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

Techniques d'enquête

Méthode de correction de l'erreur d'appartenance à une base dans les estimateurs à double base de sondage

par Dong Lin, Zhaoce Liu et Lynne Stokes

Date de diffusion : le 17 décembre 2019

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

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

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2019

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

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2019
Vol. 45, nº 3, pp. 581-605
Statistique Canada, nº 12-001-X au catalogue

Méthode de correction de l'erreur d'appartenance à une base dans les estimateurs à double base de sondage

Dong Lin, Zhaoce Liu et Lynne Stokes[1]

Résumé

Les enquêtes à double base de sondage sont utiles là où une seule base n'assure pas une couverture suffisante, mais pour les estimateurs d'un tel plan d'échantillonnage à double base, il faut connaître l'appartenance à une base pour chaque unité échantillonnée. Si cette indication ne peut être tirée de la base de sondage même, elle est souvent à recueillir auprès de l'enquêté. Si l'enquêté donne alors une indication erronée, les estimations résultantes des moyennes ou des totaux pourront s'en trouver biaisées. Nous proposons une méthode de réduction de ce biais grâce à des données exactes d'appartenance venant d'un sous-échantillon d'enquêtés. Nous examinons les propriétés de notre nouvel estimateur et le comparons à d'autres. Nous l'appliquons aux données de l'exemple qui a été à l'origine de notre étude, soit une enquête auprès des pêcheurs à la ligne pêchant à des fins récréatives qui reposait sur une base d'adresses et une base incomplète de permis de pêche.

Mots-clés : Estimateur de Hartley; correction de biais; défaut de classification.

1 Introduction

Dans les enquêtes à base de sondage unique, toutes les unités de l'échantillon sont tirées de la même base. C'est là la méthode à privilégier lorsque la base couvre la population d'intérêt. Il reste que, pour certaines applications, il n'existe pas de base de sondage unique assurant une couverture suffisante ou il peut s'avérer trop coûteux d'échantillonner à partir d'une base complète englobant des unités hors population cible. Un plan d'échantillonnage à bases multiples peut constituer une solution de rechange intéressante en pareil cas.

C'est Hartley (1962) qui a introduit le plan d'échantillonnage à base double, c'est-à-dire comportant deux bases de sondage en chevauchement. Ces bases que nous appellerons A et B couvrent ensemble la population d'intérêt U se composant de trois sous-ensembles mutuellement exclusifs et portant le nom de domaines. Le domaine a contient les unités de la base A, mais non de la base B; le domaine b contient les unités de la base B, mais non de la base A; le domaine ab contient les unités à l'intersection de A et de B. Dans les enquêtes à double base, on prélève indépendamment des échantillons sur les bases A et B. D'ordinaire, l'échantillonneur ne sait pas d'avance à quel domaine appartient telle ou telle unité, mais le vérifiera pour les unités échantillonnées dans le cadre du processus de collecte des données.

Nous disposons de plusieurs estimateurs pour combiner les renseignements des deux échantillons afin d'estimer les paramètres de U. Dans tous les cas, il faut repondérer les unités échantillonnées en fonction de l'appartenance à un domaine. Ainsi, connaître le domaine des unités échantillonnées est essentiel à l'estimation dans les plans d'échantillonnage à base double. Si nous pouvons établir fidèlement

1. Dong Lin est gestionnaire à la PenFed Credit Union, à Mclean en Virginie. Courriel : schewiz1984@msn.com. Zhaoce Liu est titulaire d'une Thèse de doctorat et postulant en statistique à l'Université Southern Methodist, à Dallas au Texas. Courriel : zhaocel@smu.edu. Lynne Stokes est professeure et présidente du Département des sciences statistiques à cette même université. Courriel : slstokes@mail.smu.edu.

l'appartenance à un domaine, nous pouvons alors élaborer des estimateurs sans biais des moyennes et des totaux. En revanche, si les estimateurs sont fondés sur des renseignements inexacts d'appartenance à un domaine, ils risquent d'être entachés d'un biais. Hartley (1962) a conçu une famille d'estimateurs pour plan à base double. Dans le présent article, nous proposons une méthode de correction de biais pour l'estimateur de Hartley lorsque les probabilités de défaut de classification sont connues ou, dans une perspective plus réaliste, lorsqu'elles peuvent être estimées à partir d'un sous-échantillon aléatoire d'enquêtés. Nous comparons cet estimateur à celui que propose Lohr (2011) dans la même optique et nous clarifions les différentes hypothèses sous-jacentes.

L'application à l'origine de cette étude était le Marine Recreational Information Program (MRIP), système de collecte de données mis en œuvre par la National Oceanic and Atmospheric Administration (NOAA) pour l'obtention d'estimations bimestrielles sur les prises de poissons selon l'espèce en eau de mer aux États-Unis. Un volet de ce programme était la Coastal Household Telephone Survey (CHTS), qui estimait le nombre de déplacements de pêche des personnes pêchant à la ligne à des fins récréatives dans chaque État côtier et dans chaque période considérée. Un rapport du National Research Council (NRC, 2006) a permis de constater les lacunes du plan de sondage initial et des améliorations ont été expérimentées dans des enquêtes pilotes. Plusieurs de ces projets pilotes comportaient un plan d'échantillonnage à base double où une base de sondage résidentielle (base de numéros de téléphone ou d'adresses pour l'ensemble des ménages d'un État) était complétée par le registre des permis de pêche à la ligne de l'État en question. La base des numéros de téléphone est peu efficace, puisqu'on a pu joindre un pêcheur à la ligne dans seulement 5 % environ des prises de contact dans l'échantillon du CHTS en région urbaine (NRC, 2006). Quant à la base des permis, elle est efficace mais incomplète, négligeant les pêcheurs relevant de diverses exemptions de permis de pêche ou s'adonnant illicitement à cette activité. Dans ce cas, on demandait à l'enquêté des données d'appartenance à un domaine en voulant savoir si chaque pêcheur de la base résidentielle était titulaire d'un permis. Andrews, Brick, Mathiowetz et Stokes (2010) ont démontré que les données d'appartenance obtenues de la sorte étaient peu sûres, car les répondants surdéclaraient ou sous-déclaraient la détention d'un permis. Le but de la présente recherche est d'établir l'effet de cette erreur sur l'estimation du nombre de déplacements et de trouver une méthode de correction.

L'étude pilote à laquelle nous avons recouru à cette fin a eu lieu par la poste. Les deux bases visées étaient respectivement constituées des adresses de tous les enquêtés d'un État, lesquelles avaient été fournies par le US Postal Service, et des permis de tous les pêcheurs à la ligne autorisés par cet État, ces dernières indications émanant de l'autorité étatique en matière de ressources naturelles. La base des adresses ne disait évidemment pas si les membres des ménages enquêtés étaient titulaires d'un permis; une question a été posée en conséquence à tous les enquêtés de l'échantillon de la base des adresses. La base des permis contenait, bien sûr, les adresses, d'où la possibilité d'établir si les pêcheurs échantillonnés dans cette base résidaient ou non dans l'État. Comme nous disposions déjà de données sur les adresses, il était relativement

facile de coupler les ménages de la base des adresses et les adresses de la base des permis. Il a toutefois été ardu de déterminer si les enquêtés de la base des adresses étaient les membres des ménages titulaires d'un permis, car aucun nom n'était demandé. Il était donc difficile de voir si un ménage se trouvait dans les deux bases ou non.

À la section 2, nous examinons l'estimation dans les plans d'échantillonnage à base double. À la section 3, nous citons des exemples pour montrer en quoi l'erreur de rattachement à un domaine influe sur l'estimation. À la section 4, nous dégageons une méthode de correction de biais pour l'estimateur à base double de Hartley et la comparons à une autre méthode (Lohr, 2011). À la section 5, nous présentons les résultats d'une simulation visant à juger de la validité de l'inférence avec des estimateurs en correction de biais. Nous illustrons enfin notre méthode à la section 6 dans une application aux données d'une des enquêtes pilotes sur les pêcheurs à la ligne de la NOAA. Une analyse suit à la section 7.

2 Estimation à base de sondage double

Depuis l'introduction des plans d'échantillonnage à base double par Hartley (1962), un grand nombre d'estimateurs ont été proposés (Fuller et Burmeister, 1972; Kalton et Anderson, 1986; Bankier, 1986; Skinner et Rao, 1996). Dans cette section, nous nous attachons à l'estimateur de Hartley, puisqu'il a été utilisé dans notre application d'étude pilote dans le cadre du MRIP. Avec la notation originale de Hartley, nous désignons par N, N_A, N_B, N_a, N_b et N_{ab} le nombre d'éléments dans U, A, B, a, b et ab respectivement. Les relations suivantes se vérifient alors : $N = N_a + N_{ab} + N_b$, $N_A = N_a + N_{ab}$ et $N_B = N_{ab} + N_b$. Nous désignons les échantillons des bases A et B par s_A et s_B et les probabilités respectives d'inclusion de l'unité i dans les deux échantillons par π_i^A et π_i^B.

Le total de population Y peut s'écrire comme la somme des totaux des trois domaines s'excluant les uns les autres :

$$Y = Y_a + Y_{ab} + Y_b, \tag{2.1}$$

où $Y_a = \sum_{i\in a} y_i$, $Y_b = \sum_{i\in b} y_i$ et $Y_{ab} = \sum_{i\in ab} y_i$. Les estimateurs du total peuvent s'écrire comme la somme des estimateurs des totaux des trois domaines :

$$\hat{Y} = \hat{Y}_a + \hat{Y}_{ab} + \hat{Y}_b. \tag{2.2}$$

L'estimateur de Hartley (1962) est

$$\hat{Y}_H = \hat{Y}_a + \theta\hat{Y}_{ab}^A + (1-\theta)\,\hat{Y}_{ab}^B + \hat{Y}_b, \tag{2.3}$$

où $\hat{Y}_{ab}^A = \sum_{i\in s_{ab}^A} w_i^A\, y_i$ désigne l'estimateur de Y_{ab} par l'information de la base A et où $\hat{Y}_{ab}^B = \sum_{i\in s_{ab}^B} w_i^B\, y_i$ est l'estimateur correspondant par l'information de la base B. s_{ab}^A (s_{ab}^B) est le sous-ensemble de s_A (s_B)

comprenant les éléments du domaine ab; $w_i^A = 1/\pi_i^A$ et $w_i^B = 1/\pi_i^B$ sont les poids d'échantillonnage selon les probabilités d'inclusion dans les bases A et B. Dans la pratique, il y a normalement repondération en fonction tant de la non-réponse que de la sous-couverture éventuelle. De même, nous estimons Y_a et Y_b comme $\hat{Y}_a = \sum_{i \in s_a} w_i^A \, y_i$ et $\hat{Y}_b = \sum_{i \in s_b} w_i^B \, y_i$. θ est un nombre entre 0 et 1 qui repondère les éléments des bases A et B. Si θ désigne une constante, la théorie prévoit une valeur optimale qui minimisera la variance de l'estimateur. Toutefois, θ dépendra de paramètres inconnus et devra donc faire l'objet d'une estimation. Un autre moyen à employer consiste à choisir une valeur de θ qui est proportionnelle à la réciproque des tailles d'échantillon en provenance de l'une et l'autre des bases. Ce sera là une constante et elle sera quasi optimale si les plans de sondage sont d'un effet semblable et que le domaine en chevauchement est petit.

Si θ est une constante, $\hat{Y}_H$ est linéaire dans les données et ses propriétés sont faciles à calculer. Si θ égale 0 ou 1, l'estimateur du domaine en chevauchement dépend des données d'une seule des deux bases. La valeur optimale de θ qui minimise la variance de $\hat{Y}_H$ (Hartley, 1962) est

$$\theta_H = \frac{V\left(\hat{Y}_{ab}^B\right) + \mathrm{Cov}\left(\hat{Y}_b^B, \hat{Y}_{ab}^B\right) - \mathrm{Cov}\left(\hat{Y}_a^A, \hat{Y}_{ab}^A\right)}{V\left(\hat{Y}_{ab}^A\right) + V\left(\hat{Y}_{ab}^B\right)}. \tag{2.4}$$

Les variances et les covariances en (2.4) sont inconnues et doivent être estimées à partir de l'échantillon si la forme optimale de θ doit être employée en (2.3). Dans ce cas, les poids résultants sont aléatoires et contribuent à la variabilité de l'estimateur. Un autre inconvénient avec la valeur optimale de θ dans $\hat{Y}_H$ est que θ_H diffère selon les variables de réponse, d'où une incohérence parmi les estimateurs de quantités apparentées. Ainsi, les estimateurs optimaux du nombre de déplacements de pêche à la ligne depuis la rive ou en bateau ne donnent pas nécessairement par addition l'estimateur optimal du nombre total de déplacements de pêche. C'est pourquoi une valeur constante de θ est fréquemment employée dans la pratique. Ainsi, une valeur $\theta = 1/2$ est quelquefois recommandée lorsque les tailles d'échantillon coïncident dans les deux bases (Lohr, 2011).

