Méthodes d'estimation sur bases de sondage multiples dans le cadre de plans de sondage à deux degrés

Guillaume CHAUVET(*), Guylène TANDEAU DE MARSAC(**)

(*)ENSAI (CREST)

(**) INSEE, Direction Régionale de Lille

Résumé

Lorsqu'on s'intéresse à une population finie, il arrive qu'il soit nécessaire de tirer des échantillons dans plusieurs bases de sondage pour représenter l'ensemble des individus. Nous nous intéressons ici au cas de deux échantillons sélectionnés selon un plan à deux degrés, avec un premier degré de tirage commun. Nous appliquons les méthodes de Hartley (1962), Bankier (1986), et Kalton et Anderson (1986), et nous montrons que ces méthodes peuvent être appliquées conditionnellement au premier degré de tirage. Nous comparons également la performance de plusieurs estimateurs dans le cadre d'une étude par simulations. Nos résultats suggèrent que le choix d'un estimateur en présence de bases de sondage multiples se fasse de façon prudente, et qu'un estimateur simple est parfois préférable même s'il n'utilise qu'une partie de l'information collectée.

Abstract

When studying a finite population, it is sometimes necessary to select samples from several sampling frames in order to represent all individuals. Here we are interested in the scenario where two samples are selected using a two-stage design, with common first-stage selection. We apply the Hartley (1962), Bankier (1986), and Kalton and Anderson (1986) methods, and we show that these methods can be applied conditional on first-stage selection. We also compare the performance of several estimators as part of a simulation study. Our results suggest that the estimator should be chosen carefully when there are multiple sampling frames, and that a simple estimator is sometimes preferable, even if it uses only part of the information collected.

Mots-clés

Enquête à extension, estimateur de Hansen-Hurwitz, estimateur de Horvitz-Thompson, sondage à deux degrés

Introduction

Lorsqu'on s'intéresse à une population finie, il arrive qu'aucune base de sondage ne la recouvre totalement et qu'il soit nécessaire de tirer des échantillons dans deux bases de sondage (ou plus) pour représenter l'ensemble des individus. Pour mettre en commun ces échantillons, de nombreuses méthodes d'estimation sur bases de sondage multiples ont été proposées (Hartley, 1962 ; Bankier, 1986 ; Kalton et Anderson, 1986 ; Mecatti, 2007 ; Rao et Wu, 2010) ; voir également les articles de revue de Lohr (2009, 2011), et les articles référencés, pour un panorama complet. Notons que la méthode de Mecatti (2007) s'inspire des travaux de Lavallée (2002, 2007) sur la Méthode généralisée du partage des poids.

Nous nous intéressons ici au cas de deux échantillons sélectionnés selon un plan à deux degrés, avec un premier degré de tirage commun. Ce cadre correspond aux enquêtes de l'INSEE avec extension : un premier échantillon de logements est sélectionné dans les communes de l'Echantillon-maître (Bourdalle et al., 2000), et un second échantillon est sélectionné et enquêté dans les communes du même Echantillon-maître afin de cibler une sous-population spécifique. On dispose de deux mesures d'enquêtes provenant de deux échantillons indépendants au deuxième degré du plan de sondage. Nous appliquons des méthodes d'estimation sur bases de sondage multiples pour la mise en commun de ces deux échantillons. Nous montrons que les estimateurs étudiés peuvent dans ce contexte être calculés conditionnellement au premier degré de tirage, ce qui simplifie leur calcul notamment pour l'estimateur optimal de Hartley (1962). Nous comparons également les performances de ces estimateurs dans le cadre d'une étude par simulations.

Ce travail est extrait du Mémoire de Master de Statistique Publique de Guylène Tandeau de Marsac. Il correspond à l'article : Chauvet et Tandeau de Marsac (2014), paru dans la revue Techniques d'Enquête de Statistique Canada.

1 Estimation pour des bases de sondage chevauchantes

On considère une population finie U sur laquelle est définie une variable d'intérêt y de valeur y_k pour l'individu k. Si on sélectionne dans U un échantillon S avec des probabilités d'inclusion π_k, l'estimateur

$$\hat{Y} = \sum_{k \in S} \pi_k^{-1} y_k$$

proposé par Narain (1951) et Horvitz et Thompson (1952) est sans biais pour le total $Y = \sum_{k \in U} y_k$ si toutes les probabilités π_k sont strictement positives.
Nous nous intéressons au cas où la population est entièrement couverte par deux bases de sondage chevauchantes U_A et U_B. En utilisant les notations de Lohr (2011), soient $a = U_A \setminus U_B$ le domaine couvert par U_A seulement ; $b = U_B \setminus U_A$ le domaine couvert par U_B seulement ; $ab = U_A \cap U_B$ le domaine couvert à la fois par U_A et U_B. On sélectionne dans U_A un échantillon S^A avec des probabilités d'inclusion $\pi_k^A > 0$. Pour tout domaine $d \subset U_A$, le sous-total

$$Y_d = \sum_{k \in d} y_k$$

est estimé sans biais par

$$\hat{Y}_d^A = \sum_{k \in S_A} d_k^A y_k 1(k \in d)$$

avec $d_k^A = (\pi_k^A)^{-1}$. On sélectionne dans U_B un échantillon S^B avec des probabilités d'inclusion $\pi_k^B > 0$. Pour tout domaine $d \subset U_B$, le sous-total Y_d est estimé sans biais par

$$\hat{Y}_d^B = \sum_{k \in S_B} d_k^B y_k 1(k \in d)$$

avec $d_k^B = (\pi_k^B)^{-1}$. L'objectif est de combiner les échantillons S^A et S^B pour obtenir une estimation de Y aussi précise que possible.

