{cases} \tag{3.3}$$

$$\mathbf{z}_{(\mathrm{RM2})i}^{*(t)} = \begin{cases} R_h^{(t-1,t)} \left\{ \left(\dfrac{\sum_{i \in s_h^{(t)}} w_i^{(t)}}{\sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}} \right) \mathbf{y}_i^{*(t-1)} + \left(1 - \left(\dfrac{\sum_{i \in s_h^{(t)}} w_i^{(t)}}{\sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)}} \right) \right) \mathbf{y}_i^{*(t)} \right\}, & \text{si } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \\ R_h^{(t-1,t)} \mathbf{y}_i^{*(t)}, & \text{si } i \in s_h^{*(t)} \setminus s_h^{*(t-1)}. \end{cases} \tag{3.4}$$

$$\mathbf{z}_{(\mathrm{RMP})i}^{*(t)} = \begin{cases} R_h^{(t-1,t)}\mathbf{y}_i^{*(t-1)}, & \text{si } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \text{ et } s_h^{*(t)} \setminus s_h^{*(t-1)} \neq \varnothing \\ R_h^{(t-1,t)}\left(\sum_{i\in s_h^{(t)}} w_i^{(t)} \Big/ \sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\right)\mathbf{y}_i^{*(t-1)}, & \text{si } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \text{ et } s_h^{*(t)} \setminus s_h^{*(t-1)} = \varnothing \\ R_h^{(t-1,t)}\left\{\mathbf{y}_i^{*(t)} - \left[\left(\sum_{i\in s_h^{(t)}\setminus s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i\in s_h^{*(t)}\setminus s_h^{*(t-1)}} w_i^{*(t)}\right)\right.\right. \\ \quad \left.\times\left(\sum_{i\in s_h^{(t)}\cap s_h^{*(t-1)}} w_i^{*(t)}\mathbf{y}_i^{*(t)} \Big/ \sum_{i\in s^{(t)}\cap s^{*(t-1)}} w_i^{*(t)}\right)\right] \\ \quad + \left[\left(\left(\sum_{i\in s_h^{(t)}} w_i^{(t)} - \sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\right) \Big/ \sum_{i\in s_h^{*(t)}\setminus s_h^{*(t-1)}} w_i^{*(t)}\right)\right. \\ \quad \left.\left.\times\left(\sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\mathbf{y}_i^{*(t-1)} \Big/ \sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\right)\right]\right\}, & \text{si } i \in s_h^{*(t)} \setminus s_h^{*(t-1)}. \end{cases} \tag{3.5}$$

où $R_h^{(t-1,t)} = \left(\sum_{i\in s_h^{(t-1)}} w_i^{(t-1)} \Big/ \sum_{i\in s_h^{(t)}} w_i^{(t)}\right)$ est un facteur de correction appliqué aux valeurs RM1, RM2 et RMP pour tenir compte de la variation relative de la taille de la population dans la strate h entre la période $t-1$ et la période t. Les autres ajustements des valeurs RM2 et RMP ont été effectués pour s'assurer que l'estimateur HT pour les « pseudo-variables auxiliaires composites » $\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)} = \sum_{h=1}^{H}\sum_{i\in s_h^{*(t)}} w_i^{*(t)}\mathbf{z}_i^{*(t)}$ à la période t soit sans biais pour les variables d'enquête clés correspondantes $\mathbf{Y}^{(t-1)}$ à la période $t-1$. Une simple preuve de l'absence de biais dans l'estimateur HT pour les « pseudo-variables auxiliaires composites » est donnée à l'annexe.

L'estimateur HT $\hat{Y}_{\mathrm{HT}}^{*(t)} = \sum_{h=1}^{H}\sum_{i\in s_h^{*(t)}} w_i^{*(t)} y_i^{*(t)}$ est équivalent à $\hat{Y}_{\mathrm{HT}}^{(t)} = \sum_{h=1}^{H}\sum_{i\in s_h^{(t)}} w_i^{(t)} y_i^{(t)}$ puisque les « pseudo-valeurs » pour la variable d'intérêt sont égales à zéro pour les unités échantillonnées additionnelles provenant des populations de « nouvelles unités » et d'« unités disparues ». De même, l'estimateur RG $\hat{Y}_{\mathrm{RG}}^{*(t)} = \sum_{h=1}^{H}\sum_{i\in s_h^{*(t)}} \tilde{w}_i^{*(t)} y_i^{*(t)}$ est équivalent à $\hat{Y}_{\mathrm{RG}}^{(t)} = \sum_{h=1}^{H}\sum_{i\in s_h^{(t)}} \tilde{w}_i^{(t)} y_i^{(t)}$ puisque les « pseudo-valeurs » pour la variable d'intérêt et les variables auxiliaires sont égales à zéro pour les unités échantillonnées additionnelles provenant des populations de « nouvelles unités » et d'« unités disparues ». Cependant, l'estimateur RM $\hat{Y}_{\mathrm{RM}}^{*(t)} = \sum_{h=1}^{H}\sum_{i\in s_h^{*(t)}} \breve{w}_i^{*(t)} y_i^{*(t)}$ n'est pas équivalent à $\hat{Y}_{\mathrm{RM}}^{(t)} = \sum_{h=1}^{H}\sum_{i\in s_h^{(t)}} \breve{w}_i^{(t)} y_i^{(t)}$ puisque les « pseudo-valeurs » pour les variables auxiliaires composites ne sont pas égales à zéro pour les unités échantillonnées additionnelles provenant des populations de « nouvelles unités » et d'« unités disparues ».

La procédure proposée d'ajout des « nouvelles unités » à la population de la période précédente et d'ajout des « unités disparues » à la population de la période courante est exécutée indépendamment à chaque période, de sorte qu'il n'y a pas d'accumulation de « nouvelles unités » et d'« unités disparues » dans la « pseudo-population » au cours du temps.

4 Étude par simulation

Une étude par simulation Monte Carlo a été réalisée afin d'examiner la performance de l'estimateur par la régression composite proposé. Dix populations artificielles ont été créées pour cette étude. En premier lieu, une population de base (population I) a été générée de manière qu'elle ait l'aspect physique des enquêtes-entreprises mensuelles types réalisées sur une période de cinq ans. Ensuite, six populations supplémentaires (populations II à VII) ont été générées en modifiant l'une des six caractéristiques clés de la population de base afin de déterminer si cette caractéristique particulière avait une incidence sur la performance de l'estimateur par la régression composite proposé. Enfin, trois populations supplémentaires (populations VIII à X) ont été générées pour examiner l'effet des variables auxiliaires sur la performance de l'estimateur par la régression composite proposé. Une brève description des dix populations artificielles figure au tableau 4.1.

Les totaux de population à la période t pour les diverses populations artificielles ont été produits en utilisant le modèle chronologique :

$$Y^{(t)} = T^{(t)} + \alpha_2 S^{(t)} + \alpha_3 I^{(t)}$$

où $T^{(t)}, S^{(t)}$ et $I^{(t)}$ sont la tendance, la composante saisonnière et la composante irrégulière de la série chronologique données par :

$$T^{(t)} = 1000 + 5(t-1) + 50(1 - \cos(\pi(t-1)/18))$$

$$S^{(t)} = 25[\sin(\pi t/6) - \cos(\pi t/6) + \cos(\pi t/3)]$$

$$I^{(t)} = 25\varepsilon^{(t)}$$

où $\alpha_2 = 1$ pour toutes les populations artificielles, sauf la population II (série à forte composante saisonnière) où $\alpha_2 = 4$, et $\alpha_3 = 1$ pour toutes les populations artificielles, sauf la population III (série à forte composante irrégulière) où $\alpha_3 = 4$, et $\varepsilon^{(t)} \sim N(0,1)$. La série originale $(T^{(t)} + S^{(t)} + I^{(t)})$, la série désaisonnalisée $(T^{(t)} + I^{(t)})$ et la tendance $(T^{(t)})$ de la série pour la population artificielle de base sont présentées à la figure 4.1.

Tableau 4.1
Description des populations artificielles

Populations artificielles	Description de la population
Population I	Série de base
Population II	Série à forte composante saisonnière
Population III	Saisie à forte composante irrégulière
Population IV	Série à fort renouvellement de la population
Population V	Série à fort renouvellement de l'échantillon
Population VI	Série à forte variation des unités
Population VII	Série à faible corrélation des unités
Population VIII	Série à corrélation de base des variables auxiliaires
Population IX	Série à corrélation forte des variables auxiliaires
Population X	Série à corrélation faible des variables auxiliaires

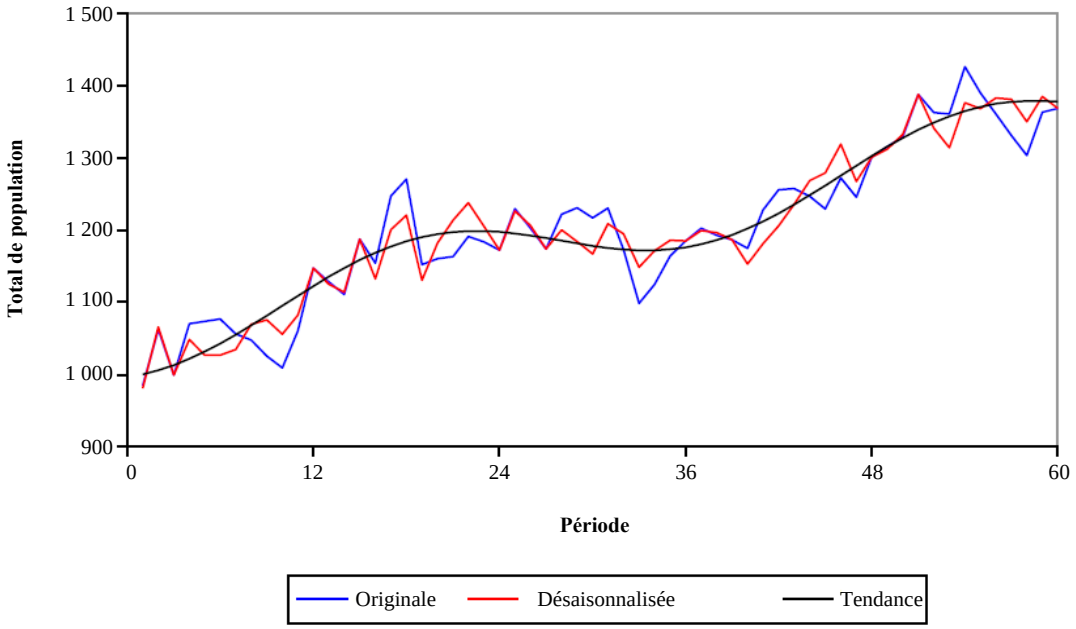


Figure 4.1 Série chronologique pour la population I

Les dix populations artificielles ont été partitionnées en cinq strates; quatre strates à tirage partiel $(h = 1,\ldots,4)$ et une strate à tirage complet $(h = 5)$. Les tailles de population de strate à la période t ont été choisies comme étant $N_h^{(t)} = N_h\left[1 + 0,5\left(T^{(t)}/T^{(1)} - 1\right)\right]$, où N_h est la population de strate pour toutes les populations artificielles à la période 1, sélectionnée pour donner une population asymétrique souvent associée à l'entreprise type.

Les taux prévus de renouvellement de la population entre la période $t-1$ et la période t, en raison de l'ajout des « nouvelles unités » et de la suppression des « unités disparues », ont été spécifiés comme étant $\alpha_4(1 - R_h)$, où R_h est la probabilité qu'une unité soit traitée comme « disparue » de la population pour la population artificielle de base à toute période. Une valeur de $\alpha_4 = 1$ a été utilisée pour toutes les populations artificielles, sauf la population IV (série à fort renouvellement de la population) où $\alpha_4 = 2$ a été utilisé. Les tailles d'échantillon de strate à la période t ont été fixées à $n_h^{(t)} = n_h$ pour les strates à tirage partiel, et à $n_h^{(t)} = N_h^{(t)}$ pour la strate à tirage complet, où n_h est la population de la strate à la période 1.

Les taux planifiés de renouvellement de l'échantillon entre la période $t-1$ et la période t ont été spécifiés comme étant $\alpha_5(1 - r_h)$, où r_h est égal à l'inverse du nombre de cycles consécutifs de l'enquête durant lesquels il est attendu qu'une unité soit incluse dans l'échantillon en l'absence d'un renouvellement de la population, pour la population artificielle de base à toute période (p. ex., un taux planifié de renouvellement de l'échantillon de 0,0417 est égal à 24 cycles d'enquête). Une valeur de $\alpha_5 = 1$ a été utilisée pour toutes les populations artificielles, sauf la population V (série à fort renouvellement de l'échantillon) où $\alpha_5 = 2$ a été utilisé. Les taux réels de renouvellement de l'échantillon dépendront de ce renouvellement planifié de l'échantillon ainsi que de tout renouvellement

non planifié de l'échantillon causé par le renouvellement de la population. Les taux prévus de renouvellement de la population et le taux planifié de renouvellement des échantillons de strate ont été sélectionnés de manière à obtenir des taux de renouvellement de la population et de l'échantillon similaires à ceux souvent observés dans les enquêtes-entreprises types.

Les moyennes de strate et les variances de population de strate à la période t ont été spécifiées respectivement comme étant $\bar{y}_h^{(t)} = 0,2\left(Y^{(t)}/N_h^{(t)}\right)$ et $S_h^{(t)2} = \alpha_6 S_h^2\left(\bar{y}_h^{(t)}/\bar{y}_h^{(t)}\right)^2$ avec $\alpha_6 = 1$ pour toutes les populations artificielles, sauf la population VI (série à forte variation des unités) où $\alpha_6 = 4$. Les corrélations de population de strate entre la période t et la période $t-k$ ont été définies en utilisant un modèle de décroissance exponentielle, $\rho\left(y_h^{(t)}, y_h^{(t-k)}\right) = \exp(-0,02\alpha_7 k)$ avec $\alpha_7 = 1$ pour toutes les populations artificielles, sauf la population VII (série à faible corrélation des unités) où $\alpha_7 = 4$. Les corrélations de population de strate entre la variable d'intérêt et la variable auxiliaire à la période t ont été définies comme étant $\rho\left(x_h^{(t)}, y_h^{(t)}\right) = 1 - \alpha_8(1-\rho_h)$ avec $\alpha_8 = 1$ pour la population VIII (série à corrélation de base des variables auxiliaires), $\alpha_8 = 0,5$ pour la population IX (série à corrélation élevée des variables auxiliaires), $\alpha_8 = 1,5$ pour la population X (série à corrélation faible des variables auxiliaires) et sans objet pour toutes les autres populations artificielles.

La variable d'intérêt $y_{hi}^{(t)}$ et les variables auxiliaires $x_{hi}^{(t)}$ pour l'unité i dans la strate h à la période t ont été générées à partir de lois lognormales multivariées de moyenne $\bar{y}_h^{(t)}$, de variance $S_h^{(t)2}$ et de coefficient de corrélation $\rho\left(y_h^{(t)}, y_h^{(t-k)}\right)$. Les caractéristiques de N_h, n_h, R_h, r_h et S_h^2 au niveau de la strate sont les valeurs présentées au tableau 4.2.

Un total de $S = 10\,000$ simulations indépendantes ont été exécutées pour chacune des dix populations artificielles. Dans chacune de ces simulations, des échantillons aléatoires stratifiés $s_h^{(t)}$ de taille $n_h^{(t)}$ ont été tirés de la population $U_h^{(t)}$ par une méthode de sélection basée sur des nombres aléatoires permanents (NAP) à chaque période, $t = 1, \ldots, 60$. À chaque période, $t > 1$, on a déterminé les « pseudo-populations », $U_h^{*(t-1)}$ et $U_h^{*(t)}$, et les « pseudo-échantillons », $s_h^{*(t-1)}$ et $s_h^{*(t)}$, et évalué les divers estimateurs RM. Ceux-ci englobaient l'estimateur RM1 $(\alpha = 0)$, l'estimateur RM2 $(\alpha = 1)$, l'estimateur RM en utilisant $\alpha = 0,25, 0,5$ et $0,75$, l'estimateur RMP et l'estimateur RMC, avec un compromis entre l'estimateur HT et l'estimateur RMP pour les populations I à VII et entre l'estimateur RG et l'estimateur RMP pour les populations VIII à X, en utilisant $\alpha = 0,25, 0,5$ et $0,75$.

Tableau 4.2
Caractéristiques des strates

h	N_h	R_h	n_h	r_h	S_h^2	ρ_h
S1	8 000	0,0150	12	0,042	0,4	0,85
S2	1 600	0,0125	18	0,042	3	0,75
S3	320	0,0100	24	0,042	20	0,65
S4	64	0,0075	30	0,000	125	0,55
S5	16	0,0025	16	0,000	625	0,95

Pour comparer la performance des divers estimateurs RM pour les estimations ponctuelles et les estimations des variations, on s'est basé sur les biais relatifs et les efficacités relatives par rapport à l'estimateur HT pour toutes les populations artificielles ainsi que par rapport à l'estimateur RG pour les

populations VIII à X. Les biais relatifs et les efficacités relatives de la variable d'intérêt y à la période t pour les estimations ponctuelles et les estimations des variations ont été calculés comme il suit :

$$\begin{aligned}
\mathrm{BR}\left(\hat{Y}^{(t)}\right) &= \frac{1}{Y^{(t)}}\left[\frac{1}{S}\sum_{s=1}^{S}\left(\hat{Y}_s^{(t)} - Y^{(t)}\right)\right] \\
\mathrm{BR}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right) &= \frac{1}{Y^{(t-1)}}\left[\frac{1}{S}\sum_{s=1}^{S}\left(\left(\hat{Y}_s^{(t)} - \hat{Y}_s^{(t-1)}\right) - \left(Y^{(t)} - Y^{(t-1)}\right)\right)\right] \\
\mathrm{ER}\left(\hat{Y}^{(t)}\right) &= \mathrm{EQM}\left(\hat{Y}_*^{(t)}\right)\Big/\mathrm{EQM}\left(\hat{Y}^{(t)}\right) \\
\mathrm{ER}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right) &= \mathrm{EQM}\left(\hat{Y}_*^{(t)} - \hat{Y}_*^{(t-1)}\right)\Big/\mathrm{EQM}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right)
\end{aligned}$$

où $\hat{Y}_s^{(t)}$ est l'estimateur pour la variable d'intérêt y à la période t pour le s^{e} échantillon de simulation, $\hat{Y}_*^{(t)}$ est l'estimateur HT ou RG de la variable d'intérêt y à la période t, et $\mathrm{EQM}\left(\hat{Y}^{(t)}\right)$ et $\mathrm{EQM}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right)$ sont les erreurs quadratiques moyennes de la variable d'intérêt y à la période t pour les estimations ponctuelles et les estimations des variations données par :

$$\begin{aligned}
\mathrm{EQM}\left(\hat{Y}^{(t)}\right) &= \frac{1}{S}\sum_{s=1}^{S}\left(\hat{Y}_s^{(t)} - Y^{(t)}\right)^2 \\
\mathrm{EQM}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right) &= \frac{1}{S}\sum_{s=1}^{S}\left(\left(\hat{Y}_s^{(t)} - \hat{Y}_s^{(t-1)}\right) - \left(Y^{(t)} - Y^{(t-1)}\right)\right)^2.
\end{aligned}$$

Les biais relatifs des estimations ponctuelles (moyenne sur 12 mois pour chacune des cinq années) pour les estimateurs RM1, RM2 et RMP pour la population I (série de base) sont présentés au tableau 4.3. Les estimateurs RM proposés (RM1-P, RM2-P, RMP-P) ont été comparés aux estimateurs RM courants (RM1-C, RM2-C, RMP-C) et aux estimateurs RM ajustés (RM1-A, RM2-A, RMP-A), où un facteur de correction a été appliqué aux valeurs RM pour tenir compte de la variation relative de la taille de la population dans la strate h entre la période $t-1$ et la période t.

Tableau 4.3
Biais relatif moyen (%) des estimations ponctuelles pour la population I

	Année 1	Année 2	Année 3	Année 4	Année 5
HT	0,024	-0,032	-0,015	-0,003	-0,005
RM1-C	-0,909	-2,871	-2,292	-2,836	-4,122
RM2-C	-0,918	-3,432	-3,449	-4,502	-6,820
RMP-C	-0,919	-3,437	-3,458	-4,515	-6,839
RM1-A	0,064	-0,129	0,002	-0,062	-0,068
RM2-A	0,169	0,024	0,039	-0,109	-0,317
RMP-A	0,152	-0,027	-0,014	-0,174	-0,410
RM1-P	0,009	-0,066	-0,040	-0,051	-0,054
RM2-P	0,022	-0,053	-0,028	-0,039	-0,034
RMP-P	0,020	-0,056	-0,030	-0,039	-0,036

Les estimateurs RM courants présentent un biais négatif important qui s'accumule au cours du temps. Même si l'estimateur RM ajusté élimine en majeure partie ces biais, les estimateurs RM2-A et RMP-A présentent encore un petit biais négatif qui s'accumule au cours du temps. Par ailleurs, les biais relatifs des estimateurs RM proposés sont négligeables et leur grandeur ne semble pas varier sur la période de cinq ans.

Le tableau 4.4 donne les biais relatifs absolus et les efficacités relatives (moyenne sur 12 mois pour chacune des cinq années) des estimateurs pour la population I (série de base). Les biais relatifs absolus moyens des estimations ponctuelles et des estimations des variations sont négligeables pour tous les estimateurs et on ne note aucune variation appréciable de la grandeur du biais relatif pour aucun des estimateurs au cours de la période de cinq ans. Pour les estimations ponctuelles, l'estimateur RM1 donne de meilleurs résultats que l'estimateur HT, tandis que les estimateurs RM2 et RMP donnent de moins bons résultats que l'estimateur HT. L'efficacité relative des estimateurs RM2 et RMP diminue considérablement au cours de la période de cinq ans, ce qui donne à penser que ces estimateurs sont susceptibles de présenter le problème de « dérive ». La présence du problème de « dérive » est évidente si l'on observe la relation entre les estimations ponctuelles au début de la première année $(t = 1)$ et celles au début de la troisième année $(t = 25)$ pour les échantillons de simulation (figure 4.2).

On peut voir qu'il existe des corrélations positives entre les estimations ponctuelles au début de la première et de la troisième années pour les estimateurs RM1, RM2, RMP et RM $(\alpha = 0,75)$, ce qui signifie que si ces estimateurs s'écartent fortement des vrais totaux de population, il est probable qu'ils s'en écarteront davantage au fil du temps. Bien que les corrélations pour l'estimateur RM1 soient plus faibles que pour l'estimateur RM2, elles sont néanmoins positives, ce qui signifie que l'estimateur RM1 n'est pas exempt du problème de dérive. Les corrélations positives ne s'observent pas pour les estimateurs HT et RMC $(\alpha = 0,25)$, donc ces estimateurs ne sont pas sujets au problème de « dérive ». De surcroît, il est clair que les estimateurs RM2, RMP et RM $(\alpha = 0,75)$ sont nettement plus variables que les estimateurs HT, RM1 et RMC $(\alpha = 0,25)$ au début de la troisième année.

Tableau 4.4
Biais relatif absolu moyen (%) et efficacité relative moyenne (%) pour la population I

	Estimations ponctuelles					Estimations des variations				
	Année 1	Année 2	Année 3	Année 4	Année 5	Année 1	Année 2	Année 3	Année 4	Année 5
					Biais relatif absolu moyen (%)					
HT	0,031	0,032	0,030	0,025	0,010	0,021	0,011	0,012	0,019	0,014
RM1	0,032	0,066	0,041	0,051	0,054	0,021	0,011	0,010	0,010	0,016
RM2	0,024	0,053	0,030	0,039	0,034	0,014	0,009	0,009	0,009	0,013
RM($\alpha = 0,25$)	0,029	0,067	0,045	0,058	0,063	0,019	0,010	0,009	0,009	0,015
RM($\alpha = 0,50$)	0,027	0,066	0,045	0,060	0,064	0,017	0,010	0,009	0,009	0,014
RM($\alpha = 0,75$)	0,025	0,061	0,040	0,054	0,055	0,016	0,009	0,009	0,009	0,014
RMP	0,023	0,056	0,032	0,040	0,036	0,014	0,009	0,009	0,009	0,013
RMC($\alpha = 0,25$)	0,027	0,041	0,025	0,018	0,011	0,016	0,009	0,010	0,009	0,014
RMC($\alpha = 0,50$)	0,028	0,036	0,028	0,021	0,010	0,018	0,008	0,011	0,010	0,014
RMC($\alpha = 0,75$)	0,029	0,033	0,029	0,024	0,010	0,019	0,008	0,011	0,014	0,014
					Efficacité relative moyenne (%)					
HT	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
RM1	122,0	126,0	118,4	112,7	114,6	137,6	132,8	132,7	134,2	133,0
RM2	92,4	74,7	57,7	47,8	45,8	223,0	203,0	206,5	206,4	204,8
RM($\alpha = 0,25$)	121,6	123,4	110,6	100,9	100,9	168,3	158,4	159,7	160,7	159,2
RM($\alpha = 0,50$)	115,3	110,0	92,8	80,9	79,3	199,0	182,8	185,6	186,0	184,3
RM($\alpha = 0,75$)	104,7	91,9	73,5	62,0	59,7	220,4	199,6	203,5	203,4	201,6
RMP	94,1	79,6	63,0	53,4	53,1	223,3	203,3	206,9	206,8	204,8
RMC($\alpha = 0,25$)	110,8	113,7	113,1	113,7	113,1	198,5	182,7	186,5	187,1	184,4
RMC($\alpha = 0,50$)	106,0	105,9	105,6	105,9	105,5	164,2	155,0	157,3	157,6	155,8
RMC($\alpha = 0,75$)	102,7	102,4	102,3	102,4	102,3	130,9	127,4	128,3	128,4	127,6

Un choix approprié de α pour les estimateurs RMC minimisera la probabilité du problème de « dérive ». Comparativement à l'estimateur RMP, cet estimateur RMC $(\alpha = 0{,}25)$ améliorera l'efficacité des estimations ponctuelles, mais réduira l'efficacité des estimations des variations. Pour les estimations des variations, l'estimateur RM1 donne d'un peu meilleurs résultats que l'estimateur HT, tandis que les estimateurs RM2 et RMP produisent de considérablement meilleurs résultats que l'estimateur HT. Dans l'ensemble, la performance de l'estimateur RMC semble être un peu meilleure que celle de l'estimateur RM. Si l'objectif est de choisir un estimateur qui n'est pas trop susceptible de présenter le problème de « dérive » et qui maximise l'efficacité des estimations des variations sans aucune perte d'efficacité relative des estimations ponctuelles, alors le « meilleur » estimateur pour cette population particulière est l'estimateur RMC avec $\alpha \approx 0{,}10$. Cet estimateur aura vraisemblablement une dérive minimale et donnera un gain d'efficacité modéré de 21,6 % pour les estimations ponctuelles et un gain d'efficacité important de 104,2 % pour les estimations des variations.

Les biais relatifs absolus moyens et les efficacités relatives moyennes des estimateurs pour les populations I à VII sont présentés au tableau 4.5. Les augmentations importantes de la saisonnalité (population II) ou de l'irrégularité (population III) de la série chronologique n'ont presque aucun effet sur la performance des divers estimateurs pour les estimations ponctuelles. Alors que de petites réductions de l'efficacité relative des estimations des variations sont observées pour les estimateurs RM2 et RMP, aucun effet n'est constaté pour l'estimateur RM1.

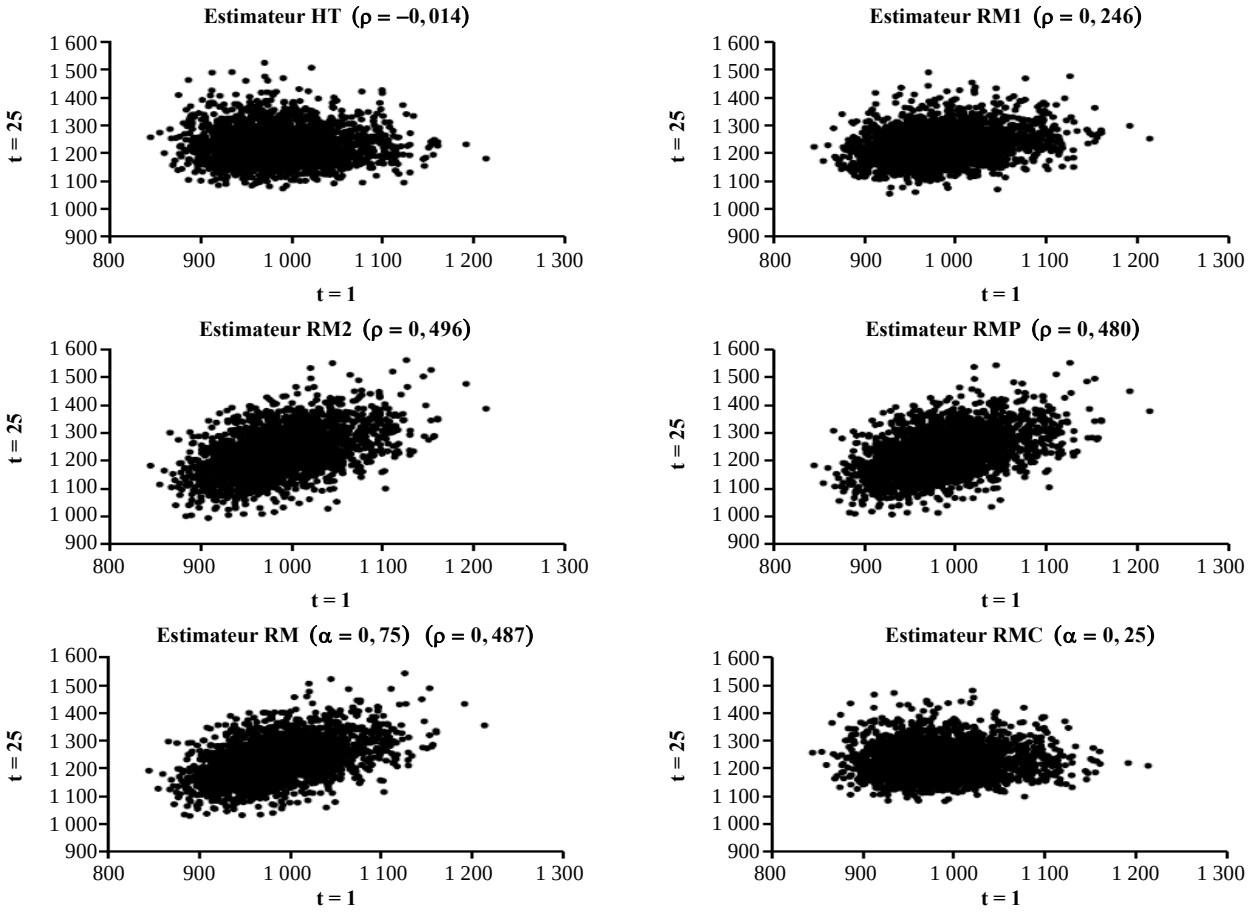


Figure 4.2 Représentation graphique de divers estimateurs pour la population I

Tableau 4.5
Biais relatif absolu moyen (%) et efficacité relative moyenne (%)

	Estimations ponctuelles							Estimations des variations						
	Pop. I	Pop. II	Pop. III	Pop. IV	Pop. V	Pop. VI	Pop. VII	Pop. I	Pop. II	Pop. III	Pop. IV	Pop. V	Pop. VI	Pop. VII
Biais relatif absolu moyen (%)														
HT	0,038	0,027	0,049	0,048	0,048	0,065	0,032	0,017	0,012	0,016	0,018	0,020	0,025	0,020
RM1	0,050	0,098	0,074	0,052	0,089	0,150	0,078	0,014	0,012	0,013	0,015	0,020	0,020	0,018
RM2	0,081	0,028	0,039	0,063	0,047	0,218	0,120	0,012	0,011	0,011	0,014	0,013	0,017	0,017
RM($\alpha = 0,25$)	0,052	0,083	0,070	0,046	0,095	0,139	0,090	0,013	0,011	0,012	0,014	0,018	0,018	0,017
RM($\alpha = 0,50$)	0,057	0,058	0,059	0,043	0,089	0,136	0,103	0,012	0,010	0,011	0,014	0,016	0,016	0,017
RM($\alpha = 0,75$)	0,066	0,038	0,047	0,050	0,069	0,160	0,111	0,012	0,010	0,011	0,014	0,014	0,016	0,017
RMP	0,074	0,032	0,045	0,065	0,055	0,223	0,124	0,012	0,011	0,011	0,014	0,013	0,017	0,017
RMC($\alpha = 0,25$)	0,034	0,023	0,046	0,049	0,049	0,059	0,034	0,012	0,010	0,012	0,015	0,015	0,018	0,017
RMC($\alpha = 0,50$)	0,037	0,025	0,048	0,049	0,050	0,064	0,033	0,014	0,011	0,014	0,017	0,017	0,023	0,019
RMC($\alpha = 0,75$)	0,038	0,026	0,048	0,048	0,049	0,065	0,032	0,015	0,012	0,015	0,018	0,019	0,025	0,019
Efficacité relative moyenne (%)														
HT	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
RM1	118,7	119,6	118,9	126,4	143,5	127,2	98,9	134,2	133,4	133,9	132,9	147,2	138,0	115,5
RM2	59,6	60,9	58,1	64,2	49,7	67,8	48,7	208,9	192,6	180,0	202,0	455,7	226,2	137,0
RM($\alpha = 0,25$)	110,8	112,0	110,4	119,8	134,2	121,5	89,2	161,6	159,3	158,5	159,0	215,0	169,3	125,7
RM($\alpha = 0,50$)	93,6	95,0	92,4	101,4	99,4	103,8	74,6	188,0	182,1	178,2	183,7	315,4	201,0	133,5
RM($\alpha = 0,75$)	75,0	76,4	73,5	80,6	69,0	83,8	60,2	206,1	194,9	186,3	200,0	424,9	222,4	137,5
RMP	65,3	66,6	63,7	76,8	52,9	74,0	53,7	209,2	194,6	183,3	202,4	454,8	225,6	137,2
RMC($\alpha = 0,25$)	112,9	111,9	112,2	114,5	151,9	112,7	107,5	188,2	183,7	181,4	184,8	347,1	193,7	134,9
RMC($\alpha = 0,50$)	105,8	105,4	105,5	107,2	123,3	105,7	104,5	158,3	156,0	154,4	156,6	223,8	160,5	126,2
RMC($\alpha = 0,75$)	102,4	102,3	102,3	103,0	109,1	102,4	102,1	128,6	127,9	127,2	128,1	149,7	129,5	114,6

Les nombres additionnels de « nouvelles unités » et d'« unités disparues » dans la population (population IV) donnent lieu à de petits gains d'efficacité relative des estimations ponctuelles pour tous les estimateurs par la régression modifiée, en raison des réductions de l'EQM de ces estimateurs. Alors que de petites pertes d'efficacité relative des estimations des variations sont constatées pour les estimateurs RM2 et RMP, aucun effet n'est observé pour l'estimateur RM1. Le fait de doubler l'ampleur du renouvellement non planifié de l'échantillon (population V) entraîne une augmentation de l'efficacité relative des estimations ponctuelles pour l'estimateur RM1, mais une diminution de cette efficacité relative pour les estimateurs RM2 et RMP. Des améliorations importantes de l'efficacité relative des estimations des variations sont observées pour tous les estimateurs par la régression modifiée, en raison d'une plus forte augmentation de l'EQM de l'estimateur HT comparativement aux estimateurs par la régression modifiée.

Une plus forte variation des valeurs déclarées par les unités (population VI) entraîne de petits gains d'efficacité relative des estimations ponctuelles pour tous les estimateurs par la régression modifiée, principalement attribuables à de plus fortes augmentations de l'EQM de l'estimateur HT comparativement aux estimateurs par la régression modifiée. Cependant, aucun effet sur l'efficacité relative des estimations des variations n'est constaté, car la grandeur des augmentations de l'EQM pour les estimateurs par la régression modifiée est semblable à celle observée pour l'estimateur HT. Une faible corrélation des valeurs déclarées par les unités au cours du temps (population VII) produit de grandes réductions de l'efficacité relative des estimations ponctuelles et des estimations des variations.

Pour l'ensemble des populations I à VII, l'estimateur RM1 donne de meilleurs résultats que les estimateurs RM2 et RMP pour les estimations ponctuelles, tandis que les estimateurs RM2 et RMP produisent de meilleurs résultats que l'estimateur RM1 pour les estimations des variations. Le « meilleur » estimateur en ce qui concerne la maximisation de l'efficacité relative des estimations des variations sans aucune perte d'efficacité relative des estimations ponctuelles est l'estimateur RMC, quoique la « meilleure » valeur de α diffère d'une population artificielle à l'autre.

Les biais relatifs absolus moyens et les efficacités relatives moyennes des estimateurs pour les populations VIII à X sont présentés au tableau 4.6. Par rapport à l'estimateur HT, l'utilisation de variables auxiliaires dans les estimateurs donne lieu à des gains importants d'efficacité relative des estimations ponctuelles et des estimations des variations pour tous les estimateurs par la régression modifiée. Le gain d'efficacité relative des estimations ponctuelles ainsi que des estimations des variations est d'autant plus important que la corrélation est forte entre la variable d'intérêt et la variable auxiliaire. Toutefois, dans le cas de l'estimateur RG, l'utilisation de variables auxiliaires dans les estimateurs produit de très petits gains d'efficacité relative des estimations ponctuelles, mais des gains modérés d'efficacité relative des estimations des variations pour la plupart des estimateurs par la régression modifiée. Le gain d'efficacité relative des estimations ponctuelles ainsi que des estimations des variations est d'autant plus faible que la corrélation est forte entre la variable d'intérêt et la variable auxiliaire.

Tableau 4.6
Biais relatif absolu moyen (%) et efficacité relative moyenne (%)

	Estimations ponctuelles			Estimations des variations		
	Pop. VIII	Pop. IX	Pop. X	Pop. VIII	Pop. IX	Pop. X
			Biais relatif absolu moyen (%)			
RG	0,021	0,014	0,020	0,010	0,008	0,011
RM1	0,042	0,041	0,044	0,016	0,015	0,016
RM2	0,032	0,026	0,031	0,014	0,013	0,014
RM($\alpha = 0,25$)	0,043	0,037	0,044	0,015	0,014	0,015
RM($\alpha = 0,50$)	0,041	0,034	0,040	0,015	0,014	0,015
RM($\alpha = 0,75$)	0,035	0,029	0,034	0,015	0,013	0,014
RMP	0,036	0,028	0,034	0,014	0,013	0,014
RMC($\alpha = 0,25$)	0,023	0,017	0,023	0,013	0,011	0,013
RMC($\alpha = 0,50$)	0,022	0,016	0,022	0,012	0,010	0,013
RMC($\alpha = 0,75$)	0,021	0,015	0,021	0,011	0,009	0,012
			Efficacité relative moyenne (%) par rapport à l'estimateur HT			
RG	256,4	428,9	183,3	169,7	215,3	140,2
RM1	258,9	421,5	191,1	166,8	198,0	150,5
RM2	265,8	436,0	194,4	218,7	247,5	202,2
RM($\alpha = 0,25$)	263,8	428,3	194,9	184,4	213,7	168,7
RM($\alpha = 0,50$)	267,6	434,7	197,4	202,5	230,5	186,9
RM($\alpha = 0,75$)	268,6	438,1	197,3	215,9	244,0	199,8
RMP	266,5	437,5	194,6	216,3	245,8	199,2
RMC($\alpha = 0,25$)	266,7	441,2	192,6	225,7	257,7	204,7
RMC($\alpha = 0,50$)	265,3	442,0	190,3	217,3	254,4	191,6
RMC($\alpha = 0,75$)	261,4	437,0	187,0	197,5	239,7	168,6
			Efficacité relative moyenne (%) par rapport à l'estimateur RG			
RG	100,0	100,0	100,0	100,0	100,0	100,0
RM1	101,0	98,3	104,2	98,3	92,0	107,4
RM2	103,7	101,6	106,1	128,9	115,0	144,3
RM($\alpha = 0,25$)	102,9	99,9	106,3	108,7	99,3	120,3
RM($\alpha = 0,50$)	104,4	101,3	107,7	119,3	107,1	133,3
RM($\alpha = 0,75$)	104,8	102,1	107,7	127,2	113,3	142,5
RMP	103,9	102,0	106,1	127,4	114,2	142,1
RMC($\alpha = 0,25$)	104,0	102,9	105,1	133,0	119,7	146,0
RMC($\alpha = 0,50$)	103,5	103,1	103,8	128,0	118,2	136,7
RMC($\alpha = 0,75$)	102,0	101,9	102,0	116,4	111,3	120,3

5 Conclusion

Le présent article décrit l'extension d'un certain nombre d'estimateurs par la régression modifiée aux enquêtes-entreprises avec bases de sondage évoluant au cours du temps en raison de l'ajout des « nouvelles entreprises » et de la suppression des « entreprises disparues ». Les résultats de l'étude par simulation indiquent que la grandeur du biais de ces divers estimateurs par la régression modifiée est négligeable. Le « meilleur » estimateur est l'estimateur par la régression modifiée de compromis qui produit d'importants gains d'efficacité des estimations ponctuelles et des estimations des variations, et élimine la probabilité du problème de « dérive » moyennant le choix approprié de α.

Remerciements

Les points de vue exprimés dans cet article sont ceux de l'auteur et ne reflètent pas nécessairement ceux de l'*Australian Bureau of Statistics* (ABS). L'auteur tient à remercier un examinateur anonyme et le rédacteur associé de leurs précieux commentaires ainsi que Dr Robert Clark de l'Université de Woollongong pour ses suggestions constructives lors d'une version antérieure de ce manuscrit.

Annexe

Les valeurs espérées de l'estimateur HT pour les « pseudo-variables auxiliaires composites » $\hat{\mathbf{Z}}_h^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{RM2})i}^{*(t)}$ à la période t sont données par :

$$
\begin{aligned}
E\left[\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)}\right] &= E\left[\sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{RM2})i}^{*(t)}\right] \\
&= \sum_{h=1}^{H} E\left[\sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{RM2})i}^{*(t)}\right] \\
&= \sum_{h=1}^{H} N_h^{(t-1)} E\left[\left(\sum_{i \in s_h^{*(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t-1)} \Big/ \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right)\right] \\
&- \sum_{h=1}^{H} N_h^{(t-1)} E\left[\left(\sum_{i \in s_h^{*(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)}\right)\right] \\
&+ \sum_{h=1}^{H} N_h^{(t-1)} \left(\sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)}} w_i^{(t)}\right) E\left[\sum_{i \in s_h^{*(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)}\right] \\
&+ \sum_{h=1}^{H} N_h^{(t-1)} \left(\sum_{i \in s_h^{(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)}} w_i^{(t)}\right) E\left[\sum_{i \in s_h^{*(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)}\right] \\
&= \sum_{h=1}^{H} N_h^{(t-1)} \bar{\mathbf{Y}}_h^{(t-1)} - \sum_{h=1}^{H} N_h^{(t-1)} \bar{\mathbf{Y}}_h^{(t)} + \sum_{h=1}^{H} N_h^{(t-1)} \bar{\mathbf{Y}}_h^{(t)} \\
&= \sum_{h=1}^{H} N_h^{(t-1)} \bar{\mathbf{Y}}_h^{(t-1)} \\
&= \sum_{h=1}^{H} \mathbf{Y}_h^{(t-1)} \\
&= \mathbf{Y}^{(t-1)}.
\end{aligned}
$$

Les valeurs espérées de l'estimateur HT pour les « pseudo-variables auxiliaires composites » $\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{RMP})i}^{*(t)}$ à la période t sont données par :

$$
\begin{aligned}
E\left[\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)}\right] &= E\left[\sum_{h=1}^{H}\sum_{i\in s_h^{*(t)}} w_i^{*(t)}\mathbf{z}_{(\mathrm{RMP})i}^{*(t)}\right] \\
&= \sum_{h=1}^{H} E\left[\sum_{i\in s_h^{*(t)}} w_i^{*(t)}\mathbf{z}_{(\mathrm{RMP})i}^{*(t)}\right].
\end{aligned}
$$

Si le « pseudo-échantillon » de la strate h à la période t ne contient pas d'unités qui n'étaient pas incluses dans le « pseudo-échantillon » de la strate h à la période $t-1$ $\left(s_h^{*(t)} \setminus s_h^{*(t-1)} = \varnothing\right)$:

$$
\begin{aligned}
E\left[\sum_{i\in s_h^{*(t)}} w_i^{*(t)}\mathbf{z}_{(\mathrm{RMP})i}^{*(t)}\right] &= N_h^{(t-1)} E\left[\left(\sum_{i\in s_h^{*(t)}\cap s_h^{*(t-1)}} w_i^{*(t)}\mathbf{y}_i^{*(t-1)} \Big/ \sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\right)\right] \\
&= N_h^{(t-1)}\overline{\mathbf{Y}}_h^{(t-1)} \\
&= \mathbf{Y}_h^{(t-1)}
\end{aligned}
$$

et si le « pseudo-échantillon » de la strate h à la période t contient des unités qui n'étaient pas incluses dans le « pseudo-échantillon » de la strate h à la période $t-1$ $\left(s_h^{*(t)} \setminus s_h^{*(t-1)} \neq \varnothing\right)$:

$$
\begin{aligned}
E\left[\sum_{i\in s_h^{*(t)}} w_i^{*(t)}\mathbf{z}_{(\mathrm{RMP})i}^{*(t)}\right] &= N_h^{(t-1)}\left(\sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)} \Big/ \sum_{i\in s_h^{(t)}} w_i^{(t)}\right) \\
&\times E\left[\sum_{i\in s_h^{*(t)}\cap s_h^{*(t-1)}} w_i^{*(t)}\mathbf{y}_i^{*(t-1)} \Big/ \sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\right] \\
&+ N_h^{(t-1)}\left(\sum_{i\in s_h^{(t)}\setminus s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i\in s_h^{(t)}} w_i^{(t)}\right) \\
&\times E\left[\sum_{i\in s_h^{*(t)}\setminus s_h^{*(t-1)}} w_i^{*(t)}\mathbf{y}_i^{*(t)} \Big/ \sum_{i\in s_h^{(t)}\setminus s_h^{*(t-1)}} w_i^{*(t)}\right] \\
&- N_h^{(t-1)}\left(\sum_{i\in s_h^{(t)}\setminus s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i\in s_h^{(t)}} w_i^{(t)}\right) \\
&\times E\left[\left(\sum_{i\in s_h^{(t)}\cap s_h^{*(t-1)}} w_i^{*(t)}\mathbf{y}_i^{*(t)} \Big/ \sum_{i\in s_h^{(t)}\cap s_h^{*(t-1)}} w_i^{*(t)}\right)\right] \\
&+ N_h^{(t-1)} E\left[\left(\sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\mathbf{y}_i^{*(t-1)} \Big/ \sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\right)\right] \\
&- N_h^{(t-1)}\left(\sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)} \Big/ \sum_{i\in s_h^{(t)}} w_i^{(t)}\right) \\
&\times E\left[\left(\sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\mathbf{y}_i^{*(t-1)} \Big/ \sum_{i\in s_h^{*(t)}\cap s_h^{(t-1)}} w_i^{*(t)}\right)\right] \\
&= N_h^{(t-1)}\overline{\mathbf{Y}}_h^{(t-1)} \\
&= \mathbf{Y}_h^{(t-1)}
\end{aligned}
$$

d'où $E\left[\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)}\right] = \sum_{h=1}^{H}\mathbf{Y}_h^{(t-1)} = \mathbf{Y}^{(t-1)}$.

Bibliographie

Australian Bureau of Statistics (ABS) (2012a). Australian demographic statistics, June Quarter 2012, Numéro de catalogue 3101.0.

Australian Bureau of Statistics (ABS) (2012b). Business indicators, June Quarter 2012, Numéro de catalogue 5676.0.

Australian Bureau of Statistics (ABS) (2012c). Counts of Australian businesses, including entries and exits, June 2008 to June 2012, Numéro de catalogue 8165.0.

Beaumont, J.-F., et Bocci, C. (2005). A refinement of the regression composite estimator in the Labour Force Survey for change estimates. *Proceedings of the Survey Methods Section,* SSC Annual Meeting, Juin 2005.

Bell, P. (1999). Comparison of alternative LFS estimators – Issues for discussion. *Methodology Advisory Committee*, Octobre 1999.

Bell, P. (2001). Comparaison d'autres estimateurs pour l'Enquête sur la population active. *Techniques d'enquête*, 27, 1, 57-68.

Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 1, 49-56.

Gambino, J., Kennedy, B. et Singh, M.P. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada : évaluation et application. *Techniques d'enquête*, 27, 1, 69-79.

Gurney, M., et Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 242-257.

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

Singh, A.C. (1994). Sampling design based estimating functions for finite population totals. Présentation invitée, *Abstracts of the Annual Meeting of the Statistical Society of Canada*, Banff, Alberta, 8 au 11 mai, 48.

Singh, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 120-129.

Singh, A.C., et Merkouris, P. (1995). Composite estimation by modified regression for repeated surveys. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 420-425.

Singh, A.C., Kennedy, B. et Wu, S. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquête*, 27, 1, 35-48.

Singh, A.C., Kennedy, B., Wu, S. et Brisebois, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 300-305.

Yansaneh, I.S., et Fuller, W.A. (1998). Méthode optimal d'estimation récursive pour les enquêtes répétitives. *Techniques d'enquête*, 24, 1, 33-42.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 101-130
Statistique Canada, n° 12-001-X au catalogue

Exploration de la récursion pour les estimateurs optimaux sous renouvellement de l'échantillon en cascade

Jan Kowalski et Jacek Wesołowski[1]

Résumé

Nous nous intéressons à l'estimation linéaire optimale des moyennes pour des éditions subséquentes d'une enquête sous renouvellement de l'échantillon, où l'évolution temporelle des échantillons est conçue selon un schéma en cascade. Depuis la publication de l'article fondamental de Patterson (1950), on sait que, si les unités n'ont pas le droit de revenir dans l'échantillon après en être sorties pendant une certaine période (pas d'intervalles dans les schémas de renouvellement), la récursion en une étape tient pour l'estimateur optimal. Cependant, dans certaines enquêtes réelles importantes, par exemple, la *Current Population Survey* aux États-Unis ou l'Enquête sur la population active dans de nombreux pays européens, les unités reviennent dans l'échantillon après en avoir été absentes pendant plusieurs éditions de l'enquête (existence d'intervalles dans les schémas de renouvellement). Le cas échéant, la question de la forme de la récurrence pour l'estimateur optimal devient considérablement plus difficile. Ce problème n'a pas encore été résolu. On a plutôt élaboré des approches sous-optimales de rechange, comme l'estimation composite K (voir, par exemple, Hansen, Hurwitz, Nisselson et Steinberg (1955)), l'estimation composite AK (voir, par exemple, Gurney et Daly (1965)) ou l'approche des séries chronologiques (voir, par exemple, Binder et Hidiroglou (1988)).

Dans le présent article, nous surmontons cette difficulté de longue date, autrement dit, nous présentons des formules de récurrence analytiques pour l'estimateur linéaire optimal de la moyenne pour des schémas de renouvellement contenant des intervalles. Ces formules sont obtenues sous certaines conditions techniques, à savoir l'HYPOTHÈSE I et l'HYPOTHÈSE II (des expériences numériques donnent à penser que ces hypothèses pourraient être universellement satisfaites). Pour atteindre l'objectif, nous élaborons une approche par opérateurs algébriques qui permet de réduire le problème de récursion pour l'estimateur linéaire optimal à deux questions : 1) la localisation des racines (éventuellement complexes) d'un polynôme Q_p défini en fonction du schéma de renouvellement (le polynôme Q_p s'exprime de façon pratique au moyen de polynômes de Tchebychev de la première espèce) et 2) le rang d'une matrice S définie en fonction du schéma de renouvellement et des racines du polynôme Q_p. En particulier, nous montrons que l'ordre de la récurrence est égal à un plus la taille de l'intervalle le plus grand dans le schéma de renouvellement. Nous donnons les formules exactes de calcul des coefficients de récurrence – naturellement, pour les utiliser il faut confirmer (dans de nombreux cas, numériquement) que les HYPOTHÈSES I et II sont satisfaites. Nous illustrons la solution à l'aide de plusieurs exemples de schémas de renouvellement tirés d'enquêtes réelles.

Mots-clés : Enquêtes répétées; renouvellement de l'échantillon; récurrence pour le BLUE de la moyenne courante; polynômes de Tchebychev; algèbre des opérateurs de translation; corrélation exponentielle.

1 Introduction

Les bureaux de la statistique et d'autres institutions utilisent fréquemment des enquêtes répétées avec renouvellement des éléments dans les échantillons. Le renouvellement préconçu de (groupes d') éléments selon une forme de schéma en cascade, c'est-à-dire des scénarios où, à chaque édition de l'enquête, l'élément (le groupe d'éléments) « le plus ancien » quitte l'échantillon et est remplacé par un nouveau, est également d'usage très répandu, mais souvent, l'information que contiennent les données d'enquête n'est pas pleinement exploitée. Cela, à son tour, entraîne la construction d'estimateurs sous-optimaux dont la variance est plus grande que le minimum réalisable. Afin d'accroître l'utilisation d'estimateurs optimaux dans les scénarios de renouvellement, dans un article fondamental, Patterson (1950) a introduit la notion

1. Jan Kowalski, École polytechnique de Varsovie, Varsovie, Pologne; Jacek Wesołowski, École polytechnique de Varsovie et Bureau central de la statistique, Varsovie, Pologne. Courriel : J.Wesolowski@mini.pw.edu.pl.

de récurrence pour calculer les meilleurs estimateurs linéaires sans biais (BLUE) de la moyenne à chaque édition de l'enquête. Les principales hypothèses étaient que les moyennes de population inconnues sont déterministes et que les réponses sont des variables aléatoires dont la variance et la structure de corrélation sont entièrement connues. Sous corrélation exponentielle et en supposant en outre que tout élément qui quitte l'échantillon n'y revient pas, Patterson a prouvé que, pour toute édition t de l'enquête, l'estimateur BLUE $\hat{\mu}_t$ de la moyenne courante μ_t au temps t (basé sur toutes les observations passées) peut être calculé à partir de la récurrence en une étape suivante :

$$\hat{\mu}_t = a_1(t)\hat{\mu}_{t-1} + \underline{r}_0^T(t)\underline{X}_t + \underline{r}_1^T(t)\underline{X}_{t-1} \tag{1.1}$$

où $\underline{X}_i$ est le vecteur des observations au temps $i = t,\ t-1$. Les formules pour les coefficients de récurrence, c'est-à-dire les nombres $a_1(t)$ et les vecteurs $\underline{r}_0(t), \underline{r}_1(t)$, étaient donnés dans cet article également. (Ici et tout au long de l'exposé, un vecteur, disons $\underline{r}$, s'entend d'une colonne, et $\underline{r}^T$ est sa transposée. Pour deux vecteurs $\underline{r} = (r_1, \ldots, r_n)$, $\underline{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$, l'expression $\underline{r}^T\underline{w} = \sum_{i=1}^n r_i w_i$ est simplement le produit scalaire de $\underline{r}$ et $\underline{w}$.)

L'hypothèse de Patterson selon laquelle *une unité qui quitte un échantillon ne revient jamais dans l'enquête* était au cœur de son approche. Si cette hypothèse est violée (c'est-à-dire, s'il existe des intervalles dans le schéma de renouvellement), on sait depuis des années que de sérieuses difficultés se posent si l'on cherche un analogue de la récurrence (1.1). Sachant cela (voir, par exemple, Yansaneh et Fuller 1998), les chercheurs ont plutôt essayé des approches de rechange : l'estimateur composite K classique a été proposé par Hansen et coll. (1955). Ses propriétés d'optimalité ont été développées dans Rao et Graham (1964) et, plus récemment, dans Ciepiela, Gniado, Wesołowski et Wojtyś (2012). La principale différence est que, au lieu de rechercher la récurrence pour l'estimateur BLUE, ces auteurs restreignent le problème d'optimalité aux estimateurs linéaires sans biais satisfaisant juste la récurrence d'ordre un, c'est-à-dire qu'ils minimisent la variance de l'estimateur basée sur l'estimateur le plus récent et les observations provenant des deux dernières éditions de l'enquête seulement. Des ajustements, appelés estimateurs composites AK, introduits dans Gurney et Daly (1965), ont été élaborés, par exemple, dans Cantwell (1988, 1990) et dans Cantwell et Caldwell (1998) – en fait dans ces articles, les auteurs introduisent la notion de plan équilibré à plusieurs niveaux, et un plan à un niveau correspond au schéma en cascade que nous considérons ici. Une autre approche basée sur l'estimateur par régression composite a été examinée dans Bell (2001), Fuller et Rao (2001), ainsi que Singh, Kennedy et Wu (2001) (avec des implications pour l'Enquête sur la population active du Canada).

La difficulté que pose l'estimation récursive dans le cas d'enquêtes répétées pour des schémas avec intervalles était soulignée dans Yansaneh et Fuller (1998), qui ont analysé les variances des estimateurs composites sous plusieurs scénarios de renouvellement de l'échantillon. Pour une description relativement à jour de l'état de l'art en la matière, le lecteur peut consulter Steel et McLaren (2008), en particulier la section IV sur les différents schémas de renouvellement et la section V sur les estimateurs composites. Des comparaisons de l'efficacité sous différents schémas en cascade figurent dans McLaren et Steel (2000) et dans Steel et McLaren (2002). Un article très récent sur l'estimation optimale sous renouvellement de l'échantillon est celui de Towhidi et Namazi-Rad (2010). Certains des articles susmentionnés traitent aussi de l'approche par série chronologique (qui n'est pas examinée dans le présent article) dans laquelle les moyennes inconnues sont traitées comme des quantités aléatoires – un aperçu de

cette approche est donné dans Binder et Hidiroglou (1988). Pour un développement plus récent de cette approche, voir, par exemple, Lind (2005).

Quant à l'approche originale de Patterson, le résultat suivant concernant la forme récursive de l'estimateur BLUE a été présenté dans Kowalski (2009), où des intervalles uniques dans le schéma de renouvellement étaient permis. Comme dans Patterson (1950), cet article était consacré à la situation « classique » dans laquelle les coefficients de l'équation (1.2) présentée plus bas peuvent dépendre de t. Trois conclusions découlant de ces travaux ont une incidence sur le présent article. Premièrement, il était suggéré que la formule (1.1) pouvait être généralisée à un scénario de renouvellement arbitraire (comprenant des intervalles) en intégrant les estimateurs optimaux et les observations provenant d'un nombre probablement plus grand (mais encore aussi petit que possible) d'éditions antérieures de l'enquête et que l'ordre de la récurrence devrait dépendre de la taille de l'intervalle le plus grand. Deuxièmement, il était observé que la corrélation exponentielle, telle que supposée dans Patterson (1950), est essentielle à l'obtention de la représentation récursive et qu'il est plausible de se limiter à la classe des schémas « en cascade ». Ces hypothèses sont toutes deux retenues plus bas. Enfin, puisque selon les simulations numériques, les coefficients de récurrence semblent être rapidement convergents quand $t \to \infty$, il a été proposé de considérer le cas « limite » de la configuration « classique », dans lequel les coefficients de récurrence ne varient pas au cours du temps.

Nous insistons sur le fait que, dans le présent article, *n'importe quel ensemble d'intervalles est permis dans le schéma de renouvellement en cascade*. L'objectif est de montrer que la récurrence

$$\hat{\mu}_t = a_1\hat{\mu}_{t-1} + \ldots + a_p\hat{\mu}_{t-p} + \underline{r}_0^T \underline{X}_t + \underline{r}_1^T \underline{X}_{t-1} + \ldots + \underline{r}_p^T \underline{X}_{t-p} \tag{1.2}$$

est vérifiée pour tout schéma de renouvellement en cascade et de trouver l'ordre de la récurrence p, les coefficients numériques $a_1, \ldots, a_p$ et les coefficients vectoriels $\underline{r}_0, \ldots, \underline{r}_p$. Soulignons que la représentation (1.2) est « stationnaire » en ce sens que ni l'ordre de la récurrence p ni les coefficients de récurrence (a_i) et $(\underline{r}_i)$ ne dépendent de t.

Notre résultat principal est la réduction du problème de récurrence à l'analyse d'un certain polynôme Q_p (de degré p, où $p-1$ est la taille de l'intervalle le plus grand dans le schéma de renouvellement) et à la question de l'obtention d'une solution unique pour un certain système linéaire d'équations, qui dépend des racines de Q_p. Heureusement, il se fait que le polynôme Q_p s'exprime de façon pratique au moyen de polynômes de Tchebychev de la première espèce. Nous fournissons une condition suffisante en ce qui concerne les propriétés de localisation des racines de Q_p pour l'existence de la forme récursive de l'estimateur BLUE d'ordre p, donnée en (1.2), et dérivons des formules explicites (exploitant les racines de Q_p) pour les coefficients de récurrence (a_i) et $(\underline{r}_i)$. Les formes des coefficients dépendent aussi de la solution unique du système linéaire susmentionné. Les résultats sont illustrés au moyen de plusieurs exemples tirés d'enquêtes réelles.

La convergence des coefficients de récurrence que nous avons observée numériquement dans de nombreux schémas « classiques » (c'est-à-dire, avec les coefficients dans l'expression analogue de (1.2) dépendant de t) de complexités diverses indique que la solution d'un tel problème de récurrence « stationnaire » devrait être universelle (en fait, cette convergence n'est prouvée formellement que dans le cas de Patterson, $p = 1$). S'il en est ainsi, elle peut être traitée comme une solution approximative pour le

scénario « classique ». Comme le lecteur le constatera, cette intuition est largement confirmée dans le présent article. Notre résultat principal n'est toujours pas universel, même dans les modèles avec corrélation exponentielle. Notre approche s'appuie fortement sur deux hypothèses (HYPOTHÈSE I et HYPOTHÈSE II ci-dessous) qui nous permettent d'affirmer que la récurrence (1.2) est vérifiée. Néanmoins, nous avons exécuté de nombreuses expériences numériques pour différents schémas de renouvellement et différentes valeurs de la corrélation qui, toutes, laissent entendre que ces hypothèses peuvent toutes deux être universellement satisfaites. Malheureusement, à l'heure actuelle, nous sommes incapables de confirmer théoriquement ces observations.

Le plan de l'article est le suivant. À la section 2, nous présentons en termes mathématiques notre modèle de travail. À la section 3, nous présentons nos deux hypothèses de base et formulons le résultat principal de l'étude. La section 4 contient des exemples d'applications du résultat principal à plusieurs scénarios de renouvellement souvent utilisés. La section 5 présente une discussion. Le corps principal de l'exposé mathématique est reporté à la section 6. Dans la première partie, 6.1, nous examinons les propriétés algébriques des opérateurs de translation. Elles sont essentielles à la preuve de la formule de récurrence qui est donnée dans la deuxième partie, 6.2, de l'annexe.

2 Modèle

Soit $\left(X_{i,j}\right)_{i,j\in\mathbb{Z}}$ une matrice doublement infinie de variables aléatoires. Du point de vue heuristique, $X_{i,j}$ représente la valeur de la variable $\mathcal{X}$ mesurée pour l'unité (le groupe de renouvellement) i lors de l'édition j de l'enquête. Nous supposons que l'espérance de $X_{i,j}$ dépend uniquement de l'édition de l'enquête et non de l'unité, c'est-à-dire

$$\mathbb{E}X_{i,j} = \mu_j, \quad \forall i, j \in \mathbb{Z}.$$

En outre, nous supposons que les corrélations entre les $X_{i,j}$ sont exponentielles en temps pour la même unité et qu'il n'existe pas de corrélation entre des unités différentes (à l'instar du modèle de Patterson (1950)), c'est-à-dire

$$\mathbb{C}\mathrm{ov}\left(X_{i,j}, X_{k,l}\right) = \rho^{|j-l|}\delta_{i,k} \quad \forall i, j, k, l \in \mathbb{Z},$$

où $|\rho| \in (0,1)$ et $\delta_{i,k} = 1$ si $i = k$, et $\delta_{i,k} = 0$ autrement. (Dans les situations pratiques, ρ est souvent compris dans $[0,1)$. Dans le cas où $\rho = 0$, les observations passées ne peuvent pas améliorer l'estimateur linéaire présent de la moyenne, de sorte que nous ne considérons pas ce cas plus loin.) Conséquemment,

$$\mathbb{V}\mathrm{ar}\, X_{i,j} = 1, \quad i, j \in \mathbb{Z}.$$

Pour tout $j \in \mathbb{Z}$, nous sommes intéressés par l'estimateur BLUE de μ_j basé sur toutes les observations disponibles provenant des éditions $i \le j$ de l'enquête. Pour un entier positif fixe N, désignons par

$$\underline{X}_j = \left(X_{j,j}, X_{j+1,j}, \ldots, X_{j+N-1,j}\right)^T$$

l'*échantillon maximal* (de taille N) lors de l'édition $j \in \mathbb{Z}$. Alors

$$\mathbb{E}\underline{X}_j = \mu_j \underline{1}, \quad j \in \mathbb{Z},$$

où $\underline{1} = (1,1,\ldots,1)^T \in \mathbb{R}^N$, et

$$\mathbb{C}\mathrm{ov}\left(\underline{X}_j, \underline{X}_{j-k}\right) = \mathbf{C}^k = \left[\mathbb{C}\mathrm{ov}\left(\underline{X}_j, \underline{X}_{j+k}\right)\right]^T, \quad j \in \mathbb{Z}, k \geq 0,$$

où $\mathbf{C}$ est une matrice de dimensions $N \times N$ de la forme

$$\mathbf{C} = \begin{bmatrix} 0 & \rho & & 0 \\ 0 & \ddots & \ddots & \\ & \ddots & \ddots & \rho \\ 0 & & 0 & 0 \end{bmatrix}.$$

Notons que $\mathbf{C}^n = 0$ pour tout $n \geq N$.

L'échantillon effectif sera défini par un *schéma en cascade*, qui est un vecteur $\underline{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^T \in \{0,1\}^N$ avec $\varepsilon_1 = \varepsilon_N = 1$. Soit

$$n = \sum_{j=1}^{N} \varepsilon_j \text{ et } h = N - n.$$

Soit H l'ensemble de zéros dans le schéma $\underline{\varepsilon}$, c'est-à-dire $j \in H$ ssi $\varepsilon_j = 0$. Manifestement, $\#H = h$. Un intervalle de taille m est un ensemble maximal de m zéros séquentiels, c'est-à-dire un ensemble qui satisfait

$$\{j, j+1, \ldots, j+m-1\} \subset H \text{ et } j-1, j+m \notin H.$$

Par conséquent, H est une union de, disons, s intervalles de tailles $m_r, r = 1,2,\ldots,s$ et $\sum_{r=1}^{s} m_r = h$.

La couverture p du schéma (voir Kowalski 2009 pour la définition équivalente) est la taille de l'intervalle le plus grand augmentée de un :

$$p = 1 + \max_{1 \leq r \leq s} m_r.$$

Lors de chaque édition de l'enquête $j \in \mathbb{Z}$, nous pouvons ne pas observer l'échantillon maximal $\underline{X}_j$, mais l'*échantillon effectif* de taille n défini par le schéma en cascade $\underline{\varepsilon}$, autrement dit le vecteur

$$\underline{Y}_j = \left(X_{j+k-1,j}, k \in \{1,\ldots,N\} \setminus H\right)^T,$$

c'est-à-dire que les valeurs des $X_{i,j}$ représentés par des zéros (intervalles) dans le schéma en cascade ε sont supprimées de l'échantillon.

Considérons l'estimateur BLUE $\hat{\mu}_t$ de la moyenne μ_t lors de l'édition $t \in \mathbb{Z}$ de l'enquête qui est basé sur les observations $\underline{Y}_j$, $j \leq t$. C'est-à-dire

$$\hat{\mu}_t = \sum_{i=0}^{\infty} \underline{\tilde{w}}_i^T \underline{Y}_{t-i}$$

avec $\underline{\tilde{w}}_i \in \mathbb{R}^n, i \geq 0,$ qui minimise $\mathbb{V}\text{ar}\hat{\mu}_t$ sous les contraintes d'absence de biais

$$\underline{\tilde{w}}_0^T \underline{1} = 1 \text{ et } \underline{\tilde{w}}_i^T \underline{1} = 0, i \geq 1.$$

Il est à la fois évident et crucial pour notre approche que, de façon équivalente,

$$\hat{\mu}_t = \sum_{i=0}^{\infty} \underline{w}_i^T \underline{X}_{t-i} \tag{2.1}$$

avec $\underline{w}_i \in \mathbb{R}^N, i \geq 0,$ minimisant $\mathbb{V}\text{ar}\hat{\mu}_t$ sous les contraintes d'absence de biais

$$\underline{w}_0^T \underline{1} = 1, \underline{w}_i^T \underline{1} = 0, i \geq 1, \tag{2.2}$$

et les contraintes de schéma en cascade

$$\underline{w}_i^T \underline{e}_j = 0 \quad \forall i \geq 0, \forall j \in H, \tag{2.3}$$

où $\underline{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ (avec 1 à la j^{e} position) est le j^{e} vecteur de la base canonique dans $\mathbb{R}^N$, $j \in H$. Notons que la contrainte (2.3) dit effectivement que les j^{es} entrées $(j \in H)$ des vecteurs $\underline{w}_i, i \geq 0,$ sont toutes nulles.

3 Récurrence

Afin de formuler notre résultat principal qui donne la récurrence exacte pour les estimateurs BLUE sous n'importe quel schéma de renouvellement de l'échantillon, nous devons introduire deux objets : un polynôme Q_p et une matrice $\mathbf{S}$. Ils ont tous deux un aspect très technique et ne possèdent pas d'interprétation heuristique directe. Néanmoins, ils semblent être d'une importance essentielle pour la formule de récurrence finale.

3.1 Polynôme Q_p

Rappelons que T_k, le k^{e} polynôme de Tchebychev de la première espèce, est défini par

$$T_k(x) = \cos(k \arccos x), \quad k = 0, 1, \ldots.$$

Définissons une fonction polynomiale d'une matrice de Toeplitz symétrique de dimensions $m \times m$ $\mathbf{T}_m$ par

$$\mathbf{T}_m = \begin{bmatrix} T_0 & T_1 & T_2 & \cdots & T_{m-2} & T_{m-1} \\ T_1 & T_0 & T_1 & \cdots & T_{m-3} & T_{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ T_{m-2} & T_{m-3} & T_{m-4} & \cdots & T_0 & T_1 \\ T_{m-1} & T_{m-2} & T_{m-3} & \cdots & T_1 & T_0 \end{bmatrix} \tag{3.1}$$

et une matrice tridiagonale inversible de dimensions $m \times m$

$$\mathbf{R}_m = \begin{bmatrix} 1+\rho^2 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1+\rho^2 \end{bmatrix}. \tag{3.2}$$

Notons que $\mathbf{R}_m$ est non singulière.

Pour un schéma en cascade $\underline{\varepsilon}$ avec tailles d'intervalles $m_1, \ldots, m_s$ et couverture p, définissons un polynôme Q_p par

$$Q_p(x) = (N-1)\left(1+\rho^2-2\rho x\right) + 1 - \rho^2 - \left(1+\rho^2-2\rho x\right)^2 \sum_{j=1}^{s} \operatorname{tr}\left(\mathbf{T}_{m_j}(x)\,\mathbf{R}_{m_j}^{-1}\right). \tag{3.3}$$

Puisque $\operatorname{tr}\left(\mathbf{T}_m(x)\,\mathbf{R}_m^{-1}\right)$ est un polynôme de degré $m-1$ en x,

$$\deg Q_p = 2 + \max_{1 \le j \le s}\left(m_j - 1\right) = p.$$

3.2 Matrice S

Considérons de nouveau un schéma en cascade $\underline{\varepsilon}$ avec couverture p et $\#(H) = h = m_1 + \ldots + m_s$. Pour les nombres complexes $d_1, \ldots, d_p$, définissons une matrice $\mathbf{S}$ de dimensions $(ph+h+1) \times p(h+1)$ au moyen de sa structure par blocs

$$\mathbf{S} = \mathbf{S}\left(d_1, \ldots, d_p\right) = \begin{bmatrix} \widetilde{\mathbf{G}}(d_1) & \widetilde{\mathbf{G}}(d_2) & \cdots & \widetilde{\mathbf{G}}(d_p) \\ \mathbf{G}(d_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(d_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}(d_p) \end{bmatrix}. \tag{3.4}$$

Les blocs $\widetilde{\mathbf{G}}(d_i)$ sont les matrices de dimensions $(h+1) \times (h+1)$

$$\widetilde{\mathbf{G}}(d) = \frac{1}{1-\rho^2}\begin{bmatrix} (N-1)(1-d\rho)+1-\rho^2 & (1-d\rho)\,\underline{1}_h^T \\ (1-d\rho)\,\underline{1}_h & \operatorname{diag}\left(\widetilde{\mathbf{H}}_{m_1},\ldots,\widetilde{\mathbf{H}}_{m_s}\right) \end{bmatrix} \tag{3.5}$$

avec $\widetilde{\mathbf{H}}_m = \widetilde{\mathbf{H}}_m(d)$ étant une matrice bidiagonale supérieure de dimensions $m \times m$

$$\widetilde{\mathbf{H}}_m(d) = \begin{bmatrix} 1 & -d\rho & & \\ & \ddots & \ddots & \\ & & \ddots & -d\rho \\ & & & 1 \end{bmatrix}. \tag{3.6}$$

Les blocs $\mathbf{G}(d_i)$ sont les matrices de dimensions $h \times (h+1)$

$$\mathbf{G}(d) = \frac{1}{1-\rho^2}\left[(1-d\rho)(d-\rho)\,\underline{1}_h, d\ \operatorname{diag}\left(\mathbf{H}_{m_1},\ldots,\mathbf{H}_{m_s}\right)\right], \tag{3.7}$$

où $\mathbf{H}_m = \mathbf{H}_m(d)$ est une matrice tridiagonale de dimensions $m \times m$

$$\mathbf{H}_m(d) = \begin{bmatrix} 1+\rho^2 & -d\rho & & \\ -\rho/d & \ddots & \ddots & \\ & \ddots & \ddots & -d\rho \\ & & -\rho/d & 1+\rho^2 \end{bmatrix}. \tag{3.8}$$

Les nombres $d_1,\ldots,d_p$ considérés plus haut sont reliés aux racines (potentiellement complexes) $x_1,\ldots,x_p$ du polynôme Q_p par la relation $2x_i = d_i + 1/d_i$, et $|d_i| < 1,\ i = 1,\ldots,p$. Certains détails supplémentaires sont donnés dans la remarque qui suit.

Remarque 3.1 *Soit* $x \in \mathbb{C}$ *telle que* $\Im x \neq 0$ ou $\Re x \notin [-1,1]$.

Alors l'équation

$$\frac{1}{2}\left(d + \frac{1}{d}\right) = x$$

en d *possède exactement deux racines, disons,* $d_+(x)$ *et* $d_-(x)$, *telles que*

$$|d_-(x)| < 1 \quad et \quad |d_+(x)| > 1.$$

Si, en outre, $\Im x = 0$, *alors* $d_+(x)$ *et* $d_-(x)$ *sont réelles.*

En désignant par x^* *la conjuguée complexe de* x *avec* $\Im x \neq 0$. *Alors*

$$d_-(x) = \left(d_-\left(x^*\right)\right)^* \quad et \quad d_+(x) = \left(d_+\left(x^*\right)\right)^*.$$

3.3 Résultat principal

Notre résultat principal donne la récursion de profondeur égale à la couverture p du schéma en cascade, ainsi que les formes analytiques des coefficients qui sont prêtes pour l'implémentation numérique. Des exemples réels de ce genre d'implémentations sont présentés à la section 4. La preuve que nous offrons (voir l'annexe) est fondée sur deux hypothèses fondamentales concernant le polynôme Q_p et la matrice $\mathbf{S}$.

HYPOTHÈSE I : Le polynôme Q_p possède des racines distinctes $x_1, \ldots, x_p \notin [-1, 1]$.

HYPOTHÈSE II : La matrice $\mathbf{S} = \mathbf{S}(d_1, \ldots, d_p)$, où $d_i = d_-(x_i), i = 1, \ldots, p$ est de plein rang.

Théorème 3.1 *Si les* HYPOTHÈSES I *et* II *sont satisfaites, alors pour tout* $t \in \mathbb{Z}$, *la récursion*

$$\hat{\mu}_t = \sum_{k=1}^{p} a_k \hat{\mu}_{t-k} + \sum_{k=0}^{p} \underline{r}_k^T \underline{X}_{t-k} \tag{3.9}$$

est vérifiée avec

$$a_k = (-1)^{k+1} \sum_{1 \le j_1 < \ldots < j_k \le p} d_{j_1} \ldots d_{j_k}, \quad k = 1, \ldots, p, \tag{3.10}$$

et

$$\underline{r}_i = \sum_{m=1}^{p} \left[\left(v_i(d_m)\mathbf{I} - v_{i-1}(d_m)\mathbf{C}^T \right) \mathbf{\Delta N}(d_m) \sum_{j \in H'} c_{j,m} \underline{e}_j \right], \quad i = 0, 1, \ldots, p,$$

où $\underline{e}_0 = \underline{1}, H' = \{0\} \bigcup H, v_0(d) = 1, v_{-1}(d) = 0,$

$$v_i(d) = d^i - \sum_{l=1}^{i} a_l d^{i-l}, \quad i = 1, \ldots, p, \tag{3.11}$$

$\Delta = \left(\mathbf{I} - \mathbf{C}\mathbf{C}^T \right)^{-1}, \mathbf{N}(d) = \mathbf{I} - d\mathbf{C}$ *et avec*

$$\underline{c} = \left[\left(c_{j,1}, j \in H' \right), \left(c_{j,2}, j \in H' \right), \ldots, \left(c_{j,p}, j \in H' \right) \right]^T$$

*étant la solution unique (elle existe en raison de l'*HYPOTHÈSE II*) du système linéaire*

$$\mathbf{S}\underline{c} = (1, 0, \ldots, 0)^T \in \mathbb{R}^{ph+h+1}.$$

En outre,

$$\mathbb{V}\mathrm{ar}(\hat{\mu}_t) = \sum_{m=1}^{p} c_{0,m}. \tag{3.12}$$

À la section suivante, nous montrons comment le résultat théorique susmentionné peut être appliqué dans plusieurs scénarios de base, en particulier, dans ceux qui sont utilisés dans les enquêtes réelles, tandis que la preuve du théorème 3.1 est donnée à la deuxième partie, 6.2, de l'annexe. Elle est fondée sur une approche basée purement sur des opérateurs algébriques qui est présentée à la première partie, 6.1, de l'annexe.

Nous insistons sur le fait que des expériences numériques intensives donnent à penser que les HYPOTHÈSES I et II peuvent être universellement satisfaites, mais qu'en ce moment, nous n'avons pas de preuve mathématique de ce fait (sauf dans les cas $p = 1, 2$ et $p = 3$ pour un schéma spécial de renouvellement de l'échantillon). Donc, les applications de la formule de récurrence données plus haut (pour $p > 2$) dans les enquêtes doivent être précédées d'une vérification numérique (qui est assez simple) que les HYPOTHÈSES I et II sont satisfaites. Des exemples sont donnés à la section 4.

4 Exemples

4.1 Scénario de Patterson, $p = 1$

Le scénario en cascade de Patterson est utilisé, par exemple, pour réaliser l'Enquête sur la population active en Australie ($N = n = 8$, voir Australian Bureau of Statistics (2002)) et au Canada ($N = n = 6$, voir Singh, Drew, Gambino et Mayda (1990)). Il n'y a pas de zéros dans le schéma, d'où $h = 0$ et le polynôme $Q_p = Q_1$, voir (3.3), ne contient pas l'opérande de somme avec la trace, c'est-à-dire

$$Q_1(x) = (N-1)\left(1 + \rho^2 - 2\rho x\right) + 1 - \rho^2.$$

Sa seule racine $x_1 = -\dfrac{1+\rho^2}{2\rho} - \dfrac{1-\rho^2}{2(N-1)\rho}$ est réelle et satisfait $|x_1| > \dfrac{1+\rho^2}{2|\rho|} > 1$, c'est-à-dire que l'HYPOTHÈSE I est satisfaite. Il donne aussi $d_1 = d_-(x_1)$ réelle de la forme

$$d_1 = \frac{N + (N-2)\rho^2 - \sqrt{\left[N + (N-2)\rho^2\right]^2 - 4(N-1)^2\rho^2}}{2(N-1)\rho}.$$

En outre, $\mathbf{S}$ définie en (3.4) est une matrice de dimensions 1×1 de la forme $\mathbf{S} = \left[(N-1)\dfrac{1-d_1\rho}{1-\rho^2} + 1\right] \neq \mathbf{0}$, c'est-à-dire que l'HYPOTHÈSE II est vérifiée trivialement. Donc, en vertu du théorème 3.1, pour tout $t \in \mathbb{Z}$, nous avons

$$\hat{\mu}_t = a_1\hat{\mu}_{t-1} + \underline{r}_0^T \underline{X}_t + \underline{r}_1^T \underline{X}_{t-1},$$

où

$$\begin{cases} a_1 = d_1 \\ \underline{r}_0 = c_{0,1}\mathbf{N}(d_1)\,\underline{1} \\ \underline{r}_1 = -c_{0,1}\mathbf{C}^T\mathbf{N}(d_1)\,\underline{1} \end{cases},$$

où

$$c_{0,1} = \frac{1}{(N-1)\dfrac{1-d_1\rho}{1-\rho^2}+1}.$$

En prenant par exemple $N = 6$ et $\rho = 0{,}9,$ nous obtenons pour tout t :

$$\hat{\mu}_t = 0{,}7942\hat{\mu}_{t-1} + \begin{bmatrix} 0{,}1765 \\ 0{,}1765 \\ 0{,}1765 \\ 0{,}1765 \\ 0{,}1765 \\ 0{,}1176 \end{bmatrix}^T \underline{X}_t + \begin{bmatrix} 0{,}0000 \\ -0{,}1588 \\ -0{,}1588 \\ -0{,}1588 \\ -0{,}1588 \\ -0{,}1588 \end{bmatrix}^T \underline{X}_{t-1}.$$

Remarque 4.1 *Patterson (1950) a considéré le même scénario dans le modèle « classique ». Il a prouvé formellement que le coefficient de récurrence $a_1(t)$ converge quand $t \to \infty$ et a montré que la limite était a_1 telle que donnée plus haut. Les vecteurs $\underline{r}_0(t)$ et $\underline{r}_1(t)$, étant des fonctions continues de $a_1(t)$, convergent vers $\underline{r}_0$ et $\underline{r}_1$, respectivement. Autrement dit, la solution « stationnaire » est en effet en harmonie avec l'asymptotique de la solution « classique ».*

4.2 Scénarios avec intervalles de taille 1, $p = 2$

Le polynôme $Q_p = Q_2,$ voir (3.3), a la forme suivante :

$$Q_2(x) = -\frac{4h\rho^2}{1+\rho^2}x^2 - 2(N-2h-1)\rho x + (N-h-1)\left(1+\rho^2\right) + 1 - \rho^2.$$

Comme $1-\rho^2 > 0,$ on voit immédiatement que son discriminant

$$\Delta = 4(N-2h-1)^2\rho^2 + 4\frac{4h\rho^2}{1+\rho^2}\left[(N-h-1)\left(1+\rho^2\right) + 1 - \rho^2\right] > 4\rho^2(N-1)^2 > 0. \quad (4.1)$$

Donc, Q_2 possède deux racines réelles uniques

$$x_\pm = \left(1+\rho^2\right)\frac{-2(N-2h-1)\rho \pm \sqrt{\Delta}}{8h\rho^2}.$$

Notons que, puisque la taille de tous les intervalles est égale à un, nous avons nécessairement $N-h-1 \geq h \geq 1.$ En utilisant ce fait et l'inégalité (4.1), nous obtenons par conséquent

$$|x_\pm| > \left(1+\rho^2\right)\frac{N-h-1}{2|\rho|} \geq \frac{1+\rho^2}{2|\rho|} > 1, \quad \text{puisque} \quad |\rho| \in (0,1).$$

Donc l'HYPOTHÈSE I du théorème 3.1 est satisfaite.

De la remarque 3.1 il découle que $d_1 = d_-(x_-) = x_- + \sqrt{x_-^2 - 1} < 0$ et $d_2 = d_-(x_+) = x_+ - \sqrt{x_+^2 - 1} > 0$ sont des nombres réels.

Puisque, dans ce cas, $s = h$ et $m_1 = \ldots = m_h = 1$, nous avons $\widetilde{\mathbf{H}}_1(d_i) = 1$ et $\mathbf{H}_1(d_i) = 1 + \rho^2$, $i = 1, 2$. Par conséquent, l'équation $\mathbf{S}\underline{c} = \underline{e}$ implique que

$$(1 - d_i\rho)(d_i - \rho)c_{0,i} + (1 + \rho^2)c_{k,i} = 0, \quad k = 1, \ldots, h, \quad i = 1, 2.$$

Donc, $c_{1,1} = c_{2,1} = \ldots = c_{h,1}$ et $c_{1,2} = c_{2,2} = \ldots = c_{h,2}$. Conséquemment, le système $\mathbf{S}\underline{c} = \underline{e}$ se réduit au système à quatre inconnues $c_{0,1}, c_{1,1}, c_{0,2}$ et $c_{1,2}$:

$$\tilde{\mathbf{S}}(c_{0,1}, c_{1,1}, c_{0,2}, c_{1,2})^T = (1, 0, 0, 0)^T$$

avec

$$\tilde{\mathbf{S}} = \frac{1}{1 - \rho^2}\begin{bmatrix} (N-1)(1 - d_1\rho) + 1 - \rho^2 & h(1 - d_1\rho) & (N-1)(1 - d_2\rho) + 1 - \rho^2 & h(1 - d_2\rho) \\ 1 - d_1\rho & 1 & 1 - d_2\rho & 1 \\ (1 - d_1\rho)(d_1 - \rho) & d_1(1 + \rho^2) & 0 & 0 \\ 0 & 0 & (1 - d_2\rho)(d_2 - \rho) & d_2(1 + \rho^2) \end{bmatrix}.$$

Pour montrer que $\tilde{\mathbf{S}}$ est non singulière, nous commençons par montrer que

$$\rho(d_1 + d_2) \geq 0. \tag{4.2}$$

À cette fin, nous notons d'abord que

$$\rho(x_- + x_+) = -(1 + \rho^2)\frac{N - 2h - 1}{2h} \leq 0. \tag{4.3}$$

En outre,

$$\begin{aligned} \rho(d_1 + d_2) &= \rho\left(x_- + x_+\sqrt{x_-^2 - 1} - \sqrt{x_+^2 - 1}\right) = \rho(x_- + x_+)\left(1 + \frac{x_- - x_+}{\sqrt{x_-^2 - 1} + \sqrt{x_+^2 - 1}}\right) \\ &= \frac{\rho(x_- + x_+)}{\sqrt{x_-^2 - 1} + \sqrt{x_+^2 - 1}}\left(\sqrt{x_-^2 - 1} + x_- + \sqrt{x_+^2 - 1} - x_+\right). \end{aligned}$$

En raison de (4.3), la dernière expression est non négative, puisque le second facteur est strictement négatif. Maintenant, nous sommes prêts à considérer le déterminant

$$\det\tilde{\mathbf{S}} = \frac{(d_2 - d_1)\rho}{(1 - \rho^2)^4}s(d_1, d_2),$$

où

$$\begin{aligned} s(d_1, d_2) &= \left(1+\rho^2\right)\left[(N-1)(1-d_1\rho)(1-d_2\rho)+\left(1-\rho^2\right)\left(1+d_1d_2\rho^2\right)\right] \\ &+ h(1-d_1\rho)(1-d_2\rho)\left(-1+(d_1+d_2)\rho+d_1d_2\rho^2-2\rho^2\right). \end{aligned}$$

Nous notons que $|d_i| < 1, i = 1, 2,$ et donc $|d_1 d_2| < 1$. Par conséquent, nous avons $1+\rho^2 > (1-d_1\rho)(1-d_2\rho) > 0, 1+d_1d_2\rho^2 > 0$. Ces inégalités ainsi que (4.2) donnent

$$\begin{aligned} s(d_1, d_2) &> (1-d_1\rho)(1-d_2\rho)\left\{(N-1)\left(1+\rho^2\right)-h\left[1+d_1d_2\rho^2+2\rho^2\right]\right\} \\ &> (1-d_1\rho)(1-d_2\rho)\left[(N-h-1)\left(1+\rho^2\right)-2h\rho^2\right] \\ &> (1-d_1\rho)(1-d_2\rho)(N-2h-1)\left(1+\rho^2\right) \\ &\geq 0. \end{aligned}$$

Conséquemment, $\det \tilde{\mathbf{S}} \neq 0$.