3 Défaut de classification dans les enquêtes à base de sondage double

Nous devons connaître le domaine de chaque unité échantillonnée pour calculer tout estimateur de plan d'échantillonnage à base double. Ainsi, (2.3) montre que, dans $\hat{Y}_H$, nous nous devons de repondérer une unité de la base résidentielle par θ ou $1 - \theta$ si l'enquêté est titulaire d'un permis, mais non dans les autres cas. Si cette indication ne peut être tirée des bases mêmes, elle doit être recueillie auprès de l'enquêté. Dans notre application, nous avons constaté qu'un certain nombre d'enquêtés ne pouvaient indiquer fidèlement s'ils étaient titulaires d'un permis de pêche ou non. Une fausse indication d'appartenance à un domaine crée un défaut de classification de domaine, lequel influe sur les propriétés des estimateurs des moyennes et des

totaux. Dans cette section, nous examinons l'effet de l'erreur de rattachement à un domaine sur le biais et la variance de l'estimateur de Hartley.

Dans le cas de la base A, l'erreur de rattachement peut se produire de deux façons. D'abord, l'enquêté échantillonné dans la base A et qui est dans le domaine ab peut se dire dans le domaine a. L'erreur peut se produire dans les deux sens si l'enquêté de la base A dans le domaine a se dit dans le domaine ab. Le domaine qu'indique l'enquêté est le domaine perçu ou déclaré qui est à distinguer du domaine réel. Ainsi, chaque enquêté appartient à un sous-groupe sur quatre selon son appartenance à l'intersection entre domaine réel et domaine perçu. Nous employons ici l'astérisque $(*)$ en exposant sur le nom de domaine pour désigner le domaine perçu et, dans ce cas, les quatre sous-groupes deviennent $a \cap a^*$, $ab \cap ab^*$, $a \cap ab^*$ et $ab \cap a^*$. De telles étiquettes figurent aussi en indice pour les paramètres, indiquant le sous-ensemble applicable dans la population. Soit A et B, par exemple, qui désignent effectivement les bases CHTS et registre des pêcheurs à la ligne. Le membre d'un ménage de la base CHTS qui n'est pas titulaire d'un permis, mais dit en détenir un, appartient au sous-groupe $a \cap ab^*$. L'univers de tous ces gens dans la base A serait désigné par $U_{a \cap ab^*}$; la taille du sous-groupe dans la population et le total et la moyenne des déplacements de pêche de ce sous-groupe seraient respectivement désignés par $N_{a \cap ab^*}$, $Y_{a \cap ab^*}$ et $\overline{Y}_{a \cap ab^*}$. Si quelqu'un est réellement titulaire d'un permis, il appartiendrait au sous-groupe $ab \cap ab^*$. Dans notre étude, l'analyste ne pourrait distinguer les domaines des deux intéressés et les placerait tous deux dans le domaine perçu en chevauchement ab^*. Le poids d'échantillonnage convenant au domaine réel mais inobservé en chevauchement (ab) s'appliquerait aux unités appartenant à $U_{ab^*} = U_{a \cap ab^*} \cup U_{ab \cap ab^*}$ au lieu de $U_{ab} = U_{ab \cap a^*} \cup U_{ab \cap ab^*}$ et les totaux estimés des unités de U_{ab^*} remplaceraient ceux d'U_{ab} en (2.2), causant le biais que nous nous proposons d'étudier ici.

L'erreur de rattachement à un domaine que nous venons de décrire pour la base A est possible pour l'une ou l'autre base dans un certain nombre d'études. Dans la nôtre, le problème ne se posait pas, parce que la base des permis comprenait les adresses. Nous savions donc si le pêcheur à la ligne inscrit à l'autorisation appartiendrait ou non à la base résidentielle de l'État. Nous introduisons cependant le cas plus général et, par conséquent, la notation définie pour la base A vaut aussi pour la base B. Une complication dans ce cas est que nous devons distinguer les cas pour le domaine perçu en chevauchement (ab^*) en fonction de la base d'appartenance des unités, et ce, parce que les unités du domaine perçu en chevauchement peuvent différer selon la base d'origine. Dans notre étude par exemple, une personne échantillonnée dans la base CHTS peut déclarer ne pas être titulaire d'un permis alors qu'elle en détient un et ne se trouverait donc pas dans ab^*. Si cette même personne était échantillonnée dans la base des permis, elle serait considérée à juste titre comme étant dans le domaine en chevauchement ab^*. Comme la confusion est possible, nous étendons la notation du domaine perçu en chevauchement pour indiquer la base d'origine de l'unité sous les formes $ab^*(A)$ et $ab^*(B)$. Dans notre étude par exemple, $\overline{Y}_{ab^*(A)}$ est le nombre moyen de déplacements pour toutes les unités de l'État (base A) qui déclareraient être titulaires d'un permis si on le leur demandait, alors qu'elles n'en détiendraient pas. Toutefois, $\overline{Y}_{ab^*(B)}$ serait la moyenne des

déplacements pour les seules unités appartenant au domaine réel ab, notre base B n'étant pas entachée d'erreurs de rattachement à un domaine.

Prenons pour $\hat{Y}_H$ une nouvelle notation qui facilitera le calcul de ses propriétés en cas de défaut de classification. Considérons d'abord le cas où aucune erreur de rattachement ne se produit. Nous définissons alors l'indicateur de domaine *réel* $\delta_i^A(a)$ avec la valeur 1 lorsque l'unité i de la base A est dans le domaine a et avec 0 dans les autres cas. Comme chaque unité se trouve dans un seul domaine réel, l'indicateur de présence de l'unité i dans le domaine ab est $1-\delta_i^A(a)$. Nous définissons de la même manière $\delta_i^B(b)$ pour la base B. S'il n'y a pas d'erreur de rattachement, $\hat{Y}_H$ peut s'écrire ainsi :

$$\begin{aligned}\hat{Y}_H &= \sum_{i=1}^{N_A} I_i^A\, w_i^A\, \delta_i^A(a)\, y_i + \theta \sum_{i=1}^{N_A} I_i^A\, w_i^A \left(1-\delta_i^A(a)\right) y_i \\ &\quad + \sum_{i=1}^{N_B} I_i^B\, w_i^B\, \delta_i^B(b)\, y_i + (1-\theta) \sum_{i=1}^{N_B} I_i^B\, w_i^B \left(1-\delta_i^B(b)\right) y_i \\ &= \sum_{i=1}^{N_A} I_i^A\, w_i^A\, x_{Ai} + \sum_{i=1}^{N_B} I_i^B\, w_i^B\, x_{Bi}, \end{aligned} \tag{3.1}$$

où I_i^A et I_i^B sont des indicateurs d'appartenance de l'unité i de la base A ou B à l'échantillon s_A ou s_B, soit $x_{Ai} = \delta_i^A(a)\, y_i + \theta\left(1-\delta_i^A(a)\right) y_i$ et $x_{Bi} = \delta_i^B(b)\, y_i + (1-\theta)\left(1-\delta_i^B(b)\right) y_i$.

Si nous comparons les effets du défaut de classification de domaine sur les propriétés de l'estimateur, nous nous bornons à examiner le cas d'espèce d'un plan d'échantillonnage aléatoire simple dans chaque base de sondage avec les tailles d'échantillon n_A et n_B. $\hat{Y}_H$ est exempt de biais. Si la correction de population finie (cpf) est négligeable, la variance de $\hat{Y}_H$ est dans ce cas

$$V\left(\hat{Y}_H\right) = N_A^2\, S_{X_A}^2 \big/ n_A + N_B^2\, S_{X_B}^2 \big/ n_B, \tag{3.2}$$

où

$$\begin{aligned} S_{X_A}^2 &= \frac{N_a}{N_A}\left(S_a^2 + \bar{Y}_a^2\right) + \theta^2 \frac{N_{ab}}{N_A}\left(S_{ab}^2 + \bar{Y}_{ab}^2\right) - \left(\frac{N_a}{N_A}\,\bar{Y}_a + \theta\,\frac{N_{ab}}{N_A}\,\bar{Y}_{ab}\right)^2, \\ S_{X_B}^2 &= \frac{N_b}{N_B}\left(S_b^2 + \bar{Y}_b^2\right) + (1-\theta)^2 \frac{N_{ab}}{N_B}\left(S_{ab}^2 + \bar{Y}_{ab}^2\right) - \left[\frac{N_b}{N_B}\,\bar{Y}_b + (1-\theta)\,\frac{N_{ab}}{N_B}\,\bar{Y}_{ab}\right]^2, \end{aligned}$$

sont les variances de population de x_{Ai} et x_{Bi}. Ici et dans ce qui suit, S_d^2 et $\bar{Y}_d$ désigneront la variance et la moyenne de la population y pour le domaine d, ce d désignant le domaine réel $(d = a, b$ ou $ab)$ ou perçu $\left(d = a^*, b^*, ab^*(A)\right.$ ou $\left.ab^*(B)\right)$.

En cas de défaut de classification de domaine, il nous faut une notation pour les indicateurs d'appartenance à un domaine perçu. Nous définissons $\eta_i^A(a)$ comme étant 1 lorsque l'unité i de la base A est dans le domaine a^* et comme 0 dans les autres cas. Comme chaque unité est dans un domaine perçu

ou l'autre, l'indicateur de présence de l'unité i dans ab^* est $1 - \eta_i^A(a)$; $\eta_i^B(b)$ est défini de la même manière pour la base B. Ainsi, l'estimateur de Hartley devient

$$\begin{aligned}\hat{Y}_H^* &= \hat{Y}_{a^*} + \theta\hat{Y}_{ab^*}^A + (1-\theta)\hat{Y}_{ab^*}^B + \hat{Y}_{b^*} \\ &= \sum_{i=1}^{N_A} I_i^A w_i^A \eta_i^A(a) y_i + \theta\sum_{i=1}^{N_A} I_i^A w_i^A \left(1-\eta_i^A(a)\right) y_i \\ &\quad + \sum_{i=1}^{N_B} I_i^B w_i^B \eta_i^B(b) y_i + (1-\theta)\sum_{i=1}^{N_B} I_i^B w_i^B \left(1-\eta_i^B(b)\right) y_i \\ &= \sum_{i=1}^{N_A} I_i^A w_i^A x_{Ai}^* + \sum_{i=1}^{N_B} I_i^B w_i^B x_{Bi}^*, \end{aligned} \tag{3.3}$$

où $x_{Ai}^* = \eta_i^A(a) y_i + \theta\left(1-\eta_i^A(a)\right) y_i$ et $x_{Bi}^* = \eta_i^B(b) y_i + (1-\theta)\left(1-\eta_i^B(b)\right) y_i$. Le biais de $\hat{Y}_H^*$ est alors :

$$\begin{aligned}\text{Biais} &= E\left(\hat{Y}_H^*\right) - Y \\ &= (1-\theta)\sum_{i=1}^{N_A}\left(\eta_i^A(a) - \delta_i^A(a)\right) y_i + \theta\sum_{i=1}^{N_B}\left(\eta_i^B(b) - \delta_i^B(b)\right) y_i \\ &= (1-\theta)\left(N_{a^*}\bar{Y}_{a^*} - N_a\bar{Y}_a\right) + \theta\left(N_{b^*}\bar{Y}_{b^*} - N_b\bar{Y}_b\right). \end{aligned} \tag{3.4}$$

À noter que $N_{a^*}\bar{Y}_{a^*} - N_a\bar{Y}_a = Y_{a^*} - Y_a = \left(Y_{a\cap a^*} + Y_{ab\cap a^*}\right) - \left(Y_{a\cap a^*} + Y_{a\cap ab^*(A)}\right) = Y_{ab\cap a^*} - Y_{a\cap ab^*(A)}$. Ainsi, le premier terme de l'expression du biais peut être positif ou négatif et grand ou petit selon le nombre relatif d'unités de la population qui perçoivent à tort appartenir ou non à la base en chevauchement, et selon leurs moyennes de réponse. La même constatation vaut pour le second terme. En théorie, les deux pourraient même s'annuler si les erreurs se produisaient dans les deux sens, mais bien sûr la chose est improbable.