1.1 Estimateur de Hartley

Hartley (1962) propose la classe d'estimateurs sans biais

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

avec θ un paramètre à déterminer. Le choix $\theta = 1/2$ conduit à donner aux échantillons S^A et S^B le même poids pour l'estimation sur le domaine intersection ab. Hartley (1962) propose de choisir le paramètre qui minimise la variance de $\hat{Y}_\theta$. Cela conduit à

$$\theta_{opt} = \frac{Cov(\hat{Y}_a^A + \hat{Y}_{ab}^B + \hat{Y}_b^B, \hat{Y}_{ab}^B - \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^B - \hat{Y}_{ab}^A)}, \quad (2)$$

que l'on peut réécrire sous la forme

$$\theta_{opt} = \frac{V(\hat{Y}_{ab}^B) + Cov(\hat{Y}_{ab}^B, \hat{Y}_b^B) - Cov(\hat{Y}_a^A, \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)} \quad (3)$$

quand les échantillons S^A et S^B sont indépendants. Comme le remarque Lohr (2007), le coefficient optimal θ_{opt} peut ne pas être compris entre 0 et 1 si un terme de covariance présent dans (3) est grand. Supposons pour simplifier que $Cov(\hat{Y}_{ab}^B, \hat{Y}_b^B) = 0$, ce qui est le cas si b et ab sont utilisés comme strates dans la sélection de S^B. Alors $\theta_{opt} > 1$ si et seulement si $Cov(\hat{Y}^A, \hat{Y}_{ab}^A) < 0$. Dans le cas où S^A est sélectionné par sondage aléatoire simple, ce sera par exemple le cas si dans U_A les faibles valeurs de la variable y sont concentrées dans le domaine ab.

En pratique, les termes de variance et de covariance sont inconnus et doivent être remplacés par des estimateurs, ce qui introduit une variabilité supplémentaire. Un autre inconvénient est que le paramètre optimal dépend de la variable d'intérêt considérée. Si des estimateurs optimaux sont calculés pour différents variables d'intérêt, les estimations peuvent souffrir d'une incohérence interne (Lohr, 2011).

1.2 Estimateur de Kalton et Anderson

Une classe plus générale d'estimateurs s'obtient en remarquant que le total Y peut se réécrire

$$Y = Y_a + \sum_{k \in ab} \theta_k y_k + \sum_{k \in ab} (1-\theta_k) y_k + Y_b,$$

avec θ_k un coefficient propre à l'individu k. Kalton et Anderson (1986) proposent le choix $\theta_k = (d_k^A + d_k^B)^{-1} d_k^B$, qui conduit à l'estimateur

$$\hat{Y}_{KA} = \sum_{k \in S_A} d_k^A m_k^A y_k + \sum_{k \in S^B} d_k^B m_k^B y_k \quad (4)$$

avec d'une part $m_k^A = 1$ si $k \in a$ et $m_k^A = \theta_k$ si $k \in ab$, d'autre part $m_k^B = 1$ si $k \in b$ et $m_k^B = 1 - \theta_k$ si $k \in ab$. Les poids d'estimation sont les mêmes quelle que soit la variable d'intérêt, ce qui assure la cohérence interne des estimations ; en revanche, l'estimateur de Kalton et Anderson est moins efficace que l'estimateur optimal de Hartley pour une variable d'intérêt donnée. Notons qu'il s'agit d'un estimateur de type Hansen-Hurwitz (1943), qui peut se réécrire sous la forme $\hat{Y}_{KA} = \sum_{k \in U} \frac{W_k}{E(W_k)}\, y_k$ en notant $W_k = 1(k \in S^A) + 1(k \in S^B)$ le nombre de fois où l'unité k est sélectionnée dans l'échantillon réunion $S^A \cup S^B$. On a en particulier $E(W_k) = \pi_k^A + \pi_k^B$.

1.3 Estimateur de Bankier

Bankier (1986) propose d'utiliser un estimateur de type Horvitz-Thompson, en calculant les probabilités d'inclusion dans l'échantillon réunion

$$\pi_k^{HT} \equiv P(k \in S^A \cup S^B) = \pi_k^A + \pi_k^B - Pr(k \in S^A \cap S^B).$$

Si les échantillons S^A et S^B sont indépendants, on obtient

$$\pi_k^{HT} = \pi_k^A + \pi_k^B - \pi_k^A \pi_k^B$$

et l'estimateur

$$\hat{Y}_{HT} = \sum_{k \in S^A \cup S^B} \frac{y_k}{\pi_k^{HT}} = \sum_{k \in S^A \cap a} \frac{y_k}{\pi_k^A} + \sum_{k \in S^B \cap b} \frac{y_k}{\pi_k^B} + \sum_{k \in (S^A \cup S^B) \cap ab} \frac{1}{\pi_k^A + \pi_k^B - \pi_k^A \pi_k^B} y_k. \quad (5)$$

