Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Autres plans de sondage : échantillonnage avec bases de sondage multiples chevauchantes

par Sharon L. Lohr

Décembre 2011

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2011
Vol. 37, N° 2, pp. 213-232
Statistique Canada, N° 12-001-X au catalogue

Autres plans de sondage : échantillonnage avec bases de sondage multiples chevauchantes

Sharon L. Lohr [1]

Résumé

Les plans de sondage et les estimateurs des enquêtes à base de sondage unique utilisés à l'heure actuelle par les organismes gouvernementaux américains ont été élaborés en réponse à des problèmes pratiques. Les programmes d'enquêtes-ménages fédéraux doivent faire face aujourd'hui à la diminution des taux de réponse et de la couverture des bases de sondage, à la hausse des coûts de collecte des données et à l'accroissement de la demande de statistiques pour des petits domaines. Les enquêtes à bases de sondage multiples, dans lesquelles des échantillons indépendants sont tirés de bases de sondage distinctes, peuvent être utilisées en vue de relever certains de ces défis. La combinaison d'une liste et d'une base de sondage aréolaire ou l'utilisation de deux bases de sondage pour échantillonner les ménages ayant une ligne de téléphone fixe et ceux ayant une ligne de téléphone mobile en sont des exemples. Nous passons en revue les estimateurs ponctuels et les ajustements de la pondération qui peuvent être utilisés pour analyser les données d'enquête à bases de sondage multiples au moyen de logiciels standard et nous résumons la construction des poids de rééchantillonnage pour l'estimation de la variance. Étant donné leur complexité croissante, les enquêtes à bases de sondage multiples obligent à résoudre des difficultés qui ne se posent pas dans le cas des enquêtes à base de sondage simple. Nous étudions le biais dû à l'erreur de classification dans les enquêtes à bases de sondage multiples et proposons une méthode pour corriger ce biais quand les probabilités d'erreur de classification sont connues. Enfin, nous discutons des travaux de recherche nécessaires en ce qui concerne les erreurs non dues à l'échantillonnage dans les enquêtes à bases de sondage multiples.

Mots clés : Correction du biais ; enquête à deux bases de sondage ; erreur de classification ; effets de mode ; échantillonnage d'événements rares ; poids d'échantillonnage ; estimation sur petits domaines.

1. Utilisation d'enquête à bases de sondage multiples

En théorie classique de l'échantillonnage fondé sur un plan de sondage, un échantillon probabiliste est tiré de la base de sondage (unique) et les probabilités d'inclusion dans le plan de sondage peuvent être utilisées pour faire des inférences au sujet de la population. Soit y_i une mesure sur l'unité i dans la population de N unités, soit $\mathcal{S}$ l'ensemble d'unités dans l'échantillon, et soit $\pi_i = P$ (l'unité i est incluse dans l'échantillon). Alors, l'estimateur de Horvitz-Thompson (1952) du total de population $Y = \sum_{i=1}^{N} y_i$ est $\hat{Y} = \sum_{i \in \mathcal{S}} w_i y_i$, où $w_i = 1/\pi_i$ est le poids de sondage. Si la base de sondage englobe toutes les unités faisant partie de la population cible, que toutes les unités échantillonnées répondent et qu'il n'y a pas d'erreur de mesure, l'estimateur de Horvitz-Thompson est sans biais pour Y.

Les difficultés pratiques posées par l'échantillonnage durant les années 1940 et les années 1950 ont mené les méthodologistes à élaborer des sondages et des estimateurs stratifiés à plusieurs degrés tels que l'estimateur de Horvitz-Thompson. Les enquêtes avec interview sur place s'appuyaient sur un échantillonnage avec probabilités inégales pour équilibrer les charges de travail des intervieweurs et réduire les variances. Les taux de réponse étaient élevés dans le cas de nombreuses enquêtes gouvernementales, de sorte que toutes les hypothèses qui sous-tendent l'estimateur de Horvitz-Thompson étaient raisonnables. Aujourd'hui, nous devons faire face à de nouvelles difficultés dans les enquêtes-ménages. Les taux de non-réponse augmentent, si bien que les estimations par sondage doivent s'appuyer davantage sur des modèles. La diversité ethnique et linguistique d'une population peut donner lieu à un sous-dénombrement et à une erreur de mesure. La diversité technologique croissante signifie qu'il est parfois préférable de recourir à des modes d'échantillonnage différents pour divers types de résidents ; on doit alors être certain que les divers modes d'échantillonnage mesurent les mêmes quantités. Les coûts de la collecte des données ont grimpé fortement, en partie à cause de l'accroissement de la non-réponse ; parallèlement, les demandes de données des organismes gouvernementaux et des chercheurs ont également augmenté considérablement.

Les enquêtes à bases de sondage multiples permettent d'obtenir une meilleure couverture de la population à un coût plus faible. Elles peuvent être utilisées comme un élément structurel d'un plan de sondage modulaire qui s'appuie sur différentes bases de sondage pour essayer de réduire les coûts et accroître la couverture. Elles permettent aussi d'utiliser efficacement les données administratives. Dans le présent article, nous décrivons divers types d'enquêtes à bases de sondage multiples et discutons de certains travaux de recherche achevés et de ceux qui pourraient être nécessaires pour que l'on puisse utiliser ces enquêtes.

1. Sharon L. Lohr, School of Mathematical and Statistical Sciences, Arizona State University, Tempe AZ 85287-1804. Courriel : sharon.lohr@asu.edu.

L'une des premières enquêtes à bases de sondage multiples (à part les premières méthodes de capture-recapture) a été menée par le Census Bureau en 1949 (Hansen, Hurwitz et Madow 1953). Dans la Sample Survey of Retail Stores, un échantillon probabiliste d'unités primaires d'échantillonnage (UPE) a été sélectionné. Dans chaque UPE, une liste des grandes entreprises de commerce de détail a été établie d'après les dossiers de l'Old Age and Survivors Insurance Bureau. Toutes les entreprises figurant sur la liste ont été incluses dans l'échantillon et un échantillon aréolaire des UPE ne figurant pas sur la liste a été sélectionné. Dans ce cas, un plan à deux bases de sondage avec *sélection* a été employé dans chaque UPE sélectionnée ; les unités figurant sur la liste ont été retirées de la base aréolaire avant l'échantillonnage. Donc, dans chaque UPE, l'estimateur du total des ventes était égal à la somme des deux estimateurs. Aucune nouvelle méthode statistique n'était nécessaire pour estimer le total des ventes dans cette enquête, puisqu'essentiellement, un échantillon stratifié était tiré dans chaque UPE : les entreprises des UPE reprises sur la liste formaient une strate et celles figurant dans la base aréolaire mais non sur la liste formaient la deuxième strate. Cette méthode d'enquête a permis de réaliser des économies, parce qu'il était relativement peu coûteux d'échantillonner les entreprises sur la liste, tout en obtenant une couverture complète grâce à l'utilisation de la base aréolaire.

De nombreuses enquêtes agricoles se sont également appuyées sur un plan à deux bases de sondage avec sélection (González-Villalobos et Wallace 1996). Dans ce type de plan, les fermes figurant sur la liste sont retirées de la base aréolaire avant que l'échantillonnage ne commence. Des économies importantes peuvent être réalisées, car l'échantillonnage d'après la liste est souvent nettement moins coûteux et que la liste contient les plus grandes entreprises.

Toutefois, dans de nombreux cas, il n'est pas forcément possible ni pratique de supprimer de la base aréolaire avant l'échantillonnage les unités qui figurent sur la liste. Au lieu de cela, dans une enquête à deux bases de sondage chevauchantes, des échantillons probabilistes indépendants sont tirés de la base de sondage A (la base de sondage aréolaire) et de la base de sondage B (la liste), comme l'illustre la figure 1. Les populations rares peuvent souvent être échantillonnées plus efficacement en se servant d'un échantillon à bases de sondage multiples (Kalton et Anderson 1986). Par exemple, dans une étude épidémiologique, la base de sondage A pourrait être utilisée pour une enquête générale sur la santé de la population, tandis que la base de sondage B pourrait être une liste des cliniques qui se spécialisent dans une maladie particulière. L'échantillon tiré de la base de sondage B devrait contenir un pourcentage élevé de personnes atteintes de la maladie d'intérêt de sorte que cet échantillonnage sera efficace ; l'échantillon tiré de la base de sondage A, quoique plus coûteux, donne une couverture complète de la population.

Dans d'autres situations, toutes les bases de sondage sont incomplètes, comme l'a considéré Hartley (1962) ; par exemple, à la figure 2, la base de sondage A pourrait être une liste de numéros de téléphone fixe et la base de sondage B, une liste de numéros de téléphone mobile. Il existe trois domaines : le domaine *a* contient les unités figurant dans la base de sondage A mais non dans la base de sondage B, le domaine *b* contient les unités figurant dans la base de sondage B mais non dans la base de sondage A, et le domaine *ab* contient les unités figurant dans les deux bases de sondage. Dans le contexte téléphonique, le domaine *a* contient les individus appartenant à un ménage ne possédant qu'un téléphone fixe, le domaine *b* contient les individus ne possédant qu'un téléphone mobile et le domaine *ab* comprend les individus qui ont à la fois un téléphone mobile et un téléphone fixe. On ne sait pas d'avance si un membre du ménage échantillonné en utilisant l'une des bases de sondage appartient aussi à l'autre (Brick, Dipko, Presser, Tucker et Yuan 2006) ; habituellement, on demande aux enquêtés des renseignements sur leur utilisation d'un téléphone mobile et d'un téléphone fixe pour déterminer le domaine auquel ils appartiennent.

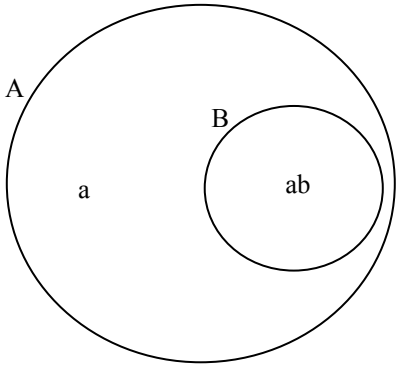

Figure 1 Un plan à deux bases de sondage dans lequel la base de sondage B est un sous-ensemble de la base de sondage A

Plus de deux bases de sondage peuvent être utilisées également, comme l'illustre la figure 3 pour une enquête à trois bases de sondage dans laquelle toutes les bases de sondage sont incomplètes. Dans cette situation, il existe sept domaines. Iachan et Dennis (1993) ont donné un exemple d'enquête à trois bases de sondage utilisée pour échantillonner la population de sans-abri, où la base de sondage A est une liste des soupes populaires, la base de sondage B est une liste des refuges et la base de sondage C correspond à des emplacements de rue. La figure 4 représente une enquête à trois bases de sondage dans laquelle la base de sondage A donne une couverture complète, tandis

que les bases de sondage chevauchantes B et C sont toutes deux incomplètes, mais moins coûteuses à échantillonner. Ce plan de sondage a été utilisé aux États-Unis pour les enquêtes du *Scientists and Engineers Statistical Data System* (SESTAT ; *National Science Foundation* 2003). Le même plan de sondage pourrait être utilisé quand A est la base de sondage pour une enquête de population générale, B est une enquête auprès des possesseurs d'un téléphone fixe et C, l'enquête auprès des possesseurs d'un téléphone mobile.

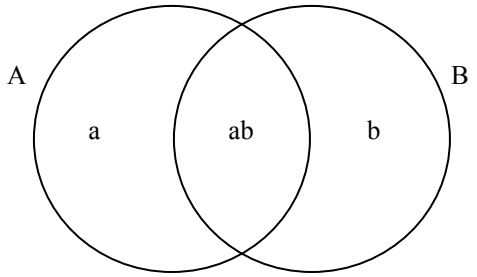

Figure 2 Les bases de sondage A et B se chevauchent, créant les trois domaines *a, b* et *ab*

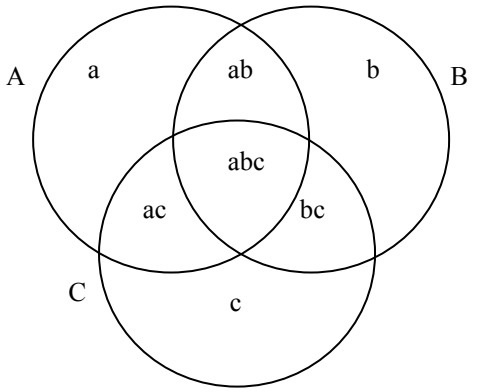

Figure 3 Les bases de sondage A, B et C sont toutes les trois incomplètes et se chevauchent

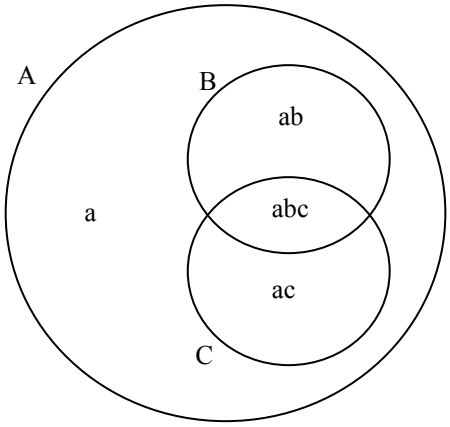

Figure 4 La base de sondage A contient l'entièreté de la population ; les bases de sondage B et C se chevauchent et sont toutes les deux contenues dans la base de sondage A

L'utilisation de plans à bases de sondage multiples dans les enquêtes-ménages offre de nombreuses possibilités, dont :

1. l'utilisation de bases de sondage multiples fondées sur des dossiers administratifs ;
2. l'échantillonnage à modes multiples (par exemple, le tirage d'échantillons indépendants à partir d'une liste de numéros de téléphone mobile et à partir d'une liste de numéros de téléphone fixe) ;
3. l'utilisation future d'Internet pour la collecte des données. Quoique Internet pose de nombreux problèmes de couverture et de spécification des domaines, son utilisation mérite d'être envisagée, en raison des économies possibles et de la facilité de recueillir et de traiter les données ;
4. l'amélioration de l'estimation sur petits domaines. Une enquête nationale peut être complétée par des enquêtes locales, plus petites, pour obtenir des estimations de plus grande précision dans les petits domaines ;
5. l'amélioration de l'estimation pour des populations rares. Une enquête de population générale peut être complétée par une enquête auprès d'un échantillon tiré d'une base de sondage dans laquelle la concentration des membres d'une population rare est élevée ;
6. l'emploi de plans de sondage modulaires. Une approche à bases de sondage multiples peut donner plus de souplesse pour la conception des enquêtes permanentes. À mesure que l'échantillonnage à partir de bases de sondage particulières devient moins cher, la répartition relative de la taille de l'échantillon entre les diverses bases de sondage peut être modifiée. L'approche modulaire donne aussi plus de souplesse pour répondre à l'évolution des besoins de données.