L'expression de la variance pour $\hat{Y}_H^*$ est semblable à celle de $V\left(\hat{Y}_H\right)$:

$$V\left(\hat{Y}_H^*\right) = N_A^2 S_{x_A^*}^2 / n_A + N_B^2 S_{x_B^*}^2 / n_B, \tag{3.5}$$

où

$$\begin{aligned}S_{x_A^*}^2 &= \frac{N_{a^*}}{N_A}\left(S_{a^*}^2 + \bar{Y}_{a^*}^2\right) + \theta^2\frac{N_{ab^*(A)}}{N_A}\left(S_{ab^*(A)}^2 + \bar{Y}_{ab^*(A)}^2\right) \\ &\quad - \left(\frac{N_{a^*}}{N_A}\bar{Y}_{a^*} + \theta\frac{N_{ab^*(A)}}{N_A}\bar{Y}_{ab^*(A)}\right)^2, \end{aligned}$$

et

$$\begin{aligned}S_{x_B^*}^2 &= \frac{N_{b^*}}{N_B}\left(S_{b^*}^2 + \bar{Y}_{b^*}^2\right) + (1-\theta)^2\frac{N_{ab^*(B)}}{N_B}\left(S_{ab^*(B)}^2 + \bar{Y}_{ab^*(B)}^2\right) \\ &\quad - \left(\frac{N_{b^*}}{N_B}\bar{Y}_{b^*} + (1-\theta)\frac{N_{ab^*(B)}}{N_B}\bar{Y}_{ab^*(B)}\right)^2 \end{aligned}$$

désignent les variances de x_{Ai}^* et x_{Bi}^*. Le domaine perçu en chevauchement ab^* reçoit aussi la notation (A) ou (B), puisqu'une même unité dans ab peut être perçue comme appartenant à des domaines différents si elle est tirée de bases différentes. Nous étudierons plus en détail à la section 3.3 l'incidence du

défaut de classification sur l'erreur quadratique moyenne (EQM) de l'estimateur de Hartley. Bien sûr, le biais de $\hat{Y}_H^*$ est 0 et $V\left(\hat{Y}_H^*\right) = V\left(\hat{Y}_H\right)$ si le domaine réel et le domaine perçu coïncident.

4 Correction de biais en cas d'erreur de rattachement à un domaine

Lohr (2011, section 6) s'est attaquée au problème qui est le nôtre, c'est-à-dire d'apporter une correction en cas de défaut de classification de domaine dans les plans d'échantillonnage à base double. Lohr a proposé d'apporter une correction à l'estimateur de Hartley en vue d'atténuer le biais par défaut de classification. Nous examinerons sa méthode et la comparerons à la méthode que nous proposons. Nous observons que, pour que la méthode de Lohr donne un estimateur sans biais de moyenne ou de total pour Y, les unités d'un même domaine doivent être à égalité de moyennes, quelle que soit leur appartenance à un domaine perçu. Cette hypothèse ne valait pas pour notre enquête auprès des pêcheurs à la ligne, car nous avons plutôt constaté que la motivation à aller à la pêche avait à voir avec le domaine perçu plutôt que réel d'appartenance en ce qui concerne le permis de pêche. Nous avons élaboré notre méthode de correction de biais dans l'optique différente d'une prise en compte de l'erreur de rattachement à un domaine. Nous supposons que les unités se rattachant à un même domaine perçu doivent être à égalité de moyennes, quelle que soit leur appartenance à un domaine réel. Nous démontrons que la méthode envisagée présente une EQM inférieure à celle de la méthode de Lohr lorsque les deux hypothèses se vérifient. Là où elles ne se vérifient pas, le choix doit se faire selon l'hypothèse qui convient.

4.1 Méthode de Lohr de correction de biais en cas de défaut de classification

Dans le modèle multinomial de Lohr pour le défaut de classification, nous posons que les unités de chacun des domaines a, ab et b ont leurs propres probabilités connues de perception d'appartenance à un de ces trois domaines. Nous prenons ces probabilités pour corriger les estimateurs et éliminer le biais tenant au défaut de classification. Dans la pratique, ces probabilités sont normalement inconnues et doivent être estimées à l'aide d'un échantillon de seconde phase pour lequel on applique une méthode coûteuse (mais exacte) de détermination d'appartenance à un domaine. Il sera question plus avant de cette méthode à la section 4.3.1. Pour l'instant, nous supposons ces probabilités connues.

Nous décrivons la méthode de Lohr pour un plan d'échantillonnage à base double, bien que l'intéressée ait décrit sa méthode pour des bases multiples. Nous définissons des vecteurs aléatoires constitués de variables aléatoires des trois domaines : $\mathbf{Y} = (Y_a, Y_{ab}, Y_b)'$, $\boldsymbol{\delta}_i^A = \left(\delta_i^A(a), \delta_i^A(ab), \delta_i^A(b)\right)'$ et $\boldsymbol{\delta}_i^B = \left(\delta_i^B(a), \delta_i^B(ab), \delta_i^B(b)\right)'$. S'il n'y a pas d'erreur de rattachement à un domaine, $\hat{\mathbf{Y}}^A = \sum_{i \in s_A} \boldsymbol{\delta}_i^A w_i^A y_i$ et $\hat{\mathbf{Y}}^B = \sum_{i \in s_B} \boldsymbol{\delta}_i^B w_i^B y_i$ sont des vecteurs d'estimateurs sans biais de totaux de domaine pour les bases A et B. Dans ce cas, l'estimateur de Hartley à θ fixe peut s'écrire comme $\hat{Y}_H = \left(\mathbf{m}^A\right)' \hat{\mathbf{Y}}^A + \left(\mathbf{m}^B\right)' \hat{\mathbf{Y}}^B$, où $\mathbf{m}^A = (1, \theta, 0)'$ et $\mathbf{m}^B = (0, 1-\theta, 1)'$.

Posons maintenant l'existence d'une erreur de rattachement. Soit $\boldsymbol{\eta}_i^A = \left(\eta_i^A(a), \eta_i^A(ab), \eta_i^A(b)\right)'$ le vecteur des indicateurs d'appartenance à un domaine perçu pour l'unité i de la base A. Ce vecteur peut se formuler comme $\boldsymbol{\eta}_i^A = \left(\mathbf{M}_i^A\right)' \boldsymbol{\delta}_i^A$, où $\mathbf{M}_i^A$ est une matrice 3×3 avec un 1 à la position (d_2, d_1) si l'unité

i du domaine d_2 est perçue comme dans le domaine d_1, et avec 0 dans les autres cas. Lorsque nous employons l'information d'appartenance au domaine perçu plutôt qu'au domaine réel, $\hat{Y}^A$ devient

$$\hat{\mathbf{Y}}^{A^*} = \sum_{i \in s_A} \boldsymbol{\eta}_i^A \, w_i^A \, y_i = \sum_{i \in s_A} \left(\mathbf{M}_i^A\right)' \boldsymbol{\delta}_i^A \, w_i^A \, y_i .$$

Définissons la matrice des probabilités de défaut de classification pour la base A :

$$\boldsymbol{\Phi}^A = \begin{pmatrix} p_{a^*|a} & p_{ab^*(A)|a} & 0 \\ p_{a^*|ab} & p_{ab^*(A)|ab} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{4.1}$$

où $p_{d_1^*|d_2}$ est la probabilité que l'unité du domaine d_2 soit perçue comme étant dans le domaine d_1. $\hat{Y}^{A^*}$ sera normalement entaché d'un biais pour les totaux de domaine de la base A lorsque les éléments hors diagonale sont non nuls, car $E\left(\hat{\mathbf{Y}}^{A^*}\right) = \left(\boldsymbol{\Phi}^A\right)' \mathbf{Y} \neq E\left(\hat{Y}^A\right)$.

Pour corriger le biais, Lohr (2011) a proposé le vecteur de repondération $\tilde{\mathbf{m}}^A = \left(\boldsymbol{\Phi}^A\right)^+ \mathbf{m}^A$, où $\left(\boldsymbol{\Phi}^A\right)^+$ est l'inverse de Moore-Penrose de $\boldsymbol{\Phi}^A$. Sans l'expliciter, cette méthode exige, pour qu'il y ait absence de biais, que l'on pose que l'appartenance au domaine réel détermine la moyenne; en d'autres termes,

$$\bar{Y}_{a \cap ab^*} = \bar{Y}_{a \cap a^*}, \; \bar{Y}_{ab \cap ab^*(A)} = \bar{Y}_{ab \cap a^*} \tag{4.2}$$

Dans ce cas, $E\left(\left(\tilde{\mathbf{m}}^A\right)' \hat{\mathbf{Y}}^{A^*}\right) = \left(\left(\boldsymbol{\Phi}^A\right)^+ \mathbf{m}^A\right)' \left(\boldsymbol{\Phi}^A\right)' \mathbf{Y} = \left(\mathbf{m}^A\right)' \mathbf{Y}$. Après avoir employé une hypothèse semblable d'égalité des moyennes, une matrice du défaut de classification et une correction pour la base B, nous obtenons l'estimateur de Lohr en correction de biais :

$$\begin{aligned} \hat{Y}_L &= \left(\tilde{\mathbf{m}}^A\right)' \hat{\mathbf{Y}}^{A^*} + \left(\tilde{\mathbf{m}}^B\right)' \hat{\mathbf{Y}}^{B^*} \\ &= \hat{Y}_{a^*} \, \phi_{11}^{A+} + \hat{Y}_{ab^*(A)} \, \phi_{21}^{A+} + \theta\left(\hat{Y}_{a^*} \, \phi_{12}^{A+} + \hat{Y}_{ab^*(A)} \, \phi_{22}^{A+}\right) \\ &\quad + (1-\theta)\left(\hat{Y}_{b^*} \, \phi_{32}^{B+} + \hat{Y}_{ab^*(B)} \, \phi_{22}^{B+}\right) + \hat{Y}_{b^*} \, \phi_{33}^{B+} + \hat{Y}_{ab^*(B)} \, \phi_{23}^{B+}, \end{aligned} \tag{4.3}$$

où ϕ_{ij}^{A+} est l'élément (i, j) de $\left(\boldsymbol{\Phi}^A\right)^+$ et où ϕ_{ij}^{B+} se définit de la même manière. Dans le cas particulier des plans d'échantillonnage à base double que nous considérons, les composantes des matrices pseudo-inverses peuvent facilement se formuler en termes exprès si $p_{ab^*(A)|a} + p_{a^*|ab} \neq 1$ et $p_{ab^*(B)|b} + p_{b^*|ab} \neq 1$:

$$\begin{aligned} \phi_{11}^{A+} &= \frac{p_{ab^*(A)|ab}}{D_A}, \; \phi_{21}^{A+} = \frac{-p_{a^*|ab}}{D_A}, \; \phi_{12}^{A+} = \frac{-p_{ab^*(A)|a}}{D_A}, \\ \phi_{22}^{A+} &= \frac{p_{a^*|a}}{D_A}, \; \phi_{33}^{B+} = \frac{p_{ab^*(B)|ab}}{D_B}, \; \phi_{23}^{B+} = \frac{-p_{b^*|ab}}{D_B}, \\ \phi_{32}^{B+} &= \frac{-p_{ab^*(B)|b}}{D_B}, \; \phi_{22}^{B+} = \frac{p_{b^*|b}}{D_B}, \end{aligned} \tag{4.4}$$

où $D_A = 1 - \left(p_{ab^*(A)|a} + p_{a^*|ab}\right)$ et $D_B = 1 - \left(p_{ab^*(B)|b} + p_{b^*|ab}\right)$. Si les probabilités de défaut de classification sont connues, cet estimateur peut se calculer et serait sans biais dans la mesure où (4.2) et les hypothèses correspondantes pour les moyennes de la base B se vérifient. Nous pouvons dégager la variance

par la même technique ayant mené à l'équation (3.2). Voir Lin (2014) aux pages 28 et 29 pour l'expression de la variance.