2 Estimation avec un premier degré de tirage commun

Nous étudions ici le cas de deux échantillons sélectionnés selon un plan à deux degrés, avec un premier degré de tirage commun. La population U est partitionnée pour obtenir une population $U_I = \{u_1, \ldots, u_M\}$ de M unités primaires. Au premier degré, on sélectionne un échantillon S_I d'unités primaires (UP) avec une probabilité de tirage π_{Ii} pour une UP u_i. Au second degré, dans chaque unité primaire $u_i \in S_I$, on sélectionne : un échantillon S_i^A dans $u_i^A \equiv u_i \cap U_A$, avec une probabilité de sélection (conditionnelle) $\pi_{k|i}^A > 0$ pour $k \in u_i^A$; un échantillon S_i^B dans $u_i^B \equiv u_i \cap U_B$, avec une probabilité de sélection (conditionnelle) $\pi_{k|i}^B > 0$ pour l'unité $k \in u_i^B$. Nous faisons les hypothèses suivantes, habituelles pour un tirage à deux degrés : le second degré de tirage au sein de l'unité primaire u_i ne dépend que de i ; entre deux unités primaires $u_i \neq u_j \in S_I$, les échantillons S_i^A et S_j^A (respectivement, S_i^B et S_j^B) sont indépendants conditionnellement à S_I (propriété d'indépendance). Nous supposons également qu'au sein de chaque unité primaire $u_i \in S_I$, les sous-échantillons S_i^A et S_i^B sont indépendants conditionnellement à S_I.

Pour un domaine $d_1 \subset U_A$, le sous-total Y_{d_1} est estimé par

$$\hat{Y}_{d_1}^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{d_1,i}^A$$

avec $d_{Ii} = (\pi_{Ii})^{-1}$ le poids de sondage de l'unité primaire u_i, avec

$$\hat{Y}^A_{d_1,i} = \sum_{k \in S_i^A} d^A_{k|i} y_k 1(k \in d_1)$$

l'estimateur du sous-total $Y_{d_1,i} = \sum_{k \in u_i} y_k 1(k \in d_1)$ sur $d_1 \cap u_i$, et $d^A_{k|i} = (\pi^A_{k|i})^{-1}$ le poids de sondage de k dans u_i^A. Pour un domaine $d_2 \subset U_B$, le sous-total Y_{d_2} est estimé par

$$\hat{Y}^B_{d_2} = \sum_{u_i \in S_I} d_{Ii} \hat{Y}^B_{d_2,i}$$

avec

$$\hat{Y}^B_{d_2,i} = \sum_{k \in S_i^B} d^B_{k|i} y_k 1(k \in d_2)$$

l'estimateur du sous-total $Y_{d_2,i}$, et $d^B_{k|i} = (\pi^B_{k|i})^{-1}$ le poids de sondage de k dans u_i^B. On obtient en particulier les estimateurs

$$\hat{Y}^A_{ab} = \sum_{u_i \in S_I} d_{Ii} \hat{Y}^A_{ab,i} \quad \text{où} \quad \hat{Y}^A_{ab,i} = \sum_{k \in S_i^A} d^A_{k|i} y_k 1(k \in ab), \tag{6}$$

$$\hat{Y}^A_{b} = \sum_{u_i \in S_I} d_{Ii} \hat{Y}^A_{b,i} \quad \text{où} \quad \hat{Y}^A_{b,i} = \sum_{k \in S_i^A} d^A_{k|i} y_k 1(k \in b), \tag{7}$$

$$\hat{Y}^B_{ab} = \sum_{u_i \in S_I} d_{Ii} \hat{Y}^B_{ab,i} \quad \text{où} \quad \hat{Y}^B_{ab,i} = \sum_{k \in S_i^B} d^B_{k|i} y_k 1(k \in ab). \tag{8}$$

2.1 Estimateur de Hartley

L'estimateur de Hartley donné en (1) peut se réécrire sous la forme

$$\hat{Y}_\theta \;=\; \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{\theta,i} \tag{9}$$

avec $\hat{Y}_{\theta,i} = \hat{Y}^A_{a,i} + \theta \hat{Y}^A_{ab,i} + (1-\theta)\hat{Y}^B_{ab,i} + \hat{Y}^B_{b,i}$ l'estimateur de Hartley du sous-total Y_i sur l'unité primaire u_i. On obtient $E(\hat{Y}_\theta | S_I) = \sum_{i \in S_I} d_{Ii} Y_i$, puis

$$V(\hat{Y}_\theta) \;=\; V\left(\sum_{i \in S_I} d_{Ii} Y_i \right) + EV(\hat{Y}_\theta | S_I). \tag{10}$$

Dans (10), le premier terme du membre de droite ne dépend pas de θ. L'estimateur optimal de Hartley peut donc se calculer en minimisant seulement le second terme. On obtient :

$$\theta_{opt|S_I} \;=\; \frac{EV(\hat{Y}^B_{ab}|S_I) + ECov(\hat{Y}^B_{ab}, \hat{Y}^B_b | S_I) - ECov(\hat{Y}^A_a, \hat{Y}^A_{ab}|S_I)}{EV(\hat{Y}^A_{ab}|S_I) + EV(\hat{Y}^B_{ab}|S_I)}, \tag{11}$$

que l'on peut estimer par

$$\hat{\theta}_{opt} \;=\; \frac{\hat{V}(\hat{Y}^B_{ab}) + \widehat{Cov}(\hat{Y}^B_{ab}, \hat{Y}^B_b) - \widehat{Cov}(\hat{Y}^A_a, \hat{Y}^A_{ab})}{\hat{V}(\hat{Y}^A_{ab}) + \hat{V}(\hat{Y}^B_{ab})} \tag{12}$$

en remplaçant chaque terme de variance et de covariance par un estimateur sans biais conditionnellement au premier degré.