Puisque $\text{rang}\,\mathbf{S} = \text{rang}\,\tilde{\mathbf{S}} + 2(h-1)$, nous obtenons $\text{rang}\,\mathbf{S} = 2(h+1)$ et donc l'HYPOTHÈSE II du théorème 3.1 est satisfaite. En outre, $\tilde{\mathbf{S}}^{-1}$ existe. Donc,

$$(c_{0,1}, c_{1,1}, c_{0,2}, c_{1,2}) = (1,0,0,0)\left[\tilde{\mathbf{S}}^{-1}\right]^T.$$

Enfin, nous concluons que la récurrence est de la forme suivante :

$$\hat{\mu}_t = a_1\hat{\mu}_{t-1} + a_2\hat{\mu}_{t-2} + \underline{r}_0^T\underline{X}_t + \underline{r}_1^T\underline{X}_{t-1} + \underline{r}_2^T\underline{X}_{t-2},$$

où

$$\begin{cases} a_1 = d_1 + d_2 \\ a_2 = -d_1d_2 \\ \underline{r}_0 = \mathbf{N}(d_1)\left[(c_{0,1}+c_{1,1})\underline{1} - c_{1,1}\underline{\varepsilon}\right] + \mathbf{N}(d_2)\left[(c_{0,2}+c_{1,2})\underline{1} - c_{1,2}\underline{\varepsilon}\right] \\ \underline{r}_1 = -\left(d_2\mathbf{I}+\mathbf{C}^T\right)\mathbf{N}(d_1)\left[(c_{0,1}+c_{1,1})\underline{1} - c_{1,1}\underline{\varepsilon}\right] - \left(d_1\mathbf{I}+\mathbf{C}^T\right)\mathbf{N}(d_2)\left[(c_{0,2}+c_{1,2})\underline{1} - c_{1,2}\underline{\varepsilon}\right] \\ \underline{r}_2 = d_2\mathbf{C}^T\mathbf{N}(d_1)\left[(c_{0,1}+c_{1,1})\underline{1} - c_{1,1}\underline{\varepsilon}\right] + d_1\mathbf{C}^T\mathbf{N}(d_2)\left[(c_{0,2}+c_{1,2})\underline{1} - c_{1,2}\underline{\varepsilon}\right] \end{cases}.$$

Par exemple, soit $N = 7, h = 2, H = \{3, 6\}$ et soit $\rho = 0{,}5$. Alors

$$Q_2(x) = -1{,}6x^2 - 2x + 5{,}75$$

et

$$\begin{cases} x_1 = -2{,}6211 \\ x_2 = 1{,}3711 \end{cases} \Rightarrow \begin{cases} d_+(x_1) = -5{,}0439 \\ d_1 = d_-(x_1) = -0{,}1983 \\ d_+(x_2) = 2{,}3091 \\ d_2 = d_-(x_2) = 0{,}4331 \end{cases} \Rightarrow \begin{cases} a_1 = 0{,}2348 \\ a_2 = 0{,}0859 \end{cases}.$$

Enfin, (3.9) prend la forme

$$\hat{\mu}_t = 0{,}2348\hat{\mu}_{t-1} + 0{,}0859\hat{\mu}_{t-2} + \begin{bmatrix} 0{,}2171 \\ 0{,}1904 \\ 0{,}0000 \\ 0{,}2171 \\ 0{,}1904 \\ 0{,}0000 \\ 0{,}1850 \end{bmatrix}^T \underline{X}_t + \begin{bmatrix} -0{,}0093 \\ -0{,}1086 \\ 0{,}0000 \\ -0{,}0093 \\ -0{,}1086 \\ 0{,}0000 \\ 0{,}0010 \end{bmatrix}^T \underline{X}_{t-1} + \begin{bmatrix} 0{,}0000 \\ 0{,}0047 \\ 0{,}0000 \\ -0{,}0476 \\ 0{,}0047 \\ 0{,}0000 \\ -0{,}0476 \end{bmatrix}^T \underline{X}_{t-2}.$$

4.3 Scénario de Szarkowski, $p = 3$

S'il existe h_2 intervalles de taille 2 et h_1 intervalles de taille 1 dans le schéma en cascade, le polynôme $Q_p = Q_3$, voir (3.3), prend la forme

$$Q_3(x) = (N-1)\left(1+\rho^2-2\rho x\right) + 1 - \rho^2 - \left(1+\rho^2-2\rho x\right)^2 \left(h_2 \frac{2\rho x + 2\left(1+\rho^2\right)}{1+\rho^2+\rho^4} + h_1 \frac{1}{1+\rho^2} \right).$$

Le scénario de Szarkowski est défini par le schéma en cascade $\underline{\varepsilon} = (1,1,0,0,1,1)^T$ (souvent noté aussi sous la forme $2-2-2$), utilisé, par exemple, par le Bureau central de la statistique de la Pologne pour réaliser l'Enquête sur la population active (connue sous l'acronyme BAEL), voir Szarkowski et Witkowski (1994) ou Popiński (2006). En fait, ce genre de scénario est utilisé également dans l'EPA d'autres pays européens. Ici, $N = 6$ et $H = \{3,4\}$. Donc $h_2 = 1, h_1 = 0$, et

$$Q_3(x) = 5\left(1+\rho^2-2\rho x\right) + 1 - \rho^2 - 2\left(1+\rho^2-2\rho x\right)^2 \frac{\rho x + 1 + \rho^2}{1+\rho^2+\rho^4}. \tag{4.4}$$

Wesołowski (2010) a prouvé que, dans ce cas, Q_3 est strictement croissant ou décroissant dans le domaine entier et possède deux racines conjuguées complexes x_1, x_2, et une racine réelle $x_3 \notin [-1,1]$, ce qui signifie que l'HYPOTHÈSE I du théorème 3.1 est vérifiée. Il a également montré dans cet article que la matrice $\mathbf{S}$, dans ce cas de dimensions 9×9, est inversible (ce qui signifie que l'HYPOTHÈSE II du théorème 3.1 est vérifiée). Donc, comme pour $p = 1,2$, la récurrence (3.9) pour le scénario de Szarkowski est toujours vérifiée.

En général, même dans le cas $p = 3$, la vérification des HYPOTHÈSES I et II du théorème 3.1 doit être effectuée numériquement, c'est-à-dire après l'attribution de la valeur du coefficient de corrélation ρ. Cependant, il convient de noter que toutes les simulations exécutées confirment l'existence de la solution. L'approximation asymptotique des paramètres du modèle « classique » a également été observée dans les expériences numériques que nous avons effectuées.

Les coefficients a_1, a_2, a_3 dépendent de $d_1 = d_-(x_1), d_2 = d_-(x_2) = d_1^*$ et $d_3 = d_-(x_3)$ de la façon suivante (voir (3.10)) :

$$\begin{cases} a_1 = d_1 + d_2 + d_3 \\ a_2 = -(d_1 d_2 + d_2 d_3 + d_1 d_3). \\ a_3 = d_1 d_2 d_3 \end{cases}$$

Pour le scénario de Szarkowski, en prenant par exemple $\rho = 0{,}7$ dans (4.4), nous obtenons

$$\begin{cases} x_1 = -0{,}5668 - 1{,}4069i \\ x_2 = -0{,}5668 + 1{,}4069i \\ x_3 = 1{,}1336 \end{cases} \Rightarrow \begin{cases} d_+(x_1) = -1{,}0368 - 3{,}1035i \\ d_1 = d_-(x_1) = -0{,}0968 + 0{,}2899i \\ d_+(x_2) = -1{,}0368 + 3{,}1035i \\ d_2 = d_-(x_2) = -0{,}0968 - 0{,}2899i \\ d_+(x_3) = 1{,}6675 \\ d_3 = d_-(x_3) = 0{,}5997 \end{cases} \Rightarrow \begin{cases} a_1 = 0{,}4060 \\ a_2 = 0{,}0227. \\ a_3 = 0{,}0560 \end{cases}$$

En raison du théorème 3.1, nous obtenons la forme suivante de (3.9) :

$$\begin{aligned} \hat{\mu}_t &= 0{,}4060\hat{\mu}_{t-1} + 0{,}0227\hat{\mu}_{t-2} + 0{,}0560\hat{\mu}_{t-3} \\ &+ \begin{bmatrix} 0{,}2862 \\ 0{,}2217 \\ 0{,}0000 \\ 0{,}0000 \\ 0{,}2862 \\ 0{,}2059 \end{bmatrix}^T \underline{X}_t + \begin{bmatrix} -0{,}0036 \\ -0{,}2004 \\ 0{,}0000 \\ 0{,}0000 \\ -0{,}0036 \\ -0{,}1984 \end{bmatrix}^T \underline{X}_{t-1} + \begin{bmatrix} -0{,}0143 \\ 0{,}0026 \\ 0{,}0000 \\ 0{,}0000 \\ -0{,}0143 \\ 0{,}0033 \end{bmatrix}^T \underline{X}_{t-2} + \begin{bmatrix} 0{,}0000 \\ 0{,}0100 \\ 0{,}0000 \\ 0{,}0000 \\ -0{,}0760 \\ 0{,}0100 \end{bmatrix}^T \underline{X}_{t-3}. \end{aligned}$$

4.4 Scénario de la CPS, $p = 9$

Considérons le scénario 4-8-4 bien connu et qui a fait l'objet de nombreuses études, pour lequel le schéma en cascade est

$$\underline{\varepsilon} = (1,1,1,1,0,0,0,0,0,0,0,0,1,1,1,1)^T$$

qui est utilisé aux États-Unis pour la *Current Population Survey*, voir U.S. Bureau of Census (2002). Dans ce cas, $N = 16, h = 8$ et $H = \{5,\ldots,12\}$. Nous ne possédons aucune preuve analytique que les HYPOTHÈSES I et II sont satisfaites dans ce scénario pour tout ρ.

Le polynôme $Q_p = Q_9$, voir (3.3), est de degré 9 et de la forme

$$Q_9(x) = 15\left(1 + \rho^2 - 2\rho x\right) + 1 - \rho^2 - \left(1 + \rho^2 - 2\rho x\right)^2 \operatorname{tr}\left(\mathbf{T}_8(x)\mathbf{R}_8^{-1}\right).$$

Par conséquent, son analyse, ainsi que l'analyse de la matrice $\mathbf{S}$ (qui est de dimensions 81×81 dans ce schéma), peut être effectuée numériquement, après avoir attribué une valeur pour ρ. Afin d'utiliser le résultat du théorème 3.1, nous devons vérifier numériquement que les HYPOTHÈSES I et II sont satisfaites pour une valeur concrète donnée de ρ. Nous avons confirmé que les hypothèses sont vérifiées pour plusieurs valeurs de ρ prises au hasard dans l'intervalle $(-1,1)$.

En prenant par exemple $\rho = 0{,}9$, nous obtenons que Q_9 possède huit racines complexes et une racine réelle de la forme

$$\begin{cases} x_1 = -0{,}7667 - 0{,}0208i \\ x_2 = -0{,}7667 + 0{,}0208i \\ x_3 = -0{,}1746 - 0{,}0320i \\ x_4 = -0{,}1746 + 0{,}0320i \\ x_5 = 0{,}4989 - 0{,}0284i \\ x_6 = 0{,}4989 + 0{,}0284i \\ x_7 = 0{,}9391 - 0{,}0121i \\ x_8 = 0{,}9391 + 0{,}0121i \\ x_9 = -1{,}0006 \end{cases} \Rightarrow \begin{cases} d_1 = d_-(x_1) = -0{,}7419 - 0{,}6220i \\ d_2 = d_-(x_2) = -0{,}7419 + 0{,}6220i \\ d_3 = d_-(x_3) = -0{,}1689 - 0{,}9532i \\ d_4 = d_-(x_4) = -0{,}1689 + 0{,}9532i \\ d_5 = d_-(x_5) = 0{,}4825 - 0{,}8389i \\ d_6 = d_-(x_6) = 0{,}4825 + 0{,}8389i \\ d_7 = d_-(x_7) = 0{,}9064 - 0{,}3335i \\ d_8 = d_-(x_8) = 0{,}9064 + 0{,}3335i \\ d_9 = d_-(x_9) = -0{,}9682 \end{cases} \Rightarrow \begin{cases} a_1 = 0{,}7429 \\ a_2 = 0{,}0019 \\ a_3 = 0{,}0023 \\ a_4 = 0{,}0029 \\ a_5 = 0{,}0037. \\ a_6 = 0{,}0049 \\ a_7 = 0{,}0066 \\ a_8 = 0{,}0088 \\ a_9 = 0{,}0119 \end{cases}$$

Le coefficient a_1 est dominant en ce qui concerne la valeur absolue. Le deuxième plus grand coefficient, a_9, est plus petit d'un ordre de grandeur, et les autres coefficients sont plus petits d'au moins deux ordres de grandeur. Les résultats pour d'autres valeurs du paramètre ρ ont un comportement similaire.

5 Discussion

Le résultat principal de l'article est une formule de récurrence explicite pour le meilleur estimateur linéaire sans biais (BLUE) de la moyenne pour n'importe quelle édition d'une enquête répétée avec tout schéma de renouvellement en cascade de l'échantillon. La principale nouveauté tient au fait de permettre des intervalles dans le schéma. Les résultats obtenus antérieurement concernaient des schémas sans intervalle ou d'autres estimateurs que les estimateurs BLUE. L'approche que nous avons élaborée s'appuie fortement sur l'algèbre des matrices et des opérateurs linéaires de dimension infinie, ainsi que sur les propriétés des polynômes de Tchebychev. Malheureusement, la formule récursive explicite que nous avons obtenue dans le théorème 3.1 requiert deux hypothèses apparemment techniques : l'HYPOTHÈSE I sur la localisation des racines d'un polynôme Q_p et l'HYPOTHÈSE II sur le rang de la matrice $\mathbf{S}$. Il convient de souligner que l'un et l'autre de ces objets, Q_p et $\mathbf{S}$, dépendent SEULEMENT de deux paramètres : le schéma de renouvellement de l'échantillon $\underline{\varepsilon}$ et le coefficient de corrélation ρ. On sait que ces deux hypothèses sont satisfaites si la couverture du schéma est $p = 1$ ou $p = 2$ pour tout scénario en cascade et $p = 3$ pour le scénario 2-2-2. On ne sait pas si elles sont satisfaites en général.

Cependant, des expériences numériques permettent de conjecturer que c'est véritablement le cas. Dans ces expériences, nous avons considéré de nombreux schémas de renouvellement d'échantillon différents. Pour chacun, nous avons considéré plusieurs valeurs de $\rho \in (-1,1)$. Après avoir choisi le schéma de renouvellement $\underline{\varepsilon}$ et la valeur de ρ, nous avons construit le polynôme Q_p et la matrice $\mathbf{S}$ respectifs. Numériquement, nous avons cherché les racines de Q_p. Ces racines étaient souvent complexes, mais quand elles étaient réelles, elles étaient situées en dehors de l'intervalle $(-1,1)$ dans toutes les expériences (c'est-à-dire que l'HYPOTHÈSE I était satisfaite). Ensuite, nous avons essayé de résoudre numériquement l'équation $\mathbf{S}\underline{c} = (1,0,\ldots,0) \in \mathbb{R}^{ph+h+1}$. De nouveau, dans tous les expériences, nous avons obtenu la solution unique, ce qui signifie que $\mathbf{S}$ était de plein rang (c'est-à-dire que l'HYPOTHÈSE II était également satisfaite). Nous pensons que les deux hypothèses sont systématiquement satisfaites, mais il est probablement difficile de donner une preuve mathématique de ces deux faits. Néanmoins, un article donnant la preuve que l'HYPOTHÈSE I est satisfaite pour tout schéma de renouvellement en cascade avec un seul intervalle de n'importe quelle taille et pour n'importe quelle valeur de $\rho \in (-1,1)$ est en préparation.

La méthode que nous proposons possède d'autres types de limites, qui sont dues aux contraintes du modèle. En particulier, dans le modèle, les corrélations sont exponentielles (comme dans le modèle original de Patterson). Cette propriété joue un rôle important dans l'argument que nous utilisons; par exemple, elle rend la matrice de covariance $\mathbf{C}$ nilpotente de degré N, c'est-à-dire que N est la plus petite valeur de j telle que $\mathbf{C}^j = 0$. En outre, on a observé (voir l'exemple 4.5 dans Kowalski 2009) que d'autres modèles de covariance peuvent donner lieu à d'importantes difficultés dans l'analyse de la formule de la variance des estimateurs. Il se peut que certains écarts raisonnables par rapport à l'hypothèse de corrélation exponentielle, par exemple, $\mathbb{C}\mathrm{ov}\left(X_{i,j}, X_{k,l}\right) = \theta + (1-\theta)\rho^{|j-l|}\delta_{i,k}$ pour $\theta \in [0,1]$ (voir Lent, Miller, Cantwell et Duff (1999), en particulier leur tableau 1, sa discussion ainsi que des références supplémentaires) aboutissent à des formules de variance solubles. Un modèle de covariance de ce genre est probablement le premier qu'il faudra examiner dans tous futurs travaux visant à étendre le modèle.

Dans le modèle, nous avons également supposé que les espérances pour une édition donnée de l'enquête sont toutes les mêmes et dépendent seulement du numéro de l'édition : $\mathbb{E}X_{i,j} = \mu_j$. Cependant, d'autres modèles pourraient présenter un intérêt, par exemple, $\mathbb{E}X_{i,j} = \mu_j + a_i$ (voir Bailar 1975). Ici les ajustements a_i peuvent être interprétés comme un biais de temps passé dans l'échantillon causé par le nombre d'éditions de l'enquête à laquelle l'unité i a participé. Naturellement, si a_i est connu, il n'y a pas de problème : il suffit d'ajuster $X_{i,j}$ en soustrayant a_i et d'utiliser l'approche que nous avons élaborée. S'il n'est pas connu, la solution opérationnelle (mais non mathématique) consisterait à ajuster les $X_{i,j}$ au moyen d'estimateurs appropriés des a_i (obtenus en dehors du modèle que nous analysons). La solution mathématique exacte est inconnue et mériterait d'être poursuivie.

Un autre aspect, qui présente un intérêt dans le modèle examiné ici, est la question de la récurrence pour l'estimateur BLUE d'une variation de la moyenne $\mu_t - \mu_{t-1}$. Nous pensons que cette question peut être approchée au moyen des méthodes élaborées dans le présent article. Néanmoins, nous nous attendons à ce que cela nécessite beaucoup d'adaptations prudentes des techniques algébriques utilisées plus haut.

Mentionnons aussi que l'horizon temporel du modèle considéré dans le présent article est infini, alors que le nombre d'éditions des enquêtes réelles est toujours fini. Comme nous l'avons déjà mentionné dans

l'introduction, les résultats que nous avons obtenus semblent être une approximation raisonnable du cas d'un horizon fini, quand les coefficients de récurrence (1.2) dépendent de t. En particulier, des expériences numériques, effectuées pour une grande gamme de valeurs de $\rho \in (-1,1)$ et divers schémas de renouvellement en cascade ε, montrent que, par exemple, la valeur des coefficients $a_i^{(t)}$ (pour l'horizon fini) était déjà à peu près la même que celle de a_i (pour l'horizon infini) pour $t \approx 10$. Nous avons observé le même comportement pour les variances des estimateurs. Néanmoins, la convergence n'a été mathématiquement établie que pour le cas $p = 1$. À l'heure actuelle, il semble aussi impossible d'obtenir les bornes analytiques pour la vitesse de convergence.

Il serait intéressant de savoir comment les estimateurs obtenus ici fonctionnent dans les enquêtes réelles. Pour répondre à cette question, il faut avoir accès à des données réelles et susciter l'intérêt des praticiens pour les solutions théoriques que nous avons proposées. Il est fort probable que les formules exactes données dans le théorème 3.1 nécessitent certains ajustements en raison des limites du modèle dont nous avons discuté.

6 Annexe

6.1 Algèbre des opérateurs de translation

Dans la première partie de l'annexe, nous introduisons et analysons un formalisme algébrique des opérateurs qui est essentiel à la preuve de notre résultat principal (donnée à la sous-section 6.2).

Pour une séquence de vecteurs $\underline{\overline{x}} = (\underline{x}_0, \underline{x}_1, \underline{x}_2, \ldots), \underline{x}_i \in \mathbb{R}^N$, définissons les translations vers la gauche et vers la droite par

$$\begin{aligned} \mathcal{L}(\underline{\overline{x}}) &= (\underline{x}_1, \underline{x}_2, \underline{x}_3, \ldots) \quad \text{translation à gauche,} \\ \mathcal{R}(\underline{\overline{x}}) &= (\underline{0}, \underline{x}_0, \underline{x}_1, \ldots) \quad \text{translation à droite.} \end{aligned}$$

Notons que $\mathcal{L}\mathcal{R} = \mathcal{I}$ (identité), mais que

$$(\mathcal{I} - \mathcal{R}\mathcal{L})\,\underline{\overline{x}} = (\underline{x}_0, \underline{0}, \underline{0}, \ldots) = \underline{x}_0 \overline{e}, \tag{6.1}$$

où $\overline{e} = (1, 0, 0, \ldots)$.

Pour toute matrice $\mathbf{A}$ de dimensions $M \times N$ définissons

$$\mathbf{A}\underline{\overline{x}} = (\mathbf{A}\underline{x}_0, \mathbf{A}\underline{x}_1, \mathbf{A}\underline{x}_2, \ldots).$$

En particulier, pour un nombre complexe (réel) a, en prenant $\mathbf{A} = a\mathbf{I}$, nous avons

$$a\underline{\overline{x}} = (a\underline{x}_0, a\underline{x}_1, a\underline{x}_2, \ldots).$$

En outre, en vertu des définitions susmentionnées, pour tout $i, j \geq 0$

$$\mathcal{R}^i \mathcal{L}^j \mathbf{A}\underline{\overline{x}} = \mathbf{A}\mathcal{R}^i \mathcal{L}^j \underline{\overline{x}}.$$

Pour une séquence constante de vecteurs $\overline{\underline{x}} = (\underline{x}, \underline{x}, \underline{x}, \ldots)$, nous avons $\mathcal{L}\overline{\underline{x}} = \overline{\underline{x}}$ et donc pour tout $i, j \geq 0$,

$$\mathcal{L}^i \mathcal{R}^j \overline{\underline{x}} = \begin{cases} \overline{\underline{x}}, & \text{pour } i \geq j, \\ \mathcal{R}^{j-i} \overline{\underline{x}}, & \text{pour } i < j. \end{cases} \tag{6.2}$$

Si $N = 1$, nous écrivons $\underline{\overline{y}} = \overline{y} = (y_0, y_1, y_2 \ldots)$, $y_i \in \mathbb{R}$, et $\mathrm{L} := \mathcal{L}, \mathrm{R} := \mathcal{R}$. Notons que, pour $\overline{y} = \left(y^n\right)_{n \geq 0}$, nous avons

$$\mathrm{L}^j \overline{y} = y^j \overline{y} \tag{6.3}$$

et donc

$$\mathrm{L}^j \mathrm{R}^i \overline{y} = \begin{cases} y^{j-i} \overline{y}, & \text{pour } j \geq i, \\ \mathrm{R}^{i-j} \overline{y}, & \text{pour } j < i. \end{cases}$$

Pour tout $\overline{y} = (y_n)_{n \geq 0}$ et tout $\overline{\underline{x}} = (\underline{x}_n)_{n \geq 0}$, définissons $\overline{y}\overline{\underline{x}} = (y_n \underline{x}_n)_{n \geq 0}$. Alors pour tous nombres complexes (réels) α, β, toutes matrices $\mathbf{A}, \mathbf{B}$, de dimensions $M \times N$, tous $i, j, k, m \geq 0$,

$$\left(\alpha \mathbf{A} \mathcal{R}^i \mathcal{L}^j + \beta \mathbf{B} \mathcal{L}^m \mathcal{R}^k\right) \overline{y}\overline{\underline{x}} = \left(\alpha \mathrm{R}^i \mathrm{L}^j \overline{y}\right)\left(\mathbf{A} \mathcal{R}^i \mathcal{L}^j \overline{\underline{x}}\right) + \left(\beta \mathrm{L}^m \mathrm{R}^k \overline{y}\right)\left(\mathbf{B} \mathcal{L}^m \mathcal{R}^k \overline{\underline{x}}\right). \tag{6.4}$$

Notons aussi que, si $\overline{\underline{x}} = (\underline{x}, \underline{x}, \ldots)$ est une séquence constante, alors

$$\mathcal{R}^i \mathcal{L}^j \overline{y}\overline{\underline{x}} = \left(\mathrm{R}^i \mathrm{L}^j \overline{y}\right) \overline{\underline{x}} \quad \text{et} \quad \mathcal{L}^j \mathcal{R}^i \overline{y}\overline{\underline{x}} = \left(\mathrm{L}^j \mathrm{R}^i \overline{y}\right) \overline{\underline{x}}. \tag{6.5}$$

Lemme 6.1 *Soit* $v_i, i = 1, \ldots, p$, *les fonctions définies en (3.11), où* $a_1, \ldots, a_p$ *sont des nombres arbitraires. Soit* $\overline{\underline{x}} = (\underline{x}, \underline{x}, \ldots)$ *et* $\overline{y} = \left(y^n\right)_{n \geq 0}$. *Alors, pour tout* $i = 1, \ldots, p$

$$\mathcal{L}^i \left(\mathcal{I} - \sum_{j=1}^{p} a_j \mathcal{R}^j\right) = \left(\mathcal{L}^p - \sum_{j=1}^{p} a_j \mathcal{L}^{p-j}\right) \mathcal{R}^{p-i}, \tag{6.6}$$

$$(\mathcal{I} - \mathcal{R}\mathcal{L}) \left(\mathcal{L}^p - \sum_{j=1}^{p} a_j \mathcal{L}^{p-j}\right) \mathcal{R}^{p-i} \overline{y}\overline{\underline{x}} = v_i(y)(\underline{x}_0, \underline{0}, \underline{0}, \ldots) \tag{6.7}$$

et

$$\left(\mathcal{L}^p - \sum_{j=1}^{p} a_j \mathcal{L}^{p-j}\right) \overline{y}\overline{\underline{x}} = v_p(y) \, \overline{y}\overline{\underline{x}}. \tag{6.8}$$

Preuve. Pour commencer, nous prouvons (6.8). En vertu de (6.4)

$$\left(\mathcal{L}^p - \sum_{j=1}^{p} a_j \mathcal{L}^{p-j}\right) \overline{y}\overline{\underline{x}} = \left(\mathrm{L}^p \overline{y}\right) \mathcal{L}^p \overline{\underline{x}} - \sum_{j=1}^{p} a_j \left(\mathrm{L}^{p-j} \overline{y}\right)\left(\mathcal{L}^{p-j} \overline{\underline{x}}\right).$$

Notons que $\mathrm{L}^k \overline{y} = y^k \overline{y}$ et $\mathcal{L}^k \underline{\overline{x}} = \underline{\overline{x}}$ pour tout $k = 0, 1, \ldots$. Par conséquent

$$\left(\mathcal{L}^p - \sum_{j=1}^{p} a_j \mathcal{L}^{p-j} \right) \overline{y}\underline{\overline{x}} = \left[\left(y^p - \sum_{m=1}^{p} a_m y^{p-m} \right) \overline{y} \right] \underline{\overline{x}}.$$

Alors (6.8) s'ensuit en vertu de la définition (3.11) pour $i = p$.

De nouveau, de (6.2), (6.4) et (6.5) il découle que

$$(\mathcal{I} - \mathcal{R}\mathcal{L}) \left(\mathcal{L}^p - \sum_{j=1}^{p} a_j \mathcal{L}^{p-j} \right) \mathcal{R}^{p-i} \overline{y}\underline{\overline{x}} = \left[(\mathrm{I} - \mathrm{RL}) \left(\mathrm{L}^p - \sum_{j=1}^{p} a_j \mathrm{L}^{p-j} \right) \mathrm{R}^{p-i} \overline{y} \right] \underline{\overline{x}}.$$

Puisque, pour tout $k \in \{0, 1, \ldots, p\}$,

$$\left(\mathrm{L}^p - \sum_{j=1}^{p} a_j \mathrm{L}^{p-j} \right) \mathrm{R}^{p-k} \overline{y} = y^k \overline{y} - \sum_{j=1}^{k} a_j y^{k-j} \overline{y} - \sum_{j=k+1}^{p} a_j \mathrm{R}^{j-k} \overline{y} = \nu_k(y) \overline{y} - \sum_{j=k+1}^{p} a_j \mathrm{R}^{j-k} \overline{y},$$

alors

$$(\mathrm{I} - \mathrm{RL}) \left(\mathrm{L}^p - \sum_{j=1}^{p} a_j \mathrm{L}^{p-j} \right) \mathrm{R}^{p-k} \overline{y} = \nu_k(y) \overline{e}$$

et donc (6.7) s'ensuit.

L'identité (6.6) s'ensuit en vertu de (6.2) puisque

$$\mathcal{L}^i \left(\mathcal{I} - \sum_{j=1}^{p} a_j \mathcal{R}^j \right) = \mathcal{L}^i - \sum_{j=1}^{p} a_j \mathcal{L}^i \mathcal{R}^j = \mathcal{L}^p \mathcal{R}^{p-i} - \sum_{j=1}^{p} a_j \mathcal{L}^{p-j} \mathcal{R}^{p-i}.$$

Lemme 6.2 *Soit $\mathcal{D}$ un opérateur sur l'espace des séquences de vecteurs dans $\mathbb{R}^N$ défini par*

$$\mathcal{D} = \mathcal{I} + \sum_{k=1}^{N-1} \left(\mathbf{C}^k \mathcal{L}^k + \left(\mathbf{C}^T \right)^k \mathcal{R}^k \right), \tag{6.9}$$

où $\mathbf{C}$ est la matrice de covariance définie à la section 2.

L'opérateur $\mathcal{D}$ est inversible et

$$\mathcal{D}^{-1} = \left(\mathcal{I} - \mathbf{C}^T \mathcal{R} \right) \mathbf{\Delta} \left(\mathcal{I} - \mathbf{C}\mathcal{L} \right). \tag{6.10}$$

Preuve. Notons que $\mathbf{I} - \mathbf{C}\mathbf{C}^T = \mathrm{diag}\left(1 - \rho^2, \ldots, 1 - \rho^2, 1\right)$. Par conséquent, $\mathbf{\Delta} = \left(\mathbf{I} - \mathbf{C}\mathbf{C}^T \right)^{-1}$ est bien définie. Notons aussi que $\sum_{k=0}^{N-1} \mathbf{C}^k \mathcal{L}^k$ est inversible et que son inverse est $\mathcal{I} - \mathbf{C}\mathcal{L}$. De même, $\sum_{k=0}^{N-1} (\mathbf{C}^T)^k \mathcal{R}^k$ est inversible et son inverse est $\mathcal{I} - \mathbf{C}^T \mathcal{R}$.

Donc,

$$\begin{aligned}\left[\left(\mathcal{I}-\mathbf{C}^T\mathcal{R}\right)\mathbf{\Delta}\left(\mathcal{I}-\mathbf{C}\mathcal{L}\right)\right]^{-1} &= \left(\mathcal{I}-\mathbf{C}\mathcal{L}\right)^{-1}\mathbf{\Delta}^{-1}\left(\mathcal{I}-\mathbf{C}^T\mathcal{R}\right)^{-1}\\ &= \left(\sum_{k=0}^{N-1}\mathbf{C}^k\mathcal{L}^k\right)\left(\mathbf{I}-\mathbf{C}\mathbf{C}^T\right)\left(\sum_{j=0}^{N-1}\left(\mathbf{C}^T\right)^j\mathcal{R}^j\right)\\ &= \sum_{k,j=0}^{N-1}\mathbf{C}^k\left(\mathbf{C}^T\right)^j\mathcal{L}^k\mathcal{R}^j-\sum_{k,j=1}^{N-1}\mathbf{C}^k\left(\mathbf{C}^T\right)^j\mathcal{L}^k\mathcal{R}^j\\ &= \mathcal{D}+\sum_{k,j=1}^{N-1}\mathbf{C}^k\left(\mathbf{C}^T\right)^j\mathcal{L}^{k-1}\left(\mathcal{L}\mathcal{R}-\mathcal{I}\right)\mathcal{R}^{j-1}\\ &= \mathcal{D}.\end{aligned}$$

6.2 Preuve de la récurrence

Preuve du théorème 3.1. Notons d'abord que, puisque les $d_1,\ldots,d_p$ sont soit réelles, soit viennent en paires conjuguées (voir la remarque 3.1), il découle de (3.10) que les $a_1,\ldots,a_p$ sont des nombres réels.

Rappelons que $\underline{e}_0 = \underline{1}$ et notons $\overline{\underline{e}}_j = \left(\underline{e}_j,\underline{e}_j,\ldots\right)$, $j \in H' = \{0\}\bigcup H$. Rappelons que la matrice diagonale $\mathbf{\Delta}$ de dimensions $N \times N$ est définie par

$$\mathbf{\Delta} = \left(\mathbf{I}-\mathbf{C}\mathbf{C}^T\right)^{-1} = \frac{1}{1-\rho^2}\operatorname{diag}\left(1,\ldots,1,1-\rho^2\right).$$

Avec $d_1,\ldots,d_p$ et $\underline{c}$ définis comme dans le théorème 3.1, soit (voir (6.10))

$$\overline{\underline{w}} = \left(\underline{w}_0,\underline{w}_1,\ldots\right) = \mathcal{D}^{-1}\sum_{m=1}^{p}\sum_{j\in H'}c_{j,m}\overline{d}_m\overline{\underline{e}}_j, \tag{6.11}$$

où $\overline{d}_m = \left(1,d_m,d_m^2,\ldots\right), m = 1,\ldots,p$. Notons que $\|\underline{w}_i\|$ (la longueur du vecteur $\underline{w}_i$) est d'ordre $\left(\max_{1\le m\le p}|d_m|\right)^i, i = 0,1,\ldots.$ D'après la remarque 3.1 et l'HYPOTHÈSE II, nous avons $\max_{1\le m\le p}|d_m| \in (0,1)$. D'où (2.1) est une définition correcte d'une série aléatoire (avec variance bornée).

Par conséquent, il suffit de montrer que :

1. La séquence $\overline{\underline{w}}$ définie en (6.11) est la séquence de poids optimaux. Pour cela, nous notons que la variance de tout estimateur linéaire $\sum_{i=0}^{\infty}\underline{u}_i^T\underline{X}_i, \underline{u}_i \in \mathbb{R}^N, i = 0,1,\ldots$ est de la forme

$$\mathbb{V}\mathrm{ar}\sum_{i=0}^{\infty}\underline{u}_i^T\underline{X}_i = \sum_{i=0}^{\infty}\underline{u}_i^T\underline{u}_i + 2\sum_{i=0}^{\infty}\sum_{k=1}^{N-1}\underline{u}_i^T\mathbf{C}^k\underline{u}_{i+k}. \tag{6.12}$$

   Nous devons montrer que $\overline{\underline{u}} = \left(\underline{u}_i\right)_{i\ge 0} := \overline{\underline{w}}$ avec $\overline{\underline{w}}$ défini en (6.11) minimise cette expression sous les contraintes (2.2) et (2.3). Puisque la variance susmentionnée en tant que fonction de $\overline{\underline{u}}$ est convexe, le problème possède la solution unique. En utilisant la méthode de Lagrange

classique, c'est-à-dire en dérivant la fonction de Lagrange (avec les multiplicateurs $(\lambda_{j,i})_{j\in H',i\geq 0}$)

$$V(\underline{\bar{u}}) = \sum_{i=0}^{\infty} \underline{u}_i^T \underline{u}_i + 2\sum_{i=0}^{\infty}\sum_{k=1}^{N-1} \underline{u}_i^T \mathbf{C}^k \underline{u}_{i+k} - 2\sum_{i=0}^{\infty}\sum_{j\in H'} \lambda_{j,i} \underline{u}_i^t \underline{e}_j,$$

par rapport à $(\underline{u}_i)_{i\geq 0}$ et en comparant les dérivées à zéro, de façon équivalente, nous devons montrer qu'il existe des nombres réels (multiplicateurs de Lagrange) $\lambda_{j,l}, j \in H', l = 0,1,\ldots,$ tels que

$$\mathcal{D}\underline{\bar{w}} = \left[\mathcal{I} + \sum_{k=1}^{N-1}\left(\mathbf{C}^k\mathcal{L}^k + \left(\mathbf{C}^T\right)^k \mathcal{R}^k\right)\right]\underline{\bar{w}} = \underline{\bar{\Lambda}}, \tag{6.13}$$

où $\underline{\bar{w}}$ est défini en (6.11) et $\underline{\bar{\Lambda}} = (\underline{\Lambda}_0, \underline{\Lambda}_1, \ldots)$ avec

$$\underline{\Lambda}_l = \sum_{j\in H'} \lambda_{j,l} \underline{e}_j, \quad l = 0,1,\ldots$$

2. Les contraintes (2.2) et (2.3) sont satisfaites pour $\underline{\bar{w}}$ défini en (6.11).
3. La récurrence de base (3.9) est vérifiée avec $\underline{\bar{w}}$ défini en (6.11), c'est-à-dire que la séquence $\underline{\bar{r}}$ définie par

$$\underline{\bar{r}} := \left(\mathcal{I} - \sum_{m=1}^{p} a_m \mathcal{R}^m\right)\underline{\bar{w}} \tag{6.14}$$

doit satisfaire

$$\mathcal{L}^{p+1}\underline{\bar{r}} = \underline{\bar{0}} \tag{6.15}$$

et pour tout $i = 0,1,\ldots,p$

$$(\mathcal{I} - \mathcal{R}\mathcal{L})\mathcal{L}^i\underline{\bar{r}} = \sum_{m=1}^{p}\left[\left(v_i(d_m)\mathbf{I} - v_{i-1}(d_m)\mathbf{C}^T\right)\mathbf{N}(d_m)\sum_{j\in H'} c_{j,m}\underline{e}_j\right]\bar{e}, \tag{6.16}$$

où $\mathbf{N}(d) = \mathbf{\Delta}(\mathbf{I} - d\mathbf{C})$.

Ad. 1. Nous allons montrer que (6.13) est vérifiée avec

$$\lambda_{j,l} = \sum_{m=1}^{p} c_{j,m} d_m^l, \quad j \in H', l = 0,1,\ldots \tag{6.17}$$

D'après la définition (6.11) de $\underline{\bar{w}}$, nous avons

$$\mathcal{D}\underline{\bar{w}} = \sum_{m=1}^{p}\sum_{j\in H'} c_{j,m}\bar{d}_m\bar{\underline{e}}_j = \left(\sum_{j\in H'}\sum_{m=1}^{p} c_{j,m} d_m^l \underline{e}_j, l = 0,1,\ldots\right).$$

Donc, par définition des $\lambda_{j,l}$, nous obtenons

$$\mathcal{D}\underline{w} = \left(\sum_{j \in H'} \lambda_{j,l} \underline{e}_j \right) = (\underline{\Lambda}_0, \underline{\Lambda}_1, \ldots) = \underline{\Lambda}.$$

Pour voir que les $\lambda_{j,l}$ définis au moyen de (6.17) sont des nombres réels, prenons d'abord les conjuguées des deux membres de $\mathbf{S}\underline{c} = \underline{e}$. Notons que

$$\mathbf{S}^* = \mathbf{S}^*\left(d_1, \ldots, d_p\right) = \mathbf{S}\left(d_1^*, \ldots, d_p^*\right).$$

Puisque les $d_1, \ldots, d_p$ sont soit réelles ou viennent en paires conjuguées (voir la remarque 3.1), l'équation $\mathbf{S}^*\underline{c}^* = \underline{e}$ implique que, pour tout $j \in H'$ et tout $m = 1, \ldots, p$, soit $\Im d_m = 0$ et alors $c_{j,m}$ est réelle ou $\Im d_m \neq 0$ et alors il existe un $n \neq m$ $\left(\text{avec } d_n^* = d_m\right)$ tel que $c_{j,n}^* = c_{j,m}$. Par conséquent, les quantités $c_{j,m} d_m^l$ dans (6.17) sont soit réelles soit viennent en paires conjuguées. Donc, il découle de (6.17) que $\lambda_{j,l}$ est réel.

Ad. 2. Notons qu'en appliquant (6.1) et (6.4) à (6.11) après une algèbre facile, nous obtenons

$$\underline{w}_0 = \sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} \mathbf{N}(d_m) \underline{e}_j$$

et

$$\underline{w}_i = \sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} d_m^{i-1} \left(d_m \mathbf{I} - \mathbf{C}^T\right) \mathbf{N}(d_m) \underline{e}_j, \quad i = 1, 2, \ldots$$

Réécrivons les contraintes (2.2) et (2.3) en utilisant les formules susmentionnées pour $\underline{w}_0$ et $\underline{w}_i, i \geq 1$. La contrainte (2.2) pour $i = 0$ avec $\underline{w}_0$ définie plus haut prend la forme

$$\sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} \underline{1}^T \mathbf{N}(d_m) \underline{e}_j = 1 \tag{6.18}$$

et pour $i \geq 1$,

$$\sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} d_m^{i-1} \underline{1}^T \left(d_m \mathbf{I} - \mathbf{C}^T\right) \mathbf{N}(d_m) \underline{e}_j = 0. \tag{6.19}$$

La contrainte (2.3) pour $i = 0$, c'est-à-dire pour $\underline{w}_0$, est de la forme

$$\sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} \underline{e}_k^T \mathbf{N}(d_m) \underline{e}_j = 0, \quad k \in H. \tag{6.20}$$

Pour $i > 0$, elle prend la forme

$$\sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} d_m^{i-1} \underline{e}_k^T \left(d_m \mathbf{I} - \mathbf{C}^T\right) \mathbf{N}(d_m) \underline{e}_j = 0, \quad k \in H. \tag{6.21}$$

Notons que la matrice de dimensions $N \times N$

$$\mathbf{N}(d) = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho d & \ddots & 0 & 0 \\ 0 & 1 & \ddots & \ddots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \ddots & \ddots & 1 & -\rho d \\ 0 & 0 & \ddots & 0 & 1-\rho^2 \end{bmatrix}$$

et $\left(d\mathbf{I} - \mathbf{C}^T\right)\mathbf{N}(d) = \frac{d}{1-\rho^2}\mathbf{H}_N(d)$ - voir (3.8). Donc, par des calculs élémentaires, nous obtenons

$$\underline{e}_k^T \mathbf{N}(d)\,\underline{e}_j = \frac{1}{1-\rho^2}\begin{cases} (N-1)(1-d\rho)+1-\rho^2, & k=j=0, \\ 1-d\rho, & k=0,\ j\in H \text{ ou } k\in H,\ j=0, \\ 1, & k=j, \\ -d\rho, & k=j-1, \\ 0, & \text{autrement,} \end{cases} \quad k, j \in H \tag{6.22}$$

et

$$\begin{aligned} &\underline{e}_k^T \left(d\mathbf{I} - \mathbf{C}^T\right)\mathbf{N}(d)\,\underline{e}_j \\ &\quad = \frac{1}{1-\rho^2}\begin{cases} (N-1)(1-d\rho)(d-\rho)+d\left(1-\rho^2\right), & k=j=0, \\ (1-d\rho)(d-\rho), & k=0,\ j\in H \text{ ou } k\in H,\ j=0, \\ -\rho, & k=j+1, \\ d\left(1+\rho^2\right), & k=j, \\ -d^2\rho, & k=j-1, \\ 0, & \text{autrement,} \end{cases} \quad k, j \in H. \end{aligned} \tag{6.23}$$

En raison de (6.22) et (6.23), les contraintes (6.18), (6.19), (6.20) et (6.21) peuvent être réécrites sous la forme matricielle

$$\begin{bmatrix} \widetilde{\mathbf{G}}(d_1) & \widetilde{\mathbf{G}}(d_2) & \cdots & \widetilde{\mathbf{G}}(d_p) \\ \bar{\mathbf{G}}(d_1) & \bar{\mathbf{G}}(d_2) & \cdots & \bar{\mathbf{G}}(d_p) \\ d_1\bar{\mathbf{G}}(d_1) & d_2\bar{\mathbf{G}}(d_2) & \cdots & d_p\bar{\mathbf{G}}(d_p) \\ \vdots & \vdots & \ddots & \vdots \\ d_1^i\bar{\mathbf{G}}(d_1) & d_2^i\bar{\mathbf{G}}(d_2) & \cdots & d_p^i\bar{\mathbf{G}}(d_p) \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \underline{c} = \bar{e}, \tag{6.24}$$

où $\widetilde{\mathbf{G}}(d)$ est définie au moyen de (3.5) et (3.6),

$$\bar{\mathbf{G}}(d) = \frac{d}{1-\rho^2}\begin{bmatrix} \mathbf{H}_{11}(d) & \mathbf{H}_{12}(d) \\ \mathbf{H}_{21}(d) & \mathbf{H}_{22}(d) \end{bmatrix}$$

avec

$$\mathbf{H}_{11}(d) = (N-1)(1-\rho d)(1-\rho/d) + 1 - \rho^2,$$
$$\mathbf{H}_{12} = \mathbf{H}_{21}^T = (1-\rho d)(1-\rho/d)\,\underline{1}_h^T,$$
$$\mathbf{H}_{22}(d) = \operatorname{diag}(\mathbf{H}_1(d), \ldots, \mathbf{H}_s(d)),$$

et les matrices $\mathbf{H}_i(d), i = 1, \ldots, s,$ sont définies en (3.8).

La matrice infinie dans le premier membre de (6.24) peut s'écrire sous la forme

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{I} & \cdots & \mathbf{I} \\ \mathbf{0} & d_1\mathbf{I} & d_2\mathbf{I} & \cdots & d_p\mathbf{I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & d_1^i\mathbf{I} & d_2^i\mathbf{I} & \cdots & d_p^i\mathbf{I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{G}}(d_1) & \widetilde{\mathbf{G}}(d_2) & \cdots & \widetilde{\mathbf{G}}(d_p) \\ \overline{\mathbf{G}}(d_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{G}}(d_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \overline{\mathbf{G}}(d_p) \end{bmatrix},$$

où $\mathbf{I} = \mathbf{I}_{h+1}$ et $\mathbf{0} = \mathbf{0}_{h+1}$ sont, respectivement, les matrices unité et nulle de dimensions $(h+1) \times (h+1)$. Notons que la première matrice dans le produit qui précède est de plein rang et peut s'écrire sous la forme

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & d_1 & d_2 & \cdots & d_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & d_1^i & d_2^i & \cdots & d_p^i \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix} \otimes \mathbf{I}_{h+1}.$$

Donc, (6.24) équivaut à

$$\begin{bmatrix} \widetilde{\mathbf{G}}(d_1) & \widetilde{\mathbf{G}}(d_2) & \cdots & \widetilde{\mathbf{G}}(d_p) \\ \overline{\mathbf{G}}(d_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{G}}(d_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \overline{\mathbf{G}}(d_p) \end{bmatrix} \underline{c} = (1, 0, \ldots, 0)^T \in \mathbb{R}^{(p+1)(h+1)}. \tag{6.25}$$

Supposons que nous prouvions que les matrices $\overline{\mathbf{G}}(d_m), m = 1, \ldots, p$ de dimensions $(h+1) \times (h+1)$ sont singulières. Notons que $d[\mathbf{H}_{21}(d), \mathbf{H}_{22}(d)] = \mathbf{G}(d)$ en raison de (3.7). Donc, la définition (3.4) de $\mathbf{S}$ implique que (6.25) équivaut à $\mathbf{S}\underline{c} = (1, 0, \ldots, 0) \in \mathbb{R}^{ph+h+1}$. Cela s'obtient en partant de (6.25) et en supprimant toutes les lignes déterminées au moyen des premières lignes des

matrices $\bar{\mathbf{G}}(d_m), m = 1, \ldots, p$. Et l'équation $\mathbf{S}\underline{c} = (1, 0, \ldots, 0)$ s'ensuit en vertu de l'HYPOTHÈSE II et de la définition de $\underline{c}$.

Par conséquent, il suffit de montrer que $\det \bar{\mathbf{G}}(d_m) = 0, m = 1, \ldots, p$. Autrement dit, nous devons vérifier que

$$0 = \det \begin{bmatrix} \mathbf{H}_{11}(d_m) & \mathbf{H}_{12}(d_m) \\ \mathbf{H}_{21}(d_m) & \mathbf{H}_{22}(d_m) \end{bmatrix}$$

pour tout $m = 1, \ldots, p$.

Notons que, si $d = d_m$, le deuxième membre peut s'écrire

$$\det \mathbf{H}_{22}(d) \det \left[\mathbf{H}_{11}(d) - \mathbf{H}_{12}(d) \mathbf{H}_{22}^{-1}(d) \mathbf{H}_{21}(d) \right]$$

et

$$\det \mathbf{H}_{22}(d) = \prod_{i=1}^{s} \det \mathbf{H}_{m_i}(d). \tag{6.26}$$

Puisque nous pouvons décomposer $\mathbf{H}_m(d)$ comme il suit

$$\mathbf{H}_m(d) = \mathbf{D}_m^{-1} \mathbf{R}_m \mathbf{D}_m, \tag{6.27}$$

où $\mathbf{D}_m = \operatorname{diag}\left(1, d, d^2, \ldots, d^{m-1}\right)$ et $\mathbf{R}_m$ est définie en (3.2), nous voyons que

$$\det \mathbf{H}_m(d) = 1 + \rho^2 + \ldots + \rho^{2m} \neq 0.$$

Or, de (6.26) il découle que $\det \mathbf{H}_{22} \neq 0$.

Par ailleurs,

$$\det \left[\mathbf{H}_{11}(d) - \mathbf{H}_{12}(d) \mathbf{H}_{22}^{-1}(d) \mathbf{H}_{21}(d) \right] = (N-1)\alpha(\rho, d) + 1 - \rho^2 - \alpha^2(\rho, d) \sum_{j=1}^{s} \underline{1}^T \mathbf{H}_{m_j}^{-1} \underline{1}, \tag{6.28}$$

où $\alpha(\rho, d) = 1 + \rho^2 - \left(d + d^{-1}\right)\rho$.

La décomposition (6.27) de $\mathbf{H}_m$ donne

$$\underline{1}^T \mathbf{H}_m^{-1} \underline{1} = \operatorname{tr}\left(\underline{1}^T \mathbf{D}_m^{-1} \mathbf{R}_m^{-1} \mathbf{D}_m \underline{1}\right) = \operatorname{tr}\left(\mathbf{D}_m \underline{1}\underline{1}^T \mathbf{D}_m^{-1} \mathbf{R}_m^{-1}\right).$$

En outre, puisque $\operatorname{tr}(\mathbf{A}) = \operatorname{tr}\left(\mathbf{A}^T\right)$,

$$\underline{1}^T \mathbf{H}_m^{-1} \underline{1} = \operatorname{tr}\left(\left(\mathbf{D}_m \underline{1}\underline{1}^T \mathbf{D}_m^{-1} \mathbf{R}_m^{-1}\right)^T\right) = \operatorname{tr}\left(\mathbf{R}_m^{-1} \mathbf{D}_m^{-1} \underline{1}\underline{1}^T \mathbf{D}_m\right) = \operatorname{tr}\left(\mathbf{D}_m^{-1} \underline{1}\underline{1}^T \mathbf{D}_m \mathbf{R}_m^{-1}\right).$$

En combinant les deux dernières expressions pour $\underline{1}^T \mathbf{H}_m^{-1} \underline{1}$, nous obtenons

$$\underline{1}^T \mathbf{H}_m^{-1} \underline{1} = \operatorname{tr}\left(\tfrac{1}{2}\left(\mathbf{D}_m \underline{1}\underline{1}^T \mathbf{D}_m^{-1} + \mathbf{D}_m^{-1} \underline{1}\underline{1}^T \mathbf{D}_m\right) \mathbf{R}_m^{-1}\right).$$

Notons que

$$\left(\mathbf{D}_m \underline{1}\underline{1}^T \mathbf{D}_m^{-1} + \mathbf{D}_m^{-1} \underline{1}\underline{1}^T \mathbf{D}_m\right)_{ij} = d^{|i-j|} + d^{-|i-j|},$$

et que

$$\tfrac{1}{2}\left(d^k + d^{-k}\right) = T_k\left(\tfrac{1}{2}\left(d + d^{-1}\right)\right), \quad k = 0,1,\ldots,$$

où (T_k) est le k^{e} polynôme de Tchebychev de la première espèce.

Donc,

$$\underline{1}^T \mathbf{H}_m^{-1} \underline{1} = \text{tr}\mathbf{T}_m(x)\mathbf{R}_m^{-1},$$

où $x = x(d) = \tfrac{1}{2}\left(d + d^{-1}\right)$ et la matrice $\mathbf{T}_m$ est définie en (3.1). En introduisant cette expression dans (6.28), nous trouvons que

$$\det\left(\mathbf{H}_{11}(d) - \mathbf{H}_{12}(d)\mathbf{H}_{22}^{-1}(d)\mathbf{H}_{21}(d)\right) = Q_p(x(d)),$$

où Q_p est le polynôme défini en (3.3). Selon l'HYPOTHÈSE I, $Q_p(x(d_m)) = 0$, donc l'égalité susmentionnée donne $\det \bar{\mathbf{G}}(d_m) = 0, m = 1,\ldots,p$. Enfin, nous concluons que les contraintes (2.2) et (2.3) sont satisfaites, et donc que la preuve du point 2 est achevée.

Ad. 3. Premièrement, nous allons montrer que, pour $\underline{r}$ défini par (6.14), l'identité (6.15) est vérifiée. Pour cela, observons que, en vertu de (6.6) pour $i = p$, (6.10) et (6.13)

$$\mathcal{L}^{p+1}\left(\mathcal{I} - \sum_{m=1}^{p} a_m \mathcal{R}^m\right)\overline{\underline{w}} = \mathcal{L}\left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\mathcal{D}^{-1}\overline{\underline{\Lambda}} = \mathcal{L}\mathcal{D}^{-1}\left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\overline{\underline{\Lambda}}.$$

Notons aussi que pour tout $j = 1,\ldots,p$, en vertu de (6.8),

$$\left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\overline{d}_j = v_p(d_j)\overline{d}_j.$$

De la définition (3.10) de $a_m, m = 1,\ldots,p$ il découle que $v_p(d_j) = 0$. En raison de la définition de Λ au moyen de (6.17), nous concluons que $\mathcal{L}^{p+1}\overline{r} = \overline{\underline{0}}$.

Afin de vérifier (6.16), commençons par noter qu'en raison de (6.10), il découle de (6.3) et (6.5) que, pour $\overline{y} = \left(y^n\right)_{n\geq 0}$ et $\overline{\underline{x}} = (\underline{x}, \underline{x}, \ldots)$

$$\mathcal{D}^{-1}\overline{y}\overline{\underline{x}} = \left(\mathcal{I} - \mathbf{C}^T\mathcal{R}\right)\mathbf{N}(y)\,\overline{y}\overline{\underline{x}}.$$

Donc, pour tout $i \geq 0$, tout d_j et $\underline{e}_{j_k}$, d'après (6.6)

$$\begin{aligned}\mathcal{L}^i\left(\mathcal{I} - \sum_{m=1}^{p} a_m \mathcal{R}^m\right)\mathcal{D}^{-1}\overline{d}_j\underline{e}_{j_k} &= \left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\mathcal{R}^{p-i}\overline{d}_j\mathbf{N}\left(d_j\right)\underline{e}_{j_k} \\ &- \left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\mathcal{R}^{p-(i-1)}\overline{d}_j\mathbf{C}^T\mathbf{N}\left(d_j\right)\underline{e}_{j_k}.\end{aligned}$$

Enfin, nous utilisons (6.7) avec $\bar{y} = \bar{d}_j, \underline{\bar{x}} = \mathbf{N}\left(d_j\right)\underline{\bar{e}}_{j_k}$ dans la première partie et avec $\bar{y} = \bar{d}_j$, $\underline{\bar{x}} = \mathbf{C}^T\mathbf{N}\left(d_j\right)\underline{\bar{e}}_{j_k}$ dans la deuxième partie de l'expression du deuxième membre de l'équation ci-dessus pour arriver à

$$(\mathcal{I} - \mathcal{R}\mathcal{L})\,\mathcal{L}^i\left(\mathcal{I} - \sum_{m=1}^{p} a_m \mathcal{R}^m\right)\mathcal{D}^{-1}\bar{d}_j\underline{\bar{e}}_{j_k} = \left(v_i\left(d_j\right)\mathbf{I} - v_{i-1}\left(d_j\right)\mathbf{C}^T\right)\mathbf{N}\left(d_j\right)\left(\underline{e}_{j_k}, \underline{0}, \underline{0}, \ldots\right).$$

Donc (6.16) est vérifiée.

Enfin, nous allons prouver la formule (3.12) de variance de l'estimateur BLUE $\hat{\mu}_t$. Pour cela, nous commençons par observer que

$$\mathbb{C}\mathrm{ov}\left(\hat{\mu}_t, \underline{X}_{t-i}\right) = \underline{w}_i + \sum_{k=1}^{N-1}\mathbf{C}^k\,\underline{w}_{i+k} + \sum_{k=1}^{i\wedge(N-1)}\left(\mathbf{C}^T\right)^k\,\underline{w}_{i-k}$$

pour tout $i = 0,1,\ldots$. Par ailleurs, en raison de (6.13), nous voyons que le deuxième membre de l'égalité qui précède est égal à $\underline{\Lambda}_i$. Autrement dit, pour tout $i = 0,1,\ldots$

$$\mathbb{C}\mathrm{ov}\left(\hat{\mu}_t, \underline{X}_{t-i}\right) = \sum_{j\in H'}\lambda_{j,i}\underline{e}_j.$$

Maintenant, écrivons

$$\mathbb{V}\mathrm{ar}\hat{\mu}_t = \sum_{i=0}^{\infty}\underline{w}_i^T\mathbb{C}\mathrm{ov}\left(\hat{\mu}_t, \underline{X}_{t-i}\right) = \sum_{i=0}^{\infty}\sum_{j\in H'}\lambda_{j,i}\underline{w}_i^T\,\underline{e}_j.$$

En raison des contraintes (2.2) et (2.3), il découle de la formule qui précède que $\mathbb{V}\mathrm{ar}\hat{\mu}_t = \lambda_{0,0}$. Donc, (3.12) découle de (6.17).

Bibliographie

Australian Bureau of Statistics (2002). Labour Force Survey sample design. *Document d'information, numéro de catalogue 6269.0*.

Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association,* 70, 23-30.

Bell, P. (2001). Comparaison d'autres estimateurs pour l'Enquête sur la population active. *Techniques d'enquête,* 27, 1, 57-68.

Binder, D.A., et Hidiroglou, M.A. (1988). Sampling in time. *Handbook of Statistics,* 6, 187-211.

Cantwell, P.J. (1988). Variance formulae for the generalized composite estimator under balanced one-level rotation plan. *SRD Research Report Census/SRD/88/26*, Bureau of the Census, Statistical Research Division, 1-16.

Cantwell, P.J. (1990). Formules de variance pour estimateurs composites dans les plans de renouvellement. *Techniques d'enquête,* 16, 1, 163-174.

Cantwell, P.J., et Caldwell, C.V. (1998). Examining the revisions in monthly retail and wholesale trade surveys under a rotation panel design. *Journal of Official Statistics,* 14, 47-54.

Ciepiela, P., Gniado, M., Wesołowski, J. et Wojtyś, M. (2012). Dynamic K-composite estimator for an arbitrary rotation scheme. *Statistics in Transition,* 13(1), 7-20.

Fuller, W.A. et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête,* 27, 1, 49-56.

Gurney, M., et Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 242-257.

Hansen, M.H., Hurwitz, W.N., Nisselson, H. et Steinberg, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association,* 50, 701-719.

Kowalski, J. (2009). Optimal estimation in rotation patterns. *Journal of Statistical Planning and Inference,* 139, 1405-1420.

Lent, J., Miller, S.M., Cantwell, P.J. et Duff, M. (1999). Effects of composite weights on some estimates from current population survey. *Journal of Official Statistics,* 15(3), 431-448.

Lind, J.T. (2005). Repeated surveys and the Kalman filter. *The Econometrics Journal,* 9, 1-10.

McLaren, C.H., et Steel, D.G. (2000). L'effet de divers plans de renouvellement sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance. *Techniques d'enquête,* 26, 2, 185-195.

Patterson, H.D. (1950). Sampling on successive occasions. *Journal of the Royal Statistical Society, Series B,* 12, 241-255.

Popiński, W. (2006). Development of the Polish Labour Force Survey. *Statistics in Transition,* 7(5), 1009-1030.

Rao, J.N.K., et Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association,* 59, 492-509.

Singh, M.P., Drew, J.D., Gambino, J.G. et Mayda, F. (1990). Méthodologie de l'enquête sur la population active du Canada 1984-1990. Statistique Canada, numéro de catalogue, 71-526.

Singh, A.C., Kennedy, B. et Wu, S. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquête,* 27, 1, 35-48.

Steel, D., et McLaren, C. (2002). In search of a good rotation pattern. *Advances in Statistics, Combinatorics and Related Areas,* Singapore, World Scientific, 309-319.

Steel, D., et McLaren, C. (2008). Design and analysis of repeated surveys. Centre for Statist. Survey Meth., Univ. Wollonong, document de travail 11-08 (2008), 1-13, http://ro.uow.edu.au/cssmwp/10.

Szarkowski, A., et Witkowski, J. (1994). The Polish Labour Force Survey. *Statistics in Transition,* 1(4), 467-483.

Towhidi, M., et Namazi-Rad, M.-R. (2010). An optimal method of estimation in rotation sampling. *Advanced Applied Statistics,* 15(2), 115-136.

U.S. Bureau of Census (2002). The Current Population Survey - Design and Methodology. Department of Commerce, document technique 63.

Wesołowski, J. (2010). Recursive optimal estimation in Szarkowski rotation scheme. *Statistics in Transition,* 11(2), 267-285.

Yansaneh, I.S., et Fuller, W.A. (1998). Méthode optimale d'estimation récursive pour les enquêtes répétitives. *Techniques d'enquête,* 24, 1, 33-42.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 131-150
Statistique Canada, n° 12-001-X au catalogue

Ajustements optimaux pour les incohérences dans les données imputées

Jeroen Pannekoek et Li-Chun Zhang[1]

Résumé

Les microdonnées imputées contiennent fréquemment des renseignements contradictoires. La situation peut découler, par exemple, d'une imputation partielle faisant qu'une partie de l'enregistrement imputé est constituée des valeurs observées de l'enregistrement original et l'autre, des valeurs imputées. Les règles de vérification qui portent sur des variables provenant des deux parties de l'enregistrement sont alors souvent enfreintes. L'incohérence peut aussi résulter d'un ajustement pour corriger des erreurs dans les données observées, aussi appelé imputation dans la vérification (*imputation in editing*). Sous l'hypothèse que l'incohérence persistante n'est pas due à des erreurs systématiques, nous proposons d'apporter des ajustements aux microdonnées de manière que toutes les contraintes soient satisfaites simultanément et que les ajustements soient minimaux selon une mesure de distance choisie. Nous examinons différentes approches de la mesure de distance, ainsi que plusieurs extensions de la situation de base, dont le traitement des données catégoriques, l'imputation totale et l'étalonnage à un macroniveau. Nous illustrons les propriétés et les interprétations des méthodes proposées au moyen de données économiques des entreprises.

Mots-clés : Règles de vérification; microdonnées cohérentes; optimisation; étalonnage.

1 Introduction

Notre propos est de rapprocher les valeurs contradictoires dans les microdonnées imputées. En guise d'exemple, considérons une petite partie d'un enregistrement provenant d'une enquête structurelle sur les entreprises présentée au tableau 1.1. Postulons deux schémas de réponse, l'un où nous observons seulement le chiffre d'affaires et l'autre où nous observons aussi l'effectif et la rémunération. Les moyens d'imputer les valeurs manquantes dans un tel enregistrement *receveur* sont nombreux et les méthodes d'ajustement que nous proposons s'appliquent quelle que soit la méthode d'imputation choisie. L'utilisation de l'imputation partielle par donneur est illustrée au tableau 1.1, où l'enregistrement *donneur* est le « plus proche voisin » issu de la même catégorie d'activité économique que l'enregistrement receveur et est le plus proche de ce dernier en ce qui concerne le chiffre d'affaires pour le schéma de réponse (I), et en ce qui concerne l'effectif, le chiffre d'affaires et la rémunération pour le schéma de réponse (II). L'imputation est dite partielle parce qu'une valeur du donneur est transférée au receveur si et uniquement si la valeur correspondante manque dans l'enregistrement receveur.

Les enregistrements de données des entreprises doivent généralement respecter un certain nombre de contraintes comptables et logiques. Dans le contexte de la vérification de la validité d'un enregistrement, ces contraintes sont appelées règles de vérification. Pour l'enregistrement choisi comme exemple ici, supposons que les trois règles de vérification suivantes sont formulées :

$a1: \quad x_1 - x_5 + x_8 = 0$ (profit = chiffre d'affaires − total des coûts)
$a2: \quad x_5 - x_3 - x_4 = 0$ (chiffre d'affaires = chiffre d'affaires principal + autre chiffre d'affaires)
$a3: \quad x_8 - x_6 - x_7 = 0$ (total des coûts = rémunération + autres coûts).

1. Jeroen Pannekoek, Statistics Netherlands, Henri Faasdreef 312, 2492 JP La Haye, Pays-Bas. Courriel : j.pannekoek@cbs.nl; Li-Chun Zhang, University of Southampton, Social Statistics and Demography, Highfield SO17 1BJ, Southampton, UK et Statistics Norway, Kongensgate 6, Pb 8131 Dep, 0033 Oslo, Norvège. Courriel : L.Zhang@soton.ac.uk.

L'imputation partielle par donneur entraîne la violation de ces trois règles de vérification, situation que nous nommons *problème de cohérence* (*de microniveau*) : pour le schéma de réponse (I), les deux premières règles de vérification portant sur le chiffre d'affaires sont enfreintes; pour le schéma de réponse (II), les trois règles de vérification sont enfreintes. Pour obtenir un enregistrement cohérent, *certaines* des huit valeurs (c'est-à-dire incluant les valeurs observées ainsi qu'imputées) doivent être modifiées. Or, dans les deux cas examinés ici, il est possible de ne remplacer que les valeurs imputées pour satisfaire à toutes les règles de vérification, donc considérons pour le moment les ajustements des valeurs imputées.

Tableau 1.1
Données, données manquantes et valeurs du donneur pour les variables d'un enregistrement d'entreprise. Effectif (nombre d'employés); chiffre d'affaires principal (chiffre d'affaires de l'activité principale); autre chiffre d'affaires (chiffre d'affaires d'autres activités); chiffre d'affaires (chiffre d'affaires total); rémunération (coûts des salaires et traitements)

Variable	Nom	Réponse (I)	Réponse (II)	Valeurs du donneur
x_1	Profit			330
x_2	Effectif		25	20
x_3	Chiffre d'affaires principal			1 000
x_4	Autre chiffre d'affaires			30
x_5	Chiffre d'affaires	950	950	1 030
x_6	Rémunération		550	500
x_7	Autres coûts			200
x_8	Total des coûts			700

Les méthodes d'ajustement habituelles, telles que l'ajustement proportionnel mis en œuvre dans le logiciel Banff (Banff Support Team 2008), sont conçues pour traiter une contrainte à la fois. Dans le cas du schéma de réponse (I), la méthode d'ajustement proportionnel pourrait se dérouler comme il suit : 1) ajuster les valeurs imputées pour le total des coûts et le profit d'un facteur 950/1 030 afin que leur somme soit égale au chiffre d'affaires observé, 2) ajuster du même facteur les valeurs imputées pour le chiffre d'affaires principal et l'autre chiffre d'affaires pour satisfaire la deuxième règle de vérification et 3) ajuster les valeurs imputées de la rémunération et des autres coûts, de nouveau du même facteur, afin que leur somme soit égale à la valeur ajustée précédente du total des coûts.

Pour le schéma de réponse (II), les étapes (1) et (2) peuvent être exécutées comme auparavant, mais l'étape (3) doit être modifiée à moins que la rémunération observée doive être « écrasée ». Notons que le total des coûts figure dans deux règles de vérification : $a1$ et $a3$. Quand le total des coûts imputé est ajusté uniquement en fonction de $a1$ à l'étape (1), l'information pertinente dans la rémunération observée est ignorée. En effet, selon les valeurs disponibles, il peut même arriver que le total des coûts soit ajusté à la baisse à l'étape (1) au point qu'il ne reste aucune solution non négative acceptable pour les autres coûts à l'étape (3). En général, l'ajustement d'une variable qui figure dans plusieurs règles de vérification en fonction de l'une d'elles seulement est non seulement sous-optimal en théorie, mais requiert aussi un

choix arbitraire de l'ordre dans lequel les règles de vérification doivent être appliquées, ce qui peut entraîner inutilement une panne de la procédure.

Sous l'hypothèse que l'incohérence n'est pas due à des erreurs systématiques, nous proposons une approche d'optimisation où toutes les contraintes sont traitées simultanément. À cette fin, il est commode d'exprimer les contraintes de vérification en notation matricielle, sous la forme $\mathbf{C}\mathbf{x} = \mathbf{d}$, où $\mathbf{C}$ est la matrice des *contraintes* (ou *restrictions*), et $\mathbf{d}$ est un vecteur constant. Pour les contraintes $a1$ à $a3$, nous avons

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 \end{pmatrix} \text{ et } \mathbf{d} = \mathbf{0}.$$

Les éléments non nuls dans une *ligne* de la matrice des contraintes identifient toutes les variables qui interviennent dans la contrainte de vérification correspondante, et les éléments non nuls dans une *colonne* de la matrice des contraintes identifient les contraintes de vérification qui font intervenir la variable correspondante.

En outre, il existe souvent des contraintes d'inégalité linéaire. Le cas le plus simple est la non-négativité de la plupart des variables économiques. Les contraintes peuvent alors être formulées sous la forme $\mathbf{C}_{\text{ég}}\mathbf{x} = \mathbf{d}_{\text{ég}}$ et $\mathbf{C}_{\text{inég}}\mathbf{x} < \mathbf{d}_{\text{inég}}$, qui correspondent aux contraintes d'égalité et d'inégalité. Pour simplifier l'exposé, nous adopterons, sans autre notation, l'expression compacte $\mathbf{C}\mathbf{x} \leq \mathbf{d}$.

Comme nous l'avons mentionné plus haut, les valeurs ne doivent ou ne devraient pas être toutes ajustées. Nous faisons donc une distinction générale entre les variables *libres* (ou ajustables) et *fixes* (non ajustables). Cela inclut comme cas particulier la situation où toutes les valeurs des données sont considérées ajustables. Nous insistons sur le fait que la distinction ne porte pas nécessairement sur celle entre les variables imputées et observées, et que l'imputation peut avoir été effectuée pour des valeurs manquantes ainsi que pour des valeurs observées incorrectes. Par exemple, certaines valeurs imputées peuvent être maintenues fixes parce qu'elles sont dérivées selon un raisonnement logique, comme dans l'imputation déductive, ou parce qu'elles ont été obtenues de sources externes qui sont considérées comme étant plus fiables. Par contre, certaines valeurs observées peuvent être considérées comme non fiables et il est permis de les modifier. Étant donné l'absence d'erreurs systématiques, une approche générale consiste à repérer les variables ajustables par « localisation des erreurs » (par exemple, de Waal, Pannekoek et Scholtus 2011), en traitant les valeurs imputées et observées comme étant aussi sujettes à erreur les unes que les autres. Néanmoins, dans la suite du texte, nous traiterons la plupart du temps les valeurs imputées comme étant ajustables et les valeurs observées, comme étant fixes, afin de faciliter l'exposé.

Étant donné les variables libres et fixes, l'enregistrement de données complet est partitionné en conséquence en sous-vecteurs $\mathbf{x}_{\text{libre}}$ et $\mathbf{x}_{\text{fixe}}$, et la matrice des contraintes, en $\mathbf{C}_{\text{libre}}$ et $\mathbf{C}_{\text{fixe}}$ contenant les colonnes de $\mathbf{C}$ qui correspondent à $\mathbf{x}_{\text{libre}}$ et $\mathbf{x}_{\text{fixe}}$, respectivement. Les contraintes pour les variables ajustables sont alors données par $\mathbf{C}_{\text{libre}}\mathbf{x}_{\text{libre}} \leq \mathbf{d} - \mathbf{C}_{\text{fixe}}\mathbf{x}_{\text{fixe}}$ ou, de façon équivalente, par

$$\mathbf{A}\mathbf{x}_{\text{libre}} \leq \mathbf{b} \tag{1.1}$$

où la matrice $\mathbf{A}$ représente les contraintes sur les variables libres et est appelée matrice *comptable*, et $\mathbf{b}$ représente le vecteur constant pour ces contraintes. Notons que la matrice des contraintes $\mathbf{C}$ est dérivée a priori d'après les règles de vérification seulement, sans référence aux données réelles, et est la même pour tous les enregistrements, tandis que la matrice comptable $\mathbf{A}$ diffère généralement d'un enregistrement à l'autre, puisque la distinction entre les variables libres et fixes varie entre les unités.

Notre stratégie en vue de remédier au problème d'incohérence de microniveau dans les données imputées consiste à apporter aux variables ajustables des ajustements qui sont minimaux selon une mesure de distance (ou de divergence) choisie, de façon que l'enregistrement ajusté satisfasse toutes les règles de vérification. Toutes les contraintes sont traitées simultanément en supposant qu'il n'y a pas d'erreurs systématiques.

Le reste de l'article est présenté comme il suit. À la section 2, nous décrivons l'approche d'optimisation. Nous considérons différentes mesures de distance (ou de divergence), ainsi que les ajustements auxquels elles donnent lieu, et nous illustrons leurs propriétés et leurs interprétations en utilisant l'exemple d'enregistrement décrit plus haut. À la section 3, nous discutons des extensions possibles de l'approche élémentaire aux ajustements fondés sur des hypothèses statistiques en plus des contraintes logiques, au traitement des données catégoriques, à l'imputation totale avec ajustement, et aux ajustements pour des contraintes d'étalonnage de macroniveau combinées aux contraintes de cohérence de microniveau. À la section 4, nous examinons les données sur la superficie des pâturages provenant du Recensement de l'agriculture de la Norvège de 2010, y compris une approche d'évaluation de l'incertitude due à la vérification. Enfin, nous concluons par un bref résumé à la section 5.

2 L'approche d'ajustement minimal

2.1 Le problème d'optimisation

Nous proposons de résoudre le problème de cohérence décrit plus haut en ajustant les variables libres simultanément et aussi peu que possible, de manière à ce que toutes les règles de vérification soient satisfaites. Représentons la partie *ajustable* de l'enregistrement *avant* l'ajustement par le vecteur $\mathbf{x}_0$ de dimension J et *après* l'ajustement, par le vecteur $\tilde{\mathbf{x}}$ de dimension J correspondant. Le problème d'optimisation peut être formulé comme suit :

$$\begin{aligned} \tilde{\mathbf{x}} &= \arg\min_{\mathbf{x}} D(\mathbf{x}, \mathbf{x}_0) \\ \text{s.c.} \quad & \mathbf{A}\tilde{\mathbf{x}} \leq \mathbf{b}, \end{aligned} \tag{2.1}$$

où $D(\mathbf{x}, \mathbf{x}_0)$ est une fonction mesurant la distance (ou divergence) entre $\mathbf{x}$ et $\mathbf{x}_0$, et $\mathbf{A}$ est la matrice comptable de dimensions $K \times J$ associée aux K contraintes sur $\tilde{\mathbf{x}}$ données en (1.1). Nous considérerons différentes fonctions D à la section 2.2.

Les conditions pour une solution du problème de minimisation (2.1) peuvent être trouvées en inspectant le lagrangien pour ce problème, lequel peut s'écrire sous la forme

$$L(\mathbf{x}, \boldsymbol{\alpha}) = D(\mathbf{x}, \mathbf{x}_0) + \boldsymbol{\alpha}^T (\mathbf{A}\mathbf{x} - \mathbf{b}) \tag{2.2}$$

où $\boldsymbol{\alpha}$ est un vecteur de dimension K de multiplicateurs de Lagrange, ou variables *duales*, avec composantes α_k, une pour chacune des K contraintes, et $\mathbf{a}_k$ est la k^{e} ligne (correspondant à la contrainte k) de la matrice comptable $\mathbf{A}_{K\times J}$. Notons qu'une contrainte de non-négativité additionnelle doit être appliquée à chaque α_k correspondant à une contrainte d'inégalité, mais non aux α_k des contraintes d'égalité.

La théorie de l'optimisation montre bien que, pour une fonction convexe $D(\mathbf{x}, \mathbf{x}_0)$ et des contraintes linéaires, la solution de (2.1) est donnée par les vecteurs $\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}}$ qui satisfont ce qu'il est convenu d'appeler les conditions de Karush-Kuhn-Tucker (KKT) (voir, par exemple, Luenberger 1984; Boyd et Vandenberghe 2004). L'une d'elles est que le gradient du lagrangien en ce qui concerne $\mathbf{x}$ est nul quand il est évalué à $\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}}$, c'est-à-dire

$$L'_{x_j}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}}) = D'_{x_j}(\tilde{\mathbf{x}}, \mathbf{x}_0) + \sum_k a_{kj}\tilde{\alpha}_k = 0, \tag{2.3}$$

où a_{kj} est l'élément (k, j) de $\mathbf{A}$, et $L'_{x_j}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}})$, le gradient de L en ce qui concerne x_j évalué à $\tilde{\mathbf{x}}$ et $\tilde{\boldsymbol{\alpha}}$, et D'_{x_j}, celui de D. L'examen de (2.3) montre comment divers choix de D mènent à différentes solutions du problème d'ajustement, auxquelles nous donnons le nom de *modèles d'ajustement*.

2.2 Fonctions de distance et modèles d'ajustement

Une fonction de distance d'usage très répandu dans de nombreux domaines de la statistique est la fonction des moindres carrés pondérés (MCP) donnée par $D(\mathbf{x}, \mathbf{x}_0) = 1/2\,(\mathbf{x} - \mathbf{x}_0)^T\,\mathbf{W}\,(\mathbf{x} - \mathbf{x}_0)$, où $\mathbf{W}$ est une matrice diagonale dont les éléments diagonaux sont w_j, pour $j = 1, ..., J$. Nous obtenons alors, à partir de (2.3), le modèle d'ajustement

$$\tilde{x}_j = x_{0,j} - \frac{1}{w_j}\sum_k a_{kj}\tilde{\alpha}_k. \tag{2.4}$$

Le critère MCP aboutit donc à des ajustements additifs : l'ajustement total de la valeur *initiale* $x_{0,j}$ est égal à la somme pondérée des ajustements qui correspondent à chacune des K contraintes. L'ajustement dû à la k^{e} contrainte dépend des éléments suivants :

- le paramètre d'ajustement (c'est-à-dire la variable duale) $\tilde{\alpha}_k$ qui décrit la grandeur de l'ajustement. Une plus petite valeur de $\tilde{\alpha}_k$ (en valeur absolue si k désigne une contrainte d'égalité) correspond à un plus petit ajustement; une valeur nulle de $\tilde{\alpha}_k$ signifie qu'aucun ajustement dû à la contrainte en question n'a lieu;
- la constante a_{kj} (c'est-à-dire un élément de la matrice comptable) qui décrit la direction et la grandeur de l'ajustement de la variable j. Souvent, a_{kj} vaut 1, -1 ou 0 et décrit alors si $x_{0,j}$ est ajustée par $\tilde{\alpha}_k$, $-\tilde{\alpha}_k$ ou ne l'est pas du tout;
- le poids w_j : les variables dont les poids sont élevés sont moins ajustées que celles dont les poids sont faibles. Le cas particulier de $w_j \equiv 1$ donne le critère des moindres carrés ordinaires

(MCO), où la quantité d'ajustement due à chaque contrainte est la même pour toutes les variables pertinentes.

Un choix particulier des poids est $w_j = 1/x_{0,j}$, pour $j = 1,...,J$, auquel cas les carrés des ajustements relatifs sont minimisés, et une grande valeur initiale (c'est-à-dire $x_{0,j}$) fait l'objet d'un plus grand ajustement qu'une valeur plus petite *en valeur absolue*. En divisant (2.4) par $x_{0,j}$, nous obtenons

$$\frac{\tilde{x}_j}{x_{0,j}} = 1 - \sum_k a_{kj}\tilde{\alpha}_k, \tag{2.5}$$

qui est un modèle d'ajustement additif pour le *ratio* entre les valeurs ajustée et non ajustée. On notera qu'il s'agit du développement en série de Taylor d'ordre un (c'est-à-dire autour de 0 pour tous les $\tilde{\alpha}_k$) de l'ajustement multiplicatif donné par

$$\frac{\tilde{x}_j}{x_{0,j}} = \prod_k \left(1 - a_{kj}\tilde{\alpha}_k\right). \tag{2.6}$$

Partant de (2.5) nous voyons que $\tilde{\alpha}_k$ détermine la variation relative de la valeur initiale $x_{0,j}$ à la valeur ajustée $\tilde{x}_j$, qui en valeur absolue est habituellement beaucoup plus petite que l'unité. Par exemple, $\tilde{\alpha}_k = \pm 0{,}2$ implique un ajustement de $|20\ \%|$ de $x_{0,j}$ si $a_{kj} = \pm 1$, ce qui est grand en pratique. Les produits des $\tilde{\alpha}_k$ sont par conséquent souvent beaucoup plus petits que les α_k proprement dits, auquel cas (2.5) devient une bonne approximation de (2.6), et l'on peut considérer l'ajustement MCP comme étant donné approximativement par le produit de tous les ajustements multiplicatifs propres aux contraintes.

L'ajustement multiplicatif par (2.6) peut changer le signe de $x_{0,j}$ si $a_{kj}\tilde{\alpha}_k > 1$ pour une certaine unité k. Les ajustements multiplicatifs qui préservent le signe de la valeur initiale $x_{0,j}$ peuvent être obtenus en utilisant la mesure de divergence de Kullback-Leibler (KL) (qui n'est pas formellement une fonction de distance) donnée par $D_{KL} = \sum_j x_j \left(\ln x_j - \ln x_{0,j} - 1\right)$. Nous avons alors, à partir de (2.3), le modèle d'ajustement

$$\tilde{x}_j = x_{0,j} \prod_k \exp\left(-a_{kj}\tilde{\alpha}_k\right). \tag{2.7}$$

L'ajustement dû à la contrainte k est égal à 1 si a_{kj} vaut 0 (c'est-à-dire aucun ajustement), il est égal à $\exp(\tilde{\alpha}_k)$ si a_{kj} vaut 1, et il est égal à $1/\exp(\tilde{\alpha}_k)$ si a_{ik} vaut -1. Puisque $1 - a_{kj}\tilde{\alpha}_k$ est l'approximation d'ordre un de $\exp\left(-a_{kj}\tilde{\alpha}_k\right)$ autour de $\tilde{\alpha}_k = 0$ si $a_{kj} \pm 1$, on peut s'attendre à ce que les critères MCP et KL donnent des ajustements similaires à condition que ceux-ci soient petits ou moyens.

2.3 Méthodes de résolution du problème d'ajustement minimal

Le problème général d'optimisation convexe (2.1) peut être résolu explicitement si la fonction d'objectif est celle des moindres carrés pondérés et qu'il existe seulement des contraintes d'égalité. Dans

ce cas, le lagrangien est $L(\mathbf{x}, \boldsymbol{\alpha}) = 1/2\,(\mathbf{x} - \mathbf{x}_0)^T\ \mathbf{W}(\mathbf{x} - \mathbf{x}_0) + \boldsymbol{\alpha}^T\ (\mathbf{A}\mathbf{x} - \mathbf{b})$, et les équations qu'il faut résoudre sont

$$L'_{\mathbf{x}}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{W}(\mathbf{x} - \mathbf{x}_0) + \mathbf{A}^T\boldsymbol{\alpha} = \mathbf{0} \tag{2.8}$$

$$L'_{\boldsymbol{\alpha}}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}. \tag{2.9}$$

En résolvant (2.8) pour trouver $\mathbf{x}$ et en substituant le résultat dans (2.9), nous obtenons

$$\tilde{\boldsymbol{\alpha}} = \left(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T\right)^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{b})$$

et alors, par substitution inverse dans (2.8), nous obtenons explicitement

$$\tilde{\mathbf{x}} = \mathbf{x}_0 - \mathbf{W}^{-1}\mathbf{A}^T\left(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T\right)^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{b}). \tag{2.10}$$

Pour d'autres fonctions d'objectif et avec des contraintes d'inégalité en général, il n'existe pas de solution explicite du problème (2.1). Cependant, de nombreux algorithmes en accès libre ou commerciaux sont disponibles pour résoudre le problème d'optimisation convexe. Pour l'application décrite dans le présent article, nous avons utilisé le langage de programmation R et appliqué le *Successive Projection Algorithm* (SPA) (ou *row action algorithm*) – voir par exemple, Censor et Zenios (1997). Le SPA est un algorithme itératif qui utilise les contraintes (lignes de la matrice comptable) une à une. En une itération, le vecteur x est ajusté séquentiellement à chacune des contraintes. L'opération d'ajustement avec une seule contrainte requiert uniquement la mise à jour des éléments du vecteur x qui interviennent dans cette contrainte (correspondant aux éléments non nuls de la ligne traitée de la matrice comptable). Une fois que toutes les contraintes ont été traitées, l'itération s'achève et la suivante commence. Pour le critère MCP, il existe un module (ou *package*) R qui met en œuvre l'algorithme SPA et est conçu spécialement pour le problème d'ajustement (van der Loo 2012).

2.4 Retour à l'exemple

Le tableau 2.1 montre les ajustements minimaux apportés à l'enregistrement du tableau 1.1 en utilisant les critères MCO, MCP et KL, respectivement. Les valeurs observées sont traitées comme fixes et inscrites en caractères gras, et les valeurs imputées sont ajustables. Pour la méthode MCP, nous utilisons $w_j = 1/x_{0,j}$, ce qui donne des résultats égaux à ceux produits par le critère KL jusqu'à la première décimale.

Pour les deux schémas de réponse, la procédure d'ajustement MCO donne pour la variable Autre chiffre d'affaires une valeur négative qui n'est pas acceptable (tableau 2.1). Quand la procédure MCO est réexécutée avec une contrainte de non-négativité pour la variable Autre chiffre d'affaires, le résultat est simplement zéro pour cette variable et 950 pour la variable Chiffre d'affaires principal en raison de la contrainte $a2$. Sans la contrainte de non-négativité, les ajustements MCO sont de -40 pour x_3 et x_4, et de -16 pour x_6 et x_7, c'est-à-dire le même ajustement pour chaque paire de variables figurant dans la même contrainte. La variable Total des coûts (x_8) fait partie de deux contraintes et son ajustement total comprend deux composantes additives. Une composante est due à la contrainte $a1$, et l'autre, à $a3$. Pour

le schéma de réponse (I), la première composante est -48 et la deuxième composante est 16, et leur somme est égale à -32 dans le tableau 2.1.

Tableau 2.1
Imputation et ajustement de l'enregistrement d'entreprise du tableau 1.1. ID : Imputation partielle par donneur sans ajustement; MCO : distance selon les moindres carrés ordinaires; MCP : distance selon les moindres carrés pondérés; KL : mesure de divergence de Kullback-Leibler; RG : ajustement par le ratio généralisé

Variable	Nom	Réponse (I)				Réponse (II)			
		ID	MCO	MCP/KL	RG	ID	MCO	MCP/KL	RG
x_1	Profit	330	282	291	304	330	260	249	239
x_2	Effectif	20	20	20	18	**25**	**25**	**25**	**25**
x_3	Chiffre d'affaires principal	1 000	960	922	922	1 000	960	922	921
x_4	Autre chiffre d'affaires	30	-10	28	28	30	-10	28	29
x_5	Chiffre d'affaires	**950**	**950**	**950**	**950**	**950**	**950**	**950**	**950**
x_6	Rémunération	500	484	470	461	**550**	**550**	**550**	**550**
x_7	Autres coûts	200	184	188	184	200	140	151	161
x_8	Total des coûts	700	668	658	646	700	690	701	711

Les ajustements MCP/KL sont plus grands, en valeur absolue, pour les grandes valeurs imputées que pour les valeurs plus petites. En particulier, l'ajustement pour Autre chiffre d'affaires n'est que de -2,3, de sorte qu'aucune valeur ajustée négative n'est produite dans ce cas, tandis que l'ajustement pour Chiffre d'affaires principal est de -77,7. On peut observer la nature multiplicative de ces ajustements car le *facteur* d'ajustement pour ces variables est égal à 0,92 (pour les deux schémas de réponse). Le facteur d'ajustement pour les variables Rémunération et Autres coûts sous le schéma de réponse (I) est égal à 0,94 dans les deux cas parce que ces variables figurent dans la même contrainte $a3$, de sorte que le ratio de leurs valeurs initiales n'est pas modifié par cet ajustement. Cependant, le ratio initial de chacune de ces variables à la variable Total des coûts n'est pas préservé, parce que le total des coûts possède un signe différent dans la contrainte $a3$ et, de surcroît, il fait aussi partie de la contrainte $a1$, si bien qu'il est sujet à deux facteurs d'ajustement.

3 Extensions possibles des problèmes d'ajustement connexes

3.1 Ajustement par le ratio généralisé

Habituellement, on utilise le modèle du ratio pour pondérer les cas dans les enquêtes auprès des entreprises en supposant que les variables économiques peuvent toutes être reliées proportionnellement à une mesure de taille commune de l'unité commerciale, voir par exemple, Särndal, Swensson et Wretman (1992). Motivés par le modèle du ratio, nous pourrions multiplier toutes les valeurs du donneur par 950/1 030 pour obtenir les valeurs imputées pour l'enregistrement pris comme exemple sous le schéma de

réponse (I), y compris la variable Effectif (x_2) pour laquelle la valeur imputée initiale de 20 n'enfreint formellement aucune contrainte. Cela montre qu'il existe peut-être des situations où, en plus des contraintes logiques et comptables, des ajustements pourraient être introduits en se fondant sur des hypothèses statistiques.

Pour le schéma de réponse (II), les variables observées Effectif (x_2), Chiffre d'affaires (x_5) et Rémunération (x_6) peuvent en principe être chacune utilisées comme variable de mesure de taille dans un modèle de ratio, de sorte que l'on ne peut pas dégager un ajustement par le ratio unique. Cependant, nous pouvons postuler l'existence d'un ratio *commun* entre les enregistrements receveur et donneur sous le modèle du ratio, et considérer les ratios observés (c'est-à-dire, 20/25 pour l'effectif, 950/1 030 pour le chiffre d'affaires et 550/500 pour la rémunération) comme les manifestations aléatoires de ce ratio commun. Donc, il semble qu'une approche plausible consiste à déterminer ce ratio commun comme étant la valeur qui minimise la variance, ou toute autre mesure de dispersion jugée appropriée, des trois ratios individuels. Enfin, dans la mesure où le ratio commun a trait aux autres variables, il devient possible d'ajuster celles-ci en utilisant l'approche du *ratio généralisé* (RG).