4.2 Autre méthode de correction de biais en cas de défaut de classification

L'estimateur de Lohr élimine le biais de l'estimateur de Hartley dans l'hypothèse où l'appartenance au domaine réel détermine la moyenne de la variable d'intérêt. Comme cette hypothèse semblait ne pas se vérifier dans le cas des données de l'enquête auprès des pêcheurs à la ligne, nous avons conçu une autre méthode de correction de biais exigeant pour notre application un jeu d'hypothèses mieux adapté.

Rappelons que $Y_{a^*}, Y_{ab^*(A)}, Y_{ab^*(B)}, Y_{b^*}$ désignent les totaux de population des domaines perçus, la notation traduisant la possibilité que l'appartenance au domaine perçu dans la base en chevauchement dépende de la base d'échantillonnage de l'unité. Nous pouvons décomposer les totaux et les tailles de domaine perçu par les étiquettes de domaine réel : $Y_{a^*} = Y_{a\cap a^*} + Y_{ab\cap a^*}$ et $N_{a^*} = N_{a\cap a^*} + N_{ab\cap a^*}$, par exemple. Nous paramétrons notre modèle en définissant les probabilités de défaut de classification en forme inverse par rapport au modèle de Lohr. En d'autres termes, soit $p_{d_1|d_2^*}$ la probabilité qu'une observation perçue comme du domaine d_2 soit en réalité du domaine d_1; nous définissons la matrice des probabilités de défaut de classification pour la base A comme

$$\mathbf{\Lambda}^A = \begin{pmatrix} p_{a|a^*} & p_{ab|a^*} & 0 \\ p_{a|ab^*(A)} & p_{ab|ab^*(A)} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Dans ce cas, nous remplaçons les hypothèses d'égalité des moyennes (4.2) par des hypothèses de détermination des moyennes par l'appartenance au domaine perçu :

$$\overline{Y}_{a\cap a^*} = \overline{Y}_{ab\cap a^*}, \overline{Y}_{a\cap ab^*(A)} = \overline{Y}_{ab\cap ab^*(A)}, \overline{Y}_{b\cap b^*} = \overline{Y}_{ab\cap b^*}, \overline{Y}_{b\cap ab^*(B)} = \overline{Y}_{ab\cap ab^*(B)} \tag{4.5}$$

Lorsque ces hypothèses se vérifient, nous avons

$$Y_a = Y_{a^*}\, p_{a|a^*} + Y_{ab^*(A)}\, p_{a|ab^*(A)}, Y_{ab} = Y_{a^*}\left(1 - p_{a|a^*}\right) + Y_{ab^*(A)}\left(1 - p_{a|ab^*(A)}\right), \tag{4.6}$$

et les expressions sont semblables pour les totaux de domaine dans le cas de la base B. Ainsi, un estimateur en correction de biais est

$$\begin{aligned} \hat{Y}_{\mathrm{CB}} &= \left(\mathbf{\Lambda}^A \mathbf{m}^A\right)' \hat{Y}^{A^*} + \left(\mathbf{\Lambda}^B \mathbf{m}^B\right)' \hat{Y}^{B^*} \\ &= \hat{Y}_{a^*}\, p_{a|a^*} + \hat{Y}_{ab^*(A)}\, p_{a|ab^*(A)} + \theta\left(\hat{Y}_{a^*}\left(1 - p_{a|a^*}\right) + \hat{Y}_{ab^*(A)}\left(1 - p_{a|ab^*(A)}\right)\right) \\ &\quad + (1-\theta)\left(\hat{Y}_{b^*}\left(1 - p_{b|b^*}\right) + \hat{Y}_{ab^*(B)}\left(1 - p_{b|ab^*(B)}\right)\right) + \hat{Y}_{b^*}\, p_{b|b^*} + \hat{Y}_{ab^*(B)}\, p_{b|ab^*(B).} \end{aligned} \tag{4.7}$$

$\hat{Y}_{\mathrm{CB}}$ est sans biais là où les hypothèses d'égalité des moyennes (4.5) se vérifient. Sa variance peut se calculer par la même technique que pour l'obtention de l'équation (3.2). Voir Lin (2014) aux pages 33 et 34 pour l'expression de la variance.

4.3 Comparaison des méthodes de correction de biais

Dans cette section, nous examinons le rendement de $\hat{Y}_L$, de $\hat{Y}_{\text{CB}}$ et de leur homologue non corrigé $\hat{Y}_H$. Nous examinons également les estimateurs en correction de biais qui peuvent s'employer lorsque les probabilités de défaut de classification sont inconnues. Nous pouvons établir ces probabilités en remplaçant les valeurs connues des probabilités en (4.3) et (4.7) par des estimateurs des probabilités de défaut à partir de sous-échantillons de phase 2. Nous prélevons un tel sous-échantillon sur l'échantillon de départ à l'aide d'une méthode exacte (et normalement coûteuse) de collecte de données en vue de vérifier l'appartenance au domaine réel. Soit $I_i^{A(2)}$ l'indicateur d'échantillon de phase 2 pour la base A, $\pi_i^{A(2)} = P\left\{I_i^{A(2)} = 1 \middle| I_i^A = 1\right\}$ sa probabilité de sélection conditionnelle et $w_i^{A(2)} = 1/\pi_i^{A(2)}$ son poids de phase 2. Nous pouvons alors élaborer des estimateurs quotient des probabilités de défaut de classification. Ainsi, $p_{a|a^*}$ peut s'estimer par

$$\hat{p}_{a|a^*}^{(2)} = \frac{\sum_{i=1}^{N_A} I_i^A I_i^{A(2)} w_i^A w_i^{A(2)} \delta_i\left(a^* \cap a\right)}{\sum_{i=1}^{N_A} I_i^A I_i^{A(2)} w_i^A w_i^{A(2)} \delta_i\left(a^*\right)} = \hat{N}_{a^* \cap a}^{(2)} \Big/ \hat{N}_{a^*}^{(2)}. \tag{4.8}$$

L'estimateur de Lohr demande que nous estimions les probabilités conditionnelles inversées dans le cas de l'échantillon de phase 2, soit $\hat{p}_{a^*|a} = \hat{N}_{a^* \cap a}^{(2)} \Big/ \hat{N}_a^{(2)}$. Nous pouvons élaborer des estimateurs semblables pour les composantes des quatre matrices de défaut de classification. Si nous procédons à la substitution de ces estimateurs pour les paramètres en (4.3) et (4.7), les estimateurs obtenus en correction de biais sont désignés par $\hat{\hat{Y}}_{\text{CB}}$ et $\hat{\hat{Y}}_L$.

Dans cette section, nous offrons plusieurs comparaisons entre les cinq estimateurs. Dans un premier exemple, nous limitons la comparaison des variances à $\hat{Y}_H$, $\hat{Y}_L$ et $\hat{Y}_{\text{CB}}$ seulement pour un scénario simple où toutes les moyennes de domaine sont égales, de sorte que les hypothèses requises au sujet des moyennes vaillent pour les deux estimateurs comparés. Le but est d'illustrer le gonflement de la variance susceptible de se produire dans $\hat{Y}_L$. Dans notre deuxième exemple, nous reproduisons partiellement la simulation présentée dans Lohr (2011), mais en ajoutant $\hat{Y}_{\text{CB}}$ et $\hat{\hat{Y}}_{\text{CB}}$. Nous comparons les erreurs quadratiques moyennes (EQM) des estimateurs lorsque les deux jeux d'hypothèses d'égalité des moyennes sont respectés. Nous examinons en outre l'effet du passage du plan de sondage de phase 2 d'un échantillon aléatoire simple à un échantillon stratifié par domaine perçu. Dans un troisième exemple, nous nous attachons à la robustesse des deux estimateurs à l'égard des hypothèses formulées en examinant en simulation le biais et l'EQM des estimateurs lorsque ni (4.2) ni (4.5) ne se vérifient. Dans tous les cas, θ est fixé à 0,5.

4.3.1 Exemple 1 : Gonflement de la variance

Nous avons comparé les variances de $\hat{Y}_{\text{CB}}$ et $\hat{Y}_L$ à l'EQM de $\hat{Y}_H$ en cas d'erreur de rattachement à un domaine, mais là où les taux en question sont connus. Nous tenons les populations pour homogènes et toutes les moyennes et les variances de domaine pour constantes : $\overline{Y}_a = \overline{Y}_{ab} = \overline{Y}_b = 2$, de sorte que, à la fois, les hypothèses d'égalité des moyennes se vérifient et $\hat{Y}_{\text{CB}}$ et $\hat{Y}_L$ soient exempts de biais. Nous posons $S_a^2 = S_{ab}^2 = S_b^2 = 1$. Dans cet exemple, les bases sont considérées comme petites et le chevauchement comme appréciable : $N_a = 3\,000$, $N_{ab} = 2\,000$, $N_b = 3\,000$. Nous prenons des échantillons aléatoires

simples de tailles $n_A = 100$ et $n_B = 50$. Nous calculons l'EQM de l'estimateur de Hartley par (3.4) et (3.5) et les variances des estimateurs en correction de biais, par les expressions dans Lin (2014) aux pages 28 et 33. Nous regardons les deux scénarios suivants :

1. La probabilité de défaut de classification pour les unités du domaine a varie de 0 à 1, et il n'existe aucune autre erreur de rattachement à un domaine.
2. La probabilité de défaut de classification pour les unités du domaine ab (en cas d'échantillonnage dans la base A) varie de 0 à 1; il n'existe aucune autre erreur de rattachement à un domaine.

Les deux parties de la figure 4.1 présentent les EQM des trois estimateurs en fonction des deux probabilités de défaut de classification. Pour chaque condition, $\hat{Y}_{\text{CB}}$ présente une EQM inférieure à celle de $\hat{Y}_L$ sur la plage de valeurs des probabilités en question, bien que les deux estimateurs soient d'un rendement à peu près semblable lorsque les probabilités sont petites. L'EQM de $\hat{Y}_{\text{CB}}$ ne varie guère sur la plage de valeurs, mais l'EQM de $\hat{Y}_L$ augmente sans borne à mesure que la probabilité approche de l'unité, et ce, parce que les composantes du pseudo-inverse (4.4) se font grandes à mesure que le défaut de classification (unique) approche de l'unité, ce qui gonfle les coefficients des estimations de sous-domaine en (4.3).

Dans la pratique, si les probabilités de défaut de classification étaient connues et dépassaient $1/2$, le processus d'identification de domaine s'inverserait simplement, si bien que ce scénario n'a pas d'importance dans la pratique. Nous verrons cependant dans le prochain exemple que, lorsque les probabilités de défaut de classification sont inconnues et doivent faire l'objet d'une estimation, le rendement de $\hat{Y}_L$ est sensible à la qualité des estimateurs de probabilité de phase 2. C'est que, même si les probabilités de défaut de classification sont modestes, les estimations correspondantes peuvent être grandes, surtout lorsque le plan d'échantillonnage de phase 2 est inefficace; le même gonflement de la variance peut alors se produire.

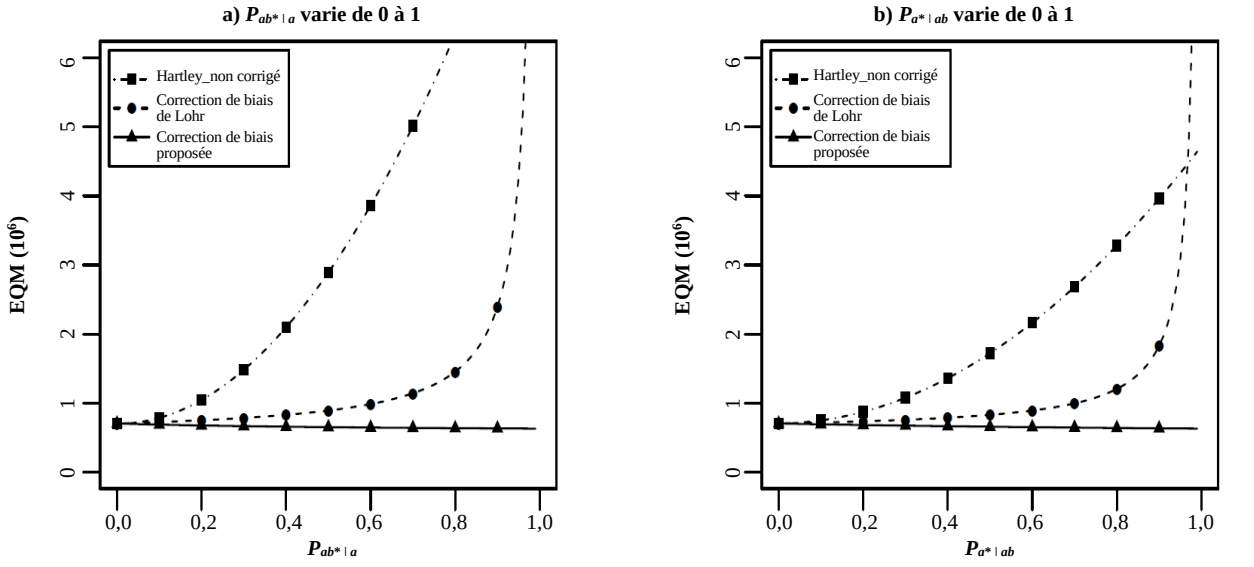

Figure 4.1 Comparaison des EQM de $\hat{Y}_H$, $\hat{Y}_L$ et $\hat{Y}_{\text{CB}}$ sur la plage des valeurs de probabilité d'un défaut unique de classification.