2.2 Estimateur de Kalton et Anderson

Avec le plan de sondage considéré, on a $d_k^A = d_{Ii} d_{k|i}^A$ pour toute unité $k \in u_i^A$, et $d_k^B = d_{Ii} d_{k|i}^B$ pour toute unité $k \in u_i^B$. L'estimateur de Kalton et Anderson donné en (4) peut donc se réécrire

$$\hat{Y}_{KA} = \sum_{i \in S_I} d_{Ii} \hat{Y}_{KA,i} \tag{13}$$

avec $\hat{Y}_{KA,i} = \sum_{k \in S^A} d_{k|i}^A m_{k|i}^A y_k + \sum_{k \in S^B} d_{k|i}^B m_{k|i}^B y_k$ l'estimateur de Kalton et Anderson du sous-total Y_i, où

$$m_{k|i}^A = \begin{cases} 1 & \text{si } k \in a \cap u_i, \\ \frac{d_{k|i}^B}{d_{k|i}^A + d_{k|i}^B} & \text{si } k \in ab \cap u_i, \end{cases} \quad \text{et} \quad m_{k|i}^B = \begin{cases} 1 & \text{si } k \in b \cap u_i, \\ \frac{d_{k|i}^A}{d_{k|i}^A + d_{k|i}^B} & \text{si } k \in ab \cap u_i. \end{cases}$$

2.3 Estimateur de Bankier

Avec le plan de sondage considéré, on a

$$\pi_k^{HT} = \pi_{Ii}(\pi_{k|i}^A + \pi_{k|i}^B - \pi_{k|i}^A \pi_{k|i}^B)$$

pour tout $k \in u_i$. L'estimateur de Bankier donné en (5) peut donc se réécrire

$$\hat{Y}_{HT} = \sum_{i \in S_I} d_{Ii} \hat{Y}_{HT,i} \tag{14}$$

avec $\hat{Y}_{HT,i} = \sum_{k \in S_i^A \cup S_i^B} \frac{y_k}{\pi_{k|i}^{HT}}$ l'estimateur de Bankier pour le sous-total Y_i, et

$$\pi_{k|i}^{HT} = \begin{cases} \pi_{k|i}^A & \text{si } k \in a, \\ \pi_{k|i}^B & \text{si } k \in b, \\ \pi_{k|i}^A + \pi_{k|i}^B - \pi_{k|i}^A \pi_{k|i}^B & \text{si } k \in ab. \end{cases}$$

Chacun des trois estimateurs étudiés s'obtient donc en appliquant la méthode d'estimation unité primaire par unité primaire, conditionnellement au premier degré. Ce résultat est particulièrement intéressant pour la méthode optimale de Hartley, puisque l'estimateur du coefficient optimal donné en (12) ne nécessite que des estimateurs de variance conditionnels au premier degré.

3 Etude par simulations

Nous utilisons des populations artificielles proposées par Saigo (2010). Nous générons deux populations, contenant chacune $M = 200$ unités primaires regroupées en $H = 4$ strates U_{Ih} de taille $M_h = 50$. Chaque unité primaire u_{hi} contient $N_{hi} = 100$ unités secondaires. Dans chaque population, nous générons pour chaque unité primaire $u_{hi} \in U_{Ih}$:

$$\mu_{hi} = \mu_h + \sigma_h v_{hi} \tag{15}$$

où les valeurs μ_h et σ_h sont celles utilisées par Saigo (2010). Le terme σ_h^2 permet de contrôler la dispersion entre les unités primaires. Les v_{hi} sont générés de façon iid selon

une loi normale centrée réduite $N(0,1)$. Pour chaque unité $k \in u_{hi}$, nous générons ensuite la valeur y_k selon le modèle

$$y_k = \mu_{hi} + \{\rho^{-1}(1-\rho)\}^{0.5}\sigma_h v_k, \tag{16}$$

où les v_k sont générés de façon iid selon une loi normale centrée réduite. Le terme de variance dans le modèle (16) permet d'obtenir un coefficient de corrélation intra-grappes approximativement égal à ρ. En particulier, plus le coefficient ρ est grand, moins les valeurs y_k sont dispersées dans les unités primaires. Nous utilisons $\rho = 0.2$ pour la population 1 et $\rho = 0.5$ pour la population 2, ce qui traduit une moindre dispersion de la variable y dans la population 2. La base de sondage U_A correspond à l'ensemble des unités secondaires, et la partie correspondante de u_{hi} est $u_{hi}^A = u_{hi}$, de taille $N_{hi}^A = N_{hi}$. On génère pour chaque unité secondaire k une valeur u_k selon une loi uniforme sur $[0,1]$. La base de sondage U_B correspond aux unités secondaires k telles que $u_k \leq 0.5$, et la partie correspondante de u_{hi} est $u_{hi}^B = u_{hi} \cap U_B$ de taille N_{hi}^B. On se trouve donc dans la situation où $ab = U_B$ et $b = \emptyset$.

Le cadre retenu dans les simulations correspond à celui des enquêtes auprès des ménages de l'INSEE, avec extension pour cibler une sous-population spécifique. Pour ces enquêtes, un échantillon S_I de communes (ou de regroupements de communes) est tout d'abord sélectionné au premier degré. Un sous-échantillon S_i^A de logements est ensuite sélectionné dans chaque $u_i \in S_I$; l'échantillon réunion $S^A = \bigcup_{u_i \in S_I} S_i^A$ représente la population entière de logements $U_A = U$. Un second sous-échantillon S_i^B de logements est ensuite sélectionné au sein d'une sous-population de chaque $u_i \in S_I$, afin de cibler une sous-population spécifique U_B (par exemple, logements situés dans une Zone Urbaine Sensible) ; l'échantillon réunion $S^B = \bigcup_{u_i \in S_I} S_i^B$ ne représente que la sous-population ciblée U_B.