La plus grande souplesse des enquêtes à bases de sondages multiples a toutefois pour prix l'accroissement de la complexité. L'information provenant des diverses enquêtes doit être combinée pour estimer les quantités de population et les estimateurs que l'on peut choisir sont nombreux. À la section 2, nous résumons les estimateurs qui ont été élaborés pour les totaux de population et décrivons comment les poids de sondage sont modifiés dans ces estimateurs ; aux sections 3 et 4, nous discutons du calage des poids et décrivons comment utiliser les progiciels d'analyse de données d'enquête avec des données provenant d'enquêtes à bases de sondage multiples. Les erreurs non dues à l'échantillonnage doivent être prises en considération en ce qui concerne leur effet dans chaque base de sondage séparément ainsi que leur effet sur les estimations calculées

d'après l'information combinée. À la section 5, nous discutons des effets de la non-réponse et des effets de mode dans les enquêtes à bases de sondage multiples.

En plus des problèmes de non-réponse, de sous-dénombrement et d'erreurs de mesure qui se posent dans le cas des enquêtes à base de sondage unique, les enquêtes à bases de sondage multiples peuvent poser le problème de la classification incorrecte dans les domaines. La modification des poids pour les estimateurs de la section 2 dépend de l'appartenance des observations aux domaines. S'il est probable que certaines observations appartenant au domaine a soient enregistrées par erreur comme appartenant au domaine ab, les estimateurs pourraient présenter un biais important. Nous étudions les effets de l'erreur de classification dans les domaines à la section 6 et proposons une nouvelle méthode pour corriger le biais dû à l'erreur de classification quand les probabilités d'erreur de classification sont connues. Enfin, à la section 7, nous examinons les questions concernant le plan de sondage et à la section 8, nous discutons des possibilités et des défis associés aux enquêtes à bases de sondage multiples.

2. Estimateurs dans les enquêtes à bases de sondage multiples chevauchantes

À la présente section, nous passons en revue les estimateurs du total de population Y pour les enquêtes à bases de sondage multiples chevauchantes, ainsi que les modifications de la pondération induite par ces estimateurs. Pour simplifier la notation, nous nous concentrons sur une enquête à deux bases de sondage à la section 2.1 et nous décrivons les extensions à des enquêtes à bases de sondage multiples à la section 2.2. Dans une enquête à deux bases de sondage, nous pouvons écrire

$$Y = Y_a + Y_{ab} + Y_b,$$

où Y_a est le total des unités de la population dans le domaine a, Y_{ab} est le total des unités de la population dans le domaine ab, et Y_b est le total des unités de la population dans le domaine b. Un cas particulier est l'estimation de la taille de population $N = N_a + N_{ab} + N_b$, dont il est question dans Haines et Pollock (1998). Nous discutons de l'estimation d'autres quantités de population que les totaux et les moyennes, et utilisons les données provenant d'enquête à bases de sondage multiples dans d'autres analyses, à la section 4.

Nous commençons par énoncer certaines propriétés souhaitables des estimateurs pour enquêtes à bases de sondage multiples.

1. Un estimateur devrait être approximativement sans biais pour la quantité de population finie correspondante.
2. Les estimateurs devraient avoir la propriété de cohérence interne : autrement dit, si $\hat{Y}_1$ estime le nombre de femmes ingénieures dans la population, $\hat{Y}_2$ estime le nombre d'hommes ingénieurs dans la population et $\hat{Y}_3$ estime le nombre total d'ingénieurs dans la population, nous devrions avoir $\hat{Y}_1 + \hat{Y}_2 = \hat{Y}_3$. La cohérence interne préserve les relations multivariées dans les données. Pratiquement parlant, la cohérence interne requiert qu'un seul ensemble de poids soit utilisé pour toutes les estimations.
3. Un estimateur devrait être efficace, c'est-à-dire avoir une faible erreur quadratique moyenne.
4. Un estimateur devrait avoir une forme permettant de le calculer au moyen des progiciels standards tels que SUDAAN ou SAS PROC SURVEYMEANS. Cela permet aux analystes de travailler avec les données sans devoir écrire ou tester de nouveaux logiciels. Du point de vue pratique, un seul fichier de données est créé pour l'enquête à bases de sondage multiples. Le fichier comprend une colonne de poids qu'il faut utiliser pour calculer les estimations ponctuelles, et il contient soit des variables décrivant les plans de sondage pour l'estimation de la variance fondée sur une formule, ou des colonnes de poids de rééchantillonnage pour l'estimation de la variance par les méthodes de rééchantillonnage.
5. Un estimateur devrait, dans la mesure du possible, être robuste aux erreurs non dues à l'échantillonnage qui pourraient survenir dans les enquêtes à bases de sondage multiples.

2.1 Estimateurs et ajustement de la pondération pour les enquêtes à deux bases de sondage

Considérons l'enquête à deux bases de sondage chevauchantes illustrée à la figure 2, où le domaine ab n'est pas vide. Un échantillon probabiliste $\mathcal{S}(A)$ de taille n_A est tiré des N_A unités dans la base de sondage A, et un échantillon probabiliste indépendant $\mathcal{S}(B)$ de taille n_B est tiré des N_B unités de la base de sondage B. L'unité i dans l'échantillon $\mathcal{S}(A)$ possède la probabilité d'inclusion π_i^A et le poids w_i^A, et l'unité i dans l'échantillon $\mathcal{S}(B)$ possède la probabilité d'inclusion π_i^B et le poids w_i^B. Les poids peuvent être l'inverse des probabilités d'inclusion, ou ils peuvent être poststratifiés afin qu'ils concordent avec les chiffres de population ; nous supposons que les estimateurs des totaux de population sont approximativement sans biais.

Alors, $E[\sum_{i\in\mathcal{S}(A)} w_i^A y_i] \approx Y_a + Y_{ab}$ et $E[\sum_{i\in\mathcal{S}(B)} w_i^B y_i] \approx Y_b + Y_{ab}$. Par conséquent, un estimateur qui combine les observations provenant des deux enquêtes avec les poids originaux, $\sum_{i\in\mathcal{S}(A)} w_i^A y_i + \sum_{i\in\mathcal{S}(B)} w_i^B y_i$, est biaisé pour le total de population Y. Si les moyennes de domaines diffèrent, l'estimateur correspondant de la moyenne de population peut également être biaisé.

Les divers estimateurs du total de population Y qui ont été proposés dans la littérature modifient la pondération de façon que les estimateurs soient approximativement sans biais. Les poids modifiés, montrés plus bas pour les divers estimateurs, sont de la forme $\tilde{w}_i^A = m_i^A w_i^A$ et $\tilde{w}_i^B = m_i^B w_i^B$. Le total de population est alors estimé par

$$\hat{Y} = \sum_{i\in\mathcal{S}(A)} \tilde{w}_i^A y_i + \sum_{i\in\mathcal{S}(B)} \tilde{w}_i^B y_i \qquad (1)$$

et la moyenne de population $\bar{Y}$ est estimée par $\hat{\bar{Y}} = \hat{Y} / \hat{N}$ où

$$\hat{N} = \sum_{i\in\mathcal{S}(A)} \tilde{w}_i^A + \sum_{i\in\mathcal{S}(B)} \tilde{w}_i^B.$$

Les estimateurs seront alors approximativement sans biais si $m_i^A \approx 1$ pour $i \in a$, $m_i^B \approx 1$ pour $i \in b$, et $m_i^A + m_i^B \approx 1$ pour $i \in ab$. Tous les estimateurs passés en revue à la présente section satisfont les critères nécessaires pour l'absence approximative de biais en l'absence d'erreurs non dues à l'échantillonnage (voir Lohr 2009).

Ajustements des poids fixes. La modification la plus simple des poids pour préserver l'absence approximative de biais, décrite par Hartley (1962), prend la forme

$$m_{i,\theta}^A = \begin{cases} 1 & \text{si } i \in a \\ \theta & \text{si } i \in ab, \end{cases} \quad m_{i,\theta}^B = \begin{cases} 1 & \text{si } i \in b \\ 1-\theta & \text{si } i \in ab, \end{cases} \qquad (2)$$

où $\theta \in [0, 1]$. En utilisant les poids modifiés $\tilde{w}_i^A = m_{i,\theta}^A w_i^A$ et $\tilde{w}_i^B = m_{i,\theta}^B w_i^B$ dans (1), l'estimateur résultant $\hat{Y}(\theta)$ peut aussi être exprimé en se servant des totaux de domaines estimés $\hat{Y}_a^A = \sum_{i\in\mathcal{S}(A), i\in a} w_i^A y_i$, $\hat{Y}_{ab}^A = \sum_{i\in\mathcal{S}(A), i\in ab} w_i^A y_i$, $\hat{Y}_{ab}^B = \sum_{i\in\mathcal{S}(B), i\in ab} w_i^B y_i$, et $\hat{Y}_b^B = \sum_{i\in\mathcal{S}(B), i\in b} w_i^B y_i$. L'estimateur

$$\begin{aligned}\hat{Y}(\theta) &= \sum_{i\in\mathcal{S}(A)} m_{i,\theta}^A w_i^A y_i + \sum_{i\in\mathcal{S}(B)} m_{i,\theta}^B w_i^B y_i \\ &= \hat{Y}_a^A + \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B \qquad (3)\end{aligned}$$

estime donc le total de domaine Y_{ab} par une moyenne pondérée de l'estimateur pour la base de sondage A, $\hat{Y}_{ab}^A$, et de l'estimateur pour la base de sondage B, $\hat{Y}_{ab}^B$.

Pour une valeur fixée de θ, l'estimateur $\hat{Y}(\theta)$ assure la cohérence interne, puisque l'on utilise le même ensemble de poids ajustés pour toutes les variables. Cet estimateur est simple à utiliser et à implémenter. Son efficacité dépend de la valeur choisie pour θ. Brick et coll. (2006) ont utilisé $\theta = 1/2$ dans leur étude d'une enquête à deux bases de sondage dans laquelle la base de sondage A est une liste de numéros de téléphone fixe et la base de sondage B, une liste de numéros de téléphone mobile, et la valeur de $\theta = 1/2$ est souvent recommandée (voir, par exemple, Mecatti 2007). Quand $\theta = 0$ ou 1, les données dans le domaine de chevauchement provenant de l'un des échantillons sont écartées et l'enquête devient une enquête à deux bases de sondage avec sélection.

Estimateurs optimaux. Hartley (1962, 1974) a proposé de choisir θ dans (3) de manière que la variance de $\hat{Y}(\theta)$ soit minimisée. La valeur de θ donnant le résultat optimal est

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

Puisque les variances et les covariances sont généralement inconnues, elles doivent être estimées d'après les données, ce qui donne

$$\hat{\theta}_H = \frac{\hat{V}(\hat{Y}_{ab}^B) + \widehat{\mathrm{Cov}}(\hat{Y}_b^B, \hat{Y}_{ab}^B) - \widehat{\mathrm{Cov}}(\hat{Y}_a^A, \hat{Y}_{ab}^A)}{\hat{V}(\hat{Y}_{ab}^A) + \hat{V}(\hat{Y}_{ab}^B)}.$$

Skinner et Rao (1996) ont montré que l'estimateur de Hartley peut être calculé en utilisant des poids ajustés. Les modifications des poids pour l'estimateur de Hartley $\hat{Y}(\hat{\theta}_H)$ sont données par (2), en substituant $\hat{\theta}_H$ à θ. Puisque $\hat{\theta}_H$ est convergent pour θ_H, l'estimateur de Hartley est asymptotiquement optimal parmi les estimateurs de la forme $\hat{Y}_a^A + \hat{Y}_b^B + \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B$. Cependant, les poids modifiés $\tilde{w}_{i,H}^A$ et $\tilde{w}_{i,H}^B$ sont des fonctions des variances et des covariances des totaux de domaines estimés. Cela a deux conséquences : 1) les poids modifiés sont des variables aléatoires et leur variabilité doit être prise en compte dans les erreurs-types des estimateurs et 2) les modifications optimales des poids varieront pour différentes variables de réponse, ce qui donnera lieu à une incohérence interne.

Fuller et Burmeister (1972) ont proposé de modifier l'estimateur de Hartley en se servant d'information additionnelle au sujet de N_{ab}, ce qui donne

$$\hat{Y}_{\mathrm{FB}}(\beta) = \hat{Y}_a^A + \hat{Y}_b^B + \beta_1\hat{Y}_{ab}^A + (1-\beta_1)\hat{Y}_{ab}^B + \beta_2(\hat{N}_{ab}^A - \hat{N}_{ab}^B).$$

Comme dans le cas de l'estimateur de Hartley, les valeurs optimales $\beta_{1\mathrm{opt}}$ et $\beta_{2\mathrm{opt}}$ sont choisies de manière à minimiser la variance de $\hat{Y}_{\mathrm{FB}}(\boldsymbol{\beta})$, et sont donc des fonctions des covariances des totaux de domaines. En substituant les estimateurs convergents $\hat{\beta}_{1\mathrm{opt}}$ et $\hat{\beta}_{2\mathrm{opt}}$, on obtient les ajustements des poids pour w_i^A et w_i^B. Lohr et Rao (2000) ont montré que l'estimateur de Fuller-Burmeister $\hat{Y}_{\mathrm{FB}}$ est celui qui, de tous les estimateurs considérés, possède la plus petite variance asymptotique. Comme dans le cas de l'estimateur de Hartley, cependant, les poids modifiés sont des variables aléatoires qui diffèrent pour des réponses différentes et, dans les plans d'échantillonnage complexes, l'estimateur de Fuller-Burmeister manque aussi de cohérence interne.