Supposons le modèle d'ajustement multiplicatif $\tilde{x}_j = x_{0,j}\delta_j$, où chaque δ_j est une manifestation aléatoire d'un ratio commun théorique. Soit la fonction de distance

$$D(\tilde{\mathbf{x}}, \mathbf{x}_0) = 1/2\left(\boldsymbol{\delta}^T\boldsymbol{\delta} - \bar{\delta}^2\right) \tag{3.1}$$

où $\boldsymbol{\delta}$ est le vecteur des δ_j et $\bar{\delta}$ est leur moyenne. Pour toutes les variables auxquelles est appliqué le ratio commun, incluant celles qui sont libres ainsi que celles qui sont fixes, nous effectuons maintenant l'ajustement en deux étapes. La première est une étape conceptuelle, où nous imaginons qu'un ajustement $\tilde{x}_j / x_{0,j}$ est apporté aux variables fixes : si $\tilde{x}_j = x_j$ est observée et fixe, alors $\delta_j = x_j / x_{0,j}$, tandis que $\delta_j = 1$ si $\tilde{x}_j$ est la valeur imputée $x_{0,j}$ mais devant être maintenue fixe pour un ajustement « supplémentaire ». À la deuxième étape, les ajustements sont effectués sur les valeurs initiales des variables libres par résolution du problème d'optimisation (2.1) en utilisant (3.1) comme fonction de distance. Cela donne les ajustements RG des trois variables libres concernées.

Une condition importante de l'approche RG est qu'au moins l'un des δ_j doit avoir trait à une variable fixe. Sinon, $\tilde{x}_j \equiv x_{0,j}$ serait une solution triviale, parce que cela donnera toujours $D = 0$. Notons que nous avons supprimé la notation J dans (3.1) et utilisé un peu abusivement les notations $\mathbf{x}_0$ et $\tilde{\mathbf{x}}$ introduites pour (2.1). Prenons le schéma de réponse (I) dans le tableau 1.1, la valeur fixe $x_5 = 950$ doit être incluse dans (3.1), ce qui donne $\delta_5 = \tilde{x}_5 / x_{0,j} = x_5 / x_{0,j} = 950/1\,030$. La résolution de (2.1) pour toutes les autres variables donne alors $\delta_j \equiv 950/1\,030$ et $D = 0$. Par contre, sans inclure δ_5, nous aurions obtenu $D = 0$ à $\delta_j = 1$ et $\tilde{x}_j = x_{0,j}$ pour $j \neq 5$.

Les ajustements RG pour le schéma de réponse (II) sont donnés au tableau 2.1. Les trois δ_j observés pour $j = 2, 5$ et 6 sont inclus dans (3.1) et maintenus fixes pour le problème d'optimisation. On voit que les résultats sont proches des ajustements MCP/KL. La variance empirique des facteurs multiplicatifs vaut 0,0270 pour les ajustements RG, 0,0276 pour les ajustements MCP/KL et 0,1434 pour les ajustements MCO. La somme relative des carrés des écarts, c'est-à-dire deux fois la distance MCP, vaut 50,6 pour les ajustements MCP/KL, 51,6 pour les ajustements RG et 78,0 pour les ajustements MCO. Enfin, la somme

non pondérée des carrés des écarts, c'est-à-dire deux fois la distance MCO, est de 20 925 pour les ajustements MCO, de 23 976 pour les ajustements MCP/KL et de 25 090 pour les ajustements RG. Donc, en ce qui concerne les trois fonctions de distance, les ajustements RG sont plus proches des ajustements MCP/KL que des ajustements MCO.

Or, les mesures de distance (ou de divergence) prises en considération à la section 2.2 pourraient être caractérisées comme étant *décomposables*, puisque la distance globale entre deux vecteurs est donnée par une somme (pondérée) des « distances » entre les composantes correspondantes. L'une des conséquences est qu'une variable ne figurant dans aucune des contraintes retiendra sa valeur initiale sous l'approche de l'ajustement minimal. Par contre, la distance (3.1) est *non décomposable*, chaque ajustement dépendant des autres ajustements. Par conséquent, même les valeurs qui n'interviennent explicitement dans aucune contrainte seront ajustées si elles sont incluses dans la fonction de distance, en raison des changements apportés aux variables qui sont liées aux contraintes. La variable Effectif dans le tableau 2.1 en est un exemple. L'approche RG offre donc la possibilité de faire des ajustements fondés sur des hypothèses statistiques en plus des contraintes logiques et comptables. En effet, si une seule variable fixe est incluse dans (3.1), les ajustements RG se réduisent à un ajustement proportionnel commun, conformément ici à la notion intuitive d'ajustement par le ratio. Si plusieurs variables fixes sont incluses, l'approche RG vise à produire une forme d'ajustements les plus uniformes en tant que généralisation du modèle de ratio unique. Pour le schéma de réponse (II) du tableau 1.1, l'approche tient compte d'un seul coup des trois ratios observés. Arriver au même résultat en formulant un modèle statistique explicite précisément pour ce schéma de réponse n'est pas aussi pratique dans des conditions de production.

3.2 Ajustements portant sur des données catégoriques

Une variable catégorique est associée à des contraintes différentes de celles d'une variable continue. Il vaut donc la peine d'examiner la mesure dans laquelle les variables catégoriques peuvent être incorporées dans l'approche d'optimisation. Nous distinguons trois types de données catégoriques que l'on rencontre fréquemment en pratique.

Premièrement, nous disons qu'une variable catégorique/discrète est *pseudo-continue* si, en pratique, elle peut être traitée comme s'il s'agissait d'une variable continue. Des exemples types de variables pseudo-continues sont l'âge, le nombre d'employés, la taille du ménage, etc. La pseudo-continuité peut avoir une incidence sur le choix du modèle d'ajustement et de la fonction de distance. Par exemple, des ajustements additifs ainsi que proportionnels peuvent être acceptables pour le nombre d'employés, tandis qu'un ajustement proportionnel de la taille du ménage ou de l'âge ne paraît pas naturel. Néanmoins, après avoir choisi le modèle d'ajustement et la fonction de distance, on peut traiter une variable pseudo-continue tout comme une vraie variable continue. Un arrondissement est nécessaire par après et son effet doit être surveillé.

Deuxièmement, nous appelons variable catégorique *nominale* une variable qui indique si une unité rentre dans une catégorie particulière. Une variable nominale avec M catégories, étiquetées $x = 1, 2, ..., M$, est associée à la contrainte

$$\prod_{m=1}^{M} (\tilde{x} - m) = 0. \tag{3.2}$$

Cependant, les étiquettes (par exemple, 1 = tomates, 2 = haricots, 3 = concombres) ne conviennent pas pour des opérations telles que l'addition, la multiplication ou l'arrondissement. En outre, une valeur nominale de 3 n'est pas plus distante de 1 que la valeur 2. Par conséquent, la contrainte (3.2) ne peut pas être prise en compte sous l'approche d'ajustement minimal qui suppose des mesures sur une échelle d'intervalle. L'ajustement d'une valeur observée qui ne satisfait pas (3.2) doit être traité en marquant cette valeur comme étant manquante, puis en imputant une valeur admissible ainsi qu'appropriée, c'est-à-dire tout comme dans le cas où la valeur manque dès le départ.

Troisièmement, une variable peut être définie comme étant nulle pour les unités qui ne sont pas admissibles. Selon que la mesure est pseudo-continue ou nominale quand l'unité est admissible, nous avons une *variable semi-continue/-nominale* ayant une probabilité non nulle d'être nulle. La différence par rapport à la pseudo-continuité susmentionnée est qu'une variable semi-continue peut nécessiter une contrainte supplémentaire de non-négativité dans la matrice comptable. Considérons alors une variable semi-nominale. En pratique, dans la conception des questionnaires, une telle variable est souvent divisée en deux, disons X_1 et X_2. Soit $X_1 = 1$ si l'unité s'adonne à une certaine activité, disons, la production de légumes en serre, et $X_1 = 0$ autrement. Soit X_2 une mesure nominale de l'activité quand $X_1 = 1$, et $X_2 = 0$ autrement. Formellement, la contrainte logique peut être exprimée par

$$(1 - \tilde{x}_1)\,\tilde{x}_2 + \tilde{x}_1 \prod_{m=1}^{M} (\tilde{x}_2 - m) = 0 \tag{3.3}$$

Considérons tous les schémas de données possibles, y compris quand une valeur manque (indiqué par « – ») :

- $(x_1, x_2) = (-, x_2)$: la valeur $\tilde{x}_1$ peut être déduite à condition que x_2 soit admissible, c'est-à-dire x_2 vaut 0 ou satisfait (3.2), sinon la situation devient le cas $(x_1, x_2) = (-, -)$ décrit plus bas.
- $(x_1, x_2) = (x_1, -)$: si $x_1 = 0$ alors $\tilde{x}_2 = 0$; si $x_1 = 1$ alors (3.3) se réduit à (3.2) mentionnée plus haut.
- $(x_1, x_2) = (-, -)$: les deux valeurs doivent être imputées par des valeurs qui satisfont (3.3).
- (x_1, x_2) : il y a violation de (3.3), par exemple si $(x_1, x_2) = (1, 0)$ ou si $x_1 = 0$ et $x_2 > 0$. Nous avons le cas $(-, x_2)$ décrit plus haut si x_2 est fixe, $(x_1, -)$ si x_1 est fixe, ou $(-, -)$ si ni l'une ni l'autre n'est fixe.

En résumé, les contraintes (3.2) et (3.3) ne peuvent pas être traitées par l'approche d'ajustement minimal avec les contraintes linéaires examinées plus haut. Elles doivent plutôt être traitées par la méthode d'imputation. Souvent, l'imputation par donneur (par exemple, le logiciel SCANCIR de Statistique Canada qui applique la méthode d'imputation par le plus proche voisin, MIPPV) peut être conçue pour imputer des données catégoriques de manière que les contraintes spécifiées par l'utilisateur soient satisfaites, voir par exemple, Bankier, Lachance et Poirier (2000).

3.3 Ajustement de l'imputation totale par donneur

Dans l'imputation totale par donneur, toutes les valeurs de l'enregistrement proviennent du donneur choisi. Cette approche offre des avantages par rapport à la modélisation conjointe de toutes les variables

cibles si celles-ci sont nombreuses. Chen et Shao (2000) établissent la cohérence de l'estimateur selon l'enquête fondé sur l'imputation par le plus proche voisin (IPPV) sous des conditions faibles. L'hypothèse clé est que la différence entre les espérances conditionnelles d'une variable cible dans un enregistrement donneur et un enregistrement receveur, sachant les variables sur lesquelles la mesure de distance est calculée, est bornée par la « distance » entre ces enregistrements. Autrement dit, si la « distance » entre eux est nulle, ils ont les mêmes espérances respectivement pour chacune des variables statistiques.

Il est donc nécessaire d'ajuster l'imputation totale par donneur quand la « distance » entre le receveur et le donneur n'est pas nulle. Pour illustrer ceci au moyen de l'enregistrement choisi comme exemple au tableau 1.1, supposons que le chiffre d'affaires (x_5) est toujours connu à partir d'une source administrative et qu'il est utilisé pour trouver le donneur, de sorte que l'imputation partielle sous le schéma de réponse (I) devienne une imputation totale. Puisque le chiffre d'affaires de l'enregistrement receveur diffère de celui de l'enregistrement donneur, la distance entre les deux enregistrements n'est pas nulle, et il paraît naturel que les valeurs du donneur soient ajustées pour tenir compte de cette différence. En effet, maintenant qu'il existe des contraintes faisant intervenir le chiffre d'affaires, des ajustements sont nécessaires de toute façon.

Posons que $\mathbf{x}$ contient les variables qui peuvent être manquantes. Posons que $\mathbf{z}$ contient les variables connues qui sont utilisées pour trouver le donneur. Soit $\mathbf{x}^* = \left(\mathbf{x}^T, \mathbf{z}^T\right)^T$ le vecteur combiné de variables. L'imputation totale (sachant $\mathbf{x}_0$) peut être considérée comme une imputation partielle du sous-vecteur manquant $\mathbf{x}$ de $\mathbf{x}^*$. L'ajustement de l'imputation totale peut être nécessaire s'il existe des règles de vérification qui font intervenir à la fois les valeurs de $\mathbf{x}$ et $\mathbf{z}$, et/ou s'il n'y a pas concordance exacte des valeurs de $\mathbf{z}$ entre le donneur et le receveur. En fait, l'imputation totale sans ajustement pourrait plutôt être considérée comme exceptionnelle en pratique.

3.4 Étalonnage de macroniveau en plus des contraintes de microniveau

Un recensement des entreprises doit faire appel à l'imputation et à la vérification afin d'obtenir un ensemble de données complet pour la production de statistiques. Ou bien, un registre statistique peut être créé en se basant sur une combinaison de données administratives et de données provenant d'une ou de plusieurs enquêtes. La vérification et l'imputation sont de nouveau nécessaires. Une caractéristique commune est que, contrairement au sondage, aucune pondération n'est nécessaire.

Durant le traitement de telles données, des contraintes d'*étalonnage* de macroniveau sont fréquemment imposées pour des raisons d'efficacité statistique et/ou de cohérence de macroniveau avec les sources externes. Une contrainte d'étalonnage est satisfaite si la somme des données complètes correspond au total d'étalonnage donné, qui peut se rapporter à différents niveaux d'agrégation, c'est-à-dire contenant des totaux pour la population ainsi que pour des sous-populations. Par exemple, certains totaux nationaux clés peuvent être estimés par une méthode appropriée et imposés comme contraintes d'étalonnage par la suite. Ou, un ensemble de contraintes d'étalonnage au niveau du domaine peut être obtenu par une technique d'estimation sur petits domaines. En outre, des contraintes d'étalonnage provenant de sources externes sont fréquentes dans les statistiques structurelles sur les entreprises – un exemple tiré du Recensement de l'agriculture de la Norvège de 2010 sera décrit à la section 4.

Les méthodes d'imputation sous contraintes d'étalonnage ont été étudiées par Beaumont (2005), Chambers et Ren (2004), Zhang (2009) et Pannekoek, Shlomo et de Waal (2013). L'approche adoptée ici

est similaire à celle suivie dans les deux premiers articles. Dans ces deux articles, une distance selon les moindres carrés pondérés entre les valeurs imputées initiales (ou les valeurs aberrantes dans le cas de Chambers et Ren 2004) et les valeurs imputées ajustées est minimisée sous la contrainte que les totaux pondérés par les poids de sondage basés sur les données ajustées soient égaux aux totaux d'étalonnage. Ici, nous supposons qu'une méthode d'imputation appropriée a été appliquée pour produire l'ensemble de données de population complet initial, qui peut ou non être étalonné. Le problème d'incohérence de microniveau implique que des ajustements de l'ensemble de données complet initial seront en général nécessaires.

Désignons par $\mathbf{X}$ l'ensemble de données complet d'intérêt, où chaque ligne correspond à un enregistrement au niveau de l'unité tel que celui du tableau 1.1, et chaque colonne correspond à une variable particulière. Soit $\mathbf{X}_0$ l'ensemble de données complet initial après imputation et $\tilde{\mathbf{X}}$ l'ensemble de données ajusté. Chaque contrainte d'étalonnage s'applique à un vecteur-colonne particulier de $\mathbf{X}$ et aux unités qui sont comprises dans son domaine. Autrement dit, cela peut s'exprimer génériquement sous la forme $\mathbf{r}^T \text{col}(\mathbf{X}) = t$, où $\text{col}(\mathbf{X})$ est le vecteur-colonne d'intérêt, et $\mathbf{r}$ est le vecteur d'indicateurs indiquant si une unité appartient au domaine d'intérêt, et t est le total d'étalonnage. De cette façon, toutes les contraintes d'étalonnage peuvent être résumées comme

$$[\mathbf{r}]^T [\text{col}(\mathbf{X})] = \mathbf{t} \tag{3.4}$$

où chaque colonne de $[\text{col}(\mathbf{X})]$ correspond à une contrainte d'étalonnage, et chaque colonne de $[\mathbf{r}]$ au vecteur d'indicateurs correspondant, et $\mathbf{t}$ est le vecteur de tous les totaux d'étalonnage. Notons la similarité entre (3.4) et (1.1). Une approche d'ajustement minimal s'ensuit en spécifiant les valeurs ajustables et fixes, ainsi que la fonction de distance (ou de divergence).

Tant les contraintes d'étalonnage que les contraintes de microniveau peuvent être considérées comme des contraintes linéaires sur le très long vecteur contenant tous les éléments de $\mathbf{X}$, $\text{vec}(\mathbf{X})$, disons. Conceptuellement, toutes les contraintes regroupées peuvent donc être exprimées sous la forme (1.1). La matrice des contraintes de cette formulation est, cependant, énorme et très éparse. Les lignes correspondant aux contraintes de microniveau peuvent contenir des valeurs non nulles concordant avec les valeurs de l'enregistrement auxquelles elles s'appliquent et des zéros pour toutes les autres valeurs de $\text{vec}(\mathbf{X})$, et les lignes correspondant aux contraintes d'étalonnage contiennent des éléments non nuls concordant uniquement avec les valeurs de $\text{vec}(\mathbf{X})$ qui contribuent au total d'étalonnage en question. En pratique, le problème d'optimisation généré par (3.4) en plus des contraintes de microniveau peut être traité en utilisant l'algorithme SPA, c'est-à-dire une contrainte à la fois et en opérant uniquement sur les éléments de $\text{vec}(\mathbf{X})$ correspondant à des éléments non nuls dans la contrainte en question, sans former effectivement cette matrice des contraintes énorme et éparse. Pour les contraintes d'étalonnage, nous devons uniquement traiter les colonnes de $[\text{col}(\mathbf{X})]$ une par une, et pour les contraintes de microniveau, nous traitons chaque enregistrement au niveau de l'unité un à la fois. Ces ajustements minimaux itératifs le long des colonnes et des lignes de $\mathbf{X}$ ressemblent à l'algorithme d'ajustement proportionnel itératif (ou de *raking*) pour l'ajustement des modèles log-linéaires sur des données de tableau de contingence et pour l'ajustement des tableaux (de contingence) sur de nouvelles marges, ce qui est formellement identique à un algorithme SPA avec les contraintes de divergence KL et d'égalité seulement.

4 Étude de cas

4.1 Imputation et ajustement des données sur les pâturages

La population cible du « questionnaire principal » du Recensement de l'agriculture de la Norvège de 2010 contient environ 45 000 unités. Les questions 22 à 24 ont trait à la superficie des pâturages :

- La question 22 demande quelles sont les unités qui possèdent des pâturages productifs.
- La question 23 demande quelle est la superficie totale des pâturages productifs en 2010.
- La question 24 demande la composition de la superficie des pâturages en fonction de la dernière fois où ils ont été ensemencés : 1) 2006 à 2010, 2) 2001 à 2005, et 3) 2000 ou antérieurement.

Désignons par $x_{0,1}, x_{0,2}$ et $x_{0,3}$ les trois catégories de superficie des pâturages déclarées à la question 24. Soit $x_0 = \sum_{j=1}^{3} x_{0,j}$ la somme qui est le sujet de la question 23. Ce total peut aussi être obtenu auprès de l'organisme gouvernemental qui administre la subvention pertinente. À l'étape de la vérification, la valeur déclarée de x_0 est remplacée par le chiffre administratif, désigné par $\tilde{x}$, et maintenue fixe par la suite. Ensuite, la réponse à la question 22 peut être inférée en sachant $\tilde{x}$ et maintenue fixe par la suite, de sorte qu'il ne reste qu'à traiter la question 24.

Ci-après nous décrivons le traitement des 34 480 unités possédant une superficie de pâturages productifs selon leurs profils d'observation respectifs (tableau 4.1, où l'indice d'unité i de toutes les variables a été omis pour faciliter l'exposé).

- 10 378 unités ont déclaré une superficie totale des pâturages conforme aux données de la source administrative : il s'agit des donneurs potentiels; aucun ajustement n'est nécessaire.
- 11 827 unités ont déclaré un total supérieur à la valeur connue : elles présentent un problème d'incohérence de microniveau. Naturellement, il pourrait aussi s'agir de valeurs manquantes si $\sum_j r_j < 3$, mais les chances sont faibles, si bien que nous supposerons qu'il n'y a pas de valeurs manquantes parmi ces unités. Toutes les valeurs observées sont ajustables, de sorte que l'équation comptable est donnée par

  $$\sum_{j;r_j=1} \tilde{x}_j = \tilde{x}.$$

  L'approche RG donne simplement l'ajustement proportionnel $\tilde{x} \Big/ \sum_{j;\, r_j=1} x_{0,j}$. Le même ajustement est donné par l'approche MCP avec $w_j = 1 \big/ x_{0,j}$ si $r_j = 1$, ainsi que par l'approche KL. Nous notons qu'il n'existe aucune raison particulière d'envisager des ajustements additifs pour ces données.
- 3 876 unités *n'*ont *pas* déclaré de superficie des pâturages d'aucune sorte, alors qu'elles possèdent une superficie de pâturages productifs selon la source administrative : il s'agit d'enregistrements avec données totalement manquantes. Le donneur qui est le plus proche voisin (PPV) est trouvé en fonction de $\tilde{x}$, dans chacune des 12 « formes d'agriculture », qui représentent une classification connue pour l'ensemble de la population. Dans le cas de donneurs PPV multiples, nous avons choisi celui pour lequel la distance physique était la plus

courte, ce qui rend l'imputation PPV entièrement déterministe, étant donné toutes les valeurs $\tilde{x}$. Enfin, un ajustement proportionnel des valeurs du donneur est effectué afin de satisfaire l'équation comptable

$$\sum_{j;r_j^*=1} \tilde{x}_j = \tilde{x}$$

où r_j^* est l'indicateur d'observation/déclaration associé au donneur.

- 3 019 unités ont déclaré des superficies de pâturages de *chacun des trois* types, mais dont la somme est inférieure au total connu : ces unités présentent un problème d'incohérence de microniveau. Un ajustement proportionnel est appliqué à toutes les valeurs déclarées en ce qui concerne l'équation comptable $\sum_{j=1}^{3} \tilde{x}_j = \tilde{x}$.

- Le dernier groupe comprend les 2 703 unités qui ont déclaré une catégorie de superficie de pâturages et les 2 677 unités qui ont déclaré deux catégories de superficie de pâturages. Manifestement, ici, le fait que le total déclaré est inférieur à la valeur connue peut être causé par des valeurs incohérentes et/ou manquantes. Pour éviter d'introduire un profil systématique dû à la vérification, nous laissons la décision dépendre du donneur. Prenons une unité ayant déclaré une seule catégorie de superficie des pâturages. Premièrement, les donneurs potentiels sont limités à ceux provenant de la même « forme d'agriculture », ainsi qu'ayant *au moins* la même catégorie de superficie des pâturages. Le donneur PPV est alors choisi parmi les donneurs potentiels de manière à minimiser

$$\max\left(\left|\tilde{x}^*/\tilde{x} - 1\right|, \left|x_j^*/\tilde{x}^* - x_{0,j}/\tilde{x}\right|_{j;r_j=1}\right)$$

où $\left(x_1^*, x_2^*, x_3^*\right)$ et $\tilde{x}^*$ sont les valeurs du donneur potentiel. Autrement dit, le donneur PPV est choisi en ce qui concerne à la fois la différence relative entre les superficies totales des pâturages et la proportion de la catégorie déclarée de superficie des pâturages par rapport au total correspondant. Soit le donneur PPV associé à $\mathbf{x}^*$ et à $\mathbf{r}^*$. Si $\sum_j r_j^* > 1 = \sum_j r_j$, alors nous supposons qu'il existe des valeurs manquantes où $r_j^* = 1$ mais $r_j = 0$; tandis que si $\sum_j r_j^* = \sum_j r_j$, alors nous supposons qu'il existe uniquement un problème d'incohérence. Les opérations d'imputation et d'ajustement restantes sont simples. Le même traitement est appliqué aux unités ayant déclaré deux catégories de superficie des pâturages, avec les modifications évidentes dues au fait que $\sum_j r_j = 2$.

Tableau 4.1
Profil d'observation parmi les unités avec superficie de pâturages productifs : $r_j = 1$ si $x_{0,j}$ est déclaré, $r_j = 0$ autrement; $j = 1, 2, 3$ pour les trois catégories de superficie de pâturages

			$\sum_j r_j x_{0,j} < \tilde{x}$			
Total	$\sum_j r_j x_{0,j} = \tilde{x}$	$\sum_j r_j x_{0,j} > \tilde{x}$	$\sum_j r_j = 0$	$\sum_j r_j = 1$	$\sum_j r_j = 2$	$\sum_j r_j = 3$
34 480	10 378	11 827	3 876	2 703	2 677	3 019

Les totaux de sous-population et de population basés sur l'imputation avec ajustement sont donnés au tableau 4.2, comparativement aux totaux des données brutes et aux totaux du fichier de recensement. Nos constatations sont les suivantes. a) Le fichier de recensement a été vérifié de la façon « conventionnelle » qui requiert beaucoup de travail manuel (environ 1,5 personne-année en tout). Par contre, ici, les procédures de vérification sont entièrement automatisées, et tout le travail (c'est-à-dire analyse exploratoire, décision concernant les traitements, programmation et traitement) a été effectué en moins de deux jours. Même si les questions concernant les superficies des pâturages ne sont qu'au nombre de 3 sur un total de 36 questions du « questionnaire principal », il est évident que l'économie de temps possible pourrait être énorme. b) Les différences entre les totaux imputés et les totaux de recensement sont faibles pour toutes les sous-populations, comparativement à celles observées entre les données brutes et les totaux de recensement. Tous les changements par rapport aux données brutes vont dans la « bonne » direction, si l'on en juge d'après les résultats du recensement. On peut conclure que les procédures de vérification automatisées ont abouti à la plupart des résultats de vérification du recensement. c) Il est possible d'ajouter des contraintes d'étalonnage. À titre d'exemple, nous avons utilisé les totaux de sous-population du fichier de recensement pour les 3 876 enregistrements avec données totalement manquantes, en plus de la superficie totale connue des pâturages pour chacun d'eux. La convergence a été atteinte en 23 itérations en utilisant le critère MCP. d) Pour les 5 380 unités pouvant contenir des données manquantes partielles, l'imputation des valeurs « manquantes » a été effectuée pour environ 25 % d'entre elles dans le cadre du traitement du recensement, tandis que la proportion est d'environ 75 % pour la procédure de vérification décrite ici. Le nombre de cas de données partiellement manquantes est probablement sous-estimé dans le fichier du recensement parce que ce nombre est fondé sur des vérifications manuelles sélectives. Quoi qu'il en soit, malgré les différences entre les traitements individuels, les totaux vérifiés sont assez proches de chacun (tableau 4.2, sous $0 < \sum_j r_j < 3$).

Tableau 4.2
Superficies totales des pâturages des sous-populations et de la population fondées sur les données brutes, l'imputation avec ajustement et les données de production du recensement (tous les chiffres $\times 10^5$)

	$\sum_j r_j x_{0,j} > \tilde{x}$			$\sum_j r_j x_{0,j} < \tilde{x}$					
				$\sum_j r_j = 3$			$0 < \sum_j r_j < 3$		
Brutes	8,20	6,95	12,76	1,40	1,45	1,53	1,33	0,86	3,05
Imput. et ajust.	5,24	4,34	8,71	1,72	1,81	1,88	2,01	1,87	3,51
Recensement	5,47	4,37	8,45	1,73	1,85	1,84	2,04	1,54	3,80
	$\sum_j r_j = 0$			$\sum_j r_j > 0$			**Total**		
Brutes	-	-	-	14,0	12,4	21,9	-	-	-
Imput. et ajust.	1,20	1,06	1,93	12,2	11,3	19,3	13,43	12,38	21,17
Recensement	1,31	1,23	1,66	12,6	11,0	19,1	13,95	12,25	20,79

4.2 Estimation approximative de l'erreur quadratique moyenne

À titre de mesure de l'incertitude des données sur la superficie des pâturages, nous utilisons ici l'erreur quadratique moyenne de prédiction (EQMP) donnée par

$$\text{EQMP}_j = E\left\{\left(\tilde{X}_j - X_j\right)^2 \middle| \mathbf{R}_U, \tilde{\mathbf{X}}_U\right\}$$

où $X_j = \sum_{i \in U} x_{ij}$ est le total de population cible et $\tilde{X}_j = \sum_{i \in U} \tilde{x}_{ij}$ est le total correspondant fondé sur l'imputation avec ajustement, pour $j = 1, 2, 3$. En outre, $\tilde{\mathbf{X}}_U = (\tilde{x}_i)_{i \in U}$ contient les totaux connus des superficies de pâturages dans la population, et $\mathbf{R}_U$ est la matrice des indicateurs de données manquantes dont la i^{e} ligne est donnée par (r_{i1}, r_{i2}, r_{i3}).

Or, même s'il est habituel de parler d'imputation lorsque l'on fait référence aux ajustements dus aux incohérences dans les microdonnées dans le cadre de la vérification de données statistiques, l'éventuelle incertitude qui y est associée est généralement « ignorée » par la suite. Cela revient à supposer que $\tilde{x}_{ij} = x_{ij}$ si $r_{ij} = 1$. Ce qu'il reste à expliquer est l'incertitude associée à l'imputation des valeurs manquantes et à l'ajustement subséquent des valeurs du donneur, sous l'hypothèse que ni l'imputation ni l'ajustement n'introduit un biais dans la valeur finale. Cela revient à supposer que $E\left(\tilde{x}_{ij} - x_{ij}\right) = 0$ si $r_{ij} = 0$. Sous ces deux hypothèses, nous avons

$$\begin{aligned}
\text{EQMP}_j &= E\left\{\left(\sum_{i \in U}\left(1 - r_{ij}\right)\tilde{x}_{ij} - \sum_{i \in U}\left(1 - r_{ij}\right)x_{ij}\right)^2\right\} \\
&= V\left(\sum_{i \in U;\, \mathbf{r}_i = \mathbf{1},\, d_{ij} \geq 1} d_{ij}\delta_{ij}x_{ij}\right) + V\left(\sum_{i \in U;\, r_{ij} = 0} x_{ij}\right) \\
&\approx \sum_{i \in U;\, \mathbf{r}_i = \mathbf{1},\, d_{ij} \geq 1} d_{ij}^2 V\left(\delta_{ij}x_{ij}\right) + \sum_{i \in U;\, r_{ij} = 0} V\left(x_{ij}\right)
\end{aligned}$$

où d_{ij} est le nombre de fois que x_{ij} est utilisée comme valeur du donneur pour l'imputation des données manquantes, et la décomposition de la variance est vérifiée à condition que les distributions des unités soient indépendantes les unes des autres. En outre, à condition que $d_{ij} \geq 1$,

$$\delta_{ij} = \sum_{k \in U;\, x_{kj}^* = x_{ij}} \tilde{x}_{kj} \Big/ \left(d_{ij}x_{ij}\right)$$

où $x_{kj}^* = x_{ij}$ signifie que x_{ij} est utilisée comme valeur du donneur pour x_{kj}, et $\tilde{x}_{kj}$ est la valeur finale après ajustement. Autrement dit, δ_{ij} est l'ajustement combiné fait à $d_{ij}x_{ij}$, où $d_{ij}x_{ij}$ aurait été la contribution de x_{ij} à $\tilde{X}_j$ par imputation s'il s'était agi d'une imputation par donneur *sans* ajustement. Notons que d_{ij} peut être traitée comme une constante dans la dernière équation (approximative) à condition que l'identification du donneur dépende uniquement de $\mathbf{R}_U$ et $\tilde{\mathbf{X}}_U$. Cela est vrai pour les 3 876 enregistrements avec données totalement manquantes, mais pas exactement pour les 5 380 unités pour lesquelles des données pourraient être partiellement manquantes. Comme il est expliqué à la section 4.1, l'identification du PPV dépend en fait aussi des valeurs observées x_{ij}. Pour cette raison, la dernière équation n'est vérifiée qu'approximativement.

Un modèle de ratio pour la variance conditionnelle de x_{ij} semble naturel ici, c'est-à-dire

$$x_{ij} = \beta_j x_i + \varepsilon_{ij} \text{ où } E\left(\varepsilon_{ij}\right) = 0 \text{ et } V\left(\varepsilon_{ij}\right) = \sigma_j^2 x_i^{\alpha_j}$$

où $\left(\beta_j, \sigma_j^2, \alpha_j\right)$ peut varier en fonction de la *composition* des superficies des pâturages, désignée par $\mathbf{q} = (1,1,1),(1,1,0),(1,0,1)$ et $(0,1,1)$, où $q_{ij} = 1$ si l'unité i possède le j^{e} type de pâturage et 0 autrement. Notons que, dans le cas de $\sum_j q_{ij} = 1$, nous avons $x_{ij} = \tilde{x}$ si $q_{ij} = 1$, de sorte que la variance conditionnelle est nulle. Les paramètres de ce modèle de ratio peuvent être estimés d'après les 10 378 donneurs potentiels satisfaisant $\sum_j r_j x_{0,j} = \tilde{x}$. L'analyse exploratoire des données montre que $\alpha_j = 2$ est un choix raisonnable dans tous les cas de sorte que, dans les calculs qui suivent, seules β_j et σ_j^2 varient en fonction du profil d'observation, désigné par $\left(\beta_{j;h}, \sigma_{j;h}^2\right)$ pour $h = 1,...,4$. Notons qu'en raison de $\alpha_j \equiv 2$, on obtiendra le même $\hat{\sigma}_{j;h}^2$ quel que soit j quand $\sum_j q_{ij} = 2$. Par exemple, si nous prenons $\mathbf{q} = (1,1,0)^T$, nous avons $\hat{\beta}_1 + \hat{\beta}_2 = 1$, de sorte que les résidus prédits « centrés réduits » sont donnés par $\hat{\varepsilon}_{i1}/\tilde{x}_i = x_{i1}/\tilde{x}_i - \hat{\beta}_1$ et $\hat{\varepsilon}_{i2}/\tilde{x}_i = x_{i2}/\tilde{x}_i - \hat{\beta}_2 = (\tilde{x}_i - x_{i1})/\tilde{x}_i - \left(1 - \hat{\beta}_1\right) = -\hat{\varepsilon}_{i1}/\tilde{x}_i$. De toute façon, nous obtenons $\hat{V}_h\left(x_{ij}\right) = \hat{\sigma}_{j;h}^2 \tilde{x}_i^2$ pour l'unité i ayant la composition h.

Le facteur d'ajustement δ_{ij} semble difficile à modéliser d'avance. Mais sa moyenne et sa variance, notées $\mu_\delta = E\left(\delta_{ij}\right)$ et $\sigma_\delta^2 = V\left(\delta_{ij}\right)$ respectivement, peuvent être estimées empiriquement *après* avoir effectué l'imputation et l'ajustement. En outre, nous supposons que δ_{ij} est indépendant de x_{ij} sachant $\tilde{x}_i$. Cela semble une hypothèse plausible, puisque le premier dépend principalement de la distribution de x dans le « voisinage » de $x = \tilde{x}$, tandis que la seconde dépend de la variation sur j étant donné que la somme est égale à $\tilde{x}$. Par exemple, asymptotiquement, à mesure que la chance de trouver un donneur dans tout voisinage arbitrairement proche tend vers l'unité, le facteur d'ajustement δ_{ij} tend vers 1 en probabilité, indépendamment des valeurs de x_{ij}. Il s'ensuit alors que, sachant la composition h, une estimation de la variance correspondante $V_h\left(\delta_{ij} x_{ij}\right)$ est donnée par

$$\hat{V}_h\left(\delta_{ij} x_{ij}\right) = \hat{\sigma}_{j;h}^2 \tilde{x}_i^2 \hat{\sigma}_\delta^2 + \left(\hat{\beta}_{j;h} \tilde{x}_i\right)^2 \hat{\sigma}_\delta^2 + \hat{\sigma}_{j;h}^2 \tilde{x}_i^2 \hat{\mu}_\delta^2.$$

Enfin, en combinant tous les éléments susmentionnés, nous obtenons une estimation approximative de l'EQMP sous la forme

$$\widehat{\text{EQMP}}_j \approx \sum_h \sum_{i \in U_h; \mathbf{r}_i = \mathbf{1}} d_{ij}^2 \hat{V}_h\left(\delta_{ij} x_{ij}\right) + \sum_h \sum_{i \in U_h; r_{ij} = 0} \hat{V}_h\left(x_{ij}\right).$$

Les résultats de l'estimation approximative de la variance sont donnés au tableau 4.3. Nous savons d'avance que le coefficient de régression du modèle de ratio doit varier en fonction de la composition de la superficie des pâturages, mais les estimations de $\sigma_{j;h}^2$ donnent à penser qu'il était raisonnable de permettre que le paramètre de variance dépende de h. La moyenne estimée de δ_{ij} est proche de l'unité pour toutes les catégories de superficie des pâturages, ne donnant donc aucun indice que les hypothèses concernant les facteurs d'ajustement ne sont pas raisonnables. La variance de δ_{ij} est clairement la plus grande pour $j = 2$, ce que reflète aussi le fait que l'EQMP estimée dans ce cas présente l'augmentation la plus importante par rapport à l'imputation PPV sans ajustement. Les racines carrées relatives de l'EQMP sont trop faibles pour expliquer les différences réelles entre les totaux de recensement et les totaux imputés (donnés au tableau 4.2). Cela illustre l'impression générale qui suit concernant l'évaluation

de l'incertitude due à la vérification. Les effets systématiques pour ce qui est des moments d'ordre un des statistiques résultantes sont habituellement les éléments qui dominent l'incertitude globale due à la vérification. Mais ils sont aussi plus difficiles à quantifier que les propriétés de variance d'ordre deux. Ici, cela concerne les deux hypothèses d'« ordre un » faites au début, c'est-à-dire $\tilde{x}_{ij} = x_{ij}$ si $r_{ij} = 1$ et $E\left(\tilde{x}_{ij} - x_{ij}\right) = 0$ si $r_{ij} = 0$. Des hypothèses plus complexes au sujet du mécanisme d'erreur des ajustements de cohérence dans la vérification sont nécessaires afin d'aller au-delà de cette approche « optimiste ».

Tableau 4.3
Estimation de la variance approximative pour l'imputation avec ajustement. REQMP : racine carrée de l'EQMP. REQMP pour l'imputation PPV sans ajustement entre parenthèses

		$j = 1$	$j = 2$	$j = 3$
$\hat{\beta}_j$	$\mathbf{q} = (1,1,1)$	0,312	0,359	0,329
	$\mathbf{q} = (1,1,0)$	0,346	0,654	-
	$\mathbf{q} = (1,0,1)$	0,407	-	0,593
	$\mathbf{q} = (0,1,1)$	-	0,567	0,433
$\hat{\sigma}_j^2$	$\mathbf{q} = (1,1,1)$	0,0248	0,0511	0,0364
	$\mathbf{q} = (1,1,0)$	0,0478	0,0478	-
	$\mathbf{q} = (1,0,1)$	0,0464	-	0,0464
	$\mathbf{q} = (0,1,1)$	-	0,0798	0,0798
	$\left(\hat{\mu}_\delta; \hat{\sigma}_\delta^2\right)$	(0,992; 0,0248)	(1,020; 0,0994)	(1,003; 0,0236)
	$\widehat{\text{REQMP}}$	3 267 (3 134)	4 190 (3 530)	3 111 (2 925)
	$\widehat{\text{REQMP}} \Big/ \sum_{i;r_{ij}=0} \tilde{x}_{ij}$	1,41 %	1,79 %	0,93 %
	$\widehat{\text{REQMP}} \Big/ \tilde{X}_j$	0,24 %	0,34 %	0,15 %

5 Sommaire

Dans le présent article, nous avons formulé une approche d'optimisation du problème d'incohérence de microniveau qui peut résulter d'erreurs de mesure et/ou de l'imputation de valeurs manquantes. Cette approche fournit une méthodologie générale qui s'étend au-delà des méthodes d'ajustement avec contrainte unique classiques, telles que l'ajustement proportionnel. Toutes les contraintes sont traitées simultanément; si une variable figure dans plus d'une contrainte, elle est ajustée en fonction de chacune d'elles. En plus d'être optimale en ce qui a trait à la fonction de distance (ou de divergence) choisie, l'approche a l'avantage pratique de ne pas nécessiter la spécification de l'ordre d'application des contraintes.

Plusieurs fonctions de distance (ou de divergence) sont analysées. Il est montré que minimiser la distance selon les moindres carrés pondérés mène à des ajustements additifs et minimiser la mesure de divergence de Kullback-Leibler aboutit à des ajustements multiplicatifs. Cependant, pour un choix particulier de poids, la solution MCP du problème d'optimisation est une approximation de la solution KL.

Les ajustements basés sur des hypothèses statistiques en plus des contraintes logiques sont introduits sous l'approche du ratio généralisé. Les ajustements RG sont considérés comme une généralisation de

l'ajustement par ratio unique sous un modèle de ratio. Tous les ratios entre les enregistrements receveur et donneur propres à une variable observée sont utilisés; une variable qui ne figure dans aucune contrainte peut également être ajustée si elle est incluse dans la fonction de distance.

Nous discutons aussi des ajustements dans les cas de données catégoriques, d'enregistrements avec données totalement manquantes et de contraintes d'étalonnage de macroniveau en plus des contraintes de cohérence de microniveau. Dans son ensemble, l'approche d'optimisation proposée est applicable à des données continues dans un certain nombre de situations.

Remerciements

Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas nécessairement les politiques de *Statistics Netherlands.*

Bibliographie

Banff Support Team (2008). *Functional Description of the Banff System for Edit and Imputation*. Rapport Technique, Statistique Canada.

Bankier, M., Lachance, M. et Poirier, P. (2000). *2001 Canadian Census Minimum Change Donor Imputation Methodology*. Document de travail 17, UN/ECE Work Session on Statistical Data Editing, Cardiff.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society, Series B* (Statistical Methodology), 67, 445-458.

Boyd, S., et Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Censor, Y., et Zenios, S.A. (1997). *Parallel Optimization*. Theory, Algorithms, and Applications. Oxford University Press, New York.

Chambers, R.L., et Ren, R. (2004). Outlier robust imputation of survey data. Dans *Proceedings of the Survey Research Methods Section,* American Statistical Association, 3336-3344.

Chen, J., et Shao, J. (2000). Biases and variances of survey estimators based on nearest neighbour imputation. *Journal of Official Statistics*, 16, 113-132.

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

Luenberger, D.G. (1984). *Linear and Nonlinear Programming, Second Edition*. Addison-Wesley, Reading.

Pannekoek, J., Shlomo, N. et de Waal, T. (2013). Calibrated imputation of numerical data under linear edit restrictions. *Annals of Applied Statistics,* 7, 1983-2006.

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

van der Loo, M. (2012). rspa: Adapt numerical records to (in)equality restrictions with the Successive Projection Algorithm. R package version 0.1-5. Disponible au : http://cran.r-project.org/web/packages/rspa/index.html.

Zhang, L.-C. (2009). *A Triple-Goal Imputation Method for Statistical Registers*. Document de travail 28, UN/ECE Work Session on Statistical Data Editing, Neuchâtel, Suisse.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 151-172
Statistique Canada, n° 12-001-X au catalogue

Traitement de la non-réponse non ignorable dans les enquêtes : une approche de modélisation par variables latentes

Alina Matei et M. Giovanna Ranalli[1]

Résumé

La non-réponse est présente dans presque toutes les enquêtes et peut fortement biaiser les estimations. On distingue habituellement la non-réponse totale et la non-réponse partielle. En notant que pour une variable d'enquête en particulier, nous avons uniquement des valeurs observées et des valeurs inobservées, nous exploitons dans la présente étude le lien entre la non-réponse totale et la non-réponse partielle. En particulier, nous supposons que les facteurs qui sous-tendent la réponse totale sont les mêmes que ceux qui sous-tendent la réponse partielle pour les variables d'intérêt choisies. Nous estimons alors les probabilités de réponse en utilisant une covariable latente qui mesure la *volonté de répondre à l'enquête* et qui peut expliquer, en partie, le comportement inconnu d'une unité en ce qui concerne la participation à l'enquête. Nous estimons cette covariable latente en nous servant de modèles à traits latents. Cette approche convient particulièrement bien pour les questions sensibles et, par conséquent, permet de traiter la non-réponse non ignorable. L'information auxiliaire connue pour les répondants et les non-répondants peut être incluse dans le modèle à variables latentes ou dans le processus d'estimation de la probabilité de réponse. L'approche peut également être utilisée quand on ne dispose pas d'information auxiliaire, et nous nous concentrons ici sur ce cas. Nous proposons une estimation au moyen d'un système de repondération basé sur la covariable latente précédente quand aucune autre information auxiliaire observée n'est disponible. Les résultats d'études par simulation en vue d'évaluer sa performance en se servant de données réelles ainsi que simulées sont encourageants.

Mots-clés : Non-réponse totale; non-réponse partielle ; modèles à traits latents; propension à répondre; modèles de Rasch.

1 Introduction

La non-réponse est un problème de plus en plus fréquent dans les enquêtes. Il s'agit d'un problème parce qu'elle se traduit par des données manquantes et, surtout, parce que ces lacunes sont une source possible de biais dans les estimations d'enquêtes. En présence de non-réponse totale, on suppose souvent qu'une probabilité de répondre à l'enquête est associée à chaque unité de la population. Cette probabilité de répondre est inconnue et plusieurs méthodes sont proposées pour l'estimer explicitement, en modélisant la propension à répondre, par exemple à l'aide de modèles de régression logistique (voir, par exemple, Kim et Kim 2007), ou implicitement, en utilisant des groupes de réponse homogènes ou, plus généralement, le calage (voir Särndal et Lundström 2005, pour un aperçu). Après avoir calculé les estimations, une méthode souvent utilisée pour traiter la non-réponse totale est la repondération : les poids de sondage des répondants sont ajustés par l'inverse de la probabilité estimée de répondre pour obtenir de nouveaux poids. L'estimation des probabilités de répondre requiert habituellement l'accès à de l'information auxiliaire, sous la forme de la valeur de certaines variables auxiliaires pour toutes les unités dans l'échantillon sélectionné au départ, ou de leur moyenne ou de leur total de population.

Dans le présent article, nous nous intéressons tout spécialement au cas où le mécanisme de création des données manquantes est non ignorable, parce que la non-réponse dépend des caractéristiques d'intérêt qui

1. Alina Matei, Institut de statistiques, Université de Neuchâtel, Pierre à Mazel 7, 2000, Neuchâtel, Suisse. Courriel : alina.matei@unine.ch et Institut de recherche et de documentation pédagogique, Neuchâtel, Suisse; M. Giovanna Ranalli, Département de sciences politiques, Université de Pérouse, Italie. Courriel : giovanna.ranalli@stat.unipg.it.

sont soit observées uniquement sur les répondants ou qui sont entièrement inobservées, ce qui mène à des données qui ne manquent pas au hasard (NMAR pour *Not Missing At Random*). Cette situation est typique des enquêtes comportant des questions sensibles (sur l'abus de drogues, les attitudes sexuelles, la politique, le revenu, etc.), mais sans s'y limiter. Diverses approches sont proposées dans la littérature sur les enquêtes pour traiter la non-réponse non ignorable. Ces approches peuvent être classées, de manière générale, en méthodes fondées sur la vraisemblance et en méthodes de repondération. Notons que toutes ces méthodes font appel à de l'information auxiliaire observée. Les problèmes relatifs aux enquêtes avec des non-réponses non ignorables sont discutés, par exemple, dans Greenlees, Reece et Zieschang (1982), Little et Rubin (1987), Beaumont (2000), Qin, Leung et Shao (2002), Zhang (2002). Copas et Farewell (1998) ont introduit dans la *National Survey of Sexual Attitudes and Lifestyles* du Royaume-Uni une variable appelée « empressement à répondre » à l'enquête, qui en principe est reliée aux probabilités de réponse totale et de réponse partielle. Les auteurs proposent une méthode d'estimation de ces probabilités en utilisant la variable susmentionnée pour obtenir des estimations sans biais des paramètres de population. Une approche fondée sur l'utilisation de variables latentes pour modéliser la non-réponse non ignorable est décrite dans Biemer et Link (2007), qui étendent les idées présentées dans Drew et Fuller (1980) et se servent d'une variable latente discrète basée sur l'historique des appels disponible pour toutes les unités de l'échantillon. La variable latente est calculée en utilisant certains indicateurs du niveau d'effort basés sur les tentatives d'appel.

Nous proposons ici une méthode de repondération afin de réduire le biais de non-réponse dans le cas d'une non-réponse non ignorable. La méthode ne requiert pas d'information auxiliaire au niveau de l'échantillon ni de la population, mais différentes hypothèses sont formulées. Premièrement, nous supposons que l'enquête fait l'objet d'une non-réponse partielle et que celle-ci affecte m variables présentant un intérêt particulier. Donc, nous pouvons définir pour chaque variable ℓ, pour $\ell = 1, \ldots, m$, un indicateur de réponse qui prend la valeur 1 si l'item ℓ est observé sur l'unité k et 0 autrement. Ensuite, nous supposons que les indicateurs de réponse sont des manifestations d'une échelle continue sous-jacente qui déterminent une variable latente qui est reliée à la propension à répondre des unités et à la variable d'intérêt. Nous pouvons calculer une telle variable latente non seulement pour les répondants, mais aussi pour toutes les unités de l'échantillon et donc l'utiliser comme variable auxiliaire dans une procédure d'estimation des probabilités de réponse. Enfin, nous pouvons utiliser le résultat de cette procédure d'estimation pour la repondération.

L'utilisation de variables latentes continues pour modéliser la non-réponse partielle est examinée dans Moustaki et Knott (2000). Dans la présente étude, nous adoptons une perspective différente et utilisons des modèles à variables latentes pour traiter la non-réponse totale non ignorable. Nous proposons d'utiliser une variable latente que nous appelons « volonté de répondre à l'enquête », en principe reliée à la probabilité de réponse totale, qui est similaire à la variable d'« empressement à répondre » définie par Copas et Farewell (1998). Comme l'ont indiqué Moustaki et Knott (2000), [*Traduction*] « la pondération en faisant appel à la modélisation par variables latentes devrait donner de bons résultats en présence de non-réponse non ignorable quand le conditionnement sur les covariables observées seulement ne suffit pas ». En outre, en l'absence de toute covariable, nous nous attendons à ce qu'un estimateur basé sur le système de pondération proposé utilisant des variables latentes donnera, en ce qui concerne la réduction du biais, de meilleurs résultats que l'estimateur naïf calculé sur l'ensemble des répondants. Moustaki et Knott (2000) proposent un système de repondération pour corriger la *non-réponse partielle* en utilisant des covariables et une ou plusieurs variables latentes. Notre principale contribution par rapport à la littérature existante tient à la construction d'un système de pondération pour traiter la *non-réponse totale et partielle*

qui est basé uniquement sur des variables latentes et peut aussi être utilisé en l'absence de toute autre covariable. Par ailleurs, notre approche diffère de celle de Copas et Farewell (1998), car ils observent leur variable « empressement à répondre » sur les répondants pour quantifier l'intérêt à répondre à l'enquête et un ensemble de covariables, tandis que nous l'inférons à partir des données.

L'article est structuré comme suit. À la section 2, nous présentons le cadre de sondage et la notation. À la section 3, nous illustrons l'estimation des probabilités de réponse. À la section 4, nous décrivons le modèle à traits latents utilisé à cette fin. À la section 5, nous décrivons l'estimateur proposé et l'estimation de sa variance. À la section 6, nous évaluons les propriétés empiriques de l'estimateur proposé au moyen d'études par simulation. Enfin, à la section 7, nous résumons nos conclusions.

2 Cadre de travail

Soit U une population finie de taille N, indicée par k variant de 1 à N. Soit s l'ensemble d'étiquettes d'échantillon, tel que $s \subset U$, tiré de la population en utilisant un plan de sondage probabiliste $p(s)$. La taille de l'échantillon est désignée par n. Soit $\pi_k = \sum_{s;s\ni k} p(s)$ la probabilité d'inclure l'unité k dans l'échantillon. Nous supposons que $\pi_k > 0, k = 1,\ldots,N$. Les unités sélectionnées dans s ne répondent pas toutes à l'enquête. Désignons par $r \subseteq s$ l'ensemble de répondants, et par $\bar{r} = s \setminus r$ l'ensemble de non-répondants. Le mécanisme de réponse est donné par la distribution $q(r|s)$ telle que, pour chaque s fixé, nous avons

$$q(r|s) \geq 0, \text{ pour tout } r \in \mathcal{R}_s \text{ et } \sum_{s\in\mathcal{R}_s} q(r|s) = 1, \text{ où } \mathcal{R}_s = \{r|r \subseteq s\}.$$

En présence de non-réponse totale, nous définissons l'indicateur de réponse $R_k = 1$ si l'unité $k \in r$ et 0 si $k \in \bar{r}$. Donc $r = \{k \in s|R_k = 1\}$. Nous supposons que ces variables aléatoires sont indépendantes l'une de l'autre et du mécanisme de sélection de l'échantillon (Oh et Scheuren 1983). Puisque seules les unités dans r sont observées, une structure de réponse est utilisée pour estimer la probabilité de répondre à l'enquête d'une unité $k \in U$, $p_k = P(k \in r|k \in s) = P(R_k = 1|k \in s)$, qui est une fonction de l'échantillon et qui doit être positive.

Supposons qu'il y ait m variables d'intérêt particulières dans l'enquête. Chaque répondant est exposé à ces m variables du questionnaire, étiquetées $\ell = 1,\ldots,m$. Supposons que l'objectif est d'estimer le total de population de certaines variables d'intérêt et, en particulier, de la variable d'intérêt y_j, c'est-à-dire $Y_j = \sum_{k=1}^{N} y_{kj}$, où y_{kj} est la valeur prise par y_j sur l'unité k. Dans le cas idéal, si la distribution des réponses $q(r|s)$ est connue, alors les p_k seront connues et disponibles pour estimer Y_j en utilisant une approche de repondération. Supposons aussi la présence d'une non-réponse partielle pour la variable y_j. Soit $r_j = \{k \text{ répond à } y_j | k \in r\}$ l'ensemble de répondants pour la variable y_j. Comme dans le cas de la non-réponse totale, nous supposons que les unités dans r_j répondent indépendamment l'une de l'autre. Soit $q_{kj} = P(k \text{ répond à } y_j | k \in r)$. L'ensemble final de poids à utiliser dans une approche entièrement repondérée pour traiter les non-réponses totale et partielle est donné par $1/(\pi_k p_k q_{kj})$, pour

tout $k \in r_j$, en supposant que $q_{kj} > 0$. Ces poids peuvent être utilisés, par exemple, en trois phases dans l'estimateur de Horvitz-Thompson (HT) suivant :

$$\hat{Y}_{j,pq,\text{réelles}} = \sum_{k \in r_j} \frac{y_{kj}}{\pi_k p_k q_{kj}}, \tag{2.1}$$

(voir Legg et Fuller 2009, pour les propriétés des estimateurs sous échantillonnage à trois phases).

Habituellement, p_k et q_{kj} sont inconnues et doivent être estimées. Un estimateur corrigé de la non-réponse est alors construit en remplaçant p_k et q_{kj} par les estimations $\hat{p}_k$ et $\hat{q}_{kj}$ dans (2.1). Les sections qui suivent fournissent des précisions à ce sujet.

3 Estimation des probabilités de réponse

3.1 Utilisation de la régression logistique pour estimer p_k

Différentes méthodes d'estimation de p_k sont proposées dans la littérature. Toutes sont fondées sur l'utilisation d'information auxiliaire connue au niveau de la population ou de l'échantillon. Si la non-réponse est non ignorable, la variable d'intérêt est elle-même la cause (ou l'une des causes) du comportement de réponse, et une covariance entre cette variable et la probabilité de réponse est produite par la voie d'une relation causale directe (voir Groves 2006). Dans un tel cas, la probabilité de réponse p_k pourrait être modélisée pour $k \in s$ par régression logistique comme suit :

$$p_k = P\left(R_k = 1 \middle| y_{kj}\right) = \frac{1}{1 + \exp\left(-\left(a_0 + a_1 y_{kj}\right)\right)}, \tag{3.1}$$

ou comme suit :

$$p_k = P\left(R_k = 1 \middle| y_{kj}, \mathbf{z}_k\right) = \frac{1}{1 + \exp\left(-\left(a_0 + a_1 y_{kj} + \mathbf{z}_k' \boldsymbol{\alpha}\right)\right)}, \tag{3.2}$$

où $\mathbf{z}_k = (z_{k1}, \ldots, z_{kt})'$ est un vecteur de valeurs prises par $t \geq 1$ covariables sur l'unité k, et a_0, a_1, et $\boldsymbol{\alpha}$ sont des paramètres.

Le biais de non-réponse dans le total non ajusté des réponses pour la variable d'intérêt y_j dépend de la covariance entre les valeurs y_{kj} et p_k (voir Bethlehem 1988). Le degré d'intérêt pour le sujet de l'enquête, comme les connaissances, les attitudes et les comportements associés à ce sujet, est un exemple de covariable qui réduit la covariance entre y_{kj} et p_k (voir Groves, Couper, Presser, Singer, Tourangeau, Acosta et Nelson 2006). L'ensemble de covariables $\mathbf{z}_k$ pourrait aussi être relié à la variable d'intérêt y_j pour réduire la variance d'échantillonnage (Little et Vartivarian 2005).

Puisque y_{kj} n'est observée que sur les répondants, les modèles (3.1) et (3.2) ne peuvent généralement pas être estimés. Donc, habituellement, les valeurs de $\mathbf{z}_k$ qui sont connues pour les répondants ainsi que

les non-répondants et qui sont reliées aux valeurs y_{kj} par une « régression que l'on espère forte » (Cassel, Särndal et Wretman 1983) sont utilisées dans le modèle suivant :

$$p_k = P(R_k = 1|\mathbf{z}_k) = \frac{1}{1 + \exp(-(a_0 + \mathbf{z}_k'\boldsymbol{\alpha}))}. \tag{3.3}$$

Alors, le maximum de vraisemblance peut servir à ajuster le modèle (3.3) en utilisant les données $(R_k, \mathbf{z}_k)$ pour $k \in s$. Cela mène à l'estimation de $\hat{a}_0$ et $\hat{\boldsymbol{\alpha}}$, et aux probabilités de réponse estimées $\hat{p}_k = 1/[1 + \exp(-(\hat{a}_0 + \mathbf{z}_k'\hat{\boldsymbol{\alpha}}))]$ à utiliser dans (2.1). Cette procédure offre une certaine protection contre le biais de non-réponse si $\mathbf{z}_k$ est un prédicteur puissant de la probabilité de réponse et/ou de la variable d'intérêt (Kim et Kim 2007).

Dans la suite de l'exposé, nous proposons un système d'ajustement par repondération s'appuyant sur une variable auxiliaire qui mesure la propension de chaque unité à participer à l'enquête. À cette fin, d'autres hypothèses concernant la structure de réponse sont introduites afin de supposer que les p_k dépendent d'une variable auxiliaire latente qui est reliée aux scores de propension de Rosenbaum et Rubin (1983). L'approche proposée peut être utilisée en l'absence d'autre information auxiliaire sur $k \in s$.

3.2 Variables latentes comme information auxiliaire

Afin d'obtenir une mesure des propensions à répondre, nous considérons le cas où une non-réponse partielle sur les variables d'intérêt est également présente. Alors, à l'instar de Chambers et Skinner (2003, page 278) selon lesquels [*Traduction*] « d'un point de vue théorique, la différence entre les non-réponses totale et partielle est superflue, la non-réponse totale étant simplement une forme extrême de non-réponse partielle », nous supposons que la non-réponse partielle sur les variables d'intérêt repose chez les répondants sur les mêmes attitudes et facteurs que ceux qui sous-tendent la non-réponse totale. Des modèles à variables latentes peuvent être utilisés pour estimer ces facteurs qui, par conséquent, peuvent servir de covariables dans un modèle de régression logistique.

Comme nous l'avons déjà mentionné, nous supposons que la non-réponse partielle touche les m variables d'intérêt particulier de l'enquête. Un deuxième indicateur de réponse est introduit pour chaque item ℓ. Pour chaque item ℓ et chaque unité k, nous définissons une variable binaire $x_{k\ell}$ qui prend la valeur 1 si l'unité k répond à l'item ℓ et 0 autrement. Soit $\mathbf{x}_k = (x_{k1}, \ldots, x_{k\ell}, \ldots, x_{km})'$ le vecteur des indicateurs de réponse de l'unité k aux m items et soit $\mathbf{y}_k = (y_{k1}, \ldots, y_{k\ell}, \ldots, y_{km})'$ le vecteur des variables étudiées pour l'unité k. Donc $y_{k\ell}$ est la valeur de la réponse de l'unité k à l'item ℓ et $x_{k\ell}$ est l'indicateur de sa réponse.

Supposons que les $x_{k\ell}$ sont reliés à une échelle continue latente sous-jacente présumée; ils sont les indicateurs d'une variable latente désignée par θ_k. De Menezes et Bartholomew (1996) appellent la variable θ_k la « tendance à répondre » à l'enquête. Nous l'appelons ici la « volonté de répondre à l'enquête » de l'unité k. Un modèle à traits latents contenant une seule variable latente est utilisé pour calculer θ_k pour chaque $k \in s$ (nous verrons plus tard de quelle façon; voir la section 4.4). Supposons que le moment que θ_k est connu pour toutes les unités de l'échantillon et, comme dans le cas de

l'information auxiliaire habituelle, qu'il peut être utilisé comme une covariable. En l'absence d'autres covariables, le modèle (3.3) se réécrit sous la forme :

$$p_k = P(R_k = 1|\theta_k) = \frac{1}{1+\exp(-(\alpha_0 + \alpha_1\theta_k))}. \quad (3.4)$$

La covariable θ_k peut être considérée comme une variable qui explique le comportement associé au sujet de l'enquête, et qui possède donc de bonnes propriétés pour réduire la covariance entre y_{kj} et p_k et, par conséquent, le biais de non-réponse. Si d'autres données auxiliaires appropriées sont disponibles, elles peuvent être insérées dans le modèle à titre de covariables supplémentaires. Maintenant, pour estimer le paramètre du modèle (3.4), la valeur de θ_k doit être disponible pour toutes les unités comprises dans l'échantillon. Les sections qui suivent expliquent en détail comment obtenir les valeurs estimées de θ_k pour les répondants ainsi que les non-répondants.

4 Calcul des propensions à répondre en utilisant des modèles à traits latents

La variable θ_k peut être calculée en utilisant un modèle à traits latents. En général, les modèles à variables latentes sont des modèles de régression multivariés qui relient des réponses continues ou catégoriques à des covariables inobservées. Un modèle à traits latents est essentiellement un modèle d'analyse factorielle pour données binaires (voir Bartholomew, Steele, Moustaki et Galbraith 2002; Skrondal et Rabe-Hesketh 2007).

Nous commençons par créer la matrice contenant les éléments $\{x_{k\ell}\}_{k\in s;\ell=1,\ldots,m}$. La figure 4.1 donne un schéma des indicateurs $x_{k\ell}$ pour les répondants et les non-répondants. Ensuite, nous supposons que les facteurs qui sous-tendent la réponse totale sont les mêmes que ceux qui sous-tendent la réponse partielle sur des variables d'intérêt choisies. Autrement dit, la non-réponse partielle est supposée non ignorable.

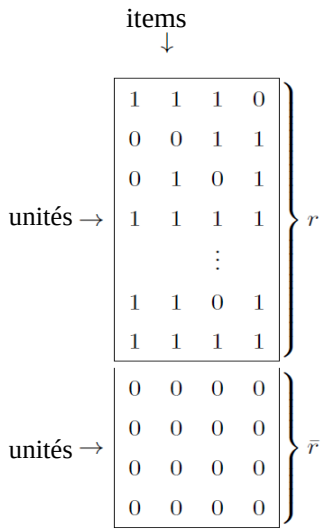


Figure 4.1 Schéma représentant les variables $x_{k\ell}$ pour les ensembles r et $\bar{r}$

Soit $q_{k\ell}$ la probabilité de réponse de l'unité k pour l'item ℓ, pour tout $\ell = 1,\ldots,m$ et $k \in r$. Comme dans le cas de la non-réponse totale, $q_{k\ell}$ est modélisée sous forme d'une fonction de la variable d'intérêt en utilisant la régression logistique comme suit :

$$q_{k\ell} = P(x_{k\ell} = 1|y_{k\ell}, \theta_k, R_k = 1) = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_{\ell 1}\theta_k + \beta_{\ell 2} y_{k\ell}))}, \quad (4.1)$$

pour $\ell = 1,\ldots,m$, et $k \in r$, où $\beta_{\ell 0}$, $\beta_{\ell 1}$ et $\beta_{\ell 2}$ sont des paramètres. Puisque $y_{k\ell}$ est connue uniquement pour les unités pour lesquelles $x_{k\ell} = 1$, $k \in r$, le modèle (4.1) ne peut pas être estimé. Comme dans le cas de la non-réponse totale, nous proposons d'estimer $q_{k\ell}$ comme une fonction d'une variable auxiliaire reliée à la variable d'intérêt, c'est-à-dire θ_k. Le modèle (4.1) se réécrit :

$$q_{k\ell} = P(x_{k\ell} = 1|\theta_k, R_k = 1) = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_{\ell 1}\theta_k))}, \quad (4.2)$$

pour $\ell = 1,\ldots,m$, et $k \in r$. Le modèle (4.2) n'est pas un modèle de régression logistique ordinaire, parce que les θ_k sont des valeurs inobservables prises par une variable latente. Les modèles à traits latents peuvent être utilisés dans ce cas pour estimer $q_{k\ell}$, θ_k et les paramètres du modèle. Notons que, dans le domaine des tests de connaissances et de la psychométrie, la modélisation à traits latents est appelée théorie des réponses aux items.

Le modèle de Rasch (Rasch 1960) est un premier modèle à traits latents simple, souvent mentionné dans la littérature psychométrique et utilisé pour analyser les données provenant d'évaluations pour mesurer des variables telles que les compétences et les attitudes. Il prend la forme suivante :

$$q_{k\ell} = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_1\theta_k))} \text{ pour } \ell = 1,\ldots,m \text{ et } k \in r. \quad (4.3)$$

Les paramètres $\beta_{\ell 0}$ sont estimés pour chaque item ℓ et reflètent le caractère extrême (la facilité) de l'item ℓ : la probabilité d'une réponse positive en tous les points de l'espace latent est d'autant plus grande que les valeurs sont grandes. Le paramètre β_1 est appelé paramètre de « discrimination » et peut être fixé à une valeur arbitraire sans incidence sur la vraisemblance, à condition de permettre que l'échelle des propensions des individus soit libre. Dans de nombreuses situations, l'hypothèse voulant que les discriminations des items soient constantes sur l'ensemble des items est trop contraignante. Le modèle logistique à deux paramètres (2PL) généralise le modèle de Rasch en permettant que les pentes varient. En particulier, le modèle 2PL suppose la forme donnée par l'équation (4.2). Les paramètres $\beta_{\ell 1}$ sont maintenant estimés pour chaque item ℓ et donnent une mesure de la quantité d'information qu'un item fournit au sujet de la variable latente θ_k. Pour arriver à l'identifiabilité du modèle (4.2), nous pouvons fixer la valeur d'un ou de plusieurs paramètres $\beta_{\ell 0}$ et $\beta_{\ell 1}$ dans le processus d'estimation. Moran (1986) a montré que, dans le modèle 2PL, tous les paramètres sont identifiables sous des conditions très générales, à condition que le nombre d'items soit supérieur à deux, et que toutes les pentes soient supposées être strictement positives. On trouve dans la littérature une généralisation supplémentaire du modèle (4.2) – le modèle 3PL – qui contient un autre paramètre, le paramètre de *pseudo-chance*, pour modéliser la

probabilité qu'un sujet pour lequel une variable latente tend vers $-\infty$ réponde à un item. Une telle extension ne paraît pas nécessaire dans le présent contexte et ne sera plus examinée.

4.1 Hypothèses dans les modèles à traits latents

Les modèles à traits latents s'appuient habituellement sur les hypothèses suivantes. La première est celle qu'il est convenu d'appeler hypothèse d'*indépendance conditionnelle*, qui postule que les réponses aux items sont indépendantes sachant la variable latente (c'est-à-dire que la variable latente rend compte de toutes les associations entre les variables observées $x_{k\ell}$). Conséquemment, sachant θ_k, la probabilité conditionnelle de $\mathbf{x}_k$ est

$$P(\mathbf{x}_k|\theta_k) = \prod_{\ell=1}^{m} P(x_{k\ell}|\theta_k).$$

Selon Bartholomew et coll. (2002, page 181) [*Traduction*] « l'hypothèse d'indépendance conditionnelle ne peut être testée qu'indirectement en vérifiant si le modèle est adéquat pour les données. Un modèle à variables latentes est considéré comme étant bien ajusté si les variables latentes expliquent la plupart de l'association entre les réponses observées ».

Une deuxième hypothèse des modèles (4.2) et (4.3) est celle de *monotonicité* : à mesure que la valeur de la variable latente θ_k augmente, la probabilité de réponse à un item augmente ou reste la même sur les intervalles de θ_k. Autrement dit, pour deux valeurs de θ_k, disons a et b, et en supposant arbitrairement que $a < b$, la monotonicité implique que $P(x_{k\ell} = 1|\theta_k = a) < P(x_{k\ell} = 1|\theta_k = b)$ pour $\ell = 1,\ldots,m$. La chance d'une réponse à chaque item est d'autant plus grande que les valeurs de θ_k sont grandes.

Enfin, la troisième hypothèse, et peut-être la plus forte, des modèles (4.2) et (4.3) est celle d'*unidimensionnalité*, impliquant qu'une variable latente unique explique complètement la volonté de l'unité k de répondre au questionnaire. Toutes ces hypothèses fondamentales impliquent que la dépendance entre les items $x_{k\ell}$ peut être expliquée par la variable latente θ_k qui représente la volonté de répondre et que la probabilité qu'une unité k réponde à une variable donnée augmente avec θ_k.

4.2 Estimation du modèle

Nous allons maintenant nous concentrer sur le modèle logistique à deux paramètres (2PL) donné en (4.2). Soit $\boldsymbol{\beta}_\ell = (\beta_{\ell 0}, \beta_{\ell 1})'$ et $\boldsymbol{\beta} = \{\boldsymbol{\beta}_\ell, \ell = 1,\ldots,m\}$. Le modèle (4.2) peut être ajusté en utilisant la méthode du maximum de vraisemblance ou une méthode bayésienne. Nous nous penchons ici sur la première. Sous l'approche du maximum de vraisemblance sont développées trois grandes méthodes, celles du maximum de vraisemblance jointe, de vraisemblance conditionnelle et de vraisemblance marginale. Ici, nous nous concentrons sur le maximum de vraisemblance marginale qui peut être appliqué pour ajuster le modèle 2PL. Cette méthode est également utilisée dans les études par simulation de la section 6. Elle consiste à maximiser la vraisemblance du modèle après avoir éliminé par intégration les θ_k en faisant l'hypothèse d'une loi commune sur ces paramètres. En particulier, on suppose que θ_k est une variable aléatoire qui suit une loi de densité de probabilité $h(\cdot)$; habituellement $\theta_k \sim N(0,1)$. On suppose aussi

que les vecteurs de réponses $\mathbf{x}_k$ sont indépendants les uns des autres et que l'hypothèse d'indépendance conditionnelle est vérifiée.

Pour un ensemble de n_r répondants ayant les vecteurs de réponses $\mathbf{x}_k, k = 1, \ldots, n_r$, la vraisemblance marginale peut être exprimée sous la forme

$$L\left(\boldsymbol{\beta}; \mathbf{x}_1, \ldots, \mathbf{x}_{n_r}\right) = \prod_{k=1}^{n_r} f\left(\mathbf{x}_k | \boldsymbol{\beta}\right),$$

où $f\left(\mathbf{x}_k | \boldsymbol{\beta}\right) = \int_{-\infty}^{\infty} g\left(\mathbf{x}_k | \theta_k, \boldsymbol{\beta}\right) h\left(\theta_k\right) d\theta_k$,

$$g\left(\mathbf{x}_k | \theta_k, \boldsymbol{\beta}\right) = \prod_{\ell=1}^{m} q_{k\ell}^{x_{k\ell}} \left(1 - q_{k\ell}\right)^{1-x_{k\ell}} = \prod_{\ell=1}^{m} \frac{\exp\left(x_{k\ell}\left(\beta_{\ell 0} + \beta_{\ell 1}\theta_k\right)\right)}{1 + \exp\left(\beta_{\ell 0} + \beta_{\ell 1}\theta_k\right)},$$

et h désigne maintenant la densité de la loi $N(0,1)$. La méthode consiste à maximiser la log-vraisemblance correspondante, donnée par

$$\log L\left(\boldsymbol{\beta}; \mathbf{x}_1, \ldots, \mathbf{x}_{n_r}\right) = \sum_{k=1}^{n_r} \log\left(f\left(\mathbf{x}_k | \boldsymbol{\beta}\right)\right),$$

par rapport à $\boldsymbol{\beta}$ en utilisant, par exemple, l'algorithme EM. Les estimations de $\beta_{\ell 0}$ et $\beta_{\ell 1}, \ell = 1, \ldots, m$ sont donc fournies. Ensuite, θ_k est estimé en utilisant la méthode de Bayes empirique en maximisant la densité a posteriori

$$h\left(\theta_k | \mathbf{x}_k\right) = \frac{g\left(\mathbf{x}_k | \theta_k, \boldsymbol{\beta}\right) h\left(\theta_k\right)}{g\left(\mathbf{x}_k\right)} \propto g\left(\mathbf{x}_k | \theta_k, \boldsymbol{\beta}\right) h\left(\theta_k\right),$$

par rapport à θ_k et en maintenant les paramètres d'item et les observations fixes. Les estimations de $q_{k\ell}$ sont obtenues en utilisant l'expression (4.2), où $\beta_{\ell 0}, \beta_{\ell 1}$ et θ_k sont remplacés par leurs estimations.

4.3 Mesures de l'adéquation du modèle

Différentes mesures d'adéquation sont proposées dans la littérature pour tester si le modèle donné en (4.2) est ajusté adéquatement aux données (voir, par exemple, Bartholomew et coll. 2002). On utilise les valeurs de marge de tableaux des réponses à double ou à triple entrée. Les écarts entre les fréquences espérées (E) et observées (O) dans ces tableaux sont mesurés en utilisant la statistique $R = (O - E)^2 / E$. Les grandes valeurs de R pour les marges d'ordre deux ou d'ordre trois détermineront des ensembles d'items pour lesquels le modèle n'est pas bien ajusté. Notons que les résidus $(O - E)^2 / E$ ne sont pas indépendants et qu'ils ne peuvent pas être totalisés pour donner une statistique de test globale qui suit une loi du khi-carré (voir Bartholomew et coll. 2002, page 186). Des indices d'adéquation des items (*item fit indexes*) (Bond et Fox 2007) peuvent être utilisés à cette fin également. En se basant sur les variables latentes et les paramètres d'item estimés, on peut calculer la réponse espérée d'une unité à un item. La similarité entre les réponses observée et espérée à un item peut être évaluée au moyen de deux statistiques d'adéquation basées sur la moyenne des carrés : la statistique d'adéquation sensible aux

valeurs aberrantes (*item outfit*) et la statistique d'adéquation pondérée par l'information (*item infit*). L'estimation produite pour l'*item outfit* est relativement plus affectée par les réponses inattendues à des items qui s'écartent du niveau mesuré d'une personne, c'est-à-dire qu'elle est surtout sensible aux réponses inattendues données par des unités à des questions auxquelles il devrait leur être relativement très facile ou très difficile de répondre. L'*item infit*, pour laquelle chaque observation est pondérée par l'information est, à l'opposé, relativement plus affectée par les réponses inattendues à des items proches du niveau mesuré de la personne, c'est-à-dire que la statistique est plus sensible à des structures inattendues de réponses données par des unités à des items qui sont approximativement ciblés sur elles en fonction de la valeur de leur variable latente. La valeur espérée pour les deux statistiques est un. Les valeurs de l'*infit* et de l'*outfit* supérieures/inférieures à un indiquent une plus grande/faible variation entre les structures de réponses observées et prédites, et un intervalle de 0,5 à 1,5 est généralement acceptable (Bond et Fox 2007).

En outre, des corrélations point-mesure (Olsson, Drasgow et Dorans 1982) peuvent être utilisées pour estimer la corrélation entre la variable latente et la réponse à un item unique. Les items pour lesquels ces mesures prennent une valeur négative ou nulle doivent être supprimés de l'analyse ou peuvent être la preuve que le concept latent n'est pas unidimensionnel. L'unidimensionnalité peut être testée en exécutant une analyse en composantes principales (ACP) des résidus standardisés pour les items (Wright 1996). De cette façon, la première composante (dimension) a déjà été éliminée, et il est possible d'examiner des dimensions, composantes ou contrastes secondaires. L'unidimensionnalité est confirmée par l'observation que la valeur propre de la première composante de l'ACP dans la matrice de corrélation des résidus est faible (habituellement inférieure à 2,0). Sinon, les poids sur le premier contraste indiquent qu'il existe des configurations contrastées dans les résidus.

Enfin, lorsque les items sont utilisés pour former une échelle, ils doivent posséder une cohérence interne. Le coefficient alpha de Cronbach peut être utilisé pour tester si les items ont la propriété de fiabilité, c'est-à-dire que s'ils mesurent tous la même chose, ils devraient être corrélés les uns aux autres.

4.4 Estimation de p_k

Deux solutions sont présentées ici pour estimer p_k en utilisant l'information provenant du modèle à traits latents. La première solution utilise la régression logistique pour estimer p_k pour tout $k \in s$, et une approche en deux étapes.

Étape 1 : Premièrement, nous fournissons une estimation $\hat{\theta}_k$ de θ_k. Pour calculer une valeur $\hat{\theta}_k$ pour $k \in \bar{r}$, nous supposons de nouveau que la non-réponse totale est simplement une forme extrême de non-réponse partielle. Donc, un non-répondant ne répond à aucun item ℓ et par conséquent $x_{k\ell} = 0$, pour tout $\ell = 1, \ldots, m$. Le calcul de $\hat{\theta}_k$ pour $k \in \bar{r}$ est traité comme suit : nous ajoutons à l'ensemble r une unité répondante fantôme $\tilde{k}$ pour laquelle $x_{\tilde{k}\ell}$ est égal à 0, pour tout $\ell = 1, \ldots, m$. Nous désignons ce nouvel ensemble par $\tilde{r} = r \cup \{\tilde{k}\}$. Nous estimons les paramètres du modèle (4.2) en utilisant toutes les unités $k \in \tilde{r}$, et nous calculons les valeurs $\hat{\theta}_k, k \in \tilde{r}$. Le modèle (4.2) permet le calcul de $\hat{\theta}_k$ pour tout $k \in \tilde{r}$. L'unité $\tilde{k}$ a une valeur estimée $\hat{\theta}_{\tilde{k}}$. Nous affectons à toutes les unités $k \in \bar{r}$ une estimation $\hat{\theta}_k$ égale à $\hat{\theta}_{\tilde{k}}$. Donc, la même valeur de $\hat{\theta}_k$ est fournie pour tout $k \in \bar{r}$. En utilisant cette méthode, chaque

unité $k \in s$ est associée à une estimation $\hat{\theta}_k$. Il s'agit de la caractéristique clé pour l'estimation des probabilités de réponse p_k donnée à l'étape suivante.

Étape 2 : Nous utilisons l'estimation $\hat{\theta}_k$, pour $k \in s$, fournie à la première étape comme une covariable dans le modèle (3.4) au lieu de la valeur inconnue de θ_k; en particulier

$$p_k = P\left(R_k = 1 \middle| \hat{\theta}_k\right) = \frac{1}{1 + \exp\left(-\left(\alpha_0 + \alpha_1 \hat{\theta}_k\right)\right)}, \text{ pour tout } k \in s. \tag{4.4}$$

Le modèle (4.4) donne les estimations $\hat{p}_k$ de p_k, pour tout $k \in s$.

L'un des arbitres a suggéré la solution suivante pour estimer p_k. Soit $S_k = \sum_{\ell=1}^{m} x_{k\ell}$, le score brut pour l'unité k, c'est-à-dire le nombre d'items auxquels l'unité k a répondu : si $k \in \bar{r}$, alors $S_k = 0$; si $k \in r$, alors $S_k > 0$. Ensuite, nous pouvons estimer p_k en modélisant $P(S_k > 0|\theta_k)$. En vertu de l'hypothèse d'indépendance conditionnelle, nous avons

$$\begin{aligned} p_k &= P(S_k > 0|\theta_k) = 1 - P(S_k = 0|\theta_k) = 1 - P\left(\bigcap_{\ell=1}^{m} (x_{k\ell} = 0|\theta_k)\right) \\ &= 1 - \prod_{\ell=1}^{m} (1 - P(x_{k\ell} = 1|\theta_k)). \end{aligned}$$

Nous avons $P(x_{k\ell} = 1|\theta_k) = P(R_k = 1|\theta_k) P(x_{k\ell} = 1|\theta_k, R_k = 1) + P(R_k = 0|\theta_k) P(x_{k\ell} = 1|\theta_k, R_k = 0) = p_k q_{k\ell}$, parce que $P(x_{k\ell} = 1|\theta_k, R_k = 0) = 0$. Par conséquent, nous obtenons

$$p_k = 1 - \prod_{\ell=1}^{m} (1 - p_k q_{k\ell}), k \in r.$$

La probabilité de réponse estimée $\hat{p}_k$, $k \in r$ s'obtient comme une solution de l'équation polynomiale

$$\hat{p}_k = 1 - \prod_{\ell=1}^{m} (1 - \hat{p}_k \hat{q}_{k\ell}).$$

Cette solution, quoique très élégante, a deux inconvénients. Si m est grand, l'équation polynomiale susmentionnée est difficile, voire impossible, à résoudre. S'il est possible de la résoudre pour une valeur modérée de m, les solutions réelles ne se trouvent pas nécessairement dans (0, 1). Cette solution n'a pas été examinée plus en détail ici.

5 L'estimateur proposé et l'estimation de sa variance

Rappelons que nous avons une variable d'intérêt particulier y_j et qu'il existe une non-réponse partielle pour cette variable. Si nous souhaitons estimer le total de population Y_j de y_j, un estimateur naïf ne comprenant de correction ni pour la non-réponse totale ni pour la non-réponse partielle est donné par

$$\hat{Y}_{j,\text{naïf}} = N \sum_{k \in r_j} \frac{y_{kj}}{\pi_k} \Big/ \sum_{k \in r_j} \frac{1}{\pi_k}. \tag{5.1}$$

La repondération des répondants aux items est aussi une approche pour traiter la non-réponse partielle. Moustaki et Knott (2000) proposent de pondérer les répondants aux items par l'inverse de la probabilité prédite de réponse à l'item $\hat{q}_{k\ell}$, en supposant que $\hat{q}_{k\ell} > 0$. Par conséquent, un poids d'ajustement possible pour les non-réponses partielle et totale associées à l'unité $k \in r_j$ est donné par $1/\left(\hat{p}_k \hat{q}_{kj}\right)$. Nous proposons d'utiliser l'estimateur sous échantillonnage à trois phases ajusté pour les non-réponses partielle et totale par repondération donné par

$$\hat{Y}_{j,pq} = \sum_{k \in r_j} \frac{y_{kj}}{\pi_k \hat{p}_k \hat{q}_{kj}}, \tag{5.2}$$

où $\hat{p}_k$ est fourni par le modèle (4.4), et $\hat{q}_{kj}$, par le modèle (4.2). Des propositions faisant appel à l'imputation des valeurs de y_{kj} pour $k \in r \setminus r_j$ pour traiter la non-réponse partielle sont également prises en considération, mais ne sont pas présentées faute d'espace. Elles peuvent être obtenues sur demande auprès des auteurs.

Les propriétés de l'estimateur proposé (5.2) dépendent des hypothèses faites au sujet des mécanismes de non-réponse totale ainsi que partielle. En particulier, l'estimateur (5.2) suppose une deuxième phase d'échantillonnage avec probabilités de réponse inconnues. Si nous ignorons l'estimation de θ_k dans le modèle (4.4), les résultats présentés dans Kim et Kim (2007) concernant la convergence de l'estimateur sous un plan échantillonnage à deux phases utilisant les probabilités de réponse estimées sont vérifiés ici, si l'on considère les estimations du maximum de vraisemblance pour les paramètres α_0 et α_1. En ignorant l'estimation de la variable latente θ_k et en utilisant les estimations du maximum de vraisemblance marginale pour les paramètres $\beta_{\ell 0}$ et $\beta_{\ell 1}$ dans le modèle (4.2), l'estimateur $\hat{Y}_{j,pq}$ sera convergent si les modèles pour les probabilités de non-réponse totale et partielle sont spécifiés correctement.

Nous pouvons considérer des méthodes de rééchantillonnage pour l'estimation de la variance de l'estimateur proposé et combiner les propositions pour l'échantillonnage à deux phases (Kim, Navarro et Fuller 2006) et pour le calage généralisé en présence de non-réponse (Kott 2006). En particulier, l'estimateur de variance par rééchantillonnage peut s'écrire comme

$$\hat{V}_r = \sum_{l=1}^{L} c_l \left(\hat{Y}_{j,pq}^{(l)} - \hat{Y}_{j,pq}\right)^2,$$

où $\hat{Y}_{j,pq}^{(l)}$ est la l^{e} version de $\hat{Y}_{j,pq}$ basée sur les observations incluses dans la l^{e} réplique, L est le nombre de répliques, c_l est un facteur associé à la réplique l déterminé par la méthode de rééchantillonnage. La l^{e} réplique de $\hat{Y}_{j,pq}$ peut s'écrire sous la forme $\hat{Y}_{j,pq}^{(l)} = \sum_{k \in r_j} w_{3k}^{(l)} y_{kj}$, où $w_{3k}^{(l)}$ désigne le poids de rééchantillonnage de la k^{e} unité dans la l^{e} réplique. Ces poids de rééchantillonnage sont calculés en utilisant une procédure en deux étapes.

Premièrement, notons que, si nous ignorons pour le moment la présence de la non-réponse partielle, l'estimateur sous échantillonnage à deux phases $\hat{Y}_{j,p} = \sum_{k\in r} w_{2k} y_{kj}$, a pour poids

$$w_{2k} = 1/(\pi_k p_k) = w_{1k} F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right),$$

avec $w_{1k} = 1/\pi_k$, $F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right) = 1 + \exp\left(-\left(\alpha_0 + \alpha_1 \hat{\theta}_k\right)\right)$ (voir l'équation (4.4)). Soit $\hat{\mathbf{z}}_1 = \sum_{k\in s} w_{1k}\mathbf{z}_{1k}$ l'estimation de première phase du total de la variable $\mathbf{z}_1$ définie comme $\mathbf{z}_{1k} = \pi_k p_k \left(1, \hat{\theta}_k\right)'$. Alors, les paramètres α_0 et α_1 sont tels que

$$\sum_{k\in r} w_{1k} F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right)\mathbf{z}_{1k} = \hat{\mathbf{z}}_1. \tag{5.3}$$

Cette procédure équivaut à obtenir des estimations non pondérées du maximum de vraisemblance, mais il est commode de la configurer comme un problème de calage généralisé non linéaire. De cette façon, il est possible d'utiliser l'approche décrite dans Kott (2006), combinée à celle décrite dans Kim et coll. (2006), pour obtenir les poids de rééchantillonnage en utilisant les étapes suivantes.

Étape 1 : Calculer l'estimation de première phase du total de $\mathbf{z}_{1k}$ en supprimant la l^{e} observation, c'est-à-dire $\hat{\mathbf{z}}_1^{(l)} = \sum_{k\in s} w_{1k}^{(l)}\mathbf{z}_{1k}$, où $w_{1k}^{(l)}$ est le poids de rééchantillonnage jackknife classique pour l'unité k dans la réplique l. Calculer les poids jackknife pour l'échantillonnage de deuxième phase en utilisant $\hat{\mathbf{z}}_1^{(l)}$ comme valeur étalon. En particulier, les $w_{2k}^{(l)}$ sont choisis comme étant $w_{2k}^{(l)} = w_{2k} w_{1k}^{(l)} F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right)/w_{1k}$ avec α_0 et α_1 tels que

$$\sum_{k\in r} w_{2k}^{(l)}\mathbf{z}_{1k} = \hat{\mathbf{z}}_1^{(l)}.$$

Cette procédure fournit des poids qui sont très similaires à ceux considérés dans Kott (2006) et peuvent être calculés en se servant des logiciels existants qui prennent en charge le calage généralisé.

La non-réponse partielle est traitée de manière similaire en considérant que $w_{3k} = 1/\left(\pi_k p_k q_{kj}\right) = w_{2k} F\left(\hat{\theta}_k; \beta_{j0}, \beta_{j1}\right)$ (comparé à l'équation (4.3)). Ici, une approximation importante consiste à supposer que, sachant $\hat{\theta}_k$, les paramètres β_{j0} et β_{j1} sont estimés en utilisant un modèle logistique classique (au lieu d'un modèle 2PL) et sont tels que

$$\sum_{k\in r_j} w_{2k} F\left(\hat{\theta}_k; \beta_{j0}, \beta_{j1}\right)\mathbf{z}_{2k} = \hat{\mathbf{z}}_2,$$

où $\hat{\mathbf{z}}_2 = \sum_{k\in r} w_{2k}\mathbf{z}_{2k}$ et $\mathbf{z}_{2k} = \pi_k p_k q_{kj}\left(1, \hat{\theta}_k\right)^T$. Un autre inconvénient est que les variables auxiliaires $\mathbf{z}_{2k}$ dépendent de j et, donc, que des ensembles de poids différents doivent être produits pour les diverses variables d'intérêt.

Étape 2 : Les poids jackknife de troisième phase sont obtenus en calculant d'abord l'estimation de deuxième phase du total de $\mathbf{z}_{2k}$ avec suppression de l'unité l en utilisant les poids provenant de l'étape 1, $\hat{\mathbf{z}}_2^{(l)} = \sum_{k \in r} w_{2k}^{(l)} \mathbf{z}_{2k}$. Alors, en utilisant $\hat{\mathbf{z}}_2^{(l)}$ comme valeur étalon, les $w_{3k}^{(l)}$ sont choisis comme étant $w_{3k}^{(l)} = w_{3k} w_{2k}^{(l)} F\left(\hat{\theta}_k ; \beta_{j0}, \beta_{j1}\right) \big/ w_{2k}$ avec β_{j0} et β_{j1} calculés au moyen de

$$\sum_{k \in r_j} w_{3k}^{(l)} \mathbf{z}_{2k} = \hat{\mathbf{z}}_2^{(l)}.$$

6 Études par simulation

Nous évaluons la performance de l'estimateur présenté à la section 5 au moyen d'une simulation Monte Carlo sous deux scénarios différents. Le premier utilise un ensemble de données réelles comme population et des variables d'intérêt qui sont toutes binaires, tandis que le second utilise des données de population simulées avec des variables d'intérêt continues. Les résultats pour le premier scénario sont présentés à la section 6.1, tandis que ceux pour le deuxième sont présentés à la section 6.2.