	Tailles d'échantillon par strate			Paramètres							
				Strate 1		Strate 2		Strate 3		Strate 4	
	m_h	n_{hi}^A	n_{hi}^B	μ_h	σ_h	μ_h	σ_h	μ_h	σ_h	μ_h	σ_h
Population 1	5 ou 25	10 ou 40	5 ou 20	200	20	150	15	120	12	100	10
Population 2	5 ou 25	10 ou 40	5 ou 20	200	10	150	7.5	120	6	100	5

TABLE 1 – Paramètres utilisés dans chaque strate pour générer les deux populations et sélectionner les échantillons

Dans chacune des deux populations ainsi constituées, on pratique plusieurs échantillonnages concurrents ; la Table 1 présente pour chaque population les huit combinaisons possibles de tailles d'échantillon par strate aux premier et second degré, ainsi que les valeurs μ_h et σ_h. Au premier degré on sélectionne indépendamment dans chaque strate U_{Ih} : soit un échantillon S_{Ih} de $m_h = 5$ unités primaires par sondage aléatoire simple ; soit un échantillon S_{Ih} de $m_h = 25$ unités primaires par sondage aléatoire simple. Au second degré, on sélectionne dans chaque $u_{hi} \in S_{Ih}$: soit un échantillon S_{hi}^A de taille $n_{hi}^A = 10$ par sondage aléatoire simple dans u_{hi}^A ; soit un échantillon S_{hi}^A de taille $n_{hi}^A = 40$ par sondage aléatoire simple dans u_{hi}^A . Au second degré, on sélectionne également dans chaque $u_{hi} \in S_{Ih}$: soit un échantillon S_{hi}^B de taille $n_{hi}^B = 5$ par sondage aléatoire simple dans u_{hi}^B ; soit un échantillon S_{hi}^B de taille $n_{hi}^B = 20$ par sondage aléatoire simple dans u_{hi}^B. On note également $f_{hi}^A = (N_{hi}^A)^{-1} n_{hi}^A$ et $f_{hi}^B = (N_{hi}^B)^{-1} n_{hi}^B$ les taux de sondage dans u_{hi}^A et u_{hi}^B.

Pour chaque échantillon, on calcule l'estimateur de Hartley donné en (9) avec soit $\theta = 1/2$ (HART1), soit pour valeur de θ l'estimateur du coefficient optimal donné en (12) (HART2), avec

$$\begin{aligned}
\hat{V}(\hat{Y}^A_{ab}) &= \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{Ih}} (N^A_{hi})^2 \frac{1-f^A_{hi}}{n^A_{hi}\,(n^A_{hi}-1)} \sum_{k\in S^A_{hi}} \left\{ y_k 1(k\in ab) - \bar{y}_{ab;S^A_{hi}} \right\}^2, \\
\hat{V}(\hat{Y}^B_{ab}) &= \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{Ih}} (N^B_{hi})^2 \frac{1-f^B_{hi}}{n^B_{hi}\,(n^B_{hi}-1)} \sum_{k\in S^B_{hi}} \left\{ y_k 1(k\in ab) - \bar{y}_{ab;S^B_{hi}} \right\}^2, \\
\widehat{Cov}(\hat{Y}^A_{a},\hat{Y}^A_{ab}) &= \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{Ih}} (N^A_{hi})^2 \frac{1-f^A_{hi}}{n^A_{hi}\,(n^A_{hi}-1)} \sum_{k\in S^A_{hi}} \left\{ y_k 1(k\in a) - \bar{y}_{a;S^A_{hi}} \right\} \left\{ y_k 1(k\in ab) - \bar{y}_{ab;S^A_{hi}} \right\},
\end{aligned}$$

en notant $\bar{y}_{d;V}$ la moyenne de la variable $y_k 1(k \in d)$ sur une partie $V \subset U$. Pour chaque échantillon, on calcule également l'estimateur de Kalton et Anderson (KALT) donné en (13), l'estimateur de Bankier (BANK) donné en (14), et l'estimateur de Horvitz-Thompson $\hat{Y}^A$ basé sur le seul échantillon S^A (HTA).

La procédure d'échantillonnage est répétée $10,000$ fois. Pour mesurer le biais d'un estimateur $\hat{Y}$, nous calculons son biais relatif de Monte Carlo

$$BR_{MC}(\hat{Y}) = \frac{E_{MC}(\hat{Y}) - Y}{Y} \times 100$$

avec $E_{MC}(\hat{Y}) = \dfrac{1}{10,000}\displaystyle\sum_{b=1}^{10,000} \hat{Y}_{(b)}$, et $\hat{Y}_{(b)}$ la valeur de l'estimateur $\hat{Y}$ pour l'échantillon b. Pour mesurer la variabilité de $\hat{Y}$, nous calculons son erreur quadratique moyenne de Monte Carlo

$$EQM_{MC}(\hat{Y}) = \frac{1}{10,000}\sum_{b=1}^{10,000} (\hat{Y}_{(b)} - Y)^2.$$

Les résultats sont donnés dans le tableau 2. Les performances de l'estimateur HTA ne dépendent pas de la taille d'échantillon n^B_{hi} choisie. Par souci de cohérence, nous indiquons donc dans le Tableau 2 les résultats obtenus dans les simulations avec $n^B_{hi} = 5$ uniquement. A tailles d'échantillon m_h et n^A_{hi} identiques, les mêmes résultats sont reportés dans le cas $n^B_{hi} = 20$.