Estimateurs du pseudo-maximum de vraisemblance (*PMV*). Afin d'obtenir la cohérence interne, Skinner et Rao (1996) ont proposé un estimateur du pseudo-maximum de vraisemblance (PMV) qui utilise les mêmes poids pour toutes les variables. Quand N_{ab} est inconnu, il est estimé par $\hat{N}_{ab}^{\mathrm{PMV}}(\theta)$, qui est la plus petite des racines de l'équation quadratique

$$\left[\frac{\theta}{N_B} + \frac{1-\theta}{N_A}\right] x^2 - \left[1 + \theta\,\frac{\hat{N}_{ab}^A}{N_B} + (1-\theta)\,\frac{\hat{N}_{ab}^B}{N_A}\right] x + \theta \hat{N}_{ab}^A + (1-\theta)\,\hat{N}_{ab}^B = 0.$$

Skinner et Rao (1996) ont proposé d'utiliser pour θ la valeur θ_P qui minimise la variance asymptotique de $\hat{N}_{ab}^{\mathrm{PMV}}(\theta)$:

$$\theta_P = \frac{N_a\, N_B\, V(\hat{N}_{ab}^B)}{N_a\, N_B\, V(\hat{N}_{ab}^B) + N_b\, N_A\, V(\hat{N}_{ab}^A)}. \qquad (4)$$

En substituant un estimateur $\hat{\theta}_P$ à θ_P, les ajustements des poids sont :

$$m_{i,P}^A = \begin{cases} \dfrac{N_A - \hat{N}_{ab}^{\mathrm{PMV}}(\hat{\theta}_P)}{\hat{N}_a^A} & \text{si } i \in a \\[2ex] \dfrac{\hat{N}_{ab}^{\mathrm{PMV}}(\hat{\theta}_P)}{\hat{\theta}_P \hat{N}_{ab}^A + (1-\hat{\theta}_P)\,\hat{N}_{ab}^B}\,\hat{\theta}_P & \text{si } i \in ab, \end{cases}$$

$$m_{i,P}^B = \begin{cases} \dfrac{N_B - \hat{N}_{ab}^{\mathrm{PMV}}(\hat{\theta}_P)}{\hat{N}_b^B} & \text{si } i \in b \\[2ex] \dfrac{\hat{N}_{ab}^{\mathrm{PMV}}(\hat{\theta}_P)}{\hat{\theta}_P \hat{N}_{ab}^A + (1-\hat{\theta}_P)\,\hat{N}_{ab}^B}\,(1-\hat{\theta}_P) & \text{si } i \in ab. \end{cases}$$

Si la valeur de θ_P ne peut pas être estimée, par exemple si les deux bases de sondage coïncident ou que le plan de la figure 1 est utilisé, on peut se servir d'un effet de plan moyen pour les deux enquêtes dans l'ajustement, comme l'ont décrit Lohr et Rao (2006). L'estimateur PMV est intérieurement cohérent ; même s'il n'est pas garanti qu'il donne l'erreur quadratique moyenne la plus petite, il est d'une très grande efficacité dans de nombreuses situations d'enquête.

Estimateurs à base de sondage unique. Bankier (1986) ainsi que Kalton et Anderson (1986) ont proposé des estimateurs de la forme (1) qui traitent toutes les observations comme si elles avaient été échantillonnées à partir d'une seule base de sondage, en fondant les poids ajustés dans le domaine d'intersection sur les probabilités d'inclusion pour chaque base de sondage. Les ajustements des poids pour l'estimateur à base de sondage unique de Kalton et Anderson (1986) sont :

$$m_{i,S}^A = \begin{cases} 1 & \text{si } i \in a \\ w_i^B \,/\, (w_i^A + w_i^B) & \text{si } i \in ab, \end{cases}$$

$$m_{i,S}^B = \begin{cases} 1 & \text{si } i \in b \\ w_i^A \,/\, (w_i^A + w_i^B) & \text{si } i \in ab. \end{cases}$$

Si $w_i^A = 1/\pi_i^A$ et $w_i^B = 1/\pi_i^B$, l'estimateur à base de sondage unique utilise $\tilde{w}_{i,S}^A = \tilde{w}_{i,S}^B = 1/(\pi_i^A + \pi_i^B)$ pour les unités dans ab. L'ajustement des poids dans le domaine ab s'appuie à la fois sur π_i^A et π_i^B. Donc, si l'on tire un échantillon aléatoire stratifié disproportionné de la base de sondage B, on doit connaître pour les unités échantillonnées dans $\mathcal{S}(A)$ la situation d'appartenance à la strate de la base de sondage B. Les poids ajustés pour l'estimateur à base de sondage unique peuvent être interprétés en fonction des probabilités d'inclusion des unités échantillonnées. Si les fractions d'échantillonnage sont faibles, $\tilde{w}_{i,S}^A$ est approximativement égale à $1/P$ (l'unité i est incluse dans l'un des échantillons). Si $\mathcal{S}(A)$ et $\mathcal{S}(B)$ sont tous deux autopondérés, l'estimateur à base de sondage unique se réduit à (3).

Les modifications des poids pour base de sondage unique sont les mêmes pour toutes les variables de réponse, de sorte que les estimateurs ont la propriété de cohérence interne. Toutefois, pour les enquêtes complexes, les estimateurs à base de sondage unique pourraient ne pas être aussi efficaces que les estimateurs optimaux ou les estimateurs PMV. Leur performance peut être améliorée par ratissage (*raking*) vers les totaux de population de la base de sondage (Skinner 1991).

Estimateurs de la pseudo-vraisemblance empirique (*PVE*). Rao et Wu (2010) ont proposé des estimateurs de la vraisemblance empirique pour les enquêtes à deux bases de sondage. En utilisant $\theta = \theta_P$, la fonction de pseudo-vraisemblance empirique (PVE) est définie par

$$\ell(\mathbf{p}_a, \mathbf{p}_{ab}^A, \mathbf{p}_{ab}^B, \mathbf{p}_b) = \frac{n_A + n_B}{N}\left[\sum_{i\in\mathcal{S}(A),\, i\in a} \frac{N_a}{\hat{N}_a}\, w_i^A \log(p_{ai}) + \sum_{i\in\mathcal{S}(A),\, i\in ab} \frac{\theta_P N_{ab}}{\hat{N}_{ab}^A}\, w_i^A \log(p_{abi}^A) + \sum_{i\in\mathcal{S}(B), i\in b} \frac{N_b}{\hat{N}_b}\, w_i^B \log(p_{bi}) + \sum_{i\in\mathcal{S}(B), i\in ab} \frac{(1-\theta_P) N_{ab}}{\hat{N}_{ab}^B}\, w_i^B \log(p_{abi}^B)\right],$$

où θ_P est donné par (4). Un estimateur $\hat{\theta}_P$ est substitué à θ_P si ce dernier est inconnu. Alors $\ell(\mathbf{p}_a, \mathbf{p}_{ab}^A, \mathbf{p}_{ab}^B, \mathbf{p}_b)$ est maximisé sous les contraintes

$$\sum\nolimits_{i\in\mathcal{S}(A), i\in a} p_{ai} = 1,\ \sum\nolimits_{i\in\mathcal{S}(A), i\in ab} p_{abi}^A = 1,$$

$$\sum\nolimits_{i\in\mathcal{S}(B), i\in b} p_{bi} = 1,\ \sum\nolimits_{i\in\mathcal{S}(B), i\in ab} p_{abi}^B = 1,$$

et

$$\sum_{i\in\mathcal{S}(A), i\in ab} p_{abi}^A y_i = \sum_{i\in\mathcal{S}(B), i\in ab} p_{abi}^B y_i. \qquad (5)$$

Quand N_{ab} est inconnu, les modifications de la pondération pour les estimateurs PVE sont

$$m_{i,\mathrm{PVE}}^{A} = \begin{cases} \dfrac{p_{ai}^{A}}{w_i^{A}}\left\{N_A - \hat{N}_{ab}^{\mathrm{PMV}}(\hat{\theta}_P)\right\} & \text{si } i \in a \\ \hat{\theta}_P \dfrac{p_{abi}^{A}}{w_i^{A}} \hat{N}_{ab}^{\mathrm{PMV}}(\hat{\theta}_P) & \text{si } i \in ab, \end{cases}$$

$$m_{i,\mathrm{PVE}}^{B} = \begin{cases} \dfrac{p_{bi}^{B}}{w_i^{B}}\left\{N_B - \hat{N}_{ab}^{\mathrm{PMV}}(\hat{\theta}_P)\right\} & \text{si } i \in b \\ (1-\hat{\theta}_P) \dfrac{p_{abi}^{B}}{w_i^{B}} \hat{N}_{ab}^{\mathrm{PMV}}(\hat{\theta}_P) & \text{si } i \in ab. \end{cases}$$

La contrainte (5) modifie les poids dans le domaine de chevauchement, de sorte que l'estimateur de Y_{ab} pour $\mathcal{S}(A)$ soit forcé d'être égal à l'estimateur de Y_{ab} pour $\mathcal{S}(B)$. Toutefois, cette contrainte produit un ensemble différent de poids pour chaque variable de réponse. L'estimateur PVE n'a donc pas de cohérence interne. Rao et Wu (2010) ont présenté une autre version de multiplicité dans laquelle les ajustements des poids ne dépendent pas de y ; en l'absence d'information auxiliaire, cet estimateur est le même que $\hat{Y}(1/2)$ dans (3).

2.2 Ajustements des poids avec trois bases de sondage ou plus

Dans le cas général, supposons que nous avons Q bases de sondage, désignées par $A_1, ..., A_Q$. Soit $\mathcal{S}(A_q)$ l'échantillon probabiliste tiré de la base de sondage A_q, pour $q = 1, ..., Q$. L'unité i dans l'échantillon $\mathcal{S}(A_q)$ a la probabilité d'inclusion $\pi_i^{A_q}$ et le poids $w_i^{A_q}$. Il existe un total de D domaines distincts.

Un estimateur pour bases de sondages multiples qui généralise (1) est de la forme

$$\hat{Y} = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}(A_q)} m_i^{A_q} w_i^{A_q} y_i,$$

où $m_i^{A_q}$ est l'ajustement de la pondération pour l'observation i dans $\mathcal{S}(A_q)$. Un estimateur à poids fixes établit les ajustements des poids $m^{(A_q,d)}$ pour chaque base de sondage et domaine sous les contraintes que $m^{(A_q,d)} \geq 0$ ($m^{(A_q,d)}$ est supposé égal à 0 si le domaine d ne fait pas partie de la base de sondage A_q) et $\sum_{q=1}^{Q} m^{(A_q,d)} = 1$ pour $d = 1, ..., D$. Alors, $m_i^{A_q} = m^{(A_q,d)}$ quand l'observation i provenant de $\mathcal{S}(A_q)$ se trouve dans le domaine d. Un choix simple, qui généralise l'estimateur à deux bases de sondage avec poids fixes $\hat{Y}(1/2)$ dans (3), se résume à prendre $m^{(A_q,d)} = [1/\text{nombre de bases de sondage qui contiennent le domaine } d]$; il s'agit de l'estimateur fondé sur la multiplicité de Mecatti (2007). D'autres choix consistent à poser que $m^{(A_q,d)} = 1$ dans exactement une base de sondage et 0 pour les autres bases de sondage, ce qui aboutit aux estimateurs avec sélection.

Nombre de propriétés du cas à deux bases de sondage s'étendent à celui de trois bases de sondage ou plus ; les versions pour bases de sondage multiples des estimateurs de la section 2.1 ont été étudiées par Hartley (1974), Lohr et Rao (2006), et Mecatti (2007). Comment les estimateurs pour bases de sondage multiples satisfont-ils aux critères établis au début de la présente section ? Tous les estimateurs – à poids fixes, optimaux, PMV, PVE et à base de sondage unique – sont approximativement sans biais pour les totaux de population quand des échantillons suffisamment grands sont tirés des bases de sondage. Les estimateurs à poids fixes, PMV et à base de sondage unique ont la propriété de cohérence interne ; par contre, les estimateurs optimaux de type Hartley et de type Fuller-Burmeister qui sont donnés dans Lohr et Rao (2006) et une extension aux bases de sondage multiples de l'estimateur PVE de Rao et Wu (2010) ne l'ont pas. Alors que les estimateurs optimaux sont asymptotiquement efficaces, ils sont souvent instables dans les échantillons de taille petite ou moyenne avec trois bases de sondage ou plus, parce que les modifications optimales estimées des poids sont des fonctions de grandes matrices de covariances estimées. Les estimateurs optimaux et les estimateurs PVE conviennent mal pour l'utilisation avec des logiciels d'analyse de données d'enquête standard, parce qu'ils requièrent un ensemble différent de poids pour chaque variable de réponse.

Nous recommandons que l'un des estimateurs avec cohérence interne – poids fixes, PMV ou base de sondage unique – soit utilisé en pratique. Lohr et Rao (2006) ont conclu que l'estimateur PMV possède une petite erreur quadratique moyenne dans de nombreuses conditions d'enquête et est donc un bon choix pour une enquête qui n'est réalisée qu'une seule fois. Dans le cas des enquêtes répétées, toutefois, on pourrait privilégier la simplicité et la transparence d'un estimateur à poids fixes. Les ajustements avec poids fixes peuvent faciliter les comparaisons d'une année à l'autre dans le cas d'une enquête annuelle où les proportions de domaine sont relativement constantes au cours du temps. Ces estimateurs se prêtent également mieux aux ajustements de la pondération pour tenir compte de la non-réponse et de la classification incorrecte dans les domaines (voir les sections 5.1 et 6.1). Si l'on peut choisir des ajustements des poids fixes qui sont proches des ajustements de poids optimaux pour les réponses importantes, peut-être en utilisant des effets de plan estimés d'après des enquêtes antérieures, l'estimateur à poids fixes aura une erreur quadratique moyenne proche de celle des estimateurs optimaux et PMV.

3. Poststratification sur les chiffres de population de contrôle

Tous les estimateurs décrits à la section 2 modifient les poids de sondage originaux. Par conséquent, certaines propriétés des poids originaux pourraient disparaître. Par exemple, si un échantillon aléatoire stratifié est tiré d'une base de sondage A, les poids modifiés n'auront pas nécessairement la propriété que la somme des poids dans une strate est égale à la taille de population de la strate.

Bankier (1986), dans l'élaboration originale des méthodes d'estimation à base de sondage unique, a proposé de caler par ratissage les poids de sondage, $\tilde{w}_{i,S}^{A}$ et $\tilde{w}_{i,S}^{B}$, sur les totaux de strate de manière que les poids ajustés $\tilde{w}_{i,S}^{A,\text{adj}}$ et $\tilde{w}_{i,S}^{B,\text{adj}}$ satisfassent

$$\sum\nolimits_{i \in S_{Ah}} (\tilde{w}_{i,S}^{A,\text{adj}} + \tilde{w}_{i,S}^{B,\text{adj}}) = N_{Ah},$$

où S_{Ah} représente les unités échantillonnées de n'importe quelle base de sondage dans la strate h de la base de sondage A, et N_{Ah} est la taille de la population de cette strate. Bankier (1986) et Skinner (1991) ont utilisé l'estimation par la méthode itérative du quotient (raking ratio) pour caler les estimateurs à base de sondage unique sur les tailles de population des bases de sondage N_A et N_B. Kott, Amrhein et Hicks (1998) ont proposé d'utiliser la méthode de calage par les moindres carrés de Deville et Särndal (1992) pour caler les totaux de population tels que les tailles de strates.