4.3.2 Exemple 2 : Efficacité relative en simulation des cinq estimateurs

Dans notre deuxième exemple, nous procédons par simulation pour comparer le rendement de $\hat{Y}_H$, $\hat{Y}_L$, $\hat{\hat{Y}}_L$, $\hat{Y}_{\text{CB}}$ et $\hat{\hat{Y}}_{\text{CB}}$ pour différents scénarios de défaut de classification, l'un où les deux hypothèses d'égalité des moyennes de domaine se vérifient et l'autre où elles ne le font pas. Les tailles de domaine choisies dans cet exemple ont été $N_a = 48\,000$, $N_{ab} = 2\,000$ et $N_b = 3\,000$, le but étant d'imiter l'enquête sur la motivation des pêcheurs à la ligne où une base (celle des adresses) était bien plus grande que l'autre (base des permis). Nous avons produit la population par le modèle $y_i \sim x_i + 1$ pour $i = 1, \ldots, N$ avec $x_i \sim$ Poisson(1), et ce, indépendamment du domaine. Ainsi, les moyennes et les variances de la population étaient approximativement les mêmes que dans le premier exemple. Nous avons simulé le prélèvement d'échantillons aléatoires simples sur chaque base avec des tailles d'échantillon $n_A = 100$ et $n_B = 50$. Pour les échantillons de phase 2, nous avons prélevé un échantillon aléatoire simple (EAS) avec quatre tailles variant de 20 % à 80 % de l'échantillon de phase 1, ce qui a donné des tailles de sous-échantillon (m_A) de 20 à 80 pour la base A et (m_B) de 10 à 40 pour la base B. Si une réplique présentait un échantillon de phase 2 de moins de deux unités pour un des sous-domaines, nous posions l'absence de défaut de classification pour ce domaine; nous avons donc élaboré des estimateurs comme s'il n'y avait pas d'erreur de rattachement pour le sous-domaine. Nous avons simulé 16 cas de défaut de classification avec chaque fois des paires des quatre cas suivants pour une même base :

1. Cas de défaut de classification pour la base A (MPA)
   a) $p_{a^*|a} = 1,\ p_{ab^*(A)|ab} = 1$
   b) $p_{a^*|a} = 0{,}9,\ p_{ab^*(A)|a} = 0{,}1,\ p_{ab^*(A)|ab} = 1$
   c) $p_{a^*|a} = 0{,}9,\ p_{ab^*(A)|a} = 0{,}1,\ p_{ab^*(A)|ab} = 0{,}9,\ p_{a^*|ab} = 0{,}1$
   d) $p_{a^*|a} = 1,\ p_{ab^*(A)|ab} = 0{,}9,\ p_{a^*|ab} = 0{,}1$
2. Cas de défaut de classification pour la base B (MPB)
   a) $p_{b^*|b} = 1,\ p_{ab^*(B)|ab} = 1$
   b) $p_{b^*|b} = 0{,}8,\ p_{ab^*(B)|b} = 0{,}2,\ p_{ab^*(B)|ab} = 1$
   c) $p_{b^*|b} = 0{,}8,\ p_{ab^*(B)|b} = 0{,}2,\ p_{ab^*(B)|ab} = 0{,}8,\ p_{b^*|ab} = 0{,}2$
   d) $p_{b^*|b} = 1,\ p_{ab^*(B)|ab} = 0{,}8,\ p_{b^*|ab} = 0{,}2$

Ainsi, nous avons examiné en simulation 64 configurations (16 cas de défaut de classification $\times$ 4 tailles d'échantillon de phase 2) et avons produit 10 000 échantillons de phase 1 pour chaque configuration et calculé les estimations correspondantes. Nous avons établi l'EQM empirique comme l'écart quadratique moyen de chaque estimation par rapport au Y réel sur les 10 000 échantillons répétés. Nous récapitulons les résultats au tableau 4.1 et aux figures 4.2 et 4.3.

Le tableau 4.1 présente les résultats des 16 cas de défaut de classification pour un taux d'échantillonnage de phase 2 de 40 % ($m_A = 40$ et $m_B = 20$). Les conclusions que nous tirons sont cohérentes pour toutes les tailles d'échantillon. Nous pouvons voir que $\hat{Y}_{\text{CB}}$ et $\hat{\hat{Y}}_{\text{CB}}$ sont moins variables que $\hat{Y}_L$ et $\hat{\hat{Y}}_L$ pour tous les cas de défaut de classification sauf s'il y a absence d'erreur. Par ailleurs, s'il y a défaut de classification uniquement dans de petits domaines (cas a ou d pour la base A), il n'y a guère d'avantage à recourir à la

correction de biais, puisque $\hat{Y}_H$ est d'un rendement égal ou supérieur à celui des estimateurs en correction de biais, d'où l'impression que cette correction pourrait ne pas être avantageuse à moins que le défaut de classification ne touche une grande partie de la population. Enfin, nous observons que $\hat{\hat{Y}}_L$ est d'un rendement pire lorsque le cas de défaut de classification c se vérifie dans la base A. Nous examinerons cet effet de plus près.

Tableau 4.1
EQM $(\times 10^7)$ pour les estimateurs à double base de sondage

		Probabilités connues			Probabilités estimées	
MPA	MPB	$\hat{Y}_{\mathrm{CB}}$	$\hat{Y}_L$	$\hat{Y}_H$	$\hat{\hat{Y}}_{\mathrm{CB}}$	$\hat{\hat{Y}}_L$
a	a	2,53	2,53	2,53	2,53	2,53
b	a	2,45	2,86	4,77	2,66	3,00
c	a	2,44	2,96	4,60	2,66	3,80
d	a	2,52	2,54	2,53	2,55	2,72
a	b	2,53	2,55	2,57	2,54	2,55
b	b	2,45	2,88	5,37	2,66	3,02
c	b	2,44	2,98	5,18	2,67	3,82
d	b	2,52	2,56	2,55	2,56	2,74
a	c	2,53	2,57	2,54	2,55	2,63
b	c	2,45	2,90	4,99	2,67	3,10
c	c	2,44	3,00	4,81	2,67	3,90
d	c	2,52	2,59	2,54	2,56	2,81
a	d	2,53	2,54	2,56	2,54	2,54
b	d	2,45	2,87	4,44	2,66	3,01
c	d	2,44	2,97	4,28	2,66	3,81
d	d	2,52	2,55	2,57	2,55	2,73

La figure 4.2 indique le rapport d'EQM entre les estimateurs en correction de biais à probabilités inconnues et à probabilités connues; la partie a) présente $\mathrm{EQM}(\hat{\hat{Y}}_{\mathrm{CB}})/\mathrm{EQM}(\hat{Y}_{\mathrm{CB}})$ et la partie b) $\mathrm{EQM}(\hat{\hat{Y}}_L)/\mathrm{EQM}(\hat{Y}_L)$. Les quatre lignes indiquent le rapport pour les quatre tailles d'échantillon de phase 2 et pour les 16 cas de défaut de classification disposés sur l'axe des x. La distance du rapport au-dessus de l'unité mesure la pénalité de variance causée par une estimation des probabilités de défaut de classification. Ce qui ressort de cette figure est l'importante pénalité pour $\hat{\hat{Y}}_L$ dans les cas de défaut de classification c et d de la base A et pour les petites tailles d'échantillon, ainsi que l'absence relative d'un tel effet pour $\hat{\hat{Y}}_{\mathrm{CB}}$. Les deux cas accusent une erreur non nulle de rattachement pour le petit domaine en chevauchement. Ainsi, l'échantillon de phase 1 a souvent moins d'unités à offrir par lesquelles estimer $p_{a^*|ab}$, ce qui entache de bruit l'estimation de la probabilité de défaut de classification. Même si la probabilité réelle de défaut n'est pas proche de l'unité pour les cas c et d $\left(p_{a^*|ab} = 0{,}1\right)$, le fait qu'elle soit inconnue et doive faire l'objet d'une estimation à partir de données clairsemées signifie que certains échantillons produisent des estimations extrêmes et gonflent la variance, comme nous l'avons vu à la figure 4.1.

Nous pouvons voir à la partie a) de la figure 4.2 que le même effet est présent pour $\hat{\hat{Y}}_{\mathrm{CB}}$, mais dans une moindre mesure. Le plus petit échantillon de phase 2 $(m_A = 20, m_B = 10)$ est d'un rendement démesurément pire que l'échantillon qui le suit en taille. La même cause joue : un échantillon aléatoire simple de phase 1 peut produire trop peu d'unités dans le domaine $ab^*(A)$. Il reste qu'il est plus simple d'améliorer l'efficacité du plan de sondage de phase 2 pour l'estimation de $p_{a|ab^*(A)}$ que de l'améliorer pour

$p_{a^*|ab}$, et ce, parce que les domaines perçus sont observables pour toutes les unités de l'échantillon de phase 1, au contraire des domaines réels, si bien que l'analyste peut contrôler par voie de stratification la taille d'échantillon dans le premier cas, mais non dans le second.

Pour voir dans quelle mesure nous en tirons un avantage, nous avons procédé à une autre simulation et étudié le rendement de $\hat{\bar{Y}}_{\text{CB}}$ lorsque le plan de sondage de phase 2 est stratifié. Nous avons choisi un plan d'échantillonnage à égalité des tailles d'échantillon pour chacune des strates de domaine perçu. Ainsi, dans la configuration de taux d'échantillonnage à 20 %, nous avons tiré $m_A/2 = 20/2 = 10$ observations des domaines a^* et ab^* (A) pour l'échantillon de phase 1. Si nous disposions de moins que $m_A/2$ observations (ou $m_B/2$) mais de plus d'une unité, tous les éléments étaient sélectionnés et le reste était tiré de l'autre domaine. Dans cette simulation, nous avons employé les 64 configurations de l'exemple précédent, la seule différence résidant dans le plan de sondage stratifié en phase 2 au lieu de l'échantillonnage aléatoire simple.

Nous présentons à la figure 4.3 les résultats de cette simulation où le rapport correspond à un effet de plan (mais pour l'EQM plutôt que la variance) pour l'échantillonnage de phase 2. Le graphique indique que, dans certaines configurations, il est possible d'apporter une amélioration significative à l'EQM de $\hat{\bar{Y}}_{\text{CB}}$ en stratifiant l'échantillon de phase 2. Le gain est particulièrement important pour la petite taille d'échantillon.