Sous les deux scénarios, nous utilisons l'échantillonnage aléatoire simple sans remise et considérons les estimateurs suivants :

- HT $= \sum_{k \in s} y_{kj} \big/ \pi_k$: l'estimateur de Horvitz-Thompson dans le cas d'une réponse complète est calculé comme valeur de référence en l'absence de non-réponse.
- $\hat{Y}_{j,\text{naïf}}$: l'estimateur naïf donné en (5.1); aucune mesure explicite n'est prise pour corriger les non-réponses totale et partielle. Notons que, sous échantillonnage aléatoire simple sans remise, il se réduit à $\hat{Y}_{j,\text{naïf}} = N \sum_{k \in r_j} y_{kj} \big/ n_{r_j}$, où n_{r_j} est la taille de l'ensemble r_j, et il est identique à l'estimateur de Horvitz-Thompson ajusté pour la non-réponse totale sous l'hypothèse de probabilités de réponse uniformes estimées par $n_{r_j} \big/ n$.
- $\hat{Y}_{j,pq}$: l'estimateur sous échantillonnage à trois phases proposé à la section 5, équation (5.2).
- $\hat{Y}_{j,pq,\text{réelles}}$: l'estimateur sous échantillonnage à trois phases qui utilise les valeurs réelles des probabilités de réponse p_k et q_{kj} est également calculé aux fins de comparaison avec $\hat{Y}_{j,pq}$ pour comprendre l'effet de l'estimation des probabilités de réponse.

Les simulations sont exécutées en R version 2.15, en utilisant le module R « `ltm` » (Rizopoulos 2006) pour ajuster les modèles à traits latents. Les mesures de performance suivantes sont calculées pour chaque estimateur, ci-après désigné génériquement par $\hat{Y}$ où le suffixe j est abandonné pour simplifier la notation (Y désigne le total de population) :

- le biais Monte Carlo

$$\mathrm{B} = E_{\mathrm{sim}}\left(\hat{Y}\right) - Y,$$

où $E_{\mathrm{sim}}\left(\hat{Y}\right) = \sum_{i=1}^{M} \hat{Y}_i \big/ M$, $\hat{Y}_i$ est la valeur de l'estimateur $\hat{Y}$ à la i^{e} exécution de la simulation et M est le nombre total d'exécutions de la simulation;

- le biais relatif

$$\mathrm{BR} = \frac{\mathrm{B}}{Y};$$

- l'écart-type Monte Carlo

$$\sqrt{\mathrm{VAR}} = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{Y}_i - E_{\mathrm{sim}}\left(\hat{Y}\right)\right)^2};$$

- l'erreur quadratique moyenne Monte Carlo

$$\mathrm{EQM} = \mathrm{B}^2 + \mathrm{VAR}.$$

6.1 Scénario de simulation 1

Nous considérons un jeu de données formé de quatre variables binaires extraites de la *British Social Attitudes Survey* de 1986 et qui concernent l'attitude à l'égard de l'avortement. Les données sont disponibles dans le module R « `ltm` » (Rizopoulos 2006). $N = 379$ personnes ont répondu aux questions suivantes après qu'on leur ait demandé si la loi devrait permettre l'avortement dans les circonstances présentées par chaque item :

1. La femme décide toute seule qu'elle ne souhaite pas garder le bébé.
2. Le couple est d'accord qu'il ne souhaite pas avoir un enfant.
3. La femme n'est pas mariée et ne souhaite pas épouser l'homme.
4. Le couple n'a pas les moyens d'avoir un autre enfant.

La variable d'intérêt y_j est choisie comme étant la deuxième $(j = 2)$ avec un total $Y_j = 225$ dans la population.

Les données sont analysées par Bartholomew et coll. (2002) à titre d'exemple de situation où l'on peut trouver une variable latente qui mesure l'attitude à l'égard de l'avortement. Au niveau de la population, nous calculons la variable latente (notée ici θ_k^a) en utilisant le modèle (4.2) sur les données $\{y_{k\ell}\}_{k=1,\ldots,N;\,\ell=1,\ldots,4}$. La corrélation entre les valeurs de $y_{k\ell}$ et θ_k^a est approximativement égale à 0,85, pour $\ell = 1,\ldots,4$. Ensuite, nous avons fixé $\theta_k = \hat{\theta}_k^a$, pour tout $k = 1,\ldots,N$.

Au niveau de la population, les probabilités de réponse totale sont générées en utilisant la structure de réponse suivante :

$$p_k = 1/(1 + \exp(-(0{,}7 + y_{k2} + \theta_k + 0{,}2\varepsilon_k))), \tag{6.1}$$

avec $\varepsilon_k \sim U(0,1)$, pour simuler une non-réponse non ignorable. La moyenne de population de p_k est égale à environ 0,74.

Pour générer les probabilités de réponse partielle au niveau de la population, le modèle utilisé est le suivant :

$$q_{k\ell} = 1/(1 + \exp(-(b_\ell \theta_k + a_\ell + y_{k\ell}))), \quad \text{pour } \ell = 1, \ldots, 4, \tag{6.2}$$

où $b_\ell = 3$, pour $\ell = 1, \ldots, 4$, tandis que a_ℓ prend différentes valeurs en fonction de ℓ; en particulier, $a_1 = 1, a_2 = 0, a_3 = -0{,}5$ et $a_4 = 1$. Le taux nominal de non-réponse partielle pour les quatre items dans la population est de 35 %, 42 %, 47 % et 31 %, respectivement.

Nous tirons $M = 10\,000$ échantillons aléatoires simples sans remise à partir de la population en utilisant deux tailles d'échantillon : $n = 50$ et $n = 100$. Dans chaque échantillon s, les unités sont classées comme étant des répondants conformément à un échantillonnage de Poisson, en utilisant les probabilités p_k calculées comme dans l'équation (6.1) et résultant en l'ensemble r. Alors, sachant r, nous construisons la matrice $\{x_{k\ell}\}_{k \in r; \ell = 1, \ldots, 4}$, où les valeurs de $x_{k\ell}$ sont tirées selon un échantillonnage de Poisson avec probabilités $q_{k\ell}$ définies en (6.2). Dans chaque simulation, le modèle (4.2) et l'ensemble de répondants r sont utilisés pour calculer la variable $\hat{\theta}_k$ pour tout $k \in s$ comme il est décrit à la section 4.4. Le modèle (4.4) est ajusté pour obtenir $\hat{p}_k$. Le taux moyen de non-réponse partielle sur les simulations pour les quatre items est de 26 %, 33 %, 38 % et 23 %, respectivement. L'estimateur de variance jackknife a été calculé comme il est décrit à la section 5 en utilisant la fonction `gencalib()` dans le module R « sampling » (Tillé et Matei 2012) et la distance logistique (Deville, Särndal et Sautory 1993).

Le tableau 6.1 donne les résultats pour $n = 50$ et $n = 100$. Comme prévu, les estimateurs HT et $\hat{Y}_{j,pq,\text{réelles}}$ ont un biais presque nul, tandis que le second présente une EQM relativement plus grande qui est due uniquement à la plus petite taille d'échantillon. L'estimateur naïf donne un biais négatif très important. Cela tient au fait que les unités dont la valeur de y_j est nulle sont moins susceptibles de répondre et que le total est clairement sous-estimé. L'estimateur $\hat{Y}_{j,pq}$ présente un bien plus petit biais que l'estimateur naïf. Notons que la performance de l'estimateur proposé est dictée principalement par le biais absolu, de sorte qu'elle ne diffère pas particulièrement lorsqu'on augmente la taille de l'échantillon, mis à part une diminution de la variance. Si nous comparons $\hat{Y}_{j,pq,\text{réelles}}$ et $\hat{Y}_{j,pq}$, nous notons que $\hat{Y}_{j,pq}$ souffre encore d'un certain biais qui provient de la spécification incorrecte de la structure de réponse (nous ne tenons pas compte des valeurs des variables d'intérêt).

Pour l'estimateur proposé, l'estimateur de variance jackknife a également été testé en examinant la couverture empirique d'un intervalle de confiance à 95 % calculé pour chaque réplique comme $\hat{Y}_{j,pq} \pm 1{,}96\sqrt{\hat{V}_r}$. Pour $n = 50$, la valeur moyenne de $\sqrt{\hat{V}_r}$ sur l'ensemble des simulations était de 54,8, tandis que pour $n = 100$, elle était de 53,3, avec un taux de couverture de l'IC à 95 % de 94,6 % et de 96,3 %, respectivement. L'estimateur par rééchantillonnage surestime l'écart-type Monte Carlo donné pour $\hat{Y}_{j,pq}$ au tableau 6.1 dans les deux cas, mais possède de bons taux de couverture.

Tableau 6.1
Résultats des simulations sous le scénario 1 – Ensemble de données sur l'avortement

Estimateur	B	$\sqrt{\text{VAR}}$	EQM	BR %
$n = 50$				
HT	0,05	24,5	600,5	< 0,1
$\hat{Y}_{j,\text{naïf}}$	-126,5	19,4	16 378,6	-56,2
$\hat{Y}_{j,pq}$	20,6	32,4	1 474,1	9,1
$\hat{Y}_{j,pq,\text{réelles}}$	0,02	35,0	1 225,0	< 0,1
$n = 100$				
HT	-0,06	16,0	255,5	< 0,1
$\hat{Y}_{j,\text{naïf}}$	-126,9	13,5	16 284,1	-56,4
$\hat{Y}_{j,pq}$	17,9	21,9	802,2	8,0
$\hat{Y}_{j,pq,\text{réelles}}$	-0,1	23,7	559,9	< 0,1

Pour étudier la performance du modèle à traits latents au niveau de la population et la corrélation entre la variable d'intérêt et la variable latente estimée, nous avons appliqué la procédure décrite plus haut en utilisant $q_{k\ell}$ définie en (6.2) pour construire la matrice $\{x_{k\ell}\}_{k=1,\ldots,N;\ell=1,\ldots,4}$ pour toutes les unités de la population. Nous avons ajusté le modèle (4.2) au niveau de la population et calculé la variable θ_k pour tout $k = 1,\ldots,N$. Le coefficient alpha de Cronbach prend la valeur de 0,83 ce qui indique une bonne cohérence interne des items. Le coefficient de corrélation entre la variable d'intérêt et la variable latente estimée prend la valeur de 0,76, ce qui indique que l'information auxiliaire latente possède un fort pouvoir de prédiction de y_{k2}, comme il l'a été prôné pour le modèle de Cassel et coll. (1983). L'inspection des marges de tableau à double entrée pour la matrice $\{x_{k\ell}\}$ donne les résidus $(O-E)^2/E$, compris entre 0,03 et 0,23. De même, les marges de tableau à triple entrée pour la matrice $\{x_{k\ell}\}$ donnent des résidus entre 0 et 1,19. Cela indique que nous n'avons aucune raison de rejeter ici le modèle à un facteur latent (4.2) (voir Bartholomew et coll. 2002, page 186).

6.2 Scénario de simulation 2

Nous générons $\{y_{k1},\ldots,y_{k6},\theta_k\}$ pour $k = 1,\ldots, N = 2\,000$ en utilisant une loi normale multivariée de moyenne 1. Le degré de corrélation entre y_ℓ et $y_{\ell'}$ est 0,8, avec $\ell, \ell' = 1,\ldots,6, \ell \neq \ell'$. Nous posons que la variable d'intérêt est y_6 et considérons divers degrés de corrélation entre ses valeurs et celles prises par θ_k, à savoir 0,3, 0,5, 0,8. Les valeurs de θ_k sont ensuite centrées et réduites afin qu'elles soient de moyenne 0 et de variance 1.

Les probabilités de réponse sont obtenues en calculant d'abord

$$p_k^\circ = 1/[1+\exp(-(0{,}5+y_{k1}+\theta_k))], \quad \text{pour } k = 1,\ldots,N, \tag{6.3}$$

puis en les rééchelonnant afin qu'elles prennent des valeurs comprises entre 0,1 et 0,9, avec une moyenne de population approximativement égale à 0,7.

Les probabilités de réponse partielle sont générées en calculant d'abord :

$$q_{k\ell}^{\circ} = 1/(1 + \exp(-(b_\ell \theta_k + a_\ell + y_{k\ell}))), \quad \text{pour } k = 1, \ldots, N \text{ et } \ell = 1, \ldots, 6, \tag{6.4}$$

où $\{a_\ell\}_{\ell=1,\ldots,6} = \{1;\ 0;\ -0{,}5;\ 1;\ 0;\ -0{,}5\}$ et $\{b_\ell\}_{\ell=1,\ldots,6} = \{1;\ 1;\ 1;\ 1{,}5;\ 1{,}5;\ 1{,}5\}$, puis en rééchelonnant les valeurs pour qu'elles soient comprises entre 0,1 et 0,95.

Nous tirons $M = 10\,000$ échantillons par échantillonnage aléatoire simple sans remise de taille $n = 200$. Pour chaque échantillon s, un ensemble de réponses r est créé en réalisant un échantillonnage de Poisson de paramètre p_k défini en (6.3). Chaque élément de la matrice $\{x_{k\ell}\}_{k\in r,\ell=1,\ldots,6}$ est généré en utilisant l'échantillonnage de Poisson de paramètre $q_{k\ell}$ défini en (6.4). Les taux de non-réponse partielle sur l'ensemble des simulations prennent approximativement les valeurs de 18 %, 28 %, 35 %, 19 %, 29 % et 34 %, pour $\ell = 1, \ldots, 6$, respectivement. Pour chaque simulation, le modèle (4.2) est utilisé pour calculer la variable $\hat{\theta}_k$ pour tout $k \in s$. Le modèle (4.4) est alors ajusté pour obtenir $\hat{p}_k$.

Tableau 6.2
Résultats des simulations sous le scénario 2 – Données continues simulées

Estimateur	B	$\sqrt{\text{VAR}}$	EQM	BR %
Coefficient de corrélation de 0,3				
HT	-0,7	131,6	17 331,2	≈ -0,0
$\hat{Y}_{j,\text{naïf}}$	825,6	177,1	713 039,3	41,0
$\hat{Y}_{j,pq}$	-227,4	188,0	87 033,0	-11,3
$\hat{Y}_{j,pq,\text{réelles}}$	48,4	231,8	56 073,2	2,4
Coefficient de corrélation de 0,5				
HT	0,1	135,0	18 220,5	≈ 0,0
$\hat{Y}_{j,\text{naïf}}$	972,6	176,2	977 009,5	50,7
$\hat{Y}_{j,pq}$	-180,0	175,5	63 552,0	-9,4
$\hat{Y}_{j,pq,\text{réelles}}$	74,8	212,7	50 844,0	3,9
Coefficient de corrélation de 0,8				
HT	-0,1	134,1	17 992,0	≈ -0,0
$\hat{Y}_{j,\text{naïf}}$	1 154,6	168,1	1 361 388,1	57,7
$\hat{Y}_{j,pq}$	-184,8	164,4	61 173,0	-9,2
$\hat{Y}_{j,pq,\text{réelles}}$	100,6	196,2	48 597,9	5,0

Le tableau 6.2 donne la performance des estimateurs pour les trois valeurs prises par le coefficient de corrélation nominal entre y_{k1} et θ_k : $0{,}3, 0{,}5$ et $0{,}8$. L'estimateur proposé est toujours capable de réduire le biais comparativement à l'estimateur naïf, même quand la corrélation entre la variable d'intérêt et la variable latente devient plus faible. Le biais relatif prend des valeurs acceptables dans la plupart des cas. Le biais mérite d'être examiné de plus près. Dans tous les cas, l'estimateur naïf surestime fortement le total. Cela n'est pas étonnant, parce que les valeurs de p_k, q_{k6}, θ_k et y_{k6} vont toutes dans la même direction. Par conséquent, dans notre échantillon de répondants, nous sommes plus susceptibles de trouver des valeurs relativement grandes de y_6, ce qui donne lieu à une surestimation pour l'estimateur naïf. Par ailleurs, $\hat{Y}_{j,pq}$ sous-estime le total, parce qu'il est fondé uniquement sur les unités observées de r_j qui ont

des valeurs relativement grandes pour y_6, mais aussi des valeurs relativement grandes pour p_k et q_{k6} et, par conséquent, qui ont à la fin un faible poids.

La matrice des valeurs de population $\{x_{k\ell}\}_{k=1,\ldots,2\,000,\ell=1,\ldots,6}$ est construite de la même façon qu'à la section 6.1 pour valider les hypothèses qui sous-tendent le modèle 2PL. Le coefficient alpha de Cronbach prend approximativement la valeur de 0,5 pour le coefficient de corrélation égal à 0,3, de 0,6 pour le coefficient de corrélation égal à 0,5, et de 0,7 pour le coefficient de corrélation égal à 0,8; l'association par paire entre les six items révèle des valeurs p plus petites que 0,01. L'inspection des marges à double entrée et à triple entrée de la matrice $\{x_{k\ell}\}$ donne des résidus $(O-E)^2/E$ qui prennent tous des valeurs inférieures à 4. Par conséquent, le modèle à un facteur latent peut être accepté et les items semblent tous mesurer le même facteur latent.

7 Discussion et conclusion

Nous avons proposé un système de repondération pour tenir compte de la non-réponse non ignorable en nous basant sur une variable auxiliaire latente. Cette variable est calculée pour chaque unité de l'échantillon au moyen d'un modèle à traits latents en supposant qu'il existe une non-réponse partielle et que les non-réponses partielle et totale cachent la même structure latente. Les probabilités de réponse totale sont alors estimées au moyen d'un modèle logistique qui utilise comme covariable le facteur latent extrait d'après les structures de réponse en utilisant un modèle à traits latents. Le système de repondération proposé est ensuite utilisé dans un estimateur sous échantillonnage à trois phases pour traiter la non-réponse, en même temps qu'une méthode de rééchantillonnage pour estimer son incertitude. L'objectif principal est de réduire le biais de non-réponse dans l'estimation du total de population. L'estimateur proposé donne de bons résultats dans nos études par simulation comparativement à l'estimateur naïf, et le gain d'efficacité est important dans certains cas. Des réductions du biais sont également observées quand la corrélation entre le facteur latent et la variable d'intérêt est modeste.

Par conception, la variable latente estimée $\hat{\theta}_k$ est reliée aux indicateurs de réponse x_{kj} pour la variable d'intérêt y_j; puisque la non-réponse est supposée non ignorable, y_{kj} et x_{kj} sont reliés également. Si la condition qui suit est vérifiée,

$$\rho^2_{y_j,x_j} + \rho^2_{\hat{\theta},x_j} > 1,$$

où les coefficients de corrélation $\rho_{y_j,x_j}, \rho_{\hat{\theta},x_j} > 0$, alors y_j et $\hat{\theta}$ sont positivement corrélés (voir Langford, Schwertman et Owens 2001). Notons que nous avons constaté que le degré minimal de corrélation entre la variable d'intérêt et la variable latente capable de réduire le biais de non-réponse était égal à 0,3 dans le scénario de simulation 2 (section 6.2). Naturellement, la réduction du biais dépend des hypothèses du modèle. Si les indicateurs de réponse ne sont pas de bons prédicteurs du comportement de réponse totale, alors nous sommes en présence d'un modèle mal spécifié et, évidemment, il pourrait ne pas y avoir de réduction du biais et une variance pourrait être introduite dans l'estimation. Néanmoins, des outils diagnostiques provenant de la théorie des réponses aux items peuvent être utilisés pour évaluer l'adéquation du modèle à traits latents employé pour estimer les valeurs de θ_k.

Nous avons considéré le cas où aucune information auxiliaire n'est disponible au niveau de l'échantillon ou de la population pour réduire le biais de non-réponse. Les covariables observées (si elles sont disponibles) et la variable latente peuvent néanmoins être utilisées ensemble dans l'estimation des probabilités de réponse. En outre, les modèles à traits latents peuvent, eux-mêmes, être ajustés en introduisant des covariables. L'introduction de covariables dans ces modèles doit être effectuée en faisant preuve d'une prudence croissante en ce qui concerne la variance.

L'estimateur proposé est un estimateur sous échantillonnage à trois phases utilisant un système de repondération basé sur $\hat{p}_k$ et $\hat{q}_{kj}$. Il est connu que de faibles valeurs de $\hat{p}_k$ et $\hat{q}_{kj}$ peuvent donner lieu à des estimateurs repondérés instables, en raison de grands poids de non-réponse. Pour résoudre ce problème, on utilise souvent en pratique la méthode du score de propension (par exemple, Eltinge et Yansaneh 1997), qui offre un bon remède contre les ajustements extrêmes des pondérations. Afin d'appliquer cette méthode dans notre cadre, les répondants à y_j doivent être groupés en différentes classes données par les quantiles de $1/\left(\hat{p}_k \hat{q}_{kj}\right)$. La dernière étape est le calcul d'un poids pour chaque classe.

Les remarques finales concernent l'hypothèse d'indépendance conditionnelle dans les modèles à traits latents. Dans la littérature sur la non-réponse, il est habituel d'utiliser l'échantillonnage de Poisson pour modéliser le comportement de réponse totale en supposant que les unités dans l'ensemble r sont sélectionnées avec des probabilités de réponse inconnues et que la réponse est indépendante d'une unité à l'autre. L'hypothèse d'indépendance conditionnelle dans les modèles à traits latents est une condition similaire appliquée aux items. Les deux hypothèses sont fortes, parfois mises en doute, et pourtant nécessaires dans le processus d'inférence statistique.

Différentes méthodes ont été élaborées dans la littérature psychométrique pour relâcher l'hypothèse d'indépendance conditionnelle. Nous citons ici l'approche d'*indépendance partielle* proposée par Reardon et Raudenbush (2006), élaborée pour le cas où les réponses à des questions antérieures déterminent si les questions ultérieures seront posées ou non, et où l'hypothèse d'indépendance conditionnelle usuelle des modèles classiques échoue. Cette approche pourrait être utilisée dans notre cadre pour le cas où $q_{k\ell}$ est définie comme $P\left(x_{k\ell} = 1 \middle| x_{kj}, \text{pour un certain } j \in \{1,\ldots,m\}, \ell \neq j, \theta_k\right)$ au lieu de $P\left(x_{k\ell} = 1 | \theta_k\right)$, $k \in r$. Une autre approche utile pour les cas où les items sont groupés est celle de la modélisation hiérarchique à traits latents. Un effet aléatoire est introduit dans un modèle à traits latents pour tenir compte de la dépendance possible des résidus due aux sources communes de variation partagées par les groupes d'items (voir, par exemple, Scott et Ip 2002). D'autres travaux de recherche devraient être effectués pour adapter ces approches dans le cadre d'enquêtes.

Remerciements

Les travaux de M. Giovanna Ranalli ont été en partie réalisés grâce au soutien du projet PRIN-SURWEY (subvention 2012F42NS8, Italie).

Bibliographie

Bartholomew, D.J., Steele, F., Moustaki, I. et Galbraith, J.I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Chapman and Hall/CRC.

Beaumont, J.-F. (2000). Une méthode d'estimation en présence de non-réponse non-ignorable. *Techniques d'enquête,* 26, 2, 145-151.

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics,* 4, 3, 251-260.

Biemer, P.P., et Link, M.W. (2007). *Evaluating and Modeling early Cooperator Effects in RDD Surveys*. New York : John Wiley & Sons, Inc.

Bond, T., et Fox, C. (2007). *Applying the Rasch model: Fundamental Measurement in the Human Sciences* (2e Éd.). Lawrence Erlbaum Associates, Inc, Mahwah, N.J.

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. Dans *Incomplete Data in Sample Surveys,* (Éds., W.G. Madow et I. Olkin), New York : Academic Press. 3, 143-160.

Chambers, R.L., et Skinner, C. (2003). *Analysis of Survey Data*. New York : John Wiley & Sons, Inc.

Copas, A.J., et Farewell, V.T. (1998). Dealing with non-ignorable non-response by using an 'enthusiasm-to-respond' variable. *Journal of the Royal Statistical Society, Series A,* 161, 385-396.

De Menezes, L.M., et Bartholomew, D.J. (1996). New developments in latent structure analysis applied to social attitudes. *Journal of Royal Statistical Society A,* 159, 213-224.

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association,* 88, 1013-1020.

Drew, J.H., et Fuller, W.A. (1980). Modeling nonresponse in surveys with callbacks. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Eltinge, J.L., et Yansaneh, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête,* 23, 1, 37-45.

Greenlees, J.S., Reece, W.S. et Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association,* 77, 251-261.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly,* 70, 5, 646-675.

Groves, R.M., Couper, M., Presser, S., Singer, E., Tourangeau, R., Acosta, G.P. et Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly,* 70, 5, 720-736.

Kim, J.K., et Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics,* 35, 501-514.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association,* 101, 473, 312-320.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête,* 32, 2, 149-160.

Langford, E., Schwertman, N. et Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician*, 55, 4, 322-325.

Legg, J.C., et Fuller, W.A. (2009). Two-phase sampling. *Handbook of Statistics*, 29, 55-70.

Little, R.J., et Vartivarian, S. (2005). La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage ? *Techniques d'enquête*, 31, 2, 175-183.

Little, R.J.A., et Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley & Sons, Inc.

Moran, P.A.P. (1986). Identification problems in latent trait models. *British Journal of Mathematical and Statistical Psychology*, 39, 2, 208-212.

Moustaki, I., et Knott, M. (2000). Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of Royal Statistical Society, Series A,* 163, 445-459.

Oh, H.L., et Scheuren, F.J. (1983). Weighting adjustments for unit non-response. Dans *Incomplete Data in Sample Surveys*, (Éds., W.G. Madow, I. Olkin et D.B. Rubin). New York : Academic Press. 2, 143-184.

Olsson, U., Drasgow, F. et Dorans, N. (1982). The polyserial correlation coefficient. *Psychometrika,* 47, 337-347.

Qin, J., Leung, D. et Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association,* 97, 193-200.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. *The Danish Institute of Educational Research*, Copenhagen.

Reardon, S.F., et Raudenbush, S.W. (2006). A partial independence item response model for surveys with filter questions. *Sociological Methodology,* 36, 1, 257-300.

Rizopoulos, D. (2006). `ltm`: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software,* 17, 5, 1-25.

Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score inobservational studies for causal effects. *Biometrika,* 70, 41-55.

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

Scott, S.L., et Ip, E.H. (2002). Empirical bayes and item-clustering effects in a latent variable hierarchical model: A case study from the national assessment of educational progress. *Journal of American Statistical Association,* 97, 459, 1-11.

Skrondal, A., et Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics,* 34, 712-745.

Tillé, Y., et Matei, A. (2012). *Sampling: Survey Sampling*. R package version 2.5.

Wright, B. (1996). Local dependency, correlations and principal components. *Rasch Meas Trans,* 10-3, 509-511.

Zhang, L.C. (2002). A method of weighting adjustment for survey data subject to nonignorable nonresponse. Document de recherche de la DACSEIS nu. 2, http://w210.ub.unituebingen.de/dbt/volltexte/2002/451.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 173-190
Statistique Canada, n° 12-001-X au catalogue

Une ou deux étapes ? Pondération par calage à partir d'une base liste complète en présence de non-réponse

Phillip S. Kott et Dan Liao[1]

Résumé

Quand un échantillon aléatoire tiré d'une base liste complète souffre de non-réponse totale, on peut faire appel à la pondération par calage sur des totaux de population pour éliminer le biais de non-réponse sous un modèle hypothétique de réponse (sélection) ou de prédiction (résultat). De cette façon, la pondération par calage peut non seulement procurer une double protection contre le biais de non-réponse, mais aussi réduire la variance. En employant une astuce simple, on peut estimer simultanément la variance sous le modèle hypothétique de prédiction et l'erreur quadratique moyenne sous la combinaison du modèle hypothétique de réponse et du mécanisme d'échantillonnage probabiliste. Malheureusement, il existe une limite pratique aux types de modèle de réponse que l'on peut supposer lorsque les poids de sondage sont calés sur les totaux de population en une seule étape. En particulier, la fonction de réponse choisie ne peut pas toujours être logistique. Cette limite ne gêne pas la pondération par calage lorsqu'elle est effectuée en deux étapes : de l'échantillon de répondants à l'échantillon complet pour éliminer le biais de réponse, et puis de l'échantillon complet à la population pour réduire la variance. Des gains d'efficacité pourraient découler de l'utilisation de l'approche en deux étapes, même si les variables de calage employées à chaque étape représentent un sous-ensemble des variables de calage de l'approche en une seule étape. L'estimation simultanée de l'erreur quadratique moyenne par linéarisation est possible, mais plus compliquée que lorsque le calage est effectué en une seule étape.

Mots-clés : Échantillonnage probabiliste; modèle de réponse; modèle de prédiction; double protection; estimation simultanée des variances.

1 Introduction

Le sondage est un outil utilisé surtout pour estimer les paramètres d'une population finie en se basant sur un échantillon de ses membres tiré aléatoirement. Les échantillons probabilistes sont assortis de poids de sondage (d'échantillonnage) qui sont souvent les inverses des probabilités de sélection des membres individuels. À condition que chaque élément de la population possède une probabilité de sélection positive, il est simple de produire un estimateur du total de population de la variable étudiée qui est sans biais par rapport au mécanisme d'échantillonnage probabiliste. Le ratio de deux estimateurs sans biais des totaux, ou toute autre fonction lisse des totaux estimés, n'est pas forcément sans biais, mais est asymptotiquement sans biais et souvent convergent puisque sa variance relative, comme son biais relatif, tend vers zéro quand la taille de l'échantillon devient arbitrairement grande.

Deville et Särndal (1992) ont introduit la pondération par calage comme outil d'ajustement des poids de sondage de façon que les sommes pondérées de certaines variables de « calage » soient égales à leurs totaux de population connus (ou mieux estimés). Si ces *équations de calage* sont vérifiées, l'erreur-type d'un total estimé pour une variable dont le total de population est inconnu est souvent réduite, tandis que l'estimation demeure quasi (c'est-à-dire asymptotiquement) sans biais sous le mécanisme d'échantillonnage probabiliste.

Bien qu'elle ait été élaborée au départ pour réduire les erreurs-types, la pondération par calage a souvent été utilisée pour éliminer le biais de sélection résultant de la non-réponse totale sous certaines

1. Phillip S. Kott, statisticien-chercheur principal, RTI International, Rockville, Maryland 20852, États-Unis. Courriel : pkott@rti.org; Dan Liao, statisticien-chercheur, RTI International, Rockville, Maryland 20852, États-Unis.

hypothèses (par exemple, Folsom 1991; Fuller, Loughin et Baker 1994; Lundström et Särndal 1999; Folsom et Singh 2000). À cette fin, on traite le fait qu'un élément sélectionné dans l'échantillon répond (ou non) à une enquête comme une phase additionnelle de l'échantillonnage aléatoire de Poisson avec probabilités de sélection inconnues, mais positives. La pondération par calage estime ces probabilités de sélection de Poisson implicitement et produit des totaux estimés qui sont presque sans biais sous le mécanisme combiné de sélection de l'échantillon et des répondants, qui est souvent appelé le « quasi-plan d'échantillonnage ». Voir Oh et Scheuren (1983).

Une *mise en garde* importante est que, si le mécanisme de sélection de l'échantillon est entièrement sous le contrôle du statisticien, le mécanisme de sélection des réponses est inconnu. Une hypothèse est émise quant à la forme particulière du mécanisme de réponse, et si cette hypothèse n'est pas vérifiée, les estimateurs peuvent être biaisés.

Une autre justification de la pondération par calage s'appuie sur un type de modélisation différent. Il est facile de montrer que la pondération par calage produit un estimateur qui est sans biais sous un modèle de prédiction (résultat) linéaire si la valeur prévue de la variable étudiée sous le modèle de prédiction est une fonction linéaire des variables de calage pourvu que les mécanismes d'échantillonnage et de réponse soient ignorables, c'est-à-dire que l'on puisse appliquer le même modèle de prédiction que l'élément de la population soit ou non échantillonné ou qu'il réponde ou non lorsqu'il est échantillonné.

Contrairement au modèle de sélection qui régit le mécanisme de réponse, il est possible que le modèle de prédiction linéaire soit vérifié pour une variable étudiée et non pour une autre. C'est la raison pour laquelle la plupart des échantillonneurs préfèrent émettre l'hypothèse d'un *modèle de sélection* lorsqu'ils corrigent la non-réponse totale. Néanmoins, il est rassurant de savoir que si *l'un ou l'autre* modèle est correct, le total estimé est quasi sans biais (c'est-à-dire qu'il possède un biais relatif qui s'évanouit asymptotiquement), une propriété que Kim et Park (2006) ont appelée « double protection » contre le biais de non-réponse.

Il est possible de simultanément éliminer le biais de sélection et réduire l'erreur-type sous le mécanisme d'échantillonnage probabiliste en une seule étape en ajustant les poids de sondage des unités répondantes afin que les totaux estimés pour un ensemble de variables de calage soient égaux aux totaux de population connus de ces unités. Néanmoins, il existe des raisons de préférer l'approche de pondération par calage en deux étapes, même quand les ensembles de variables de calage utilisés aux deux étapes sont les mêmes ou sont un sous-ensemble des variables de calage de l'approche en une étape : la première étape, de l'échantillon de répondants à l'échantillon original, élimine le biais de sélection et la deuxième étape, de l'échantillon original à la population, réduit la variance des estimateurs résultants.

Bien que Folsom et Singh (2000) et d'autres aient souligné que la pondération par calage peut aussi être utilisée pour éliminer le biais de sélection dû à une sous-couverture ou une surcouverture de la base de sondage, nous nous concentrons ici sur un échantillon à un degré tiré d'une base liste complète sans enregistrements en double. Autrement dit, nous supposons que la base de sondage est identique à la population cible (c'est-à-dire que chaque unité de la population est énumérée sur la liste de la base de sondage).

La présentation de l'article est la suivante. À la section 2, nous passons en revue certains éléments de théorie sur la pondération par calage. À la section 3, nous présentons un estimateur de variance légèrement nouveau qui, comme l'estimateur de variance décrit dans Kott (2006), peut être utilisé pour mesurer à la fois l'erreur quadratique moyenne d'un estimateur pondéré par calage sous le quasi-plan

d'échantillonnage et la variance sous le modèle de prédiction ou la combinaison du modèle de prédiction et du mécanisme d'échantillonnage original, ce qui rend sans doute la double protection contre le biais de non-réponse plus utile pour l'inférence. L'estimateur de variance donné dans Kott s'applique seulement lorsque le calage se fait sur les valeurs de population. Ici, à l'instar de Folsom et Singh (2000), nous donnons la possibilité d'effectuer le calage sur l'échantillon original.

À la section 4, nous discutons des limites de la pondération par calage en une seule étape et élaborons une théorie pour l'approche en deux étapes. Bien que notre principal objectif ici soit de faire valoir les avantages de l'utilisation de deux étapes, même lorsque des ensembles similaires de variables de calage sont employés aux deux étapes, l'estimateur par calage que nous traitons dans cette section est plus général. À la section 5, nous décrivons les résultats de certaines expériences par simulation, tandis qu'à la section 6, nous tirons quelques conclusions.

2 Pondération par calage en une étape

2.1 Pondération par calage et non-réponse totale

En l'absence de non-réponse (ou d'erreurs de base de sondage), la pondération par calage est une méthode d'ajustement des poids d'échantillonnage en vue de créer un ensemble de poids $\{w_k; k \in S\}$, asymptotiquement proche des poids de sondage originaux, $d_k = 1/\pi_k$, qui satisfont à un ensemble d'*équations de calage* (une pour chaque composante de $\mathbf{z}_k$) :

$$\sum_S w_k \mathbf{z}_k = \sum_U \mathbf{z}_k,$$

où S désigne l'échantillon, π_k désigne la probabilité de sélection dans l'échantillon de l'unité k, U désigne la population de taille N, $\mathbf{z}_k$ est un vecteur comprenant P composantes ayant chacune un total de population connu, et $\sum_A$ signifie $\sum_{k \in A}$.

Kott (2009) décrit un ensemble prudent de conditions faibles sous lesquelles $t_y = \sum_S w_k y_k$ est un estimateur quasi sans biais du total de population $T_y = \sum_U y_k$ (c'est-à-dire que le biais relatif de t_y est asymptotiquement nul). Fait plus important, on suppose que chaque probabilité $\pi_k N/n$ possède une borne inférieure positive égale à N et que la taille d'échantillon (prévue), n, devient arbitrairement grande (nous ajoutons entre parenthèses le terme « prévue » au cas où la taille d'échantillon est aléatoire).

En outre, on suppose que les quatre premiers moments de population centrés de chaque composante de $\mathbf{z}_k$ possèdent une borne supérieure, tandis que $N^{-1} \sum_U \mathbf{z}_k \mathbf{z}_k^T$ converge vers une matrice définie positive.

L'utilisation de la pondération par calage aura tendance à réduire l'erreur quadratique moyenne par rapport à l'estimateur à facteur d'extension (*expansion estimator*), $t_y^E = \sum_S d_k y_k$, quand y_k est corrélée à certaines composantes de $\mathbf{z}_k$. Cependant, il ne faut pas perdre de vue que, dans la plupart des enquêtes, les variables étudiées y_k sont nombreuses.

Un moyen simple de calculer les poids de calage consiste à le faire linéairement en utilisant la formule suivante :

$$\begin{aligned} w_k &= d_k\left[1 + \left(\sum_U \mathbf{z}_j - \sum_S d_j \mathbf{z}_j\right)^T \left(\sum_S d_j \mathbf{z}_j \mathbf{z}_j^T\right)^{-1} \mathbf{z}_k\right] \\ &= d_k\left[1 + \mathbf{g}^T \mathbf{z}_k\right]. \end{aligned}$$

Fuller et coll. (1994) et plus tard Lundström et Särndal (1999) ont soutenu que ce calage linéaire peut aussi être utilisé pour traiter la non-réponse totale. L'échantillon S est remplacé par l'échantillon de répondants R, tandis que

$$\mathbf{g} = \left[(1-\theta)\left(\sum_U \mathbf{z}_j - \sum_R d_j \mathbf{z}_j\right)^T + \theta\left(\sum_S d_j \mathbf{z}_j - \sum_R d_j \mathbf{z}_j\right)^T\right]\left(\sum_R d_j \mathbf{z}_j \mathbf{z}_j^T\right)^{-1},$$

selon que l'échantillon de répondants est *calé sur la population* $(\theta = 0)$ ou *calé sur l'échantillon original* $(\theta = 1)$. Dans l'un et l'autre cas, l'estimation est quasi sans biais sous le quasi-plan d'échantillonnage qui traite la réponse comme une deuxième phase d'échantillonnage aléatoire à condition que la probabilité de réponse de chaque unité soit de la forme :

$$p_k = 1/\left(1 + \boldsymbol{\gamma}^T \mathbf{z}_k\right), \tag{2.1}$$

et $\mathbf{g}$ est un estimateur convergent du vecteur de paramètres inconnus $\boldsymbol{\gamma}$ dans l'équation (2.1).

Le problème en ce qui concerne la fonction de réponse donnée par l'équation (2.1) est que l'estimateur implicite de p_k, $\hat{p}_k = 1/\left(1 + \mathbf{g}^T \mathbf{z}_k\right)$ peut être négatif. Une forme non linéaire de la pondération par calage permettant d'éviter cette possibilité a été proposée par Kott et Liao (2012) qui se sont fondés sur la forme exponentielle généralisée de Folsom et Singh (2000). Cette forme de calage fait appel à la méthode de Newton (approximations itératives du développement en série de Taylor) pour trouver un $\mathbf{g}$ tel que l'équation de calage (à partir d'ici, nous utilisons le terme équation de calage pour faire référence au vecteur des équations de calage des composantes) :

$$\sum_R w_k \mathbf{z}_k = \sum_R d_k \alpha\left(\mathbf{g}^T \mathbf{z}_k\right) \mathbf{z}_k = (1-\theta)\sum_U \mathbf{z}_k + \theta \sum_S d_k \mathbf{z}_k \tag{2.2}$$

est vérifiée, où $\theta = 0$ ou 1,

$$\alpha\left(\mathbf{g}^T \mathbf{z}_k\right) = \frac{\ell + \exp\left(\mathbf{g}^T \mathbf{z}_k\right)}{1 + \exp\left(\mathbf{g}^T \mathbf{z}_k\right)/u}, \tag{2.3}$$

ℓ, la borne inférieure de $\alpha(\cdot)$, est non négative (de sorte que les poids de calage sont également non négatifs), et la borne supérieure de $\alpha(\cdot)$, $u > \ell$, peut être finie ou infinie.

Bien que la *fonction d'ajustement des poids* $\alpha\left(\mathbf{g}^T \mathbf{z}_k\right)$ puisse prendre d'autres formes raisonnables, nous nous limiterons aux fonctions de la forme de l'équation (2.3). Il s'agit d'une généralisation du ratissage (*raking*) où $\ell = 0, u = \infty$, ainsi que de l'estimation implicite d'un modèle de réponse logistique, où $\ell = 1, u = \infty$. Dans l'algorithme d'ajustement proportionnel itératif original de Deming et Stephan (1940) pour le ratissage, les composantes de $\mathbf{z}_k$ ont été restreintes à des fonctions indicatrices. Nous utilisons ici le terme « ratissage » de manière plus générale pour désigner une pondération par calage avec une fonction d'ajustement des poids de la forme $\alpha\left(\mathbf{g}^T \mathbf{z}_k\right) = \exp\left(\mathbf{g}^T \mathbf{z}_k\right)$.

Quand $\ell < 1$, l'équation (2.3) devient l'ajustement par calage généralisé introduit dans Deville et Särndal (1992) et discuté plus en détail dans Deville, Särndal et Sautory (1993). Le calage généralisé permet non seulement que les composantes de $\mathbf{z}_k$ soient continues, mais aussi que l'étendue des $\alpha\left(\mathbf{g}^T \mathbf{z}_k\right)$ soit contrainte entre une valeur positive ℓ et une valeur (possiblement) finie u.

Deville et Särndal (1992) posaient comme condition que $\alpha(0) = \alpha'(0) = 1$. Puisqu'ils ne s'intéressaient pas à des échantillons avec non-réponse (ou à des bases de sondage incorrectes), $\mathbf{g}^T\mathbf{z}_k$ devait converger vers 0 et $\alpha(\mathbf{g}^T\mathbf{z}_k)$ vers 1 quand la taille d'échantillon (prévue) devenait arbitrairement grande. Cependant, lorsqu'on ajuste les poids de sondage pour corriger la non-réponse, poser que $\ell \geq 1$ est une stratégie plus raisonnable afin que la probabilité de réponse estimée implicite ne soit pas supérieure à 1.

Tandis que la définition originale de la pondération par calage donnée dans Deville et Särndal (1992) comprenait la minimisation des écarts dans R entre les w_k et d_k, mesurés par une certaine fonction de perte, des formulations ultérieures (par exemple, Estevao et Särndal 2000) ont éliminé la fonction de perte de la définition. Forcer w_k et d_k à être proches a peu de sens quand la pondération par calage est utilisée pour corriger la non-réponse totale, puisque si une unité k échantillonnée a une probabilité relativement faible de réponse, l'écart entre w_k et d_k *doit* être relativement grand.

Au lieu de supposer un modèle de réponse ayant une forme fonctionnelle particulière, une autre justification de l'utilisation de la pondération par calage comme moyen d'éliminer le biais de non-réponse totale consiste à émettre l'hypothèse d'un modèle de prédiction dans lequel la variable étudiée y_k est elle-même une variable aléatoire telle que $\mathrm{E}(y_k|\mathbf{z}_k) = \mathbf{z}_k^T\boldsymbol{\beta}$ pour un $\boldsymbol{\beta}$ inconnu, que l'unité k soit échantillonnée ou non ou qu'elle réponde ou non quand elle est échantillonnée. Kott (2006) et d'autres ont observé que l'estimateur pondéré par calage de $T_y = \sum_U y_k$ sera quasi sans biais sous le modèle de prédiction quand le calage est effectué sur la population (quand $\theta = 0$ dans l'équation (2.2)), et sous la combinaison du modèle de prédiction et du mécanisme de sélection de l'échantillon original quand le calage est effectué sur l'échantillon original (quand $\theta = 1$).

La propriété faisant qu'un estimateur pondéré par calage est dans un certain sens quasi sans biais quand un modèle hypothétique de réponse *ou* un modèle hypothétique de prédiction est vérifié a été appelée « double protection contre le biais de non-réponse » par Kim et Park (2006). Elle est appelée « double robustesse » dans la littérature biostatistique (Bang et Robins 2005) et attribuée à Robins, Rotnitzky et Zhao (1994), qui ont traité la non-réponse partielle plutôt que totale.

On suppose souvent que la distribution de $y_k|\mathbf{z}_k$ sous le modèle de prédiction est la même pour les membres de la population échantillonnés et non échantillonnés. Autrement dit, le mécanisme d'échantillonnage est considéré comme étant *ignorable*. En outre, on suppose souvent que la distribution de $y_k|\mathbf{z}_k$ est la même qu'un membre de la population réponde ou non quand il est échantillonné, c'est-à-dire que le mécanisme de réponse est également considéré comme étant ignorable (Little et Rubin 2002). Ici, nous faisons des hypothèses analogues plus faibles sous le modèle de prédiction, nommément que $\mathrm{E}(y_k|\mathbf{z}_k)$ ne dépend pas du fait que l'unité k est échantillonnée ou non ou qu'elle répond ou non quand elle est échantillonnée. Disons que les mécanismes d'échantillonnage et de réponse sont considérés comme étant « ignorable au premier moment ».

2.2 Variables instrumentales

Deville (2000) a observé que l'on peut utiliser le calage avec des variables instrumentales pour corriger le biais de non-réponse possible en émettant l'hypothèse d'un modèle de réponse qui dépend de $\mathbf{x}_k$,

$$p_k = \left[\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\right]^{-1} = \frac{1+\exp\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)/u}{\ell+\exp\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)}, \tag{2.4}$$

mais en ajustant les équations de calage avec $\mathbf{z}_k$:

$$\sum_R w_k\mathbf{z}_k = \sum_R d_k\alpha\left(\mathbf{g}^T\mathbf{x}_k\right)\mathbf{z}_k = (1-\theta)\sum_U \mathbf{z}_k + \theta\sum_S d_k\mathbf{z}_k, \tag{2.5}$$

où le $\mathbf{g}$ satisfaisant l'équation (2.5) avec $\theta = 0$ ou 1 est un estimateur convergent du vecteur de paramètres inconnus $\boldsymbol{\gamma}$ dans l'équation (2.4). Certaines conditions faibles sont nécessaires ici. Les conditions qui suivent sont suffisantes : $N^{-1}\sum_R d_k\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\mathbf{z}_k$ est un estimateur convergent et borné pour $N^{-1}\left[(1-\theta)\sum_U \mathbf{z}_k + \theta\sum_S d_k\mathbf{z}_k\right]$, $\alpha(\phi)$ est partout deux fois dérivable, et $N^{-1}\sum_R d_k\alpha'(\phi)\mathbf{z}_k\mathbf{x}_k^T$ est toujours inversible et borné quand l'échantillon devient arbitrairement grand.

Soit $R_k = 1$ quand $k \in R, 0$ autrement. Il n'est pas difficile de montrer que

$$\begin{aligned}\mathbf{g}-\boldsymbol{\gamma} = &-\left(\sum_S d_kR_k\alpha'(c_k)\mathbf{z}_k\mathbf{x}_k^T\right)^{-1}\left\{\sum_S d_kR_k\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\mathbf{z}_k - \left[(1-\theta)\sum_U \mathbf{z}_k + \theta\sum_S d_k\mathbf{z}_k\right]\right\}\\ &-\left(N^{-1}\sum_S d_kR_k\alpha'(c_k)\mathbf{z}_k\mathbf{x}_k^T\right)^{-1}\left\{N^{-1}\sum_S d_kR_k\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\mathbf{z}_k - N^{-1}\left[(1-\theta)\sum_U \mathbf{z}_k + \theta\sum_S d_k\mathbf{z}_k\right]\right\}\end{aligned}$$

pour un certain c_k compris entre $\mathbf{g}^T\mathbf{x}_k$ et $\boldsymbol{\gamma}^T\mathbf{x}_k$, comme l'ont démontré Kott et Liao (2012) quand $\mathbf{x}_k = \mathbf{z}_k$.

Deville note également que les composantes de $\mathbf{x}_k$ peuvent être des variables étudiées dont les valeurs ne sont connues que pour les répondants. Chang et Kott (2008) ont étendu la notion de la pondération par calage afin de permettre que la dimension du vecteur $\mathbf{z}_k$ soit plus grande que celle du vecteur $\mathbf{x}_k$. Nous *ne* traiterons *ni* l'une *ni* l'autre possibilité dans les sections qui suivent.

Kim et Shao (2013), en traitant la non-réponse non ignorable, désignent par « variables instrumentales » les composantes de $\mathbf{z}_k$ qui ne sont pas entièrement des fonctions des composantes de $\mathbf{x}_k$. Pour limiter toute confusion future, nous utiliserons donc le terme « variables du modèle » pour désigner les composantes de $\mathbf{x}_k$.

3 Estimation de la variance de l'estimateur par calage en une étape

À la présente section, nous posons que

$$t_y = \sum_R w_k y_k = \sum_R d_k\alpha\left(\mathbf{g}^T\mathbf{x}_k\right)y_k$$

est l'estimateur pondéré par calage de T_y, où $w_k = d_k\alpha\left(\mathbf{g}^T\mathbf{x}_k\right)$ quand $k \in R$ est le poids de calage, et w_k est défini de façon commode comme étant égal à 0 quand $k \notin R$. La fonction d'ajustement des poids $\alpha(\cdot)$ est définie implicitement par l'équation (2.4), et $\mathbf{g}$ est de nouveau choisi de façon que l'équation de calage (2.5) soit vérifiée pour $\theta = 0$ ou 1.

Nous proposons l'estimateur suivant de la variance de t_y :

$$v(t_y) = \sum_{k,j \in S} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right)\left[d_k\left(\theta \mathbf{z}_k^T \mathbf{b} + \alpha_k e_k\right)\right]\left[d_j\left(\theta \mathbf{z}_j^T \mathbf{b} + \alpha_j e_j\right)\right] + \sum_{k \in R} d_k\left(\alpha_k^2 - \alpha_k\right) e_k^2, \quad (3.1)$$

où π_{kj} est la probabilité de sélection conjointe de k et j sous le plan d'échantillonnage original, $\pi_{kk} = \pi_k = 1/d_k$, $\pi_k = \alpha\left(\mathbf{g}^T \mathbf{x}_k\right)$ quand $k \in R$ et 0 autrement,

$$\mathbf{b} = \left[\sum_R d_k \alpha'\left(\mathbf{g}^T \mathbf{x}_k\right) \mathbf{x}_k \mathbf{z}_k^T\right]^{-1} \sum_R d_k \alpha'\left(\mathbf{g}^T \mathbf{x}_k\right) \mathbf{x}_k y_k, \quad (3.2)$$

et $e_k = y_k - \mathbf{z}_k^T \mathbf{b}$. Nous montrerons que $v(t_y)$ dans l'équation (3.1) peut être quasi sans biais dans un certain sens si *soit* un modèle de réponse (section 3.1) *soit* un modèle de prédiction est vérifié (section 3.2).

L'estimateur de variance dans l'équation (5.2) de Kott (2006) est identique à $v(t_y)$ dans l'équation (3.1) quand $\theta = 0$. L'estimateur de variance dans Kim et Haziza (2014) est également similaire. Leur modèle de prédiction est plus général que le modèle de prédiction linéaire considéré ici.

Cet estimateur de variance $v(t_y)$ présuppose que le plan d'échantillonnage original est tel que chaque élément ne peut être tiré qu'une seule fois. À la section 3.1, nous voyons que, quand les probabilités de réponse sont indépendantes (Poisson), alors sous des hypothèses faibles, $v(t_y)$ est un estimateur quasi sans biais de l'erreur quadratique moyenne de t_y sous le quasi-plan d'échantillonnage, que le modèle de prédiction, $\mathrm{E}\left(y_k \mid \mathbf{x}_k, \mathbf{z}_k\right) = \mathbf{z}_k^T \boldsymbol{\beta}$, soit vérifié ou non.

À la section 3.2, nous montrons que $v(t_y)$ est un estimateur quasi sans biais pour le modèle de prédiction combiné à la variance sous le plan d'échantillonnage original de t_y en tant qu'estimateur de T_y, que le modèle de réponse donné par l'équation (2.4) soit vérifié ou non. Donc, $v(t_y)$ peut être appelé un « estimateur simultané des variances ».

3.1 Estimation de la variance sous le modèle de réponse

Pour simplifier l'exposé, nous supposerons que le modèle de réponse donné par l'équation (2.4) avec une borne supérieure u finie est vérifié. Les conditions suffisantes pour que $v(t_y)$ soit un estimateur quasi sans biais de l'erreur quadratique moyenne de t_y (en vertu desquelles le biais converge vers 0 quand la taille de l'échantillon devient arbitrairement grande) sont

$$\pi_{kj} \geq B_0 > 0 \quad (3.3)$$

$$\sum_{j=1}^{N} \left| \frac{\pi_{kj}}{\pi_k \pi_j} - 1 \right| \leq B_1 < \infty \text{ pour chaque } k, \quad (3.4)$$

$$\frac{\sum_{j=1}^{N} \psi_j^r}{N} \leq B_2 < \infty \text{ où } \psi_j \text{ est } y_j \text{ ou toute composante de } \mathbf{x}_j \text{ ou } \mathbf{z}_j, \text{ tandis que } r = 1 \text{ ou } 2, \quad (3.5)$$

et $N^{-1}\sum_R d_k\alpha'\left(\mathbf{g}^T\mathbf{x}_k\right)\mathbf{z}_k\mathbf{x}_k^T$ est de plein rang et est bornée en probabilité quand la taille de l'échantillon devient arbitrairement grande.

En vertu de cela, de $\alpha'(\phi) = (1 - \alpha(\phi)/u)\exp(\phi)/[(1 + \exp(\phi)/u)]$ étant bornée quand u est finie, et de l'inégalité de Cauchy-Schwarz $\left((\sum a_k b_k)^2 \le \sum a_k^2 \sum b_k^2\right)$, il n'est pas difficile de voir non seulement que $\mathbf{g}$ est un estimateur convergent de $\boldsymbol{\gamma}$, mais aussi que $\mathbf{b}$ dans l'équation (3.2) (qui peut être rendue sous la forme $\mathbf{b} = \left[N^{-1}\sum_R d_k\alpha'(\mathbf{g}^T\mathbf{x}_k)\mathbf{x}_k\mathbf{z}_k^T\right]^{-1} N^{-1}\sum_R d_k\alpha'(\mathbf{g}^T\mathbf{x}_k)\mathbf{x}_k y_k$) possède une limite en probabilité, que nous appellerons $\mathbf{b}^*$, que le modèle de prédiction soit vérifié ou non. En outre, $\mathbf{b} - \mathbf{b}^*$ ainsi que $\mathbf{g} - \boldsymbol{\gamma}$ sont $\mathbf{O}_p\left(1/\sqrt{n}\right)$.

Observons que

$$\begin{aligned}(t_y - T_y)/N &= \theta\left(\sum_S d_k\mathbf{z}_k^T\mathbf{b}^* - \sum_U \mathbf{z}_k^T\mathbf{b}^*\right)/N \\ &+ \left[\sum_R d_k\alpha\left(\mathbf{g}^T\mathbf{x}_k\right)e_k^* - \sum_R d_k\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)e_k^*\right]/N \\ &+ \left[\sum_R d_k\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)e_k^* - \sum_U e_k^*\right]/N,\end{aligned}$$

où $e_k^* = y_k - \mathbf{z}_k^T\mathbf{b}^*$. L'insertion de $\alpha'(\cdot)$ dans le « coefficient de régression » $\mathbf{b}$ nous permet d'ignorer la contribution du deuxième terme de cette somme, $Q = \sum_R d_k\left[\alpha\left(\mathbf{g}^T\mathbf{x}_k\right) - \alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\right]e_k^*/N$, à l'erreur quadratique moyenne sous le quasi-plan d'échantillonnage. Il en est ainsi parce que $\sum_R d_k\alpha'\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\mathbf{x}_k e_k = 0$ est vraie par définition, ce qui implique que $\sum_R d_k\alpha'\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\mathbf{x}_k e_k^*$ est $\mathrm{O}_p\left(1/\sqrt{n}\right)$ sous nos hypothèses. En outre, puisque $\alpha\left(\mathbf{g}^T\mathbf{x}_k\right) - \alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right) = \alpha'(c_k)\left(\mathbf{g} - \boldsymbol{\gamma}\right)^T\mathbf{x}_k$ est aussi $\mathrm{O}_p\left(1/\sqrt{n}\right)$, $Q = (\mathbf{g} - \boldsymbol{\gamma})^T\sum_R d_k\alpha'(c_k)\mathbf{x}_k e_k^*$ est $\mathrm{O}_p(1/n)$, qui est asymptotiquement ignorable par rapport aux deux composantes $\mathrm{O}_p\left(1/\sqrt{n}\right)$ de $(t_y - T_y)/N$.

La contribution de Q étant éliminée, un estimateur sans biais idéalisé, mais incalculable, de l'erreur quadratique moyenne sous le quasi-plan d'échantillonnage de t_y est donné par

$$v_{I1}\left(t_y\right) = \sum_{k,j\in S}\left(1 - \frac{\pi_k\pi_j}{\pi_{kj}}\right)\left[d_k\left(\theta\mathbf{z}_k^T\mathbf{b}^* + e_k^*\right)\right]\left[d_j\left(\theta\mathbf{z}_j^T\mathbf{b}^* + e_j^*\right)\right] + \sum_{k\in R}\left(\frac{d_k e_k^*}{p_k}\right)^2(1 - p_k), \quad (3.6)$$

où le premier terme du deuxième membre estime l'erreur quadratique moyenne avant la non-réponse (s'il y en a une) et le deuxième terme estime la variance ajoutée par la non-réponse.

Un estimateur quasi sans biais idéalisé de l'erreur quadratique moyenne de rechange, plus près d'être calculable, est donné par

$$v_{I2}\left(t_y\right) = \sum_{k,j\in S}\left(1 - \frac{\pi_k\pi_j}{\pi_{kj}}\right)\left[d_k\left(\theta\mathbf{z}_k^T\mathbf{b}^* + \frac{R_k}{p_k}e_k^*\right)\right]\left[d_j\left(\theta\mathbf{z}_j^T\mathbf{b}^* + \frac{R_j}{p_j}e_j^*\right)\right] + \sum_{k\in R} d_k\left(\frac{e_k^*}{p_k}\right)^2(1 - p_k), \quad (3.7)$$

où de nouveau $R_k = 1$ quand $k \in R, 0$ autrement. Puisque les $(R_k/p_k)e_k^*$ sont indépendants sous le modèle de réponse et sont de moyenne e_k^* et de variance $\left(e_k^*/p_k\right)^2 p_k(1 - p_k)$, $\mathrm{E}\left[(R_k/p_k)e_k^*\left(R_j/p_j\right)e_j^*\right] = e_k^* e_j^*$ quand $k \ne j$. Par contre, l'expression qui suit est vérifiée quand $k = j$:

$$(1-\pi_k)\,E\left[\left(d_k\,\frac{R_k}{p_k}\,e_k^*\right)^2\right] = (1-\pi_k)\left[\left(d_k e_k^*\right)^2 + \left(\frac{d_k e_k^*}{p_k}\right)^2 p_k\,(1-p_k)\right]$$

$$= (1-\pi_k)\left(d_k e_k^*\right)^2 + \left(\frac{d_k e_k^*}{p_k}\right)^2 p_k\,(1-p_k) - d_k\left(\frac{e_k^*}{p_k}\right)^2 p_k\,(1-p_k).$$

La première sommation dans le deuxième membre de l'équation (3.7) contient des termes où $k \neq j$ et des termes où $k = j$, les derniers faisant que la deuxième sommation dans (3.7) diffère de la deuxième sommation dans le deuxième membre de l'équation (3.6). Notons que l'espérance sous le modèle de réponse de $\sum_R d_k \left(e_k^*/p_k\right)^2 (1-p_k)$ dans la deuxième sommation dans le deuxième membre de (3.7) est $\sum_S d_k \left(e_k^*/p_k\right)^2 p_k\,(1-p_k)$.

Enfin, $v_{I2}\left(t_y\right)$ peut être remplacé par l'estimateur $v\left(t_y\right)$ asymptotiquement identique, mais calculable, dans l'équation (3.1) puisque $\sum_{j\in S}\left(1-\pi_k\pi_j/\pi_{kj}\right)$ est borné pour tout k sous les hypothèses (3.3) et (3.4), ce qui permet de substituer e_k et α_k à e_k^* et $1/p_k$ inconnus, respectivement (parce que $e_k^* - e_k$ et $\alpha_k - 1/p_k$ sont $\mathrm{O}_p\left(1/\sqrt{n}\right)$ pour tout k).

3.2 Estimation de la variance sous le modèle de prédiction

Les choses sont un peu plus simples quand nous supposons qu'un modèle de prédiction est vérifié mais que le modèle de réponse de l'équation (2.4) ne l'est pas nécessairement. Supposons que $\mathrm{E}\left(y_k|\mathbf{x}_k,\mathbf{z}_k\right) = \mathbf{z}_k^T\boldsymbol{\beta}$, peu importe que l'unité k soit échantillonnée ou non ou qu'elle réponde ou non quand elle est échantillonnée, et que les $\varepsilon_k = y_k - \mathbf{z}_k^T\boldsymbol{\beta}$ sont des variables aléatoires non corrélées de variance égale à $\sigma_k^2 = \mathbf{z}_k^T\boldsymbol{\eta}$, où $\boldsymbol{\eta}$ ne nécessite pas d'autres spécifications que le fait d'avoir des composantes finies.

L'erreur quadratique moyenne de t_y en tant qu'estimateur de T_y sous le modèle de prédiction est égale à la somme de la variance de prédiction de t_y en tant qu'estimateur de $T_y, \sum_R\left(w_k^2 - w_k\right)\sigma_k^2$ (voir, par exemple, Kott 2009, page 69), et du carré du biais, $\left(\sum_S \mathbf{x}_k^T\boldsymbol{\beta} - \sum_U \mathbf{x}_k^T\boldsymbol{\beta}\right)^2$, ce dernier étant égal à zéro quand $\theta = 0$. La variance combinée de t_y en tant qu'estimateur de T_y sous le modèle de prédiction et le plan d'échantillonnage original est donnée par

$$V_C = \theta\mathrm{Var}_D\left(\sum_S \mathbf{x}_k^T\boldsymbol{\beta}\right) + \mathrm{E}_D\left[\sum_S\left(w_k^2 - w_k\right)\sigma_k^2\right],$$

où l'indice inférieur D indique que l'opération (variance ou espérance) est effectuée par rapport au plan d'échantillonnage original. Rappelons que $w_k = 0$ pour $k \neq R$.

Pour voir que $v\left(t_y\right)$ dans l'équation (3.1) donne un estimateur quasi sans biais de V_C, observons d'abord que

$$e_k = y_k - \mathbf{z}_k^T\mathbf{b} = \varepsilon_k - \mathbf{z}_k^T\left[N^{-1}\sum_R d_j\alpha'\left(\mathbf{g}^T\mathbf{x}_j\right)\mathbf{x}_j\mathbf{z}_j^T\right]^{-1} N^{-1}\sum_R d_j\alpha'\left(\mathbf{g}^T\mathbf{x}_j\right)\mathbf{x}_j\varepsilon_j.$$

Soit $\delta_{kj} = 1$ quand $k = j$ et 0 autrement. Parce que les ε_k ne sont pas corrélés, et que $\mathrm{E}(\varepsilon_k^2) = \sigma_k = \mathbf{z}_k^T \boldsymbol{\eta}$, il est maintenant facile de montrer que $\mathrm{E}(e_k e_j) = \delta_{kj}\sigma_k^2 + \mathrm{O}(1/n)$ pour presque chaque paire k, j sous le modèle de prédiction quand $N^{-1}\sum_R d_k \alpha'(\mathbf{g}^T\mathbf{x}_k)\mathbf{z}_k\mathbf{x}_k^T$ converge vers une matrice inversible, et que les hypothèses (3.3), (3.4), et

$$\frac{\sum_{j=1}^{N} \psi_j^r}{N} \leq B_2 < \infty \text{ où } \psi_j \text{ désigne toute composante de } \mathbf{x}_j \text{ ou } \mathbf{z}_j, \text{ et } r = 1,\ 2,\ 3 \text{ ou } 4, \quad (3.8)$$

sont vérifiées. Observons que le changement provenant des hypothèses dans (3.5) à (3.8) fait que le biais relatif de $v(t_y)$ est un estimateur de V_C (ou $\sum_R (w_k^2 - w_k)\sigma_k^2$ quand $\theta = 0$) $\mathrm{O}(1/n)$ plutôt que $\mathrm{O}(1/\sqrt{n})$.

4 Pondération par calage en deux étapes

4.1 Pondération par calage en deux étapes

En pratique, les composantes de $\mathbf{x}_k$ sont souvent des identificateurs d'appartenance à un groupe de type 0/1, et les groupes sont mutuellement exclusifs et exhaustifs. Dans cette situation, $\mathbf{g}^T\mathbf{x}_k$ ne peut prendre que P valeurs. *Presque* toute fonction d'ajustement des poids, $\alpha(\mathbf{g}^T\mathbf{x}_k)$, donnera des résultats équivalents. La fonction linéaire, $\alpha(\mathbf{g}^T\mathbf{x}_k) = 1 + \mathbf{g}^T\mathbf{x}_k$, de Lundström et Särndal (1999) en est un exemple.

Une fonction d'ajustement des poids d'usage répandu qui, parfois, *ne peut pas* être utilisée (noter le mot « presque » en italiques dans le paragraphe précédent) est $\alpha(\mathbf{g}^T\mathbf{x}_k) = 1 + \exp(\mathbf{g}^T\mathbf{x}_k)$, qui suppose que la réponse est une fonction logistique de $\mathbf{x}_k$. Le problème est que cette fonction d'ajustement des poids ne peut pas retourner des valeurs plus petites que l'unité. Nous avons mentionné à la section précédente que, parfois, on peut avoir besoin que α_k soit plus petit que 1. Une routine qui essaie d'utiliser $\alpha(\mathbf{g}^T\mathbf{x}_k) = 1 + \exp(\mathbf{g}^T\mathbf{x}_k)$ et d'ajuster les équations de calage échouera.

Cela peut poser problème en particulier quand on émet l'hypothèse d'un modèle de réponse logistique et que l'on essaie de le caler sur la population en une seule étape. Il pourrait exister une composante de $\mathbf{z}_k$, disons z_{ka}, qui est toujours non négative, mais l'échantillon original et l'ensemble de réponses sont tels que $\sum_R d_k z_{ka} > \sum_U z_{ka}$ même si $\sum_R d_k z_{ka}$ ne peut pas excéder $\sum_S d_k z_{ka}$. Donc, le calage sur la population échouera toujours, parce qu'aucun α_k ne peut être plus petit que 1.

Le calage sur l'échantillon original, par contre, ne doit pas échouer, puisque $\sum_R d_k z_{ka} \leq \sum_S d_k z_{ka}$. Cela suggère que l'on effectue d'abord le calage sur l'échantillon original, ce qui élimine le biais de réponse si le modèle hypothétique de réponse est vérifié, puis sur la population, ce qui élimine le biais de réponse si le modèle de prédiction est vérifié. Estevao et Särndal (2002) discutent de divers moyens de procéder au calage par étapes, mais nous nous concentrons sur une seule méthode ici.

Un deuxième avantage de la pondération par calage en deux étapes tient au fait qu'elle peut être réalisée même si les variables de calage utilisées aux deux étapes sont les mêmes ou sont un

sous-ensemble de celles utilisées dans la méthode en une seule étape. Cela se produit quand le modèle de réponse est vérifié et que le modèle de prédiction linéaire n'est qu'approximativement vrai. Une certaine version ou estimation « optimale » peut alors être utilisée à la deuxième étape de pondération par calage pour accroître l'efficacité. Rao (1994) a introduit la notion d'estimateur par la régression optimal. Il a été mis sous forme de pondération par calage et discuté plus en détail dans Bankier (2002) et dans Kott (2009, section 4.2). Des renseignements détaillés sur la façon dont cela peut être fait sont fournis aux sections 4.2 et 5.

4.2 Estimation et estimation de la variance sous calage en deux étapes

À la présente sous-section, nous commençons par décrire un estimateur par calage en deux étapes assez général d'un total, puis nous abordons l'estimation de sa variance. La première étape de pondération par calage, qui est effectuée sur l'échantillon original, emploie $\mathbf{x}_{1k}$ comme vecteur des variables du modèle de réponse et $\mathbf{z}_{1k}$ comme vecteur de calage. Chacun possède P_1 composantes. La fonction d'ajustement des poids est de la forme décrite à l'équation (2.4) où $\mathbf{g}_1$ remplace maintenant $\mathbf{g}$. L'équation de calage est $\sum_R d_k \alpha\left(\mathbf{g}_1^T \mathbf{x}_{1k}\right) \mathbf{z}_{1k} = \sum_S d_k \mathbf{z}_{1k}$.

La deuxième étape de la pondération par calage, qui est effectuée sur la population, emploie $\mathbf{x}_{2k}$ et $\mathbf{z}_{2k}$, chacun ayant P_2 composantes. Le biais de non-réponse sous le modèle de réponse est éliminé à la première étape. Comme fonction d'ajustement des poids pour la deuxième étape, nous proposons d'utiliser

$$h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right) = \frac{\ell_k + \exp\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right)}{1 + \exp(\mathbf{g}_2^T \mathbf{x}_{2k})/u_k}, \tag{4.1}$$

où l'on peut fixer $u_k > \ell_k > 0$ presque à sa guise (mais voir plus bas). Le deuxième membre de l'équation (4.1) peut varier sur les unités k (et peut donc dépendre de d_k et α_k), pourtant $h_k(0) = h_k'(0) = 1$, ce qui la rend asymptotiquement indistinguable de la fonction linéaire : $1 + \mathbf{g}_2^T \mathbf{x}_{2k}$. Pour simplifier, nous désignerons $h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right)$ et $h_k'\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right)$, par h_k et h_k', respectivement. Du point de vue d'un quasi-plan d'échantillonnage, les deux fonctions sont asymptotiquement identiques à l'unité. La deuxième équation de calage est $\sum_S d_k h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right) \mathbf{z}_{2k} = \sum_U \mathbf{z}_{2k}$. Comme cette équation doit être vérifiée, il existe des limites aux choix disponibles pour u_k et ℓ_k dans l'équation (4.1).

Un bon estimateur simultané des variances pour $t_y = \sum_R w_k y_k = \sum_R d_k \alpha\left(\mathbf{g}_1^T \mathbf{x}_{1k}\right) h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right) y_k$ est (comme nous le verrons)

$$\begin{aligned} v\left(t_y\right) = & \sum_{k,j \in S} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right)\left[d_k\left(\mathbf{z}_{1k}^T \mathbf{b}_1 + \alpha_k h_k e_{1k}\right)\right]\left[d_j\left(\mathbf{z}_{1j}^T \mathbf{b}_1 + \alpha_j h_j e_{1j}\right)\right] \\ & + \sum_{k \in R} d_k\left(h_k^2 \alpha_k^2 - h_k \alpha_k\right) e_{1k}^2, \end{aligned} \tag{4.2}$$

où

$$e_{2k} = y_k - \mathbf{z}_{2k}^T \left(\sum_S d_j \alpha_j h_j' \mathbf{x}_{2j} \mathbf{z}_{2j}^T\right)^{-1} \sum_S d_j \alpha_j h_j' \mathbf{x}_{2j} y_j, \tag{4.3}$$

$$\mathbf{b}_1 = \left(\sum_S d_f \alpha_f' \mathbf{x}_{1f} \mathbf{z}_{1f}^T\right)^{-1} \sum_S d_f \alpha_f' h_f \mathbf{x}_{1f} e_{2f}, \tag{4.4}$$

et

$$e_{1k} = e_{2k} - \mathbf{x}_{1k}^T \mathbf{b}_1. \tag{4.5}$$

Soit maintenant $\mathbf{x}_k$ le vecteur composé des composantes non en double de $\mathbf{x}_{1k}$ et $\mathbf{x}_{2k}$, et définissons $\mathbf{z}_k$ de manière analogue. Les conditions suffisantes pour que (4.2) soit un estimateur simultané des variances comprennent les composantes correspondantes de l'équation (4.1) selon que le modèle de réponse de l'équation (2.4) est vérifié avec $\mathbf{x}_{1k}$ remplaçant $\mathbf{x}_k$ ou que le modèle de prédiction est $\mathrm{E}(y_k | \mathbf{x}_k, \mathbf{z}_k) = \mathbf{z}_{2k}^T \boldsymbol{\beta}_2$, que l'unité k soit ou non échantillonnée ou réponde ou non si elle est échantillonnée, et les $\varepsilon_{2k} = y_k - \mathbf{z}_{2k}^T \boldsymbol{\beta}_2$ sont des variables aléatoires non corrélées de variances égales à $\sigma_{2k}^2 = \mathbf{z}_{2k}^T \boldsymbol{\eta}_2$, où $\boldsymbol{\eta}_2$ ne doit pas être spécifié outre le fait que ses composantes doivent être finies. Maintenant, $N^{-1} \sum_R d_k \alpha' (\mathbf{g}_1^T \mathbf{x}_{1k}) \mathbf{z}_{1k} \mathbf{x}_{1k}^T$ ainsi que $N^{-1} \sum_R d_k h_k' (\mathbf{g}_2^T \mathbf{x}_{2k}) \mathbf{z}_{2k} \mathbf{x}_{2k}^T$ sont considérées comme étant de plein rang et bornées quand la taille de l'échantillon devient arbitrairement grande.

L'estimateur de variance donné par l'équation (4.2) est presque le même que l'estimateur donné en (3.1) : $\mathbf{x}_k$ a été remplacé par $\mathbf{x}_{1k}$ et $\mathbf{z}_k$, par $\mathbf{z}_{1k}$, tandis que $h_k e_{2k}$ se substitue à y_k (nous parlerons sous peu d'une petite différence). Observons que e_{2k} est effectivement une expression du « résidu » de la deuxième étape de pondération par calage. Ce résidu est multiplié par la fonction d'ajustement des poids h_k, qui est asymptotiquement égale à l'unité dans la perspective fondée sur le quasi-plan d'échantillonnage et à une constante du point de vue du modèle de prédiction. Le produit est alors utilisé pour créer le « coefficient de régression » de la première étape $\mathbf{b}_1$ dans l'équation (4.4) et ses « résidus » connexes e_{1k} dans l'équation (4.5). Nous effectuons la régression de la deuxième étape pour commencer, parce que $t_y - T_y = \sum_R w_k y_k - \sum_U y_k = \sum_R w_k e_{2k} - \sum_U e_{2k}$.

C'est pour estimer le modèle de prédiction de t_y en tant qu'estimateur de $T_y, \sum_S (w_k^2 - w_k) \sigma_{2k}^2$, que la dernière apparition de h_k dans le deuxième membre de l'équation (4.2) n'est pas élevée au carré, comme elle le serait si $h_k e_{2k}$ se substituait à y_k partout. Du point de vue d'un quasi-plan, h_k est asymptotiquement identique à l'unité, de sorte que, qu'elle soit élevée au carré ou non ne fait asymptotiquement aucune différence.

Notons que les h_j' ont été insérées dans l'équation (4.3) pour la même raison que α' a été inséré dans $\mathbf{b}$ dans l'équation (3.1). Cependant, comme les h_j' sont asymptotiquement égales à l'unité, elles ne sont pas vraiment nécessaires (et ne remplissent aucune fonction du point de vue d'un modèle de prédiction). Un argument similaire s'applique aux h_f dans l'équation (4.4) : elles sont asymptotiquement égales à l'unité du point de vue du quasi-plan d'échantillonnage (et font partie d'une estimation de 0 du point de vue du modèle de prédiction).

5 Quelques simulations

Comme dans Kott et Liao (2012), nous avons créé une population synthétique, U, d'hôpitaux à partir du fichier de données à grande diffusion DAWN de 2008. Après avoir créé U, nous avons tiré indépendamment 3 600 échantillons aléatoires simples stratifiés de taille 400 de U en utilisant les

définitions des strates du fichier de données à grande diffusion. Ces définitions incorporent l'information sur l'emplacement et la propriété de l'hôpital (publique ou privée) qui n'est pas fournie directement dans le fichier.

Nous avons fixé les tailles des échantillons de strate de façon qu'elles soient approximativement proportionnelles à une mesure de taille q_k, mais jamais inférieures à quatre. Pour q_k, nous avons utilisé le nombre annuel de visites au service d'urgence associées à la consommation de drogues, qui était toujours positif. Dans le fichier DAWN, une variable de taille est en fait associée à chaque hôpital figurant dans la base de sondage, à savoir le nombre de visites au service d'urgence durant une année antérieure selon l'*American Hospital Association.* Malheureusement, cette variable n'était pas incluse dans le fichier de données à grande diffusion. Dans nos simulations, les poids de sondage variaient entre 4,375 et 48, ce qui nous a permis de traiter les facteurs de correction pour population finie comme étant ignorables dans l'estimation de la variance.

Comme dans notre article original, nous avons généré un échantillon de répondants R pour chaque échantillon simulé selon un tirage de Bernoulli à partir de la fonction logistique :

$$p_k = (1 + \exp(3{,}735 - 0{,}4\log(q_k)))^{-1}, \tag{5.1}$$

Nous avons également créé des échantillons de répondants de rechange en utilisant

$$p_k = \left(1 + \exp\left(0{,}597 - 0{,}005q_k^{1/2}\right)\right)^{-1}. \tag{5.2}$$

Les modèles de réponse ont tous deux produit des taux de réponse globaux non pondérés d'environ 54 %, ce qui est similaire à la situation réelle du fichier DAWN, où la réponse prend aussi la forme d'une fonction légèrement croissante de la variable de taille. Notons que $\alpha_k = 1/p_k$ est borné même si ni l'une ni l'autre probabilité ne peut être exprimée par l'équation (2.4) avec une borne supérieure u finie.

Comme dans l'étude précédente, nous nous sommes concentrés sur l'estimation des totaux de population pour trois variables étudiées. Les nombres annuels de visites au service d'urgence liées à la consommation de drogues avec réaction pharmaceutique indésirable et de celles résultant en un décès ont été extraits du fichier de données à grande diffusion. Puisque ces variables étaient approximativement linéaires en notre mesure de taille, la troisième variable « étudiée » a été construite artificiellement. Il s'agissait de la mesure de taille (nombre de visites annuelles au service d'urgence liées à la consommation de drogues) élevée à la puissance 1,3.

Nous avons étudié huit estimateurs et estimations de leur variance. Les résultats sont résumés au tableau 5.1. Les deux premiers comportaient le calage sur l'échantillon original seulement (équation (2.5) avec $\theta = 1$), en supposant que la réponse était de forme logistique en le logarithme de la mesure de taille. Nous avons employé l'équation (2.3) avec $\mathbf{x}_k = (1 \ \log(q_k))^T$. Le premier estimateur utilisait $\mathbf{z}_k = (1 \ \log(q_k))^T$ comme vecteur de calage, tandis que le deuxième utilisait $\mathbf{z}_k = (1 \ q_k)^T$, qui était davantage en harmonie avec un modèle de prédiction raisonnable, du moins pour les réactions indésirables et les décès.

Nos troisième et quatrième estimateurs comportaient le calage sur l'échantillon et sur la population en une seule étape (équation (2.5) avec $\theta = 1$, puis $\theta = 0$) en utilisant $\mathbf{x}_k = \mathbf{z}_k = (1 \ \log(q_k) \, q_k)^T$. Ils

étaient conçus pour être quasiment sans biais si le modèle de réponse logistique en $(1\ \log(q_k))^T$ ou le modèle de prédiction linéaire en $(1\ q_k)^T$ étaient vérifiés.

Tableau 5.1
Sommaire de l'exercice de simulation (tous les résultats sont exprimés en pourcentage %)

Estimateur	t_{y1}	t_{y2}	t_{y3}	t_{y4}	t_{y5}	t_{y6}	t_{y7}	t_{y8}
Calage sur l'échantillon								
Variables du modèle de réponse : $\mathbf{x}_{1k}$	$(1\ \log(q_k))^T$	$(1\ \log(q_k))^T$	$(1\ \log(q_k)\ q_k)^T$	-	$(1\ \log(q_k))^T$	$(1\ \log(q_k))^T$	$(1\ \log(q_k))^T$	$(1\ \log(q_k))^T$
Variables de calage : $\mathbf{z}_{1k}$	$(1\ \log(q_k))^T$	$(1\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	-	$(1\ \log(q_k))$	$(1\ q_k)^T$	$(1\ \log(q_k))^T$	$(1\ q_k)^T$
Calage sur la population								
Variables du modèle de réponse : $\mathbf{x}_{2k}$	-	-	-	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$f_k\ (1\ \log(q_k)\ q_k)^T$	$f_k\ (1\ \log(q_k)\ q_k)^T$
Variables de calage : $\mathbf{z}_{2k}$	-	-	-	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$
Réponse vraie : $p_k = 1/\{1 + \exp[3{,}735 + 0{,}4 \log(q_k)]\}$								
Réactions indésirables								
Biais relatif de t_y	-0,07	0,06	-0,11	-0,13	-0,02	-0,07	0,10	0,09
REQM relative de t_y	4,97	3,98	4,01	2,45	2,51	2,57	2,40	2,39
Biais relatif de $v(t_y)$	8,60	12,59	12,52	6,24	6,76	6,16	6,76	6,48
Décès								
Biais relatif de t_y	-0,17	0,06	-0,20	-0,26	-0,20	-0,30	0,04	-0,07
REQM relative de t_y	11,75	11,39	11,56	11,07	11,28	11,36	10,91	10,91
Biais relatif de $v(t_y)$	-1,34	-0,48	-0,90	-0,76	-1,00	-0,60	-0,12	-0,28
$(\text{Taille})^{1,3}$								
Biais relatif de t_y	-0,16	-0,05	0,08	0,09	0,04	0,06	-0,02	0,01
REQM relative de t_y	6,92	5,07	5,06	0,95	1,05	1,12	0,89	0,89
Biais relatif de $v(t_y)$	10,01	18,49	17,47	-2,26	-3,41	-3,32	0,51	-2,12
Réponse vraie : $p_k = 1/\{1 + \exp[0{,}597 + 0{,}005\ q_k^{1/2}]\}$								
Réactions indésirables								
Biais relatif de t_y	2,87	-0,26	0,08	0,04	0,48	0,53	0,15	0,07
REQM relative de t_y	5,90	3,97	4,00	2,35	2,43	2,45	2,33	2,35
Biais relatif de $v(t_y)$	-18,22	11,63	11,95	9,90	8,82	7,35	7,19	6,67
Décès								
Biais relatif de t_y	1,24	-1,88	0,47	0,36	1,03	1,20	-0,58	-0,67
REQM relative de t_y	11,42	11,01	11,41	10,95	11,18	11,26	10,69	10,72
Biais relatif de $v(t_y)$	5,30	3,00	6,27	6,24	5,65	5,06	6,21	5,90
$(\text{Taille})^{1,3}$								
Biais relatif de t_y	5,17	1,05	-0,07	-0,05	-0,31	-0,36	0,01	0,08
REQM relative de t_y	9,11	5,31	5,05	0,85	0,97	1,01	0,80	0,82
Biais relatif de $v(t_y)$	-26,83	11,70	17,09	8,23	0,29	-3,98	5,17	2,90

$f_k = d_k \alpha_k - 1 = (d_k / \hat{p}_k) - 1$

Il n'est pas surprenant de constater que l'erreur quadratique moyenne relative (empirique) du quatrième estimateur est toujours plus faible que celle du troisième. La raison en est assez évidente si l'on examine l'équation (3.1) et que l'on considère la conséquence du fait que θ est égal à 0 (calage sur la population) plutôt qu'à 1 (calage sur l'échantillon).

Les cinquième à huitième estimateurs ont été calés en deux étapes. Pour les cinquième et septième estimateurs, on a employé la pondération par calage utilisée pour le premier estimateur à la première étape, tandis que pour les sixième et huitième, on a employé la pondération par calage du deuxième estimateur. Pour les cinquième et sixième estimateurs, on a utilisé $\mathbf{z}_{2k} = \mathbf{x}_{2k} = (1 \ \log(q_k)\, q_k)^T$ à la deuxième étape, tandis que les septième et huitième étaient quasi pseudo-optimaux (Kott 2011) en utilisant $\mathbf{z}_{2k} = (1 \ \log(q_k)\, q_k)^T$ et $\mathbf{x}_{2k} = (d_k \alpha_k - 1)\mathbf{z}_{2k}$ à la deuxième étape. Pour les quatre estimateurs, on a employé les fonctions d'ajustement des poids individuels suivantes :

$$h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right) = \frac{1}{d_k \alpha_k} + \left(1 - \frac{1}{d_k \alpha_k}\right) \exp\left[\frac{\mathbf{g}_2^T \mathbf{x}_{2k}}{1 - \dfrac{1}{d_k \alpha_k}}\right].$$

Comme l'a montré Kott (2011), ces $h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right)$ sont asymptotiquement identiques à la fonction d'ajustement des poids, $1 + \mathbf{g}_2^T \mathbf{x}_{2k}$, quand $\mathbf{g}_2^T \mathbf{x}_{2k} = \mathrm{O}_P\left(1/\sqrt{n}\right)$, mais empêchent tout poids w_k de devenir inférieur à l'unité. Chacune est une version de l'équation (4.1) avec $\ell_k = 1/(d_k \alpha_k), c = 1,$ et $u = \infty.$

Comme le taux de non-réponse n'était pas élevé, nous n'avons pas eu de problème à calculer les troisième et quatrième estimateurs quel qu'était l'échantillon de répondants simulés utilisé. L'erreur quadratique moyenne relative du quatrième estimateur était systématiquement légèrement plus grande que celle des septième et huitième estimateurs, dans lesquels était incorporé un calage quasi pseudo-optimal à la deuxième étape. Curieusement, cela n'était pas le cas pour la comparaison du quatrième estimateur aux cinquième et sixième estimateurs qui, bien que comprenant les deux étapes, n'intégraient pas le calage quasi pseudo-optimal.

Il convient de souligner que, même si le deuxième estimateur possédait systématiquement une plus petite erreur quadratique moyenne relative que le premier, du fait qu'il était davantage en harmonie avec un modèle de prédiction raisonnable (même pour $q_k^{1,3}$, la variable étudiée paraissait plus près d'être linéaire en q_k qu'en $\log(q_k)$), les autres paires analogues (cinquième c. sixième et septième c. huitième) ne présentaient aucun schéma évident de supériorité. Cela tient au fait que ce sont les résidus de la deuxième étape qui sont effectivement modélisés dans l'équation (4.4) et non les valeurs de y.

La production de la non-réponse au moyen de l'équation (5.2) plutôt que (5.1) ne semble pas avoir beaucoup d'effet sur les résultats, sauf en ce qui concerne les biais relatifs du premier estimateur. Tant pour les réactions indésirables que pour la $(\text{taille})^{1,3}$, le biais relatif de cet estimateur est supérieur à 40 % de l'erreur quadratique moyenne relative. Il en est vraisemblablement ainsi parce que les deux modèles qui pouvaient être utilisés pour justifier cet estimateur (la réponse est logistique en le logarithme de la mesure de taille et la variable étudiée est linéaire en le logarithme de la mesure de taille) n'ont pas tenu. Il n'est donc pas étonnant, puisque le biais relatif représente une telle part de l'erreur quadratique moyenne relative dans ces deux situations, que $v(t_k)$ sous-estime fortement l'erreur quadratique moyenne. Nulle part ailleurs le biais relatif de $v(t_k)$ n'est supérieur à 15 %.

Il semble que même notre variable artificielle, $(\text{taille})^{1,3}$, s'approchait suffisamment de la linéarité en la mesure de taille pour que le biais ne soit jamais un problème pour tout autre estimateur que le premier.

Le premier estimateur lui-même avait un biais relatif négligeable quand la réponse était un modèle logistique du logarithme de la mesure de taille, comme on le suppose.

6 Conclusion

À la section 4, nous avons mentionné deux raisons de préférer la pondération par calage en deux étapes : rendre l'ajustement implicite d'un modèle de réponse logistique plus facile et intégrer le calage presque quasi-optimal. Un avantage secondaire du calage en deux étapes est une estimation plus efficace du modèle de réponse à la première étape, puisque aucune erreur d'échantillonnage ne fausse l'estimation. Cette propriété est utile si l'on veut analyser les causes de la non-réponse totale en tant que fin en soi.

Nous concédons, cependant, que la réduction de l'erreur quadratique moyenne en utilisant les deux étapes était modeste dans nos expériences par simulation à la section 5. En outre, nous ne pouvons nier l'attrait pratique de la simplicité du calage en une seule étape.

Lorsqu'on utilise la pondération par calage pour corriger la non-réponse quand les réponses ne manquent pas au hasard comme il est décrit dans Chang et Kott (2008) et dans Kott et Chang (2010), des gains d'efficacité vraisemblablement importants découlent d'une deuxième étape où n'interviennent que des variables de calage et des fonctions des variables de calage comme variables du modèle.

Quand les facteurs de correction pour population finie peuvent être ignorés, le rééchantillonnage offre une approche beaucoup plus simple d'estimation de la variance que l'équation (3.7), même si l'on peut laisser tomber la deuxième sommation dans le deuxième membre dans cette situation. Une autre option intéressante est la version « contractée » de l'équation (4.2) qui ignore l'effet de la première étape de calage :

$$\tilde{v}\left(t_y\right) = \sum_{k,j\in S}\left(1-\frac{\pi_k\pi_j}{\pi_{kj}}\right)\left[w_k e_{2k}\right]\left[w_j e_{2j}\right] + \sum_{k\in R} d_k\left(h_k^2\alpha_k^2 - h_k\alpha_k\right)e_{2k}^2.$$

Cet estimateur estime manifestement la variance du modèle de prédiction si ce modèle est vérifié. Une version de cet estimateur – avec la deuxième sommation supprimée – a donné de bons résultats dans nos expériences par simulation (résultats non présentés). Une certaine prudence est de rigueur avant de tirer une conclusion trop catégorique de ce résultat, puisque le modèle linéaire n'était jamais très loin d'être vérifié dans nos investigations.