Pour l'estimateur PMV, Lohr et Rao (2000) ont recommandé de combiner d'abord les échantillons, puis d'utiliser les méthodes de calage pour l'ajustement sur le total de population, ainsi que sur les totaux de population des bases de sondage distinctes. En présence de non-réponse et si l'on utilise un estimateur à poids fixes, Brick, Cervantes, Lee et Norman (2011) ont conclu qu'il est préférable de poststratifier d'abord les échantillons individuels, puis de combiner les échantillons. Dans certaines situations, il est plus efficace de faire la poststratification avant ainsi qu'après avoir combiné les échantillons ; dans d'autres, la poststratification peut accroître le biais (voir section 6). Les décisions concernant la poststratification doivent être fondées sur l'erreur quadratique moyenne, qui comprend les effets des erreurs non dues à l'échantillonnage plutôt que simplement la variance d'échantillonnage.

4. Analyse des données d'enquête à bases de sondage multiples à l'aide de logiciels pour données d'enquête

4.1 Estimation ponctuelle avec des logiciels pour données d'enquête

Seuls les ajustements de poids maintenant la cohérence interne peuvent être utilisés dans les logiciels d'analyse de données d'enquête s'il existe plusieurs variables d'intérêt. Chacune des méthodes maintenant la cohérence interne présentée à la section 2.1 produisent un vecteur de poids ajustés pour chaque échantillon. Ces vecteurs peuvent ensuite être concaténés pour former un vecteur de poids $\tilde{\mathbf{w}} = [\tilde{w}_i^{A_1}, i \in \mathcal{S}(A_1), ..., \tilde{w}_i^{A_Q}, i \in \mathcal{S}(A_Q)]$. Soit $\mathbf{y}$ le vecteur correspondant d'observations, formé par la concaténation des observations provenant des échantillons $\mathcal{S}(A_1)$ à $\mathcal{S}(A_Q)$. Alors, $\hat{Y} = \tilde{\mathbf{w}}'\mathbf{y}$. Du point de vue de l'utilisateur, une fois que les poids modifiés sont construits, la procédure suivie pour trouver les estimations ponctuelles des totaux et des moyennes de population est la même que dans le cas d'une enquête à base de sondage unique.

Les poids modifiés provenant d'une procédure avec maintien de la cohérence interne peuvent être utilisés pour estimer toute quantité de population. Soit $F(y)$ la fonction de répartition pour la population, avec

$$F(y) = \sum_{i=1}^{N} I(y_i \le y) \Big/ N,$$

où $I(y_i \le y) = 1$ si $y_i \le y$ et 0 autrement. Dans une enquête à base de sondage unique, $F(y)$ est estimée par la fonction de répartition empirique

$$\hat{F}(y) = \sum\nolimits_{i \in \mathcal{S}} w_i I(y_i \le y) \Big/ \sum\nolimits_{i \in \mathcal{S}} w_i.$$

Les poids modifiés peuvent être utilisés pour estimer $F(y)$ dans une enquête à bases de sondage multiples :

$$\hat{F}(y) = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}(A_q)} \tilde{w}_i^{A_q} I(y_i \le y) \Bigg/ \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}(A_q)} \tilde{w}_i^{A_q}.$$

Le dénominateur est approximativement sans biais pour N, et le numérateur est approximativement sans biais pour $\sum_{i=1}^{N} I(y_i \le y)$. Toute fonctionnelle de la fonction de répartition peut alors être estimée en utilisant $\hat{F}(y)$: la moyenne, $\int y dF(y)$, la médiane m satisfaisant $F(m) \approx 1/2$, ou toute autre quantité.

Puisque les estimateurs avec poids de sondage modifiés sont approximativement sans biais pour les moyennes et les totaux de population, ils sont aussi approximativement sans biais pour les fonctions lisses des moyennes de population, tels que les ratios et les coefficients de régression. Toute quantité de population qui pourrait être estimée en utilisant les poids provenant d'une enquête à base de sondage unique peut être estimée de manière analogue en utilisant le vecteur de poids ajustés pour l'enquête à bases de sondage multiples.

4.2 Estimation de la variance avec les logiciels pour données d'enquête

Les plans de sondage doivent être connus pour calculer les erreurs-types. L'estimation de la variance est simple

pour l'estimateur (3), dans lequel les ajustements des poids ne dépendent pas des données. Dans cette situation,

$$V[\hat{Y}(\theta)] = V\left[\sum_{i\in\mathcal{S}(A)} \tilde{w}_i^A y_i\right] + V\left[\sum_{i\in\mathcal{S}(B)} \tilde{w}_i^B y_i\right],$$

où $\tilde{w}_i^A$ et $\tilde{w}_i^B$ sont définis sous (2). Créer l'ensemble de données par concaténation des observations provenant de $\mathcal{S}(A)$ et $\mathcal{S}(B)$ comme à la section 4.1, en utilisant $\tilde{w}_i^A$ et $\tilde{w}_i^B$ comme poids. Définir la variable de stratification pour l'échantillon combiné comme étant la combinaison des catégories données par la variable indicatrice de la base de sondage, la variable de stratification de la base de sondage A et la variable de stratification de la base de sondage B. Définir la variable de groupement de premier degré pour l'échantillon combiné de la même façon que pour la combinaison des catégories des variables de groupement des bases de sondage individuelles. Ensuite, un logiciel d'analyse de données d'enquête standard peut être utilisé pour estimer les moyennes et les totaux de population en se servant des poids modifiés, et pour estimer les variances en se servant des variables de stratification et des variables de groupement provenant des échantillons combinés.

L'estimation de la variance est plus compliquée quand les modifications des poids m_i^A ou m_i^B dépendent de quantités qui sont estimées d'après l'échantillon, comme dans l'estimateur PMV, ou quand l'échantillon combiné est poststratifié ou calé sur des quantités de population. Les méthodes de linéarisation, du jackknife et du bootstrap peuvent alors être employées pour estimer les variances.

Voici maintenant un résumé des méthodes qui peuvent être utilisées pour estimer la variance si les UPE provenant des bases de sondage sont sélectionnées indépendamment. Quand les échantillons provenant des diverses bases de sondage ont des UPE en commun, d'autres méthodes doivent être utilisées. Si, par exemple, les UPE sont tirées de la population et qu'un plan à deux bases de sondage est utilisé dans chaque UPE sélectionnée, les estimateurs ponctuels des totaux d'UPE peuvent être calculés en utilisant l'une des méthodes décrites à la section 2. Dans ce cas, les méthodes de rééchantillonnage classiques peuvent être utilisées pour calculer un estimateur de variance avec remplacement.

Sous certaines conditions de régularité, les méthodes de linéarisation et du jackknife sont convergentes pour l'estimation de la variance d'une caractéristique de population τ qui peut s'écrire sous la forme $\tau = g(\mathbf{A}, \mathbf{B})$, où $\mathbf{A}$ est un vecteur de totaux de population provenant de la base de sondage A, $\mathbf{B}$ est un vecteur de totaux de population provenant de la base de sondage B, et g est une fonction deux fois continuellement dérivable (Skinner et Rao 1996 ; Lohr et Rao 2000). Le vecteur $\mathbf{A}$ est estimé d'après $\mathcal{S}(A)$ par $\hat{\mathbf{A}}$, avec la matrice de covariance estimée $\hat{\boldsymbol{\Sigma}}_A$; de même, $\hat{\mathbf{B}}$ estime $\mathbf{B}$ d'après $\mathcal{S}(B)$, avec $\hat{V}(\hat{\mathbf{B}}) = \hat{\boldsymbol{\Sigma}}_B$. L'estimateur par linéarisation de la variance de $\hat{\tau} = g(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ est

$$\hat{V}_L(\hat{\tau}) = g_A' \hat{\boldsymbol{\Sigma}}_A g_A + g_B' \hat{\boldsymbol{\Sigma}}_B g_B,$$

où g_A est le vecteur de dérivées partielles de $g(\mathbf{A}, \mathbf{B})$ par rapport aux composantes de $\mathbf{A}$ et g_B est le vecteur correspondant de dérivées partielles provenant de la base de sondage B. Demnati, Rao, Hidiroglou et Tambay (2007) ont obtenu les estimateurs de la variance par linéarisation pour les enquêtes à bases de sondage multiples en prenant les dérivées d'une fonction des poids plutôt que des moyennes. Les méthodes de linéarisation requièrent que les dérivées soient calculées séparément pour chaque estimateur pris en considération, et ces calculs peuvent être fastidieux. Pour cette raison, il est parfois préférable d'utiliser les méthodes de rééchantillonnage quand des enquêtes à bases de sondage multiples sont adoptées.

Supposons que l'on tire un échantillon stratifié à plusieurs degrés comprenant H strates de la base de sondage A, où la strate h contient $\tilde{n}_h^A$ unités primaires d'échantillonnage. Un échantillon stratifié à plusieurs degrés indépendant comprenant L strates est tiré de la base de sondage B, où la strate l contient $\tilde{n}_l^B$ unités primaires d'échantillonnage. L'estimateur jackknife de la variance peut être calculé en créant un total de $\sum_{h=1}^{H} \tilde{n}_h^A + \sum_{l=1}^{L} \tilde{n}_l^B$ colonnes de poids de rééchantillonnage (Lohr et Rao 2000). Les poids de rééchantillonnage pour la colonne correspondant à la suppression de l'UPE i de la strate h dans $\mathcal{S}_A$ sont donnés par :

$$\tilde{w}_{k(hi)}^A = \begin{cases} \dfrac{\tilde{n}_h^A}{\tilde{n}_h^A - 1}\tilde{w}_k^A & \text{si l'unité } k \text{ est dans la strate } h \text{ mais non dans l'UPE } i\,; \\ 0 & \text{si l'unité } k \text{ est dans l'UPE } i \text{ de la strate } h\,; \\ \tilde{w}_k^A & \text{si l'unité } k \text{ est dans la strate } g \neq h. \end{cases}$$

Le coefficient jackknife pour cette colonne est le multiplicateur $(\tilde{n}_h^A - 1) / \tilde{n}_h^A$. La colonne de poids de rééchantillonnage correspondant à la suppression de l'UPE j provenant de la strate l dans $\mathcal{S}_B$ est formée de la même façon, avec le coefficient jackknife $(\tilde{n}_l^B - 1) / \tilde{n}_l^B$. Lorsqu'il y a plus de deux bases de sondage, des colonnes supplémentaires de poids de rééchantillonnage correspondant aux UPE supprimées provenant de ces échantillons sont ajoutées. Les poids pour une méthode bootstrap d'estimation de la variance (voir Lohr 2007) peuvent être définis similairement.

Les méthodes d'estimation de la variance par rééchantillonnage de bases de sondage multiples peuvent être utilisées avec les progiciels d'analyse de données d'enquête standard qui permettent l'utilisation des poids de rééchantillonnage. Au besoin, chaque colonne de poids de

rééchantillonnage peut être poststratifiée sur les totaux de population et de base de sondage, de sorte que la post-stratification soit prise en compte dans l'estimation de la variance.

L'une des difficultés que posent les méthodes d'estimation de la variance par rééchantillonnage tient au fait qu'un très grand nombre de colonnes de poids de rééchantillonnage peut être nécessaire si un échantillon aléatoire simple ou un échantillon aléatoire stratifié est tiré de l'une des bases de sondage. Dans le cas du bootstrap, nous avons constaté que, pour certaines enquêtes, au moins 500 itérations bootstrap sont requises pour les estimations de la variance dans le cas d'enquêtes à deux bases de sondage, ce qui de nouveau peut être excessif. Il se peut que l'estimation de la variance pour des strates combinées, comme il est discuté dans Lu, Brick et Sitter (2006), soit utilisable dans le cas des enquêtes à bases de sondage multiples en vue de réduire le nombre de répliques nécessaires.

5. Erreurs non dues à l'échantillonnage

Les enquêtes à bases de sondage multiples ont souvent une meilleure couverture de la population que les enquêtes à bases de sondage unique. Quand toutes les bases de sondage sont incomplètes, comme à la figure 3, n'importe laquelle des bases de sondage A, B ou C, si elle était utilisée comme seule base de sondage, donnerait lieu à un sous-dénombrement grave. Les plans à bases de sondage multiples font en sorte que toutes les unités qui se trouvent dans les bases de sondage chevauchantes ont une probabilité d'inclusion positive.

Comme toutes les autres enquêtes, les enquêtes à bases de sondage multiples sont sujettes à des erreurs non dues à l'échantillonnage. Elles souffrent de non-réponse, qui peut être différente dans les diverses bases de sondage. Même si l'union des bases de sondage peut produire une meilleure couverture qu'une base de sondage unique, il pourrait persister un sous-dénombrement de la population cible. Les estimateurs pour les enquêtes à bases de sondage multiples sont également sensibles aux erreurs de classification dans les domaines et aux biais qui pourraient résulter de différentes méthodes ou de différents modes de collecte dans les enquêtes composantes. Nous discutons de la non-réponse et des effets de mode à la présente section, et nous étudions les effets de l'erreur de classification dans les domaines à la section 6.

5.1 Non-réponse

Dans toute enquête, la non-réponse peut introduire un biais dans les estimations des totaux de population et d'autres quantités. Des taux de non-réponse différents dans les échantillons provenant des deux bases de sondage peuvent avoir une incidence sur les estimations ponctuelles du total de population données à la section 2 ; en outre, la non-réponse peut affecter les ajustements des poids prescrits par certaines méthodes.

Kennedy (2007) a discuté d'un problème qui a été constaté quand la base de sondage A est une liste de numéros de téléphone fixe et que la base de sondage B est une liste de numéros de téléphone mobile : les unités dans le domaine d'intersection ab qui ont répondu sur un téléphone mobile différaient de celles qui l'ont fait sur un téléphone fixe. Par exemple, on a estimé que 18 % des unités dans le domaine d'intersection avaient de 18 à 25 ans dans l'échantillon tiré de la base de sondage B, tandis que 8 % seulement des unités de l'intersection avaient de 18 à 25 ans en se servant de l'échantillon de la base de sondage A. La différence a été attribuée à la non-réponse : on a pensé que les personnes qui utilisent principalement un téléphone mobile (et donc sont difficiles à joindre dans le cadre d'une enquête fondée sur des numéros de téléphone fixe) ont tendance à être plus jeunes. Kennedy (2007) a suggéré un calage (*raking*) en se servant de l'usage relatif estimé du téléphone (c'est-à-dire si la plupart des appels sont faits sur un téléphone fixe ou sur un téléphone mobile).