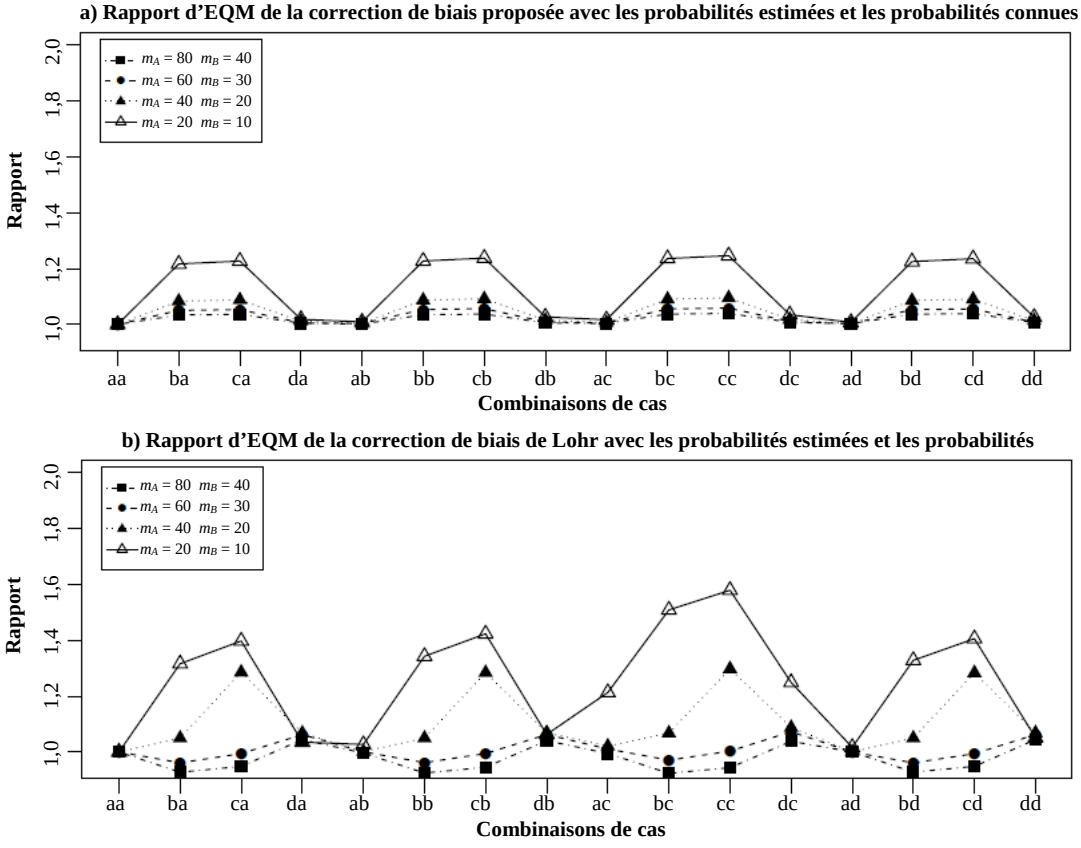

Figure 4.2 Rapport d'EQM en simulation pour les estimateurs à probabilités inconnues de défaut de classification : a) EQM($\hat{\bar{Y}}_{\text{CB}}$) / EQM($\hat{Y}_{\text{CB}}$) et b) EQM($\hat{\bar{Y}}_L$) / EQM($\hat{Y}_L$).

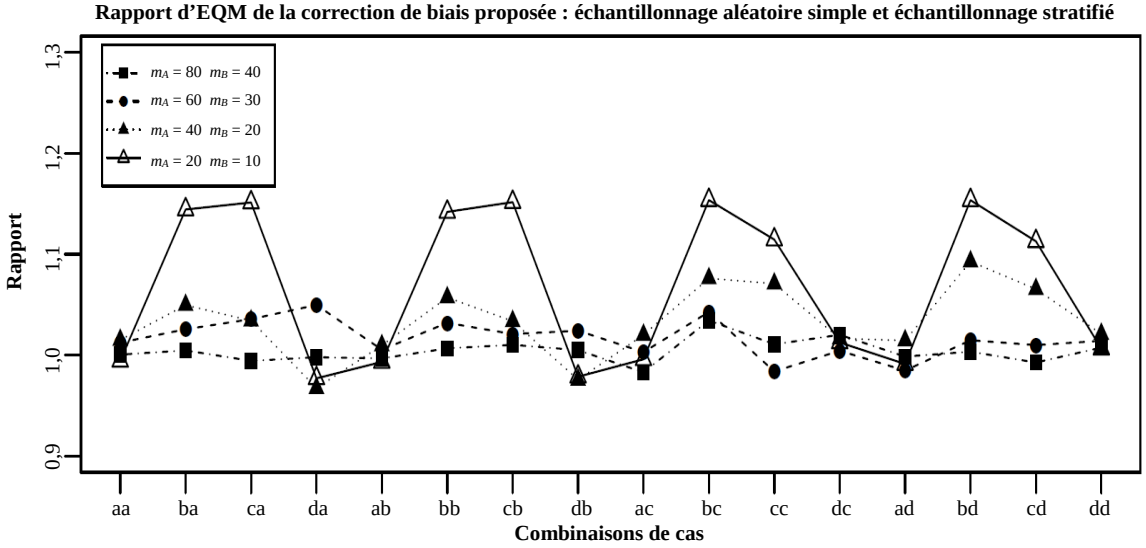

Figure 4.3 Rapport d'EQM en simulation pour les estimateurs avec deux plans de sondage de phase 2 : $EQM_{srs}(\hat{\hat{Y}}_{CB}) / EQM_{str}(\hat{\hat{Y}}_{CB})$.

4.3.3 Exemple 3 : Robustesse en cas de violation des hypothèses relatives aux moyennes

Les deux estimateurs en correction de biais demandent des hypothèses d'égalité des moyennes comme garantie d'absence de biais dans le cas des probabilités connues de défaut de classification. Ce sont des hypothèses qui diffèrent. Lin (2014) présente aux pages 43 à 47 des expressions du biais de $\hat{Y}_L$ et $\hat{Y}_{CB}$ lorsque les hypothèses en question ne se vérifient pas. Dans notre exemple, nous étudions par simulation la taille du biais pour $\hat{\hat{Y}}_L$ et $\hat{\hat{Y}}_{CB}$, ainsi que pour $\hat{Y}_L$ et $\hat{Y}_{CB}$. Les scénarios de simulation dans cet exemple sont semblables aux scénarios qui précèdent sauf pour les moyennes. Nous nous sommes attachés à seulement 3 des 16 cas de défaut de classification, c'est-à-dire à ceux où le défaut de classification de domaine se présente uniquement pour la base de grande taille (indiquée en premier lieu) : *ba*, *ca* et *da*.

Nous avons simulé les populations de sorte qu'une des quatre moyennes de sous-domaine soit d'environ 3 ($y \sim x_i + 2$, avec $x_i \sim$ Poisson(1)) et que les autres demeurent à environ 2 ($y \sim x_i + 1$). Dans chaque cas donc, les hypothèses d'égalité des moyennes (4.2) et (4.5) n'étaient pas respectées pour les deux estimateurs en correction de biais. Nous avons calculé le biais empirique, la variance et l'EQM de chacun des cinq estimateurs à partir de 10 000 échantillons répétés pour chacune des 48 configurations (3 cas de défaut de classification × 4 cas de violation de l'hypothèse d'égalité des moyennes × 4 tailles d'échantillon de phase 2). Nous livrons les résultats pour un sous-ensemble représentatif de ces configurations à la figure 4.4.

Cette figure présente les diagrammes de quartiles des estimations en simulation de $\hat{Y}_H$, $\hat{\hat{Y}}_{CB}$ et $\hat{\hat{Y}}_L$ pour une des tailles d'échantillon de phase 2 ($m_A = 40$ et $m_B = 20$). Les colonnes du diagramme indiquent les résultats pour les quatre cas de violation de l'hypothèse d'égalité des moyennes et les lignes, pour les trois

cas de défaut de classification. Le trait horizontal dans chaque diagramme correspond au total réel de population. La figure montre que tant $\hat{\hat{Y}}_{\mathrm{CB}}$ que $\hat{\hat{Y}}_{L}$ ont un biais inférieur à celui de $\hat{Y}_{H}$ pour toutes les configurations. Ainsi, de formuler la mauvaise hypothèse au sujet de la moyenne est mieux pour le biais que de supposer une absence d'erreur de rattachement à un domaine. Pour les configurations considérées, il semblerait que $\hat{\hat{Y}}_{\mathrm{CB}}$ en particulier n'est pas trop sensible à de légères violations de l'hypothèse d'égalité des moyennes.

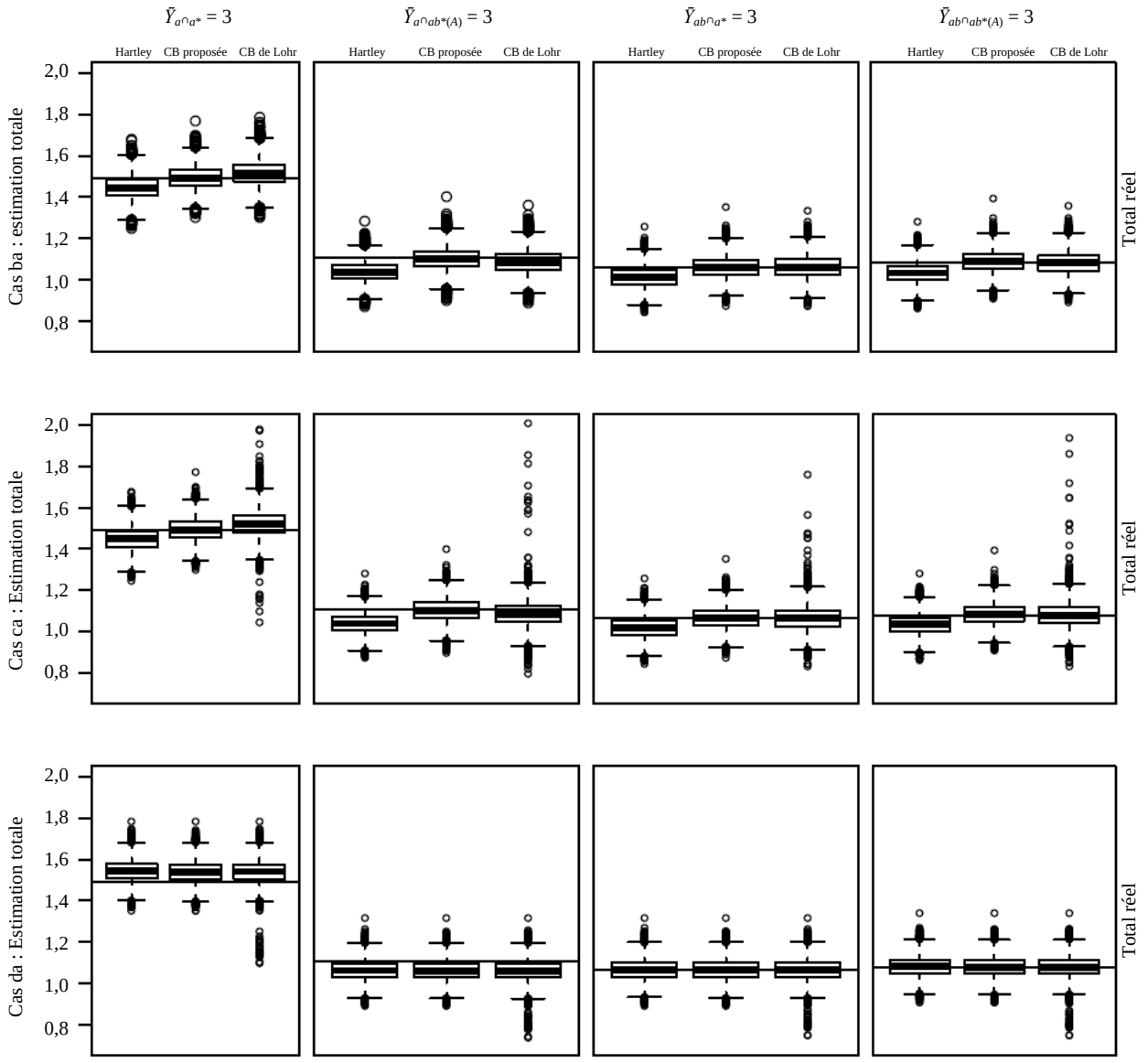

Figure 4.4 Estimation du total $(\times 10^5)$ pour différents cas de violation de l'hypothèse d'égalité des moyennes.