Tous les estimateurs sont virtuellement sans biais. L'estimateur HART2 donne les meilleurs résultats en termes d'erreur quadratique moyenne, comme on pouvait s'y attendre. L'estimateur HTA donne des résultats quasiment équivalents. Ce résultat s'explique par le fait que le coefficient optimal est proche de 1 (dans les simulations, $\hat{\theta}_{opt}$ est compris entre 0.80 et 1.06), et que dans ce cas la formule (1) montre que les estimateurs HART2 et HTA sont très proches : nous présentons en Annexe des conditions générales sous lesquelles cette propriété est approximativement vérifiée. Parmi les trois autres estimateurs, HART1 donne les meilleurs résultats, avec une erreur quadratique moyenne plus faible ou équivalente à celle de KALT et BANK dans 11 cas sur 16.

Pour chaque estimateur, toutes choses égales par ailleurs, l'erreur quadratique moyenne est plus faible dans la population 2 que dans la population 1. Ce résultat provient du fait

				HART1		HART2		KALT		BANK		HTA	
Pop.	m_h	n_{hi}^A	n_{hi}^B	BR (%)	EQM $\times 10^9$	BR (%)	EQM $\times 10^9$	BR (%)	EQM $\times 10^9$	BR (%)	EQM $\times 10^9$	BR (%)	EQM $\times 10^9$
1	5	10	5	0.05	7.76	0.01	5.70	0.05	7.79	0.06	8.56	0.04	5.75
1	5	10	20	0.01	7.57	−0.05	5.57	0.03	11.36	0.04	12.75	0.04	5.75
1	5	40	5	0.01	5.01	−0.02	4.51	−0.02	4.57	−0.02	4.81	−0.02	4.52
1	5	40	20	0.00	4.65	−0.01	4.33	0.00	4.66	0.00	5.22	−0.02	4.52
1	25	10	5	−0.03	1.19	−0.02	0.78	−0.03	1.20	−0.02	1.34	−0.01	0.78
1	25	10	20	−0.01	1.17	0.00	0.78	−0.03	1.94	−0.03	2.22	−0.01	0.78
1	25	40	5	0.00	0.62	0.01	0.51	0.00	0.52	0.00	0.57	0.01	0.51
1	25	40	20	0.02	0.58	0.01	0.51	0.02	0.58	0.02	0.68	0.01	0.51
2	5	10	5	0.00	3.59	0.01	1.15	0.00	3.56	0.02	4.38	0.01	1.15
2	5	10	20	0.00	3.60	−0.02	1.15	0.00	7.38	0.00	8.76	0.01	1.15
2	5	40	5	0.00	1.48	0.01	1.07	0.00	1.13	0.01	1.35	0.01	1.07
2	5	40	20	0.00	1.49	−0.01	1.09	0.00	1.49	0.00	2.03	0.01	1.07
2	25	10	5	0.00	0.63	0.00	0.14	0.00	0.63	0.00	0.78	0.00	0.14
2	25	10	20	0.00	0.62	0.00	0.13	0.00	1.38	0.00	1.67	0.00	0.14
2	25	40	5	0.00	0.20	0.00	0.12	0.00	0.13	0.00	0.18	0.00	0.12
2	25	40	20	0.00	0.20	0.00	0.12	0.00	0.20	0.01	0.31	0.00	0.12

TABLE 2 – Biais relatif et Erreur Quadratique Moyenne de 5 estimateurs

que la variance due au premier degré de tirage, qui est la même pour chaque estimateur et vaut

$$V\left(\sum_{i\in S_I} d_{Ii}Y_i\right) \;=\; \sum_{h=1}^{H} M_h^2\left(\frac{1}{m_h}-\frac{1}{M_h}\right)S^2_{Y;U_{Ih}}, \tag{17}$$

est plus grande dans la population 1 : le terme de dispersion $S^2_{Y;U_{Ih}} = (M_h-1)^{-1}\sum_{u_i\in U_{Ih}}(Y_i-\bar{Y}_{U_{Ih}})^2$ augmente avec σ_h^2 et, dans une moindre mesure, augmente quand ρ diminue. L'erreur quadratique moyenne diminue pour chaque estimateur quand le nombre m_h d'unités primaires tirées dans chaque strate augmente, car dans ce cas le terme de variance commun donné en (17) diminue. De façon analogue, l'erreur quadratique moyenne diminue pour chaque estimateur quand n^A augmente, car dans ce cas la variance due au second degré de tirage diminue. Pour les estimateurs HART1 et HART2, l'erreur quadratique moyenne est stable quand n^B augmente, et de façon plus surprenante pour les estimateurs KALT et BANK l'erreur quadratique moyenne augmente quand n^B augmente. Ce résultat quelque peu contre-intuitif est dû à la conjonction de deux faits. D'une part, la contribution de l'échantillon S^B à la variance due au second degré de tirage est faible : l'augmentation de n^B peut diminuer cette variance, mais même dans ce cas la réduction globale de variance est marginale. D'autre part, dans le cas des estimateurs KALT et BANK, la contribution de l'échantillon S^A à la variance due au second degré de tirage augmente quand n^B augmente.