Brick et coll. (2011) ont proposé deux méthodes pour l'ajustement de la non-réponse dans les enquêtes téléphoniques à deux bases de sondage (lignes mobiles/lignes fixes) avec des estimateurs à poids fixes. Ils ont considéré les conditions dans lesquelles le domaine de chevauchement possède deux groupes : les ménages qui reçoivent tous leurs appels, ou presque tous, sur des téléphones mobiles (mobile principalement) et les autres ménages dans le domaine de chevauchement (fixe principalement). La première méthode, qui ne requiert pas d'estimation externe des totaux de contrôle, consiste à fixer la valeur de θ dans l'estimateur avec ajustement des poids fixes de manière à réduire le biais de non-réponse en utilisant les taux de réponse pour les ménages se servant d'un téléphone mobile principalement et d'un téléphone fixe principalement dans chaque échantillon. La deuxième méthode nécessite des totaux de contrôle de poststratification pour les groupes à téléphone mobile principalement et téléphone fixe principalement dans le domaine chevauchant, N_{1ab} et N_{2ab}, et estime le total de population dans le domaine ab par

$$\sum_{g=1}^{2}\left[\theta_g \frac{N_{gab}}{\hat{N}^A_{gab}} \hat{Y}^A_{gab} + (1-\theta_g)\frac{N_{gab}}{\hat{N}^B_{gab}} \hat{Y}^B_{gab}\right],$$

où $\hat{Y}^A_{gab}$ représente le total estimé du groupe g dans le domaine ab provenant de $\mathcal{S}(A)$, les autres totaux sont définis de la même façon et $0 \le \theta_g \le 1$ pour $g = 1, 2$.

5.2 Effets de mode

Dans certains cas, les bases de sondage multiples peuvent aussi signifier modes multiples. De Leeuw (2008)

a comparé les avantages et les inconvénients de divers modes d'échantillonnage et a résumé la recherche empirique sur les biais dus au mode. Les personnes peuvent donner des réponses différentes lorsqu'on leur présente les questions sous une forme visuelle ou lorsqu'on les leur présente sous forme auditive, ce qui donne lieu à un biais de mode. Les effets de mode qui se produisent dans les enquêtes à base de sondage unique auront lieu dans les enquêtes à bases de sondage multiples. Si différents modes sont utilisés dans différentes bases de sondage, il est difficile de faire la distinction entre les effets de mode et d'autres erreurs non dues à l'échantillonnage.

Bon nombre d'estimateurs pour enquête à bases de sondage multiples combinent les estimations provenant des domaines de chevauchement, et ces méthodes reposent sur l'hypothèse que les estimateurs de Y_{ab} provenant des enquêtes composantes estiment tous deux la même quantité. Si, par contre, l'enquête à base de sondage A est menée sur place, tandis que l'enquête à base de sondage B est menée par téléphone, il est possible qu'un recensement du domaine ab d'après la base de sondage B donne un total de domaine différent du recensement d'après la base de sondage A.

Un moyen d'étudier les effets de mode consiste à réaliser l'enquête à base de sondage B en utilisant un échantillon scindé, par exemple interviewé en partie sur place et en partie par téléphone, mais cela réduirait l'économie due à l'utilisation de deux bases de sondage. Un pré-test minutieux peut atténuer les effets de mode. Des travaux de recherche sont nécessaires dans ce domaine ; le même problème d'effets de mode se pose évidemment dans les enquêtes à base de sondage unique, tel que l'American Community Survey, dans laquelle le suivi des cas de non-réponse est effectué selon un mode différent de celui utilisé pour l'échantillon original (voir Citro et Kalton 2007). Les méthodes présentées dans Leeuw, Hox et Dillman (2008) pour la conception des enquêtes dans le cas de modes de collecte multiples s'appliquent aussi au cas des enquêtes à bases de sondage multiples.

Vannieuwenhuyze, Loosveldt et Molenberghs (2011) ont présenté une méthode pour distinguer les effets de mode des effets de sélection quand une enquête à mode unique complémentaire est disponible. Ils ont toutefois constaté que la méthode requiert l'hypothèse forte selon laquelle les erreurs de couverture et de non-réponse sont équivalentes dans les deux enquêtes. Si cette hypothèse est satisfaite pour une enquête à deux bases de sondage, de sorte que les échantillons dans le domaine de chevauchement provenant des bases de sondage A et B représentent la même population, et que la classification par domaine est correcte, l'effet de mode peut être estimé d'après le domaine de chevauchement comme étant $D_{ab} = \hat{Y}_{ab}^{A} - \hat{Y}_{ab}^{B}$. Une différence qui diffère de manière significative de 0 indique la présence d'un effet de mode s'il n'existe pas d'autres erreurs non dues à l'échantillonnage. Sinon, une grande valeur de D_{ab} ne fournit pas d'information sur la cause de la différence ; une expérimentation est nécessaire pour faire la distinction entre les causes possibles.

6. Erreur de classification dans les domaines et correction du biais

Dans les estimateurs examinés à la section 2, les poids des observations sont construits en se basant sur l'appartenance à un domaine. Donc, dans l'estimateur $\hat{Y}(\theta)$ donné par (3), le multiplicateur du poids d'une observation provenant de l'échantillon tiré de la base de sondage A est 1 si l'observation est dans le domaine a, et est θ si l'observation est dans le domaine ab, afin de tenir compte de la multiplicité de l'échantillonnage.

En pratique, l'appartenance au domaine n'est pas toujours claire. Dans le cas de la figure 1, on peut ne pas savoir si un répondant dans une base de sondage aréolaire appartient aussi à la liste. Si la base de sondage A est une base aréolaire et que la base de sondage B est une base de sondage en ligne, par exemple, le seul moyen de déterminer si une personne échantillonnée à partir de la base de sondage A figure aussi dans la base de sondage B pourrait consister à demander à la personne si elle a accès à Internet, et la personne pourrait ne pas donner la réponse correcte.

Si l'on recourt à l'appariement ou au couplage d'enregistrements pour déterminer l'appartenance à une base de sondage, un appariement imparfait peut également produire une classification incorrecte des observations. Lesser et Kalsbeek (1999) ont discuté des erreurs non dues à l'échantillonnage qui ont lieu dans les enquêtes à deux bases de sondage menées par le U.S. National Agricultural Statistics Service. Une erreur de classification dans les domaines peut avoir lieu si une ferme échantillonnée dans la base aréolaire est classée incorrectement en ce qui concerne son appartenance à la liste. Dans les enquêtes téléphoniques à deux bases de sondage téléphone fixe/téléphone cellulaire, il est difficile de déterminer si une personne présente dans une base de sondage est également présente dans l'autre (Kennedy 2007). Une personne rejointe dans un échantillon de numéros de téléphone fixe peut aussi posséder un téléphone mobile, mais rarement prendre des appels sur le téléphone mobile. Bien qu'elle appartienne techniquement au domaine de chevauchement, cette personne ne peut pour ainsi dire pas être rejointe dans l'enquête utilisant les téléphones mobiles. Dans certaines enquêtes téléphone fixe/téléphone mobile, on demande aux répondants d'indiquer les parts relatives d'usage du téléphone mobile ou du téléphone fixe, mais une erreur de classification peut avoir lieu.

En pratique, nous nous attendons à ce que les erreurs de classification dans les domaines soient reliées aux réponses d'intérêt ; nous nous attendons aussi à ce que, dans de nombreuses situations, l'erreur de classification soit plus probablement dans certaines directions. Dans les enquêtes longitudinales à deux bases de sondage, l'erreur de classification dans les domaines peut avoir des effets plus importants que dans les enquêtes transversales (Lu et Lohr 2010). Dans certaines situations, l'indicateur de domaine peut manquer ou ne pas être disponible. Clark, Winglee et Liu (2007) ont étudié des méthodes de régression logistique et des méthodes d'appariement d'enregistrements pour prédire le domaine d'une observation pour laquelle manque l'information sur le domaine.

6.1 Corrections du biais dû à l'erreur de classification

Si l'erreur de classification dans les domaines est importante, chaque méthode de modification des poids de sondage en vue de les ajuster pour tenir compte de la multiplicité peut produire des estimations biaisées des quantités de population. À la présente section, nous établissons une correction pour le biais dû à l'erreur de classification dans les domaines de l'estimateur à poids fixes de la section 2.2 quand les probabilités d'erreur de classification sont connues. Soit le D-vecteur $\boldsymbol{\delta}_i^{A_q}$ qui désigne l'appartenance vraie au domaine pour l'observation i de la base de sondage A_q, contenant un 1 en position d si l'observation i est dans le domaine d, et 0 autrement. Soit $\mathbf{Y} = (Y_1, ..., Y_D)'$ le vecteur des totaux de population pour les D domaines. Pour une enquête à deux bases de sondage chevauchantes, $\mathbf{Y} = (Y_a, Y_{ab}, Y_b)'$; pour une enquête à trois bases de sondage, $\mathbf{Y} = (Y_a, Y_{ab}, Y_{ac}, Y_{abc}, Y_b, Y_{bc}, Y_c)'$. S'il n'y a pas d'erreur de classification dans les domaines,

$$\hat{\mathbf{Y}}^{A_q} = \sum\nolimits_{i \in \mathcal{S}(A_q)} \boldsymbol{\delta}_i^{A_q} w_i^{A_q} y_i$$

est l'estimateur correspondant de $\mathbf{Y}$ provenant de $\mathcal{S}(A_q)$. Pour le vecteur d'ajustement des poids fixes $\mathbf{m}^{A_q} = (m^{(A_q,1)}, ..., m^{(A_q,D)})'$ dans la base de sondage A_q, satisfaisant $\sum_{q=1}^{Q} \mathbf{m}^{(A_q,d)} = 1$, alors $E[\sum_{q=1}^{Q} (\mathbf{m}^{A_q})' \hat{\mathbf{Y}}^{A_q}] = Y$.

Maintenant, supposons qu'il existe une erreur de classification. Soit $\boldsymbol{\eta}_i^{A_q}$ la classification observée de l'observation i dans $\mathcal{S}$. Nous pouvons écrire $\boldsymbol{\eta}_i^{A_q} = (\mathbf{M}_i^{A_q})' \boldsymbol{\delta}_i^{A_q}$, où $\mathbf{M}_i^{A_q}$ est une matrice de dimensions $D \times D$ contenant un 1 en position (d, e) si l'observation i dans le vrai domaine d est classée (in)correctement comme appartenant au domaine e, et 0 autrement.

Pour tenir compte de l'erreur de classification différentielle à l'intérieur des domaines, nous postulons une structure dans laquelle les probabilités d'erreur de classification peuvent différer selon la sous-population dans une base de sondage. Dans une enquête téléphone fixe/téléphone mobile, on pourrait savoir que la probabilité d'une erreur de classification est plus élevée pour certains groupes d'âge que pour d'autres. Chambers, Chipperfield, Davis et Kovačević (2008) ont utilisé une approche de groupement similaire pour corriger les erreurs de couplage d'enregistrements. Supposons que la population peut être divisée en G groupes, $g = 1, ..., G$, dans lesquels les probabilités d'erreur de classification sont connues pour chaque base de sondage A_q. Soit $\phi_g^{A_q}(d, e)$ la probabilité qu'une observation dans le groupe g avec domaine vrai d soit classée dans le domaine e dans l'échantillon $\mathcal{S}(A_q)$, et soit $\boldsymbol{\Phi}_g^{A_q}$ la matrice de dimension $D \times D$ dont les entrées sont $\phi_g^{A_q}(d, e)$. Pour l'observation i appartenant au groupe g et au vrai domaine d, supposons que la ligne d de $\mathbf{M}_i^{A_q}$ est générée sous forme d'une variable aléatoire multinominale de taille 1 avec les probabilités dans la ligne d de la matrice des erreurs de classification attendues $\boldsymbol{\Phi}_g^{A_q}$, et que tous les $\mathbf{M}_i^{A_q}$ sont indépendants les uns des autres et des variables d'inclusion dans l'échantillon. Nous avons donc G matrices de probabilités d'erreur de classification pour la base de sondage A_q, $\boldsymbol{\Phi}_1^{A_q}, ..., \boldsymbol{\Phi}_G^{A_q}$. Désignons le vecteur de totaux de population pour le groupe g par $\mathbf{Y}(g) = \sum_{i=1}^{N} \boldsymbol{\delta}_i^{A_q} \chi_i(g) y_i$, où $\chi_i(g) = 1$ si l'observation i est dans le groupe g et 0 autrement.

Pour les classifications dans les domaines observées $\boldsymbol{\eta}_i^{A_q}$, l'estimateur pondéré par les poids de sondage du vecteur des totaux de population dans le groupe g est

$$\begin{aligned}\hat{\mathbf{Y}}^{A_q}(\text{mis}, g) &= \sum_{i \in \mathcal{S}(A_q)} \boldsymbol{\eta}_i^{A_q} \chi_i(g) w_i^{A_q} y_i \\ &= \sum_{i \in \mathcal{S}(A_q)} (\mathbf{M}_i^{A_q})' \boldsymbol{\delta}_i^{A_q} \chi_i(g) w_i^{A_q} y_i,\end{aligned}$$

de sorte que $E[\hat{\mathbf{Y}}^{A_q}(\text{mis}, g)] = (\boldsymbol{\Phi}_g^{A_q})' \mathbf{Y}(g)$.

Considérons maintenant un nouveau vecteur d'ajustements des poids $\tilde{\mathbf{m}}_g^{A_q} = (\tilde{m}_g^{(A_q,1)}, ..., \tilde{m}_g^{(A_q,D)})'$ pour le groupe g dans la base de sondage A_q. Alors

$$E\left[\sum_{g=1}^{G} \sum_{q=1}^{Q} (\tilde{\mathbf{m}}_g^{A_q})' \hat{\mathbf{Y}}^{A_q}(\text{mis}, g)\right] = \sum_{g=1}^{G} \sum_{q=1}^{Q} (\boldsymbol{\Phi}_g^{A_q} \tilde{\mathbf{m}}_g^{A_q})' \mathbf{Y}(g).$$

Puisque $\sum_{g=1}^{G} \sum_{q=1}^{Q} (\mathbf{m}^{A_q})' \mathbf{Y}(g) = Y$, le biais sera éliminé sous ce modèle quand

$$\tilde{\mathbf{m}}_g^{A_q} = (\boldsymbol{\Phi}_g^{A_q})^{+} \mathbf{m}^{A_q}, \tag{6}$$

où $(\boldsymbol{\Phi}_g^{A_q})^{+}$ est l'inverse de Moore-Penrose de $\boldsymbol{\Phi}_g^{A_q}$, obtenue en prenant l'inverse des lignes et des colonnes non nulles de $\boldsymbol{\Phi}_g^{A_q}$.