Enfin, un certain nombre d'hypothèses ont été faites pour simplifier l'exposé. Le lecteur que cela intéresse peut étendre les résultats à une d_k non bornée ou à des fonctions d'ajustement des poids plus générales et qui ne sont pas nécessairement bornées, ou permettre que les erreurs du modèle de prédiction soient corrélées à l'intérieur des unités primaires d'échantillonnage. Quand N augmente plus rapidement que n, l'hypothèse selon laquelle $\sigma_k^2 = \mathbf{z}_k^T\boldsymbol{\eta}$ peut parfois être abandonnée. Voir, par exemple, Kott (2009, page 69).

Remerciements

Le présent article a été préparé à l'occasion du *Symposium on the Analysis of Survey Data and Small Area Estimation* organisé en l'honneur du 75e anniversaire du professeur J.N.K. Rao et parrainé par le

Fields Institute for Research in Mathematical Sciences. Les auteurs remercient les organisateurs de la conférence de les avoir invités à présenter cet article et l'Institut de son généreux financement de la conférence sans lequel le présent article n'aurait jamais été rédigé. Ils remercient également plusieurs rédacteurs et examinateurs de leurs commentaires utiles.

Bibliographie

Bang, H., et Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-972.

Bankier, M. (2002). Regression estimators for the 2001 Canadian Census. Présenté à l'International Conference in Recent Advances in Survey Sampling.

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

Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal total are known. *Annals of Mathematical Statistics,* 11, 427-444.

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. Dans *COMPSTAT: Proceedings in Computational Statistics, 14th Symposium, Utrecht, The Netherlands,* (Éds., J.G. Bethlehem et P.G.M. Van der Heidjen), Heidelberg : Physica Verlag, 65-76.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 87, 418, 376-382.

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association,* 88, 1013-1020.

Estevao, V.M., et Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics,* 18, 233-255.

Estevao, V.M., et Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the American Statistical Association, Social Statistics Section,* 197-202.

Folsom, R.E., et Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the American Statistical Association, Survey Research Methods Section,* disponible en ligne au http://www.amstat.org/sections/srms/Proceedings/, 598-603.

Fuller, W.A., Loughin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête,* 20, 1, 79-89.

Kim, J.K., et Haziza, D. (2014). Doubly robust inference with missing survey data. *Statistica Sinica,* 24, 375-394.

Kim, J.K., et Park, H. (2006). Imputation using response probability. *Canadian Journal of Statistics,* 34, 1-12.

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

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête,* 32, 2, 149-160.

Kott, P.S. (2009). Calibration weighting: Combining probability samples and linear prediction models. Dans *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis,* (Éds., D. Pfeffermann et C.R. Rao), New York : Elsevier.

Kott, P.S. (2011). A nearly pseudo-optimal method for keeping calibration weights from falling below unity in the absence of nonresponse or frame errors. *Pakistan Journal of Statistics,* 27, 391-396.

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

Kott, P.S., et Liao, D. (2012). Comparing weighting methods when adjusting for logistic unit Nonresponse. Présenté au Federal Committee on Survey Methodology Research Conference, disponible en ligne au http://www.fcsm.sites.usa.gov/files/2014/05/Kott_2012FCSM_III-B.pdf.

Little, R.J., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2e Éd.), New York : John Wiley & Sons, Inc.

Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics,* 15, 305-327.

Oh, H.L., et Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys,* (Éds., W.G. Madow, I. Olkin et D.B. Rubin), New York : Academic Press, 2.

Rao, J.N.K. (1994). Estimation of totals and distributing functions using auxiliary information at the estimation stage. *Journal of Official Statistics,* 10, 153-165.

Robins J.M., Rotnitzky A. et Zhao L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association,* 89, p. 846-866.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 191-204
Statistique Canada, n° 12-001-X au catalogue

La pertinence du suivi dans la collecte des données pour le système d'assurance de la qualité du Recensement de la population et du logement du Portugal

Paula Vicente, Elizabeth Reis et Álvaro Rosa[1]

Résumé

La mise en œuvre des opérations du Recensement de la population et du logement du Portugal est gérée par une structure hiérarchique dans laquelle Statistique Portugal se situe au sommet et les institutions gouvernementales locales, à la base. Quand le recensement a lieu, tous les 10 ans, Statistique Portugal demande aux administrations locales de collaborer avec lui à l'exécution et à la surveillance des opérations sur le terrain au niveau local. À l'étape de l'essai pilote du Recensement de 2011, on a demandé aux administrations locales une collaboration supplémentaire, à savoir répondre à un sondage sur la perception du risque, qui avait pour objectif de recueillir des renseignements en vue de concevoir un instrument d'assurance de la qualité pour surveiller les opérations du recensement. Le taux de réponse espéré au sondage était de 100 %, mais à l'échéance de la collecte des données, près du quart des administrations locales n'avaient pas répondu et il a donc été décidé de procéder à un suivi par la poste. Dans le présent article, nous examinons si nous aurions pu tirer les mêmes conclusions sans le suivi qu'avec celui-ci, et nous évaluons son influence sur la conception de l'instrument d'assurance de la qualité. La comparaison des réponses pour un ensemble de variables de perception a révélé que les réponses des administrations locales avant ou après le suivi ne différaient pas. Cependant, la configuration de l'instrument d'assurance de la qualité a changé lorsque l'on a inclus les réponses au suivi.

Mots-clés : Assurance de la qualité; sondages auprès des administrations locales; suivis; carte d'alerte.

1 Introduction

Le dernier Recensement de la population et du logement du Portugal a eu lieu en mars 2011. Il s'agissait d'une opération statistique coûteuse et de grande envergure comportant des prises de contact en personne, porte à porte, pour la distribution et la collecte des questionnaires imprimés partout dans le pays. La tâche principale de toute opération de recensement est de dénombrer tous les habitants et de déterminer où ils vivent, sans omettre personne (Waite 2007). Cependant, le bon accomplissement d'une telle tâche peut être compromis par divers facteurs, notamment la performance des ressources humaines participantes, le degré de coopération des citoyens, et les caractéristiques particulières des régions et des populations qui sont dénombrées. Des données fiables ne peuvent être obtenues qu'au moyen de processus valables et rigoureux, raison pour laquelle le recensement s'appuie sur un système d'assurance de la qualité (AQ) complet qui est conçu et mis en œuvre tout au long des opérations de recensement proprement dites (Wroth-Smith, Abbott, Compton et Benton 2011).

Avant 2011, le système d'AQ des opérations de recensement s'appuyait sur des procédures nationales normalisées, c'est-à-dire des normes, des indicateurs, des processus et des sous-processus définis au niveau national, si bien que toutes les régions utilisaient les mêmes activités d'AQ pour les besoins de la surveillance. Bien que le Portugal soit un petit pays, il présente une grande diversité géographique et démographique caractérisée par des régions fortement urbanisées ainsi que des régions rurales, des régions à très forte densité de population ainsi que des villages quasiment abandonnés et déserts, de même que des

1. Paula Vicente, Elizabeth Reis et Álvaro Rosa, Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisbonne, Portugal, ISCTE-IUL, Av. Forças Armadas, 1649-026 Lisbonne, Portugal. Courriel : paula.vicente@iscte.pt.

régions habitées principalement par des personnes âgées et d'autres, par des personnes beaucoup plus jeunes. Cette diversité est susceptible d'avoir une incidence sur la mise en œuvre d'un recensement, car les problèmes, les difficultés et le risque d'échec ne sont pas uniformes, mais varient plutôt en fonction des caractéristiques particulières de la population et des régions où le recensement est effectué. Compte tenu de cette situation, une nouvelle perspective a été adoptée pour le Recensement de 2011 : le système d'AQ a été remanié afin de l'adapter aux caractéristiques locales particulières des régions géographiques et des populations (Statistics Portugal 2007). Ce changement a rendu nécessaire l'établissement de la carte des risques d'échec sur le territoire du Portugal, ce qui a mené à l'élaboration d'une carte d'alerte (*Map of Alert*) (Statistics Portugal 2010).

Le Portugal est subdivisé administrativement en 303 municipalités englobant 4 260 *freguesias* (administrations locales). (La *freguesia* est la plus petite région administrative/gouvernementale au Portugal. Chaque municipalité comprend un ensemble de *freguesias*. La *freguesia* est l'équivalent d'une paroisse civile.) Cette organisation sert de base à la mise en œuvre des opérations du recensement : la *freguesia* est le niveau le plus bas de la hiérarchie de coordination des opérations; au-dessus d'elle vient la coordination municipale, puis la coordination régionale et, enfin, la coordination nationale au sommet. Le Bureau du recensement de Statistique Portugal est chargé de la coordination stratégique et nationale des opérations complètes. Statistique Portugal nomme des délégués régionaux chargés de la coordination régionale; les présidents des municipalités sont responsables de la coordination municipale, et enfin, les présidents des *junta de freguesia* (PJF). Les PJF sont chargés de la coordination au niveau de la *freguesia* (le *junta de freguesia* est l'organe directeur de chaque *freguesia* et est administré par le président du *junta de freguesia*).

La carte d'alerte est une carte détaillée du territoire portugais au niveau des *freguesias*, dans laquelle un code de couleur est attribué à chaque *freguesia* pour indiquer le risque potentiel d'échec des opérations de recensement : rouge (risque élevé), orange (risque moyen) et vert (risque faible). Par risque d'échec, nous entendons d'éventuels problèmes de couverture, c'est-à-dire omettre de dénombrer certaines personnes ou en dénombrer d'autres en double. Dresser la carte des 4 260 *freguesias* en fonction de leur risque d'échec devrait permettre aux coordonnateurs municipaux de savoir d'avance quelles *freguesias* nécessiteront des activités d'AQ particulières afin de soutenir efficacement les opérations sur le terrain. Cela permettrait de cibler les ressources sur les *freguesias* que l'on sait poser un risque élevé d'échec des opérations. Donc, les *freguesias* vertes ou oranges pourraient être traitées en appliquant les procédures d'AQ standard, tandis que des procédures spéciales seraient conçues et mises en œuvre pour répondre aux caractéristiques locales particulières des *freguesias* rouges. Ces procédures spéciales pourraient inclure l'affectation d'agents recenseurs plus chevronnés aux régions les plus difficiles, le contrôle plus régulier du travail des agents recenseurs, ou la vérification du travail d'une proportion d'agents recenseurs supérieure aux 5 % habituels.

L'information sur les caractéristiques des populations, des logements et des régions susceptibles de causer des difficultés de couverture pour le recensement (par exemple, l'existence de sans-logis, de personnes appartenant à des groupes minoritaires ou de régions comptant de nombreux logements inoccupés (Groves 1989, page 137, Groves et Couper 1998, page 176)) était nécessaire pour dresser la carte d'alerte. Ce genre d'information aurait pu être tirée des données du Recensement de 2001, mais comme celles-ci risquaient d'être dépassées, on a décidé de recueillir l'information requise au moyen d'un sondage expédié par la poste à tous les PJF. Il était essentiel d'obtenir la coopération de chacun des

4 260 PJF afin d'être certains que chaque *freguesia* soit classée selon un niveau de risque dans la carte d'alerte.

Le questionnaire du sondage sur la perception du risque a été envoyé au début d'octobre 2010. L'échéance pour la collecte des données établie à l'interne par l'équipe de recherche était le milieu de décembre 2010, mais comme les répondants ont tendance à remettre à plus tard la réponse aux sondages par la poste, on leur a demandé de retourner le questionnaire dûment rempli dans le mois suivant sa réception. Plus de la moitié des *freguesias* (58 %) l'ont fait dans ce délai; après cette période, les réponses ont continué d'arriver, mais à un rythme plus lent. À l'échéance, 77 % des *freguesias* avaient retourné le questionnaire, mais le nombre de questionnaires arrivant à la fin avait baissé fortement. Malgré le bon taux de réponse (Dillman, Smyth et Christian 2009), l'objectif visant à obtenir les données auprès de toutes les *freguesias* était loin d'être atteint. Terminer la collecte des données à la mi-décembre aurait signifié qu'un niveau de risque ne serait pas attribué à près du quart des *freguesias*, réduisant ainsi considérablement l'efficacité de la carte d'alerte en tant qu'instrument d'assurance de la qualité. Un envoi par la poste de suivi a donc été adressé aux *freguesias* non répondantes le 16 décembre. En plus d'accroître la taille de l'échantillon, on espérait améliorer la qualité de l'information pour la conception de la carte d'alerte. En fait, on craignait que les non-répondants soient des *freguesias* dont les caractéristiques posaient problème pour le recensement, de sorte que l'ampleur réelle du code rouge serait sous-représentée dans la carte. Il est bien connu que la demande de renseignements personnels ou délicats dans les questionnaires augmente le risque de non-réponse (par exemple, Groves, Fowler Jr, Couper, Lepkowski, Singer et Tourangeau 2004, page 224) et, même si les renseignements demandés dans le sondage sur la perception du risque n'étaient pas de nature personnelle (c'est-à-dire concernant le PJF lui-même), ils communiquaient des informations que les PJF pourraient hésiter à partager. Ceux-ci pourraient juger les questions sur l'existence de sans-logis, de zones sans éclairage public, ou de chaussées non revêtues de tarmacadam dans les régions qu'ils gouvernent comme exagérément délicates, et donc entraîner leur non-participation au sondage. L'envoi par la poste de suivi avait aussi pour objectif de réduire au minimum l'effet de la non-réponse sur la classification des *freguesias* selon le niveau de risque.

La carte d'alerte a été utilisée pour la première fois dans le cadre du Recensement de la population et du logement du Portugal de 2011, mais Statistique Portugal a l'intention de l'adopter comme instrument d'AQ permanent pour les futures opérations de recensement. L'étude décrite dans le présent article examine l'effet du suivi sur le taux de réponse et les résultats du sondage sur la perception du risque, et évalue la mesure dans laquelle les réponses au suivi ont modifié la configuration de la carte d'alerte, c'est-à-dire la classification selon le niveau de risque.

La méthode est présentée à la section 2. Les résultats sont donnés à la section 3. Enfin, une discussion est fournie à la section 4.

2 Méthode

Le sondage sur la perception du risque a eu lieu à l'étape de l'essai pilote du Recensement de la population et du logement du Portugal de 2011 (l'essai pilote, qui était la dernière étape préparatoire du Recensement de 2011, s'est déroulé pendant presque l'entièreté de 2010). L'objectif était de recueillir des renseignements sur les caractéristiques particulières des *freguesias* susceptibles d'entraver le

dénombrement exhaustif et exact des personnes et des logements. Le sondage avait pour population cible les *freguesias* du Portugal (N = 4 260). Les présidents des *juntas de freguesia* ont été choisis comme répondants, parce qu'ils sont en contact étroit avec la population et ont une connaissance approfondie des problèmes de la région qu'ils gouvernent.

Le questionnaire comprenait deux modules de questions (le questionnaire est présenté à la figure A.1 de l'annexe). Le premier module contenait les questions sur l'âge du répondant, son niveau d'études, la durée de l'occupation du poste de président du *junta de freguesia*, et la fréquence d'utilisation d'un ordinateur et d'Internet, ainsi que l'identification de la *freguesia* et de la municipalité. Le deuxième module comprenait des questions sur les caractéristiques des *freguesias* susceptibles d'avoir une incidence sur la mise en œuvre du recensement. Ce module comprenait quatre sections. La première section contenait un ensemble de six questions sur les caractéristiques de la population de la *freguesia*. Les répondants devaient se servir d'une échelle de cinq points variant de « peu » à « beaucoup » pour répondre à chacune des six questions. La deuxième section contenait un ensemble de six questions sur les caractéristiques des immeubles et des régions de la *freguesia*. De nouveau, une échelle de cinq points variant de « peu » à « beaucoup » devait être utilisée pour répondre à chacune des questions. La section suivante contenait deux questions sur le recrutement des agents recenseurs auxquelles il fallait répondre en se servant d'une échelle de cinq points variant de « difficile » à « facile ». Le questionnaire se terminait par une question sur la perception globale de la mise en œuvre du Recensement de 2011 dans la *freguesia*.

Statistique Portugal possède une liste à jour des adresses postales de tous les *Juntas de Freguesia* qui a été utilisée comme base de sondage. Le premier envoi par la poste a été adressé aux 4 260 PJF, de sorte que le sondage sur la perception du risque était davantage un recensement qu'un sondage. L'envoi par la poste comprenait un questionnaire, une enveloppe affranchie pour le retour du questionnaire et une lettre d'introduction. La lettre et le questionnaire étaient imprimés sur du papier à l'en-tête du Recensement de 2011 et de Statistique Portugal, responsable de la mise en œuvre et de la coordination du sondage. Puisque la réponse au sondage n'était pas obligatoire, l'importance de celui-ci a été soulignée dans la lettre d'invitation dans le but d'améliorer le taux de coopération (par exemple, Porter 2004, Dillman et coll. 2009) : la lettre expliquait que le sondage concernait les opérations du Recensement de 2011 et que les réponses des PJF étaient indispensables à la qualité de ces opérations au niveau tant local que national. En outre, l'importance de la réponse était mise en relief par le fait que la demande venait de Statistique Portugal.

À toutes les *freguesias* qui n'avaient pas retourné le questionnaire le 15 décembre 2010, on a adressé un envoi par la poste de suivi contenant un deuxième exemplaire du questionnaire, une lettre d'introduction insistant sur l'importance de la réponse et une enveloppe affranchie pour le retour du questionnaire. La collecte des données a pris fin à la mi-février 2011.

3 Résultats

Pour les besoins de l'analyse, nous considérons deux « groupes » de réponses : le groupe initial et le groupe final. Le groupe initial englobe les *freguesias* qui ont retourné le questionnaire avant la date du suivi; le groupe final comprend toutes les *freguesias* qui ont répondu au sondage, c'est-à-dire le groupe initial plus les *freguesias* qui ont retourné le questionnaire après le suivi. Les deux groupes ne sont pas mutuellement exclusifs.

L'analyse débute par une description des résultats de l'envoi par la poste. Nous examinons les taux de réponse (global et par région) et la répartition géographique des *freguesias* auxquelles a pu être attribué un niveau de risque (dans la version initiale ainsi que dans la version finale de la carte d'alerte). Pour les analyses par région, nous utilisons la classification NUTS II du territoire du Portugal, qui comporte six régions : le Nord, le Centre, Lisbonne, l'Alentejo, l'Algarve et l'Archipel de Madère et des Açores. À la deuxième étape, une analyse en composantes principales des réponses des PJF est réalisée dans le but de réduire la dimensionnalité des données et de déterminer les dimensions latentes du risque. Cette analyse est effectuée pour les deux groupes de réponses. Enfin, nous évaluons la classification des *freguesias* selon le niveau de risque dans la carte d'alerte initiale ainsi que dans la carte finale. Les *freguesias* qui n'ont pas répondu au sondage sur la perception du risque (appelées non-répondants) sont décrites en fonction de leur répartition géographique.

3.1 Analyse des taux de réponse

La figure 3.1 donne la répartition des nombres de questionnaires reçus par jour sur l'ensemble de la période de collecte (du 10 octobre 2010 quand les premiers questionnaires ont été reçus jusqu'à l'échéance finale du 16 février 2011). On observe deux pics de réponses, le premier environ un mois après le premier envoi par la poste et le deuxième, quelques jours après l'envoi par la poste de suivi. Presque aucun questionnaire n'était reçu au moment de l'envoi par la poste de suivi, ce qui porte à croire qu'aucun autre questionnaire n'aurait été reçu sans le deuxième envoi par la poste.

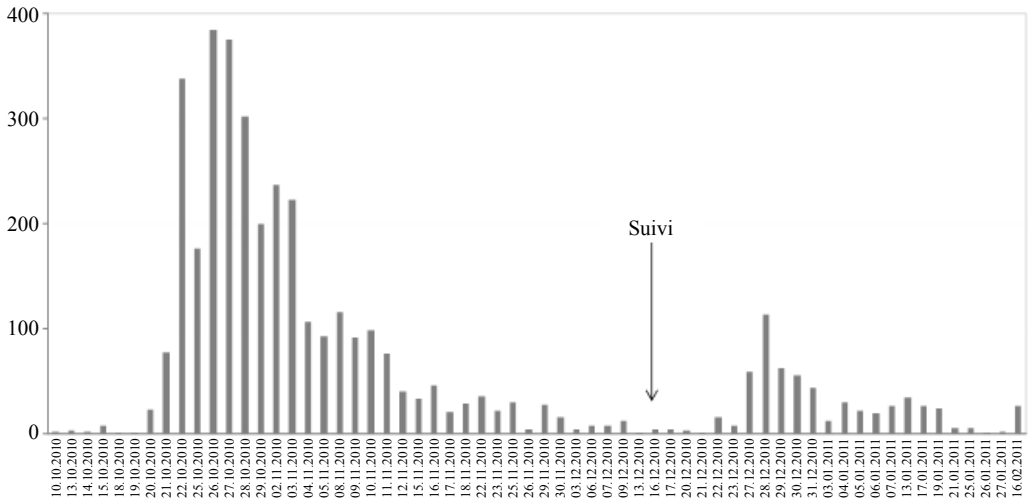


Figure 3.1 Nombre de questionnaires reçus par jour

D'un total de 4 260 questionnaires expédiés lors du premier envoi postal, 2 457 ont été retournés dûment remplis dans le délai recommandé (un mois), 816 ont été reçus après cette période, mais avant l'envoi par la poste de suivi, et 609 ont été retournés dûment remplis après la date de suivi. Des 4 260 *freguesias*, 378 n'ont pas répondu. Cette absence de réponse a été considérée comme un refus, puisqu'il est peu probable que ces questionnaires n'aient pas été livrés, car une liste d'adresses mise à jour a été utilisée pour l'expédition par la poste. Le taux de réponse global au sondage, calculé en pourcentage

de *freguesias* qui ont répondu au questionnaire sur le nombre total de *freguesias* dans la population, était de 91,1 % (tableau 3.1).

Tableau 3.1
Résultats de l'envoi des questionnaires par la poste

	N	%
Freguesias ayant retourné le questionnaire dans le mois suivant l'envoi	2 457	57,7
Freguesias ayant retourné le questionnaire après un mois et avant l'envoi par la poste de suivi	816	19,2
Freguesias ayant retourné le questionnaire après l'envoi par la poste de suivi	609	14,3
Freguesias n'ayant pas retourné le questionnaire	378	8,9
Questionnaires envoyés	4 260	100,0
Total des *freguesias* ayant retourné le questionnaire	3 882	91,1

Le tableau 3.2 donne le taux de réponse par région pour le groupe initial et le groupe final. Le taux de réponse à l'envoi par la poste initial variait de 71 % dans le Nord à 88,1 % dans l'Algarve; le taux de réponse final variait de 87,3 % dans le Nord à 96,4 % dans l'Algarve. L'envoi par la poste de suivi a permis un accroissement du taux de réponse global ainsi que du taux de réponse dans chaque région, mais il a été plus efficace dans le Nord que dans les autres régions. La participation au sondage a augmenté de 16,3 % dans le Nord, comparativement à environ 6 % dans la région de l'Archipel de Madère et des Açores.

Tableau 3.2
Taux de réponse par région selon le groupe de réponses (%)

Région	Initial	Final
Nord	71,0	87,3
Centre	79,3	91,4
Lisbonne	84,3	95,3
Alentejo	83,1	96,0
Algarve	88,1	96,4
Archipel de Madère et des Açores	86,7	93,3
Global	76,8	91,1

Le tableau 3.3 donne la répartition géographique des *freguesias* auxquelles un niveau de risque a été attribué dans la carte d'alerte initiale et dans la carte d'alerte finale. Plus de 40 % des *freguesias* sont situées dans le Nord et environ 26 % sont situées dans le Centre. Si l'on compare la répartition finale à celle de toutes les *freguesias* dans la population, les différences les plus importantes s'observent pour les régions de Lisbonne (13,1 % c. 7,0 %, ce qui signifie que la région de Lisbonne est surreprésentée dans la carte d'alerte) et le Centre (26,1 % c. 30,6 %, ce qui signifie que la région du Centre est sous-représentée dans la carte d'alerte). La répartition géographique des *freguesias* auxquelles un niveau de risque a été attribué dans la carte finale est très semblable à celle de la carte initiale.

Pour ce qui est des *freguesias* non répondantes, plus de la moitié sont situées dans le Nord et environ une sur quatre est située dans le Centre. Les autres régions comptent moins de 10 % de *freguesias* auxquelles aucun niveau de risque n'a été attribué. Ce schéma s'observe pour le groupe initial ainsi que le groupe final.

Tableau 3.3
Répartition géographique des *freguesias* auxquelles un niveau de risque est attribué et des non-répondants dans la carte d'alerte, selon le groupe de réponses et les *freguesias* dans la population (%)

Région	*Freguesias* auxquelles un niveau de risque est attribué		Non-répondants		Population
	Initial	Final	Initial	Final	
Nord	44,0	46,1	59,5	63,2	46,6
Centre	26,7	26,1	23,1	23,8	30,6
Lisbonne	13,7	13,1	8,4	6,1	7,0
Alentejo	7,7	7,5	5,2	2,6	8,9
Algarve	2,3	2,1	1,0	0,9	2,0
Archipel de Madère et des Açores	5,6	5,1	2,8	3,4	4,9
N =	3 264†	3 873†	987	378	4 260

† Un niveau de risque n'a pas pu être attribué à neuf *freguesias* du groupe initial parce qu'il n'y avait pas de réponse à la question sur l'identification de la *freguesia*.

3.2 Analyse des réponses des PJF

Afin de simplifier la structure des données d'enquête et de cerner les dimensions possibles du risque associé aux opérations de recensement, on a procédé à deux analyses en composantes principales (ACP). L'une de ces ACP portait sur les cinq questions au sujet des caractéristiques des PJF (âge, niveau d'études, durée de l'occupation du poste de président du *junta de freguesia*, fréquence de l'utilisation d'un ordinateur et fréquence de l'utilisation d'Internet), et l'autre ACP portait sur les questions avec échelle de Likert au sujet des caractéristiques des *freguesias* et du recrutement des agents recenseurs (sections 1 à 3 du questionnaire). Une valeur propre plus grande que un a été choisie comme critère pour extraire les composantes. Le tableau 3.4 donne le nombre de composantes principales (CP) et le pourcentage de la variance totale qu'elles expliquent en se basant sur une rotation varimax. Les deux ACP ont été exécutées sur le groupe initial et sur le groupe final de *freguesias*.

Les résultats révèlent que les réponses obtenues auprès des *freguesias* initiales ont la même structure, en ce qui a trait aux dimensions latentes du risque, que les réponses du groupe final de *freguesias*. L'indicateur d'adéquation de l'échantillon pour l'ACP sur les caractéristiques des PJF était raisonnablement bon (KMO > 0,6) dans les ensembles de données sur les *freguesias* initial ainsi que final. Dans les deux ensembles de données, l'analyse a donné lieu à l'extraction de deux composantes principales représentant environ 77 % de la variance des données. Les CP ont été nommées : CP_A – Compétences des PJF et CP_B – Expérience des PJF.

Tableau 3.4
Caractéristiques des analyses en composantes principales selon le groupe de réponses

Caractéristique analysée	Initial	Final
ACP sur les caractéristiques des PJF		
Mesure de Kaiser-Meyer-Olkin de l'adéquation de l'échantillon	0,687	0,685
CP extraites	2	2
Variance expliquée	77,3 %	77,2 %
ACP sur les caractéristiques des *freguesias*		
Mesure de Kaiser-Meyer-Olkin de l'adéquation de l'échantillon	0,693	0,696
CP extraites	5	5
Variance expliquée	61,4 %	61,3 %

L'indicateur d'adéquation de l'échantillon pour l'ACP sur les questions avec échelle de Likert était également raisonnablement bon (KMO > 0,6) pour les deux ensembles de données. Tant dans l'ensemble de données initial que dans l'ensemble de données final, cinq CP ont été extraites, représentant près de 61 % de la variance des données, à savoir : CP_1 – Population difficile à joindre, CP_2 – Agents recenseurs possédant les compétences appropriées et disponibles pour travailler au recensement, CP_3 – Population âgée, CP_4 – Régions désertes et CP_5 – Régions à taux élevé de logements habitables vacants.

En ce qui concerne l'opinion globale quant à la difficulté de mise en œuvre des opérations du Recensement de 2011 (question de la section 4 du questionnaire), la réponse de près des deux tiers des répondants était supérieure au point médian de l'échelle dans les deux groupes de réponses. Pour le groupe initial, 67,8 % des répondants ont choisi le niveau « 4 » ou « 5 » sur l'échelle de réponse comparativement à 67,5 % pour le groupe final (tableau 3.5).

Tableau 3.5
Opinion globale au sujet du recensement selon le groupe de réponses (%)

	Initial	Final
1 – « difficile »	1,7	1,7
2	3,8	3,8
3	26,7	27,0
4	38,4	37,9
5 – « facile »	29,4	29,6

3.3 Classification du niveau de risque des *freguesia*s

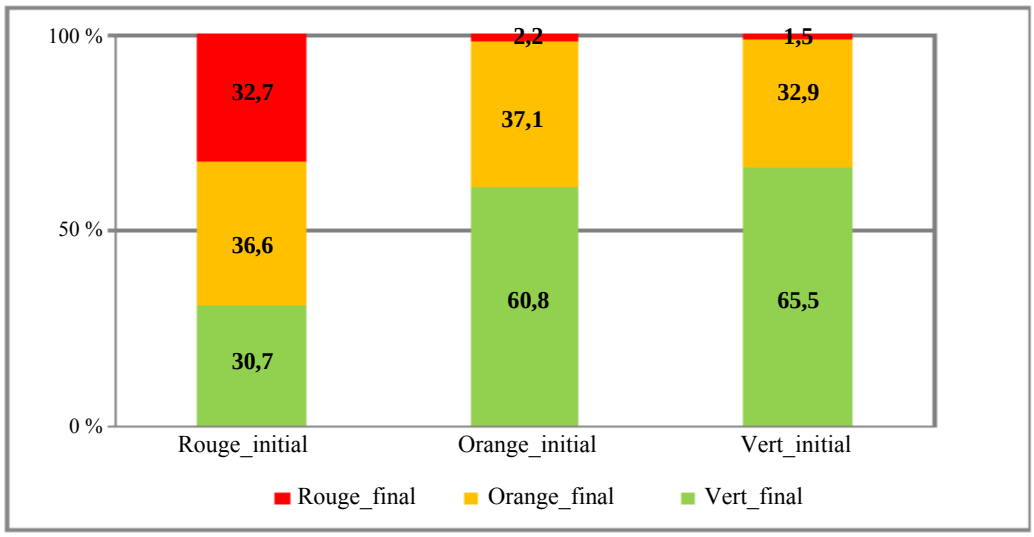
Les sept dimensions du risque dégagées des deux ACP ont ensuite été utilisées comme entrée dans un modèle de mélange fini et soumises à une classification automatique pour produire une segmentation des *freguesias* (les détails et les données de sortie de cette analyse ne sont pas présentés, mais peuvent être consultés sur ISCTE-IUL (2011)). La segmentation est effectuée pour les groupes initial et final de *freguesias*. Le résultat de la segmentation est présenté dans la carte d'alerte dans laquelle les *freguesias* sont représentées en rouge, en orange ou en vert (la carte d'alerte finale est présentée à la figure A.2 en annexe. Les taches noires représentent les *freguesias* auxquelles aucun niveau de risque n'a été attribué en raison de la non-réponse). Le tableau 3.6 résume la classification des niveaux de risque des *freguesias* dans les versions initiale et finale de la carte.

Tableau 3.6
Classification des niveaux de risque dans la carte d'alerte selon le groupe de réponses (%)

Niveau de risque	Initial (n = 3 264)	Final (n = 3 873)	Δ%
Risque élevé (rouge)	6,4	3,7	– 42,2
Risque moyen (orange)	53,3	33,9	– 36,4
Risque faible (vert)	40,3	62,4	+ 54,8

Dans la carte d'alerte initiale, la couleur dominante est l'orange (53,3 % des *freguesias* sont évaluées comme présentant un risque moyen). La part de *freguesias* à risque élevé n'est que de 6,4 %. Dans la carte finale, le vert prédomine (62,4 % des *freguesias* sont classées comme présentant un faible risque) et moins de 4 % de *freguesias* reçoivent le code de couleur rouge. L'ajout des réponses de suivi aux réponses initiales a donné lieu à une modification de la configuration de la carte d'alerte, surtout une augmentation du pourcentage de *freguesias* considérées comme présentant un risque faible (+ 54,8 %).

Nous avons ensuite analysé la façon dont les réponses de suivi ont modifié la classification des niveaux de risque des *freguesias* initiales. Les réponses des 3 264 *freguesias* initiales ont permis d'attribuer un code de couleur à chacune d'elles et d'établir la version initiale de la carte d'alerte. Après avoir intégré les réponses des *freguesias* de suivi, la carte d'alerte a été remaniée : non seulement il était possible d'attribuer un code de couleur à un plus grand nombre de *freguesias*, mais la couleur attribuée au départ aux *freguesias* initiales changeait aussi dans certains cas. Des 3 264 *freguesias* initiales, environ 50 % ont reçu une couleur différente dans la carte d'alerte finale. La figure 3.2 donne les changements globaux de la classification des niveaux de risque des *freguesias* initiales après l'intégration des réponses des *freguesias* de suivi.



(Base : n = 3 264)

Figure 3.2 Classification des niveaux de risque dans la carte d'alerte finale selon la classification des niveaux de risque dans la carte d'alerte initiale

Les *freguesias* qui avaient reçu le code de couleur vert au départ (vert_initial) avaient tendance à demeurer vertes (vert_final) après la prise en compte des réponses au suivi (65,5 %). Seulement 32,9 % des *freguesias* colorées en vert au départ sont passées à l'alerte orange (orange_final), et 1,5 % sont passées à l'alerte rouge (rouge_final). Quant aux *freguesias* qui au départ avaient reçu la couleur orange (orange_initial), les réponses au suivi ont entraîné le passage de 60,8 % d'entre elles à la couleur verte (vert_final); seulement 37,1 % sont demeurées orange (orange_final) et une minorité de 2,2 % sont passées à l'alerte rouge (rouge_final). Le changement le plus important causé par les réponses au suivi s'observe pour le groupe de *freguesias* ayant reçu la couleur rouge : seulement 32,7 % des *freguesias* rouges initiales (rouge_initial) sont demeurées dans la catégorie à risque élevé (rouge_final), et la majorité sont passées à l'orange (36,6 %) ou au vert (30,7 %).

Enfin, nous avons analysé la classification des niveaux de risque par région, et comparé les cartes initiale et finale (tableau 3.7).

Tableau 3.7
Classification des niveaux de risque par région selon le groupe de réponses (%)

Région	Niveau de risque	Initial	Final	Δ%
Nord	Risque élevé	4,2	0,8	–81,0
	Risque moyen	52,7	46,1	–12,5
	Faible risque	43,1	53,1	+23,2
Centre	Risque élevé	3,7	0,3	–91,9
	Risque moyen	54,8	18,6	–66,1
	Faible risque	41,5	81,2	+95,7
Lisbonne	Risque élevé	19,1	20,3	+6,3
	Risque moyen	45,3	24,1	–46,8
	Faible risque	35,6	55,6	+56,2
Alentejo	Risque élevé	4,0	1,0	–75,0
	Risque moyen	65,9	5,0	–92,4
	Faible risque	30,1	94,0	+212,3
Algarve	Risque élevé	17,0	29,8	+75,3
	Risque moyen	43,4	38,1	–12,2
	Faible risque	39,6	32,1	–18,9
Archipel de Madère et des Açores	Risque élevé	5,2	1,5	–71,2
	Risque moyen	56,0	60,9	+8,8
	Faible risque	38,8	37,6	–3,1

Lisbonne et l'Algarve sont les régions présentant le pourcentage le plus élevé de *freguesias* codées en rouge (19,1 % et 17,0 %, respectivement). Cette tendance s'observe dans la carte d'alerte initiale ainsi que finale. Les réponses au suivi ont entraîné une réduction du pourcentage de *freguesias* codées en rouge dans toutes les régions, sauf celles de Lisbonne et de l'Algarve pour lesquelles la carte d'alerte finale présente des pourcentages plus élevés de *freguesias* en rouge que la carte initiale. En ce qui concerne le pourcentage de *freguesias* à risque faible, les réponses au suivi ont causé une augmentation dans toutes les régions, sauf l'Algarve et l'Archipel de Madère et des Açores dans lesquelles une diminution a été constatée. En outre, le pourcentage de *freguesias* oranges a diminué dans toutes les régions après l'ajout des réponses au suivi, sauf pour l'Archipel de Madère et des Açores.

4 Discussion

Les résultats présentés plus haut montrent clairement que l'envoi par la poste de suivi a été utile et a eu un effet positif sur le taux de réponse au sondage sur la perception du risque, ainsi que sur la conception de la carte d'alerte.

Bien qu'il ait été impossible d'atteindre le taux de réponse cible de 100 % au sondage sur la perception du risque, le taux de réponse élevé – 91,1 % – n'a pu être obtenu que grâce à l'envoi par la poste de suivi. Le taux de réponse variait selon la région, mais le suivi a permis de l'accroître dans chacune d'elles. Le Nord affichait le taux de réponse le plus faible pour le groupe initial – 71 % – ainsi qu'après le suivi – 87,3 %. Plusieurs facteurs pourraient être à l'origine de ce résultat. Premièrement, les PJF du Nord

demeurent en fonction plus longtemps que partout ailleurs dans le pays. La durée moyenne de l'occupation du poste de PJF est de 8,6 années dans le Nord alors que la moyenne nationale est de 7,8 années. En outre, tandis que le 90^{e} centile de la distribution de la « durée de la période en tant que président » est de 20 ans dans le Nord, il ne dépasse pas 17 ans dans les autres régions. Cela signifie que les PJF en fonction dans le Nord ont plus d'expérience de gouvernance et sont probablement capables de mieux évaluer l'effet des caractéristiques particulières de leur *freguesia* sur le recensement. Il se pourrait que ces PJF aient eu le sentiment que leur *freguesia* ne poserait pas de problèmes pour le recensement et n'ont donc pas pris la peine de répondre au questionnaire. Un autre fait qui pourrait avoir joué un rôle dans le taux de réponse plus faible dans le Nord est que le principal parti de l'opposition est celui qui a recueilli le plus grand nombre de votes dans le Nord lors de la dernière élection parlementaire, de sorte que le manque de coopération des PJF pourrait avoir été une forme de censure à l'encontre du gouvernement central, parce qu'ils savaient que le sondage avait été demandé par le bureau officiel de la statistique du pays. Enfin, le Nord est la région comptant le plus grand nombre de *freguesias*, près de 2 000, ce qui rend l'obtention d'un taux de réponse de 100 % plus difficile que dans les régions plus petites, comme l'Algarve, qui compte moins de 90 *freguesias*.

Les réponses au suivi ont entraîné des changements de la classification des niveaux de risque des *freguesias*. Contrairement aux premières attentes, le scénario de codes de couleur de la carte d'alerte finale ne posait pas davantage problème que celui de la carte initiale. Non seulement le pourcentage de *freguesias* avec un code d'alerte rouge était plus faible dans la carte finale, mais le pourcentage de *freguesias* ayant un code vert avait augmenté. Par conséquent, en plus d'accroître le nombre de *freguesias* auxquelles était attribué un niveau de risque sur la carte (pour passer de 3 264 à 3 873 *freguesias*), l'envoi par la poste de suivi a permis de « corriger » la classification de certaines *freguesias*, à savoir celles classées au départ comme présentant un risque élevé, qui pour la plupart ont été recodées en orange ou en vert après avoir tenu compte de l'ensemble de données incluant les réponses au suivi.

Ces résultats montrent qu'il est important que les administrations locales s'investissent davantage et participent activement aux futures éditions du sondage. La stratégie de prise de contact adoptée pour le sondage sur la perception du risque comprenait l'envoi et le retour du questionnaire par la poste, mais d'autres approches pourraient être envisagées dans l'avenir, notamment l'ajout d'autres modes de collecte, comme Internet. En outre, les stratégies de prise de contact pourraient être personnalisées en fonction des caractéristiques particulières des régions. Comme le Nord affichait le taux de réponse le plus faible, une stratégie comportant un plus grand nombre de prises de contact de suivi (par la poste, par téléphone ou par courriel) pourrait être adoptée dans cette région, et une stratégie de contact et de re-contact moins agressive pourrait être appliquée dans les autres régions. Enfin, il convient de souligner que la carte administrative du Portugal a changé en 2013 et que le nombre total de *freguesias* est maintenant réduit à environ 3 000. Ce nouveau format d'organisation aura certainement un effet favorable sur le prochain sondage sur la perception du risque, puisque le plus petit nombre de PJF simplifiera la mise en œuvre de la stratégie de prise de contact et facilitera l'interrogation exhaustive de toutes les *freguesias*.

Remerciements

Le présent article fait partie du projet *Programa de Controlo e Avaliação da Qualidade dos Censos 2011*, un projet mené de concert par Statistique Portugal et l'*Instituto Universitário de Lisboa* (ISCTE-IUL).

Annexe

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

CENSOS 2011

Sondage sur la perception du risque

Questionnaire à l'intention des présidents des *juntas de freguesia* utilisé pour l'essai pilote du Recensement de 2011

IDENTIFICATION

Freguesia : _____

Municipalité : _____

Nom : _____ Âge : _____

Niveau d'études :

Inférieur au niveau de base ☐ Niveau de base (9 années obligatoires) ☐ Secondaire ☐ Université ☐

Depuis combien de temps êtes-vous président de ce *junta de freguesia*? : _____ années

Fréquence de l'utilisation d'un ordinateur : Rarement ☐ Plusieurs fois par jour ☐ Plusieurs fois par semaine ☐ Tous les jours ☐

Fréquence de l'utilisation d'Internet : Rarement ☐ Plusieurs fois par jour ☐ Plusieurs fois par semaine ☐ Tous les jours ☐

PERCEPTION AU SUJET DES CARACTÉRISTIQUES DE LA *FREGUESIA*

Servez-vous de l'échelle de 1 à 5 pour répondre aux questions qui suivent concernant la *freguesia*. Marquez d'un X le chiffre correspondant à votre choix.

1 POPULATION							
1. Existence de personnes âgées (âge ≥ 65 ans)	Peu	1	2	3	4	5	Beaucoup
2. Existence de personnes analphabètes (ne sachant pas lire ou écrire)	Peu	1	2	3	4	5	Beaucoup
3. Existence de personnes vivant dans des quartiers à logements sociaux	Peu	1	2	3	4	5	Beaucoup
4. Existence d'émigrants	Peu	1	2	3	4	5	Beaucoup
5. Existence d'immigrants	Peu	1	2	3	4	5	Beaucoup
6. Existence de sans-logis	Peu	1	2	3	4	5	Beaucoup
2 LOGEMENTS ET RÉGIONS							
1. Existence de régions à prédominance de copropriétés protégées	Peu	1	2	3	4	5	Beaucoup
2. Existence de régions à prédominance de résidences secondaires ou d'été	Peu	1	2	3	4	5	Beaucoup
3. Existence de régions à prédominance de bâtiments résidentiels construits récemment	Peu	1	2	3	4	5	Beaucoup
4. Existence de régions d'accès difficile (p. ex., pas de routes revêtues de tarmacadam, par d'éclairage, …)	Peu	1	2	3	4	5	Beaucoup
5. Existence de régions à logements dispersés	Peu	1	2	3	4	5	Beaucoup
6. Existence de régions principalement dortoirs	Peu	1	2	3	4	5	Beaucoup
3 RESSOURCES HUMAINES							
1. Dans quelle mesure sera-t-il difficile de recruter des agents recenseurs ayant les compétences voulues?	Difficile	1	2	3	4	5	Facile
2. Dans quelle mesure sera-t-il difficile de recruter des agents recenseurs ayant du temps disponible?	Difficile	1	2	3	4	5	Facile
4 OPINION GÉNÉRALE AU SUJET DU RECENSEMENT							
Dans quelle mesure sera-t-il difficile de mettre en œuvre le Recensement de 2011 dans la *freguesia*?	Difficile	1	2	3	4	5	Facile

Figure A.1 Questionnaire sur la perception du risque

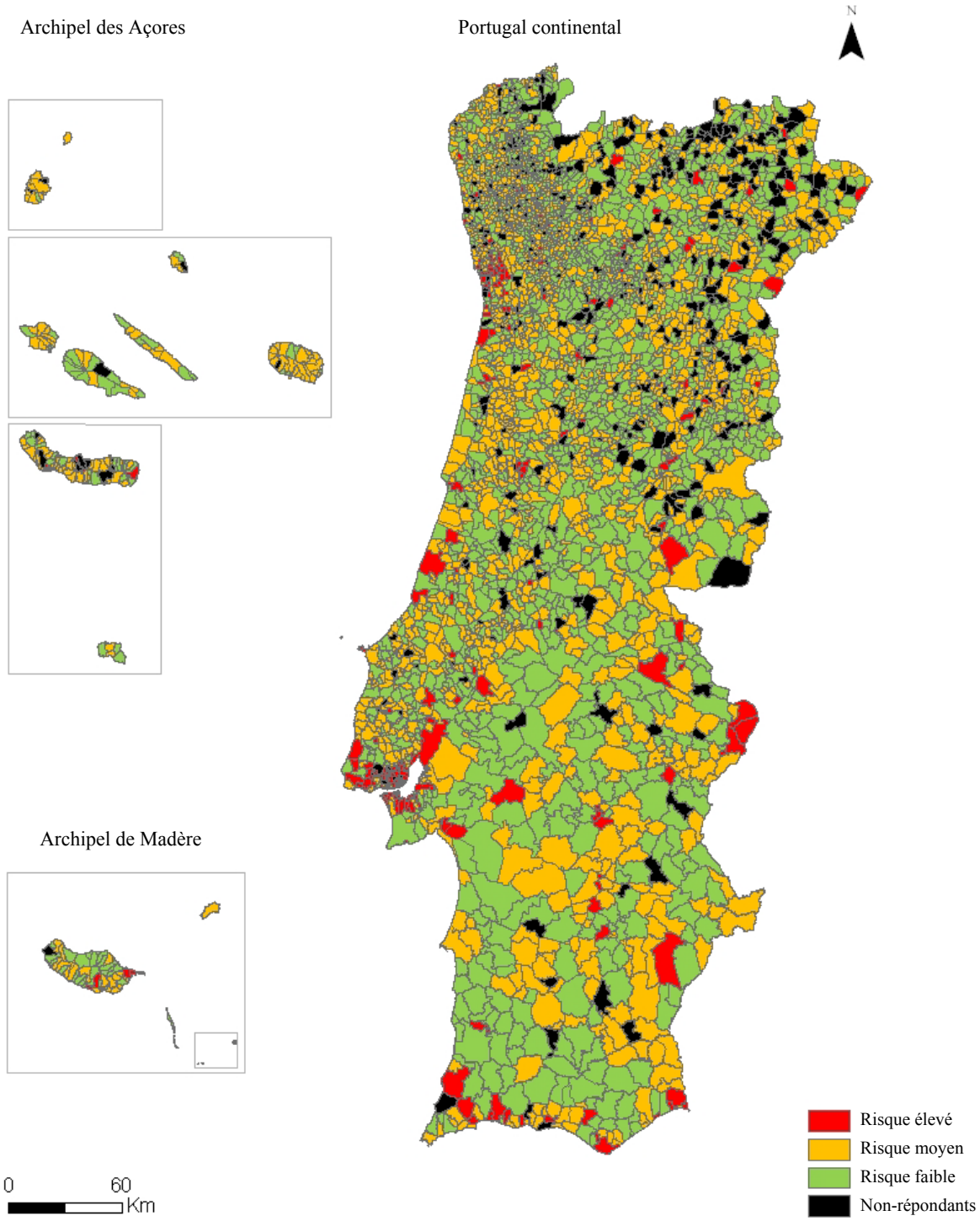


Figure A.2 Carte d'alerte finale

Bibliographie

Dillman, D., Smyth, J. et Christian, L. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 3ième Édition. New Jersey : Wiley.

Groves, R. (1989). *Survey Errors and Survey Costs*. New York : Wiley-Interscience.

Groves, R., et Couper, M. (1998). *Non-response in Household Interview Surveys*. New York : Wiley-Interscience.

Groves, R., Fowler Jr, F., Couper, M., Lepkowski, J., Singer, E. et Tourangeau, R. (2004). *Survey Methodology*. New York : Wiley-Interscience.

ISCTE-IUL (2011). Censos 2011-sistema de indicadores de alerta. (Document non-publié).

Porter, S. (2004). Raising response rates: What works? *New Directions for Institutional Research*, 121, 5-21.

Statistics Portugal (2007). *Programa de Acção para os Censos 2011*. Census Office, Statistics Portugal.

Statistics Portugal (2010). *Plano de Controlo e Avaliação da Qualidade Censos 2011 – Controlo do Processo Produtivo*. Census Office, Statistics Portugal.

Waite, P. (2007). *State, Local and Tribal Governments Benefit by Early Participation in the 2010 Census*. US Census Bureau Press Release.

Wroth-Smith, J., Abbott, O., Compton, G. et Benton, P. (2011). Quality assuring the 2011 Census population estimates. *Population Trends,* 143, 13-21.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 205-224
Statistique Canada, n° 12-001-X au catalogue

Mesure de l'emploi temporaire. Les données d'enquête ou de registre disent-elles la vérité ?

Dimitris Pavlopoulos et Jeroen K. Vermunt[1]

Résumé

L'une des principales variables de l'Enquête sur la population active des Pays-Bas est celle indiquant si un enquêté possède un emploi permanent ou temporaire. Le but de notre étude est de déterminer l'erreur de mesure de cette variable en appariant l'information tirée de la partie longitudinale de cette enquête à des données de registre uniques provenant de l'organisme de gestion des assurances sociales pour salariés des Pays-Bas (UVW). Contrairement aux approches antérieures visant à comparer des ensembles de données de ce genre, nous tenons compte du fait que les données de registre contiennent aussi des erreurs et que l'erreur de mesure qu'elles présentent est vraisemblablement corrélée dans le temps. Plus précisément, nous proposons d'estimer l'erreur de mesure dans ces deux sources en utilisant un modèle de Markov caché étendu au moyen de deux indicateurs observés du type de contrat d'emploi. Selon nos résultats, aucune des deux sources ne doit être considérée comme étant exempte d'erreur. Pour les deux indicateurs, nous constatons que les travailleurs titulaires d'un contrat d'emploi temporaire sont souvent classés incorrectement comme ayant un contrat d'emploi permanent. En particulier, dans le cas des données de registre, nous observons que les erreurs de mesure sont fortement autocorrélées, car les erreurs commises à une période ont tendance à se répéter. En revanche, lorsque l'enregistrement est correct, la probabilité qu'une erreur soit commise à la période suivante est presque nulle. Enfin, nous constatons que les contrats d'emploi temporaire sont plus répandus que ne le laisse supposer l'Enquête sur la population active, tandis que les taux de transition entre les contrats d'emploi temporaire et permanent sont nettement moins élevés que ne le suggèrent les deux ensembles de données.

Mots-clés : Contrats d'emploi temporaire; erreur de mesure; modèle de Markov caché; données de registre.

1 Introduction

La question de l'emploi temporaire occupe une place de plus en plus importante dans le débat économique et politique. Les contrats d'emploi temporaire permettent aux employeurs de contourner les règlements stricts en matière d'embauche et de licenciement (Bentolila et Bertola 1990; Booth 1997; Cahuc et Postel-Vinay 2002) et parfois même les règlements concernant la rigidité des salaires (OECD 2002). Surtout durant les récessions économiques, les employeurs recourent aux contrats d'emploi temporaire pour ajuster leur effectif en fonction des fluctuations de la demande de produits.

Les Pays-Bas font figure de pionniers en matière de souplesse d'emploi depuis le début des années 1990. La souplesse contractuelle est une caractéristique importante du marché du travail néerlandais. L'emploi temporaire a augmenté fortement, pour passer de 5,9 % en 1991 à 17,1 % en 2010 (OECD 2012), tandis que la contribution de la croissance de l'emploi temporaire à la croissance de l'emploi total a été de 9,9 points de pourcentage de 1990 à 2000 (OECD 2002). Les employeurs adoptent habituellement une stratégie d'effectif à « capacité minimale » (Sels et Van Hootegem 2001), qui consiste à offrir des contrats permanents à leurs travailleurs « de base » et des contrats temporaires aux autres travailleurs afin de pouvoir ajuster leur effectif en période de ralentissement économique.

Jusque récemment, les statistiques sur les contrats d'emploi temporaire aux Pays-Bas reposaient exclusivement sur des données d'enquêtes auprès des ménages et d'enquêtes sur la population active, mais

1. Dimitris Pavlopoulos, Université libre d'Amsterdam, département de sociologie, De Boelelaan 1081, 1081 HV Amsterdam, Pays-Bas. Courriel : d.pavlopoulos@vu.nl; Jeroen K. Vermunt, Université de Tilburg, département de méthodologie et de statistique, CP 90153, 5000 LE Tilburg, Pays-Bas. Courriel : j.k.vermunt@tilburguniversity.edu.

il existe aujourd'hui des données de registre de haute qualité qui peuvent compléter les données d'enquête, voire les remplacer. La première comparaison des deux sources de données a révélé certains chiffres gravement divergents quant à la taille de l'emploi temporaire. En 2009, la part de tous les types de contrats d'emploi temporaire était de 15,4 % selon l'Enquête sur la population active (EPA), tandis qu'elle était de 23,6 % selon les données du « Polisadministratie » (PA), qui sont des données de registre fournies par l'organisme de gestion des assurances sociales pour salariés (UWV) (Hilbers, Houwing et Kösters 2011). Comme la taille de l'emploi temporaire est un élément très important à prendre en considération dans l'élaboration des politiques relatives au marché du travail, *Statistics Netherlands* a entrepris de résoudre les divergences entre les deux sources de données. Une analyse plus approfondie des données ne s'est pas avérée très prometteuse. Les résultats préliminaires indiquent que 15,6 % des personnes titulaires d'un contrat d'emploi permanent selon l'EPA semblent posséder un contrat d'emploi temporaire selon le PA, tandis que 18,3 % des personnes titulaires d'un contrat d'emploi temporaire d'une durée inférieure à un an selon l'EPA semblent posséder un contrat d'emploi permanent selon le PA (Mars 2011). Bien qu'une part des incohérences puisse être expliquée par les définitions un peu différentes de l'emploi temporaire dans les deux sources de données, de grands écarts persistent même quand on utilise un échantillon apparié et que l'on sélectionne les cas pour lesquels il n'existe pas de différence de définition.

Comme le laissent entendre des études antérieures, l'erreur de mesure peut expliquer les incohérences constatées entre les données d'enquête et les données de registre. Dans le cas des données d'enquête, l'erreur de mesure est reconnue comme étant une source importante de biais (Rodgers, Brown et Duncan 1993; Pischke 1995; Bollinger 1996; Rendtel, Langeheine et Berntsen 1998; Bound, Brown et Mathiowetz 2001; Biemer 2011). L'erreur de mesure relative au type de contrat d'emploi n'a fait l'objet d'aucune étude jusqu'à présent, mais des travaux de recherche sur d'autres caractéristiques du marché du travail, dont l'activité, les salaires, le nombre d'heures de travail, l'industrie et la profession, montrent que les données d'enquête peuvent contenir de grandes quantités d'erreurs de mesure, susceptibles de biaiser sévèrement les résultats des analyses statistiques. Ainsi, Biemer (2004) soutient que dans les éditions de 1992 à 1994 de la *Current Population Survey*, 20,9 % des enquêtés en chômage ont été classés incorrectement dans d'autres catégories. Gottschalk (2005) indique que les deux tiers des réductions de salaire nominal sans changement d'emploi étaient dues à des erreurs de mesure. En particulier, 17 % des travailleurs déclarent une réduction de salaire nominal d'une année à l'autre tout en restant au service du même employeur. Cependant, si l'on neutralise l'effet de l'erreur de mesure, la proportion de travailleurs qui demeurent avec le même employeur et font face à des réductions de salaire nominal annuelles ne dépasse pas 4 % à 5 %. En s'appuyant sur l'étude de validation de la *Panel Study of Income Dynamics* (PSID), Mathiowetz (1992) constate que la concordance entre les registres des entreprises et les réponses aux enquêtes pour ce qui est de la catégorie professionnelle est de 87,3 %. Brown et Medoff (1996) observent une corrélation de 0,82 entre les registres des entreprises et les réponses aux enquêtes en ce qui a trait à la taille de l'établissement, et une corrélation de 0,86 en ce qui concerne la taille de l'entreprise.

La recherche sur l'erreur de mesure est nettement moins riche dans le cas des données de registre que dans celui des données d'enquête. Les données de registre sont habituellement traitées comme étant exemptes d'erreur et utilisées comme « norme de référence » quand elles sont comparées aux données d'enquête. Ainsi, la plupart de la recherche fondée sur l'étude de validation de la PSID fait appel à cette hypothèse (Duncan et Hill 1985; Rodgers et coll. 1993; Bound, Brown, Duncan et Rodgers 1994;

Pischke 1995). Toutefois, certains travaux montrent aussi que l'hypothèse de la « norme de référence » n'est pas toujours plausible. Kapteyn et Ypma (2007) étudient l'erreur de mesure dans la rémunération et, quoiqu'ils retiennent l'hypothèse que les données de registre sont exemptes d'erreur, ils tiennent compte de l'erreur dans l'appariement des données d'enquête aux données de registre. En particulier, ils supposent qu'un enregistrement dans le registre est identique à un enregistrement dans l'enquête avec une certaine probabilité. Ils concluent que l'introduction de cette source supplémentaire d'erreur modifie le profil de l'erreur de mesure dans l'enquête. Abowd et Stinson (2005) comparent la rémunération déclarée dans le cadre de la *Survey of Income and Program Participation* (SIPP) et les données des *Detailed Earnings Records* (DER). Ils constatent que l'erreur de mesure est plus importante dans les données administratives des DER (20 % à 27 %) que dans les données de la SIPP (13 % à 15 %). En comparant les mêmes ensembles de données, Gottschalk et Huynh (2010) soutiennent que l'erreur de mesure peut biaiser fortement les mesures de l'inégalité des revenus.

L'objectif du présent article consiste à estimer la grandeur de l'erreur de mesure concernant le type de contrat d'emploi dans l'EPA des Pays-Bas. À cette fin, nous apparions les données de l'enquête à des données provenant du registre PA. Les données de registre sont traitées comme n'étant pas exemptes d'erreur, car nous modélisons simultanément l'erreur de mesure dans les deux sources. Nous utilisons un modèle de Markov caché étendu au moyen de deux indicateurs du type de contrat (temporaire ou permanent), chacun provenant de l'une de nos sources de données.

Le restant de l'article est présenté comme il suit. À la section 2, nous nous étendons davantage sur le problème de la mesure de l'emploi temporaire au Pays-Bas en présentant les renseignements pertinents sur les deux sources de données, ainsi que certaines statistiques descriptives. À la section 3, nous décrivons le modèle de Markov caché utilisé pour l'étude. À la section 4, nous discutons des résultats de l'analyse. Enfin, à la section 5, nous présentons les conclusions de notre étude.

2 Description des deux sources de données

Les deux sources de données sur les contrats d'emploi temporaire sont l'Enquête sur la population active (en néerlandais : *Enquête Beroepsbevolking*) administrée par *Statistics Netherlands* (en néerlandais : *Centraal Bureau voor de Statistiek* – CBS) et l'ensemble de données « Polisadministratie » de l'organisme de gestion des assurances sociales pour salariés (UWV). L'EPA est une enquête trimestrielle à panel rotatif sur les caractéristiques du marché du travail qui est représentative de la population néerlandaise de 15 ans et plus. L'enquête a été lancée en 1987, mais sa composante longitudinale a été introduite en 1999. Depuis 1999, les enquêtés sont interviewés durant cinq vagues consécutives du panel, ce qui permet d'étudier les évolutions individuelles à court terme sur le marché du travail. L'information recueillie a trait à la situation au moment de l'interview. Les interviews sont réparties assez uniformément au cours du trimestre.

Les erreurs de mesure du type de contrat dans le cadre de l'EPA sont, comme cela est habituellement le cas dans les enquêtes, le résultat d'erreurs de déclaration par les enquêtés ou d'erreurs d'enregistrement des réponses par les enquêteurs. Le recours à des interviews par personne interposée est une source supplémentaire d'erreur. Habituellement, dans l'EPA, un seul membre du ménage fournit les réponses pour tous les membres du ménage inclus dans l'échantillon, ce qui accroît l'erreur de mesure. Dans notre

échantillon de l'EPA, 40,1 % des observations proviennent d'interviews par personne interposée. Une autre cause possible d'erreur de mesure tient au fait que les travailleurs peuvent confondre le contrat d'emploi juridique et le contrat d'emploi implicite ou psychologique avec leur employeur. En particulier chez les cohortes de jeunes travailleurs parmi lesquels les contrats flexibles sont très répandus et dans les secteurs où la mobilité d'emploi est grande et où les conditions d'emploi évoluent, comme le secteur de la santé, les travailleurs peuvent déclarer qu'ils ont un contrat permanent en se basant sur les promesses de l'employeur, alors qu'en réalité, ils sont employés aux termes d'un contrat temporaire.

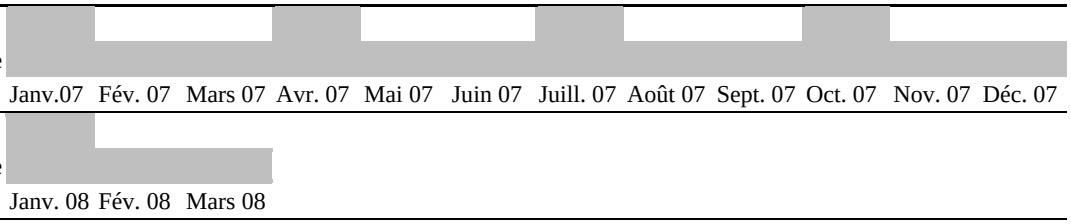
Le PA est un ensemble de données de registre unique contenant l'information sur le marché du travail et le revenu de tous les travailleurs assurés au Pays-Bas. Cet ensemble de données est construit en recueillant et en appariant l'information provenant de diverses sources, comme le Bureau de l'impôt (en néerlandais : *Belastingdienst*) – y compris des données provenant des déclarations de revenu des particuliers aux fins de l'impôt (en néerlandais : *jaaropgave*), les déclarations fournies par les agences d'emploi temporaire (en néerlandais : *weekaanleveringen*) et le registre de la population (en néerlandais : *Gemeentelijke BasisAdministratie persoonsgegevens* - GBA). Le PA est administré par l'organisme de gestion des assurances sociales pour salariés des Pays-Bas (UWV).

L'UWV a tout intérêt à maintenir le haut niveau de qualité et d'exactitude du PA, car cette source de données est utilisée par plusieurs institutions gouvernementales. Par exemple, les cotisations de sécurité sociale, les allocations de logement (en néerlandais : *huurtoeslag*), et les indemnités de soins de santé (en néerlandais : *zorgtoeslag*) sont déterminées en se servant de l'information provenant de cet ensemble de données. Afin d'améliorer la qualité des donnés, le PA a fait l'objet de plusieurs révisions depuis 2006. Il n'y a pas de données manquantes car les employeurs sont obligés de transmettre les relevés d'information fiscale. Cependant, alors que l'ensemble de données contient des renseignements mensuels, les employeurs transmettent habituellement l'information pertinente une fois par an (il est impossible d'extraire le moment de la transmission). Cette situation peut donner lieu à des erreurs pour la période comprise entre deux transmissions consécutives, surtout en ce qui concerne la mesure du type de contrat, qui n'est manifestement pas la variable la plus importante pour les utilisateurs du PA. Par conséquent, il est probable que, si une erreur est commise concernant le type de contrat, elle persistera jusqu'au moment où l'employeur soumettra le rapport suivant à l'UWV. Cela signifie que l'on peut s'attendre à ce que l'erreur de mesure dans le PA soit autocorrélée.

Pour les besoins de notre étude, nous choisissons des participants à l'EPA qui ont été interviewés pour la première fois durant le premier trimestre de 2007. Puisque nous nous concentrons sur les personnes ayant un emploi, nous avons gardé dans l'échantillon les personnes âgées de 25 à 55 ans. Après l'application de la contrainte d'âge, nous avons obtenu une taille d'échantillon de 11 632 personnes. Pour toutes ces personnes, *Statistics Netherlands* a apparié l'information provenant de l'EPA à l'information mensuelle provenant du PA en utilisant le numéro de sécurité sociale des individus. Le niveau d'appariement atteint était de 98 % et toutes les incohérences pertinentes ont été résolues (l'appariement et le contrôle de la qualité ont été effectués par *Statistics Netherlands*). Notre ensemble de données finales se présente sous la forme d'un fichier personne-mois pour 11 632 personnes avec 15 observations pour la période de janvier 2007 à mars 2008 contenant l'information complète provenant du PA et l'information partiellement observée (5 observations – une réponse tous les trois mois) provenant de l'EPA. L'ensemble de données appariées est illustré au tableau 2.1. Cet ensemble de données de panel n'est pas équilibré pour l'EPA, car nos données d'enquête souffrent d'une certaine attrition. Plus précisément, des

11 632 personnes qui avaient répondu à la première interview, 9 970 restaient dans l'échantillon de l'EPA pour la deuxième interview, 9 113 pour la troisième, 8 953 pour la quatrième et 8 629 pour la dernière interview. Dans le cas des données du PA pour cet échantillon, il n'y a pas d'attrition, de sorte que l'échantillon est entièrement équilibré.

Tableau 2.1
Une illustration de notre exemple



Nota : Le tableau illustre la correspondance entre le panel rotatif de l'EPA et les observations mensuelles provenant du registre Polisadministratie. Il a trait aux personnes qui ont été interviewées le premier mois de chaque trimestre. Une cellule ombrée en gris indique une observation valide.

La principale variable d'intérêt de notre étude est le type de contrat d'emploi, qui peut prendre trois valeurs : permanent, temporaire et « autre ».

Le type de contrat est déterminé d'après l'emploi principal, donc l'information sur les autres emplois qu'une personne pourrait occuper n'est pas prise en compte. Les personnes qui n'ont pas d'emploi rémunéré sont classées dans la catégorie « autre ». Il convient de souligner que cette dernière catégorie est assez hétérogène, car elle englobe, entre autre, les catégories de travail autonome, de chômage et d'études à temps plein. Cependant, il est nécessaire d'ajouter cette catégorie dans notre analyse, car dans les modèles de Markov, les classes latentes doivent être mutuellement exclusives et exhaustives.

Le tableau 2.2 présente la distribution des types de contrats observée pour le premier mois de la période de référence selon les données d'enquête et les données de registre. Les écarts les plus importants ont lieu pour les pourcentages de personnes possédant des contrats d'emploi permanent et temporaire, et sont moindre pour la catégorie « autre ». Selon les données d'enquête, en janvier 2007, 8 % de la population active détenait un contrat d'emploi temporaire, tandis que selon les données du registre, la proportion était plus élevée (12,3 %).

Tableau 2.2
Distribution des types de contrats selon l'enquête et selon le registre

	Enquête	Registre
Permanent	0,659	0,602
Temporaire	0,080	0,123
Autre	0,261	0,275
Total	1,0	1,0
Cas	3 887	11 632

Nota : Ces distributions de fréquence ont trait au premier mois de la période de référence, janvier 2007. L'échantillon de l'EPA est plus petit que l'échantillon du PA car seulement 3 887 participants à l'EPA ont été interviewés pour la première fois en janvier 2007. Les autres participants ont été interviewés en février et en mars 2007.

Le tableau 2.3 donne le croisement du type de contrat selon les deux sources pour l'échantillon groupé. Ce tableau confirme les écarts importants entre les deux sources de données signalés par *Statistics Netherlands*. Ces écarts ont trait principalement aux personnes qui sont enregistrées comme travaillant aux termes d'un contrat temporaire. Plus précisément, 50,2 % de personnes qui, selon les données du registre, ont un contrat d'emploi temporaire semblent posséder un contrat d'emploi permanent selon l'enquête. Des divergences plus faibles, mais néanmoins existantes, se dégagent pour les personnes qui sont inscrites comme ayant un contrat d'emploi permanent ou comme appartenant à une autre catégorie.

Les incohérences de classification des personnes présentées au tableau 2.3 ont des conséquences importantes en ce qui concerne les transitions entre les différents états d'emploi. Le tableau 2.4 donne les taux de transition sur trois mois pour les cas assortis d'une observation valide provenant de l'EPA. Ce tableau indique que les données du registre contiennent un plus grand nombre de transitions que les données d'enquête. En particulier, parmi les personnes ayant un contrat d'emploi temporaire au mois $t - 3$, 5,7 % ont un contrat d'emploi permanent au mois t selon les données d'enquête et 8,5 % sont dans cette situation selon les données du registre.

Tableau 2.3
Tableau croisé du type de contrat selon l'enquête et selon le registre

Données du registre	Données d'enquête			
	Permanent	**Temporaire**	**Autre**	**Total**
Permanent	0,944	0,039	0,017	1,0
Temporaire	0,502	0,437	0,061	1,0
Autre	0,081	0,030	0,889	1,0
Total	0,667	0,087	0,246	1,0
Cas	32 225	4 216	11 856	48 297

Nota : Les distributions de fréquence sont calculées pour l'échantillon groupé. Le total général représente le nombre d'enregistrements de l'EPA inclus dans notre analyse dans l'échantillon groupé.

Tableau 2.4
Transitions sur trois mois observées dans l'EPA et dans le PA

Transitions observées d'après les données de l'enquête				
			Contrat en *t*	
		Permanent	**Temporaire**	**Autre**
Contrat en *t*-3	Permanent	0,981	0,009	0,010
	Temporaire	0,057	0,889	0,054
	Autre	0,017	0,035	0,948
	Total	0,674	0,089	0,237
Transitions observées d'après les données du registre				
			Contrat en *t*	
		Permanent	**Temporaire**	**Autre**
Contrat en *t*-3	Permanent	0,967	0,018	0,015
	Temporaire	0,085	0,860	0,055
	Autre	0,018	0,036	0,946
	Total	0,624	0,128	0,247

Nota : Il s'agit des taux de transition sur une période de trois mois et pour 34 820 cas de notre échantillon groupé. Ces cas proviennent des participants à l'EPA qui figurent au moins deux fois dans notre échantillon.

3 Modèle de Markov caché utilisé pour estimer l'erreur de mesure du type de contrat

Le modèle que nous utilisons pour estimer l'erreur dans la mesure du type de contrat est un modèle de Markov caché ou latent. Ce modèle a été utilisé pour estimer l'erreur de mesure de variables provenant des enquêtes sur l'emploi (voir, entre autre, van der Pol et Langeheine 1990; Rendtel et coll. 1998; Bassi, Hagenaars, Croon et Vermunt 2000; Biemer et Bushery 2000; Biemer 2011; Pavlopoulos, Muffels et Vermunt 2012). Notre application diffère des applications susmentionnées en ce sens que nous avons deux mesures au lieu d'une seule pour la variable de résultat; autrement dit, le type de contrat d'après le registre PA et d'après l'EPA. D'autres exemples d'applications de modèles de Markov latents spécifiées pour de multiples variables de réponse figurent dans Langeheine (1994), Paas, Vermunt et Bijmolt (2007), Bartolucci, Lupparelli et Montanari (2009) et Manzoni, Vermunt, Luijkx et Muffels (2010).

Soit C_{it} et E_{it} l'état observé de la personne i au point dans le temps t selon le registre et selon l'enquête, respectivement, où $i = 1, \ldots, N$ et $t = 0, \ldots, T$. Pour tenir compte du fait que E_{it} n'est observé que tous les trois mois, nous utilisons la variable indicatrice δ_{it} qui est égale à 1 si l'information de l'enquête était disponible pour le mois concerné et 0 autrement. En plus des mesures provenant du registre et de l'enquête, le modèle de Markov caché contient une variable non observable représentant le type de contrat réel d'une personne au point t dans le temps. Nous désignons cet état latent par X_{it}. Notons que C_{it}, E_{it} et X_{it} peuvent prendre trois valeurs représentant les catégories de contrat permanent, temporaire et autre. Nous désignons une catégorie particulière de ces variables par c_t, e_t, et x_t, respectivement.

Le diagramme de cheminement pour le modèle de Markov caché d'intérêt est illustré à la figure 3.1. Pour simplifier, ce diagramme de cheminement se rapporte uniquement aux personnes qui sont entrées dans l'échantillon de l'EPA durant un mois particulier. Pour cette raison, des quatre observations qui sont illustrées dans le diagramme, seules celles des mois $t-3$ et t ne manquent pas pour l'EPA. Comme le montre la figure, le type de contrat latent X_{it} suit un processus de Markov d'ordre 1; autrement dit, le contrat réel au point dans le temps t, X_{it}, est indépendant du contrat au point dans le temps t', $X_{it'}$, pour $t' < t-1$, conditionnellement à l'état au temps $t-1, X_{i(t-1)}$. Une autre hypothèse est que les états observés sont indépendants l'un de l'autre à l'intérieur et entre les points dans le temps, ce que nous appelons l'hypothèse d'indépendance locale ou l'hypothèse d'erreurs de classification indépendantes (ECI). On peut aussi constater que E_{it} est observé seulement tous les trois points dans le temps.

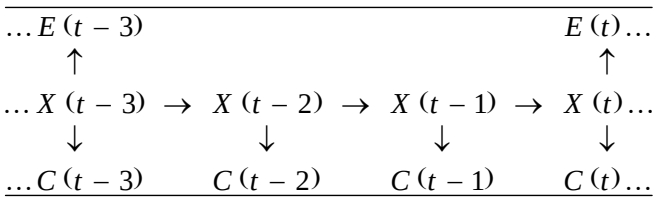


Figure 3.1 Diagramme de cheminement pour le modèle de Markov caché avec deux indicateurs (partiellement) observés

Comme il est mentionné à la section précédente, nous utilisons les données pour 15 mois, ce qui signifie que t va de 0 à $T = 14$. La probabilité de suivre un certain cheminement observé sur la période de $T + 1$ mois peut être exprimée comme il suit :

$$P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E}_i = \mathbf{e}_i) = \sum_{x_0=1}^{3}\sum_{x_1=1}^{3}\ldots\sum_{x_T=1}^{3} P(X_{i0} = x_0)\prod_{t=1}^{T} P(X_{it} = x_t | X_{i(t-1)} = x_{t-1}) \prod_{t=0}^{T} P(C_{it} = c_t | X_{it} = x_t)\prod_{t=0}^{T} P(E_{it} = e_t | X_{it} = x_t)^{\delta_{it}} . \quad (3.1)$$

Les probabilités pertinentes qui figurent dans cette équation sont les probabilités de l'état initial $P(X_{i0} = x_0)$, les probabilités de transition propre à la période $P(X_{it} = x_t | X_{i(t-1)} = x_{t-1})$, les probabilités d'erreur de mesure pour le registre $P(C_{it} = c_t | X_{it} = x_t)$, et les probabilités d'erreur de mesure pour l'enquête $P(E_{it} = e_t | X_{it} = x_t)$.

Jusqu'à présent, nous avons supposé que l'erreur de mesure n'était pas corrélée entre les divers points dans le temps – c'est-à-dire que l'hypothèse ECI est vérifiée – ce qui pourrait ne pas être raisonnable dans le cas de notre application. Avant tout, comme il est mentionné à la section précédente, l'erreur de mesure dans les données du registre est vraisemblablement autocorrélée; autrement dit, l'existence d'une erreur de classification entre X_{it} et C_{it} au point dans le temps t augmente la probabilité que la même erreur soit présente au point dans le temps $t + 1$. Cela tient au fait que les erreurs que les employeurs commettent dans leurs registres ne sont pas corrigées jusqu'à ce qu'ait lieu un contrôle ordinaire. Dans les données d'enquête, surtout parce qu'elles sont prospectives plutôt que rétrospectives, il n'y a aucune raison de penser qu'il existe une structure d'erreur autocorrélée « directe » similaire. Par contre, les erreurs dans les données d'enquête peuvent être corrélées dans le temps parce que la probabilité de commettre une erreur peut varier d'un groupe de personnes à l'autre, ce qui est parfois appelé erreur de mesure différentielle. En particulier, l'erreur de mesure dans les données d'enquête est vraisemblablement plus importante dans les secteurs où la mobilité est fréquente et où il existe une ambiguïté quant aux accords conclus entre les employeurs et les travailleurs, par exemple dans le secteur de la santé. De surcroît, les erreurs peuvent être plus importantes pour les jeunes travailleurs qui se soucient moins d'une relation de long terme avec l'employeur et qui, par conséquent, pourraient avoir une vision moins claire que les enquêtés plus âgés des modalités officielles de leur contrat. La figure 3.2 représente le diagramme de cheminement du modèle lorsque l'on corrige l'hétérogénéité et l'autocorrélation possibles de l'erreur de mesure, où V représente les variables observées qui introduisent une corrélation dans le temps de l'erreur de mesure dans les données d'enquête.

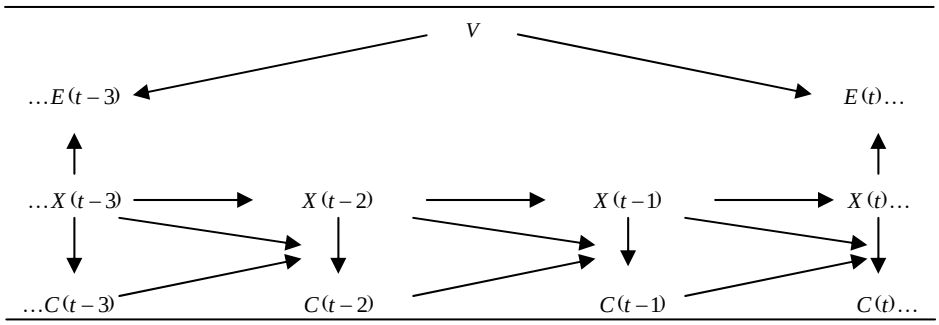


Figure 3.2 Diagramme de cheminement pour le modèle de Markov caché avec deux indicateurs et des erreurs corrélées

Comme il est également important de tenir compte de l'hétérogénéité dans la partie structurelle d'un modèle de Markov (Shorrocks 1976), le modèle est étendu au moyen de variables observées – éventuellement variant en fonction du temps – qui ont une incidence sur l'état initial et les probabilités de transition latentes, en suivant l'approche de Vermunt, Langeheine et Böckenholt (1999). Nous désignons ces variables de contrôle par $\mathbf{Z}_{it}$. Cependant, ces variables de contrôle observées ne peuvent pas traduire complètement l'hétérogénéité présente dans les probabilités de transition latentes, car celles-ci peuvent également être affectées par des traits personnels non observés, tels que la motivation et les compétences. En suivant l'approche la plus classique dans le cadre des modèles de Markov caché, nous corrigeons l'hétérogénéité non observée en supposant que la population est constituée d'un petit nombre de classes latentes dont l'état initial et les probabilités de transition diffèrent (Poulsen 1990). De cette façon, nous évitons les hypothèses de distribution de la variable latente peu attrayantes qui sont adoptées dans les modèles à effets aléatoires continus (Heckman et Singer 1984; Vermunt 1997). Le nombre de classes latentes K peut être déterminé en utilisant des indices d'ajustement du modèle.

Dans notre modèle de Markov caché mixte, la probabilité conjointe d'avoir un cheminement vers un état observé particulier conditionnellement aux valeurs du prédicteur peut s'exprimer par l'équation :

$$
\begin{aligned}
P\left(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E}_i = \mathbf{e}_i \middle| \mathbf{V}_i, \mathbf{Z}_i\right) \;=\; & \sum_{k=1}^{K} \sum_{x_0=1}^{3} \sum_{x_1=1}^{3} \ldots \sum_{x_T=1}^{3} \pi_k P\left(X_{i0} = x_0 \middle| \mathbf{Z}_{i0}, k\right) \\
& \prod_{t=1}^{T} P\left(X_{it} = x_t \middle| X_{i(t-1)} = x_{t-1}, \mathbf{Z}_{it}, k\right) \\
& P\left(C_{i0} = c_0 \middle| X_{i0} = x_0\right) \qquad (3.2) \\
& \prod_{t=1}^{T} P\left(C_{it} = c_t \middle| X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1}\right) \\
& \prod_{t=0}^{T} P\left(E_{it} = e_t \middle| X_{it} = x_t, \mathbf{V}_{it}\right)^{\delta_{it}} ,
\end{aligned}
$$

qui spécifie un modèle de mélange fini comportant K classes latentes pour tenir compte de l'hétérogénéité non observée dans l'état latent initial et dans les probabilités de transition latentes. π_k est la probabilité d'appartenir à la classe latente $k, \mathbf{V}_{it}$ est le vecteur de covariables ayant une incidence sur l'erreur de mesure dans les données d'enquête (âge et interview par personne interposée), et $\mathbf{Z}_{it}$ est le vecteur des covariables ayant une incidence sur les probabilités de transition latentes (sexe, âge, niveau de scolarité et pays d'origine). $\mathbf{Z}_{i0}$ est le vecteur des valeurs de ces covariables au point dans le temps initial.

Contrairement à l'équation 3.1, dans l'équation 3.2, les probabilités d'erreur dans les données d'enquête peuvent dépendre des covariables $(\mathbf{V}_{it})$. Les effets des covariables sur ces probabilités d'erreur sont modélisés en utilisant un modèle logit. En outre, les probabilités d'erreur dans les données du registre peuvent dépendre du type de contrat observé décalé et du type de contrat réel décalé. Notons que $X_{i(t-1)}$ et $C_{i(t-1)}$ peuvent prendre trois valeurs, ce qui implique qu'il existe neuf (3 fois 3) ensembles différents de

probabilités d'erreur dans les données du registre, un pour chaque combinaison possible de contrat observé et de contrat latent décalés. Comme cela n'a pas de sens d'estimer toutes ces probabilités d'erreur librement, nous avons utilisé un modèle restreint. Plus spécifiquement, nous définissons un modèle logit pour $P(C_{it} = c_t | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1})$ de la forme $\alpha_{c_t,x_t} + \beta_{c_t,c_{t-1},x_t,x_{t-1}}$, où $\beta_{c_t,c_{t-1},x_t,x_{t-1}}$ est un paramètre libre quand $c_t = c_{t-1} \neq x_t = x_{t-1}$ (quand la même erreur est commise entre des points dans le temps adjacents) et égal à 0 autrement. Ce modèle, qui contient six paramètres supplémentaires comparativement à un modèle sans effets décalés sur les probabilités d'erreur de classification, exprime que la vraisemblance de commettre une erreur particulière dépend du fait que *la même erreur* a été ou non commise au point dans le temps précédent. Des structures d'erreur corrélées contraintes similaires ont été utilisées par Manzoni et coll. (2010) dans un modèle de Markov latent pour des réponses recueillies rétrospectivement.

Pour restreindre l'état initial et les probabilités de transition latentes, nous nous servons aussi de modèles logit, tandis que pour les transitions latentes, nous utilisons des modèles qui séparent les coefficients selon l'état d'origine. Le même ensemble de covariables ($\mathbf{Z_{i0}}$ et $\mathbf{Z_{it}}$, respectivement) est introduit dans les modèles pour l'estimation de l'état initial et des probabilités de transition latentes. Notons que le modèle de Markov caché mixte décrit dans l'équation 3.2 repose sur l'hypothèse d'un processus de Markov d'ordre 1 pour les états réels conditionnellement aux valeurs individuelles des covariables et des effets non observés constants dans le temps, mais que cette hypothèse ne doit pas être vérifiée après la marginalisation sur les valeurs des covariables et les classes latentes. Un simple modèle de Markov de premier ordre ne conviendrait pas pour les transitions dans l'emploi, particulièrement au niveau mensuel, parce qu'il existe une dépendance à l'égard de la durée dans le chômage. Par exemple, il est peu probable qu'une personne en chômage durant les mois 3 à 9 ait la même probabilité de se trouver dans un état particulier sur le marché du travail au mois 10 qu'une personne qui n'a été en chômage que durant le mois 9. Cependant, dans un modèle de Markov caché, le biais de l'erreur de classification due à la violation de l'hypothèse de Markov est minime. En utilisant des simulations, Biemer et Bushery (2000) montrent que, même dans les cas d'une violation grave de l'hypothèse de Markov, dans un modèle de Markov caché, le biais de l'estimation de l'erreur de classification dans la catégorie du chômage ne dépasse pas 3 %.

Les estimations du maximum de vraisemblance des paramètres du modèle sont obtenues en utilisant une variante de l'algorithme espérance-maximisation (EM) (Dempster, Laird et Rubin 1977) appelée algorithme *forward-backward* ou algorithme de Baum-Welch (Baum, Petrie, Soules et Weiss 1970). Nous utilisons une extension de cet algorithme pour modèles de Markov latents mixtes avec covariables comme il est décrit, entre autres, dans Vermunt, Tran et Magidson (2008) et Pavlopoulos et coll. (2012). À l'étape E, nous calculons la log-vraisemblance prévue pour les données complètes, ce qui comprend le calcul des probabilités a posteriori marginales pertinentes pour les classes latentes et les états latents. À l'étape M, les paramètres du modèle sont mis à jour en utilisant des algorithmes standard pour l'analyse par régression logistique, où les probabilités a posteriori marginales sont utilisées comme pondérations. Cet algorithme est implémenté dans le programme Latent GOLD (Vermunt et Magidson 2008), qui fournit aussi les erreurs-types pour les paramètres du modèle (d'autres programmes populaires pour l'estimation des modèles de Markov latents sont MPLUS, LEM et PANMARK).

Les valeurs manquantes en raison de la conception de l'enquête (parce que les participants ne sont interviewés qu'une fois tous les trois mois) manquent complètement au hasard (MCAR pour *Missing*

Completely At Random). Les valeurs manquantes dues à l'attrition sont traitées comme des données manquant au hasard (MAR pour *Missing At Random*). Plus précisément, de la manière standard dans la procédure d'estimation du MV, nous maximisons la log-vraisemblance pour les données observées incomplètement, ce qui s'obtient en éliminant par intégration les valeurs manquantes. Cette procédure est valide sous MAR.

Comme le plan de sondage de l'EPA est complexe, nous avons utilisé dans le modèle les poids de sondage de l'enquête, à savoir une pondération unique par observation. Ces poids sont employés dans une procédure d'estimation du pseudo MV, où les erreurs-types sont ajustées pour tenir compte de la pondération en utilisant un estimateur par linéarisation (Skinner, Holt et Smith 1989). Comme il s'agit de pondérations trimestrielles, elles ne conviennent pas pour estimer les totaux de population au niveau mensuel. Cependant, comme nous utilisons l'information provenant du registre pour tous les participants à l'EPA qui sont entrés dans l'enquête durant un trimestre particulier, ces pondérations sont appropriées pour l'estimation des modèles de Markov cachés.

4 Résultat pour les données de l'EPA et du PA appariées

En tout, nous avons estimé neuf modèles qui sont présentés au tableau 4.1. Tous ces modèles sont des modèles de Markov cachés d'ordre 1 avec deux variables indicatrices pour le type de contrat comme il est décrit à la section 3. Les probabilités d'erreur sont homogènes dans le temps. Les probabilités de transition (latentes) sont supposées être hétérogènes dans le temps; autrement dit, les logit des transitions peuvent dépendre du temps et du carré du temps. Ces modèles sont aussi des modèles de mélange finis qui comprennent trois classes latentes pour tenir compte de l'effet de l'hétérogénéité non observée dans l'état latent initial et dans les probabilités de transition latentes. Ce nombre de classes latentes a été choisi en comparant des variantes des modèles B" et C comprenant différents nombres de classes latentes (les résultats de ces tests sont disponibles sur demande).

Les modèles A', A" et A spécifient des erreurs de classification indépendantes (ECI) pour l'enquête, le registre et les deux ensembles de données, respectivement. Le modèle B' spécifie l'erreur dans l'enquête de façon qu'elle dépende des covariables V_{it} d'âge et d'interview par personne interposée, le modèle B" spécifie des erreurs autocorrélées dans le registre, tandis que le modèle B combine ces deux spécifications. Les modèles C' et C" étendent le modèle B" en introduisant les prédicteurs Z_{it} (sexe, âge, niveau de scolarité et pays d'origine) pour les transitions, d'une part, et pour l'état initial ainsi que les transitions, d'autre part. Le modèle C étend le modèle B en introduisant les mêmes prédicteurs.

Le tableau 4.1 donne les valeurs de la log-vraisemblance, du critère d'information bayésien (BIC) et du critère d'information d'Akaike (AIC), ainsi que le nombre de paramètres pour neuf des modèles qui ont été estimés en utilisant les données de l'EPA et du PA appariées. Dans tous les modèles, les probabilités de transition (latentes) sont supposées être hétérogènes dans le temps; autrement dit, les logit des transitions peuvent dépendre du temps et du carré du temps.

Le modèle A spécifie que les données de l'enquête ainsi que les données du registre contiennent des erreurs de classification (indépendantes). Comme la qualité de l'ajustement de ce modèle aux données est meilleure que celle des modèles A' et A" restreints, qui supposent que seule l'enquête (Modèle A') ou seul le registre (Modèle A") contient des erreurs, nous concluons qu'il s'agit d'une preuve que les sources contiennent toutes deux des erreurs de classification.

Dans les modèles B', B" et B, l'hypothèse ECI est relâchée pour l'enquête, pour le registre, ainsi que pour l'enquête et le registre, respectivement. Plus précisément, l'erreur de mesure dans les données d'enquête peut dépendre de l'âge de l'enquêté et du fait que l'information a été obtenue ou non par personne interposée, et l'erreur de mesure dans les données du registre peut dépendre du type de contrat latent et observé décalé. Cette dernière condition est obtenue en estimant un jeu distinct de probabilités d'erreur pour la répétition de *la même erreur* sur les diverses éditions de l'enquête ou du registre. Les versions restreintes du modèle B sont estimées également pour examiner si la violation de l'hypothèse ECI s'applique à l'erreur de mesure des données d'enquête seulement (modèle B') ou des données du registre seulement (modèle B"). Le fait que le modèle B" soit mieux ajusté que les modèles A et B' indique que l'hypothèse ECI doit être relâchée pour l'indicateur de données du registre. Le modèle B améliore légèrement l'ajustement comparativement au modèle B", ce qui indique que l'hypothèse ECI pour la variable indicatrice d'enquête doit aussi être relâchée dans un modèle sans prédicteurs pour les transitions et pour l'état initial.