5 Inférence pour $\hat{\hat{Y}}_{\mathrm{CB}}$

Lohr (2011) fait observer à la section 4 que l'inférence est simple pour les estimateurs à double base de sondage et à pondération non aléatoire lorsqu'on utilise un logiciel d'enquête standard. Tel est le cas tant

pour $\hat{Y}_L$ que pour $\hat{Y}_{\mathrm{CB}}$. Dans le second cas, la pondération élaborée pour les unités des quatre domaines est $\tilde{w}_i^A = w_i^A \left(p_{a|a^*} + \theta\left(1 - p_{a|a^*}\right)\right)$ et $\tilde{w}_i^A = w_i^A \left(p_{a|ab^*(A)} + \theta\left(1 - p_{a|ab^*(A)}\right)\right)$ pour les unités des domaines a^* et ab^* échantillonnées dans la base A; elle est $\tilde{w}_i^B = w_i^B \left((1-\theta)\left(1 - p_{b|b^*}\right) + p_{b|b^*}\right)$ et $\tilde{w}_i^B = w_i^B \left((1-\theta)\left(1 - p_{b|ab^*(B)}\right) + p_{b|ab^*(B)}\right)$ pour les unités des domaines b^* et ab^* échantillonnées dans la base B. Nous pouvons alors calculer $\hat{Y}_{\mathrm{CB}}$ et son erreur-type en fournissant au logiciel les fichiers contenant les données et les poids des deux bases. S'il s'agit cependant d'estimer les probabilités de défaut de classification comme pour $\hat{\hat{Y}}_{\mathrm{CB}}$, les variances sont gonflées, ainsi que le fait voir la partie a) de la figure 4.2. Dans ce cas, nous pourrions appliquer les méthodes de linéarisation, jackknife ou bootstrap pour tenir compte de cette variance majorée.

Lin (2014) présente à la section 5.2.2 une expression approchée de la variance pour $\hat{\hat{Y}}_{\mathrm{CB}}$ avec échantillonnage aléatoire simple aussi bien en phase 1 qu'en phase 2 en se fondant sur la méthode de linéarisation. Toutefois, il faut un codage spécial pour appliquer cette méthode, et nous devrions l'adapter à des plans de sondage complexes aux deux phases. Ainsi, nous avons choisi d'étudier l'exactitude d'une méthode approchée qui fait abstraction du surcroît de variabilité créé par l'estimation des probabilités de défaut de classification et qui produit des intervalles de confiance par ce même logiciel d'enquête standard. La seule condition préalable à ce traitement est que nous ayons à calculer les probabilités estimées de défaut de classification à partir de l'échantillon de phase 2 et à les substituer aux probabilités connues dans les expressions de la pondération qui précèdent.

Pour éprouver cette méthode, nous avons simulé une population aux caractéristiques semblables à celles de l'exemple 2 à la section précédente. Nous avons examiné un sous-ensemble des cas de défaut de classification examinés dans cette section, à savoir *db*, *cb* et *ca*. Nous avons fixé à $n_A = 400$ et $n_B = 200$ les tailles des échantillons de phase 1; nous avons choisi trois taux d'échantillonnage de phase 2 allant de 5 % à 20 % (c'est-à-dire de $m_A = 20$ et $m_B = 10$ à $m_A = 80$ et $m_B = 40$) en échantillonnage aléatoire simple et en échantillonnage stratifié. Nous avons produit 10 000 échantillons répétés pour chaque configuration. Nous avons estimé les probabilités de défaut de classification avec chaque échantillon et les avons introduites dans les expressions de la pondération qui précèdent. À des fins de comparaison, nous avons aussi produit des poids avec les probabilités connues de défaut de classification et vérifié le rendement des intervalles de confiance en fonction de $\hat{Y}_{\mathrm{CB}}$. Nous avons fait tous les calculs avec le paquet *survey* en langage R. Nous avons tiré de chaque échantillon répété les estimations d'erreur-type produites par le logiciel, ainsi que l'intervalle de confiance à 95 % pour le total de population (par la procédure jackknife standard du paquet *survey*).

Nous récapitulons les résultats au tableau 5.1. Les colonnes intitulées *Var.sim.* présentent la variance de chaque estimateur calculée à partir des 10 000 répliques, ce qui sera notre meilleure évaluation des variances réelles. Les colonnes intitulées *Var.est.* font la moyenne sur les 10 000 échantillons répétés des estimations produites par le logiciel de la variance de $\hat{Y}_{\mathrm{CB}}$ et $\hat{\hat{Y}}_{\mathrm{CB}}$. Les colonnes intitulées *Taux succès* font voir la

proportion des répliques pour lesquelles l'intervalle de confiance comprend le total réel. Les parties a) et b) livrent respectivement les résultats des plans d'échantillonnage aléatoire simple et d'échantillonnage stratifié de phase 2.

Tableau 5.1
Estimation de la variance et étendue de l'intervalle de confiance

			Probabilités connues			Probabilités estimées								
						$m_A = 80$ $m_B = 40$			$m_A = 40$ $m_B = 20$			$m_A = 20$ $m_B = 10$		
		cas	Var. sim.	Var. est.	Taux succès	Var. sim.	Var. est.	Taux succès	Var. sim.	Var. est.	Taux succès	Var. sim.	Var. est.	Taux succès
a) Méthode d'échantillonnage aléatoire simple $(\times 10^6)$	EAS	db	6,42	6,45	0,95	6,60	6,45	0,94	7,00	6,45	0,94	7,14	6,46	0,94
		cb	6,36	6,26	0,94	7,42	6,28	0,93	8,46	6,31	0,91	11,6	6,34	0,85
		ca	6,26	6,27	0,95	7,20	6,29	0,93	8,41	6,33	0,91	11,8	6,37	0,85
b) Méthode de stratification $(\times 10^6)$	STR	db	6,35	6,46	0,95	6,59	6,46	0,95	6,91	6,45	0,94	7,54	6,45	0,93
		cb	6,18	6,27	0,95	6,65	6,27	0,94	7,16	6,27	0,93	8,38	6,27	0,91
		ca	6,50	6,28	0,94	6,94	6,28	0,94	7,00	6,28	0,93	7,86	6,29	0,92

Si nous comparons les colonnes *Var.sim.* et *Var.est.*, nous voyons que la variance de $\hat{\hat{Y}}_{\mathrm{CB}}$ est sous-estimée en moyenne sur l'ensemble des configurations. Comme on pouvait s'y attendre, la sous-estimation est pire pour la taille d'échantillon la plus petite et le plan de sondage inefficace (EAS). C'est pourquoi l'étendue de l'intervalle de confiance est moindre que sa valeur nominale pour la plupart des configurations. La sous-couverture demeure toutefois légère (moins de 5 %) dans tous les cas sauf pour la plus petite taille d'échantillon et l'EAS de phase 2. Comme les intervalles de confiance avec $\hat{Y}_{\mathrm{CB}}$ comportaient leurs valeurs nominales, nous pouvons en conclure que la sous-couverture était entièrement due à l'estimation des probabilités de défaut de classification. Nous pensons que, dans l'inférence, on peut sans danger ne pas tenir compte du surcroît de variation tenant à l'estimation de ces probabilités si on emploie un plan efficace de sondage de phase 2, à condition que les tailles d'échantillon ne soient pas très petites. À en juger par notre simulation et si nous estimons les probabilités de défaut de classification à partir d'au moins 10 unités dans chaque domaine perçu, les probabilités de couverture ne déviaient pas de plus de quelques points de pourcentage. C'est ce qu'on peut faire avec une taille moindre d'échantillon dans l'ensemble pour un échantillon stratifié de phase 2 par opposition à un échantillon aléatoire simple, et c'est pourquoi nous recommandons un plan de sondage stratifié pour la phase 2.

6 Exemple : Enquête auprès des pêcheurs à la ligne

Nous illustrons l'utilisation de la méthode proposée de correction de biais par son application à une enquête postale à double base de sondage auprès des pêcheurs à la ligne de Caroline du Nord en 2009. C'était là une enquête pilote permettant de vérifier plusieurs changements apportés à un programme permanent de collecte de données de la NOAA sur l'activité récréative de pêche à la ligne en eau de mer. Le degré d'activité était défini par le nombre de déplacements de pêche pendant une certaine période.

L'enquête comportait une base d'adresses en Caroline du Nord et une base de permis avec les noms et les adresses des pêcheurs à la ligne titulaires d'un permis d'un des types de permis offerts. La population visée se composait des pêcheurs à la ligne pêchant à des fins récréatives en eau salée dans cet État, quel que soit leur lieu de résidence. La période de pêche considérée était le cycle 6 (novembre et décembre) en 2009. Les deux bases se caractérisaient ensemble par une certaine sous-couverture, parce que les pêcheurs non titulaires de permis dont le domicile se trouvait à l'extérieur de la Caroline du Nord n'étaient pas compris dans l'union des deux bases.

6.1 Plan de sondage

La base des adresses, qui avait été obtenue du US Postal Service, visait tous les ménages de Caroline du Nord. La base des permis comprenait tous les gens recensés dans la base de données sur les pêcheurs à la ligne autorisés en Caroline du Nord, et ce, à la date du tirage de la base en question, soit plusieurs jours avant l'envoi postal des questionnaires. Des échantillons indépendants d'adresses ont été tirés des deux bases. On a estimé le degré d'activité de pêche en Caroline du Nord dans le cycle 6 en 2009 à l'aide de l'estimateur de Hartley et avec $\theta = 1/2$. Dans la base des adresses, l'échantillonnage a été complexe et en deux phases. Dans la base des permis, il a pu se faire en une seule phase. Dans cette application, les unités étaient à risque de défaut de classification seulement si elles étaient choisies dans la plus grande des deux bases, celle des adresses, puisqu'on ignorait si les membres des ménages en question étaient titulaires d'un permis de pêche. Les analystes savaient que tous les gens sélectionnés dans la base des permis détenaient un permis en règle pendant le cycle. Ils savaient aussi si les intéressés avaient ou non une adresse en Caroline du Nord, puisque leur adresse figurait dans la base.

L'échantillonnage dans la base des adresses s'est fait en deux phases. On a d'abord prélevé un échantillon aléatoire de 1 800 adresses et on l'a stratifié par région. Les strates d'adresses étaient définies comme les comtés côtiers et non côtiers de Caroline du Nord avec un échantillon de 900 unités dans chaque cas. Dans un questionnaire de sélection, on demandait si tout membre du ménage était allé pêcher en eau salée dans les 12 derniers mois. L'échantillon de seconde phase comprenait un pêcheur choisi au hasard dans chaque ménage ayant déclaré un membre pêcheur en première phase. Un second pêcheur a été choisi au sein des ménages déclarant plus d'un pêcheur adulte. Si on a échantillonné en deux phases, c'est pour éviter d'envoyer un long questionnaire aux ménages non pêcheurs, d'où une diminution du coût et une augmentation du taux de réponse.

La base des permis était la base de données sur les autorisations de pêche en Caroline du Nord. Tous les gens figurant dans la base de données le jour du tirage de la base de sondage et autorisés à pêcher pendant la période considérée (cycle 6 en 2009) étaient inclus. On a traité préalablement la base pour qu'elle se prête à l'échantillonnage. On a retranché les enregistrements multiples comportant les mêmes données de base (nom, date de naissance et adresse), tout comme les pêcheurs dont l'enregistrement indiquait qu'ils avaient moins de 18 ans. La base des permis a été divisée en trois strates pour la région côtière, la région non côtière

et la région extérieure à la Caroline du Nord. On a trié le fichier par adresse et prélevé un échantillon systématique de 450 pêcheurs sur chaque strate. L'échantillonnage dans la base des permis s'est fait en une seule phase à l'aide d'un questionnaire identique au questionnaire de seconde phase pour l'échantillon de la base des adresses. Comme pour cette dernière base, on a tiré un échantillon supplémentaire des adresses pour lesquelles plusieurs pêcheurs étaient autorisés dans la base.

Le questionnaire commun aux deux bases comportait une question sur la détention d'un permis de pêche récréative en eau salée en Caroline du Nord. Avec cette question, on voulait déterminer à quel domaine appartenaient les unités choisies dans la base des adresses. Les analystes ont toutefois observé qu'un certain nombre de répondants de la base des permis déclaraient ne pas avoir de permis, ce qui devait les mettre sur la piste d'une erreur de rattachement à un domaine. C'est pourquoi on a lancé une opération pour établir à quel domaine appartenaient réellement les unités de la base des adresses. Nous visions à un appariement intégral des adresses échantillonnées avec les données de la base des permis. En dernière étape, un préposé à l'appariement devait voir si le membre d'une adresse appariée serait le pêcheur autorisé d'après les données disponibles de la base des permis et les réponses à l'enquête. Comme une telle opération était vorace en temps, on a cherché une solution de rechange. Le but était de concevoir pour le stade opérationnel de l'enquête des méthodes de détermination de l'appartenance à un domaine réel pour seulement un sous-ensemble de l'échantillon. Comme nous avions effectivement accès au domaine réel pour tout l'échantillon, nous avons toutefois pu examiner les probabilités de défaut de classification et les moyennes de sous-domaine, de même que comparer les résultats des $\hat{\hat{Y}}_{\mathrm{CB}}$ à une estimation venant des données « réelles ».