Le remplacement des ajustements $\mathbf{m}^{A_q}$ par $\tilde{\mathbf{m}}_g^{A_q}$ élimine le biais sous le modèle d'erreur de classification

multinomial, mais augmente la variance. Partant de la base de sondage A_q,

$$V\left[\sum_{g=1}^{G} (\tilde{\mathbf{m}}_g^{A_q})' \hat{\mathbf{Y}}^{A_q}(\text{mis}, g)\right]$$

$$= E\left[V\left(\sum_{g=1}^{G} \sum_{i\in\mathcal{S}(A_q)} \{(\mathbf{\Phi}_g^{A_q})^+ \mathbf{m}^{A_q}\}' (\mathbf{M}_i^{A_q})' \right.\right.$$

$$\left.\left. \boldsymbol{\delta}_i^{A_q} \chi_i(g) w_i^{A_q} y_i \mid \mathcal{S}(A_1), ..., \mathcal{S}(A_Q) \right)\right]$$

$$+ V\left[E\left(\sum_{g=1}^{G} \sum_{i\in\mathcal{S}(A_q)} \{(\mathbf{\Phi}_g^{A_q})^+ \mathbf{m}^{A_q}\}' (\mathbf{M}_i^{A_q})' \right.\right.$$

$$\left.\left. \boldsymbol{\delta}_i^{A_q} \chi_i(g) w_i^{A_q} y_i \mid \mathcal{S}(A_1), ..., \mathcal{S}(A_Q)\right)\right]$$

$$= \sum_{g=1}^{G} [(\mathbf{\Phi}_g^{A_q})^+ \mathbf{m}^{A_q}]' E\left[\sum_{i\in\mathcal{S}(A_q)} \chi_i(g)(w_i^{A_q} y_i)^2 \right.$$

$$\left. \{\text{diag}[(\mathbf{\Phi}_g^{A_q})'\boldsymbol{\delta}_i^{A_q}] - (\mathbf{\Phi}_g^{A_q})'\boldsymbol{\delta}_i^{A_q}(\boldsymbol{\delta}_i^{A_q})'\mathbf{\Phi}_g^{A_q}\}\right] (\mathbf{\Phi}_g^{A_q})^+ \mathbf{m}^{A_q}$$

$$+ V\left[\sum_{i\in\mathcal{S}(A_q)} \{\mathbf{m}^{A_q}\}' \boldsymbol{\delta}_i^{A_q} w_i^{A_q} y_i\right].$$

Le deuxième terme est la variance de la contribution de la base de sondage A_q quand les unités sont classées correctement. Le premier terme est nul uniquement si $\mathbf{\Phi}_g^{A_q}$ est diagonale pour tout g, c'est-à-dire qu'il n'y a pas d'erreur de classification.

Les ajustements des poids donnés par (6) peuvent être étendus au cas dans lequel les poids fixes originaux $\mathbf{m}^{A_q}$ varient pour les groupes, à condition que $\sum_{q=1}^{Q} m_g^{(A_q, d)} = 1$ pour chaque domaine. Notons que la méthode de correction du biais exposée à la présente section est proposée uniquement pour les estimateurs à poids fixes et non pour les estimateurs PMV, PVE ou optimaux dans lesquels les poids fondés sur la multiplicité dépendent des données. La correction du biais dépend de la spécification correcte des probabilités d'erreur de classification. Si ces dernières sont estimées d'après une autre enquête, les méthodes d'opérations des enquêtes doivent être semblables.

6.2 Étude en simulation

Lohr et Rao (2006) ont constaté dans des études en simulation que l'estimateur PMV avait une plus petite erreur quadratique moyenne que les autres estimateurs en présence d'erreur de classification aléatoire, mais cela tient en grande partie à la variance plus faible que cet estimateur. Pour étudier la sensibilité des estimateurs à d'autres formes d'erreur de classification dans les domaines, nous avons exécuté une étude en simulation pour des enquêtes à deux et à trois bases de sondage. La population du domaine d a été générée en utilisant le modèle $y_{ij} = \mu_d + \alpha_i + \varepsilon_{ij}$ pour $i = 1, ..., N_d$ et $j = 1, ..., 5$, avec $\alpha_i \sim N(0, 1)$ et $\varepsilon_{ij} \sim N(0, 1)$ générée indépendamment, puis des échantillons probabilistes ont été tirés de cette population.

Pour l'étude à deux bases de sondage, les moyennes de domaine sont $\mu_a = -1$, $\mu_{ab} = 0$, $\mu_b = 2$ et les facteurs de la simulation sont :

1. Taille de l'échantillon : 100 ou 200 provenant de chaque base de sondage.
2. Échantillon en grappes ou échantillon aléatoire simple tiré de la base de sondage A. Un échantillon en grappes a été tiré en sélectionnant un échantillon aléatoire simple de $n_A/5$ des groupes utilisés pour générer la population.
3. Probabilités d'erreur de classification pour la base de sondage A (toutes les probabilités non énumérées sont égales à 0) :
   a. $\phi_{aa}^A = 1$, $\phi_{ab,ab}^A = 1$ (pas d'erreur de classification) ;
   b. $\phi_{aa}^A = 0{,}9$, $\phi_{a,ab}^A = 0{,}1$, $\phi_{ab,ab}^A = 1$;
   c. $\phi_{aa}^A = 0{,}9$, $\phi_{a,ab}^A = 0{,}1$, $\phi_{ab,ab}^A = 0{,}9$, $\phi_{ab,a}^A = 0{,}1$;
   d. $\phi_{aa}^A = 1$, $\phi_{ab,ab}^A = 0{,}9$, $\phi_{ab,a}^A = 0{,}1$.
4. Probabilités d'erreur de classification pour la base de sondage B :
   a. $\phi_{bb}^B = 1$, $\phi_{ab,ab}^B = 1$ (pas d'erreur de classification) ;
   b. $\phi_{bb}^B = 0{,}8$, $\phi_{b,ab}^B = 0{,}2$, $\phi_{ab,ab}^B = 1$;
   c. $\phi_{bb}^B = 0{,}8$, $\phi_{b,ab}^B = 0{,}2$, $\phi_{ab,ab}^B = 0{,}9$, $\phi_{ab,b}^B = 0{,}1$;
   d. $\phi_{bb}^B = 1$, $\phi_{ab,ab}^B = 0{,}8$, $\phi_{ab,b}^B = 0{,}2$.
5. Tailles de population : $N_a = N_b = N_{ab} = 25\,000$; $N_a = N_b = 10\,000$, $N_{ab} = 55\,000$; $N_a = 25\,000$, $N_{ab} = 40\,000$, $N_b = 10\,000$.

Dix mille répliques ont été exécutées pour chaque combinaison de facteurs, ce qui donne à l'estimation Monte Carlo du biais une erreur-type d'environ 100. Nous avons étudié tous les estimateurs décrits à la section 2, y compris $\hat{Y}(1/2)$, $\hat{Y}(2/3)$, et $\hat{Y}(1)$ provenant de l'équation (3). Nous avons également examiné les estimateurs poststratifiés qui pourraient être employés quand les chiffres de population de domaine N_a, N_{ab} et N_b sont connus : les estimateurs avec l'indice inférieur « post1 » appliquent la poststratification aux deux échantillons pour commencer puis combinent les échantillons, tandis que les estimateurs avec l'indice inférieur « post2 » combinent les échantillons d'abord, puis font la poststratification sur les chiffres de population de domaine. Les estimateurs corrigés du biais $\hat{Y}(1/2)_{bc}$ et $\hat{Y}(2/3)_{bc}$ modifient les poids fixes initiaux correspondant à $\theta = 1/2$ et $\theta = 2/3$ en utilisant (6). Avec le profil d'erreur de classification (b) dans la base de sondage A, par exemple, les ajustements des poids corrigés du biais pour $\hat{Y}(1/2)_{bc}$ sont $\tilde{m}_i^A = 19/18$ pour i classé

dans a et $\tilde{m}_i^A = 1/2$ pour i classé dans ab ; pour le profil (c), les ajustements des poids corrigés du biais sont 17/16 et 7/16, respectivement. L'estimateur pour base de sondage unique est omis de ces tableaux puisqu'il est le même que $\hat{Y}(1/2)$ ou $\hat{Y}(2/3)$; l'estimateur pour base de sondage unique calé par ratissage sur les totaux de population N_A et N_B est désigné par $\hat{Y}_{\text{SF, rat}}$. Les tableaux 1 et 2 donnent les résultats pour $n_A = 100$, $n_B = 100$, $N_a = N_{ab} = N_b = 25\,000$, et un échantillon aléatoire simple tiré de la base de sondage A ; les tableaux 3 et 4 donnent les résultats pour $n_A = 200$, $n_B = 100$, $N_a = N_{ab} = N_b = 25\,000$, et un échantillon en grappes tiré de la base de sondage A. Les profils généraux des résultats sont similaires pour les autres simulations qui ne sont pas présentées ici.

Premièrement, considérons les estimateurs à poids fixes. Les estimateurs corrigés du biais réduisent le biais comme prévu ; dans tous les cas étudiés avec erreur de classification, le biais empirique provenant des estimateurs corrigés du biais étaient inférieurs à 200 en valeur absolue, ce qui est compris dans la marge d'erreur. Bien que l'écart-type pour les estimateurs corrigés du biais soit plus élevé que pour les estimateurs non corrigés, dans la plupart des cas, les erreurs quadratiques moyennes sont comparables

L'estimateur avec sélection $\hat{Y}(1)$, qui élimine les unités provenant de la base de sondage B du domaine ab, ne présente pas de biais d'erreur de classification quand les unités de la base de sondage B sont classées correctement. Il ne présente pas non plus de biais dans les tableaux 1 et 3 dans le cas du profil (d) d'erreur de classification dans la base de sondage B, parce que les observations classées incorrectement dans le domaine b au lieu du domaine ab ont une moyenne nulle ; pour différents ensembles de moyennes de domaine, le profil (d) ne crée pas de biais. Pour les autres cas, l'estimateur avec sélection est celui dont le biais est le plus important. Pour chaque profil d'erreur de classification, l'estimateur avec sélection possède une plus grande erreur quadratique moyenne parce que des données sont écartées. Si les moyennes de domaine sont similaires, l'erreur de classification dans les domaines pourrait ne pas causer de biais appréciable, mais le rejet des observations provenant du domaine ab dans $\mathcal{S}(B)$ augmenterait fortement l'erreur quadratique moyenne.

La poststratification sur les totaux de domaine en présence d'erreurs de classification augmente souvent le biais au lieu de le réduire. Considérons la ligne 4 du tableau 1, où 20 % des observations de $\mathcal{S}(B)$ appartenant au domaine ab sont classées par erreur dans le domaine b. Les poids des observations qui appartiennent réellement au domaine b, de moyenne 2, sont réduits de 500 à environ 417, ce qui cause un biais dans les versions poststratifiées de $\hat{Y}(1/2)$. L'effet de la poststratification sur l'erreur quadratique moyenne est mixte et dépend du fait que la réduction de la variance obtenue par poststratification est supérieure ou non au biais supplémentaire qui peut être introduit. Le calage sur les totaux des bases de sondage N_A et N_B, dans $\hat{Y}_{\text{SF, rat}}$, a le même effet que la poststratification sur le biais d'erreur de classification.

Pour les échantillons aléatoires simples des tableaux 1 et 2, les estimateurs PMV et PVE présentent souvent un nettement plus grand biais que les estimateurs à poids fixes non corrigés. Pour ces méthodes, les contributions relatives des deux bases de sondage dépendent des variances estimées de $\hat{N}_{ab}^A$ et $\hat{N}_{ab}^B$, les poids de domaine dépendent de $\hat{N}_{ab}^{\text{PMV}}$, et ces deux facteurs interagissent de façon complexe selon la structure de l'erreur de classification. Pour le profil d'erreur de classification (d) dans l'une ou l'autre base de sondage, $\hat{N}_{ab}^{\text{PMV}}$ est trop petit parce que les observations dans le domaine ab sont classées incorrectement ; par conséquent, les poids des observations comprises dans les domaines non chevauchants sont trop grands. Une version poststratifiée de l'estimateur PMV présente les problèmes de biais des estimateurs à poids fixes poststratifiés. En forçant les estimateurs de Y_{ab} à être égaux, l'estimateur PVE peut aggraver le biais. Par exemple, dans la situation de la ligne 3 du tableau 1, avec une classification correcte pour la base de sondage A et le profil (c) pour la base de sondage B, le biais de PVE est supérieur de 50 % au biais de PMV. Dans ce cas, l'estimateur PVE tire l'estimateur sans biais $\hat{Y}_{ab}^A$ provenant de $\mathcal{S}(A)$ vers l'estimateur biaisé provenant de la base de sondage B. Les estimateurs optimaux ont aussi un biais important.

Quand un échantillon en grappes est tiré de la base de sondage A comme aux tableaux 3 et 4, les profils de biais sont similaires. Quand il n'y a pas d'erreur de classification, les EQM des estimateurs optimaux et PMV sont plus faibles que celles de $\hat{Y}(2/3)$, parce qu'elles tiennent compte du plan de sondage. Toutefois, en présence d'erreur de classification, l'avantage en ce qui concerne l'EQM est réduit à cause de l'accroissement du biais.

Pour étudier l'erreur de classification dans une enquête à trois bases de sondage, nous avons tiré des échantillons aléatoires simples de chaque base de sondage et nous avions les classifications correctes pour les bases de sondage B et C. Le tableau 5 donne les résultats pour une simulation avec trois bases de sondage et un échantillon aléatoire simple de taille 200 provenant de chacune des bases. La population a été générée avec $N_d = 10\,000$ dans chaque domaine et les moyennes de domaine $\mu_a = 1$, $\mu_{ab} = 2$, $\mu_{ac} = 3$, $\mu_{abc} = 4$, $\mu_b = 5$, $\mu_{bc} = 6$, $\mu_c = 7$. Dans cette simulation, les bases de sondages B et C sont classées correctement et les profils d'erreur de classification pour la base de sondage A sont donnés dans le tableau. Nous avons aussi étudié d'autres moyennes de domaine, tailles de domaine de population et tailles d'échantillon en nous servant d'un plan factoriel ; les

résultats pour les autres conditions ont produit un profil similaire et ne sont pas présentés ici. L'estimateur fondé sur la multiplicité $\hat{Y}_{\text{ave}}$, avec $m_i = 1$ pour $i \in \{a, b, c\}$, $m_i = 1/2$ pour $i \in \{ab, ac, bc\}$, et $m_i = 1/3$ pour $i \in abc$, est optimal quand il n'y a pas d'erreur de classification et il est égal à l'estimateur pour base de sondage unique non ratissé. Les autres estimateurs à poids fixes étudiés sont $\hat{Y}_{2/3}$, avec $m^{(A,a)} = m^{(B,b)} = m^{(C,c)} = 1$, $m^{(A,ab)} = m^{(A,ac)} = m^{(A,abc)} = 2/3$, $m^{(B,ab)} = m^{(C,ac)} = 1/3$, et $m^{(B,abc)} = m^{(C,abc)} = 1/6$, et l'estimateur avec sélection $\hat{Y}_{\text{scr}}$, avec $m^{(A,a)} = m^{(B,b)} = m^{(C,c)} = m^{(A,ab)} = m^{(A,ac)} = m^{(A,abc)} = m^{(B,bc)} = 1$.