Dans le cas de KALT, l'estimateur peut se réécrire

$$\hat{Y}_{KA} \;=\; \sum_{h=1}^{H}\frac{M_h}{m_h}\sum_{i\in S_{Ih}}\hat{Y}_{KA,i}$$

avec

$$\hat{Y}_{KA,i} = \frac{1}{f^A_{hi}} \sum_{k \in S^A_i} m^A_{k|i} y_k + \frac{1}{f^A_{hi} + f^B_{hi}} \sum_{k \in S^B_i} y_k \quad \text{et} \quad m^A_{k|i} = \begin{cases} 1 & \text{si } k \in a \cap u_i, \\ \frac{f^A_{hi}}{f^A_{hi}+f^B_{hi}} & \text{si } k \in ab \cap u_i. \end{cases} \tag{18}$$

Dans (18), la dispersion de la variable $m^A_{k|i}$ (donc celle de $m^A_{k|i}\ y_k$) augmente quand le facteur $\dfrac{f^A_{hi}}{f^A_{hi} + f^B_{hi}}$ s'éloigne de 1. Or, ce facteur est proche de 1 quand f^B_{hi} est faible devant f^A_{hi} (donc n^B petit devant n^A), mais s'en éloigne quand n^B augmente. Notons que la variance (conditionnelle à S_I) du second terme de $\hat{Y}_{KA,i}$ est égale à

$$V\left(\frac{1}{f^A_{hi} + f^B_{hi}} \sum_{k \in S^B_i} y_k \middle| S_I \right) \quad = \quad (N^A_{hi})^2 N^B_{hi} \times \frac{n^B_{hi}(N^B_{hi} - n^B_{hi})}{(N^B_{hi} n^A_{hi} + N^A_{hi} n^B_{hi})^2} \times S^2_{u^B_{hi}}$$

avec $S^2_{u^B_{hi}} = (N^B_{hi} - 1)^{-1} \sum_{k \in u^B_{hi}} (y_k - \bar{y}_{u^B_{hi}})^2$. Cette variance ne décroît pas forcément quand n^B_{hi} augmente. Par exemple, l'un des cas considéré dans les simulations correspond à $N^A_{hi} = 100$, $N^B_{hi} \simeq 50$ et $n^A_{hi} = 40$. Dans ce cas, le terme $\dfrac{n^B_{hi}(N^B_{hi} - n^B_{hi})}{(N^B_{hi} n^A_{hi} + N^A_{hi} n^B_{hi})^2}$ atteint sa valeur maximale pour $n^B_{hi} = 11$.

Dans le cas de BANK, l'estimateur peut se réécrire

$$\hat{Y}_{HT} \quad = \quad \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{i \in S_{Ih}} \hat{Y}_{HT,i}$$

avec

$$\hat{Y}_{HT,i} = \sum_{k \in S^A_i \cup S^B_i} \frac{y_k}{\pi^{HT}_{k|i}} \quad \text{et} \quad \pi^{HT}_{k|i} = \begin{cases} f^A_{hi} & \text{si } k \in a, \\ f^A_{hi} + f^B_{hi}(1 - f^A_{hi}) & \text{si } k \in ab. \end{cases} \tag{19}$$

Dans (19), la dispersion de la variable $\pi^{HT}_{k|i}$ augmente quand le facteur $f^B_{hi}(1 - f^A_{hi})$ augmente. Or, ce facteur est proche de 0 quand n^B_{hi} (et donc f^B_{hi}) est faible, mais s'accroît quand n^B_{hi} augmente.

4 Conclusion

Nous avons étudié les estimateurs de Hartley (1962), de Kalton et Anderson (1986) et de Bankier (1986) pour mettre en commun les échantillons issus de deux vagues d'enquête. Nous avons plus particulièrement étudié le cas où un échantillon représente la population entière (échantillon complètement représentatif), alors que le second n'en représente qu'une partie (échantillon partiellement représentatif). Dans le cadre considéré dans les simulations (voir également l'Annexe pour un cadre plus général), l'utilisation de l'échantillon partiellement représentatif ne permet pas de gagner en précision : si sa taille augmente, la précision des estimateurs de la classe de Hartley reste stable ou s'améliore légèrement, alors que la précision des estimateurs de Kalton et Anderson et de Bankier se dégrade. L'estimateur optimal de Hartley lui-même, bien que plus complexe à calculer, offre une précision qui n'est que légèrement améliorée par rapport à l'estimateur

de Horvitz-Thompson classique calculé sur l'échantillon complètement représentatif. Bien que notre étude par simulations soit limitée, ces résultats suggèrent d'être prudents dans le choix d'un estimateur en présence de bases de sondage multiples, et qu'un estimateur simple est parfois préférable, même s'il n'utilise qu'une partie de l'information collectée.

Remerciements

Les auteurs remercient un éditeur associé de Techniques d'Enquête et un arbitre pour leur lecture attentive et leurs remarques qui ont permis d'améliorer significativement l'article, et David Haziza pour des discussions utiles.

Références

[1] Bankier, M.D. (1986). Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys. Journal of the American Statistical Association, 81, p.1074-1079.

[2] Bourdalle, G., Christine, M., et Wilms, L. (2000). Échantillons maître et emploi. Série INSEE Méthodes, 21, p. 139-173.

[3] Chauvet G., Tandeau de Marsac, G. (2014). Méthodes d'estimation sur bases de sondage multiples dans le cadre de plans de sondage à deux degrés. Techniques d'enquête, 40, n° 2, p. 367-378, Statistique Canada, n° 12-001-X.

[4] Hansen, M.H., et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics, 14, p. 333-362.

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

[6] 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, p. 663-685.

[7] Kalton, G., et Anderson, D.W. (1986). Sampling Rare Populations. Journal of the Royal Statistical Society, A, 149, p. 65-82.

[8] Lavallée, P. (2002). Le Sondage Indirect, ou la Méthode généralisée du partage des poids. Éditions de l'Université de Bruxelles (Belgique) et Éditions Ellipses (France).

[9] Lavallée, P. (2007). Indirect Sampling. New York : Springer.