Tableau 4.1
Mesures de l'ajustement pour huit modèles estimés au moyen de données de l'EPA et du PA appariées

Modèle	Log-vraisem-blance	BIC (LV)	AIC (LV)	Paramètres	L^2	dl	Valeur p
A': Enquête ECI	-286 814	574 118	573 716	44	240 543,4	69 327	1,6e-18 454
A": Registre ECI	-454 196	908 882	908 480	44	575 307,7	69 327	8,5e-78 021
A: Les deux ECI	-284 413	569 384	568 926	50	235 742,1	69 321	4,8e-17 717
B': A + enquête non-ECI	-283 573	567 748	567 254	54	426 966,7	69 317	6,6e-50 302
B": A + registre non-ECI	-246 054	492 732	492 220	56	435 025,8	69 315	2,9e-51 771
B: A + les deux non-ECI	-246 000	492 669	492 120	60	477 741,8	69 311	7,6e-59 639
C': B" + prédicteurs des transitions	-245 282	491 590	490 748	92	486 186,8	69 279	1,8e-61 222
C": B" + prédicteurs état initial et transitions	-241 990	485 140	484 189	104	479 603,4	69 267	4,9e-60 003
C: B + prédicteurs état initial et transitions	-242 006	485 217	484 229	108	479 635,2	69 263	1,2e-60 010

Nota : Les modèles A', A" et A spécifient des erreurs de classification indépendantes (ECI) pour l'enquête, le registre et les deux ensembles de données, respectivement. Le modèle B' spécifie une erreur dans l'enquête qui dépend de l'âge et de l'interview par personne interposée, le modèle B" spécifie des erreurs autocorrélées dans le registre, tandis que le modèle B combine ces deux spécifications. Les modèles C' et C" étendent le modèle B" en introduisant le sexe, l'âge, le niveau de scolarité et le pays d'origine comme prédicteurs des transitions, ainsi que de l'état initial et des transitions, respectivement. Le modèle C étend le modèle B en introduisant les mêmes prédicteurs. Tous les modèles sont des modèles de mélange finis comportant trois classes latentes pour corriger l'hétérogénéité non observée dans l'état latent initial et dans les probabilités de transition latentes. En outre, tous les modèles supposent que les probabilités de transition latentes sont hétérogènes dans le temps. En particulier, nous conditionnons les probabilités de transition latentes sur une tendance linéaire pour le mois de l'observation ainsi que sur son carré.

Enfin, nous étendons les modèles B" et B en incluant des covariables (sexe, âge, niveau de scolarité et pays d'origine) dans les modèles pour les probabilités de transition latentes et d'état initial latent (modèle C" et C, respectivement). Le modèle C' est une version restreinte du modèle C". Les variables explicatives ne peuvent avoir une incidence que sur les probabilités de transition latentes. Le fait que la qualité de l'ajustement du modèle C" soit meilleure que celles du modèle B" et du modèle C' indique que les covariables ont un effet significatif à la fois sur les transitions et sur les états initiaux. Le fait, d'après deux des trois mesures, que la qualité de l'ajustement du modèle C est moins bonne que celle du modèle C"

signifie que l'hypothèse d'ECI dans les données d'enquête devrait être gardée dans le modèle comprenant des variables explicatives pour les transitions et pour l'état initial (comme le montrent les résultats du modèle C, la taille de l'erreur de mesure dans les données d'enquête ne varie que légèrement avec l'âge et l'interview par personne interposée. Il s'agit d'une preuve supplémentaire en faveur du maintien de l'hypothèse d'ECI pour l'indicateur d'enquête. En fait, les estimations de la taille de l'erreur de mesure dans les données d'enquête ainsi que dans les données de registre et les estimations des probabilités de transition latentes sont très semblables pour les modèles C, C' et C". Cela montre que les résultats de notre modèle sont robustes aux petites erreurs de spécification du modèle). Dans la suite de l'exposé, nous présentons les estimations calculées d'après le modèle C" (les estimations calculées d'après les modèles C et C' sont disponibles sur demande).

Nous avons examiné diverses options de modèle non-ECI. En particulier, nous avons cherché à déterminer si l'erreur de mesure dans les données d'enquête diffère pour des secteurs où la mobilité de contrat et d'emploi est grande, comme le secteur de la santé, mais cela ne s'est pas avéré être le cas. Pour les données de registre, nous avons examiné des spécifications restreintes de rechange pour les erreurs corrélées, mais celles-ci ont donné des modèles dont la qualité de l'ajustement était pire que celle des modèles du tableau 4.1.

Examinons maintenant la quantité d'erreur de classification dans les deux sources de données. Selon l'équation 3.2, pour les données d'enquête et de registre, elle est représentée par les probabilités $P(E_{it} = e_{it} | X_{it} = x_t)$ et $P(C_{it} = c_{it} | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1})$, respectivement. Les estimations d'après le modèle C" sont présentées aux tableaux 4.2 et 4.3. En particulier, le tableau 4.2 montre que l'état de contrat permanent et l'état « autre » sont mesurés de façon très précise dans l'EPA, car presque toutes les personnes dans ces états sont classées correctement. Ce fait est indiqué par les grandes probabilités sur la diagonale principale du tableau. Une certaine erreur est observée pour les personnes qui possèdent en réalité un contrat d'emploi temporaire. 12,5 % de ces personnes déclarent qu'elles ont un contrat d'emploi temporaire, tandis que 4,2 % déclarent être dans une autre situation.

Tableau 4.2
Taille de l'erreur de mesure dans les données d'enquête selon le modèle C"

	Contrat observé à la période t		
Contrat latent en t	**Permanent**	**Temporaire**	**Autre**
Permanent	0,998	0,001	0,002
Temporaire	0,125	0,832	0,042
Autre	0,004	0,005	0,991

Nota : Les erreurs-types sont toujours inférieures à 0,0001.

Le tableau 4.3 donne les probabilités estimées d'erreur de mesure pour les données de registre, qui, selon l'équation 3.2, dépendent des états observé et latent décalés. En raison des restrictions imposées (voir la section 3), nous avons estimé des paramètres (logit) d'erreur distincts pour la répétition de la même erreur entre les mois $t-1$ et t. Ces situations correspondent aux cellules ombrées dans le tableau 4.3. Comme le montre le tableau, les erreurs de mesure sont fortement autocorrélées; autrement dit, si une erreur a été commise au mois $t-1$ et qu'il était possible de répéter la même erreur (si la personne était restée dans le même état latent), l'erreur persistait presque certainement au mois t. Par exemple, si une personne possédant un contrat d'emploi permanent au mois $t-1$ était enregistrée par erreur comme ayant un contrat d'emploi temporaire et qu'elle possédait encore un contrat d'emploi

permanent au mois t, elle avait une probabilité de 0,968 d'être de nouveau enregistrée incorrectement comme ayant un contrat d'emploi temporaire au temps t. Pour les cinq autres erreurs possible, la probabilité d'une erreur de mesure persistante est un peu plus faible, mais n'est jamais inférieure à 0,84.

Un tableau différent se dégage lorsqu'aucune erreur n'est commise au temps $t-1$ ou quand une personne change d'état latent entre $t-1$ et t, et que, par conséquent, aucune répétition de l'erreur n'est possible. Dans ces cas, les données de registre sont presque exemptes d'erreur. Par exemple, quand une personne a été enregistrée correctement comme ayant un contrat d'emploi permanent au temps $t-1$ et qu'elle possède un contrat d'emploi temporaire au temps t, le type de contrat est enregistré correctement comme étant temporaire au temps t avec une probabilité de 0,930. En pratique, cela signifie que l'enregistrement initial du type de contrat est crucial pour le registre PA. Si cet enregistrement est correct, on peut avoir pleinement confiance dans le type de contrat enregistré pour la personne jusqu'à ce qu'un changement réel de situation sur le marché du travail ait eu lieu. En revanche, si le type de contrat de la personne est enregistré incorrectement au départ, cette erreur persistera presque certainement jusqu'à ce que le type de contrat de la personne change.

Tableau 4.3
Probabilités conditionnelles d'erreur de mesure dans les données du registre selon le modèle C"

Contrat observé en $t-1$	Contrat latent en t	Contrat latent en $t-1$	Contrat observé en t		
			Permanent	Temporaire	Autre
Permanent	Permanent	Permanent	0,986	0,009	0,004
Permanent	Permanent	Temporaire	0,986	0,009	0,004
Permanent	Permanent	Autre	0,986	0,009	0,004
Permanent	Temporaire	Permanent	0,045	0,930	0,025
Permanent	Temporaire	Temporaire	0,968	0,032	0,001
Permanent	Temporaire	Autre	0,045	0,930	0,025
Permanent	Autre	Permanent	0,005	0,005	0,990
Permanent	Autre	Temporaire	0,005	0,005	0,990
Permanent	Autre	Autre	0,913	0,000	0,087
Temporaire	Permanent	Permanent	0,027	0,973	0,000
Temporaire	Permanent	Temporaire	0,986	0,009	0,004
Temporaire	Permanent	Autre	0,986	0,009	0,004
Temporaire	Temporaire	Permanent	0,045	0,930	0,025
Temporaire	Temporaire	Temporaire	0,045	0,930	0,025
Temporaire	Temporaire	Autre	0,045	0,930	0,025
Temporaire	Autre	Permanent	0,005	0,005	0,990
Temporaire	Autre	Temporaire	0,005	0,005	0,990
Temporaire	Autre	Autre	0,001	0,842	0,157
Autre	Permanent	Permanent	0,039	0,000	0,961
Autre	Permanent	Temporaire	0,986	0,009	0,004
Autre	Permanent	Autre	0,986	0,009	0,004
Autre	Temporaire	Permanent	0,045	0,930	0,025
Autre	Temporaire	Temporaire	0,005	0,099	0,896
Autre	Temporaire	Autre	0,045	0,930	0,025
Autre	Autre	Permanent	0,005	0,005	0,990
Autre	Autre	Temporaire	0,005	0,005	0,990
Autre	Autre	Autre	0,005	0,005	0,990

Nota : Les erreurs-types sont systématiquement inférieures à 0,0001.

Afin d'estimer la quantité globale d'erreurs dans les données de registre, nous utilisons la probabilité a posteriori de posséder un type particulier de contrat latent à chaque point dans le temps. Cette probabilité

est estimée pour toutes les personnes comprises dans notre échantillon au moyen du modèle de Markov caché. Ces estimations sont assez exactes, car l'erreur de classification n'est que de 0,016. Les moyennes de ces probabilités sur l'ensemble des personnes et des points dans le temps sont présentées au tableau 4.4. En comparant les probabilités sur les diagonales principales des tableaux 4.1 et 4.4, nous voyons que l'erreur est plus grande dans l'indicateur de registre que dans l'indicateur d'enquête. En particulier, les personnes qui travaillent réellement aux termes d'un contrat temporaire ont une probabilité de 0,237 d'être enregistrées comme ayant un contrat permanent (0,125 dans les données d'enquête) et une probabilité de 0,079 d'être enregistrées comme se trouvant dans l'état « autre » dans le PA (0,042 dans les données d'enquête). Il existe également une erreur de classification pour les personnes qui possèdent réellement un contrat d'emploi permanent, car elles ont une probabilité de 0,081 d'être enregistrées comme des travailleurs temporaires et une probabilité de 0,031 d'être enregistrées comme étant dans une autre situation.

Tableau 4.4
Taille de l'erreur de mesure dans les données de registre selon le modèle C"

	Contrat observé en *t*		
Contrat latent en *t*	**Permanent**	**Temporaire**	**Autre**
Permanent	0,888	0,081	0,031
Temporaire	0,237	0,684	0,079
Autre	0,032	0,017	0,951

Nota : Ces probabilités sont égales à la moyenne des probabilités a posteriori d'avoir un type particulier de contrat latent telles qu'elles sont estimées au moyen du modèle C" avec une erreur de classification de 0,016.

Nous nous intéressons non seulement à l'erreur de mesure proprement dite, mais aussi à la grandeur de son effet sur l'estimation de la taille de l'emploi temporaire. En utilisant de nouveau la moyenne des probabilités a posteriori d'avoir un type particulier de contrat latent, nous estimons la taille de l'emploi temporaire aux Pays-Bas. Au tableau 4.5, nous comparons la taille de l'emploi temporaire estimée au moyen du modèle de Markov caché avec les distributions observées du type de contrat selon l'EPA et selon le PA. La probabilité a posteriori moyenne d'avoir un contrat temporaire est de 10,9 % et est comprise entre les valeurs obtenues d'après l'EPA et d'après le PA.

Tableau 4.5
Taille moyenne de l'emploi temporaire selon le modèle C"

	Observée		Latente
	Enquête	**Registre**	
Permanent	0,667	0,597	0,634
Temporaire	0,087	0,130	0,109
Autre	0,246	0,273	0,257
Cas	48 297	174 480	174 480

Nota : Les probabilités latentes sont égales à la moyenne des probabilités a posteriori d'avoir un type particulier de contrat latent telles qu'elles sont estimées au moyen du modèle C" avec une erreur de classification de 0,016.

Le tableau 4.6 donne l'évolution de la taille de l'emploi temporaire d'après les deux sources de données et d'après le modèle de Markov caché. Ce tableau confirme la constatation selon laquelle la taille de l'emploi temporaire calculée d'après notre modèle est comprise entre celles obtenues d'après les données du registre et d'après les données de l'enquête. On voit aussi que, durant la période de référence,

la proportion d'employés temporaires a augmenté. La faible baisse observée dans les données de registre en janvier 2008 (mois 13) comparativement à décembre 2007 (mois 12) s'explique par le fait que de nombreux contrats temporaires se terminent le 31 décembre et que, de surcroît, certains de ces contrats sont convertis en contrats d'emploi permanent. La fluctuation un peu plus importante de la taille de l'emploi temporaire selon les données d'enquête est due au fait que les participants à l'EPA sont interviewés tous les trois mois et que les diverses estimations mensuelles proviennent donc partiellement de différents participants à l'enquête.

Il est important d'examiner non seulement la variation agrégée, mais aussi la variation au niveau individuel, c'est-à-dire la probabilité d'une transition d'un emploi temporaire à un emploi permanent et inversement. Ces probabilités de transition sont présentées au tableau 4.7. En particulier, ce tableau donne les probabilités de transition latentes (moyennes) obtenues d'après le modèle C". Les probabilités de transition ont trait à une période de trois mois et une moyenne est calculée sur les 12 périodes de trois mois dans nos données. Si nous comparons les résultats du tableau 4.7 à ceux du tableau 2.4, nous voyons que les probabilités de transition latentes sont beaucoup plus faibles que celles découlant des données de registre et des données d'enquête. D'après les probabilités de transition latentes, 3,2 % des personnes ayant un contrat d'emploi temporaire possédaient un contrat d'emploi permanent trois mois plus tard, mais selon les données d'enquête et de registre, ces pourcentages étaient de 5,7 % et 8,5 %, respectivement. Cela montre que l'erreur de mesure donne lieu à une inflation de la taille des probabilités de transition. Une telle inflation serait clairement attendue quand les erreurs sont indépendantes dans le temps (Hagenaars 1990, 1994). Si les erreurs ne sont pas indépendantes dans le temps, comme dans notre cas, l'attente est moins claire, car les erreurs pourraient soit accroître soit réduire les probabilités de transition, selon la nature et la taille de l'association. La même tendance à la sous-estimation de la stabilité peut être observée pour l'état de contrat permanent : 98,1 % et 96,7 % de personnes sont restées dans cet état selon les données de l'enquête et les données du registre, respectivement, alors que la stabilité réelle était de 98,7 %.

Tableau 4.6
Évolution de la proportion d'employés temporaires durant la période de janvier 2007 à mars 2008

	Source		
Mois	**Enquête**	**Registre**	**Latente**
1	0,080	0,123	0,102
2	0,082	0,124	0,103
3	0,085	0,123	0,102
4	0,084	0,128	0,103
5	0,084	0,129	0,103
6	0,090	0,129	0,104
7	0,089	0,130	0,105
8	0,087	0,131	0,106
9	0,091	0,135	0,110
10	0,087	0,134	0,112
11	0,088	0,135	0,114
12	0,091	0,135	0,114
13	0,090	0,131	0,116
14	0,089	0,131	0,118
15	0,096	0,132	0,121

Nota : Les données d'enquête comprennent des observations trimensuelles par personne, tandis que les données de registre comprennent des observations mensuelles par personne. Les probabilités latentes sont égales à la moyenne des probabilités a posteriori d'avoir un type particulier de contrat latent telles qu'elles sont estimées au moyen du modèle C" avec une erreur de classification de 0,016.

Tableau 4.7
Transitions observées sur 3 mois dans l'EPA et le PA et transitions latentes selon le modèle C

Transitions latentes				
		Permanent	**Temporaire**	**Autre**
Contrat en t-3	Permanent	0,987	0,006	0,007
	Temporaire	0,032	0,931	0,037
	Autre	0,009	0,030	0,961
	Total	0,634	0,110	0,256

Nota: Les probabilités latentes sont égales à la moyenne des probabilités a posteriori d'avoir un type particulier de contrat latent telles qu'elles sont estimées au moyen du modèle C" avec une erreur de classification de 0,016.

5 Conclusion

Dans le présent article, nous avons étudié l'erreur de mesure du type de contrat d'emploi dans l'EPA des Pays-Bas en appariant les données de sa composante longitudinale pour la période de 2007 au début de 2008 à un ensemble de données de registre unique, c'est-à-dire le PA. Nous avons appliqué plusieurs modèles de Markov cachés dans lesquels le type de contrat réel est traité comme un état latent et dans lesquels les données provenant de l'enquête et du registre servent d'indicateurs observés du contrat réel d'une personne. Nous avons modélisé l'erreur de mesure dans les deux sources de données en tenant compte du fait que l'erreur dans le registre est corrélée entre les périodes.

Nos résultats montrent que les données du registre contiennent plus d'erreurs que les données d'enquête et qu'elles ne peuvent donc pas être utilisées comme normes de référence. Cependant, l'amélioration de l'enregistrement initial dans les données du registre peut accroître considérablement leur qualité, car l'erreur de mesure dans l'indicateur du type de contrat provenant de cet ensemble de données est autocorrélée.

L'erreur de mesure résulte en une sous-estimation du pourcentage de personnes possédant un contrat d'emploi temporaire. Dans l'EPA, ce pourcentage est de 8,9 %, mais après correction de l'erreur de mesure, il passe à 10,9 %. Un autre effet de l'erreur de mesure est qu'elle produit des probabilités de transition fortement surestimées. Selon l'EPA et le PA, la probabilité de transition entre l'emploi temporaire et l'emploi permanent pour une période de trois mois est de 5,7 % et de 8,5 %, respectivement, tandis que la probabilité de transition latente correspondante n'est que de 3,2 %. Cette constatation est particulièrement importante pour les responsables de l'élaboration des politiques aux Pays-Bas, car elle indique clairement que la mobilité de l'emploi temporaire vers l'emploi permanent est nettement moindre qu'on ne le pensait au départ.

Les résultats de la présente étude demeurent relativement stables pour les diverses spécifications de modèle que nous avons évaluées. Cela montre qu'ils sont robustes aux petites erreurs de spécification du modèle. Cependant, ils restent quelque peu dépendants des hypothèses du modèle. D'autres applications et tests de sensibilité permettront de vérifier plus en profondeur la validité de nos résultats. De futures études pourraient se concentrer sur des tests de sensibilité en faisant appel à des simulations Monte Carlo.

Remerciements

Les auteurs sont reconnaissants à l'endroit de *Statistics Netherlands* pour l'accès aux données de cet article. Les auteurs remercient aussi Frank van der Pol, Wendy Smits, Ruben van Gaalen et les participants aux conférences de l'ESPE et de l'EALE ainsi que les participants du groupe de recherche SILC de l'Université libre d'Amsterdam pour leurs pertinents commentaires et suggestions. La contribution de Jeroen Vermunt a été supportée par la *Netherlands Organization for Scientific Research* (NWO) [Numéro de subvention VICI 453-10-002].

Bibliographie

Abowd, J.M., et Stinson, M.H. (2005). Estimating measurement error in SIPP annual job earnings: A comparison of census survey and SSA administrative data. Papier technique, *U.S. Census Bureau*.

Bartolucci, F., Lupparelli, M. et Montanari, G.E. (2009). Latent markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *Annals of Applied Statistiscs*, 3(2), 611-636.

Bassi, F., Hagenaars, J.A., Croon, M.A. et Vermunt, J.K. (2000). Estimating true changes when categorical panel data are affected by uncorrelated and correlated classification errors. *Sociological Methods and Research*, 29(2), 230-268.

Baum, L.E., Petrie, T., Soules, G. et Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1), 164-171.

Bentolila, S., et Bertola, G. (1990). Firing costs and labour demand: How bad is eurosclerosis? *The Review of Economic Studies*, 57(3), 381-402.

Biemer, P. (2004). Une analyse de l'erreur de classification pour les questions sur l'emploi révisées de la Current Population Survey. *Techniques d'enquête*, 30, 2, 141-155.

Biemer, P. (2011). *Latent Class Analysis of Survey Error*. New Jersey : John Wiley & Sons, Inc.

Biemer, P.P., et Bushery, J.M. (2000). Validité de l'analyse markovienne de structure latente pour l'estimation de l'erreur de classification des données sur la population active. *Techniques d'enquête*, 26, 2, 157-171.

Bollinger, C.R. (1996). Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics*, 73(2), 387-399.

Booth, A.L. (1997). An analysis of firing costs and their implications for unemployment policy. Dans *Unemployment Policy*, (Éds., D.J. Snower et G. de la Dehesa). Cambridge : Cambridge University Press.

Bound, J., Brown, C., Duncan, G.J. et Rodgers, W.L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 12(3), 345-368.

Bound, J., Brown, C. et Mathiowetz, N. (2001). Measurement error in survey data. Dans *Handbook of econometrics*, (Éds., J.J. Heckman et E. Leamer), Amsterdam : Elsevier, 5, 3705-3843.

Brown, C., et Medoff, J.L. (1996). Employer characteristics and work environment. *Annales D'Économie et de Statistique*, 41, 275-298.

Cahuc, P., et Postel-Vinay, F. (2002). Temporary jobs, employment protection and labor market performance. *Labour Economics*, 9(1), 63-91.

Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1), 1-38.

Duncan, G.J., et Hill, D.H. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*, 3(3), 508-522.

Gottschalk, P. (2005). Downward nominal-wage flexibility: Real or measurement error. *Review of Economics and Statistics*, 87(3), 556-568.

Gottschalk, P., et Huynh, M. (2010). Are earlings inequality and mobility overstated? The impact of non-classical measurement error. *Review of Economics and Statistics*, 92(2), 302-315.

Hagenaars, J.A. (1990). *Categorical Longitudinal Data Log-Linear Panel, Trend and Cohort Analysis.* Newbury Park, CA : Sage Publications.

Hagenaars, J.A. (1994). Latent variables in log-linear models of repeated observations. Dans *Latent Variable Analysis: Applications for Developmental Research*, (Éds., A. von Eye et C.C. Clogg). Thousand Oaks, CA : Sage Publications, 329-352.

Heckman, J.J., et Singer, B.L. (1984). A method for minimising the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52 (2), 271-320.

Hilbers, P., Houwing, H. et Kösters, L. (2011). De flexibele schil – overeenkomsten en verschillen tussen CBS- en UWV-cijfers [the flexible periphery – similarities and differences between CBS and UWV-data]. Dans *Sociaaleconomische Trends, 2e Kwartaal 2011 [Socioeconomic Trends, 2nd Trimester 2011]*, (Éds., B. Hermans et coll.). Den Haag/Heerlen : Statistics Netherlands, 26-33.

Kapteyn, A., et Ypma, J.Y. (2007). Measurement error and misclassification: A comparison of survey and register data. *Journal of Labor Economics*, 25(3), 513-551.

Langeheine, R. (1994). Latent variable markov models. Dans *Latent Variables Analysis. Applications for Developmental Research*, (Éds., A. von Eye et C. Clogg). Thousand Oaks, Californie : Sage Publications, 373-395.

Manzoni, A., Vermunt, J.K., Luijkx, R. et Muffels, R. (2010). Memory bias in retrospectively collected employment careers: A model-based approach to correct for measurement error. *Sociological Methodology*, 40(1), 39-73.

Mars, G. (2011, décembre). *Cijfers over Flexibele Arbeidsrelaties - Confrontatie Van Bronnen en Defiities [Figures on Flexible Labour Relations - Confrontation of Sources and Definitions]*. Statistics Netherlands, rapport nr SAH-2011-H11. La Haye/Heerlen.

Mathiowetz, N.A. (1992). Errors in reports of occupations. *Public Opinion Quarterly*, 56(3), 352-355.

OECD (2002). *Employment Outlook* 2002. Paris : Auteur.

OECD (2012). *Country Statistical Profiles*. Base de données de l'OECD : consultée le 16/12/2012 à partir de http://stats.oecd.org/.

Paas, L.J., Vermunt, J.K. et Bijmolt, T.H. (2007). Discrete-time discrete-state latent markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society, Series A*, 170(4), 955-974.

Pavlopoulos, D., Muffels, R. et Vermunt, J.K. (2012). How real is mobility between low pay, high pay and non-employment. *Journal of Royal Statistical Society, Series A*, 175(3), 749-773.

Pischke, J.-S. (1995). Measurement error and earnings dynamics: Some estimates from the PSID validation study. *Journal of Business and Economic Statistics*, 13(3), 305-314.

Poulsen, C.S. (1990). Mixed markov and latent markov modelling applied tobrand choice behaviour. *International Journal of Research in Marketing*, 7(1), 5-19.

Rendtel, U., Langeheine, R. et Berntsen, R. (1998). The estimation of poverty dynamics using different measurements of household income. *Review of Income and Wealth*, 44(1), 81-98.

Rodgers, W.L., Brown, C. et Duncan, G.J. (1993). Errors in survey reports of earnings, hours worked, and hourly wages. *Journal of the American Statistical Association*, 88(3), 345-368.

Sels, L., et Van Hootegem, G. (2001). Seeking the balance between flexibility and security: A rising issue in the low countries. *Work, Employment and Society*, 15(2), 327-352.

Shorrocks, A.F. (1976). Income mobility and the markov assumption. *Economic Journal*, 86, 566-578.

Skinner, C.J., Holt, D. et Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Wiley.

van der Pol, F., et Langeheine, R. (1990). Mixed markov latent class models. *Sociological Methodology*, 20, 213-247.

Vermunt, J.K. (1997). *Log-Linear Models for Event Histories*. Londres : SAGE publications.

Vermunt, J.K., et Magidson, J. (2008). *LG - Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*. Belmont Massachussetts : Statistical Innovations Inc.

Vermunt, J.K., Langeheine, R. et Böckenholt, U. (1999). Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24, 178-205.

Vermunt, J.K., Tran, B. et Magidson, J. (2008). Latent class models in longitudinal research. Dans *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, (Éd., S. Menard). Burlington, MA : Elsevier, 373-385.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 225-247
Statistique Canada, n° 12-001-X au catalogue

Cadre généralisé pour la détermination des probabilités d'inclusion optimales dans les plans de sondage à un degré pour des enquêtes à plusieurs variables et plusieurs domaines

Piero Demetrio Falorsi et Paolo Righi[1]

Résumé

L'article décrit un cadre généralisé de calcul des probabilités d'inclusion optimales dans divers contextes d'enquête dans lesquels il est requis de diffuser des estimations d'enquête d'une précision préétablie pour de multiples variables et domaines d'intérêt. Le cadre permet de définir des plans de sondage stratifiés classiques ou incomplets. Les probabilités d'inclusion optimales sont obtenues en minimisant les coûts au moyen d'un algorithme qui garantit l'établissement de bornes pour les erreurs d'échantillonnage au niveau du domaine, en supposant que les variables d'appartenance au domaine sont disponibles dans la base de sondage. Les variables cibles sont inconnues, mais peuvent être prédites au moyen de modèles de superpopulation appropriés. L'algorithme tient compte correctement de l'incertitude de ces modèles. Certaines expériences basées sur des données réelles montrent les propriétés empiriques de l'algorithme.

Mots-clés : Répartition optimale; stratification multidimensionnelle; estimations de domaine; échantillonnage équilibré.

1 Introduction

Les enquêtes menées dans le contexte de la statistique officielle produisent fréquemment un grand nombre d'estimations qui ont trait à différents paramètres d'intérêt ainsi qu'à des domaines d'estimation d'un niveau de détail très élevé. Lorsque des variables indicatrices de domaine sont disponibles pour chaque unité d'échantillonnage figurant dans la base de sondage, le concepteur du plan de sondage peut essayer de sélectionner un échantillon dans lequel la taille de chaque domaine est fixée. Dans ces conditions, il est possible d'obtenir des estimations directes pour chaque domaine et de contrôler les erreurs d'échantillonnage au niveau du domaine. Nous présentons ici un cadre *unifié* et *général* pour définir les *probabilités d'inclusion optimales* pour les *plans d'échantillonnage à un degré* lorsqu'on connaît les variables d'appartenance au domaine à l'étape de l'établissement du plan. Il pourrait s'agir du scénario le plus fréquent dans les enquêtes auprès des établissements et dans d'autres contextes d'enquête, comme les enquêtes agricoles ou les enquêtes sociales si les domaines sont de nature géographique (par exemple, type de municipalité, région, province, etc.). La progression croissante de l'intégration des données des registres administratifs et des bases de sondage pourrait aussi rendre l'approche présentée ici plus applicable aux enquêtes sociales. La proposition pourrait être utile pour la planification d'un sondage de deuxième phase optimal si l'on a recueilli les données sur les variables d'appartenance au domaine à la première phase.

Le problème de l'établissement de plans de sondage optimaux a été abordé dans certains articles récents. Gonzalez et Eltinge (2010) donnent un aperçu intéressant des approches en vue de définir des stratégies d'échantillonnage optimales. Le problème d'optimisation est habituellement traité dans le

1. Piero Demetrio Falorsi, FAO, Viale delle Terme di Caracalla, Roma. Courriel : piero.falorsi@fao.org; Paolo Righi, ISTAT Via C. Balbo 16, 00184 Roma. Courriel : parighi@istat.it.

contexte de l'échantillonnage stratifié avec taille d'échantillon fixe dans chaque strate. La répartition optimale sous échantillonnage stratifié pour une population univariée est bien décrite dans la littérature sur l'échantillonnage (Cochran 1977). Dans les cas multivariés, où plus d'une caractéristiques doivent être mesurées sur chaque unité échantillonnée, la répartition optimale pour les caractéristiques individuelles est de peu d'intérêt pratique, à moins que les diverses caractéristiques étudiées soient fortement corrélées. Il en est ainsi parce qu'une répartition optimale pour une caractéristique est généralement loin de l'être pour les autres. La multidimensionalité du problème mène à la définition d'une méthode de répartition de compromis (Khan, Mati et Ahsan 2010) associée à une perte de précision comparativement aux répartitions optimales individuelles. Plusieurs auteurs ont discuté de divers critères permettant d'obtenir une répartition de compromis réalisable – voir, par exemple, Kokan et Khan (1967), Chromy (1987), Bethel (1989), Falorsi et Righi (2008), Falorsi, Orsini et Righi (2006) et Choudhry, Rao et Hidiroglou (2012).

Récemment, certains articles ont porté sur la recherche des probabilités d'inclusion optimales sous échantillonnage équilibré (Tillé et Favre 2005; Chauvet, Bonnéry et Deville 2011), une classe générale de plans d'échantillonnage qui inclut les plans d'échantillonnage stratifiés comme cas particuliers. Plus précisément, Chauvet et coll. (2011) proposent l'adoption de l'algorithme du point fixe pour définir les probabilités d'inclusion optimales. Néanmoins, les articles susmentionnés n'abordent pas le cas où les variables d'équilibrage dépendent des probabilités d'inclusion et ne présentent qu'une solution partielle au problème dû au fait que la variance d'échantillonnage est une fonction **implicite** des probabilités d'inclusion. Choudhry et coll. (2012) propose un algorithme de répartition optimale pour les estimations de domaine sous échantillonnage stratifié (si les domaines d'estimation ne recoupent pas les strates). Leur algorithme représente un cas particulier de l'approche que nous proposons. Les conditions méthodologiques illustrées ici représentent une amélioration considérable par rapport à la version antérieure de la méthodologie décrite dans Falorsi et Righi (2008) qui ne tenait compte que du cas où les valeurs des variables d'intérêt étaient connues et où la mesure de la précision était exprimée par la variance sous le plan; en outre, la version antérieure ne tenait pas compte du fait que la variance sous le plan, bornée dans le problème d'optimisation, est une fonction implicite des probabilités d'inclusion. Le présent article porte sur le cas plus réaliste où les variables d'intérêt ne sont pas connues et doivent être estimées. En outre, il traite explicitement le problème découlant du fait que les variances anticipées sont des fonctions implicites des probabilités d'inclusion. Le nouvel algorithme d'optimisation peut être exécuté facilement, parce qu'il est fondé sur une décomposition générale de la mesure de la précision. Nous proposons un plan d'échantillonnage général qui englobe la plupart des plans d'échantillonnage à un degré adoptés dans les enquêtes réelles, par exemple l'échantillonnage aléatoire simple sans remise (EASSR), l'EASSR stratifié, l'échantillonnage PPT stratifié, les plans avec stratification incomplète, etc. Le cadre est fondé sur l'utilisation conjointe de *plans d'échantillonnage équilibrés* (Deville et Tillé 2004) qui, suivant les différentes définitions des équations d'équilibrage, représentent une vaste gamme de plans d'échantillonnage et de *modèles de superpopulation pour la prédiction* des valeurs inconnues des variables d'intérêt. La présentation de l'article est la suivante. À la section 2, nous exposons les définitions et la notation. À la section 3 et à la section 4, nous illustrons le plan d'échantillonnage et la variance anticipée. À la section 5, nous décrivons l'algorithme utilisé pour définir les probabilités d'inclusion optimales. À la section 6, nous illustrons les propriétés empiriques de l'algorithme au moyen de certaines expériences fondées sur des données réelles sur les entreprises. Enfin, à la section 7, nous présentons les conclusions.

2 Définitions et notation

À la présente section, nous exposons les concepts du ***domaine d'estimation*** et du ***domaine planifié*** qui jouent un rôle clé dans le cadre présenté ici.

Soit U la population de référence de N éléments et soit $U_d\ (d = 1,\ldots,D)$ un ***domaine d'estimation***, c'est-à-dire une sous-population générique de U contenant N_d éléments, pour laquelle des estimations distinctes doivent être calculées. Soit y_{rk} la valeur de la $r^{\text{e}}\ (r = 1,\ldots,R)$ variable d'intérêt attachée à la k^{e} unité de population et soit γ_{dk} l'indicateur d'appartenance au domaine pour l'unité k défini par

$$\gamma_{dk} = \begin{cases} 1 & \text{si } k \in U_d \\ 0 & \text{autrement} \end{cases}. \tag{2.1}$$

Nous supposons que les valeurs de γ_{dk} sont disponibles dans la base de sondage et que plus d'une valeur de $\gamma_{dk}\ (d = 1,\ldots,D)$ peut être égale à 1 pour chaque unité k; par conséquent, les domaines d'estimation peuvent se chevaucher.

Les paramètres d'intérêt sont les $D \times R$ totaux de domaine

$$t_{(dr)} = \sum_{k\in U} y_{rk}\gamma_{dk}\ (r = 1,\ldots,R; d = 1,\ldots,D). \tag{2.2}$$

Soit $p(\cdot)$ un plan d'échantillonnage sans remise à un degré et $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_k,\ldots,\pi_N)'$ le vecteur de dimension N des probabilités d'inclusion. Soit s l'échantillon sélectionné avec la probabilité $p(s)$. Désignons par $U_h\ (h = 1,\ldots,H)$ la sous-population de taille $N_h = \sum_{k\in U_h}\delta_{hk}$ où $\delta_{hk} = 1$ si $k \in U_h$ et $\delta_{hk} = 0$ autrement.

Nous nous concentrons que les plans d'échantillonnage à taille fixe qui sont ceux qui satisfont

$$\sum_{k\in s}\boldsymbol{\delta}_k = \mathbf{n}, \tag{2.3}$$

où $\boldsymbol{\delta}_k = (\delta_{1k},\ldots,\delta_{hk},\ldots,\delta_{Hk})'$ et $\mathbf{n} = (n_1,\ldots,n_h,\ldots,n_H)'$ est le vecteur de nombres entiers définissant les tailles d'échantillon fixées au moment de l'établissement du plan d'échantillonnage. Puisque la taille d'échantillon n_h, qui correspond à U_h, ne varie pas d'une sélection d'échantillon à l'autre, la sous-population U_h sera appelée ***domaine planifié*** dans la suite de l'exposé. Une condition nécessaire, mais non suffisante, pour s'assurer que (2.3) soit satisfaite est que le vecteur $\boldsymbol{\pi}$ soit tel que

$$\sum_{k\in U}\pi_k\boldsymbol{\delta}_k = \mathbf{n}. \tag{2.4}$$

Dans notre configuration, les domaines planifiés peuvent se chevaucher; par conséquent, l'unité k peut posséder plus d'une valeur $\delta_{hk} = 1$ (pour $h = 1,\ldots,H$). Supposons que les valeurs de δ_{hk} sont connues et disponibles dans la base de sondage pour toutes les unités de la population. Supposons en outre que la matrice $(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_k,\ldots,\boldsymbol{\delta}_N)'$ de dimensions $N \times H$ n'est pas singulière.

Les domaines planifiés et leur relation avec les domaines d'estimation jouent un rôle central dans notre cadre généralisé. Nous supposons que les domaines d'estimation peuvent être définis comme un agrégat de domaines planifiés complets, de sorte que la taille d'échantillon *prévue* dans le d^{e} domaine

d'estimation U_d, disons n_d, peut être obtenue sous forme d'un agrégat simple des tailles d'échantillon prévues des domaines planifiés inclus. Enfin, soit $\hat{t}_{(dr)}$ l'estimateur d'Horvitz-Thompson (HT) de $t_{(dr)}$ avec

$$\hat{t}_{(dr)} = \sum_{k \in s} \frac{1}{\pi_k} y_{rk} \gamma_{dk}. \tag{2.5}$$

Un exemple tiré des enquêtes-entreprises. Supposons que l'on doive calculer les estimations d'enquête séparément en considérant trois types de domaines, à savoir la *région* (20 modalités), l'*activité économique* (2 modalités : biens ou services) et la *taille de l'entreprise* (3 modalités : petite, moyenne ou grande). Autrement dit, il existe $D = 20 + 2 + 3 = 25$ domaines d'estimation chevauchants possibles. Les domaines planifiés peuvent être définis selon différentes options.

Option 1. Le domaine planifié unique U_h est défini par une intersection spécifique des catégories des domaines d'estimation. Dans ce cas, $H = 20 \times 2 \times 3 = 120$ domaines planifiés sont définis. Ils représentent une partition particulière de U. Les domaines planifiés ne se chevauchent pas et $\sum_h \delta_{hk} = 1$.

Option 2. Les domaines planifiés U_h coïncident avec les domaines d'estimation. Par conséquent, $H = D = 25$ et les $\boldsymbol{\delta}'_k$ sont définis comme des vecteurs contenant trois 1, de sorte que $\sum_h \delta_{hk} = 3$. Rappelons que les domaines planifiés se chevauchent.

Option 3. Les domaines planifiés U_h sont définis par *i*) la région selon l'activité économique et *ii*) l'activité économique selon la taille d'entreprise; alors, $H = (20 \times 2) + (2 \times 3) = 46$ avec $\sum_h \delta_{hk} = 2$.

D'autres relations intermédiaires entre les domaines d'estimation et les domaines planifiés sont possibles.

Soulignons que les domaines planifiés représentent le fondement pour la définition de classes plus générales de plans d'échantillonnage. Par exemple, les ***plans d'échantillonnage stratifiés*** requièrent que les domaines planifiés ne se chevauchent pas, car $\sum_h \delta_{hk} = 1$ et que chaque U_h est désignée comme étant une strate. Par conséquent, l'option 1 de l'exemple qui précède nous mène à définir un plan d'échantillonnage stratifié. En outre, les strates définies comme dans l'option 1 servent de fondement à ce que l'on appelle un « plan d'échantillonnage stratifié multimensionnel » (Winkler 2001).

Si $\sum_h \delta_{hk} > 1$, les tailles d'échantillon des domaines planifiés définies dans l'option 1 (strates) ne sont pas strictement contrôlées. Néanmoins, elles restent contrôlées à un niveau agrégé. Dans l'option 2 de l'exemple susmentionné, les tailles d'échantillon sont contrôlées uniquement pour les domaines d'estimation; par contre, dans l'option 3, les tailles d'échantillon sont contrôlées pour les sous-ensembles de deux partitions différentes, définies par *i*) la région selon l'activité économique et *ii*) l'activité économique selon la taille d'entreprise. En nous basant sur la définition de Winkler, nous désignons les plans utilisant ces types de domaines planifiés comme étant des ***plans d'échantillonnage stratifiés multidimensionnels incomplets*** **(ESI)**.

3 Échantillonnage

Soit $\mathbf{z}_k$ un vecteur de variables auxiliaires disponible pour toutes les unités $k \in U$. Un plan d'échantillonnage $p(s)$ est dit équilibré sur les variables auxiliaires si, et seulement si, il satisfait les ***équations d'équilibrage*** suivantes

$$\sum_{k \in s} \frac{\mathbf{z}_k}{\pi_k} = \sum_{k \in U} \mathbf{z}_k \tag{3.1}$$

pour chaque échantillon s tel que $p(s) > 0$ (Deville et Tillé 2004). Selon les variables auxiliaires et les probabilités d'inclusion, l'équation (3.1) peut être exactement ou approximativement satisfaite dans chaque échantillon possible; par conséquent, un plan d'échantillonnage équilibré n'existe pas toujours. En spécifiant

$$\mathbf{z}_k = \pi_k \boldsymbol{\delta}_k, \tag{3.2}$$

les équations (3.1) deviennent

$$\sum_{k \in s} \boldsymbol{\delta}_k = \sum_{k \in U} \pi_k \boldsymbol{\delta}_k. \tag{3.3}$$

Dans ce cas, les équations d'équilibrage stipulent que la taille d'échantillon réalisée dans chaque sous-population U_h est égale à la taille prévue. Dans différents contextes, Ernst (1989) et Deville et Tillé (2004; page 905 Section 7.3) ont prouvé que *i*) sous la spécification (3.2) et *ii*) si le vecteur des tailles prévues d'échantillon, données par $\mathbf{n} = \sum_{k \in U} \pi_k \boldsymbol{\delta}_k$, ne contient que des nombres entiers, alors un plan d'échantillonnage équilibré existe toujours. La spécification (3.2) définit des plans d'échantillonnage qui garantissent le respect de l'équation (2.4) sur laquelle nous souhaitons nous concentrer. Deville et Tillé (2004, pages 895 et 905), Deville et Tillé (2005, page 577) et Tillé (2006, page 168) ont montré que plusieurs plans d'échantillonnage habituels peuvent être considérés comme des cas particuliers de l'échantillonnage équilibré, en définissant de manière appropriée les vecteurs $\boldsymbol{\pi}$ et $\boldsymbol{\delta}_k$ de l'équation (3.2). Ces problèmes sont illustrés à la remarque 4.2 et à la section 6. Des échantillons équilibrés peuvent être tirés par la méthode du cube (Deville et Tillé 2004). Cette méthode facilite grandement la sélection sous des plans d'échantillonnage stratifiés incomplets en permettant de contourner les inconvénients de calcul des méthodes fondées sur des algorithmes de programmation linéaire (Lu et Sitter 2002). La méthode du cube satisfait exactement les équations (3.1) quand la spécification (3.2) est vérifiée et que $\mathbf{n}$ est un vecteur de nombres entiers. Dans les cas de l'EASSR et de l'EASSRS, on peut utiliser les méthodes classiques de sélection de l'échantillon, ainsi que la méthode du cube. Deville et Tillé (2005) proposent une approximation de la variance pour l'estimateur HT sous-échantillonnage équilibré

$$E_p \left(\hat{t}_{(dr)} - t_{(dr)}\right)^2 \cong [N/(N-H)]\left[\sum_{k \in U} (1/\pi_k - 1)\, \eta^2_{(dr)k}\right] \tag{3.4}$$

où E_p désigne l'espérance d'échantillon et

$$\eta_{(dr)k} = y_{rk}\gamma_{dk} - \pi_k \boldsymbol{\delta}'_k \left[\mathbf{A}(\boldsymbol{\pi})\right]^{-1} \sum_{j \in U} \pi_j \left(1/\pi_j - 1\right) \boldsymbol{\delta}_j y_{rk} \gamma_{dk} \tag{3.5}$$

avec

$$\mathbf{A}(\boldsymbol{\pi}) = \sum_{j \in U} \boldsymbol{\delta}_j \boldsymbol{\delta}'_j \pi_j (1 - \pi_j). \tag{3.6}$$

Les résultats de simulations donnés récemment dans Breidt et Chauvet (2011) confirment que l'équation (3.4) représente une bonne approximation de la variance d'échantillonnage quand les équations d'équilibrage sont satisfaites exactement. L'estimation de la variance est étudiée dans Deville et Tillé (2005).

4 Variance anticipée

Avant l'échantillonnage, les valeurs de y_{rk} ne sont pas connues et la variance exprimée par la formule (3.4) ne peut pas être utilisée pour planifier la précision de l'échantillonnage à la phase d'élaboration du plan. En pratique, il est nécessaire d'obtenir des valeurs substitutives ou de prédire les valeurs y_{rk} en se basant sur des modèles de superpopulation qui exploitent l'information auxiliaire. La disponibilité croissante d'information auxiliaire (obtenue par intégration des registres administratifs et des bases de sondage) facilite l'usage des prédictions. Sous inférence fondée sur un modèle, on suppose que les valeurs de y_{rk} sont la réalisation d'un modèle de superpopulation M. Le modèle que nous étudions est de la forme suivante :

$$\begin{cases} y_{rk} = f_r(\mathbf{x}_k; \boldsymbol{\beta}_r) + u_{rk} \\ E_M(u_{rk}) = 0 \ \ \forall k; \ E_M(u_{rk}^2) = \sigma_{rk}^2; \ E_M(u_{rk}, u_{rl}) = 0 \ \ \forall k \neq l \end{cases}, \tag{4.1}$$

où $\mathbf{x}_k$ est un vecteur de variables explicatives (disponibles dans la base de sondage), $\boldsymbol{\beta}_r$ est un vecteur de coefficients de régression et $f_r(\mathbf{x}_k; \boldsymbol{\beta}_r)$ est une fonction connue, u_{rk} est le terme d'erreur et $E_M(\cdot)$ désigne l'espérance sous le modèle. Les paramètres $\boldsymbol{\beta}_r$ et les variances σ_{rk}^2 sont supposés connus, quoiqu'en pratique ils sont habituellement estimés. Le modèle (4.1) est spécifique à une variable, et l'on peut utiliser différents modèles pour différentes variables sans créer de difficultés supplémentaires. Comme mesure de l'incertitude, nous considérons la *variance anticipée* (VA) (Isaki et Fuller 1982) :

$$\mathrm{VA}(\hat{t}_{(dr)}) = E_M E_p (\hat{t}_{(dr)} - t_{(dr)})^2. \tag{4.2}$$

Une expression générale pour la VA sous des modèles linéaires a été établie par Nedyalkova et Tillé (2008). Leur formulation s'obtient en considérant une fonction linéaire $f_r(\cdot)$ et un ensemble unique de variables auxiliaires, $\mathbf{x}_k$, utilisé à la fois pour la prédiction des valeurs de y et pour l'équilibrage de l'échantillon. Dans notre contexte, nous avons introduit $\mathbf{x}_k$ et $\mathbf{z}_k = \pi_k \boldsymbol{\delta}_k$, en soulignant que les variables auxiliaires peuvent être différentes pour la prédiction et l'équilibrage. Les variables $\mathbf{x}_k$ doivent être aussi prédictives de y_{rk} que possible, tandis que les variables $\mathbf{z}_k$ jouent un rôle instrumental dans le contrôle des tailles d'échantillon pour les sous-populations.

Dans le contexte considéré ici, en insérant la variance approximative (3.4) dans l'équation (4.2), nous obtenons l'expression approximative de la VA :

$$\mathrm{VAA}(\hat{t}_{(dr)}) = [N/(N-H)] \sum_{k \in U} (1/\pi_k - 1) E_M(\eta_{(dr)k}^2), \tag{4.3}$$

où les termes $\eta^2_{(dr)k}$ de (3.4) sont remplacés par $E_M\left(\eta^2_{(dr)k}\right)$. En définissant

$$\tilde{y}_{rk} = f_r\left(\mathbf{x}_k;\boldsymbol{\beta}_r\right), \tag{4.4}$$

nous pouvons reformuler l'équation (4.3) sous la forme

$$\mathrm{VAA}\left(\hat{t}_{(dr)}\right) = [N/(N-H)]\left[\sum_{k\in U}\frac{1}{\pi_k}\left(\tilde{y}^2_{rk}+\sigma^2_{rk}\right)\gamma_{dk} - \sum_{k\in U}\left(\tilde{y}^2_{rk}+\sigma^2_{rk}\right)\gamma_{dk} - \mathrm{VAA}_{3(dr)}\right], \tag{4.5}$$

où la troisième composante de variance de $\mathrm{VAA}\left(\hat{t}_{(dr)}\right)$ est

$$\begin{aligned}\mathrm{VAA}_{3(dr)} &= \sum_{k\in U}(1-\pi_k)\,a_{(dr)k}(\boldsymbol{\pi})\left[2\tilde{y}_{rk}\gamma_{dk} - \pi_k a_{(dr)k}(\boldsymbol{\pi})\right] \\ &+ \sum_{k\in U}(1-\pi_k)\left[2b_{(dr)k}(\boldsymbol{\pi}) - \pi_k c_{(dr)k}(\boldsymbol{\pi})\right]\end{aligned} \tag{4.6}$$

et $a_{(dr)k}(\boldsymbol{\pi})$, $b_{(dr)k}(\boldsymbol{\pi})$ et $c_{(dr)k}(\boldsymbol{\pi})$ sont des nombres réels définis respectivement par les équations (A1.4), (A1.7) et (A1.8) de l'annexe A1.

Remarque 4.1. L'expression (4.5) est une formule dont le calcul est laborieux mais, à toute fin pratique, ce calcul peut être simplifié au moyen d'une légère approximation à la hausse en posant que $b_{(dr)k}(\boldsymbol{\pi}) = c_{(dr)k}(\boldsymbol{\pi}) = 0$ dans (4.6). La preuve est donnée à l'annexe A3. Une approximation à la hausse est un choix prudent dans ces conditions, puisqu'il évite le risque de définir une taille d'échantillon insuffisante pour la précision attendue.

Remarque 4.2. Le plan EASSRS est obtenu si les domaines planifiés définissent une partition unique de la population (Option 1 de l'exemple à la section 2) et que le modèle (4.1) est spécifié de façon que les valeurs prédites soient $\tilde{y}_{rk} = \bar{Y}_{rh}$ avec $\sigma^2_{rk} = \sigma^2_{rh}$ (pour $k \in U_h$). La VAA devient

$$\mathrm{VAA}\left(\hat{t}_{(dr)}\right) = [N/(N-H)]\sum_{d=1}^{D}\sum_{h\in H_d}\sigma^2_{rh}N_h\left(N_h/n_h - 1\right), \tag{4.7}$$

où H_d est l'ensemble de domaines planifiés inclus dans U_d (voir l'annexe A4). Notons que l'expression (4.7) concorde avec le *résultat 2* de Nedyalkova et Tillé (2008), sauf pour le terme $N/(N-H)$. Si $[N/(N-H)](1/N_h) \approx 1/(N_h - 1)$, l'expression (4.7) approximerait la variance de l'estimation HT sous le plan EASSRS. Il est prouvé que l'approximation susmentionnée est vraie quand le nombre de domaines H reste petit comparativement à la taille globale de la population N, et que les tailles de domaine N_h sont grandes.

5 Détermination des probabilités d'inclusion optimales

Le vecteur des valeurs de π est déterminé en résolvant le problème d'optimisation suivant :

$$\begin{cases}\mathrm{Min}\left(\sum_{k\in U}\pi_k c_k\right) & \\ \mathrm{VAA}\left(\hat{t}_{(dr)}\right) \le \bar{V}_{(dr)} & (d=1,\ldots,D;\, r=1,\ldots,R), \\ 0 < \pi_k \le 1 & (k=1,\ldots,N)\end{cases} \tag{5.1}$$

où c_k est le coût de la collecte de l'information auprès de l'unité k et $\bar{V}_{(dr)}$ est un seuil de variance fixe correspondant à $\hat{t}_{(dr)}$. Le système (5.1) minimise le coût prévu en s'assurant que les variances anticipées soient bornées et que les probabilités d'inclusion soient comprises entre 0 et 1. Si toutes les valeurs de c_k sont des constantes égales à 1, le problème (5.1) minimise la taille d'échantillon. Nous notons que, dans le problème (5.1), les variances σ^2_{rk} figurant dans $\mathrm{VAA}(\hat{t}_{(dr)})$ sont traitées comme étant connues; en pratique, elles doivent être estimées. À la section 6, nous procédons à une évaluation empirique afin d'étudier la sensibilité de la taille d'échantillon globale en utilisant différentes valeurs estimées de σ^2_{rk}.

Pour résoudre (5.1), nous réarrangeons les contraintes d'inégalité afin d'obtenir

$$\sum_{k\in U} \frac{\left(\tilde{y}^2_{rk} + \sigma^2_{rk}\right)\gamma_{dk}}{\pi_k} \le \frac{N-H}{N}\bar{V}_{(dr)} + \sum_{k\in U}\left(\tilde{y}^2_{rk} + \sigma^2_{rk}\right)\gamma_{dk} + \mathrm{VAA}_{3(dr)}. \qquad (5.2)$$

En fixant de manière appropriée les valeurs de $\mathrm{VAA}_{3(dr)}$, le problème d'optimisation devient un problème linéaire convexe séparé (PLCS) classique (Boyd et Vandenberghe 2004). La figure 5.1 illustre le diagramme de cheminement de l'algorithme (un logiciel prototype dans lequel est mis en œuvre l'algorithme est disponible à l'adresse http://www.istat.it/it/strumenti/metodi-e-software/software), qui est structuré en deux boucles emboîtées : la **boucle externe** (BE) et la **boucle interne** (BI). Les deux boucles sont mises à jour en suivant un schéma d'algorithme *du point fixe*. La convergence sous certaines approximations est démontrée à l'annexe A2.

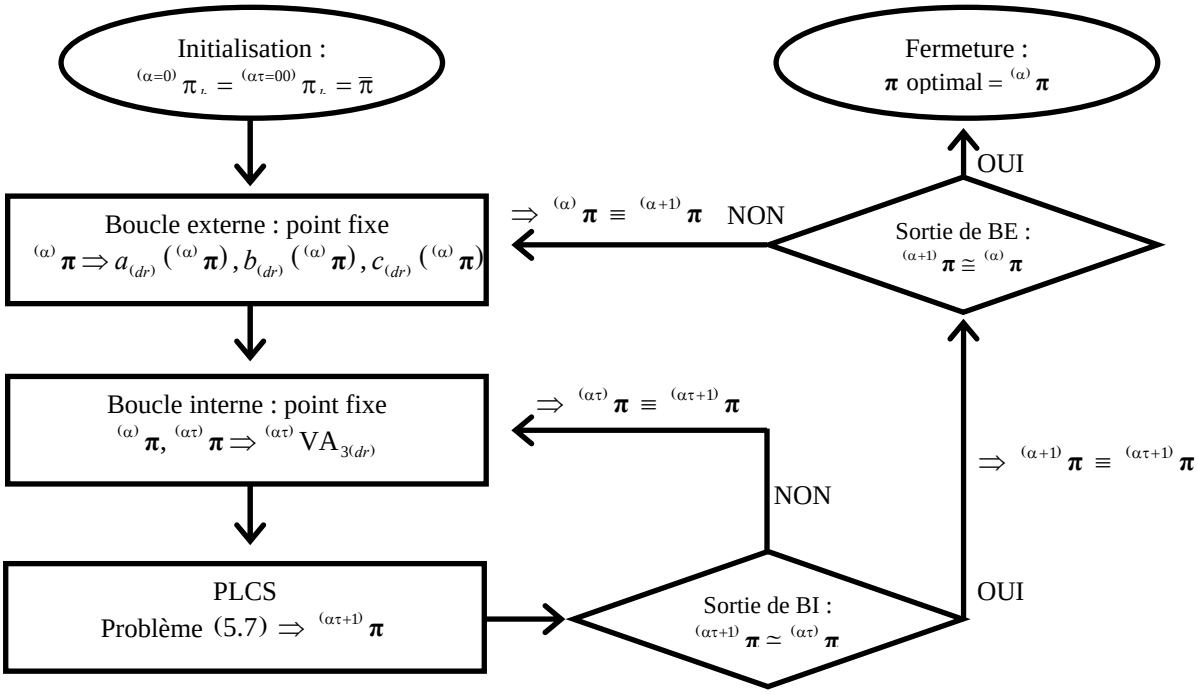


Figure 5.1 Diagramme de cheminement de l'algorithme

Initialisation. À l'itération $\alpha = 0$ de la BE, fixer $^{(\alpha=0)}\boldsymbol{\pi} = \left\{{}^{(\alpha=0)}\pi_k = \bar{\pi}; k = 1,\ldots,N\right\}$ avec $0 < \bar{\pi} \le 1$. Un choix raisonnable est $\bar{\pi} = 0,5$. À l'itération $\tau = 0$ de la boucle interne, fixer $^{(\alpha\tau=0)}\boldsymbol{\pi} = {}^{(\alpha)}\boldsymbol{\pi}$. Fixer le vecteur de dimension N, $\boldsymbol{\varepsilon}$, de faibles valeurs positives.

Boucle externe

- ***Fixation des valeurs pour la boucle interne***. Conformément aux expressions (A1.4), (A1.7) et (A1.8) données à l'annexe A1, les valeurs scalaires réelles suivantes sont calculées

$$a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = \boldsymbol{\delta}_k' \left[\mathbf{A}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]^{-1} \sum_{j\in U} \boldsymbol{\delta}_j \tilde{y}_{rj} \gamma_{dj} \left(1 - {}^{(\alpha)}\pi_j\right), \tag{5.3}$$

$$b_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = \boldsymbol{\delta}_k' \left[\mathbf{A}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]^{-1} \boldsymbol{\delta}_k \sigma_{rk}^2 \gamma_{dk} \left(1 - {}^{(\alpha)}\pi_k\right), \tag{5.4}$$

$$c_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = \pi_k^2 \boldsymbol{\delta}_k' \left[\mathbf{A}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]^{-1} \left[\sum_{j\in U} \boldsymbol{\delta}_j \boldsymbol{\delta}_j' \sigma_{rj}^2 \gamma_{dj} \left(1 - {}^{(\alpha)}\pi_j\right)^2\right] \left[\mathbf{A}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]^{-1} \boldsymbol{\delta}_k. \tag{5.5}$$

- ***Lancement de la boucle interne***. La boucle interne est exécutée jusqu'à la convergence.
- ***Mise à jour ou sortie.*** Si le vecteur ${}^{(\alpha+1)}\boldsymbol{\pi}$ est tel que $\left|{}^{(\alpha+1)}\boldsymbol{\pi} - {}^{(\alpha)}\boldsymbol{\pi}\right| > \boldsymbol{\varepsilon}$, alors la boucle externe est itérée en mettant à jour le vecteur ${}^{(\alpha)}\boldsymbol{\pi}$ avec ${}^{(\alpha+1)}\boldsymbol{\pi}$. Si $\left|{}^{(\alpha+1)}\boldsymbol{\pi} - {}^{(\alpha)}\boldsymbol{\pi}\right| \leq \boldsymbol{\varepsilon}$, alors la bouche externe se ferme et ${}^{(\alpha)}\boldsymbol{\pi}$ représente la solution donnant les valeurs optimales du problème donné par le système (5.1).

Boucle interne

- ***Fixation des valeurs pour le PLCS***. Les valeurs suivantes sont calculées :

$$\begin{aligned} {}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)} &= \sum_{k\in U} \left(1 - {}^{(\alpha\tau)}\pi_k\right) a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) \left[2\tilde{y}_{rk}\gamma_{dk} - {}^{(\alpha\tau)}\pi_k a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right] \\ &+ \sum_{k\in U} \left(1 - {}^{(\alpha\tau)}\pi_k\right) \left[2b_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) - {}^{(\alpha\tau)}\pi_k c_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]. \end{aligned} \tag{5.6}$$

conformément à l'expression (A1.7) à l'annexe A1.

- ***Résolution du PLCS***. En considérant que les valeurs de ${}^{(a\tau)}\mathrm{VAA}_{3(dr)}$ sont fixes, ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ s'obtient en résolvant, au moyen d'un algorithme standard pour un PLCS classique, le problème d'optimisation suivant :

$$\begin{cases} \mathrm{Min}\left(\sum_{k\in U} {}^{(\alpha\tau+1)}\pi_k c_k\right) \\ \sum_{k\in U} \dfrac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{{}^{(\alpha\tau+1)}\pi_k} \leq \dfrac{N-H}{N}\bar{V}_{(dr)} + \sum_{k\in U}\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk} + {}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)}. \\ 0 < {}^{(\alpha\tau+1)}\pi_k \leq 1 \qquad (k = 1,\ldots,N) \end{cases} \tag{5.7}$$

- ***Mise à jour ou sortie.*** Si le vecteur ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ est tel que $\left|{}^{(\alpha\tau+1)}\boldsymbol{\pi} - {}^{(\alpha\tau)}\boldsymbol{\pi}\right| > \boldsymbol{\varepsilon}$, alors la boucle interne est itérée en mettant à jour le vecteur ${}^{(\alpha\tau)}\boldsymbol{\pi}$ avec ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$. Si $\left|{}^{(\alpha\tau+1)}\boldsymbol{\pi} - {}^{(\alpha\tau)}\boldsymbol{\pi}\right| \leq \boldsymbol{\varepsilon}$, alors la boucle interne se ferme et le vecteur mis à jour ${}^{(\alpha+1)}\boldsymbol{\pi}$ pour la boucle externe est donnée par ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$.

Remarque 5.1. Le problème du système (5.7) peut être résolu par l'algorithme proposé dans Falorsi et Righi (2008, section 3.1) qui représente une légère modification de l'algorithme de Chromy (1987), élaboré au départ pour la répartition optimale multivariée sous des plans EASSRS et mis en œuvre dans des outils logiciels standard (voir par exemple le logiciel Mauss-R disponible à l'adresse : http://www3.istat.it/strumenti/metodi/software/campione/mauss_r/). Ou bien, le PLCS peut être traité en se servant de la procédure NLP de SAS comme l'ont proposé Choudhry et coll. (2012).

Remarque 5.2. L'algorithme fait la distinction entre le vecteur ${}^{(\alpha)}\pi_k$ (mis à jour dans la boucle externe) et le vecteur ${}^{(\alpha\tau)}\pi_k$ (mis à jour dans la boucle interne). L'innovation de l'algorithme proposé tient précisément à cette particularité. Si cette distinction entre les probabilités d'inclusion n'est pas faite, c'est-à-dire si ${}^{(\alpha\tau)}\boldsymbol{\pi} = {}^{(\alpha)}\boldsymbol{\pi}$, nous avons observé dans plusieurs expériences que les solutions itérées du PLCS pour chaque boucle externe ne convergent pas vers un point stationnaire.

Remarque 5.3. Après la phase d'optimisation, dans laquelle le vecteur $\boldsymbol{\pi}$ est défini comme étant la solution du problème du système (5.1), une *phase de calage* est exécutée (Falorsi et Righi 2008) afin d'obtenir les probabilités d'inclusion calées, ${}_{\text{cal}}\pi_k$, qui modifient marginalement le vecteur $\boldsymbol{\pi}$ *optimal* afin de satisfaire $\sum_{k\in U} {}_{\text{cal}}\pi_k\boldsymbol{\delta}_k = \mathbf{n}$, où $\mathbf{n}$ est un vecteur de nombres entiers. L'utilisation de l'algorithme d'ajustement proportionnel itératif généralisé (Dykstra et Wollan 1987) permet de s'assurer que toutes les probabilités d'inclusion calées sont comprises dans l'intervalle $(0,1]$.

6 Évaluations empiriques

Plusieurs simulations ont été exécutées sur des ensembles de données réelles et de données simulées pour étudier les propriétés empiriques de la stratégie d'échantillonnage proposée. Ici, nous montrons les résultats obtenus pour un seul exercice portant sur des données réelles se rapportant à la population d'entreprises de 1999 dont le nombre d'employés était compris entre 1 et 99 et qui appartenaient au secteur des Activités informatiques (code à deux chiffres de la *Nomenclature statistique des activités économiques dans la Communauté européenne, Rév. 1*, dont l'acronyme est NACE). Nous avons effectué trois expériences. L'expérience (a) avait pour but de vérifier si la répartition obtenue au moyen de l'algorithme proposé convergeait vers la solution de l'algorithme de Chromy sous le plan EASSRS. L'expérience (b) visait à comparer les tailles d'échantillon du plan EASSRS classique avec celles du plan d'échantillonnage stratifié incomplet (ESI), dans lequel les strates définies par classification croisée étaient des sous-populations non planifiées; cette expérience consistait à étudier le risque de fardeau statistique dû à la sélection répétée lors de différentes éditions de l'enquête. Enfin, l'expérience (c) avait pour objet de mesurer les discordances entre le coefficient de variation (CV) prévu calculé par l'algorithme et le CV empirique obtenu par une simulation Monte Carlo.

Dans les trois expériences, les valeurs de c_k ont été fixées uniformément à 1. La variance anticipée obtenue conformément à l'approximation proposée à la remarque 4.1 a également été calculée.

La taille de la population choisie pour les expériences était de $N = 10\,392$ entreprises. Les domaines d'intérêt définissaient deux partitions de la population cible, à savoir la *région géographique*, avec 20 domaines marginaux (DOM1), et le *groupe d'activités économiques* (code à 3 chiffres de la NACE

avec 6 groupes distincts) *selon la classe de taille* (définie en fonction du nombre d'employés : $1 = 1-4; 2 = 5-9; 3 = 10-19;\ \ 4 = 20-99$), avec 24 domaines marginaux (DOM2). Le nombre global de domaines marginaux était égal à 44, tandis que le nombre de strates formées par classification croisée ou de strates multidimensionnelles ayant une taille de population non nulle était de 360. La valeur modale de la distribution des tailles de population était de 1, et 29,17 % des strates formées par classification croisée ne contenaient au plus que 2 unités. Ce type de strate représente un problème critique dans le contexte des approches d'échantillonnage stratifiées classiques. En effet, pour calculer des estimations de variance sans biais, ces strates doivent être à tirage complet (afin qu'elles ne contribuent pas à la variance des estimations), alors que la règle de répartition exigerait un moins grand nombre d'unités et, en général, un nombre non entier d'unités échantillonnées. Le *coût de la main-d'œuvre* et la *valeur ajoutée* étaient les variables d'intérêt pour lesquelles les données sont fournies par une source administrative pour chaque unité de la population. Habituellement, les deux variables ont une distribution fortement asymétrique.

Pour toutes les études empiriques, les estimations cibles étaient les 88 totaux au niveau du domaine (2 variables fois 44 domaines marginaux). Dans chaque expérience, les probabilités d'inclusion ont été déterminées en fixant la variance $\bar{V}_{(dr)} = (0{,}1 t_{(dr)})^2$ dans (5.1), ce qui équivaut à fixer à 10 % le niveau accepté maximal du CV en pourcentage des estimations au niveau du domaine.

Étude empirique (a). La première expérience tenait compte de la partition DOM1. Ces domaines représentaient à la fois les domaines *planifiés* et les domaines *d'estimation*. Puisque les domaines planifiés définissaient une partition de la population d'intérêt, ils pouvaient également être considérés comme des strates dans les plans d'échantillonnage classiques. Le modèle de travail prédictif était donné par

$$\begin{cases} y_{rk} = \alpha_d + u_{rk} \ \ \forall k \in U_d \ (d = 1, \ldots, 20) \\ E_M(u_{rk}) = 0,\ E_M\left(u_{rk}^2\right) = \sigma_{rd}^2 \ \ \forall k \in U_d;\ \ E_M(u_{rk}, u_{rl}) = 0 \ \ \forall k \neq l \end{cases}, \tag{6.1}$$

où α_d est un effet fixe et les variances dans la superpopulation σ_{rd}^2 étaient estimées au moyen de la variance résiduelle du modèle prédictif dans chaque région. L'algorithme proposé à la section 5 a été exécuté en utilisant trois valeurs initiales distinctes des probabilités d'inclusion $\bar{\pi}$, égales à 0,01, 0,50 et 0,99, respectivement. Les valeurs initiales des probabilités d'inclusion n'avaient aucune incidence sur la solution finale, mais celle-ci était obtenue à la suite d'un nombre différent d'itérations. Nous constatons que le nombre global de boucles internes était de 17 pour $\bar{\pi} = 0{,}01$. La convergence a été obtenue avec 13 boucles internes pour $\bar{\pi} = 0{,}50$; 14 boucles internes ont été nécessaires pour $\bar{\pi} = 0{,}99$. Cependant, après la neuvième itération, les trois tailles d'échantillon étaient relativement similaires (figure 6.1). Dans l'expérience, les tailles d'échantillon globales étaient de 3 105 pour la répartition de Chromy servant de référence et de 3 110 pour la méthode proposée ici. Cependant, les différences entre les deux tailles d'échantillonnage au niveau du domaine étaient des nombres fractionnaires qui étaient toujours inférieurs à 1, et la différence relative absolue la plus importante était inférieure à 0,3 %. Cela met en relief le fait que l'algorithme proposé définit en fait les mêmes tailles d'échantillon de domaine que celles calculées pour la répartition de référence. En ce qui concerne la convergence, les valeurs initiales des probabilités d'inclusion n'ont aucune incidence sur la solution finale, quoique celle-ci soit obtenue moyennant des nombres différents d'itérations.

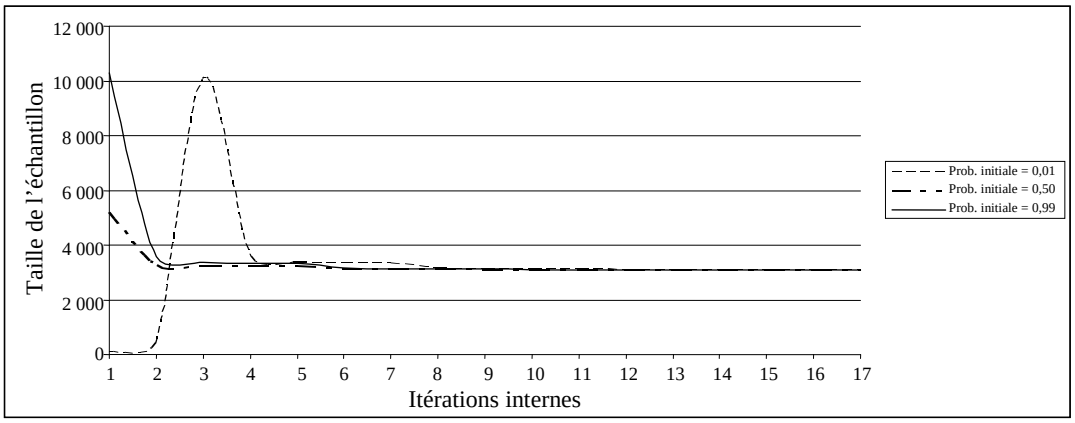


Figure 6.1 Convergence de l'algorithme avec différentes probabilités d'inclusion initiales dans l'étude empirique (a)

Des résultats similaires ont été obtenus quand les domaines d'intérêt étaient définis par la partition DOM2.

Études empiriques (b). Soit U_{d_1} une région particulière $(d_1 = 1,\ldots,20)$ de DOM1, et soit U_{d_2} (avec $d_2 = 1,\ldots,24$) un groupe d'activités économiques particulier selon la classe de taille d'entreprise de la partition DOM2. Nous avons utilisé deux modèles de prédiction, M_1 et M_2. En se référant à la notation des modèles ANOVA, M_1 est le modèle saturé donné par

$$\begin{cases} y_{rk} = \alpha_{d_1} + \lambda_{d_2} + (\alpha\lambda)_{d_1 d_2} + u_{rk} \;\; \forall k \in U_{d_1} \cap U_{d_2} \\ E_M(u_{rk}) = 0,\; E_M\left(u_{rk}^2\right) = \sigma^2_{r(d_1 d_2)} \;\; \forall k \in U_{d_1} \cap U_{d_2};\;\; E_M(u_{rk}, u_{rl}) = 0 \;\; \forall k \neq l \end{cases}, \qquad (6.2)$$

dans lequel α_{d_1} et λ_{d_2} sont les effets principaux, reliés aux domaines U_{d_1} et U_{d_2}, respectivement, et où $(\alpha\lambda)_{d_1 d_2}$ est l'effet d'interaction. Les variances de modèle $\sigma^2_{r(d_1 d_2)}$ ont été estimées par la méthode des moindres carrés ordinaires en calculant les variances des termes résiduels au niveau $U_{d_1} \cap U_{d_2}$. Le modèle M_2 est identique au modèle M_1 sans le facteur d'interaction. Le tableau 6.1 montre la qualité de l'ajustement des deux modèles.

Tableau 6.1
Qualité de l'ajustement des modèles utilisés pour la prédiction

Modèle	Qualité de l'ajustement R^2 %	
	Coût de la main-d'œuvre	Valeur ajoutée
Modèle M_1 (expression 6.2)	68,1	64,1
Modèle M_2 (expression 6.2 sans les interactions)	65,1	61,0

Dans le cas du modèle M_1, nous avons considéré trois répartitions différentes pour l'EASSRS : *i*) aucune contrainte de taille d'échantillon de strate n'est imposée; *ii*) au moins une unité échantillonnée par strate est requise (pour obtenir des estimations ponctuelles sans biais); *iii*) au moins deux unités

échantillonnées par strate sont requises (pour obtenir des estimations de variance sans biais) pour toutes les strates ayant une taille de population de deux entreprises ou plus. Les deux premières répartitions sont plutôt théoriques, puisque dans toutes les enquêtes-entreprises réalisées par l'Institut national de statistique de l'Italie, la sélection d'au moins deux unités par strate est requise. Les résultats de l'expérience sont présentés plus bas au tableau 6.2. Seuls les résultats pour le cas où les probabilités d'inclusion initiales étaient égales à $\bar{\pi} = 0{,}50$ sont examinés ici; des tailles d'échantillon identiques ont été obtenues pour les autres valeurs initiales des probabilités d'inclusion, avec un processus de convergence un peu plus lent. Les trois plans EASSRS comptaient 716,6, 944 et 1 042 unités d'échantillonnage, respectivement. Le plan d'échantillonnage stratifié incomplet (ESI) a donné 936 unités pour le modèle M_1, tandis qu'il a donné 991 unités pour le modèle M_2. Le meilleur résultat donné par le modèle M_1 comparativement au modèle M_2 tenait au fait que son ajustement était meilleur. Enfin, les plans ESI ont aidé à aborder la question du fardeau statistique des entreprises répondantes. En effet, si l'on suppose que les probabilités d'inclusion restent fixes pour les différentes éditions de l'enquête, leurs distributions peuvent être utilisées pour évaluer le fardeau statistique dans les enquêtes répétées. Le tableau 6.2 montre que le nombre d'entreprises sélectionnées avec certitude lors de chaque édition de l'enquête était de 175 pour le troisième plan EASSRS, tandis que 30 et 40 entreprises ont été sélectionnées avec certitude sous le premier et le deuxième plan ESI, respectivement. L'analyse des tailles (mesurées par l'effectif) des entreprises incluses dans l'échantillon avec certitude montre que, dans le cas du troisième plan EASSRS, la taille moyenne était égale à 20,6. Dans certains cas, des entreprises comptant deux employés étaient incluses dans l'échantillon sélectionné avec certitude. Inversement, nous constatons que dans le cas du premier et du deuxième plan ESI, la taille minimale des entreprises était de 17 et 16 employés, respectivement, et que la taille moyenne était supérieure à 40 unités.

Tableau 6.2
Tailles d'échantillon et répartition des entreprises incluses avec certitude dans l'échantillon, pour différents plans d'échantillonnage

Plan d'échantillonnage		Taille de l'échantillon	Entreprises sélectionnées avec certitude		
			Nombre	Nombre d'employés	
				Moyen	Minimum
Stratifié classique avec le modèle M_1	Pas de contrainte de taille d'échantillon de strate	716,6	10	47,0	23,0
	Au moins une unité échantillonnée par strate	944,0	119	24,0	2,0
	Au moins deux unités échantillonnées par strate	1 042,0	175	20,6	2,0
Échantillonnage stratifié incomplet avec le modèle M_1		936,0	30	50,1	17,0
Échantillonnage stratifié incomplet avec le modèle M_2 sans interactions		991,0	40	42,9	16,0

Enfin, pour évaluer la sensibilité de la solution, nous avons répété l'expérience artificiellement et modifié les valeurs de $\tilde{y}_{rk}$ et $\tilde{\sigma}^2_{rk}$ dans le problème d'optimisation (5.1). En particulier, nous avons augmenté les valeurs prédites de $\tilde{\sigma}^2_{rk}$ de 20 % et 120 % respectivement, et diminué de 20 % les valeurs de $\tilde{y}_{rk}$ prédites par le modèle M_1. Comme prévu, les tailles d'échantillon ont augmenté, mais le plan EASSRS avec au moins une unité échantillonnée par strate et le premier plan ESI ont défini approximativement les mêmes tailles d'échantillon (tableau 6.3).

Tableau 6.3
Tailles d'échantillon avec valeurs prévues modifiées des prédictions du modèle (4.1)

Plan d'échantillonnage		Taille de l'échantillon		
		$\tilde{\sigma}^2_{rk}$ augmenté de 20 %	$\tilde{\sigma}^2_{rk}$ augmenté de 120 %	$\tilde{y}_{rk}$ diminué de 20 %
EASSRS avec modèle M_1	Aucune contrainte de taille d'échantillon de strate	821,0	1 269,0	993,8
	Au moins une unité échantillonnée par strate	1 035,0	1 472,0	1 206,0
	Au moins deux unités échantillonnées par strate	1 125,0	1 536,0	1 283,0
Plan ESI avec modèle M_1		1 039,7	1 460,9	1 207,5

Étude empirique (c). Nous avons utilisé le modèle de prédiction linéaire hétéroscédastique M_3 :

$$\begin{cases} y_{rk} = \alpha_r + \varphi_r x_k + u_{rk} \\ E_M(u_{rk}) = 0, \;\; E_M(u_{rk}^2) = \sigma_{rk}^2 = \sigma_r^2 x_k \;\; \forall k \in U; \; E_M(\varepsilon_{rk}, \varepsilon_{rl}) = 0 \;\; \forall k \neq l \end{cases}, \qquad (6.3)$$

où x_k est le *nombre d'employés* dans la k^{e} entreprise, et α_r et φ_r sont les paramètres de régression. Notons que le nombre d'employés est disponible dans la base de sondage en Italie.

Nous avons calculé deux estimations différentes de la variance du modèle :
a) $\tilde{\sigma}^2_{rk} = 1/N_{(X=x_k)} \sum_{k \in U_{(X=x_k)}} (y_{rk} - \mathrm{A}_r - \mathrm{F}_r x_k)^2$ et b) $\tilde{\sigma}^2_{rk} = \tilde{\sigma}^2_r x_k$, dans lesquelles $\tilde{\sigma}^2_r = 1/(N-2) \sum_{k \in U} [(y_{rk} - \mathrm{A}_r - \mathrm{F}_r x_k)/x_k]^2$, où $U_{(X=x)}$ est la population d'entreprises, de taille $N_{(X=x)}$, pour laquelle la variable X prend la valeur x; A_r et F_r sont les estimations de α_r et φ_r, respectivement, par les moindres carrés pondérés pour la population dénombrée complète. La somme des variances de modèle obtenue par la méthode (a) était plus faible que celle obtenue par la méthode (b). Cela a été reflété par les tailles d'échantillon calculées. La première répartition définit une taille d'échantillon global de 927 unités, tandis que la deuxième répartition définit une taille d'échantillon de 951. Nous avons tiré successivement 1 000 échantillons pour chacune des répartitions et avons calculé les ratios $\mathrm{RCV}(\hat{t}_{(dr)}) = \mathrm{CVP}(\hat{t}_{(dr)})/\mathrm{CVS}(\hat{t}_{(dr)})$, avec $\mathrm{CVP}(\hat{t}_{(dr)}) = \left[\sqrt{\mathrm{VAA}(\hat{t}_{(dr)})}/\hat{t}_{(dr)}\right]100$ représentant le CV prévu (%) et

$$\mathrm{CVS}(\hat{t}_{(dr)}) = 100 \sqrt{(1/I)\left[\sum_{i=1}^{I} \hat{t}^i_{(dr)} - (1/I)\sum_{i=1}^{I} \hat{t}^i_{(dr)}\right]^2} \Big/ (1/I)\sum_{i=1}^{I} \hat{t}^i_{(dr)}$$

représentant le CV simulé (ou empirique), obtenu comme résultat de la simulation, en désignant par $\hat{t}^i_{(dr)}$ l'estimation HT dans la i^{e} itération et $I = 1\,000$. Par souci de concision, seuls les principaux résultats de la répartition (b) sont présentés à la figure 6.2 pour DOM1 et DOM2, respectivement, pour les deux variables d'intérêt. En examinant la figure de gauche, nous remarquons que la simulation produit généralement un CV plus petit que le CV prévu, ce qui donne un ratio RCV plus grand que 1 pour les deux variables. Une exception a lieu, pour la valeur ajoutée dans un domaine de DOM1.

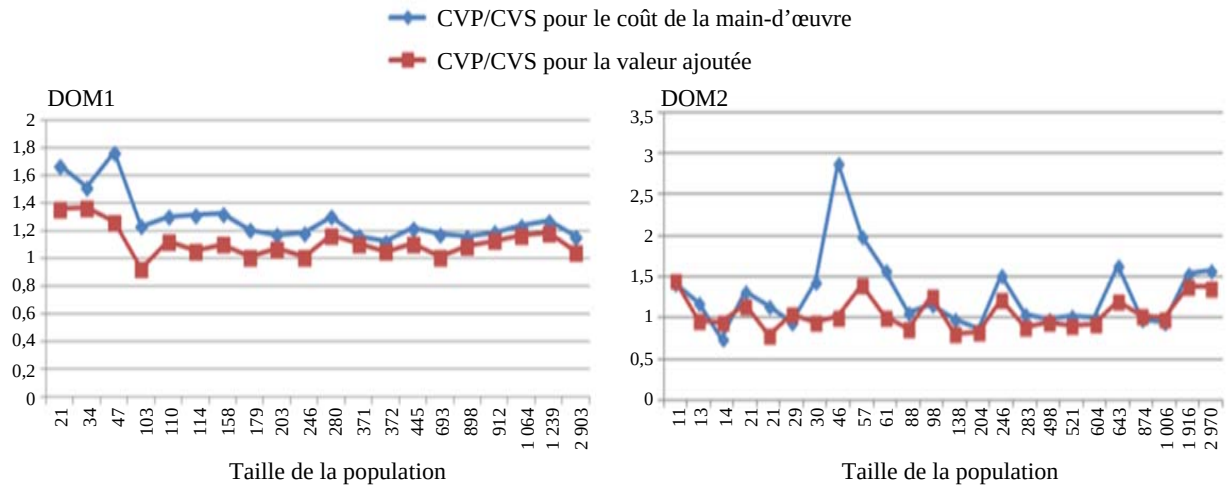


Figure 6.2 RCV selon la taille de la population pour le coût de la main-d'œuvre et la valeur ajoutée

La valeur de RCV inférieure à 1 peut être expliquée par l'augmentation des tailles d'échantillon de domaine en raison de l'étape de calage. Nous constatons qu'en général, ces divergences sont observées dans des domaines dont la taille de population est petite; donc, l'étape de calage peut avoir un effet non négligeable. La figure de droite présente des données empiriques plus articulées et conflictuelles. Premièrement, nous constatons que les RCV sont souvent plus grands que 1 ou très proches de 1. Néanmoins, dans trois domaines, la variable de valeur ajoutée possède un CV simulé égal à 11,5 %, 12,0 % et 12,3 %, respectivement. Dans ces cas rares, et certains autres (coût de la main-d'œuvre dans deux domaines), les divergences sont en harmonie avec les constatations de Deville et Tillé (2005) quant aux propriétés empiriques de l'approximation de la variance pour l'échantillonnage équilibré.

7 Conclusion

L'article décrit une nouvelle approche en vue de déterminer les probabilités d'inclusion optimales dans divers contextes d'enquête caractérisés par la nécessité de diffuser des estimations d'enquête d'une précision préétablie, pour de multiples variables et domaines d'intérêt.

La principale contribution de l'article a trait au calcul pratique de ces probabilités au moyen d'un nouvel algorithme, qui convient pour un plan d'échantillonnage multidimensionnel général dans lequel l'échantillonnage stratifié classique représente un cas particulier. L'approche proposée, l'algorithme et le calcul final sont orientés domaine et variable.

Dans notre cadre, les variables indicatrices d'appartenance à un domaine sont supposées connues, tandis que les variables d'intérêt sont inconnues. La procédure est alors appliquée aux valeurs prédites des caractéristiques d'intérêt au moyen d'un modèle de superpopulation, et l'algorithme permet de tenir compte de l'incertitude du modèle; cela reflète le fait que les valeurs des variables d'intérêt sont inconnues. En utilisant la variance anticipée comme mesure de la précision de l'estimateur, cette approche

permet de contourner les limites des algorithmes standard utilisés pour la répartition des échantillons, dans lesquels les variables d'intérêt dictant la solution sont supposées connues.

L'algorithme proposé exploite une procédure standard, mais présente certaines innovations en matière de calcul qui pourraient être utiles pour faire face à la complexité qui découle du fait que les variances anticipées sont des fonctions implicites des probabilités d'inclusion. L'algorithme a été testé sur des données simulées et des données d'enquête réelles afin d'évaluer sa performance et ses propriétés. Les résultats d'un petit ensemble d'expériences sont présentés ici. Ils confirment une amélioration, en ce qui concerne l'efficacité, de la stratégie d'échantillonnage. Une généralisation naturelle du cas examiné ici peut être élaborée en considérant que les indicateurs de domaine et d'autres variables indépendantes quantitatives sont connus à l'étape de l'élaboration du plan d'échantillonnage. Nous notons que la variance anticipée en ne tenant compte que des indicateurs de domaine est plus grande que la variance anticipée de ce cas plus général. Donc, notre solution représente une borne supérieure (et d'une certaine robustesse) de la solution à la phase de l'élaboration du plan. En outre, la solution algorithmique peut être adaptée facilement à cette situation plus générale.

Remerciements

La présente étude a été financée par le partenariat de la Stratégie mondiale pour l'amélioration des statistiques agricoles et rurales : http://www.fao.org/economic/ess/ess-capacity/strategie-mondiale/fr/.

Annexe A1

VA de l'estimateur HT

Considérons le résidu $\eta_{(dr)k}$ tel qu'il est exprimé par l'équation (3.5), et remplaçons le terme y_{rk} par $\tilde{y}_{rk} + u_{rk}$, ce qui nous donne

$$\eta_{(dr)k} = (\tilde{y}_{rk} + u_{rk})\gamma_{dk} - \pi_k \boldsymbol{\delta}'_k [\mathbf{A}(\boldsymbol{\pi})]^{-1} \sum_{j\in U} \pi_j \boldsymbol{\delta}_j (\tilde{y}_{rj} + u_{rj})\gamma_{dj} (1/\pi_j - 1). \quad \text{(A1.1)}$$

Les moindres prédictions pondérées de $\tilde{y}_{rk}\gamma_{dk}$ et $u_{rk}\gamma_{dk}$, avec les prédicteurs $\pi_k \boldsymbol{\delta}_k$ et les pondérations $1/\pi_k - 1$, sont

$$\hat{\tilde{y}}_{(dr)k} = \pi_k a_{(dr)k} \quad \text{(A1.2)}$$

et

$$\hat{u}_{(dr)k} = \pi_k \boldsymbol{\delta}'_k [\mathbf{A}(\boldsymbol{\pi})]^{-1} \sum_{j\in U} \pi_j \boldsymbol{\delta}_j u_{rj}\gamma_{dj} (1/\pi_j - 1), \quad \text{(A1.3)}$$

avec

$$a_{(dr)k}(\boldsymbol{\pi}) = \boldsymbol{\delta}'_k [\mathbf{A}(\boldsymbol{\pi})]^{-1} \sum_{j\in U} \pi_j \boldsymbol{\delta}_j \tilde{y}_{rj}\gamma_{dk} (1/\pi_j - 1). \quad \text{(A1.4)}$$

En utilisant les formules (A1.2) et (A1.3), l'expression (A1.1) peut être reformulée sous la forme $\eta_{(dr)k} = (\tilde{y}_{rk} + u_{rk})\gamma_{dk} - [\hat{\tilde{y}}_{(dr)k} + \hat{u}_{(dr)k}]$. Par conséquent, l'espérance sous le modèle de $\eta^2_{(dr)k}$ est

$$E_M\left(\eta_{(dr)k}^2\right) = \left(\tilde{y}_{rk}\gamma_{dk} - \hat{\tilde{y}}_{(dr)k}\right)^2 + E_M\left[\left(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k}\right)^2\right] + \text{termes de moyenne nulle,} \quad \text{(A1.5)}$$

car $E_M\left(u_{rk}\right) = 0$. En outre,

$$E_M\left[\left(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k}\right)^2\right] = \sigma_{rk}^2\gamma_{dk} + E_M\left(\hat{u}_{(dr)k}\right)^2 - 2E_M\left(u_{rk}\gamma_{dk}, \hat{u}_{(dr)k}\right), \quad \text{(A1.6)}$$

où $E_M\left(u_{rk}\gamma_{dk}\hat{u}_{(dr)k}\right) = \pi_k b_{(dr)k}\left(\boldsymbol{\pi}\right)$ et $E_M\left(\hat{u}_{(dr)k}\right)^2 = \pi_k^2 c_{(dr)k}\left(\boldsymbol{\pi}\right)$, avec

$$b_{(dr)k}\left(\boldsymbol{\pi}\right) = \boldsymbol{\delta}_k'\left[\mathbf{A}\left(\boldsymbol{\pi}\right)\right]^{-1}\boldsymbol{\delta}_k\sigma_{rk}^2\gamma_{dk}\left(1-\pi_k\right) \quad \text{(A1.7)}$$

et

$$c_{(dr)k}\left(\boldsymbol{\pi}\right) = \boldsymbol{\delta}_k'\left[\mathbf{A}\left(\boldsymbol{\pi}\right)\right]^{-1}\left[\sum_{j\in U}\boldsymbol{\delta}_j\boldsymbol{\delta}_j'\sigma_{rj}^2\gamma_{dj}\left(1-\pi_j\right)^2\right]\left[\mathbf{A}\left(\boldsymbol{\pi}\right)\right]^{-1}\boldsymbol{\delta}_k. \quad \text{(A1.8)}$$

L'expression (4.5) est obtenue facilement en insérant les expressions provenant de (A1.2) à (A1.8) dans l'équation (4.3).