Tableau 1
Biais estimé pour l'erreur de classification sous deux bases de sondage, avec $n_A = n_B = 100$ et un échantillon aléatoire simple tiré de chaque base de sondage. PEA et PEB désignent les profils d'erreur de classification pour les bases de sondage A et B

PEA	PEB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	$\hat{Y}_H$	$\hat{Y}_{\text{FB}}$	$\hat{Y}_{\text{PMV}}$	$\hat{Y}_{\text{PVE}}$	$\hat{Y}_{\text{BU, rat}}$
a	a	-194	-87	-87	-194	-215	-215	-258	-68	10	-121	-119	-163
a	b	-5 015	4 145	4 529	5	-6 678	17	-10 002	-5 417	1 248	2 486	1 542	2 361
a	c	-5 142	-1 118	-898	-128	-6 823	-138	-10 185	-5 413	-2 583	-1 650	-2 482	-1 690
a	d	-57	-8 430	-8 431	-47	-69	-55	-92	30	-6 576	-6 723	-6 725	-6 795
b	a	1 163	-1 238	-1 290	-82	748	-82	-82	1 355	-2 376	-2 631	-2 551	-2 704
b	b	-3 724	3 040	3 264	43	-5 784	65	-9 905	-3 967	-920	-30	-850	-100
b	c	-3 882	-2 192	-2 187	-124	-5 977	-136	-10 167	-3 954	-4 319	-3 821	-4 477	-3 853
b	d	1 322	-9 445	-9 621	92	917	104	108	1 600	-8 219	-8 720	-8 531	-8 879
c	a	1 366	1 315	1 312	123	969	140	174	1 530	1 529	1 325	1 355	1 276
c	b	-3 729	5 456	5 948	51	-5 801	64	-9 945	-4 216	2 096	3 500	2 391	3 355
c	c	-3 797	235	512	-15	-5 868	2	-10 011	-4 089	-1 377	-417	-1 318	-466
c	d	1 285	-7 072	-7 212	56	873	60	48	1 535	-4 665	-5 131	-4 976	-5 222
d	a	-120	2 134	2 134	-111	-132	-126	-155	32	3 710	3 535	3 538	3 470
d	b	-4 979	6 497	7 086	34	-6 620	65	-9 901	-5 599	4 339	5 928	4 788	5 697
d	c	-5 152	1 174	1 644	-137	-6 835	-152	-10 200	-5 622	310	1 626	578	1 540
d	d	90	-5 999	-5 998	107	98	119	114	193	-2 964	-3 116	-3 120	-3 155

Tableau 2
$\sqrt{\text{EQM}}$ estimée pour l'erreur de classification sous deux bases de sondage, avec $n_A = n_B = 100$ et un échantillon aléatoire simple tiré de chaque base de sondage. PEA et PEB désignent les profils d'erreur de classification pour les bases de sondage A et B

PEA	PEB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	$\hat{Y}_H$	$\hat{Y}_{\text{FB}}$	$\hat{Y}_{\text{PMV}}$	$\hat{Y}_{\text{PVE}}$	$\hat{Y}_{\text{BU, rat}}$
a	a	9 646	7 917	7 910	9 646	9 729	9 729	10 304	9 677	8 151	8 081	8 115	8 075
a	b	10 602	9 351	9 531	9 926	11 531	10 197	14 181	11 157	8 212	8 377	8 198	8 311
a	c	10 779	8 622	8 603	10 071	11 715	10 402	14 376	11 243	8 817	8 514	8 720	8 508
a	d	9 789	11 719	11 704	9 674	9 884	9 795	10 432	9 819	10 979	10 978	11 003	11 007
b	a	9 623	8 182	8 185	9 718	9 686	9 766	10 307	9 780	8 446	8 447	8 444	8 459
b	b	9 955	9 054	9 137	9 995	10 949	10 212	14 069	10 489	8 074	7 913	7 995	7 898
b	c	10 146	9 014	9 014	10 160	11 197	10 489	14 404	10 616	9 443	9 108	9 448	9 114
b	d	9 868	12 600	12 716	9 826	9 952	9 927	10 567	10 023	12 063	12 284	12 188	12 371
c	a	9 843	8 185	8 180	9 887	9 853	9 877	10 341	9 991	8 516	8 417	8 442	8 402
c	b	10 049	10 113	10 396	10 039	11 029	10 229	14 127	10 662	8 520	8 863	8 529	8 778
c	c	10 247	8 701	8 718	10 254	11 233	10 534	14 306	10 799	8 762	8 527	8 669	8 516
c	d	10 021	10 861	10 936	9 966	10 068	10 016	10 579	10 177	10 113	10 211	10 168	10 240
d	a	9 795	8 127	8 121	9 734	9 845	9 788	10 343	9 829	9 158	9 024	9 042	8 991
d	b	10 718	10 601	10 970	10 001	11 602	10 258	14 149	11 358	9 461	10 157	9 595	9 986
d	c	10 847	8 558	8 650	10 099	11 769	10 426	14 387	11 424	8 674	8 707	8 608	8 664
d	d	9 945	10 070	10 057	9 778	10 019	9 885	10 510	9 986	9 458	9 412	9 449	9 417

Comme dans l'étude avec deux bases de sondage, les estimateurs corrigés du biais sont approximativement sans biais. L'estimateur avec sélection est également approximativement sans biais puisque $\mathcal{S}(A)$ seulement est classifié incorrectement. Les autres estimateurs présentent tous un biais important avec au moins certains des profils d'erreur de classification. Pour les conditions de simulation du tableau 5, les estimateurs poststratifiés, à base de sondage unique (BU) avec ratissage, de Hartley et PMV présentent un grand biais mais ont néanmoins une erreur quadratique moyenne plus petite que les estimateurs à poids fixes et corrigés du biais ; ce classement en fonction de l'EQM ne tient pas dans certaines des autres conditions de simulation.

Mecatti (2007) ainsi que Rao et Wu (2010) soutiennent que l'estimateur à poids fixes fondé sur la multiplicité $\hat{Y}_{\text{ave}}$ est sans biais si la seule erreur de classification est une erreur entre les domaines qui appartiennent au même nombre de bases de sondage. Classer incorrectement les observations du domaine *ab* dans le domaine *ac* (profil c) ne produit pas de biais, parce que l'ajustement des poids dans les deux domaines est $1/2$. Toutefois, en pratique, on s'attendrait à ce que le profil (c), avec deux erreurs d'appartenance à un domaine (ne pas déclarer l'appartenance à la base de sondage B et déclarer correctement l'appartenance à la base de sondage C), soit moins susceptible de survenir en pratique que l'erreur de classification d'une observation appartenant à ab comme appartenant à a ou à abc ; $\hat{Y}_{\text{ave}}$ peut être très sensible aux dernières formes d'erreur de classification. Bien qu'un estimateur à poids fixes ne soit pas sensible à l'erreur de classification dans les domaines dans lesquels les ajustements des poids sont égaux, dans ces simulations, chaque estimateur à poids fixes présente un biais important pour certains profils d'erreur de classification.

Les tableaux 1 à 5 montrent que chaque estimateur décrit à la section 2 peut présenter un biais grave dû à l'erreur de classification dans les domaines. Nous recommandons d'étudier la portée éventuelle de l'erreur de classification entre les domaines durant la phase de pré-test de l'enquête afin que cette information puisse être utilisée dans le plan de sondage. Si les probabilités d'erreur de classification sont connues avec précision, il pourrait être possible de choisir un estimateur à poids fixes qui n'est pas sensible à la forme présumée de l'erreur de classification. Quand on ne peut pas trouver d'estimateur robuste à l'erreur de classification ou quand cet estimateur est inefficace, les estimateurs à poids fixes peuvent être ajustés afin de réduire le biais. Il convient de souligner que les poids corrigés du biais proposés à la section 6.1 sont sensibles aux probabilités d'erreur de classification entrées. Ils ne tiennent pas compte d'autres erreurs non dues à l'échantillonnage, telle que la non-réponse ; l'application des ajustements des poids pour tenir compte de l'erreur de classification décrit à la section 6.1 suivie par les ajustements des poids pour la non-réponse décrits dans Brick et coll. (2011) peut produire des poids finaux qui ne corrigent ni l'erreur de classification ni la non-réponse. En présence à la fois d'erreur de classification dans les domaines et de non-réponse, des ajustements des poids traitant simultanément les deux problèmes sont nécessaires.

Tableau 3
Biais estimé pour l'erreur de classification sous deux bases de sondage, avec n_A = 200, n_B = 100, un échantillon en grappes tiré de la base de sondage A et un échantillon aléatoire simple tiré de la base de sondage B. PEA et PEB désignent les profils d'erreur de classification pour les bases de sondage A et B

PEA	PEB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	$\hat{Y}_H$	$\hat{Y}_{\text{FB}}$	$\hat{Y}_{\text{PMV}}$	$\hat{Y}_{\text{PVE}}$	$\hat{Y}_{\text{BU, rat}}$
a	a	-148	-142	-139	-148	-155	-155	-170	-312	63	-119	-172	-184
a	b	-5 090	4 199	4 599	-72	-6 774	-84	-10 144	-4 976	1 210	3 615	2 181	1 025
a	c	-5 069	-1 088	-851	-72	-6 759	-96	-10 139	-4 800	-1 994	177	-1 136	-3 216
a	d	-39	-8 379	-8 383	-35	-63	-58	-111	-237	-5 757	-5 909	-5 961	-6 996
b	a	1 168	-1 221	-1 258	-79	768	-63	-32	1 395	-1 690	-1 663	-2 514	-3 170
b	b	-3 716	2 979	3 236	60	-5 784	79	-9 918	-2 815	-86	1 776	-346	-2 087
b	c	-3 704	-2 108	-2 074	73	-5 771	92	-9 905	-2 561	-2 970	-1 410	-3 267	-5 814
b	d	1 317	-9455	-9 610	95	926	123	144	1 609	-7 285	-7 317	-7 938	-9 498
c	a	1 179	1 281	1 304	-66	772	-58	-41	1 486	1 831	1 652	943	840
c	b	-3 879	5 545	6 087	-118	-5 971	-126	-10 156	-2 972	3 532	4 597	2 405	1 683
c	c	-3 811	318	636	-44	-5 893	-42	-10 058	-2 671	110	1 128	-784	-2 328
c	d	1 423	-6 858	-6 973	191	1 022	206	220	1 824	-4 328	-4 014	-4 516	-5 624
d	a	-33	2 282	2 290	-28	-35	-32	-40	-148	3 627	3 138	3 103	3 728
d	b	-4 974	6 514	7 123	46	-6 660	30	-10 033	-4 863	4 768	6 274	4 742	4 549
d	c	-4 951	1 412	1 883	80	-6 621	84	-9 961	-4 682	1 357	2 863	1 451	388
d	d	42	-5 987	-5 991	53	40	52	37	-126	-2 899	-2 780	-2 791	-3 317

Tableau 4
$\sqrt{\text{EQM}}$ estimée pour l'erreur de classification sous deux bases de sondage, avec n_A = 200, n_B = 100, un échantillon en grappes tiré de la base de sondage A et un échantillon aléatoire simple tiré de la base de sondage B. PEA et PEB désignent les profils d'erreur de classification pour les bases de sondage A et B

PEA	PEB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	$\hat{Y}_H$	$\hat{Y}_{\text{FB}}$	$\hat{Y}_{\text{PMV}}$	$\hat{Y}_{\text{PVE}}$	$\hat{Y}_{\text{BU, rat}}$
a	a	10 916	8 912	8 899	10 916	11 092	11 092	11 879	10 975	9 250	9 155	10 109	9 418
a	b	11 786	10 186	10 324	11 157	12 743	11 503	15 463	12 253	8 906	9 391	10 123	9 231
a	c	11 983	9 575	9 537	11 409	12 922	11 814	15 600	12 395	9 575	9 279	10 391	10 039
a	d	11 042	12 357	12 375	10 941	11 250	11 173	12 056	11 051	11 591	11 605	12 229	12 053
b	a	10 698	9 133	9 154	10 872	10 921	11 049	11 875	10 823	9 255	9 151	10 195	9 766
b	b	10 957	9 803	9 867	11 071	12 033	11 413	15 262	11 215	8 681	8 748	9 610	9 182
b	c	11 115	9 860	9 846	11 272	12 172	11 675	15 361	11 306	9 721	9 252	10 558	10 970
b	d	10 988	13 269	13 408	11 046	11 222	11 262	12 143	11 084	12 484	12 347	13 279	13 598
c	a	10 995	9 090	9 073	11 106	11 187	11 254	12 028	11 125	9 309	9 190	9 798	9 389
c	b	11 104	10 779	11 015	11 090	12 162	11 380	15 348	11 430	9 450	9 724	9 754	9 144
c	c	11 155	9 425	9 400	11 189	12 234	11 600	15 424	11 389	9 219	9 064	9 868	9 658
c	d	10 922	11 328	11 421	10 896	11 121	11 091	11 929	11 017	10 759	10 456	11 151	11 182
d	a	11 011	9 080	9 045	10 920	11 181	11 103	11 913	11 041	9 873	9 579	10 375	10 135
d	b	11 838	11 357	11 669	11 164	12 723	11 453	15 299	12 337	10 258	10 848	11 009	10 403
d	c	11 804	9 334	9 371	11 159	12 707	11 548	15 298	12 224	9 349	9 507	10 102	9 442
d	d	11 179	10 839	10 854	10 989	11 355	11 199	12 059	11 195	10 440	10 302	10 916	10 519

Tableau 5
Biais estimé et $\sqrt{\text{EQM}}$ pour l'erreur de classification dans une enquête à trois bases de sondage, avec $n_A = n_B = n_C = 200$ et un échantillon aléatoire simple tiré de chaque base de sondage. PEA désigne le profil d'erreur de classification pour la base de sondage A. Le profil (a) ne présente aucune erreur de classification ; (b) $\phi^A_{aa} = 0{,}8$, $\phi^A_{a,ab} = 0{,}1$, $\phi_{a,abc} = 0{,}1$, $\phi^A_{ab,ab} = 1$, $\phi^A_{ac,ac} = 1$, $\phi^A_{abc,abc} = 1$; (c) $\phi^A_{aa} = 1$, $\phi^A_{ab,ab} = 0{,}9$, $\phi^A_{ab,ac} = 0{,}1$, $\phi^A_{ac,ac} = 1$, $\phi^A_{abc,abc} = 1$; (d) $\phi^A_{aa} = 1$, $\phi^A_{ab,ab} = 0{,}9$, $\phi^A_{ab,abc} = 0{,}1$, $\phi^A_{ac,ac} = 1$, $\phi^A_{abc,abc} = 1$; (e) $\phi^A_{aa} = 1$, $\phi^A_{ab,ab} = 0{,}8$, $\phi^A_{ab,a} = 0{,}1$, $\phi_{ab,abc} = 0{,}1$, $\phi^A_{ac,ac} = 1$, $\phi^A_{abc,abc} = 1$