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

[11] Lohr, S.L. (2009). Multiple frame surveys. In Handbook of Statistics, Sample Surveys : Design, Methods and Applications, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam :

North Holland, Vol. 29A, p. 71-88.

[12] Lohr, S.L. (2011). Alternative survey sample designs : Sampling with multiple overlapping frames. Survey Methodology, 37, p. 197-213.

[13] Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. Survey Methodology, 33, p. 151-157.

[14] Narain, R.D. (1951). On sampling without replacement with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 3, p. 169-175.

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

[16] Saigo, H (2010). Comparing four bootstrap methods for stratified three-Stage sampling. Journal of Official Statistics, Vol. 26, No. 1, 2010, p. 193-207.

Annexe : Comparaison entre l'estimateur optimal de Hartley et l'estimateur de Horvitz-Thompson

Nous reprenons le cadre et les notations de la Section 3 : les échantillons S^A et S^B sont sélectionnés selon un plan à deux degrés avec un premier degré de tirage commun. On utilise un sondage aléatoire simple stratifié au premier degré, et un sondage aléatoire simple au second degré dans chaque unité primaire. La base de sondage U_A correspond à la population entière, alors que la base de sondage U_B ne recouvre qu'une partie de la population.

Dans le cas de l'estimateur optimal de Hartley, la formule (11) donne

$$\theta_{opt|S_I} = \frac{EV(\hat{Y}^B_{ab}|S_I) - ECov(\hat{Y}^A_a, \hat{Y}^A_{ab}|S_I)}{EV(\hat{Y}^B_{ab}|S_I) + EV(\hat{Y}^A_{ab}|S_I)}.$$

Après un peu de calcul, nous obtenons

$$EV(\hat{Y}^A_{ab}|S_I) = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{u_{hi}\in U_{Ih}} (N_{hi})^2 \frac{1-f^A_{hi}}{n^A_{hi}} \left\{ \frac{N^B_{hi}-1}{N_{hi}-1} S^2_{u^B_{hi}} + \frac{N^B_{hi}(N_{hi}-N^B_{hi})(\bar{y}_{u^B_{hi}})^2}{N_{hi}(N_{hi}-1)} \right\}, \quad (20)$$

$$-ECov(\hat{Y}^A_a, \hat{Y}^A_{ab}|S_I) = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{u_{hi}\in U_{Ih}} (N_{hi})^2 \frac{1-f^A_{hi}}{n^A_{hi}} \left\{ \frac{N^B_{hi}(\bar{y}_{u^B_{hi}})(N_{hi}\bar{y}_{u_{hi}} - N^B_{hi}\bar{y}_{u^B_{hi}})}{N_{hi}(N_{hi}-1)} \right\}$$

avec $\bar{y}_{u_{hi}} = (N_{hi})^{-1} \sum_{k\in u_{hi}} y_k$, $\bar{y}_{u^B_{hi}} = (N^B_{hi})^{-1} \sum_{k\in u^B_{hi}} y_k$ et $S^2_{u^B_{hi}} = (N^B_{hi}-1)^{-1} \sum_{k\in u^B_{hi}} (y_k - \bar{y}_{u^B_{hi}})^2$.

L'estimateur de Horvitz-Thompson basé sur le seul échantillon S^A et l'estimateur optimal de Hartley coïncident si le coefficient $\theta_{opt|S_I}$ est égal à 1, ce qui est le cas si $EV(\hat{Y}^A_{ab}|S_I) = -ECov(\hat{Y}^A_a, \hat{Y}^A_{ab}|S_I)$. Cette condition sera en particulier vérifiée si dans (20) les termes entre accolades coïncident pour chaque unité primaire u_{hi}. On aura donc $\theta_{opt|S_I} \simeq 1$ si

$$\forall u_{hi} \in U_I \quad \frac{N_{hi}(N^B_{hi}-1)}{N^B_{hi}} \frac{S^2_{u^B_{hi}}}{\bar{y}_{u^B_{hi}}(N_{hi}\bar{y}_{u_{hi}} - N^B_{hi}\bar{y}_{u^B_{hi}})} + \frac{(N_{hi}-N^B_{hi})\bar{y}_{u^B_{hi}}}{N_{hi}\bar{y}_{u_{hi}} - N^B_{hi}\bar{y}_{u^B_{hi}}} \simeq 1. \quad (21)$$

Supposons que la valeur moyenne de y soit approximativement la même dans les bases U_A et U_B pour chaque unité primaire, i.e. que $\forall u_{hi} \in U_I$ $\bar{y}_{u^B_{hi}} \simeq \bar{y}_{u_{hi}}$. Alors la condition (21) sera approximativement vérifiée si $\forall u_{hi} \in U_I$ $cv^2_{u^B_{hi}}$ est proche de 0, avec $cv_{u^B_{hi}} = \sqrt{S^2_{u^B_{hi}}}/\bar{y}_{u^B_{hi}}$.

En résumé, l'estimateur de Horvitz-Thompson basé sur le seul échantillon S^A et l'estimateur optimal de Hartley seront proches si au sein de chaque unité primaire u_{hi} : (a) la valeur moyenne de y est peu différente entre les deux bases, et (b) la variable y est faiblement dispersée au sein de u^B_{hi}. Dans les simulations, la condition (a) est approximativement respectée car la répartition des individus entre les bases de sondage U_A et U_B se fait complètement aléatoirement ; la condition (b) est approximativement respectée avec des valeurs de $cv^2_{u^B_{hi}}$ variant de 0.02 à 0.10 pour la population 1, et de 0.001 à 0.005 pour la population 2.