Annexe A2

Convergence de l'algorithme

Le problème d'optimisation (5.1) est résolu par deux *itérations du point fixe* emboîtées. Étant donné un vecteur $\mathbf{x}$ de dimension q inconnu, l'itération du point fixe choisit une valeur supposée initiale ${}^{(0)}\mathbf{x}$. Puis, l'algorithme calcule des itérés subséquents selon ${}^{(\tau+1)}\mathbf{x} = \mathbf{g}\left({}^{(\tau)}\mathbf{x}\right)$, avec $\tau = 1, 2, \ldots$, où $\mathbf{g}(\cdot)$ est un système de q équations de mise à jour. La fonction multivariée $\mathbf{g}$ possède un point fixe dans un domaine $Q \subseteq \Re^q$ si $\mathbf{g}$ applique Q dans Q. Soit $J_{\mathbf{g}}\left(\mathbf{x}\right)$ la matrice jacobéenne de la dérivée partielle première de $\mathbf{g}$ évaluée à $\mathbf{x}$. S'il existe une constante $\rho < 1$ telle que, dans une norme matricielle naturelle, $\left\|J_{\mathbf{g}}\left(\mathbf{x}\right)\right\| \leq \rho, \mathbf{x} \in Q$, $\mathbf{g}$ possède un point fixe unique $\mathbf{x}^* \in Q$, et l'itération du point fixe est garantie de converger vers $\mathbf{x}^*$ pour toute valeur supposée initiale choisie dans Q. En ce qui concerne l'algorithme proposé, la convergence de la boucle interne (BI) et de la boucle externe (BE) est obtenue quand les termes ${}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)}$ convergent vers le point fixe. Cela signifie que les vecteurs ${}^{(\alpha)}\boldsymbol{\pi}$ et ${}^{(\alpha\tau)}\boldsymbol{\pi}$ ne changent pas dans les itérations de la BE et de la BI. Dans la démonstration qui suit, nous considérons la méthode proposée par Chromy (1987) pour résoudre le PLCS du système (5.7), et nous formulons certaines hypothèses raisonnables, à savoir : 1) $\hat{u}_{(dr)k} \cong 0$; 2) $[N/(N-H)] \cong 1$; 3) $\hat{\tilde{y}}_{rk} \cong \tilde{y}_{rk}$; 4) ${}^{(\alpha)}\pi_k \cong {}^{(\alpha\tau)}\Delta\,{}^{(\alpha\tau)}\pi_k$ avec $0 < {}^{(\alpha\tau)}\Delta \leq 1$; 5) $c_k \cong \bar{c}$. L'hypothèse (1) correspond à l'approximation à la hausse de la variance anticipée, donnée à la remarque 4.1, et implique que $b_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = c_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right) = 0$. L'hypothèse (3) implique que $a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\tilde{y}_{rk}\gamma_{dk} \cong \tilde{y}_{rk}^2\gamma_{dk}\,/\,{}^{(\alpha)}\pi_k$. L'hypothèse (4) énonce que la structure des probabilités d'inclusion demeure à peu près constante dans les différentes itérations de la BI. L'hypothèse devient raisonnable compte tenu du fait que l'équation de mise à jour A2.2 qui suit (d'une probabilité d'inclusion donnée) est essentiellement déterminée par le seuil de variance qui requiert la taille d'échantillon la plus grande. Il est plausible d'émettre l'hypothèse que ce seuil demeure plus ou moins le même dans les itérations de la BI subséquentes d'une BE donnée.

Preuve de la convergence de la boucle interne. En reformulant l'expression (4.6) conformément aux hypothèses (1) à (4),

$$ {}^{(\alpha\tau+1)}\mathrm{VAA}_{3(dr)} = \sum_{k\in U}\left[\left(\frac{1}{{}^{(\alpha\tau+1)}\pi_k}-1\right)\left(2\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta}-\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta^2}\right)\right]. \tag{A2.1}$$

En considérant que, dans le problème (5.7), les valeurs de ${}^{(a\tau)}\mathrm{VAA}_{3(dr)}$ sont fixes, chaque valeur du vecteur ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ s'obtient comme une solution du PLCS avec l'algorithme de Chromy. Désignons par $\alpha\tau v^*$ l'itération de l'algorithme de Chromy durant laquelle il converge, où ${}^{(\alpha\tau v^*+1)}\boldsymbol{\pi} \cong {}^{(\alpha\tau v^*)}\boldsymbol{\pi}$. Alors, la BI met à jour la probabilité générique conformément à l'expression

$$ {}^{(\alpha\tau+1)}\pi_k = \left[\sum_{(dr)} {}^{(\alpha\tau v^*+1)}\phi_{(dr)}\frac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{\bar{c}}\right]^{1/2}, \tag{A2.2}$$

où le deuxième terme du membre de droite représente la formule de mise à jour de l'algorithme de Chromy, et $\sum_{(dr)}$ représente $\sum_{d=1}^{D}\sum_{r=1}^{R}$, et ${}^{(\alpha\tau v^*+1)}\phi_{(dr)}$ est le multiplicateur de Lagrange généralisé, où

$$\begin{aligned} {}^{(\alpha\tau v^*+1)}\phi_{(dr)} &= {}^{(\alpha\tau v^*)}\phi_{(dr)}\left[\frac{{}^{(\alpha\tau v^*)}V_{(dr)}}{\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)}}\right]^2, \\ {}^{(\alpha\tau v^*)}V_{(dr)} &= \sum_{k\in U}\frac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{{}^{(\alpha\tau v^*)}\pi_k} \end{aligned} \tag{A2.3}$$

et

$$\ddot{V}_{(dr)} = \bar{V}_{(dr)} + \sum_{k\in U}\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}.$$

La théorie de Kuhn-Tucker énonce que ${}^{(\alpha\tau v^*)}\phi_{(dr)}\left[{}^{(\alpha\tau v^*)}V_{(dr)} - \left(\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{AV}_{3(dr)}\right)\right] = 0$; par conséquent, ${}^{(\alpha\tau v^*+1)}\phi_{(dr)} = {}^{(\alpha\tau v^*)}\phi_{(dr)}$ et ${}^{(\alpha\tau v^*+1)}\phi_{(dr)} > 0$ si et seulement si ${}^{(\alpha\tau v^*)}V_{(dr)}/\left(\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{AV}_{3(dr)}\right) = 1$. Chromy affirme que peu de ${}^{(\alpha\tau v^*)}\phi_{(dr)}$ (pour $r = 1,\ldots,R; d = 1,\ldots,D$) sont plus grands que zéro, et que dans la plupart des cas, une seule valeur est strictement positive. En notant ${}^{(\alpha\tau)}\mathbf{VAA}_3 = \left({}^{(\alpha\tau)}\mathrm{VAA}_{3(11)},\ldots,{}^{(\alpha\tau)}\mathrm{VAA}_{3(1R)},\ldots,{}^{(\alpha\tau)}\mathrm{VAA}_{3(DR)}\right)'$, nous définissons ${}^{(\alpha\tau+1)}\mathbf{VAA}_3 = \mathbf{g}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)$ comme étant le système de $D\times R$ équations de mise à jour, où l'équation $(\overline{dr})$ générique du système

$$\begin{aligned} g_{(\overline{dr})}\left({}^{(\alpha\tau)}\mathbf{VAA_3}\right) &\cong \sum_{k\in U}\left(2\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta}-\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta^2}\right) \\ &\times \left\{\left[\sum_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)}\left[\frac{{}^{(\alpha\tau v^*)}V_{(dr)}}{\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{VAA}_{3(dr)}}\right]^2\frac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{\bar{c}}\right]^{-1/2}-1\right\}, \end{aligned} \tag{A2.4}$$

s'obtient en insérant l'expression (A2.2) dans (A2.1). Si l'on obtient la convergence, alors dans la dernière itération, ${}^{(\alpha\tau+1)}\mathbf{VAA}_3 \cong {}^{(\alpha\tau)}\mathbf{VAA}_3$. La fonction de l'équation (A2.4) est continue et dérivable. En outre,

elle s'applique sur l'intervalle des valeurs possibles de $\mathrm{VAA}_{3(dr)}$. Alors, la BI converge si la condition qui suit est satisfaite :

$$\left\| J_{\mathbf{g}} \left(\mathbf{VAA}_3 \right) \right\| \leq 1. \tag{A2.5}$$

La matrice jacobienne est semi-définie positive, et un résultat bien connu énonce que $\text{trace}\left(J_{\mathbf{g}} J_{\mathbf{g}}'\right) \leq \text{trace}\left(J_{\mathbf{g}}\right)^2$. En considérant la norme de Frobenius $\left\|J_{\mathbf{g}}\right\|_F = \sqrt{\text{trace}\left(J_{\mathbf{g}} J_{\mathbf{g}}'\right)}$, elle devient $\left\|J_{\mathbf{g}}\right\|_F \leq \text{trace}\left(J_{\mathbf{g}}\right)$. Donc, nous pouvons tenir compte de la trace de la matrice jacobienne pour vérifier la condition (A2.5). Soit $g'_{(\overline{dr})} = \partial g_{(\overline{dr})} \left({}^{(\alpha\tau-1)}\mathbf{VAA}_{3(dr)} \big/ \partial\, {}^{(\alpha\tau-1)}\mathrm{VAA}_{3(\overline{dr})} \right)$ l'élément $(\overline{dr})$ de la diagonale de $J_{\mathbf{g}}\left(\mathbf{VAA}_3\right)$. En utilisant la condition de Kuhn-Tucker ${}^{(\alpha\tau v^*)}V_{(dr)} \big/ \left(\ddot{V}_{(dr)} + {}^{(\alpha\tau)}\mathrm{AV}_{3(dr)} \right) = 1$,

$$\begin{aligned} g'_{(\overline{dr})} &= \sum\nolimits_{k\in U} \left(2\frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta} - \frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta^2} \right) \left[\sum\nolimits_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)} \frac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{\overline{c}} \right]^{-3/2} \\ &\times {}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} \frac{1}{{}^{(\alpha\tau v^*)}V_{(\overline{dr})}} \frac{\left(\tilde{y}_{\overline{rk}}^2 + \sigma_{\overline{rk}}^2\right)\gamma_{\overline{dk}}}{\overline{c}}. \end{aligned}$$

Puisque dans de nombreux cas, ${}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} = 0$ (Chromy 1987), l'élément $g'_{(\overline{dr})}$ respectif est nul. Quand ${}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} > 0$, alors

$$g'_{(\overline{dr})} \leq \sum\nolimits_{k\in U} \left(2\frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta} - \frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta^2} \right) \left[{}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} \frac{\left(\tilde{y}_{\overline{rk}}^2 + \sigma_{\overline{rk}}^2\right)\gamma_{\overline{dk}}}{\overline{c}} \right]^{-3/2} \times {}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} \frac{1}{{}^{(\alpha\tau v^*)}V_{(\overline{dr})}} \frac{\left(\tilde{y}_{\overline{rk}}^2 + \sigma_{\overline{rk}}^2\right)\gamma_{\overline{dk}}}{\overline{c}}$$

$$= \sum\nolimits_{k\in U} \left(2\frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta} - \frac{\tilde{y}_{\overline{rk}}^2 \gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta^2} \right) \frac{1}{\sqrt{{}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} \frac{\left(\tilde{y}_{\overline{rk}}^2 + \sigma_{\overline{rk}}^2\right)\gamma_{\overline{dk}}}{\overline{c}}\, {}^{(\alpha\tau v^*)}V_{(\overline{dr})}}}$$

$$\leq \sum\nolimits_{k\in U} \frac{\frac{\tilde{y}_{\overline{rk}}\gamma_{\overline{dk}}}{{}^{(\alpha\tau+1)}\Delta}\left(2 - \frac{1}{{}^{(\alpha\tau+1)}\Delta}\right)}{\sqrt{\overline{c}\, {}^{(\alpha\tau v^*)}\phi_{(\overline{dr})}\gamma_{\overline{dk}}\, {}^{(\alpha\tau v^*)}V_{(\overline{dr})}}} << 1.$$

Par conséquent, la $\text{trace}\left(J_{\mathbf{g}}\right)$ doit être inférieure à 1.

Preuve de la convergence de la boucle externe. Soit ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ la solution du problème de point fixe de la BI; alors, la BE met à jour le vecteur ${}^{(\alpha)}\boldsymbol{\pi}$ avec ${}^{(\alpha+1)}\boldsymbol{\pi} = {}^{(\alpha\tau+1)}\boldsymbol{\pi}$. Sous les conditions (1), (2) et (3),

$${}^{(\alpha+1)}\mathrm{VAA}_{3(dr)} = \sum\nolimits_{k\in U} \left(\frac{1}{{}^{(\alpha\tau+1)}\pi_k} - 1 \right) \tilde{y}_{rk}^2 \gamma_{dk}. \tag{A2.6}$$

En insérant l'expression (A2.2) dans la formule (A2.6) quand la BI converge, le système de $D \times R$ équations de mise à jour de ${}^{(\alpha+1)}\mathbf{VAA}_3$ est donné par ${}^{(\alpha+1)}\mathbf{VAA}_3 = \mathbf{j}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)$, où l'équation générique de $\mathbf{j}$ est

$$
\begin{aligned}
{}^{(\alpha+1)}\mathrm{VAA}_{3(dr)} &= j_{(\bar{d}\bar{r})}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right) \\
&= \sum\nolimits_{k\in U} \tilde{y}_{\bar{r}k}^2 \gamma_{\bar{d}k} \left(\left[\sum\nolimits_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)} \left[\frac{{}^{(\alpha\tau v^*)}V_{(dr)}}{\ddot{V}_{(\bar{d}\bar{r})} + {}^{(\alpha\tau)}\mathrm{VAA}_{3(\bar{d}\bar{r})}} \right]^2 \frac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{\bar{c}} \right]^{-1/2} - 1 \right).
\end{aligned}
\tag{A2.7}
$$

En notant que ${}^{(\alpha)}\mathbf{VAA}_3 = {}^{(\alpha\tau=0)}\mathbf{VAA}_3$, le système **j** peut être exprimé sous une forme récursive

$$
{}^{(\alpha+1)}\mathbf{VAA}_3 \cong \mathbf{j}\left(\mathbf{g}\left({}^{(\alpha\tau-1)}\mathbf{VAA}_3\right)\right) = \mathbf{j}\left(\mathbf{g}\left(\mathbf{g}\left(.....\mathbf{g}\left({}^{(\alpha\tau=0)}\mathbf{VAA}_3\right)\right)\right)\right) = \mathbf{f}\left({}^{(\alpha)}\mathbf{VAA}_3\right),
$$

avec $\mathbf{f}(\cdot) = \mathbf{j}(\mathbf{g}(\mathbf{g}(.....\mathbf{g}(\cdot))))$ en tant que système de $D \times R$ équations de mise à jour de ${}^{(\alpha+1)}\mathrm{VAA}_3$, par rapport aux valeurs antérieures de la BE, ${}^{(\alpha)}\mathbf{VAA}_3$. Pour démontrer la convergence de la BE, il est nécessaire de démontrer que la norme jacobienne $\|J_{\mathbf{f}}(\mathbf{VAA}_3)\|$ est inférieure à 1. En utilisant les résultats classiques de l'algèbre matricielle,

$$
\|J_{\mathbf{f}}(\mathbf{VAA}_3)\| \le \left\|J_{\mathbf{j}}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)\right\| \times \left\|J_{\mathbf{g}}\left({}^{(\alpha\tau-1)}\mathbf{VAA}_3\right)\right\| \times \ldots \times \left\|J_{\mathbf{g}}\left({}^{(\alpha\tau=0)}\mathbf{VAA}_3\right)\right\|,
$$

où la norme générique $\|J_{\mathbf{g}}(\cdot)\|$ est inférieure à 1 (voir la preuve de convergence de la BI). Soit $j'_{(\bar{d}\bar{r})}$ l'élément $(\bar{d}\bar{r})$ de la diagonale de $J_{\mathbf{j}}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)$. Il est donné par

$$
\begin{aligned}
j'_{(\bar{d}\bar{r})} &= \sum\nolimits_{k\in U} \tilde{y}_{\bar{r}k}^2 \gamma_{\bar{d}k} \left[\sum\nolimits_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)} \frac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{\bar{c}} \right]^{-3/2} \\
&\times {}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})} \frac{1}{{}^{(\alpha\tau v^*)}V_{(\bar{d}\bar{r})}} \frac{\left(\tilde{y}_{\bar{r}k}^2 + \sigma_{\bar{r}k}^2\right)\gamma_{\bar{d}k}}{\bar{c}}.
\end{aligned}
\tag{A2.8}
$$

Par conséquent, nous avons

$$
\begin{aligned}
j'_{(\bar{d}\bar{r})} &\le \sum\nolimits_{k\in U} \tilde{y}_{\bar{r}k}^2 \gamma_{\bar{d}k} \left[{}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})} \frac{\left(\tilde{y}_{\bar{r}k}^2 + \sigma_{\bar{r}k}^2\right)\gamma_{\bar{d}k}}{\bar{c}} \right]^{-3/2} {}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})} \frac{1}{{}^{(\alpha\tau v^*)}V_{(\bar{d}\bar{r})}} \frac{\left(\tilde{y}_{\bar{r}k}^2 + \sigma_{\bar{r}k}^2\right)\gamma_{\bar{d}k}}{\bar{c}} \\
&= \frac{1}{{}^{(\alpha\tau v^*)}V_{(\bar{d}\bar{r})}} \sum\nolimits_{k\in U} \tilde{y}_{\bar{r}k}^2 \gamma_{\bar{d}k} \left[{}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})} \frac{\left(\tilde{y}_{\bar{r}k}^2 + \sigma_{\bar{r}k}^2\right)\gamma_{\bar{d}k}}{\bar{c}} \right]^{-1/2}.
\end{aligned}
$$

L'inégalité qui suit est vérifiée

$$
j'_{(\bar{d}\bar{r})} < \frac{\sum_{k\in U} \tilde{y}_{\bar{r}k}\gamma_{\bar{d}k}}{\sqrt{\bar{c}\,{}^{(\alpha\tau v^*)}\phi_{(\bar{d}\bar{r})}}\,{}^{(\alpha\tau v^*)}V_{(\bar{d}\bar{r})}} << 1.
$$

Donc, la norme $\left\|J_{\mathbf{j}}\left({}^{(\alpha\tau)}\mathbf{VAA}_3\right)\right\| < 1$, et par conséquent la BE converge.

Annexe A3

Preuve que l'approximation de la remarque 4.1 est à la hausse

Puisque $\hat{u}_{(dr)k}$ est la prédiction par les moindres carrés pondérés de $u_{rk}\gamma_{dk}$, en utilisant une valeur différente de $\hat{u}_{(dr)k}$, telle que $\hat{u}_{(dr)k} = 0$, nous obtenons

$$\sum_{k\in U}(1/\pi_k - 1)\,E_M\left[(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k})^2\right] \leq \sum_{k\in U}(1/\pi_k - 1)\,E_M\left[(u_{rk}\gamma_{dk} - 0)^2\right],$$

où $E_M\left[(u_{rk}\gamma_{dk} - 0)^2\right] = \sigma^2_{rk}\gamma_{dk}$. En remplaçant les termes $E_M\left[(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k})^2\right]$ par $\sigma^2_{rk}\gamma_{dk}$ dans l'expression (A1.5), la VAA (4.3) est surestimée. L'approximation $\hat{u}_{(dr)k} = 0$ implique que $b_{(dr)k}(\boldsymbol{\pi}) = c_{(dr)k}(\boldsymbol{\pi}) = 0$. Enfin, nous soulignons que, dans la plupart des cas, la hausse est légère, puisque les $\hat{u}_{(dr)k}$ sont obtenus au moyen des variables $\mathbf{z}_k$ qui ont généralement un pouvoir prédictif très faible pour les valeurs de $u_{rk}\gamma_{dk}$ (voir la section 4). Dans ces situations, $\hat{u}_{(dr)k} \cong (1/N)\sum_{k\in U} u_{rk}\gamma_{dk} \cong 0$. Donc $E_M(u_{rk}\gamma_{dk}\hat{u}_{(dr)k}) \cong 0$ et $E_M(\hat{u}_{(dr)k})^2 \cong 0$.

Annexe A4

Preuve de l'expression (4.7)

Dans ce cas, chaque vecteur $\boldsymbol{\delta}_k$ contient $H - 1$ éléments nuls et 1 élément égal à 1 (correspondant à la population planifiée à laquelle l'unité k appartient). Étant donné les valeurs d'entrée, la procédure d'optimisation $\pi_k = \pi_h$ pour $k \in U_h$. Sous l'hypothèse susmentionnée, $[\mathbf{A}(\boldsymbol{\pi})]^{-1}$ est une matrice diagonale dont le hh^{e} élément est donné par $[\mathbf{A}_{hh}(\boldsymbol{\pi})]^{-1} = \left[N_h\pi_h^2(1/\pi_h - 1)\right]^{-1}$. En considérant que $\tilde{y}_{rk} = \bar{Y}_{rh}$, les expressions (A1.2) et (A1.3) peuvent être reformulées, respectivement, sous la forme

$$\hat{\bar{y}}_{(dr)k} = \pi_h\boldsymbol{\delta}'_k[\mathbf{A}(\boldsymbol{\pi})]^{-1}N_h\pi_h(1/\pi_h - 1)\bar{Y}_{rh} = \bar{Y}_{rh}. \tag{A4.1}$$

$$\hat{u}_{(dr)k} = \pi_h\boldsymbol{\delta}'_k[\mathbf{A}(\boldsymbol{\pi})]^{-1}\pi_h(1/\pi_h - 1)\sum_{j\in U}u_{rj} = (\pi_h N_h)^{-1}\sum_{j\in U_h}u_{rj}, \tag{A4.2}$$

mais $\sum_{j\in U_h} u_{rj} = 0$ en tant que somme des résidus d'un modèle de régression.

En utilisant les formules (A4.1) et (A4.2), l'expression (4.5) est donnée par

$$\begin{aligned}\text{VAA}(\hat{t}_{(dr)}) &= [N/(N-H)]\sum_h\left(\frac{1}{\pi_h} - 1\right)\sum_{k\in U_h}E_M(u_{rk}\gamma_{dk})^2 \\ &= [N/(N-H)]\sum_{d=1}^{D}\sum_{h\in H_d}\sigma^2_{rh}N_h(N_h/n_h - 1),\end{aligned}$$

puisque que $\pi_h = n_h/N_h$, et l'expression (4.7) peut être obtenue.

Bibliographie

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête*, 15, 1, 49-60.

Boyd, S., et Vanderberg, L. (2004). *Convex Optimization*. Cambridge University Press.

Breidt, F.J., et Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 479-487.

Chauvet, G., Bonnéry, D. et Deville, J.-C. (2011). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141, 984-994.

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*, 18, 1, 25-32.

Chromy, J. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 194-199.

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

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128, 569-591.

Dykstra R. et Wollan P. (1987). Finding I-projections subject to a finite set of linear inequality constraints, *Applied Statistics*, 36, 377-383.

Ernst, L.R. (1989). Further applications of linear programming to sampling problems. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 625-631.

Falorsi, P.D., et Righi, P. (2008). Une approche d'échantillonnage équilibré pour des plans de sondage à stratification multidimensionnelle pour l'estimation pour petits domaines. *Techniques d'enquête*, 34, 2, 247-259.

Falorsi, P.D., Orsini, D. et Righi, P. (2006). Balanced and coordinated sampling designs for small domain estimation. *Statistics in Transition*, 7, 1173-1198.

Gonzalez, J.M., et Eltinge, J.L. (2010). Optimal survey design: A review. *Section on Survey Research Methods – JSM 2010*, Octobre.

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

Khan, M.G.M., Mati, T. et Ahsan, M.J. (2010). An optimal multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach. *Journal of Official Statistics*, 26, 695-708.

Kokan, A., et Khan, S. (1967). Optimum allocation in multivariate surveys: An analytical solution. *Journal of the Royal Statistical Society, Series B*, 29, 115-125.

Lu, W., et Sitter, R.R. (2002). Méthode pratique de stratification multiple par programmation linéaire. *Techniques d'enquête*, 28, 2, 215-224.

Nedyalkova, D., et Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95, 521-537.

Tillé, Y. (2006). *Sampling Algorithms*. Springer-Verlag, New York.

Tillé, Y., et Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74, 31-37.

Winkler, W.E. (2001). Multi-way survey stratification and sampling. *Research Report Series*, Statistics #2001-01. Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 249-276
Statistique Canada, n° 12-001-X au catalogue

Une méthode d'estimation efficace pour l'échantillonnage matriciel

Takis Merkouris[1]

Résumé

L'échantillonnage matriciel, aussi appelé échantillonnage avec questionnaire fractionné ou scindé, est un plan d'échantillonnage qui consiste à diviser un questionnaire en sous-ensembles de questions, éventuellement chevauchants, puis à administrer chaque sous-ensemble à un ou à plusieurs sous-échantillons aléatoires d'un échantillon initial. Ce type de plan, de plus en plus attrayant, répond aux préoccupations concernant les coûts de la collecte, le fardeau de réponse et la qualité des données, mais réduit le nombre d'unités échantillonnées auxquelles les questions sont posées. Un concept élargi du plan d'échantillonnage matriciel comprend l'intégration d'échantillons provenant d'enquêtes distinctes afin de rationaliser les opérations d'enquête et d'accroître la cohérence des données de sortie. Dans le cas de l'échantillonnage matriciel avec sous-ensembles chevauchants de questions, nous proposons une méthode d'estimation efficace qui exploite les corrélations entre les items étudiés dans les divers sous-échantillons afin d'améliorer la précision des estimations de l'enquête. La méthode proposée, fondée sur le principe de la meilleure estimation linéaire sans biais, produit des estimateurs par régression optimale composites des totaux de population en utilisant un scénario approprié de calage des poids d'échantillonnage de l'échantillon complet. Une variante de ce scénario de calage, d'usage plus général, produit des estimateurs par régression généralisée composites qui sont également très efficaces sur le plan des calculs.

Mots-clés : Meilleur estimateur linéaire sans biais; calage; estimateur composite; estimateur par régression généralisée; échantillonnage matriciel non emboîté; questionnaire fractionné.

1 Introduction

L'échantillonnage matriciel est un plan d'échantillonnage selon lequel un long questionnaire est divisé en sous-ensembles de questions (items), éventuellement chevauchants, puis à administrer chaque sous-ensemble de questions à un ou à plusieurs sous-échantillons aléatoires distincts d'un échantillon initial. Sous ses diverses formes, ce plan peut servir différents objectifs, dont réduire la longueur et le coût du processus d'enquête et répondre aux préoccupations que soulève un long questionnaire en ce qui concerne le fardeau de réponse et la qualité des données. L'échantillonnage matriciel a été appliqué ou étudié dans divers domaines, principalement ceux de l'évaluation pédagogique et des études de santé publique. Un examen des travaux de recherche antérieurs sur l'échantillonnage matriciel, accompagné d'une discussion des problèmes que pose sa mise en œuvre dans les enquêtes, est présenté dans Gonzalez et Eltinge (2007). Pour des travaux récents sur les plans de sondage et l'estimation sous échantillonnage matriciel, motivés par les avantages potentiels de ce genre de plans d'échantillonnage dans les enquêtes à grande échelle, consulter Raghunathan et Grizzle (1995), Thomas, Raghunathan, Schenker, Katzoff et Johnson (2006), Gonzalez et Eltinge (2008), Chipperfield et Steel (2009, 2011), ainsi que les bibliographies connexes. Parmi les nombreux plans d'échantillonnage matriciel étudiés dans la littérature, nous distinguons quatre plans principaux qui diffèrent quant au nombre de sous-échantillons et au nombre de sous-questionnaires (chevauchants ou non) administrés à chaque sous-échantillon.

1. Takis Merkouris, Département de statistiques, Université d'économie et de commerce d'Athènes, Patision 76, Athènes 10434, Grèce. Courriel : merkouris@aueb.gr.

a) Différents ensembles (non chevauchants) de questions sont administrés à différents sous-échantillons.

b) Un ensemble de questions de base additionnel est administré à tous les sous-échantillons traités selon le plan (a). Il existe plusieurs raisons d'inclure un ensemble d'items de base dans tous les sous-échantillons : une grande précision peut être nécessaire pour certains items d'intérêt particulier; certains autres items (par exemple les caractéristiques démographiques) définissent les sous-populations et peuvent être utilisés dans des tableaux croisés des résultats de l'enquête; la corrélation des items de base avec le reste des items peut être utilisée pour améliorer la précision des estimations pour tous les items.

c) Une variante du plan (a) comportant un sous-échantillon additionnel auquel est administré le questionnaire complet. Elle peut être considérée comme une généralisation du plan d'échantillonnage à deux phases. La raison qui motive ce plan est de permettre l'analyse de l'interaction entre les ensembles de questions, en obtenant les réponses à toutes les questions auprès des unités de l'échantillon additionnel, et de permettre une estimation plus efficace.

d) Une extension du plan (c), dans laquelle l'ensemble de questions de base est administré à tous les sous-échantillons. Ce plan englobe toutes les caractéristiques des trois plans précédents.

L'une des tendances actuelles en ce qui concerne la planification des enquêtes consiste à appliquer une variante de l'échantillonnage matriciel dans laquelle un certain nombre d'enquêtes distinctes avec chevauchement du contenu sont intégrées en vue de rationaliser les opérations d'enquête, d'harmoniser le contenu des enquêtes, d'accroître la cohérence des données et d'améliorer l'estimation. Dans ce cadre d'échantillonnage matriciel non classique, les enquêtes distinctes peuvent être réalisées auprès de sous-échantillons d'un grand échantillon principal ou auprès d'échantillons indépendants tirés de la même population. Des plans d'échantillonnage de ce type sont étudiés activement ou mis en œuvre par divers organismes statistiques; voir, par exemple, l'intégration des enquêtes auprès des ménages de l'Office of National Statistics du Royaume-Uni (Smith 2009) et de l'Australian Bureau of Statistics (2011). Bien qu'une telle intégration puisse être considérée comme le processus inverse du fractionnement d'un questionnaire, la structure du plan de sondage en ce qui concerne la collecte des différents sous-ensembles d'éléments de données auprès de différents échantillons est essentiellement la même que dans le cadre classique. Dans le cas particulier où les échantillons provenant des diverses enquêtes sont indépendants, éventuellement issus de plans d'échantillonnage différents, les plans (b), (c) et (d) pourraient être caractérisés comme des plans d'échantillonnage matriciel non emboîté. Il convient de souligner que les avantages de l'échantillonnage matriciel ne dépendent pas toujours de l'utilisation de sous-échantillons (nécessairement dépendants) d'un échantillon initial. Dans certaines situations, il pourrait être plus pratique d'utiliser des échantillons indépendants, même s'il se peut que le chevauchement des échantillons soit négligeable.

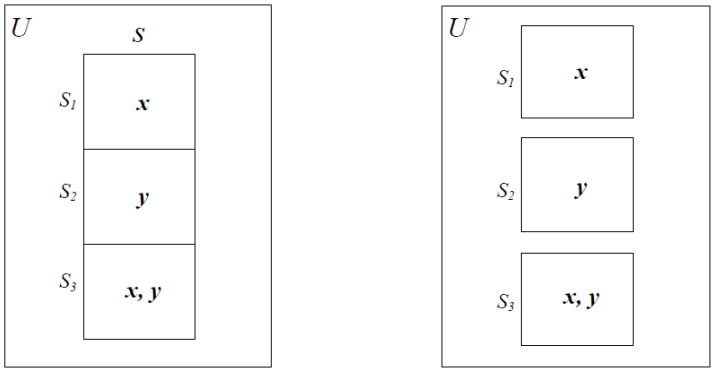
Dans le présent article, nous abordons le problème de l'estimation sous échantillonnage matriciel, c'est-à-dire la perte de précision des estimations de l'enquête, attribuable au fait que les éléments de données ne sont pas tous recueillis auprès de toutes les unités de l'échantillon. Dans le cas de l'échantillonnage matriciel non classique du paragraphe précédent, le problème d'estimation consiste à améliorer la précision des estimations pour chaque enquête composante. Pour les plans d'échantillonnage

matriciel (b), (c) et (d), qui comprennent un chevauchement des sous-ensembles de questions, une tâche d'estimation double consiste à combiner les données sur les items communs provenant des différents sous-échantillons pour améliorer l'estimation, et à exploiter les corrélations entre les items étudiés dans les divers sous-échantillons pour rendre l'estimation plus efficace pour tous les items. À cette fin, Raghunathan et Grizzle (1995) ainsi que Thomas et coll. (2006) ont exploré l'estimation avec imputation des valeurs manquantes causées par les items omis dans chaque sous-questionnaire. Gonzalez et Eltinge (2008) ont considéré l'estimation en utilisant un simple ajustement des poids qui combine les données sur les items communs. Dans le cas particulier du plan non emboîté (b), le problème d'estimation associé à la combinaison de données provenant d'échantillons indépendants a également été traité dans la littérature; voir, par exemple, Renssen et Nieuwenbroek (1997), Houbiers (2004), Merkouris (2004, 2010), Wu (2004), ainsi que Kim et Rao (2012). Le plan non emboîté (d) a été étudié dans Renssen (1998). Nous proposons une méthode d'estimation efficace, s'appuyant sur le principe de la meilleure estimation linéaire sans biais, qui produit des estimateurs par régression optimale composites des totaux au moyen d'une procédure de calage appropriée des poids d'échantillonnage de l'échantillon combiné, quand les probabilités d'inclusion de deuxième ordre dans l'échantillon sont connues. Une variante de cette procédure de calage, d'application plus générale, produit des estimateurs par régression généralisée composites qui, pour certaines conditions d'échantillonnage, sont des estimateurs par régression optimale. La méthode exploite les corrélations des items entre les sous-échantillons pour améliorer l'efficacité des estimateurs, même pour les items étudiés dans tous les sous-échantillons. Elle est également très commode sur le plan opérationnel, car elle produit des estimations pour tous les items au niveau de la population ou du domaine moyennant une simple adaptation du système de calage classique utilisé couramment par les organismes statistiques. Nous présentons ici la méthode en étudiant en détail les plans principaux (c) et (d). Les adaptations à des plans plus généraux sont relativement simples.

À la section 2 et à la section 3, nous décrivons la méthode proposée pour le plan (c). À la section 4, nous décrivons l'application de la méthode au plan (d). À la section 5, nous traitons l'estimation par domaine. À la section 6, nous présentons une étude par simulation. Enfin, à la section 7, nous concluons par une discussion.

2 Estimation composite par régression optimale pour le plan (c)

Une méthode d'estimation générale pour l'échantillonnage matriciel est illustrée pour le plan (c) dans les conditions les plus simples comportant trois échantillons S_1, S_2 et S_3 avec plans arbitraires et tailles $n_1, n_2, n_3,$ qui peuvent être des sous-échantillons d'un échantillon initial de taille $n = n_1 + n_2 + n_3$ pour une population étiquetée $U = 1,\ldots,k,\ldots,N,$ ou qui peuvent être tirés indépendamment de U. Un vecteur de dimension p de variables $\mathbf{x}$ et un vecteur de dimension q de variables $\mathbf{y}$ sont étudiés dans S_1 et $S_2,$ respectivement, et les deux vecteurs sont étudiés dans S_3. Ces deux modes d'échantillonnage matriciel, illustrés à la figure 2.1, seront appelés ci-après échantillonnage matriciel emboîté et non emboîté, respectivement, par analogie avec l'échantillonnage à deux phases emboîté et non emboîté (Hidiroglou 2001).



Plan emboîté Plan non emboîté

Figure 2.1 Plan (c) d'échantillonnage matriciel emboîté et non emboîté

Nous désignons par $\mathbf{w}_i$ le vecteur de poids de sondage pour l'échantillon $S_i, i = 1,2,3,$ et par $\mathbf{X}_i$ et $\mathbf{Y}_i$, les matrices d'échantillon de $\mathbf{x}$ et $\mathbf{y}$, l'indice inférieur indiquant l'échantillon. Nous obtenons les simples estimateurs de Horvitz-Thompson (HT) $\hat{\mathbf{X}}_1 (= \mathbf{X}_1'\mathbf{w}_1)$ et $\hat{\mathbf{X}}_3$ du total de population $\mathbf{t}_\mathbf{x}$ de $\mathbf{x}$, en utilisant S_1 et S_3, respectivement, et les simples estimateurs HT $\hat{\mathbf{Y}}_2$ et $\hat{\mathbf{Y}}_3$ du total $\mathbf{t}_\mathbf{y}$ de $\mathbf{y}$, en utilisant S_2 et S_3. Pour obtenir une estimation plus efficace des totaux $\mathbf{t}_\mathbf{x}$ et $\mathbf{t}_\mathbf{y}$, nous recherchons des estimateurs composites qui combinent toute l'information sur $\mathbf{x}$ et $\mathbf{y}$ disponible dans les trois échantillons. Ces estimateurs composites, qui sont les meilleurs estimateurs linéaires sans biais (BLUE), c'est-à-dire les combinaisons linéaires sans biais à variance minimale des quatre estimateurs $\hat{\mathbf{X}}_1, \hat{\mathbf{Y}}_2, \hat{\mathbf{X}}_3$ et $\hat{\mathbf{Y}}_3$, sont notés $\hat{\mathbf{X}}^B$ et $\hat{\mathbf{Y}}^B$, et donnés sous forme matricielle par

$$\begin{pmatrix} \hat{\mathbf{X}}^B \\ \hat{\mathbf{Y}}^B \end{pmatrix} = \boldsymbol{\mathcal{P}} \begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \\ \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix}, \tag{2.1}$$

où $\boldsymbol{\mathcal{P}} = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1}$, la matrice $\mathbf{W}$ satisfait $E\left[(\hat{\mathbf{X}}_1', \hat{\mathbf{Y}}_2', \hat{\mathbf{X}}_3', \hat{\mathbf{Y}}_3')'\right] = \mathbf{W}(\mathbf{t}_\mathbf{x}', \mathbf{t}_\mathbf{y}')'$ et contient des entrées 1 et 0, et $\mathbf{V}$ est la matrice de variance-covariance de $(\hat{\mathbf{X}}_1', \hat{\mathbf{Y}}_2', \hat{\mathbf{X}}_3', \hat{\mathbf{Y}}_3')'$. Cette méthode d'estimation a été proposée par Chipperfield et Steel (2009), qui ont fourni des expressions analytiques de l'estimateur BLUE pour les scalaires x et y sous échantillonnage matriciel non emboîté, en supposant que l'échantillonnage est aléatoire simple et que $\mathbf{V}$ est connue. Ce type d'approche de l'estimation composite a également été étudié dans un différent contexte d'enquête; voir Wolter (1979), Jones (1980) et Fuller (1990). En général, le calcul de l'estimateur BLUE donné par (2.1) n'est vraiment pas pratique, car le calcul d'une matrice estimée $\mathbf{V}$ (et de son inverse) dans $\boldsymbol{\mathcal{P}}$ est assez laborieux, surtout si le nombre de variables ou les tailles des échantillons sont grands; ce calcul serait prohibitif si les estimations pour des sous-populations étaient également requises. Naturellement, le problème devient plus difficile quand un plus grand nombre d'échantillons sont utilisés.

Voici une formulation plus pratique de cette procédure d'estimation. Premièrement, nous exprimons les estimateurs composites donnés par (2.1) explicitement comme des combinaisons linéaires des estimateurs HT $\hat{\mathbf{X}}_1, \hat{\mathbf{Y}}_2, \hat{\mathbf{X}}_3$ et $\hat{\mathbf{Y}}_3$, c'est-à-dire

$$\begin{aligned}\hat{\mathbf{X}}^B &= \mathbf{B}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + \mathbf{B}_{2\mathbf{x}}\hat{\mathbf{Y}}_2 + \mathbf{B}_{3\mathbf{x}}\hat{\mathbf{X}}_3 + \mathbf{B}_{4\mathbf{x}}\hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Y}}^B &= \mathbf{B}_{1\mathbf{y}}\hat{\mathbf{X}}_1 + \mathbf{B}_{2\mathbf{y}}\hat{\mathbf{Y}}_2 + \mathbf{B}_{3\mathbf{y}}\hat{\mathbf{X}}_3 + \mathbf{B}_{4\mathbf{y}}\hat{\mathbf{Y}}_3.\end{aligned}$$

La condition d'absence de biais, $E(\hat{\mathbf{X}}^B) = \mathbf{t}_\mathbf{x}$ et $E(\hat{\mathbf{Y}}^B) = \mathbf{t}_\mathbf{y}$, implique que $\mathbf{B}_{3\mathbf{x}} = \mathbf{I} - \mathbf{B}_{1\mathbf{x}}$, $\mathbf{B}_{4\mathbf{x}} = -\mathbf{B}_{2\mathbf{x}}$ et $\mathbf{B}_{4\mathbf{y}} = \mathbf{I} - \mathbf{B}_{2\mathbf{y}}$, $\mathbf{B}_{3\mathbf{y}} = -\mathbf{B}_{1\mathbf{y}}$. Donc, $\boldsymbol{\mathcal{P}}$ et $\mathbf{W}$ peuvent être exprimées sous la forme

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} \mathbf{B}_{1\mathbf{x}} & \mathbf{B}_{2\mathbf{x}} & \mathbf{I} - \mathbf{B}_{1\mathbf{x}} & -\mathbf{B}_{2\mathbf{x}} \\ \mathbf{B}_{1\mathbf{y}} & \mathbf{B}_{2\mathbf{y}} & -\mathbf{B}_{1\mathbf{y}} & \mathbf{I} - \mathbf{B}_{2\mathbf{y}} \end{pmatrix}, \quad \mathbf{W}' = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{I} \end{pmatrix},$$

respectivement, et les deux estimateurs composites possèdent nécessairement la forme de régression

$$\begin{aligned}\hat{\mathbf{X}}^B &= \hat{\mathbf{X}}_3 + \mathbf{B}_{1\mathbf{x}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{2\mathbf{x}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3) \\ \hat{\mathbf{Y}}^B &= \hat{\mathbf{Y}}_3 + \mathbf{B}_{1\mathbf{y}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{2\mathbf{y}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3).\end{aligned} \tag{2.2}$$

Alors, en écrivant que $\boldsymbol{\mathcal{P}} = (\boldsymbol{\mathcal{B}}, \mathbf{I} - \boldsymbol{\mathcal{B}})$, en notation évidente pour la matrice $\boldsymbol{\mathcal{B}}$, nous pouvons exprimer (2.1) comme

$$\begin{pmatrix} \hat{\mathbf{X}}^B \\ \hat{\mathbf{Y}}^B \end{pmatrix} = \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \end{pmatrix} + (\mathbf{I} - \boldsymbol{\mathcal{B}}) \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} + \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix}, \tag{2.3}$$

le deuxième membre de (2.3) étant la forme matricielle de (2.2). Le problème consistant à trouver la valeur optimale (minimisant la variance) de $\boldsymbol{\mathcal{P}}$ de l'estimateur BLUE en (2.1) se réduit alors au problème consistant à trouver la matrice optimale $\boldsymbol{\mathcal{B}}$ en (2.3). La matrice optimale estimée $\hat{\boldsymbol{\mathcal{B}}}^o$ est donnée par

$$\hat{\boldsymbol{\mathcal{B}}}^o = -\widehat{\mathrm{Cov}}\left(\begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix}, \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix} \right) \left[\hat{V} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix} \right]^{-1}, \tag{2.4}$$

et, quand les trois échantillons sont indépendants, elle se réduit à

$$\hat{\boldsymbol{\mathcal{B}}}^o = \hat{V} \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} \left[\hat{V} \begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \end{pmatrix} + \hat{V} \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} \right]^{-1}. \tag{2.5}$$

Compte tenu de (2.3), avec un tel $\hat{\boldsymbol{\mathcal{B}}}^o$ optimal, le BLUE estimé en (2.1) faisant intervenir la matrice estimée $\hat{\mathbf{V}}$, et avec $\hat{\boldsymbol{\mathcal{P}}} = (\hat{\boldsymbol{\mathcal{B}}}^o, \mathbf{I} - \hat{\boldsymbol{\mathcal{B}}}^o)$, est un type particulier d'estimateur par régression multivariée optimale. Pour la forme de l'estimateur par régression optimale ordinaire (un seul échantillon) et une discussion pertinente, voir Montanari (1987) et Rao (1994).

En exprimant la variance estimée de l'estimateur HT d'un total (voir, par exemple, Särndal, Swensson et Wretman 1992, page 43) sous une forme quadratique avec matrice définie non négative associée $\boldsymbol{\Lambda}^0 = \{(\pi_{kl} - \pi_k\pi_l)/\pi_k\pi_l\pi_{kl}\}$, où π_k, π_{kl} sont les probabilités d'inclusion d'ordre un et d'ordre deux, on peut montrer, après certaines opérations algébriques sur les matrices, que

$$\hat{\mathcal{B}}^{o} = (\mathcal{X}_3'\boldsymbol{\Lambda}^0\mathcal{X})(\mathcal{X}'\boldsymbol{\Lambda}^0\mathcal{X})^{-1}, \tag{2.6}$$

où

$$\mathcal{X} = \begin{pmatrix} -\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & -\mathbf{Y}_2 \\ \mathbf{X}_3 & \mathbf{Y}_3 \end{pmatrix} \tag{2.7}$$

est la matrice de plan de dimensions $n \times (p+q)$ correspondant à l'estimateur par régression (2.3), $\mathcal{X}_3$ est la matrice $\mathcal{X}$ dans laquelle les éléments des deux premières lignes sont fixés à zéro, et $\boldsymbol{\Lambda}^0$ est associée à l'échantillon combiné $S = S_1 \cup S_2 \cup S_3$, qui se réduit dans l'échantillonnage non emboîté à la matrice diagonale par blocs $\text{diag}\{\boldsymbol{\Lambda}_i^0\}$ avec $\boldsymbol{\Lambda}_i^0$ associée à l'échantillon S_i. Pour le plan d'échantillonnage emboîté, les probabilités définissant $\boldsymbol{\Lambda}^0$ sont les produits des probabilités d'inclusion dans S et des probabilités de sous-échantillonnage conditionnelles (sur S). Avec cette matrice optimale estimée $\hat{\mathcal{B}}^{o}$, le BLUE estimé en (2.3), appelé estimateur par régression optimale composite (ROC) et désigné par $\hat{\mathcal{X}}^{\text{ROC}}$, s'écrit de manière compacte sous la forme $\hat{\mathcal{X}}^{\text{ROC}} = \hat{\mathcal{X}}_3 - \hat{\mathcal{B}}^{o}\hat{\mathcal{X}}[= (\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}^{o'})'\mathbf{w}]$, où $\mathbf{w} = (\mathbf{w}_1', \mathbf{w}_2', \mathbf{w}_3')'$ est le vecteur des poids de sondage de l'échantillon combiné S. Il s'avère que l'estimateur ROC est, en fait, égal à la somme des résidus de la régression pour l'échantillon pondérée, et que $\hat{\mathcal{B}}^{o}$ minimise la forme quadratique $(\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}^{o'})'\boldsymbol{\Lambda}^0(\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}^{o'})$ en ces résidus, ce qui est la variance approximative (en grand échantillon) estimée de $\hat{\mathcal{X}}^{\text{ROC}}$.

Or, en écrivant $\hat{\mathcal{X}}^{\text{ROC}}$ sous la forme $\hat{\mathcal{X}}^{\text{ROC}} = \mathcal{X}_3'[\mathbf{w} + \boldsymbol{\Lambda}^0\mathcal{X}(\mathcal{X}'\boldsymbol{\Lambda}^0\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{w})]$, il apparaît que l'estimateur ROC possède la forme d'un estimateur par calage (avec le vecteur de totaux de calage $\mathbf{0} = (\mathbf{0}', \mathbf{0}')'$ de dimension $(p+q)$), dont les composantes satisfont les contraintes $\hat{\mathbf{X}}_1^{\text{ROC}} = \hat{\mathbf{X}}_3^{\text{ROC}}$ et $\hat{\mathbf{Y}}_2^{\text{ROC}} = \hat{\mathbf{Y}}_3^{\text{ROC}}$, c'est-à-dire que les estimations calées du même total provenant de deux échantillons différents sont égales. En effet, le vecteur

$$\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}^0\mathcal{X}(\mathcal{X}'\boldsymbol{\Lambda}^0\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{w}), \tag{2.8}$$

est le vecteur des poids calés qui minimise la distance au sens des moindres carrés généralisés $(\mathbf{c} - \mathbf{w})'(\boldsymbol{\Lambda}^0)^{-1}(\mathbf{c} - \mathbf{w})$ tout en satisfaisant les contraintes $\mathbf{X}_1'\mathbf{c}_1 = \mathbf{X}_3'\mathbf{c}_3$ et $\mathbf{Y}_2'\mathbf{c}_2 = \mathbf{Y}_3'\mathbf{c}_3$, où le sous-vecteur $\mathbf{c}_i$ correspond à l'échantillon S_i. Cela découle d'un résultat général pour le cas avec un seul échantillon, selon lequel le calage au moyen de la mesure de distance par les moindres carrés généralisés peut faire intervenir une matrice définie positive de dimensions $n \times n$ arbitraire $\mathbf{R}$ au lieu de $\boldsymbol{\Lambda}^0$; voir Andersson et Thorburn (2005).

Nous pouvons maintenant écrire l'estimateur ROC formellement sous la forme d'un estimateur par calage, $\hat{\mathcal{X}}^{\text{ROC}} = \mathcal{X}_3'\mathbf{c}$, et, en utilisant le sous-vecteur de poids calés $\mathbf{c}_3$, pour l'échantillon S_3 seulement, nous obtenons les composantes de $\hat{\mathcal{X}}^{\text{ROC}}$ directement sous les formes linéaires simples

$$\hat{\mathbf{X}}^{\text{ROC}} = \mathbf{X}_3'\mathbf{c}_3 = \sum\nolimits_{S_3} c_k\mathbf{x}_k; \quad \hat{\mathbf{Y}}^{\text{ROC}} = \mathbf{Y}_3'\mathbf{c}_3 = \sum\nolimits_{S_3} c_k\mathbf{y}_k,$$

comme dans la pratique courante des enquêtes. Toutefois, une décomposition du vecteur $\mathbf{c}$ basée sur le lemme général ci-après concernant le calage donne une expression analytique de $\hat{\mathbf{X}}^{\text{ROC}}$ et $\hat{\mathbf{Y}}^{\text{ROC}}$ de la

forme (2.2), qui renseigne sur la structure et l'efficacité de l'estimateur ROC. La preuve du lemme est donnée en annexe.

Lemme 1 *Soit* $\mathcal{X}$ *une matrice de plan de dimensions* $n \times (p+q)$ *et de plein rang écrite sous forme partitionnée* $(\boldsymbol{X}, \boldsymbol{\Psi})$, *avec le vecteur correspondant de totaux de calage* $\mathbf{t}_{\mathcal{X}} = (\mathbf{t}_{\boldsymbol{X}}', \mathbf{t}_{\boldsymbol{\Psi}}')'$, *et soit* $\mathbf{R}$ *toute matrice définie positive de dimensions* $n \times n$. *Alors, le vecteur de poids calés* $\mathbf{c} = \mathbf{w} + \mathbf{R}\mathcal{X}(\mathcal{X}'\mathbf{R}\mathcal{X})^{-1}(\mathbf{t}_{\mathcal{X}} - \mathcal{X}'\mathbf{w})$, *obtenu par la procédure de calage utilisant la mesure de distance* $(\mathbf{c}-\mathbf{w})'\mathbf{R}^{-1}(\mathbf{c}-\mathbf{w})$ *et la contrainte* $\mathcal{X}'\mathbf{c} = \mathbf{t}_{\mathcal{X}}$ *peut être décomposé comme il suit*

$$\mathbf{c} = \mathbf{w} + \mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}[\mathbf{t}_{\boldsymbol{X}} - \boldsymbol{X}'\mathbf{w}] + \mathbf{L}_{\boldsymbol{X}}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{L}_{\boldsymbol{X}}\boldsymbol{\Psi})^{-1}[\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}], \tag{2.9}$$

où $\mathbf{L}_{\boldsymbol{X}} = \mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{X}})$ *avec* $\mathbf{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\mathbf{R}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{R}$, *et* $\mathbf{L}_{\boldsymbol{\Psi}} = \mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Psi}})$ *avec* $\mathbf{P}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\mathbf{R}$. *Le vecteur* $\mathbf{c}$ *peut s'écrire*

$$\mathbf{c} = \mathbf{c}_{\boldsymbol{\Psi}} + \mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}[\mathbf{t}_{\boldsymbol{X}} - \boldsymbol{X}'\mathbf{c}_{\boldsymbol{\Psi}}], \tag{2.10}$$

où le vecteur

$$\mathbf{c}_{\boldsymbol{\Psi}} = \mathbf{w} + \mathbf{R}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}[\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}]$$

est généré par le calage des poids de sondage ne faisant intervenir que $\boldsymbol{\Psi}$ *et* $\mathbf{t}_{\boldsymbol{\Psi}}$. *Par symétrie,*

$$\mathbf{c} = \mathbf{c}_{\boldsymbol{X}} + \mathbf{L}_{\boldsymbol{X}}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{L}_{\boldsymbol{X}}\boldsymbol{\Psi})^{-1}[\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{c}_{\boldsymbol{X}}], \tag{2.11}$$

où

$$\mathbf{c}_{\boldsymbol{X}} = \mathbf{w} + \mathbf{R}\boldsymbol{X}(\boldsymbol{X}'\mathbf{R}\boldsymbol{X})^{-1}[\mathbf{t}_{\boldsymbol{X}} - \boldsymbol{X}'\mathbf{w}].$$

Or, si $\mathcal{X}$ est tel qu'en (2.7), avec le vecteur correspondant de totaux de calage $\mathbf{t}_{\mathcal{X}} = (\mathbf{0}', \mathbf{0}')'$, et si $\mathbf{R} = \boldsymbol{\Lambda}^0$, alors il découle de (2.9) que (2.8) peut s'écrire sous la forme

$$\mathbf{c} = \mathbf{w} + \mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X}(\mathbf{X}'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X})^{-1}[\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3] + \mathbf{L}_{\mathbf{X}}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{L}_{\mathbf{X}}\boldsymbol{\Psi})^{-1}[\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3],$$

et donc

$$\begin{aligned} \hat{\mathbf{X}}^{\mathrm{ROC}} &= \mathbf{X}_3'\mathbf{c}_3 = \hat{\mathbf{X}}_3 + \hat{\mathbf{B}}_{1\mathbf{x}}^{o}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \hat{\mathbf{B}}_{2\mathbf{x}}^{o}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3) \\ &= \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\hat{\mathbf{X}}_1 + (\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o})\hat{\mathbf{X}}_3 + \hat{\mathbf{B}}_{2\mathbf{x}}^{o}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3), \end{aligned} \tag{2.12}$$

en notation évidente pour $\hat{\mathbf{B}}_{1\mathbf{x}}^{o}$ et $\hat{\mathbf{B}}_{2\mathbf{x}}^{o}$. Une expression similaire s'obtient pour $\hat{\mathbf{Y}}^{\mathrm{ROC}}$. On voit en examinant (2.12) que l'estimateur ROC $\hat{\mathbf{X}}^{\mathrm{ROC}}$ de $\mathbf{t}_{\mathbf{x}}$ est approximativement (pour les grands échantillons) sans biais, et tire son efficacité de la combinaison des deux estimateurs élémentaires $\hat{\mathbf{X}}_1$ et $\hat{\mathbf{X}}_3$ (mise en commun de l'information provenant des échantillons S_1 et S_3) et de l'emprunt d'information provenant de l'échantillon S_2 grâce à la corrélation entre $\mathbf{x}$ et $\mathbf{y}$. Compte tenu de (2.10), l'estimateur $\hat{\mathbf{X}}^{\mathrm{ROC}}$ prend la forme de rechange

$$\begin{aligned}\hat{\mathbf{X}}^{\mathrm{ROC}} &= \mathbf{X}_3'\mathbf{c}_{3\boldsymbol{\Psi}} + \mathbf{X}_3'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X}(\mathbf{X}'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X})^{-1}\left[\mathbf{X}_1'\mathbf{c}_{1\boldsymbol{\Psi}} - \mathbf{X}_3'\mathbf{c}_{3\boldsymbol{\Psi}}\right] \\ &= \hat{\mathbf{X}}_3^{\mathrm{RO}} + \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\left[\hat{\mathbf{X}}_1^{\mathrm{RO}} - \hat{\mathbf{X}}_3^{\mathrm{RO}}\right] \\ &= \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\hat{\mathbf{X}}_1^{\mathrm{RO}} + \left(\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\right)\hat{\mathbf{X}}_3^{\mathrm{RO}},\end{aligned} \tag{2.13}$$

où $\hat{\mathbf{X}}_i^{\mathrm{RO}} = \hat{\mathbf{X}}_i + \mathbf{X}_i'\boldsymbol{\Lambda}^0\boldsymbol{\Psi}(\boldsymbol{\Psi}'\boldsymbol{\Lambda}^0\boldsymbol{\Psi})^{-1}\left(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right)$ représente les estimateurs par régression optimale (RO) incorporant l'effet de régression du dernier terme en (2.12).

Dans le cas de l'échantillonnage matriciel non emboîté, $\boldsymbol{\Lambda}^0 = \mathrm{diag}\{\boldsymbol{\Lambda}_i^0\}$, $\hat{\mathbf{X}}_1^{\mathrm{RO}} = \hat{\mathbf{X}}_1$, $\hat{\mathbf{X}}_3^{\mathrm{RO}} = \hat{\mathbf{X}}_3 + \widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)\left[\hat{V}(\hat{\mathbf{Y}}_2) + \hat{V}(\hat{\mathbf{Y}}_3)\right]^{-1}\left[\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right]$, dont la variance approximative estimée est $\widehat{\mathrm{AV}}(\hat{\mathbf{X}}_3^{\mathrm{RO}}) = \hat{V}(\hat{\mathbf{X}}_3) - \widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)\left[\hat{V}(\hat{\mathbf{Y}}_2) + \hat{V}(\hat{\mathbf{Y}}_3)\right]^{-1}\widehat{\mathrm{Cov}}'(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)$, et $\hat{\mathbf{B}}_{1\mathbf{x}}^{o} = \widehat{\mathrm{AV}}(\hat{\mathbf{X}}_3^{\mathrm{RO}})\left[\hat{V}(\hat{\mathbf{X}}_1) + \widehat{\mathrm{AV}}(\hat{\mathbf{X}}_3^{\mathrm{RO}})\right]^{-1}$ est le coefficient qui minimise la variance $\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{ROC}})$. La forme explicite $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o} = \hat{V}(\hat{\mathbf{X}}_1)[\hat{V}(\hat{\mathbf{X}}_1) + \hat{V}(\hat{\mathbf{X}}_3) - \widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3) \times [\hat{V}(\hat{\mathbf{Y}}_2) + \hat{V}(\hat{\mathbf{Y}}_3)]^{-1}\widehat{\mathrm{Cov}}'(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)]^{-1}$ indique clairement que le terme $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o}$ est d'autant plus grand que la corrélation entre $\mathbf{x}$ et $\mathbf{y}$ est forte, , et que plus de poids est donné à la composante moins variable $\hat{\mathbf{X}}_3^{\mathrm{RO}}$. Dans cette connexion, on peut montrer facilement que $\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{ROC}})$ satisfait

$$\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{ROC}})\left[\hat{V}(\hat{\mathbf{X}}_1)\right]^{-1} = \hat{\mathbf{B}}_{1\mathbf{x}}^{o} < \mathbf{I}, \quad \widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{ROC}})\left[\widehat{\mathrm{AV}}(\hat{\mathbf{X}}_3^{\mathrm{RO}})\right]^{-1} = \mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o} < \mathbf{I}.$$

Ces inégalités sont également vérifiées pour toute combinaison linéaire des composantes de chacun des estimateurs concernés. L'efficacité de l'estimateur par régression optimale composite $\hat{\mathbf{X}}^{\mathrm{ROC}}$ dépasse d'une valeur correspondant aux quantités montrées l'efficacité de chacune de ses deux composantes $\hat{\mathbf{X}}_1$ et $\hat{\mathbf{X}}_3^{\mathrm{RO}}$, l'efficacité dépendant de la force de la corrélation entre $\mathbf{x}$ et $\mathbf{y}$. L'estimateur $\hat{\mathbf{X}}^{\mathrm{ROC}}$ est également plus efficace que l'estimateur $\tilde{\mathbf{X}}^{\mathrm{ROC}} = \tilde{\mathbf{B}}_{1\mathbf{x}}^{o}\hat{\mathbf{X}}_1 + \left(\mathbf{I} - \tilde{\mathbf{B}}_{1\mathbf{x}}^{o}\right)\hat{\mathbf{X}}_3$, avec $\tilde{\mathbf{B}}_{1\mathbf{x}}^{o} = \hat{V}(\hat{\mathbf{X}}_3)\left[\hat{V}(\hat{\mathbf{X}}_1) + \hat{V}(\hat{\mathbf{X}}_3)\right]^{-1}$, qui n'incorpore pas l'information sur $\mathbf{y}$ (n'emprunte pas d'information à l'échantillon S_2) et dont la variance estimée est $\widehat{\mathrm{AV}}(\tilde{\mathbf{X}}^{\mathrm{ROC}}) = \hat{V}(\hat{\mathbf{X}}_1)\left[\hat{V}(\hat{\mathbf{X}}_1) + \hat{V}(\hat{\mathbf{X}}_3)\right]^{-1}\hat{V}(\hat{\mathbf{X}}_3)$. En effet, en écrivant la variance $\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{ROC}}) = \hat{V}(\hat{\mathbf{X}}_1)\hat{\mathbf{B}}_{1\mathbf{x}}^{o}$ sous la forme $\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{ROC}}) = \hat{V}(\hat{\mathbf{X}}_1)\left[\hat{V}(\hat{\mathbf{X}}_1) + \hat{V}(\hat{\mathbf{X}}_3)\right]^{-1}\hat{V}(\hat{\mathbf{X}}_3)\mathbf{E}$, où $\mathbf{E} = \mathbf{E}_1\mathbf{E}_2$ avec $\mathbf{E}_1 = [\mathbf{I} - (\hat{V}(\hat{\mathbf{X}}_3))^{-1}\widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)[\hat{V}(\hat{\mathbf{Y}}_2) + \hat{V}(\hat{\mathbf{Y}}_3)]^{-1}\widehat{\mathrm{Cov}}'(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)]$ et $\mathbf{E}_2 = [\mathbf{I} - [\hat{V}(\hat{\mathbf{X}}_1) + \hat{V}(\hat{\mathbf{X}}_3)]^{-1}\widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)[\hat{V}(\hat{\mathbf{Y}}_2) + \hat{V}(\hat{\mathbf{Y}}_3)]^{-1}\widehat{\mathrm{Cov}}'(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)]^{-1}$, et en notant que $\mathbf{E} \leq \mathbf{I}$, il s'ensuit que

$$\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{ROC}})\left[\widehat{\mathrm{AV}}(\tilde{\mathbf{X}}^{\mathrm{ROC}})\right]^{-1} = \mathbf{E} \leq \mathbf{I},$$

c'est-à-dire que l'emprunt d'information à S_2 réduit la variance de l'estimateur composite de $\mathbf{t}_{\mathbf{x}}$ d'un facteur $\mathbf{E}$, qui dépend de la force de la corrélation entre $\mathbf{x}$ et $\mathbf{y}$. Il est facile de vérifier que, pour deux variables scalaires x et y sous échantillonnage aléatoire simple, ce résultat se réduit au résultat analytique analogue sur l'efficacité de l'estimateur BLUE donné dans Chipperfield et Steel (2009, page 231). Dans ce cas simple, $E = [n_1 + n_3]\left[n_3 + n_2(1-\rho^2)\right]/\left[(n_1 + n_3)(n_2 + n_3) - n_1 n_2 \rho^2\right]$, où ρ est la corrélation entre x et y. En guise d'exemple, en supposant que les tailles d'échantillon sont égales et que la corrélation $\rho = 0,7$, le gain d'efficacité est de 13,96 %.

Dans le cas de l'échantillonnage matriciel emboîté, les deux estimateurs en (2.13) sont $\hat{\mathbf{X}}_i^{\mathrm{RO}} = \hat{\mathbf{X}}_i + \widehat{\mathrm{Cov}}\left(\hat{\mathbf{X}}_i, \hat{\boldsymbol{\Psi}}\right)[\hat{V}(\hat{\boldsymbol{\Psi}})]^{-1}\left[\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right]$, et $\hat{\mathbf{B}}_{1x}^{o} = [\widehat{\mathrm{AV}}(\hat{\mathbf{X}}_3^{\mathrm{RO}}) - \widehat{\mathrm{AC}}(\hat{\mathbf{X}}_1^{\mathrm{RO}}, \hat{\mathbf{X}}_3^{\mathrm{RO}})][\widehat{\mathrm{AV}}(\hat{\mathbf{X}}_1^{\mathrm{RO}}) + \widehat{\mathrm{AV}}(\hat{\mathbf{X}}_3^{\mathrm{RO}}) - 2\widehat{\mathrm{AC}}(\hat{\mathbf{X}}_1^{\mathrm{RO}}, \hat{\mathbf{X}}_3^{\mathrm{RO}})]^{-1}$, où AC désigne la covariance approximative. Dans ce cas, en plus de la corrélation $\rho_{x3,y3}$ entre $\hat{\mathbf{X}}_3$ et $\hat{\mathbf{Y}}_3$ dans l'échantillon S_3, l'efficacité de $\hat{\mathbf{X}}^{\mathrm{ROC}}$ dépend des corrélations $\rho_{x1,x3}, \rho_{y2,y3}, \rho_{y2,x3}$ des estimateurs dues à la dépendance des sous-échantillons. Si x et y sont univariées et en émettant l'hypothèse simplificatrice que les plans de sondage sont identiques pour les trois sous-échantillons (comme dans le fractionnement égal de l'échantillon complet), nous obtenons certains indices au moyen des expressions simples $\widehat{\mathrm{AV}}(\hat{X}^{\mathrm{ROC}}) = V(\hat{X}_3)[2(1-\rho_{x1,x3}^2)(1-\rho_{y2,y3}) - (\rho_{x3,y3} - \rho_{y2,x3})^2] / [4(1-\rho_{x1,x3})(1-\rho_{y2,y3}) - (\rho_{x3,y3} - \rho_{y2,x3})^2]$, et $\widehat{\mathrm{AV}}(\tilde{X}^{\mathrm{ROC}}) = V\left(\hat{X}_3\right)(1+\rho_{x1,x3})/2$. Manifestement, l'estimateur $\tilde{X}^{\mathrm{ROC}}$, qui ne tient pas compte de l'information sur y, n'est plus efficace que la moyenne simple des estimateurs sur un seul échantillon de t_x que si la corrélation $\rho_{x1,x3}$ est négative. L'efficacité de $\hat{X}^{\mathrm{ROC}}$ par rapport à $\tilde{X}^{\mathrm{ROC}}$

$$\frac{\widehat{\mathrm{AV}}(\hat{X}^{\mathrm{ROC}})}{\widehat{\mathrm{AV}}(\tilde{X}^{\mathrm{ROC}})} = \frac{4\left(1-\rho_{x1,x3}^2\right)\left(1-\rho_{y2,y3}\right) - 2\left(\rho_{x3,y3} - \rho_{y2,x3}\right)^2}{4\left(1-\rho_{x1,x3}^2\right)\left(1-\rho_{y2,y3}\right) - \left(1+\rho_{x1,x3}\right)\left(\rho_{x3,y3} - \rho_{y2,x3}\right)^2}$$

dépend du signe et de la grandeur de $\rho_{x1,x3}$ et de la grandeur de $\left|\rho_{x3,y3} - \rho_{y2,x3}\right|$.

Bien que la procédure de calage, avec le vecteur de poids calés (2.8), facilite considérablement le calcul de l'estimateur par régression optimale composite pour tout total d'intérêt, la matrice $\boldsymbol{\Lambda}^0$ rend les calculs extrêmement exigeants, particulièrement dans le cas de l'échantillonnage emboîté où les sous-échantillons dépendent les uns des autres et $\boldsymbol{\Lambda}^0$ n'est donc pas $\mathrm{diag}\{\boldsymbol{\Lambda}_i^0\}$. En outre, les probabilités π_{kl} ne sont pas connues pour la plupart des plans d'échantillonnage. Un estimateur par régression composite de rechange dont les calculs sont très rapides est élaboré à la section suivante.

3 Estimation composite par régression généralisée pour le plan (c)

Une variante très commode sur le plan des calculs, mais généralement sous-optimale, de $\hat{\boldsymbol{\mathcal{B}}}^{o}$ en (2.6) s'obtient en remplaçant la matrice $\boldsymbol{\Lambda}^0$ par la « matrice de pondération » diagonale $\boldsymbol{\Lambda}$ dont la ik^{e} entrée diagonale est w_{ik}/q_{ik}, où les $\{w_{ik}\}$ sont les poids de sondage de S_i et les $\{q_{ik}\}$ sont des constantes positives. Cela donne l'estimateur par régression généralisée composite (RGC) multivariée de $\left(\mathbf{t}_{\mathbf{x}}', \mathbf{t}_{\mathbf{y}}'\right)'$

$$\begin{pmatrix} \hat{\mathbf{X}}^{\mathrm{RGC}} \\ \hat{\mathbf{Y}}^{\mathrm{RGC}} \end{pmatrix} = \hat{\boldsymbol{\mathcal{B}}}\begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \end{pmatrix} + (\mathbf{I} - \hat{\boldsymbol{\mathcal{B}}})\begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} + \hat{\boldsymbol{\mathcal{B}}}\begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix}, \tag{3.1}$$

où $\hat{\boldsymbol{\mathcal{B}}} = (\boldsymbol{\mathcal{X}}_3'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}})(\boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}})^{-1}$ est le coefficient de régression de la matrice associée. Pour une discussion approfondie de l'estimateur par régression généralisée dans le cas d'un seul échantillon, voir Särndal et coll. (1992, chapitre 6). L'estimateur RGC peut s'écrire de manière compacte sous la forme $\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RGC}} = \hat{\boldsymbol{\mathcal{X}}}_3 - \hat{\boldsymbol{\mathcal{B}}}\hat{\boldsymbol{\mathcal{X}}}\left[= \left(\boldsymbol{\mathcal{X}}_3 - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}}'\right)' \mathbf{w}\right]$, c'est-à-dire la somme pondérée des résidus de régression de

l'échantillon. Le coefficient $\hat{\mathcal{B}}$ est optimal au sens des moindres carrés généralisés, c'est-à-dire qu'il minimise la forme quadratique $(\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}')' \boldsymbol{\Lambda}(\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}')$ dans ces résidus. Comme l'estimateur ROC, l'estimateur RGC peut aussi être obtenu dans la forme de calage comme $\mathcal{X}_3'\mathbf{c}$, où le vecteur $\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}\mathcal{X}(\mathcal{X}'\boldsymbol{\Lambda}\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{w})$ minimise la distance au sens des moindres carrés généralisés $(\mathbf{c} - \mathbf{w})'\boldsymbol{\Lambda}^{-1}(\mathbf{c} - \mathbf{w})$ et satisfait les contraintes $\hat{\mathbf{X}}_1^{\mathrm{RGC}} = \hat{\mathbf{X}}_3^{\mathrm{RGC}}$ et $\hat{\mathbf{Y}}_2^{\mathrm{RGC}} = \hat{\mathbf{Y}}_3^{\mathrm{RGC}}$. Cela étend au présent contexte l'équivalence bien connue de l'estimation par régression généralisée et de l'estimation par calage (Deville et Särndal 1992) dans le cas d'un échantillon unique. Or, en utilisant le sous-vecteur de poids calés $\mathbf{c}_3$, pour l'échantillon S_3 seulement, nous obtenons les estimateurs composites donnés en (3.1) sous les formes linéaires simples $\hat{\mathbf{X}}^{\mathrm{RGC}} = \mathbf{X}_3'\mathbf{c}_3$ et $\hat{\mathbf{Y}}^{\mathrm{RGC}} = \mathbf{Y}_3'\mathbf{c}_3$. En utilisant le lemme 1 et la structure diagonale de $\boldsymbol{\Lambda}$, il s'avère que $\hat{\mathbf{X}}^{\mathrm{RGC}}$ peut s'écrire

$$\hat{\mathbf{X}}^{\mathrm{RGC}} = \hat{\mathbf{B}}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + (\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}})\hat{\mathbf{X}}_3^{\mathrm{RG}}, \tag{3.2}$$

où $\hat{\mathbf{X}}_3^{\mathrm{RG}} = \hat{\mathbf{X}}_3 + \mathbf{X}_3'\boldsymbol{\Lambda}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\boldsymbol{\Lambda}\boldsymbol{\Psi})^{-1}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3)$ est l'analogue par régression généralisée (RG) de $\hat{\mathbf{X}}_3^{\mathrm{RO}}$. Le coefficient de régression de la matrice $\hat{\mathbf{B}}_{1\mathbf{x}}$ s'écrit explicitement sous la forme $\hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_3'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X}(\mathbf{X}_1'\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X})^{-1}$, où $\mathbf{X}_3'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X} = \mathbf{X}_3'\boldsymbol{\Lambda}_3\mathbf{X}_3 - \mathbf{X}_3'\boldsymbol{\Lambda}_3\mathbf{Y}_3(\mathbf{Y}_2'\boldsymbol{\Lambda}_2\mathbf{Y}_2 + \mathbf{Y}_3'\boldsymbol{\Lambda}_3\mathbf{Y}_3)^{-1}$ $\mathbf{Y}_3'\boldsymbol{\Lambda}_3\mathbf{X}_3$. Si $\mathbf{x}$ et $\mathbf{y}$ n'étaient pas corrélées, ou si l'information sur $\mathbf{y}$ n'était pas utilisée dans l'estimation de $\mathbf{t}_{\mathbf{x}}$, on aurait alors $\hat{\mathbf{X}}_3^{\mathrm{RG}} = \hat{\mathbf{X}}_3$ et $\hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_3'\boldsymbol{\Lambda}_3\mathbf{X}_3(\mathbf{X}_1'\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\boldsymbol{\Lambda}_3\mathbf{X}_3)^{-1}$. Mais l'estimateur RG $\hat{\mathbf{X}}_3^{\mathrm{RG}}$ est généralement plus efficace que l'estimateur HT $\hat{\mathbf{X}}_3$, et puisque $\mathbf{X}_1'\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X} < \mathbf{X}_1'\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\boldsymbol{\Lambda}_3\mathbf{X}_3$ (dans le classement par ordre partiel des matrices définies non négatives), il est clair que plus de poids est attribué à $\hat{\mathbf{X}}_3^{\mathrm{RG}}$ dans (3.2), par la voie de $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_1'\boldsymbol{\Lambda}_1\mathbf{X}_1(\mathbf{X}_1'\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X})^{-1}$, qu'il n'aurait été donné à l'estimateur composant $\hat{\mathbf{X}}_3$ dans l'estimateur composite simple ne faisant intervenir que l'information sur $\mathbf{x}$. Cela donne à penser que l'estimateur RGC donné en (3.2), dans lequel est intégrée l'information provenant de l'échantillon S_2, est un estimateur plus efficace. L'efficacité de $\hat{\mathbf{X}}^{\mathrm{RGC}}$ est également suggérée par son expression de rechange, obtenue en utilisant (2.11), $\hat{\mathbf{X}}^{\mathrm{RGC}} = \tilde{\mathbf{X}}^{\mathrm{RGC}} + \mathbf{X}_3'\mathbf{L}_{\mathbf{X}}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{L}_{\mathbf{X}}\boldsymbol{\Psi})^{-1}[\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3^{\mathrm{RG}}]$, où $\tilde{\mathbf{X}}^{\mathrm{RGC}} = \hat{\mathbf{X}}_3 + \mathbf{X}_3'\boldsymbol{\Lambda}\mathbf{X}$ $(\mathbf{X}'\boldsymbol{\Lambda}\mathbf{X})^{-1}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) = \tilde{\mathbf{B}}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + (\mathbf{I} - \tilde{\mathbf{B}}_{1\mathbf{x}})\hat{\mathbf{X}}_3$ est l'estimateur par régression composite de $\mathbf{t}_{\mathbf{x}}$ en utilisant l'information sur $\mathbf{x}$ provenant de S_1 et S_3.

En général, l'estimateur RGC $(\hat{\mathbf{X}}^{\mathrm{RGC}}, \hat{\mathbf{Y}}^{\mathrm{RGC}})$ plus simple à calculer, comprenant le coefficient $\hat{\mathcal{B}}$, est moins efficace que l'estimateur par régression optimale composite $(\hat{\mathbf{X}}^{\mathrm{ROC}}, \hat{\mathbf{Y}}^{\mathrm{ROC}})$ qui fait intervenir le coefficient optimal estimé $\hat{\mathcal{B}}^o$ et possède la même variance asymptotique que l'estimateur BLUE donné en (2.3); la perte d'efficacité peut être plus importante dans le cas de l'échantillonnage matriciel emboîté, pour lequel la matrice $\boldsymbol{\Lambda}^0$ n'est pas diagonale par blocs. Par ailleurs, $(\hat{\mathbf{X}}^{\mathrm{ROC}}, \hat{\mathbf{Y}}^{\mathrm{ROC}})$ peut être instable pour les petits échantillons, quand le nombre de degrés de liberté disponibles pour l'estimation de $\hat{\mathcal{B}}^o$ est faible, ce qui est particulièrement le cas dans l'échantillonnage matriciel emboîté; pour une discussion de la stabilité relative de l'estimateur par régression optimale par opposition à la régression généralisée dans le cas d'un seul échantillon, voir Rao (1994) ou Montanari (1998). Pour certaines stratégies d'échantillonnage, décrites dans le théorème qui suit, $\hat{\mathcal{B}} = \hat{\mathcal{B}}^o$ et l'estimateur RGC coïncide avec l'estimateur ROC et, asymptotiquement, avec l'estimateur BLUE; la preuve est donnée en annexe.

Théorème 1 *Considérons les stratégies d'échantillonnage suivantes.*

Plan d'échantillonnage non emboîté

a) *Pour chacun des trois échantillons* S_1, S_2 *et* S_3, *supposons que l'on procède à un échantillonnage aléatoire simple stratifié sans remise (EASSTR) avec fraction d'échantillonnage* $f_{ih} = n_{ih}/N_{ih}$ *dans la strate* h *de l'échantillon* i, $h = 1,\ldots,H_i$ *et que* N_{ih} *désigne la taille de strate, et spécifions les constantes* q_{ik} *dans* $\mathbf{\Lambda}_i$ *sous la forme* $q_{ik} = (n_{ih} - 1)/N_{ih}(1 - f_{ih})$ *pour toutes les unités de la strate* h. *En outre, supposons que, dans chaque échantillon, les unités sont triées par strate, et considérons la matrice de plan augmentée* $\mathcal{Z} = (\mathcal{X}, \mathbf{D})$ *donnée en (2.7), où* $\mathbf{D}$ *est la matrice diagonale par blocs* $\text{diag}\{\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3\}$ *et* $\mathbf{D}_i$ *est la matrice diagonale* $\text{diag}\{\mathbf{1}_{i1},\ldots,\mathbf{1}_{ih},\ldots,\mathbf{1}_{iH_i}\}$, *avec l'élément diagonal* $\mathbf{1}_{ih}$ *correspondant à un vecteur de valeurs un pour toutes les unités de la strate* h *dans l'échantillon* S_i, *et considérons le vecteur augmenté correspondant de totaux de calage* $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}', \mathbf{0}', \mathbf{N}_1', \mathbf{N}_2', \mathbf{N}_3')'$, *où* $\mathbf{N}_i$ *est le vecteur des tailles des strates pour l'échantillon* S_i.

b) *Pour chacun des trois échantillons* S_1, S_2 *et* S_3, *supposons que l'on procède à un échantillonnage de Poisson stratifié et spécifions les constantes* q_{ik} *dans les entrées de* $\mathbf{\Lambda}_i$ *sous la forme* $q_{ik} = \pi_{ihk}/(1 - \pi_{ihk})$ *pour les unités de la strate* h, *où* π_{ihk} *est la probabilité d'inclusion de l'unité* k *dans la strate* h *de la* i^{e} *enquête.*

Plan d'échantillonnage emboîté

a') *Supposons qu'un échantillon aléatoire simple stratifié initial* S *est découpé par strate en trois sous-échantillons aléatoires simples* S_1, S_2 *et* S_3. *Spécifions les fractions d'échantillonnage* f_{ih}, *les constantes* q_{ik} *dans* $\mathbf{\Lambda}_i$, *la matrice de plan* $\mathcal{Z} = (\mathcal{X}, \mathbf{D})$ *et le vecteur des totaux de calage* $\mathbf{t}_{\mathcal{Z}}$ *comme à la partie* (a).

b') *Supposons qu'un échantillon de Poisson stratifié initial* S *est découpé aléatoirement par strate en trois sous-échantillons* S_1, S_2 *et* S_3, *avec probabilités d'inclusion inégales pour les unités de chaque sous-échantillon. Spécifions les constantes* q_{ik} *dans* $\mathbf{\Lambda}_i$ *sous la forme* $q_{ik} = \pi_{ihk}/(1 - \pi_{ihk})$ *pour les unités de la strate* h, *où* π_{ihk} *est la probabilité d'inclusion marginale de l'unité* k *dans la strate* h *pour le* i^{e} *sous-échantillon.*

Sous chacune des stratégies (a) *et* (b), *la procédure de calage avec la matrice* $\mathbf{\Lambda}$ *dans la mesure de distance au sens des moindres carrés donne l'estimateur RGC donné en (3.1) avec* $\hat{\mathcal{B}} = \hat{\mathcal{B}}^o$, *ce qui implique que l'estimateur RGC correspond à l'estimateur ROC. Pour* (a') *et* (b'), *cette constatation est vérifiée approximativement quand les fractions d'échantillonnage dans les strates sont approximativement nulles.*

Corollaire 1 *Le résultat du théorème 1 est également vérifié pour les versions non stratifiées de chacun des quatre plans d'échantillonnage. Pour l'échantillonnage aléatoire simple sans remise (EAS), en particulier, la matrice* $\mathbf{D}$ *se réduit à la matrice diagonale* $\text{diag}\{\mathbf{1}_1, \mathbf{1}_2, \mathbf{1}_3\}$ *ayant pour* i^{e} *élément*

diagonal unitaire de dimension n_i *le vecteur* $\mathbf{1}_i$, *et le vecteur des totaux de calage est alors* $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}', \mathbf{0}', N, N, N)'$.

Corollaire 2 *Dans le cas de l'échantillonnage non emboîté, quand le plan d'échantillonnage pour chacun des trois échantillons est l'un des plans décrits en* (a) *et* (b) *ou l'une de leurs versions non stratifiées, mais qu'il n'est pas le même pour tous les échantillons, le résultat du théorème 1 est vérifié à condition que la matrice* $\mathbf{D}$ *dans* $\mathcal{Z}$ *et le vecteur* $\mathbf{t}_{\mathcal{Z}}$ *soient réduits de manière à correspondre uniquement aux échantillons pour lesquels est utilisé l'EAS ou l'EASSTR.*

Le scénario de calage étendu dans le théorème $1(a, a')$ comprend le calage sur les tailles de strate (ou sur la taille de population dans la version EAS) grâce à l'inclusion d'une ordonnée à l'origine pour chaque strate dans la matrice de plan $\mathcal{X}$. Aucune autre information que celle supposé pour le plan d'échantillonnage (a) ou (a') n'est utilisée, et la forme de l'estimateur RGC résultant demeure la même qu'en (3.1) parce que les estimations HT des tailles de la population et des strates sont exactes. L'effet de ce calage étendu (avec les valeurs spécifiées de q_{ik}) se limite à la conversion du coefficient RGC $\hat{\mathcal{B}}$ en le coefficient optimal $\hat{\mathcal{B}}^o$ et, donc, de l'estimateur RGC en l'estimateur ROC. L'importance pratique de cette conversion réside dans l'exécution de l'estimation par régression optimale composite selon la procédure de calage beaucoup plus simple de l'estimation par régression généralisée.

Le sous-échantillonnage comme à la partie (a'), en fixant a priori les tailles d'échantillon, est une procédure naturelle en échantillonnage matriciel comportant le fractionnement d'un questionnaire. Par contre, dans le scénario de sous-échantillonnage de la partie (b'), n_i est la taille d'échantillon prévue de S_i, la taille réelle étant aléatoire. Des probabilités de sous-échantillonnage inégales peuvent être déterminées de manière adaptative pour accroître l'efficacité; voir Gonzalez et Eltinge (2008).

Les résultats du théorème 1 pourraient être étendus à d'autres plans d'échantillonnage, comme l'échantillonnage aléatoire simple à deux degrés stratifié sous échantillonnage matriciel non emboîté. Cependant, il ne serait pas plus facile d'apporter les ajustements requis aux matrices $\mathbf{\Lambda}_i$ que d'utiliser directement les matrices $\mathbf{\Lambda}_i^0$ dans le calage pour obtenir l'estimateur par régression optimale composite.

Pour les plans d'échantillonnage autres que ceux supposés dans le théorème 1, la valeur de q_{ik} dans les entrées de $\mathbf{\Lambda}_i$ doit être fixée à $q_{ik} = \tilde{n}_i/(\tilde{n}_1 + \tilde{n}_2 + \tilde{n}_3)$, où $\tilde{n}_i = n_i/d_i$, d_i désigne l'effet de plan, afin de tenir compte des différences de taille effective d'échantillon entre les trois échantillons. Si le même plan est utilisé pour tous les échantillons, alors $\tilde{n}_i = n_i$. La justification de cet ajustement s'appuie sur l'argument donné dans Merkouris (2010) pour un problème similaire d'estimation par régression composite.

4 Estimation composite pour le plan d'échantillonnage matriciel (d)

4.1 Ensemble de variables de base dont les totaux sont connus

Nous commençons par discuter d'un cas particulier du plan d'échantillonnage matriciel (d) dans lequel les totaux sont connus pour les variables qui sont communes aux trois échantillons. Dans ces conditions

d'échantillonnage très réalistes, on recueille aussi auprès de tous les échantillons l'information sur le même vecteur de variables auxiliaires $\mathbf{z}$ pour lequel le vecteur des totaux de population $\mathbf{t}_\mathbf{z}$ est connu. À titre d'illustration, considérons de nouveau trois échantillons, comme à la figure 2.1 (mais avec $\mathbf{z}$ ajouté dans tous les sous-échantillons). Alors, l'estimateur RGC $\hat{\mathbf{X}}^{\mathrm{RGC}}$ donné en (3.1) peut être augmenté au moyen des termes de régression ordinaires $\hat{\mathbf{B}}_{3\mathbf{x}}(\mathbf{t}_\mathbf{z} - \hat{\mathbf{Z}}_1) + \hat{\mathbf{B}}_{4\mathbf{x}}(\mathbf{t}_\mathbf{z} - \hat{\mathbf{Z}}_2) + \hat{\mathbf{B}}_{5\mathbf{x}}(\mathbf{t}_\mathbf{z} - \hat{\mathbf{Z}}_3)$, où $\hat{\mathbf{Z}}_i, i = 1,2,3$ est l'estimateur HT de $\mathbf{t}_\mathbf{z}$ fondé sur l'échantillon S_i; nous procédons de façon similaire pour $\hat{\mathbf{Y}}^{\mathrm{RGC}}$. Cet estimateur est plus efficace, car il incorpore de l'information additionnelle, et il est généré par une procédure de calage qui comprend les trois contraintes supplémentaires $\hat{\mathbf{Z}}_i^{\mathrm{RGC}} = \mathbf{t}_\mathbf{z}$, et possède la matrice de plan $\boldsymbol{\mathcal{X}}$ donnée en (2.7) augmentée au moyen de la matrice diagonale par blocs $\mathbf{Z} = \mathrm{diag}\{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3\}$. Dans le cas le plus simple où les matrices d'échantillon $\mathbf{Z}_i$ se réduisent à la colonne de valeurs unitaires $\mathbf{1}_i$ (avec total correspondant de la taille de la population), le scénario de calage est celui spécifié dans le corollaire 1 susmentionné. Comme il est montré dans la preuve du prochain théorème, une application du lemme 1 à la procédure actuelle de calage, avec la matrice de plan partitionnée $(\boldsymbol{\mathcal{X}}, \mathbf{Z}), \mathbf{R} = \boldsymbol{\Lambda}$ et les totaux de calage $(\mathbf{0}', \mathbf{0}', \mathbf{t}_\mathbf{z}', \mathbf{t}_\mathbf{z}', \mathbf{t}_\mathbf{z}')'$, donne une forme RGC modifiée de (3.1) avec les estimateurs RG incorporant l'information sur $\mathbf{z}$ à la place des estimateurs HT. Cela s'écrit de manière compacte sous la forme $\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RG}} - \hat{\boldsymbol{\mathcal{B}}}\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RG}}$, où $\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RG}} = \hat{\boldsymbol{\mathcal{X}}}_3 + \boldsymbol{\mathcal{X}}_3'\boldsymbol{\Lambda}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z})^{-1}(\mathbf{t}_{(\mathbf{z})} - \hat{\mathbf{Z}})$, avec $\mathbf{t}_{(\mathbf{z})} = (\mathbf{t}_\mathbf{z}', \mathbf{t}_\mathbf{z}', \mathbf{t}_\mathbf{z}')'$, et $\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RG}}$ sont exprimés de manière similaire, et où $\hat{\boldsymbol{\mathcal{B}}} = [\boldsymbol{\mathcal{X}}_3'\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})\boldsymbol{\mathcal{X}}]$ $[\boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})\boldsymbol{\mathcal{X}}]^{-1}$ avec $\mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Lambda}$.