	PEA	$\hat{Y}_{\text{ave}}$	$\hat{Y}_{\text{ave, post1}}$	$\hat{Y}_{\text{ave, post2}}$	$\hat{Y}_{\text{ave, }bc}$	$\hat{Y}_{2/3}$	$\hat{Y}_{2/3,\, bc}$	$\hat{Y}_{\text{scr}}$	$\hat{Y}_H$	$\hat{Y}_{\text{PMV}}$	$\hat{Y}_{\text{BU, rat}}$
	a	-8	31	28	-8	5	5	20	9	-26	-208
	b	-938	-1 409	-1 478	57	-586	77	107	-2 039	-5 676	-5 624
biais	c	-26	-485	-508	-26	6	6	6	-324	-825	-957
	d	-231	-514	-557	104	108	108	85	-326	-1 321	-1 438
	e	704	287	247	34	697	27	-4	1 488	1 420	1 193
	a	9 003	4 419	4 410	9 003	10 013	10 013	13 108	7 990	7 281	7 293
	b	8 961	4 711	4 730	8 955	9 952	9 953	13 092	8 085	9 107	9 074
$\sqrt{\text{EQM}}$	c	9 119	4 432	4 422	9 119	10 140	10 140	13 238	8 112	7 396	7 422
	d	8 894	4 405	4 405	8 893	9 874	9 874	12 919	7 957	7 414	7 433
	e	9 088	4 438	4 424	9 059	10 071	10 046	13 180	8 254	7 621	7 581

7. Problèmes de plan de sondage

Comme il est mentionné à la section 1, les plans à bases de sondage multiples peuvent donner une meilleure couverture et une meilleure précision qu'une enquête à base de sondage unique de coût équivalent. Toutefois, le problème d'élaboration du plan de sondage est plus complexe que dans le cas d'une enquête à base de sondage unique, puisqu'un plan de sondage qui est optimal pour la base de sondage A et pour la base de sondage B séparément pourraient ne pas l'être pour l'échantillon combiné. De même, un plan de sondage qui est optimal quand l'estimateur $\hat{Y}(1/2)$ est utilisé peut ne pas l'être pour $\hat{Y}_{\text{PMV}}$.

Hartley (1962, 1974) a obtenu des plans de sondage optimaux pour l'estimateur $\hat{Y}(\hat{\theta}_H)$ quand un échantillon aléatoire simple est tiré de chaque base de sondage. Les tailles optimales d'échantillons n_A et n_B dépendent des coûts relatifs de l'échantillonnage dans les deux bases de sondage ainsi que des moyennes et des variances de la variable de réponse à l'intérieur des domaines. Cochran (1977, pages 144-145) a décrit l'enquête à base de sondage double de la figure 1 dans son chapitre sur l'échantillonnage stratifié. Dans cette situation, N_a et N_{ab} peuvent être connus, surtout si la base de sondage B est une liste. Les domaines a et ab sont traités comme des strates ; il existe un échantillon provenant de la strate a et deux échantillons

indépendants provenant de la strate *ab*. Le problème de plan de sondage peut être approché comme un plan de sondage stratifié.

En général, le plan optimal est une fonction des variances d'échantillonnage et des erreurs non dues à l'échantillonnage dans chaque base de sondage, ainsi que de l'estimateur choisi. Biemer (1984) ainsi que Lepkowski et Groves (1986) ont discuté de plans de sondage pour la situation de la figure 1 quand un échantillon à plusieurs degrés est tiré de chaque base de sondage en utilisant l'estimateur de Hartley $\hat{Y}(\hat{\theta}_H)$. Lepkowski et Groves (1986) ont tenu compte de la variabilité due à l'intervieweur et du biais dû au mode, ainsi que de l'erreur d'échantillonnage pour évaluer la précision des divers plans ; de plus grandes tailles d'échantillons sont affectées aux bases de sondage dont le biais de mode est plus faible. Brick (2010) a obtenu les répartitions optimales en présence de non-réponse et constaté que tenir compte de la non-réponse pour affecter les ressources aux deux bases de sondage peut accroître considérablement l'efficacité dans les enquêtes à base de sondage double avec sélection ainsi qu'avec chevauchement.

L'un des avantages d'un plan à bases de sondage multiples est sa souplesse ; il convient bien pour une approche modulaire du plan de sondage. Dans certaines situations, il pourrait être pratique de tirer un échantillon initial de la population générale (base de sondage A dans la figure 4). La conception des échantillons provenant des bases de sondage B et C, qui correspondent à des sous-populations d'intérêt, peut alors être déterminée en se servant de l'information figurant dans l'échantillon de la base de sondage A. Par exemple, si l'échantillon provenant de la base de sondage A produit un trop petit nombre d'ingénieurs, la taille de l'échantillon tiré d'une liste de membres d'une association d'ingénieurs peut être augmentée en conséquence.

Rao (2003) a proposé d'utiliser des enquêtes à bases de sondage multiples pour améliorer l'exactitude des estimations sur petits domaines dans les sous-groupes d'intérêt. Dans cette application, des enquêtes supplémentaires peuvent être réalisées à l'aide de bases de sondage à forte concentration des sous-groupes d'intérêt. À mesure que les besoins de recherche évoluent, les ressources peuvent être réaffectées entre les enquêtes supplémentaires sans changer le plan de l'enquête principale. Une enquête sur la victimisation s'appuyant sur une base aréolaire nationale peut être complétée par des enquêtes locales sur la victimisation ; à mesure que les profils de victimisation changent, les enquêtes locales peuvent être réalisées auprès d'échantillons de diverses tailles ou être transférées vers d'autres régions géographiques.

La plupart des plans de sondage ont pour objectif d'être efficaces pour les réponses importantes, mais dans certaines situations, un plan qui est efficace pour une réponse est inefficace pour d'autres. Pour une enquête dans laquelle chacune des quatre réponses d'intérêt était fortement corrélée avec l'une des variables de stratification possibles (mais pas nécessairement corrélée aux autres variables de stratification), Skinner, Holmes et Holt (1994) ont utilisé une enquête à bases de sondage multiples avec quatre échantillons stratifiés indépendants tirés d'une base de sondage commune. Chaque échantillon a été stratifié en utilisant la variable de stratification qui était corrélée avec l'une des réponses d'intérêt et était très efficace pour la réponse en question. Dans l'estimation, l'information provenant des quatre échantillons a été combinée.

Les enquêtes à bases de sondage multiples peuvent aussi être utilisées conjuguées à des méthodes d'échantillonnage séquentiel ou adaptatif afin d'améliorer le rendement pour une population rare ou difficile à joindre tels que les nouveaux immigrants. Par exemple, un plan de sondage stratifié à plusieurs degrés pourrait être employé pour la base de sondage A, tandis qu'un plan de sondage en grappes adaptatif (Thompson 2002) pourrait être utilisé pour la base de sondage B. Les estimations de domaine peuvent être calculées séparément pour les deux plans de sondage, puis combinées en utilisant les méthodes de la section 2. Dans cette situation, les bases de sondage A et B peuvent se chevaucher entièrement, de sorte que l'erreur de classification dans les domaines ne sera pas un problème.

8. Conclusion

Dans le présent article, nous avons résumé un problème que pose l'utilisation des méthodes à bases de sondage multiples pour les enquêtes-ménages américaines. Les plans à bases de sondage multiples offrent d'énormes possibilités d'améliorer l'efficacité de la collecte des données dans les enquêtes-ménages. Ils peuvent accroître la couverture en combinant des bases de sondage incomplètes, améliorer l'exactitude des estimations pour des sous-groupes ou populations rares, et accroître la souplesse et la capacité d'adaptation de la collecte des données fédérales. Les enquêtes à bases de sondage multiples peuvent faciliter l'échantillonnage de populations difficiles à joindre, tels que les nouveaux immigrants ou les ménages avec des nourrissons ; une enquête de population générale peut être combinée avec un plan de sondage adaptatif ou avec une liste de naissances.

Dans de nombreux cas, les enquêtes à bases de sondage multiples peuvent fournir des estimations plus précises des quantités de population sans accroître les coûts de collecte des données, mais le plan et l'estimateur doivent être choisis prudemment afin de réaliser ces économies. Une enquête à

bases de sondage multiples peut, comme les autres enquêtes, présenter une non-réponse, des effets de mode et des erreurs de mesure. En outre, à moins que toutes les bases de sondage correspondent à la population complète, les estimateurs pour enquête à bases de sondage multiples peuvent être sensibles à l'erreur de classification dans les domaines. Une correction pour l'erreur de classification a été donnée à la section 6, mais une étude plus approfondie de ces défis est nécessaire. Les effets de l'erreur de classification dans les domaines, de la non-réponse et du biais de mode peuvent être confondus. Une expérience peut aider à débrouiller ces effets. Nous étudions à l'heure actuelle la relation entre ces trois types d'erreurs non liées à l'échantillonnage. Chaque forme d'erreur non due à l'échantillonnage a une incidence sur l'exactitude des estimateurs pour bases de sondage multiples, et les erreurs non dues à l'échantillonnage attendues doivent être intégrées dans un plan de sondage optimal.

Remerciements

Cette étude a été financée en partie par la National Science Foundation aux termes des subventions SES-0604373 et DRL-0909630. Une partie de la matière du présent article a été présentée à l'assemblée annuelle de 2010 de la Société de statistique de l'Italie et publiée dans les actes de cette conférence. L'auteur remercie les examinateurs de leurs commentaires utiles.

Bibliographie

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

Biemer, P.P. (1984). Methodology for optimal dual frame sample design. Bureau of the Census SRD Research Report CENSUS/SRD/RR-84/07.

Brick, J.M. (2010). Dual frame landline and cell phone surveys. Article présenté à l'Annual meeting of the Statistical Society of Canada, Québec.

Brick, J.M., Cervantes, I.F., Lee, S. et Norman, G. (2011). Erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double. *Techniques d'enquête*, 37, 1, 1-16.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. et Yuan, Y. (2006). Nonresponse bias in a dual frame survey of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Chambers, R., Chipperfield, J., Davis, W. et Kovačević, M. (2008). Inference based on estimating equations and probability-linked data. University of Wollongong Centre for Statistical & Survey Methodology Working Paper 18-09.

Citro, C.F., et Kalton, G., Éds. (2007). *Using the American Community Survey*: *Benefits and Challenges*. Washington, D.C. : National Academies Press.

Clark, J., Winglee, M. et Liu, B. (2007). Handling imperfect overlap determination in a dual-frame survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3233-3238.

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

de Leeuw, E. (2008). Choosing the method of data collection. Dans *International Handbook of Survey Methodology*, (Éds., E. de Leeuw, J. Hox et D. Dillman). New York : Lawrence Erlbaum, 113-135.

de Leeuw, E., Hox, J. et Dillman, D. (2008). Mixed-mode surveys: When and why. Dans *International Handbook of Survey Methodology*, (Éds., E. de Leeuw, J. Hox et D. Dillman). New York : Lawrence Erlbaum, 299-316.

Demnati, A., Rao, J.N.K., Hidiroglou, M.A. et Tambay, J.-L. (2007). On the allocation and estimation for dual frame survey data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2938-2945.

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

González-Villalobos, A., et Wallace, M.A. (1996). *Multiple Frame Agriculture Surveys*. Rome : Food and Agriculture Organization of the United Nations. Vols. 1 et 2.

Haines, D.E., et Pollock, K.H. (1998). Combinaison de bases multiples pour estimer la taille et les chiffres de la population. *Techniques d'enquête*, 24, 81-91.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York : John Wiley & Sons, Inc. Volume 1.

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

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, Séries C, 36, 99-118.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Iachan, R., et Dennis, M.L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics*, 9, 747-764.

Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, Séries A, 149, 65-82.

Kennedy, C. (2007). Evaluating the effects of screening for telephone service in dual frame RDD surveys. *Public Opinion Quarterly*, 71, 750-771.

Kott, P.S., Amrhein, J.F. et Hicks, S.D. (1998). Échantillonnage et estimation à partir de bases de sondage listes multiples. *Techniques d'enquête*, 24, 3-10.

Lepkowski, J.M., et Groves, R.M. (1986). A mean squared error model for multiple frame, mixed mode survey design. *Journal of the American Statistical Association*, 81, 930-937.

Lesser, V.M., et Kalsbeek, W.D. (1999). Nonsampling errors in environmental surveys. *Journal of Agricultural*, *Biological*, *and Environmental Statistics*, 4, 473-488.

Lohr, S.L. (2007). Recent developments in multiple frame surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3257-3264.

Lohr, S.L. (2009). Multiple frame surveys. Dans *Handbook of Statistics*, *Sample Surveys*: *Design*, *Methods and Applications*, (Éds., D. Pfeffermann et C.R. Rao). Amsterdam : North Holland, Vol. 29A, 71-88.

Lohr, S.L., et Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Lu, W., Brick, J.M. et Sitter, R. (2006). Algorithms for constructing combined strata variance estimators. *Journal of the American Statistical Association*, 101, 1680-1692.

Lu, Y., et Lohr, S.L. (2010). L'estimation des flux bruts dans les enquêtes à base de sondage double. *Techniques d'enquête*, 36, 13-24.

Mecatti, F. (2007). Un estimateur à base de sondage unique fondé sur la multiplicité pour les sondages à bases multiples *Techniques d'enquête*, 33, 171-178.

National Science Foundation (2003). *SESTAT*: *Design and Methodology*, http://srsstats.sbe.nsf.gov/docs/techinfo.html.

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

Rao, J.N.K., et Wu, C. (2010). Pseudo-empirical likelihood inference for dual frame surveys. *Journal of the American Statistical Association*, 105, 1494-1503.

Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J., Holmes, D.J. et Holt, D. (1994). Multiple frame sampling for multivariate stratification. *Revue Internationale de Statistique*, 62, 333-347.

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

Thompson, S.K. (2002). *Sampling Techniques*, 2e Éd. New York : John Wiley & Sons, Inc.

Vannieuwenhuyze, J., Loosveldt, G. et Molenberghs, G. (2011). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74, 1027-1045.