Certaines unités observées de la base des permis commettaient des erreurs de déclaration concernant la détention d'un permis, mais cela ne causait pas d'erreur de classification de domaine pour la base des permis, parce que l'état réel de détention était connu. Dans le cas de la base des permis, un défaut de classification de domaine était possible seulement s'il était impossible de déterminer en toute précision si le ménage résidait à l'intérieur ou à l'extérieur de la Caroline du Nord. On ne peut exclure que de telles erreurs se produisent. Si on échantillonne un ménage ayant une adresse extérieure dans la base des permis et une adresse intérieure dans la base des adresses, l'attribution d'un domaine sera faussée. Nous pensions cependant que les cas seraient trop peu fréquents pour être pris en considération, aussi avons-nous traité les probabilités de défaut de classification comme si elles étaient nulles pour les unités de la base des permis dans notre analyse.

6.2 Analyse d'échantillon

Nous présentons séparément par strate au tableau 6.1 les taux de défaut de classification de domaine pour l'échantillon de la base des adresses. Dans ce cas, le domaine ab^* comprend les unités de la base des adresses qui déclarent être titulaires d'un permis, tandis que le domaine a^* comprend les unités déclarant ne pas détenir de permis. Les pêcheurs ayant déclaré ne pas en détenir présentent un taux d'erreur d'environ 5 % dans les deux strates. Les unités ayant déclaré en détenir un ont des taux d'erreur extrêmement élevés

et, à cet égard, les unités des comtés non côtiers ont plus souvent tort que raison! Nous signalons que les enquêtés de la base des adresses pour qui ces estimations sont présentées appartiennent à l'échantillon de seconde phase de cette base. Ainsi, ils ont été sélectionnés parce que leur ménage comptait au moins un membre ayant pêché dans les 12 derniers mois. C'est pourquoi une très forte proportion des intéressés étaient des pêcheurs à comparer à la proportion dans l'ensemble de la population.

Tableau 6.1
Taux de défaut de classification calculé à partir de l'échantillon entier (base des adresses, cycle 6, 2009)

	Proportion des unités se déclarant sans permis qui sont $\left(\hat{p}_{ab\mid a^*}\right)$	Proportion des unités se déclarant avec permis qui ne sont pas $\left(\hat{p}_{a\mid ab^*}\right)$
Strate côtière	0,04	0,46
Strate non côtière	0,06	0,63

Nous avons également examiné les hypothèses d'égalité des moyennes à l'aide des données de l'échantillon de la base des adresses. Le tableau 6.2 décrit le degré moyen d'activité estimé dans chacune des quatre catégories de domaine et l'appartenance à un domaine perçu. Les colonnes classent les enquêtés dans les domaines perçus et les lignes, dans les domaines réels. Le tableau indique que le comportement de pêche des répondants correspond à ce qu'ils disent être leur domaine perçu de permis par opposition à leur domaine réel. Nous croyons donc que, dans les données d'enquête sur les pêcheurs à la ligne, l'hypothèse d'égalité des moyennes de la méthode que nous proposons est plus raisonnable que l'hypothèse correspondante de Lohr.

Tableau 6.2
Moyenne estimée du nombre de déplacements de pêche (ET) par sous-domaine pour la base des adresses de Caroline du Nord dans le cycle 6 en 2009

$\bar{y}$ pour les sous-domaines	déclaration de non-détention de permis $\left(a^*\right)$	déclaration de détention de permis $\left(ab^*\right)$
déclaration vraie de non-détention de permis (a)	0,34 (0,14)	0,88 (0,41)
déclaration vraie de détention de permis (ab)	0,35 (0,46)	0,98 (0,24)

Les données de l'échantillon comportaient des poids attribués par les concepteurs de l'enquête qui rendaient compte du plan de sondage complexe et de la correction de non-réponse. Comme les probabilités de défaut de classification de domaine variaient selon les strates, nous avons repondéré séparément par strate comme nous le décrivons à la section 5 en prenant des estimations individuelles de défaut de classification pour chaque domaine de la base des adresses. Nous avons supposé que ce défaut était nul dans le cas de la base des permis. Nous avons calculé six estimations du degré d'activité de pêche au tableau 6.3 :

1) estimateur de Hartley non corrigé (*Valeur non corrigée* au tableau) : nous nous sommes reportés à l'appartenance au domaine perçu pour calculer le total avec l'estimateur de Hartley comme en (3.3);

2) estimateur en sous-échantillonnage à 20 %, 40 % et 100 % : nous avons sous-échantillonné les unités de chaque strate de l'échantillon de phase 1 de la base des adresses; nous nous sommes reportés aux domaines réels pour estimer les probabilités de défaut de classification et nous avons repondéré en fonction des probabilités estimées;
3) estimateur de Hartley corrigé (*Valeur réelle*) : nous nous sommes reportés à l'appartenance au domaine réel vérifiée à l'aide de l'opération d'appariement pour estimer le nombre total par l'estimateur de Hartley avec la pondération d'origine comme en (3.1); on doit y voir la meilleure estimation disponible, puisqu'elle n'exige aucune hypothèse quant à l'absence de biais.

La première ligne contient les cinq estimations, la deuxième une estimation du biais dans chaque cas et la troisième la racine carrée de la somme de la variance estimée et du biais au carré. Le biais à la ligne 2 est la différence entre chaque estimation et l'estimation de Hartley corrigée (colonne *Valeur réelle*). L'algorithme d'appariement des adresses est indubitablement imparfait, ce qui veut dire que l'estimateur « Valeur réelle » pourrait encore être entaché d'un biais s'ajoutant à sa variabilité d'échantillonnage. Si nous y voyons notre meilleure évaluation du biais, nous constatons que, après application de la méthode de correction de biais, le biais estimé se réduit de $211K$ à $40K - 80K$ avec l'estimateur en correction de biais. La différence entre $\hat{Y}_H$ et $\hat{\hat{Y}}_{\mathrm{CB}}$ dans un sous-échantillonnage à 100 % pourrait traduire l'insuffisance des hypothèses requises d'égalité des moyennes. Nous abaissons d'environ $70K$ la racine estimée de l'erreur quadratique moyenne (REQM) en employant une méthode à réduction de biais au lieu de l'estimateur de Hartley non corrigé.

Tableau 6.3
Total estimé des déplacements de pêche (base des adresses, cycle 6, 2009)

	Valeur non corrigée	Sous-échantillonnage à 20 % $m_A = 36$	Sous-échantillonnage à 40 % $m_A = 71$	Sous-échantillonnage à 100 %	Valeur réelle
Estimation	731 430	889 860	863 488	905 947	942 360
Biais	210 930	52 500	78 872	36 413	0
REQM	244 531	181 809	180 311	176 954	213 966

7 Analyse

Une importante observation qui se dégage de cette application est que les enquêtés pourraient éprouver de la difficulté à déclarer en toute précision l'appartenance à un domaine, même lorsque celle-ci se définit par une simple notion comme celle de la détention d'un permis de pêche. Cela pose un problème tout particulièrement dans le cas d'une estimation à double base de sondage. Quand l'appartenance à un domaine peut être tirée d'autres sources à un coût infime, la solution est toute trouvée, mais quand le coût est élevé, il peut être bon de recourir à une méthode de correction de biais. Une autre observation importante pour notre application est que la seule méthode déjà disponible dont nous avions connaissance exigeait comme hypothèse tacite que les unités soient homogènes à l'intérieur des domaines réels. On pourrait y voir une

hypothèse naturelle et anodine pour bien des applications, mais elle semblait ne pas se vérifier pour la question clé de notre questionnaire.

Dans cette étude, ce qui expliquait l'erreur de rattachement à un domaine était que le répondant était incapable ou peu désireux de donner le bon renseignement recherché. Il se pouvait par ailleurs que la base elle-même accuse des erreurs. Dans l'un et l'autre cas, cerner le mécanisme à l'origine de l'erreur pouvait aider à juger de l'hypothèse relative aux moyennes qui était la plus vraisemblable. Dans notre application, il semble évident rétrospectivement que quelqu'un qui admet aller fréquemment à la pêche hésiterait à avouer le faire sans permis à un organisme gouvernemental faisant enquête.

Nombreuses sont les applications de l'estimation à double base de sondage qui visent à améliorer l'efficacité plutôt que la couverture, comme dans notre application où nous avons inclus la base des permis pour réduire le coût de prise de contact avec les pêcheurs à la ligne. Dans ce cas, l'appartenance à l'une des bases de sondage permet probablement de prévoir les variables de réponse clés de l'enquête et, par conséquent, les moyennes des réponses peuvent varier amplement selon les sous-groupes de la population. Ce n'est pas toujours le cas cependant. Si l'une ou l'autre des bases a directement à voir avec le thème de l'enquête même, on peut estimer plus probable que le domaine réel détermine la réponse moyenne, voire que les quatre sous-groupes de la population présentent la même moyenne. (Dans ce dernier cas, aucune des deux méthodes de correction de biais ne serait à exclure.) Supposons que les deux bases de la présente enquête comprennent les téléphones terrestres et les téléphones cellulaires et qu'on demande, par exemple, à un enquêté échantillonné dans la base des téléphones cellulaires s'il dispose d'un téléphone terrestre de manière à établir s'il appartient au domaine en chevauchement. On peut s'attendre à ce que les réponses à cette question soient entachées d'une certaine erreur de mesure, mais il serait improbable que le fait que le répondant *dise* qu'il a un téléphone terrestre soit plus révélateur de sa motivation à pêcher à la ligne que le fait qu'il *dispose* réellement d'un téléphone terrestre. (En fait, l'aspect de la possession réelle aurait probablement plus à voir avec la propension à la pêche, parce que ces deux facteurs sont en corrélation avec l'âge.) On aura intérêt à consulter des experts du thème de l'enquête pour prévoir quelle hypothèse se vérifiera relativement aux moyennes, mais il sera nécessaire à la fin d'examiner les données de l'enquête même avant de prendre une décision en matière de correction de biais.

Le coût d'une correction de biais est une augmentation de la variance. La pénalité est importante, surtout là où il n'y a guère d'information disponible sur les taux de défaut de classification. S'il est possible d'établir que les erreurs de rattachement sont peu importantes, soit que les domaines exposés à l'erreur ne représentent qu'une légère fraction de la population soit que le taux d'erreur demeure très bas, la correction de biais pourrait alors ne pas en valoir la peine. Calculer des estimateurs en correction de biais est simple avec un logiciel d'enquête une fois qu'on dispose d'estimations sur le défaut de classification. Les estimations de variance correspondantes, jointes à la différence entre les estimations mêmes avec et sans correction de biais, peuvent éclairer notre choix.

Dans notre recherche, nous nous sommes donné comme contexte l'estimateur à double base de sondage de Hartley, parce que c'était là l'estimateur utilisé dans notre application. Un grand nombre d'estimateurs du genre sont disponibles, et tous exigent une connaissance de l'appartenance au domaine. Il serait possible

d'adapter d'autres estimateurs comme celui de Fuller-Burmeister à l'aide de méthodes comme celles que nous avons décrites. Pour d'autres comme les estimateurs de pseudo-maximum de vraisemblance, il faudrait envisager une méthode différente.

Bibliographie

Andrews, R., Brick, J.M., Mathiowetz, N. et Stokes, L. (2010). Pilot test of a dual frame two-phase mail survey of anglers in North Carolina. Rapport final de la National Oceanic and Atmospheric Administration.

Bankier, M. (1986). Estimators based on several stratied samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Fuller, W., et Burmeister, L. (1972). Estimators for samples selected from two overlapping frames. Dans *Proceedings of the Social Statistics Section,* American Statistical Association, 245-249.

Hartley, H. (1962). Multiple frame surveys. Dans *Proceedings of the Social Statistics Section,* American Statistical Association, 203-206.

Kalton, G., et Anderson, D. (1986). Pilot test of a dual frame two-phase mail survey of anglers in North Carolina. *Journal of the Royal Statistical Society, Series A*, 149, 65-82.

Lin, D. (2014). Dissertation: *Measurement Error in Dual Frame Estimation*.

Lohr, S. (2011). Autres plans de sondage : échantillonnage avec bases de sondage multiples chevauchantes. *Techniques d'enquête*, 37, 2, 213-232. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11608-fra.pdf.

National Research Council (NRC) (2006). Review of recreational sheries survey methods. Rapport technique.

Skinner, C., et Rao, J. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.