Le remplacement de $\boldsymbol{\Lambda}$ par $\boldsymbol{\Lambda}^0$ dans la procédure de calage donne l'estimateur par régression optimale composite, écrit de manière compacte sous la forme $\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RO}} - \hat{\boldsymbol{\mathcal{B}}}^o\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RO}}$, avec les estimateurs par régression optimale incorporant l'information sur $\mathbf{z}$ à la place des estimateurs RG, et avec $\hat{\boldsymbol{\mathcal{B}}}^o = [\boldsymbol{\mathcal{X}}_3'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}^0)\boldsymbol{\mathcal{X}}][\boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}^0)\boldsymbol{\mathcal{X}}]^{-1}$ où $\mathbf{P}_\mathbf{Z}^0 = \mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Lambda}^0$. En notant que $(\mathbf{I} - \mathbf{P}_\mathbf{Z}^0)\boldsymbol{\mathcal{X}}_3$ est la matrice des résidus correspondant à $\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RO}}$, et que $\boldsymbol{\mathcal{X}}_3'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}^0)\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}_3'(\mathbf{I} - \mathbf{P}_\mathbf{Z}^0)'$ $\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}^0)\boldsymbol{\mathcal{X}} = \widehat{\mathrm{AC}}(\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RO}}, \hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RO}})$, et de même pour $\widehat{\mathrm{AV}}(\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RO}})$, il s'ensuit que

$$\hat{\boldsymbol{\mathcal{B}}}^o = -\widehat{\mathrm{AC}}\left[\begin{pmatrix}\hat{\mathbf{X}}_3^{\mathrm{RO}}\\ \hat{\mathbf{Y}}_3^{\mathrm{RO}}\end{pmatrix}, \begin{pmatrix}\hat{\mathbf{X}}_1^{\mathrm{RO}} - \hat{\mathbf{X}}_3^{\mathrm{RO}}\\ \hat{\mathbf{Y}}_2^{\mathrm{RO}} - \hat{\mathbf{Y}}_3^{\mathrm{RO}}\end{pmatrix}\right]\left[\widehat{\mathrm{AV}}\begin{pmatrix}\hat{\mathbf{X}}_1^{\mathrm{RO}} - \hat{\mathbf{X}}_3^{\mathrm{RO}}\\ \hat{\mathbf{Y}}_2^{\mathrm{RO}} - \hat{\mathbf{Y}}_3^{\mathrm{RO}}\end{pmatrix}\right]^{-1}, \tag{4.1}$$

par analogie avec (2.4), ou avec (2.5) sous échantillonnage non emboîté. Donc, $\hat{\boldsymbol{\mathcal{B}}}^o$ est optimal au sens de la minimisation de la variance approximative de l'estimateur $\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RO}} - \hat{\boldsymbol{\mathcal{B}}}^o\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RO}}$, qui est alors asymptotiquement équivalent à l'estimateur BLUE. Un estimateur de rechange, d'optimalité plus faible, prend la forme $\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RG}} - \hat{\boldsymbol{\mathcal{B}}}^{wo}\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RG}}$, où le coefficient $\hat{\boldsymbol{\mathcal{B}}}^{wo} = \left[\boldsymbol{\mathcal{X}}_3'(\mathbf{I} - \mathbf{P}_\mathbf{Z})'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z})\boldsymbol{\mathcal{X}}\right]$ $\left[\boldsymbol{\mathcal{X}}'(\mathbf{I} - \mathbf{P}_\mathbf{Z})'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z})\boldsymbol{\mathcal{X}}\right]^{-1}$ possède la forme (4.1), mais avec des estimateurs RG remplaçant les estimateurs RO. Cet estimateur, qui ne diffère de l'estimateur RGC qu'en ce qui concerne le coefficient de régression, est optimal au sens restreint où il est le composite des estimateurs RG incorporant l'information sur $\mathbf{z}$ qui possède une variance approximative minimale. En général, cet estimateur composite ne peut pas être obtenu sous forme d'estimateur par calage. Le théorème qui suit donne les conditions sous lesquelles l'estimateur RGC est optimal dans l'un des deux sens dans le cas de

l'échantillonnage matriciel non emboîté; la preuve est donnée en annexe. La version avec échantillonnage emboîté du théorème, ainsi que les scénarios de sous-échantillonnage et la preuve tels qu'au théorème 1, sont omis par souci de concision.

Théorème 2 *Considérons les stratégies d'échantillonnage qui suivent.*

a) *Pour chacun des trois échantillons* S_1, S_2 *et* S_3, *supposons un EAS avec les fractions d'échantillonnage* $f_i = n_i/N$, *et spécifions toutes les constantes* q_{ik} *dans* $\mathbf{\Lambda}_i$ *sous la forme* $q_{ik} = (n_i - 1)/N(1 - f_i)$. *Considérons la matrice de plan augmentée* $\mathcal{Z} = (\mathcal{X}, \mathbf{Z})$ *en (2.7), où* $\mathbf{Z} = \text{diag}\{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3\}$, *avec le vecteur augmenté correspondant de totaux de calage* $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}', \mathbf{0}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}')'$. *En outre, supposons que* $\mathbf{Z}_i\mathbf{h}_i = \mathbf{1}$ *pour les vecteurs constants* $\mathbf{h}_i$.

Alors, la procédure de calage donne l'estimateur RGC comme étant $\hat{\mathcal{X}}_3^{\text{RG}} - \hat{\mathcal{B}}\hat{\mathcal{X}}^{\text{RG}} = \hat{\mathcal{X}}_3^{\text{RG}} - \hat{\mathcal{B}}^{wo}\hat{\mathcal{X}}^{\text{RG}}$, *c'est-à-dire que l'estimateur RGC est le composite optimal des estimateurs RG incorporant l'information sur* $\mathbf{z}$.

b) *Pour chacun des trois échantillons* S_1, S_2 *et* S_3, *supposons un EASSTR avec la fraction d'échantillonnage* $f_{ih} = n_{ih}/N_{ih}$ *dans la strate* h *de l'échantillon* $i, h = 1, \ldots, H_i$, *et que* N_{ih} *désigne la taille de strate, et spécifions les constantes dans* $\mathbf{\Lambda}_i$ *sous la forme* $q_{ik} = (n_{ih} - 1)/N_h(1 - f_{ih})$ *pour toutes les unités de la strate* h. *En outre, supposons que, dans chaque échantillon, les unités sont triées par strate, et considérons la matrice de plan augmentée* $\mathcal{Z} = (\mathcal{X}, \mathbf{Z}, \mathbf{D})$ *donnée en (2.7), avec le vecteur augmenté correspondant de totaux de calage* $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}', \mathbf{0}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}', \mathbf{N}_1', \mathbf{N}_2', \mathbf{N}_3')'$. *Les définitions de* $\mathbf{D}$ *et* $\mathbf{N}_i$ *sont les mêmes qu'auparavant.*

Alors, la procédure de calage donne l'estimateur RGC sous la forme $\hat{\mathcal{X}}_3^{\text{RO}} - \hat{\mathcal{B}}^{o}\hat{\mathcal{X}}^{\text{RO}}$, *c'est-à-dire que l'estimateur RGC est le composite optimal des estimateurs par régression optimale incorporant l'information sur* $\mathbf{z}$.

c) *Pour chacun des trois échantillons* S_1, S_2 *et* S_3, *supposons un échantillonnage de Poisson stratifié et spécifions les constantes* q_{ik} *dans les entrées de* $\mathbf{\Lambda}_i$ *sous la forme* $q_{ik} = \pi_{ihk}/(1 - \pi_{ihk})$ *pour les unités de la strate* h.

Alors, la procédure de calage, avec $\mathcal{Z}$ *et* $\mathbf{t}_{\mathcal{Z}}$ *comme en* (a), *donne l'estimateur RGC sous la forme* $\hat{\mathcal{X}}_3^{\text{RG}} - \hat{\mathcal{B}}\hat{\mathcal{X}}^{\text{RG}} = \hat{\mathcal{X}}_3^{\text{RO}} - \hat{\mathcal{B}}^{o}\hat{\mathcal{X}}^{\text{RO}}$, *c'est-à-dire que les estimateurs RG et RO sont identiques, et que l'estimateur RGC est le composite optimal des estimateurs par régression optimale incorporant l'information sur* $\mathbf{z}$.

La condition $\mathbf{Z}_i\mathbf{h}_i = \mathbf{1}$ en (a) du théorème 2 est habituellement satisfaite quand le vecteur $\mathbf{z}$ contient des variables catégoriques. Des résultats analogues aux corollaires 1 et 2 de la section précédente sont également vérifiés pour les parties (b) et (c) du théorème 2. Ici aussi, pour des plans d'échantillonnage

autres que ceux supposés au théorème 2, la valeur $q_{ik} = \tilde{n}_i/(\tilde{n}_1 + \tilde{n}_2 + \tilde{n}_3)$ doit être utilisée dans les entrées de $\mathbf{\Lambda}$.

Enfin, par analogie avec (3.2) et avec la décomposition appropriée du vecteur de poids calés $\mathbf{c}$, l'estimateur composite $\hat{\mathbf{X}}^{\mathrm{RGC}}$ prend maintenant la forme

$$\hat{\mathbf{X}}^{\mathrm{RGC}} = \hat{\mathbf{B}}_{1\mathbf{x}}\hat{\mathbf{X}}_1^{\mathrm{RG}} + (\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}})\hat{\mathbf{X}}_3^{\mathrm{RG}},$$

où $\hat{\mathbf{X}}_1^{\mathrm{RG}}$ et $\hat{\mathbf{X}}_3^{\mathrm{RG}}$ sont les estimateurs RG utilisant l'information sur $\mathbf{z}$ provenant de S_1, et l'information sur $\mathbf{y}$ et $\mathbf{z}$ provenant de S_2 et S_3, respectivement, et $\hat{\mathbf{B}}_{1\mathbf{x}}$ est le coefficient de régression de la matrice correspondante. L'expression pour $\hat{\mathbf{Y}}^{\mathrm{RGC}}$ est similaire. Naturellement, $\hat{\mathbf{X}}^{\mathrm{RGC}}$ et $\hat{\mathbf{Y}}^{\mathrm{RGC}}$ peuvent être obtenus directement au moyen de ce vecteur $\mathbf{c}$ modifié sous les simples formes linéaires $\hat{\mathbf{X}}^{\mathrm{RGC}} = \mathbf{X}_3'\mathbf{c}_3$ et $\hat{\mathbf{Y}}^{\mathrm{RGC}} = \mathbf{Y}_3'\mathbf{c}_3$.

4.2 Ensemble de variables de base dont les totaux sont inconnus

Examinons maintenant le cas du plan d'échantillonnage matriciel (d) dans lequel les totaux pour les variables $\mathbf{z}$ qui sont communes aux trois échantillons sont inconnus. Dans ces conditions, l'estimation comprend la construction d'un estimateur composite du vecteur des totaux $\mathbf{t}_\mathbf{z}$. En harmonie avec la formulation de la section 2, les estimateurs composites de $\mathbf{t}_\mathbf{x}, \mathbf{t}_\mathbf{y}$ et $\mathbf{t}_\mathbf{z}$ qui sont les meilleures combinaisons linéaires sans biais des estimateurs HT $\hat{\mathbf{X}}_1, \hat{\mathbf{Z}}_1, \hat{\mathbf{Y}}_2, \hat{\mathbf{Z}}_2, \hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3, \hat{\mathbf{Z}}_3$ sont donnés par

$$\begin{aligned}
\hat{\mathbf{X}}^B &= \mathbf{B}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + (\mathbf{I} - \mathbf{B}_{1\mathbf{x}})\hat{\mathbf{X}}_3 + \mathbf{B}_{3\mathbf{x}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3) + \mathbf{B}_{2\mathbf{x}}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_3) + \mathbf{B}_{4\mathbf{x}}(\hat{\mathbf{Z}}_2 - \hat{\mathbf{Z}}_3) \\
\hat{\mathbf{Y}}^B &= \mathbf{B}_{3\mathbf{y}}\hat{\mathbf{Y}}_2 + (\mathbf{I} - \mathbf{B}_{3\mathbf{y}})\hat{\mathbf{Y}}_3 + \mathbf{B}_{1\mathbf{y}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{2\mathbf{y}}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_3) + \mathbf{B}_{4\mathbf{y}}(\hat{\mathbf{Z}}_2 - \hat{\mathbf{Z}}_3) \qquad (4.2) \\
\hat{\mathbf{Z}}^B &= \mathbf{B}_{2\mathbf{z}}\hat{\mathbf{Z}}_1 + \mathbf{B}_{4\mathbf{z}}\hat{\mathbf{Z}}_2 + (\mathbf{I} - \mathbf{B}_{2\mathbf{z}} - \mathbf{B}_{4\mathbf{z}})\hat{\mathbf{Z}}_3 + \mathbf{B}_{1\mathbf{z}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{3\mathbf{z}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3).
\end{aligned}$$

Les estimateurs en (4.2) peuvent s'écrire sous la forme de régression matricielle

$$\begin{pmatrix} \hat{\mathbf{X}}^B \\ \hat{\mathbf{Y}}^B \\ \hat{\mathbf{Z}}^B \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Z}}_3 \end{pmatrix} + \mathcal{B}\begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Z}}_2 - \hat{\mathbf{Z}}_3 \end{pmatrix}, \qquad (4.3)$$

avec la matrice minimisant la variance des coefficients donnée par $\mathcal{B} = -\mathrm{Cov}(\mathbf{u}_3, \mathbf{u}_{12} - \mathbf{u}_3^\star)[V(\mathbf{u}_{12} - \mathbf{u}_3^\star)]^{-1}$, où $\mathbf{u}_3 = (\hat{\mathbf{X}}_3', \hat{\mathbf{Y}}_3', \hat{\mathbf{Z}}_3')'$, $\mathbf{u}_3^\star = (\hat{\mathbf{X}}_3', \hat{\mathbf{Z}}_3', \hat{\mathbf{Y}}_3', \hat{\mathbf{Z}}_3')'$, $\mathbf{u}_{12} = (\hat{\mathbf{X}}_1', \hat{\mathbf{Z}}_1', \hat{\mathbf{Y}}_2', \hat{\mathbf{Z}}_2')'$. Avec les matrices de covariance et de variance estimées, nous obtenons la matrice optimale estimée $\hat{\mathcal{B}}^o$, et (4.3) devient alors un estimateur par régression multivariée optimale. Alors, en procédant comme à la section 2, on peut montrer que

$$\hat{\mathcal{B}}^o = (\mathcal{X}_{3-}'\mathbf{\Lambda}^0\mathcal{X})(\mathcal{X}'\mathbf{\Lambda}^0\mathcal{X})^{-1},$$

où

$$\mathcal{X} = \begin{pmatrix} -\mathbf{X}_1 & -\mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{Y}_2 & -\mathbf{Z}_2 \\ \mathbf{X}_3 & \mathbf{Z}_3 & \mathbf{Y}_3 & \mathbf{Z}_3 \end{pmatrix} \qquad (4.4)$$

est la matrice de plan correspondant à l'estimateur par régression (4.3), $\mathcal{X}_{3-}$ est la matrice $\mathcal{X}$ dont la deuxième colonne est éliminée et dont les deux premières lignes sont fixées égales à zéro, et $\mathbf{\Lambda}^0$ est telle qu'il est défini à la section 2.

Le remplacement de la matrice $\mathbf{\Lambda}^0$ par la matrice de pondération $\mathbf{\Lambda}$ donne le coefficient de régression généralisée $\hat{\mathcal{B}} = (\mathcal{X}'_{3-}\mathbf{\Lambda}\mathcal{X})(\mathcal{X}'\mathbf{\Lambda}\mathcal{X})^{-1}$, et (4.3) devient l'estimateur RGC de $(\mathbf{t}'_{\mathbf{x}}, \mathbf{t}'_{\mathbf{y}}, \mathbf{t}'_{\mathbf{z}})'$

$$\begin{pmatrix} \hat{\mathbf{X}}^{\mathrm{RGC}} \\ \hat{\mathbf{Y}}^{\mathrm{RGC}} \\ \hat{\mathbf{Z}}^{\mathrm{RGC}} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Z}}_3 \end{pmatrix} + \hat{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Z}}_2 - \hat{\mathbf{Z}}_3 \end{pmatrix}. \tag{4.5}$$

L'estimateur (4.5) peut être obtenu de manière commode par une procédure de calage qui donne un vecteur de poids calés pour l'échantillon combiné S de la forme $\mathbf{c} = \mathbf{w} + \mathbf{\Lambda}\mathcal{X}(\mathcal{X}'\mathbf{\Lambda}\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{w})$, comme auparavant, mais qui satisfait maintenant la contrainte supplémentaire $\hat{\mathbf{Z}}_1^{\mathrm{RGC}} = \hat{\mathbf{Z}}_2^{\mathrm{RGC}} = \hat{\mathbf{Z}}_3^{\mathrm{RGC}}$. L'expression (4.5) est alors obtenue simplement comme $\mathcal{X}'_{3-}\mathbf{c}$, fondé sur l'échantillon S_3.

L'expression explicite (4.2), qui ne diffère pour les variantes de la régression optimale et de la régression généralisée que par la forme des coefficients linéaires, montre que les estimateurs composites de $\mathbf{t}_{\mathbf{x}}$ et $\mathbf{t}_{\mathbf{y}}$ sont plus efficaces que leurs analogues dans le plan d'échantillonnage matriciel (c), équation (2.2), parce qu'ils incorporent l'information sur les variables communes $\mathbf{z}$, en supposant que la corrélation avec $\mathbf{x}$ et $\mathbf{y}$ est non nulle. L'expression pour l'estimateur composite de $\mathbf{t}_{\mathbf{z}}$ est particulièrement remarquable : elle comprend une combinaison linéaire des trois estimateurs HT de $\mathbf{t}_{\mathbf{z}}$ dérivée des trois échantillons, ainsi que les deux termes de régression impliquant une efficacité additionnelle par la voie de la corrélation de $\mathbf{z}$ avec $\mathbf{x}$ et $\mathbf{y}$. On s'attendrait à ce que les termes additionnels soient nuls, parce qu'une combinaison optimale des trois estimateurs devrait intégrer toute l'information sur $\mathbf{z}$ disponible dans les trois échantillons. Cependant, en général, les coefficients associés ne sont pas nuls. Sous échantillonnage non emboîté, les conditions dans lesquelles ces coefficients sont nuls sont données par la proposition qui suit, dont la preuve figure en annexe. Le résultat devrait également être vérifié sous échantillonnage emboîté.

Proposition 1 *Les coefficients* $\mathbf{B}_{1\mathbf{z}}$ *et* $\mathbf{B}_{3\mathbf{z}}$ *dans l'estimateur* $\hat{\mathbf{Z}}^B$ *en (4.2) sont nuls uniquement si*

$$\begin{aligned} [V(\hat{\mathbf{Z}}_1)]^{-1}\mathrm{Cov}(\hat{\mathbf{X}}_1, \hat{\mathbf{Z}}_1) &= [V(\hat{\mathbf{Z}}_3)]^{-1}\mathrm{Cov}(\hat{\mathbf{X}}_3, \hat{\mathbf{Z}}_3) \\ [V(\hat{\mathbf{Z}}_2)]^{-1}\mathrm{Cov}(\hat{\mathbf{Y}}_2, \hat{\mathbf{Z}}_2) &= [V(\hat{\mathbf{Z}}_3)]^{-1}\mathrm{Cov}(\hat{\mathbf{Y}}_3, \hat{\mathbf{Z}}_3). \end{aligned} \tag{4.6}$$

Cela peut se produire seulement si les trois échantillons sont sélectionnés selon des plans identiques, y compris des tailles d'échantillon égales, ou s'ils sont sélectionnés selon le même plan avec probabilités d'inclusion égales pour toutes les unités, mais pas nécessairement la même taille d'échantillon.

En notant que les quantités dans chaque membre des équations (4.6) sont les coefficients de régression, suivant la proposition 1, les termes de l'estimateur $\hat{\mathbf{Z}}^B$ incorporant la corrélation de $\mathbf{z}$ avec $\mathbf{x}$ et $\mathbf{y}$ sont nuls uniquement si l'effet de la régression de $\mathbf{x}$ et $\mathbf{y}$ sur $\mathbf{z}$ est identique dans les échantillons S_1 et S_3 et

dans les échantillons S_2 et S_3, respectivement. L'essence de cette constatation est que l'estimation de $\mathbf{t}_\mathbf{z}$ en utilisant uniquement l'information sur $\mathbf{z}$ provenant des trois échantillons, mais en ignorant l'information sur $\mathbf{x}$ et $\mathbf{y}$, sera sous-optimale lorsque l'effet de régression de $\mathbf{x}$ et $\mathbf{y}$ sur $\mathbf{z}$ diffère dans les divers échantillons. L'efficacité de $\hat{\mathbf{Z}}^B$ par rapport à l'estimateur composite $\tilde{\mathbf{Z}}^B$ qui utilise uniquement l'information sur $\mathbf{z}$ a pu être évaluée dans les conditions simples comprenant les scalaires x, y et z, l'échantillonnage aléatoire simple pour S_1 et S_3, et l'échantillonnage de Bernoulli pour S_2, et des taux d'échantillonnage égaux pour les trois échantillons. Alors, seule la première équation de (4.6) est vérifiée. Après de nombreuses opérations algébriques fastidieuses, l'efficacité de $\hat{\mathbf{Z}}^B$ par rapport à $\tilde{\mathbf{Z}}^B$ a été dérivée comme étant $\left[V(\tilde{\mathbf{Z}}^B) - V(\hat{\mathbf{Z}}^B)/V(\tilde{\mathbf{Z}}^B)\right] = G/H$, avec

$$
\begin{aligned}
G &= 2\left(r_{xz}^2 - 1\right)\left(r_{yz} cv_y - cv_z\right)^2 \\
H &= \left(cv_z^2 + 1\right)\left(\left(12 - 9r_{yz}^2\right)r_{xz}^2 - 3r_{xy}\left(2r_{yz}r_{xz} - 1\right) + 12\left(r_{yz}^2 - 1\right)\right)cv_z^2 cv_y^2 \\
&+ 2\left(r_{xy}^2 + r_{yz}^2\right)cv_y^2 + 8\left(r_{xz}^2 - 1\right)cv_y^2 - 4r_{yz}r_{xy}r_{xz}cv_y^2 \\
&+ 6\left(r_{xz}^2 - 1\right)cv_z\left(cv_z - 2r_{yz}cv_y\right)
\end{aligned}
$$

où r_{xy}, r_{xz} et r_{yz} désignent les coefficients de corrélation dans la population, et cv_y, cv_z désigne les coefficients de variation. Même si, dans ce scénario, l'écart par rapport aux conditions de la proposition 1 est minime, différentes configurations des valeurs admissibles pour $r_{xy}, r_{xz}, r_{yz}, cv_y$ et cv_z montrent que le gain d'efficacité peut être considérable, palliant l'inefficacité de l'estimateur HT de $\mathbf{t}_\mathbf{z}$ basé sur l'échantillon de Bernoulli S_2. Par exemple, quand $r_{xy} = 0,3, r_{xz} = 0,3, r_{yz} = 0,3$ et $cv_y = 0,1, cv_z = 0,6$, le gain d'efficacité est de 23 %. Dans le cas de l'estimateur par régression optimale composite $\hat{\mathbf{Z}}^{\mathrm{ROC}}$, avec les coefficients estimés $\hat{\mathbf{B}}_{1\mathbf{z}}^o$ et $\hat{\mathbf{B}}_{3\mathbf{z}}^o$, les coefficients de régression donnés en (4.6) sont estimés, et donc les égalités en (4.6) ne seront jamais vérifiées exactement à cause des différences entre les échantillons. Il en va de même pour l'estimateur RGC $\hat{\mathbf{Z}}^{\mathrm{RGC}}$, pour lequel les équations formellement identiques à (4.6) sont données en fonction des coefficients de la régression généralisée pour l'échantillon.

En ce qui concerne l'efficacité de l'estimateur RGC (4.5), un analogue exact du théorème 1 est vérifié dans les présentes conditions, avec les mêmes stratégies d'échantillonnage que celles pour lesquelles l'estimateur RGC correspond à l'estimateur par régression optimale et, asymptotiquement, à l'estimateur BLUE.

L'estimation composite pour un scénario d'échantillonnage matriciel faisant intervenir un ensemble de variables de base avec des totaux connus ainsi qu'inconnus peut être exécutée en utilisant le scénario de calage étendu évident.

5 Estimation par domaine

Les estimateurs composites pour les domaines (sous-populations) d'intérêt peuvent être obtenus facilement en utilisant les poids calés dérivés aux sections précédentes, c'est-à-dire par sommation des valeurs pondérées d'une variable sur tout domaine $U_d \subset U$. Par exemple, en désignant par $\mathbf{X}_{id}$ la matrice $\mathbf{X}_i$, pour l'échantillon S_i, en fixant les entrées de la k^{e} ligne égale à 0 si $k \notin U_d$, l'estimateur RGC du

total de domaine $\mathbf{t}_{\mathbf{x}d}$ basé sur les poids de S_3 calés selon le scénario du plan (c) (voir la section 3) est donné par

$$\hat{\mathbf{X}}_{3d}^{\mathrm{RGC}} = \mathbf{X}'_{3d}\mathbf{c}_3 = \hat{\mathbf{X}}_{3d}^{\mathrm{RG}} + \mathbf{X}'_{3d}\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X}(\mathbf{X}'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X})^{-1}\left[\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3^{\mathrm{RG}}\right],$$

où $\hat{\mathbf{X}}_{3d}^{\mathrm{RG}} = \hat{\mathbf{X}}_{3d} + \mathbf{X}'_{3d}\boldsymbol{\Lambda}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\boldsymbol{\Lambda}\boldsymbol{\Psi})^{-1}\left(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right)$ et l'indice inférieur d indique le domaine. L'estimateur RGC $\hat{\mathbf{X}}_{1d}^{\mathrm{RGC}}$ basé sur l'échantillon S_1 s'obtient de la même manière. Cependant, contrairement à l'estimateur au niveau de la population (3.2) qui résulte du calage de deux estimateurs l'un sur l'autre au niveau de la population, les estimateurs $\hat{\mathbf{X}}_{1d}^{\mathrm{RGC}}$ et $\hat{\mathbf{X}}_{3d}^{\mathrm{RGC}}$ ne sont pas construits comme des composites de deux estimateurs de domaine basés sur les échantillons S_1 et S_3, et ils ne sont pas identiques. De surcroît, même si $\hat{\mathbf{X}}_{1d}^{\mathrm{RGC}}$ et $\hat{\mathbf{X}}_{3d}^{\mathrm{RGC}}$ incorporent l'information sur $\mathbf{x}$ provenant des échantillons S_1 et S_3, leur construction (non personnalisée au niveau du domaine) peut comporter une certaine perte d'efficacité.

Une simple modification de la procédure de calage qui aboutit à une estimation composite efficace pour tous les totaux d'intérêt comprend l'augmentation de la matrice de plan au moyen de colonnes définies au niveau de chaque domaine pour les variables pertinentes. Donc, pour le plan (c), l'estimation du total de domaine $\mathbf{t}_{\mathbf{x}d}$ comprend l'augmentation de la matrice de plan $\mathcal{X}$ en (2.7) au moyen de la colonne $(-\mathbf{X}'_{1d}, \mathbf{0}', \mathbf{X}'_{3d})'$. L'estimateur résultant, $\tilde{\mathbf{X}}_d^{\mathrm{RGC}}$, peut s'écrire sous la forme

$$\begin{aligned}\tilde{\mathbf{X}}_d^{\mathrm{RGC}} &= \hat{\mathbf{X}}_{3d} + \hat{\mathbf{B}}_{1\mathbf{x}d}\left(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3\right) + \hat{\mathbf{B}}_{2\mathbf{x}d}\left(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right) + \hat{\mathbf{B}}_{3\mathbf{x}d}\left(\hat{\mathbf{X}}_{1d} - \hat{\mathbf{X}}_{3d}\right) \\ &= \hat{\mathbf{B}}_{1\mathbf{x}d}\tilde{\mathbf{X}}_{1d}^{\mathrm{RG}} + \left(\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}d}\right)\tilde{\mathbf{X}}_{3d}^{\mathrm{RG}},\end{aligned} \tag{5.1}$$

où $\tilde{\mathbf{X}}_{1d}^{\mathrm{RG}}$ et $\tilde{\mathbf{X}}_{3d}^{\mathrm{RG}}$ sont maintenant les estimateurs RG de domaine incorporant l'effet de régression des deuxième et troisième termes de (5.1). L'ajout dans (5.1) d'un autre terme comprenant la différence $\hat{\mathbf{Y}}_{2d} - \hat{\mathbf{Y}}_{3d}$ pourrait ne pas améliorer appréciablement l'efficacité de $\tilde{\mathbf{X}}_d^{\mathrm{RGC}}$, mais sera nécessaire si l'estimation du total de domaine $\mathbf{t}_{\mathbf{y}d}$ est également requise. Dans toute situation particulière, l'augmentation de la matrice de plan $\mathcal{X}$ ne concerne que les composantes de $\mathbf{x}$ ou de $\mathbf{y}$ pour lesquelles les estimations par domaine sont nécessaires. Un inconvénient éventuel de cette procédure est le fardeau de calcul supplémentaire, qui augmente avec le nombre de domaines et de variables pour lesquels l'estimation par domaine est requise.

Une autre approche, qui pourrait être plus appropriée quand les estimations par domaine d'intérêt sont nombreuses, consiste à produire séparément les estimations par domaine en exécutant le calage composite uniquement au niveau du domaine. Pour le total de domaine $\mathbf{t}_{\mathbf{x}d}$, cela donnera l'estimateur RGC de domaine, par analogie avec l'estimateur RGC de population (3.2),

$$\breve{\mathbf{X}}_d^{\mathrm{RGC}} = \breve{\mathbf{B}}_{1\mathbf{x}d}\hat{\mathbf{X}}_{1d} + \left(\mathbf{I} - \breve{\mathbf{B}}_{1\mathbf{x}d}\right)\breve{\mathbf{X}}_{3d}^{\mathrm{RG}},$$

où $\breve{\mathbf{B}}_{1\mathbf{x}d} = \mathbf{X}'_{3d}\mathbf{L}_{\boldsymbol{\Psi}_d}\mathbf{X}_d\left(\mathbf{X}'_d\mathbf{L}_{\boldsymbol{\Psi}_d}\mathbf{X}\right)_d^{-1}$ et $\breve{\mathbf{X}}_{3d}^{\mathrm{RG}} = \hat{\mathbf{X}}_{3d} + \mathbf{X}'_{3d}\boldsymbol{\Lambda}\boldsymbol{\Psi}_d\left(\boldsymbol{\Psi}'_d\boldsymbol{\Lambda}\boldsymbol{\Psi}_d\right)^{-1}\left(\hat{\mathbf{Y}}_{2d} - \hat{\mathbf{Y}}_{3d}\right)$. L'efficacité de l'estimateur conjoint $\left(\breve{\mathbf{X}}_d^{\mathrm{RGC}}, \breve{\mathbf{Y}}_d^{\mathrm{RGC}}\right)$ par rapport à l'estimateur $\left(\hat{\mathbf{X}}_d^{\mathrm{RGC}}, \hat{\mathbf{Y}}_d^{\mathrm{RGC}}\right)$ peut être vérifiée sous les conditions de la proposition suivante (dont la preuve est donnée en annexe).

Proposition 2 *Sous les scénarios d'échantillonnage du théorème 1,*

$$\widehat{\mathrm{AV}}\begin{pmatrix}\breve{\mathbf{X}}_{3d}^{\mathrm{RGC}} \\ \breve{\mathbf{Y}}_{3d}^{\mathrm{RGC}}\end{pmatrix} < \widehat{\mathrm{AV}}\begin{pmatrix}\hat{\mathbf{X}}_{3d}^{\mathrm{RGC}} \\ \hat{\mathbf{Y}}_{3d}^{\mathrm{RGC}}\end{pmatrix}.$$

Produire séparément les estimations de domaine, par calage composite au niveau du domaine, a notamment pour inconvénient la perte de cohérence entre les estimations au niveau de la population et au niveau du domaine.

Les considérations qui précèdent s'étendent à l'estimation par domaine dans le cas du plan d'échantillonnage matriciel (d).

6 Une étude par simulation

Nous avons réalisé une simulation pour étudier les propriétés relatives des divers estimateurs composites pour la version emboîtée du plan (c) élémentaire. Les valeurs des variables scalaires corrélées x et y ont été tirées d'une loi log-normale bivariée de moyenne (μ_x, μ_y) et de variance (σ_x^2, σ_y^2). Nous avons fixé $\mu_x = 3,\ \mu_y = 5,$ quatre combinaisons de variances (σ_x^2, σ_y^2) (5 et 10) et avons considéré trois valeurs de la corrélation $\rho(x, y)$ (0,5, 0,7, 0,9). Les variances $\sigma_x^2 = 5,\ \sigma_x^2 = 10$ impliquent une asymétrie de 2,65 et de 4,33, respectivement, tandis que les variances $\sigma_y^2 = 5,\ \sigma_y^2 = 10$ impliquent une asymétrie de 1,43 et de 2,15, respectivement. Pour chacune de ces 12 configurations, nous avons créé une population de taille $N = 1\ 000\ 000.$ De chacune des 12 populations, nous avons tiré un échantillon aléatoire simple S de taille $n = 5\ 000$ sans remise, et l'avons divisé en trois sous-échantillons aléatoires simples (S_1, S_2, S_3) selon deux répartitions différentes, à savoir $(n_1 = 2\ 000, n_2 = 2\ 000, n_3 = 1\ 000)$ et $(n_1 = 1\ 500,$ $n_2 = 1\ 500, n_3 = 2\ 000),$ la deuxième répartition donnant des échantillon combinés plus grands $S_1 \cup S_3$ et $S_2 \cup S_3.$ Donc, un total de 24 configurations de simulation ont été créées. Pour chaque configuration, nous avons calculé les estimateurs HT des totaux t_x et t_y en utilisant l'échantillon complet $S,$ ainsi que l'estimateur HT de t_x en utilisant S_1 et $S_3,$ et l'estimateur HT de t_y en utilisant S_2 et $S_3.$ Pour les estimateurs HT basés sur deux sous-échantillons, nous avons employé la méthode simple de combinaison de deux sous-échantillons (Gonzales et Eltinge 2008) par un ajustement de la pondération faisant intervenir la probabilité de sélection d'une unité de population dans S_1 ou dans S_3 et dans S_2 ou dans $S_3.$ En outre, pour t_x ainsi que $t_y,$ nous avons calculé les estimateurs RGC et ROC. Chaque configuration d'échantillonnage de simulation a été répétée 10 000 fois.

Le biais simulé (en pourcentage) de tous les estimateurs était inférieur à 0,05 %, excepté pour deux configurations comprenant $\sigma_x^2 = 10,$ avec l'asymétrie de population associée de 4,33, pour lesquelles les plus grandes valeurs observées de 0,14 % et 0,17 % correspondent aux estimateurs RGC et ROC pour $t_x,$ respectivement, sous la répartition d'échantillon (2 000, 2 000, 1 000), et tombent à 0,10 % et 0,13 % sous la répartition plus favorable (1 500, 1 500, 2 000). Donc, les efficacités relatives des estimateurs sont évaluées en utilisant leurs variances sous le plan de sondage simulé.

Le tableau 6.1 montre l'efficacité des estimateurs composites RGC et ROC par rapport aux estimateurs HT qui utilisent $S_1 \cup S_3$ et $S_2 \cup S_3.$ La mesure de cette efficacité relative est la différence relative entre les variances en pourcentage [V(RGC)-V(HT)]/V(HT) et [V(ROC)-V(HT)]/V(HT). Une valeur négative de cette mesure indique le gain d'efficacité obtenu avec les deux estimateurs composites. La perte d'efficacité simulée des estimateurs HT de t_x ainsi que t_y due au fait de ne pas utiliser l'échantillon complet $S,$ qui n'est pas présentée au tableau 6.1, est très proche de la perte nominale pour

l'EAS, c'est-à-dire 66,8 % pour la répartition (2 000, 2 000, 1 000) et 43,1 % pour la répartition (1 500, 1 500, 2 000).

Tableau 6.1
Différences relatives (en pourcentage) entre les variances de RGC et ROC par rapport à HT pour x et y, basées sur 10 000 échantillons simulés avec deux répartitions d'échantillons différentes

(n1, n2, n3)	(2 000; 2 000; 1 000)				(1 500; 1 500; 2 000)			
	x		*y*		*x*		*y*	
	RGC	**ROC**	**RGC**	**ROC**	**RGC**	**ROC**	**RGC**	**ROC**
$\sigma_x^2 = 5\ \ \sigma_y^2 = 5$								
$\rho = 0,5$	-2,24	-6,86	26,39	-6,23	-5,19	-6,29	12,59	-6,52
$\rho = 0,7$	-11,90	-14,75	10,21	-13,96	-12,78	-13,24	0,25	-13,13
$\rho = 0,9$	-24,89	-28,57	-12,49	-28,10	-21,55	-23,37	-14,55	-23,03
$\sigma_x^2 = 5\ \ \sigma_y^2 = 10$								
$\rho = 0,5$	-0,27	-6,75	6,50	-6,26	-3,94	-6,60	0,50	-6,44
$\rho = 0,7$	-11,47	-14,56	-6,29	-14,04	-12,87	-13,51	-9,51	-13,10
$\rho = 0,9$	-28,14	-28,42	-25,74	-28,23	-23,70	-23,54	-22,07	-23,09
$\sigma_x^2 = 10\ \ \sigma_y^2 = 5$								
$\rho = 0,5$	-4,57	-6,51	28,64	-6,17	-5,90	-5,98	17,57	-6,44
$\rho = 0,7$	-11,29	-14,37	16,08	-13,92	-11,66	-12,90	6,69	-13,00
$\rho = 0,9$	-20,32	-28,09	-2,46	-28,19	-18,46	-22,97	-6,97	-22,91
$\sigma_x^2 = 10\ \ \sigma_y^2 = 10$								
$\rho = 0,5$	-4,79	-6,49	8,54	-6,13	-6,06	-6,22	3,41	-6,34
$\rho = 0,7$	-13,27	-14,28	-2,57	-13,95	-13,27	-13,15	-6,00	-12,93
$\rho = 0,9$	-26,01	-28,06	-20,37	-28,21	-22,18	-23,17	-18,48	-22,89

Pour la variable x, l'utilisation de l'estimateur RGC sous une faible corrélation $\rho = 0,5$ et avec la répartition (2 000, 2 000, 1 000) donne un gain d'efficacité qui varie de 0,27 % à 4,79 % pour les quatre configurations de variance différentes; ce gain reflète la quantité d'information perdue qui est récupérée par l'estimateur RGC. Un gain important est réalisé pour $\rho = 0,7$, variant de 11,29 % à 13,27 %, et un gain encore plus important pour $\rho = 0,9$, variant de 20,32 % à 28,14 %. Avec la répartition d'échantillon (1 500, 1 500, 2 000), l'estimateur RGC donne de meilleurs résultats pour $\rho = 0,5$, et $\rho = 0,7$, mais non pour $\rho = 0,9$. Un gain supplémentaire est produit par l'estimateur ROC, qui est plus efficace que l'estimateur RGC dans toutes les configurations sauf deux (où les estimateurs sont aussi efficaces l'un que l'autre, voir la colonne 7). L'efficacité de l'estimateur ROC par rapport à l'estimateur HT est proche de la valeur nominale pour l'efficacité de l'EAS, qui est de 6,25, 13,92 et 28,12 pour $\rho = 0,5$, $\rho = 0,7$, $\rho = 0,9$, respectivement, pour la répartition (2 000, 2 000, 1 000) et de 6,417, 13,186 et 23,30 pour la répartition (1 500, 1 500, 2 000); voir la quantité E à l'avant-avant dernier paragraphe de la section 2. Comme prévu, l'estimateur RGC concurrence mieux l'estimateur ROC lorsque la corrélation et la taille d'échantillon augmentent.

Pour la variable y, l'estimateur RGC est inférieur à l'estimateur HT au niveau de corrélation $\rho = 0{,}5$ et, dans la moitié des configurations simulées au niveau $\rho = 0{,}7$; voir les valeurs positives dans les colonnes 4 et 8. Cette inefficacité de l'estimateur RGC varie de 6,50 % (pour $\rho = 0{,}7$) à 28,64 % (pour $\rho = 0{,}5$) pour la répartition d'échantillon (2 000, 2 000, 1 000), et se réduit pour varier de 0,25 % (pour $\rho = 0{,}7$) à 17,57 % (pour $\rho = 0{,}5$) pour la répartition d'échantillon (1 500, 1 500, 2 000). Cela s'explique par la plus grande asymétrie de x (la variable x étant utilisée comme variable auxiliaire de y dans la régression); les niveaux plus faibles d'inefficacité sont observés pour $\sigma_y^2 = 10$, quand la différence d'asymétrie entre x et y est la plus petite. Par ailleurs, au niveau de corrélation $\rho = 0{,}9$ et avec la répartition (2 000, 2 000, 1 000), le gain d'efficacité de l'estimateur RGC par rapport à l'estimateur HT varie de 2,46 % (quand la différence d'asymétrie est la plus grande) à 25,74 % (quand la différence d'asymétrie est la plus petite), avec des niveaux d'efficacité similaires observés pour la répartition (1 500, 1 500, 2 000). L'estimateur ROC est plus efficace que l'estimateur RGC dans toutes les configurations, l'efficacité relative étant proche de l'efficacité nominale pour l'EAS (même efficacité qu'avec x). Pour y aussi, l'estimateur RGC concurrence mieux l'estimateur ROC lorsque la corrélation et la taille d'échantillon augmentent.

Cette étude empirique limitée, qui simule essentiellement la version EAS du théorème $1(a')$, confirme la théorie sur l'efficacité de l'estimateur optimal ROC, même pour une modeste taille d'échantillon, et montre l'utilité des deux estimateurs composites RGC et ROC pour ce qui est de récupérer partiellement l'information perdue en raison du fractionnement du questionnaire complet. Elle montre aussi que le pratique estimateur RGC n'est pas toujours un bon substitut de l'estimateur ROC quand les échantillons sont petits et que la corrélation entre x et y est faible.

7 Discussion

La méthode d'estimation proposée pour l'échantillonnage matriciel comprend un calage en une étape des poids de l'échantillon combiné. Les estimations des totaux pour toutes les variables peuvent être obtenues en utilisant uniquement les unités de l'échantillon S_3 et leurs poids calés qui incorporent toute l'information disponible provenant des trois échantillons. Ces poids pourraient être utilisés pour calculer d'autres statistiques pondérées, dont des moyennes, des ratios, des quantiles et des coefficients de régression. Lorsque les probabilités d'inclusion d'ordre deux sont connues, y compris les probabilités d'inclusion interéchantillons dans le cas emboîté, la procédure de calage de la section 2 peut produire des estimateurs par régression optimale composites et leurs variances, mais les calculs sont très difficiles. Pour des configurations d'échantillonnage générales, le scénario de calage beaucoup plus simple de la section 3 produit facilement des estimateurs par régression généralisée composites, qui, pour certaines stratégies d'échantillonnage, sont des estimateurs par régression optimale.

L'estimation de la variance d'un estimateur RGC peut, en principe, être fondée sur la méthode de linéarisation de Taylor de l'estimateur par régression généralisée (voir, par exemple, Särndal et coll. 1992, pages 235 et 237). Cette approche requiert des calculs qui pourraient ne pas être pratiques, voire même possibles, pour des plans d'échantillonnage complexes, parce que les probabilités d'inclusion d'ordre deux sont rarement connues. Les méthodes de rééchantillonnage pour l'estimation de la variance, telles que la méthode du jackknife ou la méthode du bootstrap (voir, par exemple, Rust et Rao 1996), peuvent être

appliquées aux estimateurs RGC des sections précédentes. Ainsi, la méthode du jackknife, habituellement utilisée dans les enquêtes avec plan d'échantillonnage stratifié à plusieurs degrés, pourrait être utilisée pour répéter les procédures de calage qui donnent lieu aux estimateurs RGC. Pour le plan d'échantillonnage non emboîté, il est nécessaire d'appliquer la méthode du jackknife à l'échantillon combiné, en traitant les trois échantillons indépendants comme des superstrates d'échantillon contenant les strates de l'échantillon. La procédure de rééchantillonnage s'appliquerait alors à l'échantillon combiné trié par échantillon et par strate dans chaque échantillon, pour produire les répliques des poids calés définis aux sections précédentes. Le nombre total de strates utilisées dans la procédure de rééchantillonnage par le jackknife est le nombre total de strates dans les trois échantillons, chaque réplique comprenant toutes les strates. Les fichiers de microdonnées à grande diffusion peuvent contenir les poids de rééchantillonnage calés pour permettre aux utilisateurs d'estimer facilement la variance. À cette fin également, seuls les poids de rééchantillonnage pour S_3 doivent être inclus, ce qui permet de réaliser une importante économie de stockage de données dans ces fichiers de microdonnées. Le cas du plan d'échantillonnage emboîté est plus compliqué. Des investigations plus poussées dans cette direction seront le sujet d'une étude distincte.

La méthode d'estimation décrite s'adapte facilement aux plans d'échantillonnage matriciel comprenant plus de deux sous-questionnaires ou plus de trois sous-échantillons, ce qui fait ressortir la puissance opérationnelle de la procédure de calage. Dans chaque cas, l'étape cruciale consiste à déterminer la matrice de plan $\mathcal{X}$. De tels plans peuvent comporter des scénarios plus complexes en ce qui concerne le nombre de sous-questionnaires administrés aux divers sous-échantillons. Toutes les estimations composites peuvent alors être obtenues en utilisant uniquement les valeurs des variables pondérées provenant du nombre minimal de sous-échantillons qui, combinés, contiennent tous les items.

Remerciements

L'auteur remercie le rédacteur, le rédacteur associé et deux examinateurs de leurs commentaires et suggestions qui lui ont permis d'améliorer considérablement le manuscrit.

Annexe

Preuve du lemme 1

Pour la matrice partitionnée $\mathcal{X} = (\mathbf{X}, \boldsymbol{\Psi})$, le vecteur $\mathbf{c} = \mathbf{w} + \mathbf{R}\mathcal{X}(\mathcal{X}'\mathbf{R}\mathcal{X})^{-1}(\mathbf{t}_{\mathcal{X}} - \mathcal{X}'\mathbf{w})$ prend la forme

$$\begin{aligned}\mathbf{c} &= \mathbf{w} + (\mathbf{R}\mathbf{X}, \mathbf{R}\boldsymbol{\Psi})\begin{pmatrix}\mathbf{X}'\mathbf{R}\mathbf{X} & \mathbf{X}'\mathbf{R}\boldsymbol{\Psi}\\ \boldsymbol{\Psi}'\mathbf{R}\mathbf{X} & \boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{t}_{\mathbf{X}} - \mathbf{X}'\mathbf{w}\\ \mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}\end{pmatrix}\\ &= \mathbf{w} + (\mathbf{R}\mathbf{X}\mathbf{A}_{11} + \mathbf{R}\boldsymbol{\Psi}\mathbf{A}_{21})(\mathbf{t}_{\mathbf{X}} - \mathbf{X}'\mathbf{w}) + (\mathbf{R}\mathbf{X}\mathbf{A}_{12} + \mathbf{R}\boldsymbol{\Psi}\mathbf{A}_{22})(\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}),\end{aligned}$$

où, découlant de l'algèbre des matrices partitionnées, $\mathbf{A}_{11} = [\mathbf{X}'\mathbf{R}\mathbf{X} - \mathbf{X}'\mathbf{R}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\mathbf{R}\mathbf{X}]^{-1} = [\mathbf{X}'\mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Psi}})\mathbf{X}]^{-1}$ avec $\mathbf{P}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\mathbf{R}$, $\mathbf{A}_{22} = [\boldsymbol{\Psi}'\mathbf{R}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\Psi}']^{-1}$ avec $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}$, $\mathbf{A}_{12} = -(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{R}\boldsymbol{\Psi})\mathbf{A}_{22}$ et $\mathbf{A}_{21} = -(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}'\mathbf{R}\mathbf{X})\mathbf{A}_{11}$. Alors, l'équation (2.9) s'ensuit sans

difficulté. Pour prouver l'équation (2.10), nous posons que $\mathbf{c}_\Psi = \mathbf{w} + \mathbf{R}\Psi(\Psi'\mathbf{R}\Psi)^{-1}(\mathbf{t}_\Psi - \Psi'\mathbf{w})$, de sorte que $(\mathbf{X}'\mathbf{R}\Psi)(\Psi'\mathbf{R}\Psi)^{-1}(\mathbf{t}_\Psi - \Psi'\mathbf{w}) = \mathbf{X}'\mathbf{c}_\Psi - \mathbf{X}'\mathbf{w}$, et nous utilisons la forme de rechange $\mathbf{A}_{22} = (\Psi'\mathbf{R}\Psi)^{-1} + (\Psi'\mathbf{R}\Psi)^{-1}(\Psi'\mathbf{R}\mathbf{X})\mathbf{A}_{11}(\mathbf{X}'\mathbf{R}\Psi)(\Psi'\mathbf{R}\Psi)^{-1}$ pour écrire $\mathbf{c}$ susmentionné sans le deuxième terme sous la forme

$$\begin{aligned}
\mathbf{w} &+ \mathbf{R}\Psi\mathbf{A}_{22}(\mathbf{t}_\Psi - \Psi'\mathbf{w}) - \mathbf{R}\mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{R}\Psi)\mathbf{A}_{22}(\mathbf{t}_\Psi - \Psi'\mathbf{w}) \\
&= \mathbf{w} + \left[\mathbf{R}\Psi(\Psi'\mathbf{R}\Psi)^{-1} + \mathbf{R}\Psi(\Psi'\mathbf{R}\Psi)^{-1}(\Psi'\mathbf{R}\mathbf{X})\mathbf{A}_{11}(\mathbf{X}'\mathbf{R}\Psi)(\Psi'\mathbf{R}\Psi)^{-1}\right](\mathbf{t}_\Psi - \Psi'\mathbf{w}) \\
&- \mathbf{R}\mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}\left[\mathbf{I} + (\mathbf{X}'\mathbf{R}\Psi)(\Psi'\mathbf{R}\Psi)^{-1}(\Psi'\mathbf{R}\mathbf{X})\mathbf{A}_{11}\right](\mathbf{X}'\mathbf{R}\Psi)(\Psi'\mathbf{R}\Psi)^{-1}(\mathbf{t}_\Psi - \Psi'\mathbf{w}) \\
&= \mathbf{c}_\Psi + \mathbf{R}\Psi(\Psi'\mathbf{R}\Psi)^{-1}(\Psi'\mathbf{R}\mathbf{X})\mathbf{A}_{11}(\mathbf{X}'\mathbf{c}_\Psi - \mathbf{X}'\mathbf{w}) \\
&- \mathbf{R}\mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}\left[\mathbf{I} + (\mathbf{X}'\mathbf{R}\Psi)(\Psi'\mathbf{R}\Psi)^{-1}(\Psi'\mathbf{R}\mathbf{X})\mathbf{A}_{11}\right](\mathbf{X}'\mathbf{c}_\Psi - \mathbf{X}'\mathbf{w}) \\
&= \mathbf{c}_\Psi + \mathbf{R}\Psi(\Psi'\mathbf{R}\Psi)^{-1}(\Psi'\mathbf{R}\mathbf{X})\mathbf{A}_{11}(\mathbf{X}'\mathbf{c}_\Psi - \mathbf{X}'\mathbf{w}) \\
&- \mathbf{R}\mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}\left[\mathbf{I} + (\mathbf{X}'\mathbf{R}\mathbf{X} - \mathbf{A}_{11}^{-1})\mathbf{A}_{11}\right](\mathbf{X}'\mathbf{c}_\Psi - \mathbf{X}'\mathbf{w}) \\
&= \mathbf{c}_\Psi + \left[\mathbf{R}\Psi(\Psi'\mathbf{R}\Psi)^{-1}(\Psi'\mathbf{R}\mathbf{X}) - \mathbf{R}\mathbf{X}\right]\mathbf{A}_{11}(\mathbf{X}'\mathbf{c}_\Psi - \mathbf{X}'\mathbf{w}) \\
&= \mathbf{c}_\Psi - \mathbf{R}(\mathbf{I} - \mathbf{P}_\Psi)\mathbf{X}\left[\mathbf{X}'\mathbf{R}(\mathbf{I} - \mathbf{P}_\Psi)\mathbf{X}\right]^{-1}(\mathbf{X}'\mathbf{c}_\Psi - \mathbf{X}'\mathbf{w}).
\end{aligned}$$

L'ajout à cela du deuxième terme de $\mathbf{c}$ provenant de (2.9) donne (2.10) sous la forme explicite

$$\mathbf{c}_\Psi + \mathbf{R}(\mathbf{I} - \mathbf{P}_\Psi)\mathbf{X}\left[\mathbf{X}'\mathbf{R}(\mathbf{I} - \mathbf{P}_\Psi)\mathbf{X}\right]^{-1}(\mathbf{t}_\mathbf{X} - \mathbf{X}'\mathbf{c}_\Psi).$$

Preuve du théorème 1

a) Le calage avec la matrice de plan $\mathcal{Z} = (\mathcal{X}, \mathbf{D})$ et le vecteur de totaux $\mathbf{t}_\mathcal{Z} = (\mathbf{0}', \mathbf{N}')'$, avec $\mathbf{0} = (\mathbf{0}', \mathbf{0}')'$, $\mathbf{N} = (\mathbf{N}_1', \mathbf{N}_2', \mathbf{N}_3')'$, donne le vecteur de poids calés $\mathbf{c} = \mathbf{w} + \mathbf{\Lambda}\mathcal{Z}(\mathcal{Z}'\mathbf{\Lambda}\mathcal{Z})^{-1}(\mathbf{t}_\mathcal{Z} - \mathcal{Z}'\mathbf{w})$, qui, en vertu du lemme 1, s'écrit sous la forme $\mathbf{c} = \mathbf{c}_\mathbf{D} + \mathbf{L}_\mathbf{D}\mathcal{X}(\mathcal{X}'\mathbf{L}_\mathbf{D}\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{c}_\mathbf{D})$, où $\mathbf{c}_\mathbf{D} = \mathbf{w} + \mathbf{\Lambda}\mathbf{D}(\mathbf{D}'\mathbf{\Lambda}\mathbf{D})^{-1}(\mathbf{N} - \mathbf{D}'\mathbf{w})$ et $\mathbf{L}_\mathbf{D} = \mathbf{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{D})$, avec $\mathbf{P}_\mathbf{D} = \mathbf{D}(\mathbf{D}'\mathbf{\Lambda}\mathbf{D})^{-1}\mathbf{D}'\mathbf{\Lambda}$. Dans le cas de l'EASSTR avec $f_{ih} = n_{ih}/N_{ih}$, $\mathbf{D}'\mathbf{w} = \hat{\mathbf{N}} = \mathbf{N}$ et, donc $\mathbf{c} = \mathbf{w} + \mathbf{L}_\mathbf{D}\mathcal{X}(\mathcal{X}'\mathbf{L}_\mathbf{D}\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{w})$. Alors, compte tenu de (2.8), afin de montrer que $\hat{\mathcal{B}} = \hat{\mathcal{B}}^o$, il suffit de montrer que $\mathbf{L}_\mathbf{D} = \mathbf{\Lambda}^0$. Pour l'EASSTR, il est facile de montrer que $\mathbf{\Lambda}^0 = \text{diag}\{\lambda_{ih}(\mathbf{I} - \mathbf{P}_{\mathbf{1}ih})\}$, où $\lambda_{ih} = N_{ih}^2(1 - f_{ih})/[n_{ih}(n_{ih} - 1)]$ et $\mathbf{P}_{\mathbf{1}ih} = \mathbf{1}_{ih}(\mathbf{1}_{ih}'\mathbf{1}_{ih})^{-1}\mathbf{1}_{ih}'$. Ensuite, observons que la matrice $\mathbf{P}_\mathbf{D}$ est diagonale avec pour ih^{e} entrée $\mathbf{1}_{ih}(\mathbf{1}_{ih}'\mathbf{\Lambda}_{ih}\mathbf{1}_{ih})^{-1}\mathbf{1}_{ih}'\mathbf{\Lambda}_{ih} = \mathbf{P}_{\mathbf{1}ih}$, parce que les éléments de $\mathbf{\Lambda}_{ih}$ sont constants. Comme cet élément constant est $w_{ik}/q_{ik} = (N_{ih}/n_{ih})[N_{ih}(1 - f_{ih})/(n_{ih} - 1)] = \lambda_{ih}$, nous obtenons $\mathbf{L}_\mathbf{D} = \text{diag}\{\mathbf{\Lambda}_{ih}(\mathbf{I} - \mathbf{P}_{\mathbf{1}ih})\} = \mathbf{\Lambda}^0$, c.q.f.d.

b) Pour l'échantillonnage de Poisson, $\mathbf{\Lambda}_i^0 = \text{diag}\{(1 - \pi_{ihk})/\pi_{ihk}^2\}, h = 1, \ldots, H_i$. La preuve découle immédiatement de l'observation que, avec les constantes spécifiées q_{ik} dans les entrées de $\mathbf{\Lambda}_i$, nous avons $\mathbf{\Lambda}_i = \mathbf{\Lambda}_i^0$.

a') Pour simplifier, laissons tomber l'indice inférieur de strate. Le sous-échantillonnage aléatoire simple est effectué séquentiellement avec des tailles fixes n_1, n_2 et n_3. On peut montrer que les probabilités d'inclusion marginales d'ordre un et d'ordre deux pour S_i sont $\pi_{ik} = n_i/N$ et $\pi_{ikl} = n_i(n_i-1)/[N(N-1)]$, comme si S_i était tiré directement de U. Un argument combinatoire montre que la probabilité d'inclusion d'ordre deux conditionnelle (sachant S) pour S_i et S_j est $\pi_{ikjl|S} = n_i n_j/[n(n-1)]$ et donc que la probabilité d'inclusion marginale est $\pi_{ikjl} = n_i n_j/[N(N-1)]$. Pour $k = l$, $\pi_{ikjk} = 0$. Alors $\Delta_{kl} = \pi_{ikjl} - \pi_{ik}\pi_{jl} = n_i n_j/[N^2(N-1)]$ et $\Delta_{kk} = -n_i n_j/N^2$. Donc $\Delta_{kl} \approx 0$, pour $k, l \in U$ quand les fractions d'échantillonnage sont faibles, et donc $\boldsymbol{\Lambda}^0 \approx \text{diag}\{\boldsymbol{\Lambda}_i^0\}$. L'optimalité de l'estimateur RGC découle alors du théorème 1 (a).

b') Attribuer aléatoirement les unités de S aux trois sous-échantillons, avec une taille de sous-échantillon prévue fixe, implique que l'inclusion des unités est effectuée indépendamment à l'intérieur des sous-échantillons et entre les sous-échantillons. Puisque, dans l'échantillonnage de Poisson, les unités de U sont également incluses dans S indépendamment, $\Delta_{kl} = \pi_{ikjl} - \pi_{ik}\pi_{jl} = 0$ et $\Delta_{kk} = -\pi_{ik}\pi_{jl}$. Δ_{kk} est approximativement nul pour les petites fractions d'échantillonnage, et alors $\boldsymbol{\Lambda}^0 \approx \text{diag}\{\boldsymbol{\Lambda}_i^0\}$. L'optimalité de l'estimateur RGC découle alors du théorème 1(*b*).

Preuve du théorème 2

Nous partons de l'expression de l'estimateur RGC. En vertu du lemme 1, avec la matrice de plan partitionnée $(\boldsymbol{\mathcal{X}}, \mathbf{Z})$ et $\mathbf{R} = \boldsymbol{\Lambda}$, le vecteur de poids calés $\mathbf{c}$ peut être écrit sous la forme $\mathbf{c} = \mathbf{c}_\mathbf{Z} + \mathbf{L}_\mathbf{Z}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\mathbf{L}_\mathbf{Z}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}'\mathbf{c}_\mathbf{Z})$, où $\mathbf{c}_\mathbf{Z} = \mathbf{w} + \boldsymbol{\Lambda}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z})^{-1}(\mathbf{t}_{(\mathbf{z})} - \mathbf{Z}'\mathbf{w})$ et $\mathbf{L}_\mathbf{Z} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})$. Alors $\hat{\boldsymbol{\mathcal{X}}}_3^{\text{RG}} = \boldsymbol{\mathcal{X}}_3'\mathbf{c}_\mathbf{Z} = \hat{\boldsymbol{\mathcal{X}}}_3 + \boldsymbol{\mathcal{X}}_3'\boldsymbol{\Lambda}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z})^{-1}(\mathbf{t}_{(\mathbf{z})} - \hat{\mathbf{Z}})$ et $\hat{\boldsymbol{\mathcal{X}}}^{\text{RG}} = \hat{\boldsymbol{\mathcal{X}}} + \boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z})^{-1}(\mathbf{t}_{(\mathbf{z})} - \hat{\mathbf{Z}})$. Il s'ensuit que l'estimateur RGC est donné par $\boldsymbol{\mathcal{X}}_3'\mathbf{c} = \hat{\boldsymbol{\mathcal{X}}}_3^{\text{RG}} - \hat{\boldsymbol{\mathcal{B}}}\hat{\boldsymbol{\mathcal{X}}}^{\text{RG}}$, où $\hat{\boldsymbol{\mathcal{B}}} = [\boldsymbol{\mathcal{X}}_3'\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})\boldsymbol{\mathcal{X}}]$ $[\boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})\boldsymbol{\mathcal{X}}]^{-1}$.

a) Puisque $\mathbf{P}_\mathbf{Z} = \text{diag}\{\mathbf{P}_{\mathbf{Z}_i}\}$ et, pour l'EAS, $\boldsymbol{\Lambda}^0 = \text{diag}\{\lambda_i(\mathbf{I} - \mathbf{P}_{\mathbf{1}i})\}$, où $\lambda_i = N^2(1 - f_i)/[n_i(n_i - 1)]$ et $\mathbf{P}_{\mathbf{1}i} = \mathbf{1}_i(\mathbf{1}_i'\mathbf{1}_i)^{-1}\mathbf{1}_i'$, nous avons $\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}) = \text{diag}\{\lambda_i(\mathbf{I} - \mathbf{P}_{\mathbf{1}i})(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i})\}$. Or, par hypothèse $\mathbf{1} = \mathbf{Z}_i\mathbf{h}_i$, de sorte que $\mathbf{1}'\mathbf{P}_{\mathbf{Z}_i} = \mathbf{1}'$ et donc $\mathbf{P}_{\mathbf{1}i}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i}) = \mathbf{0}$. Par conséquent, $\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}) = \text{diag}\{\lambda_i(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i})\}$ et, puisque les matrices $\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i}$ sont idempotentes, $(\mathbf{I} - \mathbf{P}_\mathbf{Z})'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}) = \text{diag}\{\lambda_i(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i})\}$. Mais $\lambda_i = w_{ik}/q_{ik}$, où $w_{ik} = N/n_i$ et les q_{ik} sont les constantes spécifiées dans les entrées de $\boldsymbol{\Lambda}_i$. Il s'ensuit que $(\mathbf{I} - \mathbf{P}_\mathbf{Z})'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}) = \text{diag}\{\boldsymbol{\Lambda}_i(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i})\} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})$ et donc $\hat{\boldsymbol{\mathcal{B}}} = \hat{\boldsymbol{\mathcal{B}}}^{wo}$, de sorte que $\hat{\boldsymbol{\mathcal{X}}}_3^{\text{RG}} - \hat{\boldsymbol{\mathcal{B}}}\hat{\boldsymbol{\mathcal{X}}}^{\text{RG}} = \hat{\boldsymbol{\mathcal{X}}}_3^{\text{RG}} - \hat{\boldsymbol{\mathcal{B}}}^{wo}\hat{\boldsymbol{\mathcal{X}}}^{\text{RG}}$.

b) En vertu du lemme 1, avec la matrice de plan partitionnée $\boldsymbol{\mathcal{Z}} = (\boldsymbol{\mathcal{X}}, \mathbf{Z}, \mathbf{D})$ et le vecteur de totaux $\mathbf{t}_\mathcal{Z} = (\mathbf{0}', \mathbf{t}_{(\mathbf{z})}', \mathbf{N}')'$, le vecteur de poids calés $\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}})^{-1}(\mathbf{t}_\mathcal{Z} - \boldsymbol{\mathcal{Z}}'\mathbf{w})$ peut

s'écrire sous la forme $\mathbf{c} = \mathbf{c}_{\mathbf{D}} + \mathbf{L}_{\mathbf{D}}(\boldsymbol{\mathcal{X}}, \mathbf{Z})\left[(\boldsymbol{\mathcal{X}}, \mathbf{Z})' \mathbf{L}_{\mathbf{D}}(\boldsymbol{\mathcal{X}}, \mathbf{Z})\right]^{-1}\left[(\mathbf{0}', \mathbf{t}'_{(\mathbf{z})})' - (\boldsymbol{\mathcal{X}}, \mathbf{Z})' \mathbf{c}_{\mathbf{D}}\right]$, où $\mathbf{c}_{\mathbf{D}} = \mathbf{w} + \boldsymbol{\Lambda}\mathbf{D}(\mathbf{D}'\boldsymbol{\Lambda}\mathbf{D})^{-1}(\mathbf{N} - \mathbf{D}'\mathbf{w})$ et $\mathbf{L}_{\mathbf{D}} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_{\mathbf{D}})$, avec $\mathbf{P}_{\mathbf{D}} = \mathbf{D}(\mathbf{D}'\boldsymbol{\Lambda}\mathbf{D})^{-1}\mathbf{D}'\boldsymbol{\Lambda}$. Mais, comme il est montré dans la preuve du théorème 1 (a), $\mathbf{c}_{\mathbf{D}} = \mathbf{w}$ et $\mathbf{L}_{\mathbf{D}} = \boldsymbol{\Lambda}^0$. Donc, $\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}^0(\boldsymbol{\mathcal{X}}, \mathbf{Z})\left[(\boldsymbol{\mathcal{X}}, \mathbf{Z})' \boldsymbol{\Lambda}^0(\boldsymbol{\mathcal{X}}, \mathbf{Z})\right]^{-1}\left[(\mathbf{0}', \mathbf{t}'_{(\mathbf{z})})' - (\boldsymbol{\mathcal{X}}, \mathbf{Z})' \mathbf{w}\right]$. Ensuite, en appliquant de nouveau le lemme 1, maintenant avec $\mathbf{R} = \boldsymbol{\Lambda}^0$ et la matrice de plan $(\boldsymbol{\mathcal{X}}, \mathbf{Z})$, nous obtenons $\mathbf{c} = \mathbf{c}_{\mathbf{Z}} + \mathbf{L}^0_{\mathbf{Z}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\mathbf{L}^0_{\mathbf{Z}}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}'\mathbf{c}_{\mathbf{Z}})$, où $\mathbf{c}_{\mathbf{Z}} = \mathbf{w} + \boldsymbol{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z})^{-1}(\mathbf{t}_{(\mathbf{z})} - \mathbf{Z}'\mathbf{w})$ et $\mathbf{L}^0_{\mathbf{Z}} = \boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}^0_{\mathbf{Z}})$. Alors, il s'ensuit que l'estimateur RGC est $\boldsymbol{\mathcal{X}}'_3\mathbf{c} = \boldsymbol{\mathcal{X}}'_3\mathbf{c}_{\mathbf{Z}} - \boldsymbol{\mathcal{X}}'_3\mathbf{L}^0_{\mathbf{Z}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\mathbf{L}^0_{\mathbf{Z}}\boldsymbol{\mathcal{X}})^{-1}\boldsymbol{\mathcal{X}}'\mathbf{c}_{\mathbf{Z}} = \hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RO}} - \hat{\boldsymbol{\mathcal{B}}}^o\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RO}}$, en les expressions évidentes pour $\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RO}}, \hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RO}}$ et $\hat{\boldsymbol{\mathcal{B}}}^o$.

c) Il a été montré dans la preuve du théorème 1 que $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^0$. Clairement, il est alors vérifié que $\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RG}} = \hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RO}}$, $\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RG}} = \hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RO}}$ et $\hat{\boldsymbol{\mathcal{B}}} = \hat{\boldsymbol{\mathcal{B}}}^o$, et donc $\hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RG}} - \hat{\boldsymbol{\mathcal{B}}}\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RG}} = \hat{\boldsymbol{\mathcal{X}}}_3^{\mathrm{RO}} - \hat{\boldsymbol{\mathcal{B}}}^o\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{RO}}$.

Preuve de la proposition 1

Toutes les matrices qui apparaissent dans cette preuve sont définies au niveau de la population. Le partitionnement de la matrice $\boldsymbol{\mathcal{X}}$ donnée en (4.4) sous la forme $(\mathbf{Z}, \boldsymbol{\Psi})$, où $\mathbf{Z}$ est constituée des deuxième et quatrième colonnes, et $\boldsymbol{\Psi}$, du reste, et en appliquant le lemme 1 avec $\mathbf{R} = \boldsymbol{\Lambda}^0 = \{(\pi_{kl} - \pi_k\pi_l)/\pi_k\pi_l\}$, nous obtenons le vecteur de poids calés décomposé de la forme

$$\mathbf{c} = \mathbf{w} + \mathbf{L}^0_{\boldsymbol{\Psi}}\mathbf{Z}\left(\mathbf{Z}'\mathbf{L}^0_{\boldsymbol{\Psi}}\mathbf{Z}\right)^{-1}[\mathbf{0} - \mathbf{Z}'\mathbf{w}] + \mathbf{L}^0_{\mathbf{Z}}\boldsymbol{\Psi}\left(\boldsymbol{\Psi}'\mathbf{L}^0_{\mathbf{Z}}\boldsymbol{\Psi}\right)^{-1}[\mathbf{0} - \boldsymbol{\Psi}'\mathbf{w}],$$

où $\mathbf{L}^0_{\mathbf{Z}} = \boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}^0_{\mathbf{Z}})$ avec $\mathbf{P}^0_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Lambda}^0$. L'estimateur $\hat{\mathbf{Z}}^B$ donné en (4.2) s'obtient sous la forme $\mathbf{Z}'_{3-}\mathbf{c}$, où $\mathbf{Z}_{3-} = (\mathbf{0}', \mathbf{0}', \mathbf{Z}'_3)'$. Les deux derniers termes de (4.2) sont consolidés dans le terme $\mathbf{Z}'_{3-}\mathbf{L}^0_{\mathbf{Z}}\boldsymbol{\Psi}\left(\boldsymbol{\Psi}'\mathbf{L}^0_{\mathbf{Z}}\boldsymbol{\Psi}\right)^{-1}[\mathbf{0} - \boldsymbol{\Psi}'\mathbf{w}]$. Ces deux termes disparaissent uniquement si $\mathbf{Z}'_{3-}\mathbf{L}^0_{\mathbf{Z}}\boldsymbol{\Psi}(= \mathbf{Z}'_{3-}\boldsymbol{\Lambda}^0\boldsymbol{\Psi} - \mathbf{Z}'_{3-}\boldsymbol{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Lambda}^0\boldsymbol{\Psi}) = \mathbf{0}$. Premièrement, nous obtenons facilement $\mathbf{Z}'_{3-}\boldsymbol{\Lambda}^0\boldsymbol{\Psi} = \left(\mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{X}_3, \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Y}_3\right)$ et $\mathbf{Z}'_{3-}\boldsymbol{\Lambda}^0\mathbf{Z} = \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Z}_3\,(\mathbf{I}, \mathbf{I})$, ainsi que

$$\mathbf{Z}'\boldsymbol{\Lambda}^0\boldsymbol{\Psi} = \begin{pmatrix} \mathbf{Z}'_1\boldsymbol{\Lambda}^0_1\mathbf{X}_1 + \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{X}_3 & \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Y}_3 \\ \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{X}_3 & \mathbf{Z}'_2\boldsymbol{\Lambda}^0_2\mathbf{Y}_2 + \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Y}_3 \end{pmatrix},$$

et

$$\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z} = \begin{pmatrix} \mathbf{Z}'_1\boldsymbol{\Lambda}^0_1\mathbf{Z}_1 + \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Z}_3 & \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Z}_3 \\ \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Z}_3 & \mathbf{Z}'_2\boldsymbol{\Lambda}^0_2\mathbf{Z}_2 + \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Z}_3 \end{pmatrix}.$$

Ensuite, nous écrivons

$$(\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z})^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}' & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}' & \mathbf{E}^{-1} \end{pmatrix},$$

où $\mathbf{E} = \mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}$ et $\mathbf{F} = \mathbf{A}^{-1}\mathbf{B}$. Il s'ensuit alors que $\mathbf{Z}'_{3-}\boldsymbol{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z})^{-1} = (\mathbf{B}\mathbf{A}^{-1} + \mathbf{B}\mathbf{F}\mathbf{E}^{-1}\mathbf{F}' - \mathbf{B}\mathbf{E}^{-1}\mathbf{F}', \mathbf{B}(\mathbf{I} - \mathbf{F})\mathbf{E}^{-1}) = ((\mathbf{D} - \mathbf{B})\mathbf{E}^{-1}\mathbf{F}', \mathbf{B}(\mathbf{I} - \mathbf{F})\mathbf{E}^{-1})$. En utilisant les expressions analytiques $\mathbf{B} = \mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Z}_3$,

$\mathbf{D} = \mathbf{Z}_2'\boldsymbol{\Lambda}_2^0\mathbf{Z}_2 + \mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Z}_3$, $\mathbf{F} = \left(\mathbf{Z}_1'\boldsymbol{\Lambda}_1^0\mathbf{Z}_1 + \mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Z}_3\right)^{-1}\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Z}_3$ et $\mathbf{E} = \mathbf{Z}_2'\boldsymbol{\Lambda}_2^0\mathbf{Z}_2 + \mathbf{Z}_1'\boldsymbol{\Lambda}_1^0\mathbf{Z}_1\mathbf{F}$, nous obtenons après certaines opérations algébriques

$$\mathbf{Z}_{3-}'\boldsymbol{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z})^{-1} = \mathbf{K}^{-1}\left[\left(\mathbf{Z}_1'\boldsymbol{\Lambda}_1^0\mathbf{Z}_1\right)^{-1}, \left(\mathbf{Z}_2'\boldsymbol{\Lambda}_2^0\mathbf{Z}_2\right)^{-1}\right],$$

où $\mathbf{K} = \left(\mathbf{Z}_1'\boldsymbol{\Lambda}_1^0\mathbf{Z}_1\right)^{-1} + \left(\mathbf{Z}_2'\boldsymbol{\Lambda}_2^0\mathbf{Z}_2\right)^{-1} + \left(\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Z}_3\right)^{-1}$. Nous pouvons obtenir sans trop de difficulté

$$\begin{aligned}\mathbf{Z}_{3-}'\mathbf{L}_{\mathbf{Z}}^0\boldsymbol{\Psi} &= \mathbf{Z}_{3-}'\boldsymbol{\Lambda}^0\boldsymbol{\Psi} - \mathbf{Z}_{3-}'\boldsymbol{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Lambda}^0\boldsymbol{\Psi} \\ &= \mathbf{K}^{-1}\left[\left(\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Z}_3\right)^{-1}\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{X}_3 - \left(\mathbf{Z}_1'\boldsymbol{\Lambda}_1^0\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1'\boldsymbol{\Lambda}_1^0\mathbf{X}_1,\right. \\ &\quad \left.\left(\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Z}_3\right)^{-1}\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Y}_3 - \left(\mathbf{Z}_2'\boldsymbol{\Lambda}_2^0\mathbf{Z}_2\right)^{-1}\mathbf{Z}_2'\boldsymbol{\Lambda}_2^0\mathbf{Y}_2\right].\end{aligned}$$

Il s'ensuit que $\mathbf{Z}_{3-}'\mathbf{L}_{\mathbf{Z}}^0\boldsymbol{\Psi} = (\mathbf{0},\mathbf{0})$ uniquement si $\left(\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Z}_3\right)^{-1}\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{X}_3 = \left(\mathbf{Z}_1'\boldsymbol{\Lambda}_1^0\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1'\boldsymbol{\Lambda}_1^0\mathbf{X}_1$ et $\left(\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Z}_3\right)^{-1}\mathbf{Z}_3'\boldsymbol{\Lambda}_3^0\mathbf{Y}_3 = \left(\mathbf{Z}_2'\boldsymbol{\Lambda}_2^0\mathbf{Z}_2\right)^{-1}\mathbf{Z}_2'\boldsymbol{\Lambda}_2^0\mathbf{Y}_2$. Mais ces deux équations sont identiques aux équations données en (4.6). Puisque dans $\left(\mathbf{Z}_i'\boldsymbol{\Lambda}_i^0\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i'\boldsymbol{\Lambda}_i^0\mathbf{X}_i$ toutes les matrices sont définies au niveau de la population, avec l'indice inférieur $i = 1,3$ indiquant l'enquête, cette quantité n'est constante pour les diverses enquêtes que si la matrice particulière au plan $\boldsymbol{\Lambda}_i^0$ est constante, ou que $\boldsymbol{\Lambda}_i^0$ diffère d'une enquête à l'autre d'un multiple constant (dépendant de la taille de l'échantillon). Cela demeure également vrai pour $\left(\mathbf{Z}_i'\boldsymbol{\Lambda}_i^0\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i'\boldsymbol{\Lambda}_i^0\mathbf{Y}_i$, $i = 2,3$, ce qui achève la preuve.

Preuve de la proposition 2

Sous le scénario d'échantillonnage (a) du théorème 1, le calage composite au niveau de la population avec la matrice de plan $\boldsymbol{\mathcal{Z}} = (\boldsymbol{\mathcal{X}}, \mathbf{D})$ et le vecteur de totaux $\mathbf{t}_{\boldsymbol{\mathcal{Z}}} = (\mathbf{0}', \mathbf{N}')'$ produit l'estimateur de domaine RGC conjoint de $(\mathbf{t}_{\mathbf{x}d}', \mathbf{t}_{\mathbf{y}d}')'$ fondé sur les poids de S_3 et s'écrit sous la forme $\hat{\boldsymbol{\mathcal{X}}}_{3d}^{\mathrm{RGC}} = \hat{\boldsymbol{\mathcal{X}}}_{3d} + \hat{\boldsymbol{\mathcal{B}}}_d\left(\mathbf{t}_{\boldsymbol{\mathcal{Z}}} - \hat{\boldsymbol{\mathcal{Z}}}\right)$, où $\hat{\boldsymbol{\mathcal{B}}}_d = \boldsymbol{\mathcal{X}}_{3d}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}})^{-1}$. La matrice associée des résidus de régression est $\boldsymbol{\mathcal{X}}_{3d} - \boldsymbol{\mathcal{Z}}\hat{\boldsymbol{\mathcal{B}}}_d'$, qui peut aussi s'écrire $(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})\boldsymbol{\mathcal{X}}_{3d}$, avec $\mathbf{P}_{\boldsymbol{\mathcal{Z}}} = \boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}})^{-1}\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}$. Alors, $\widehat{\mathrm{AV}}\left(\hat{\boldsymbol{\mathcal{X}}}_{3d}^{\mathrm{RGC}}\right) = \boldsymbol{\mathcal{X}}_{3d}'(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})\boldsymbol{\mathcal{X}}_{3d}$. Ensuite, rappelons que, d'après la preuve du théorème 1, $\boldsymbol{\Lambda}^0 = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_{\mathbf{D}})$, avec $\mathbf{P}_{\mathbf{D}} = \mathbf{D}(\mathbf{D}'\boldsymbol{\Lambda}\mathbf{D})^{-1}\mathbf{D}'\boldsymbol{\Lambda}$, et notons que $\mathbf{D} = \boldsymbol{\mathcal{Z}}\mathbf{H}$ pour une matrice constante appropriée $\mathbf{H}$. Il est facile de montrer que $\mathbf{P}_{\mathbf{D}}\mathbf{P}_{\boldsymbol{\mathcal{Z}}} = \mathbf{P}_{\mathbf{D}}$. Il s'ensuit alors que $\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}}) = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})$ et $(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}}) = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})$. Donc, $\widehat{\mathrm{AV}}\left(\hat{\boldsymbol{\mathcal{X}}}_{3d}^{\mathrm{RGC}}\right) = \boldsymbol{\mathcal{X}}_{3d}'\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})\boldsymbol{\mathcal{X}}_{3d}$. Or, le calage composite au niveau du domaine fait intervenir la matrice de plan $\boldsymbol{\mathcal{Z}}_d = (\boldsymbol{\mathcal{X}}_d, \mathbf{D})$; il n'est pas nécessaire de restreindre $\mathbf{D}$ au domaine U_d. L'estimateur RGC résultant est $\check{\boldsymbol{\mathcal{X}}}_{3d}^{\mathrm{RGC}} = \hat{\boldsymbol{\mathcal{X}}}_{3d} + \check{\boldsymbol{\mathcal{B}}}_d\left(\mathbf{t}_{\boldsymbol{\mathcal{Z}}_d} - \hat{\boldsymbol{\mathcal{Z}}}_d\right)$, où $\check{\boldsymbol{\mathcal{B}}}_d = \boldsymbol{\mathcal{X}}_{3d}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}}_d(\boldsymbol{\mathcal{Z}}_d'\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}}_d)^{-1}$. Comme pour l'estimateur $\hat{\boldsymbol{\mathcal{X}}}_{3d}^{\mathrm{RGC}}$ susmentionné, on peut montrer que $\widehat{\mathrm{AV}}\left(\check{\boldsymbol{\mathcal{X}}}_{3d}^{\mathrm{RGC}}\right) = \boldsymbol{\mathcal{X}}_{3d}'\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}_d})\boldsymbol{\mathcal{X}}_{3d}$, où $\mathbf{P}_{\boldsymbol{\mathcal{Z}}_d} = \boldsymbol{\mathcal{Z}}_d(\boldsymbol{\mathcal{Z}}_d'\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}})_d^{-1}\boldsymbol{\mathcal{Z}}_d'\boldsymbol{\Lambda}$. Alors $\widehat{\mathrm{AV}}\left(\hat{\boldsymbol{\mathcal{X}}}_{3d}^{\mathrm{RGC}}\right) - \widehat{\mathrm{AV}}\left(\check{\boldsymbol{\mathcal{X}}}_{3d}^{\mathrm{RGC}}\right) =$

$\mathcal{X}'_{3d}\boldsymbol{\Lambda}\left(\mathbf{P}_{\mathcal{Z}_d}-\mathbf{P}_{\mathcal{Z}}\right)\mathcal{X}_{3d}$. En notant que $\mathcal{X}'_{3d}\boldsymbol{\Lambda}\mathcal{Z}=\mathcal{X}'_{3d}\boldsymbol{\Lambda}\mathcal{Z}_d$, nous pouvons écrire $\mathbf{P}_{\mathcal{Z}}=\mathcal{Z}_d\left(\mathcal{Z}'\boldsymbol{\Lambda}\mathcal{Z}\right)^{-1}\mathcal{Z}'_d\boldsymbol{\Lambda}$. Il est alors trivial de montrer que $\left(\mathbf{P}_{\mathcal{Z}_d}-\mathbf{P}_{\mathcal{Z}}\right)=\left(\mathbf{P}_{\mathcal{Z}_d}-\mathbf{P}_{\mathcal{Z}}\right)^2$, et puisque la matrice $\boldsymbol{\Lambda}$ est diagonale avec entrées positives, il s'ensuit que $\mathcal{X}'_{3d}\boldsymbol{\Lambda}\left(\mathbf{P}_{\mathcal{Z}_d}-\mathbf{P}_{\mathcal{Z}}\right)\mathcal{X}_{3d}>\mathbf{0}$ et donc $\widehat{\mathrm{AV}}\left(\breve{\mathcal{X}}_{3d}^{\mathrm{RGC}}\right)<\widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}_{3d}^{\mathrm{RGC}}\right)$.

Sous les conditions de la partie (b), $\boldsymbol{\Lambda}=\boldsymbol{\Lambda}^0$ et l'estimateur de domaine RGC est identique à l'estimateur de domaine ROC $\hat{\mathcal{X}}_{3d}^{\mathrm{ROC}}=\hat{\mathcal{X}}_{3d}-\hat{\mathcal{B}}_d^0\hat{\mathcal{X}}$, où $\hat{\mathcal{B}}_d^0=\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}\left(\mathcal{X}'\boldsymbol{\Lambda}^0\mathcal{X}\right)^{-1}$. La matrice associée aux résidus de régression est $\left(\mathbf{I}-\mathbf{P}_{\mathcal{X}}\right)\mathcal{X}_{3d}$, avec $\mathbf{P}_{\mathcal{X}}=\mathcal{X}\left(\mathcal{X}'\boldsymbol{\Lambda}^0\mathcal{X}\right)^{-1}\mathcal{X}'\boldsymbol{\Lambda}^0$. Alors, $\widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}_{3d}^{\mathrm{ROC}}\right)=\mathcal{X}'_{3d}\left(\mathbf{I}-\mathbf{P}_{\mathcal{X}}\right)'\boldsymbol{\Lambda}^0\left(\mathbf{I}-\mathbf{P}_{\mathcal{X}}\right)\mathcal{X}_{3d}=\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\left(\mathbf{I}-\mathbf{P}_{\mathcal{X}}\right)\mathcal{X}_{3d}$. Par ailleurs, pour l'estimateur $\breve{\mathcal{X}}_{3d}^{\mathrm{ROC}}=\breve{\mathcal{X}}_{3d}-\breve{\mathcal{B}}_d^0\hat{\mathcal{X}}$, où $\hat{\mathcal{B}}_d^0=\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}_d\left(\mathcal{X}'_d\boldsymbol{\Lambda}^0\mathcal{X}_d\right)^{-1}$ nous avons $\widehat{\mathrm{AV}}(\breve{\mathcal{X}}_{3d}^{\mathrm{ROC}})=\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0(\mathbf{I}-\mathbf{P}_{\mathcal{X}_d})\mathcal{X}_{3d}$, avec $\mathbf{P}_{\mathcal{X}_d}=\mathcal{X}_d\left(\mathcal{X}'_d\boldsymbol{\Lambda}^0\mathcal{X}_d\right)^{-1}\mathcal{X}'_d\boldsymbol{\Lambda}^0$. Alors, $\widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}_{3d}^{\mathrm{ROC}}\right)-\widehat{\mathrm{AV}}\left(\breve{\mathcal{X}}_{3d}^{\mathrm{ROC}}\right)=\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\left(\mathbf{P}_{\mathcal{X}_d}-\mathbf{P}_{\mathcal{X}}\right)\mathcal{X}_{3d}$. Notons que $\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}_d=\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}_{3d}$ et, puisque $\boldsymbol{\Lambda}^0$ est diagonale, $\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}=\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}_{3d}$. Il s'ensuit que $\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\left(\mathbf{P}_{\mathcal{X}_d}-\mathbf{P}_{\mathcal{X}}\right)\mathcal{X}_{3d}=\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\left(\mathbf{P}_{\mathcal{X}_d}-\mathbf{P}_{\mathcal{X}}\right)^2\mathcal{X}_{3d}$ et donc $\widehat{\mathrm{AV}}\left(\breve{\mathcal{X}}_{3d}^{\mathrm{ROC}}\right)<\widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}_{3d}^{\mathrm{ROC}}\right)$.

Pour les parties (a') et (b'), la preuve est la même qu'en (a) et (b), compte tenu de la preuve du théorème 1.

Bibliographie

Andersson, P.G., et Thorburn, D. (2005). Une distance de calage optimale menant à un estimateur par la régression optimal. *Techniques d'enquête*, 31, 1, 103-107.

Australian Bureau of Statistics (2011). Household Expenditure Survey and Survey of Income and Housing, Guide d'utilisateur, Australie, 2009-10 (numéro du cat. 6503.0).

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

Chipperfield, J.O., et Steel, D.G. (2011). Efficiency of split questionnaire surveys. *Journal of Statistical Planning and Inference*, 141, 1925-1932.

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

Fuller, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 2, 177-190.

Gonzalez, J.M., et Eltinge, J.L. (2007). Multiple matrix sampling: A review. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3069-3075.

Gonzalez, J.M., et Eltinge, J.L. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3069-3075.

Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 2, 157-169.

Houbiers, M. (2004). Towards a social statistical database on unified estimates at Statistics Netherlands. *Journal of Official Statistics,* 20, 55-75.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Serie B,* 42, 221-226.

Kim, J.K., et Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika,* 99, 1, 85-100.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association,* 99, 1131-1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for more efficient small domain estimation. *Journal of the Royal Statistical Society, Serie B,* 72, 27-48.

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.

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

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics,* 10, 153-165.

Renssen, R.H. (1998). Utilisation de méthodes d'appariement statistique dans l'estimation de calage. *Techniques d'enquête,* 24, 2, 185-199.

Renssen, R.H., et Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association,* 92, 368-375.

Rust, K.F., et Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research,* 5, 283-310.

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

Smith, P. (2009). Survey harmonization in official household surveys in the United Kingdom. *Proceedings of the ISI World Statistical Congresses,* Dublin.

Thomas, N., Raghunathan, T.E., Schenker, N., Katzoff, M.J. et Johnson, C.L. (2006). Une évaluation des méthodes d'échantillonnage matriciel à l'aide de données provenant de la « National Health and Nutrition Examination Survey ». *Techniques d'enquête,* 32, 2, 241-257.

Wolter, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association,* 74, 604-613.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics,* 32, 15-26.

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 30, No. 4, 2014

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 31, No. 1, 2015

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 43, No. 1, March/mars 2015

CONTENTS TABLE DES MATIÈRES

Volume 43, No. 2, June/juin 2015

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 (rte@statcan.gc.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(·) et log(·), 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.