Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 119-131
Statistique Canada

Comparaison d'estimateurs de totaux de population obtenus par sondages successifs à deux degrés à l'aide de l'information auxiliaire

F.C. OKAFOR[1]

RÉSUMÉ

Singh et Srivastava (1973) ont élaboré un estimateur linéaire non biaisé de moyennes de population qui pourrait être utilisé dans des sondages successifs à l'aide de plusieurs variables auxiliaires dont les moyennes de population connues ne changent pas d'une période à l'autre. Dans le présent document, trois estimateurs composites T_1, T_2 et T_3, utilisant chacun une variable auxiliaire dont la moyenne de population connue change d'une période à l'autre, sont présentés pour l'estimation du total de population de la période donnée. Les estimateurs proposés sont comparés à l'estimateur habituel, T_0, et à l'estimateur habituel de sondages successifs, T', du total de population de la période donnée sans l'aide de l'information auxiliaire. Nous observons que l'utilisation conjuguée de l'information auxiliaire et d'une méthode par sondages successifs ne produit pas toujours uniformément un gain d'efficacité par rapport à T_0 ou T'. Toutefois, quand ils ont été appliqués à une enquête visant à estimer la taille moyenne des arbres dans des plantations de teck, les estimateurs T_1, T_2 et T_3 se sont avérés plus efficaces que T_0 ou que T'.

MOTS CLÉS: Périodes successives; appariement partiel; variable auxiliaire.

1. INTRODUCTION

La théorie et la pratique du sondage d'une même population à des moments différents - qu'on appelle sondages échelonnés ou sondages successifs - ont été beaucoup étudiées par certains statisticiens d'enquête. Les principaux objectifs des sondages successifs sont d'estimer des paramètres de population (par exemple des totaux, des moyennes, des ratios de population, etc.) pendant la période la plus récente et d'estimer les variations dans ces paramètres d'une période à l'autre.

La théorie des sondages successifs a été proposée pour la première fois par Jessen (1942). Beaucoup d'autres auteurs ont depuis étudié la question, en particulier en ce qui concerne l'estimation de moyennes de population; notamment Singh (1968), Abraham et coll. (1969), Kathuria et Singh (1971) et Kathuria (1976), pour n'en nommer que quelques-uns.

Singh (1968) a été le premier à étendre la théorie des sondages à un seul degré aux sondages échelonnés à deux degrés. Il a utilisé un plan de sondage dans lequel, à la deuxième période, une fraction λ des unités de sondage du premier degré (USPD) choisies à la période précédente est retenue, en plus des unités de sondage correspondants du second degré (USSD) et d'une fraction μ ($\lambda + \mu = 1$) choisie de nouveau. Il a ensuite obtenu un estimateur non biaisé à variance minimum de la moyenne de population de la période donnée.

Abraham et coll. (1969) ont considéré le cas où un appariement partiel des unités était effectué aux deux degrés. Les unités étaient choisies à l'aide d'une méthode de sondage aléatoire simple et sans remise (SASSR). Kathuria (1975) a modifié cette façon de procéder en utilisant une méthode de sélection des USPD avec probabilités proportionnelles à la taille et avec remise (PPTAR) et proposé un estimateur linéaire composite pour estimer la moyenne de population de la période donnée.

[1] F.C. Okafor, Département de statistique, Université d'Ibadan, Ibadan, Nigeria.

Quand une variable auxiliaire est très corrélée à la caractéristique étudiée, on peut améliorer l'estimation de la moyenne (ou du total) de population de cette caractéristique en utilisant la variable auxiliaire. Shivtar Singh et Srivastava (1973) ont utilisé de l'information auxiliaire pour améliorer l'estimateur de Singh (1968). Ils ont obtenu un estimateur linéaire non biaisé de la moyenne de population de la période la plus récente en utilisant des variables auxiliaires dont les moyennes de population étaient connues et ne changeaient pas d'une période à l'autre. Kathuria (1978) a poussé davantage dans cette voie en supposant que la moyenne de population de la variable auxiliaire n'est pas connue. Il a utilisé une technique d'échantillonnage double (ou sondage à deux phases) pour estimer d'abord la moyenne de population de la variable auxiliaire et ensuite la moyenne de population de la caractéristique à l'étude.

Dans leurs ouvrages, Shivtar Singh et coll. (1973) et Kathuria (1978) ont supposé qu'il était possible d'obtenir l'information nécessaire sur les variables auxiliaires des répondants ou des unités déclarantes (UD). Cela n'est généralement pas le cas. Il peut arriver que le caractère délicat de la question ou le refus pur et simple des répondants de fournir toute information faussent l'information sur la variable auxiliaire au point de la rendre inutile. Il se peut aussi que l'information sur la variable auxiliaire ne puisse être recueillie parce que la question qui aurait permis de l'obtenir n'est pas incluse dans le questionnaire.

Shivtar Singh et coll. ont également supposé que le total de population connu de la variable auxiliaire est le même à toutes les périodes. Il se peut que cela ne soit pas vrai en pratique. Si le total de population de la caractéristique principale varie d'une période à l'autre, il y a tout lieu de penser que le total de population de toute autre variable qui serait corrélée à la caractéristique principale variera également.

Dans le présent document, trois estimateurs composites de total de population utilisant de l'information auxiliaire et un plan de sondage comportant des sondages successifs à deux degrés sont proposés. Les performances des trois estimateurs sont comparées empiriquement entre elles; les trois estimateurs ont également été appliqués à une enquête visant à estimer la taille moyenne des arbres dans des plantations de teck.

2. SONDAGE POUR DEUX PÉRIODES

Pour les trois estimateurs proposés, nous supposons que le total de population de la variable auxiliaire change à la deuxième période.

Les estimateurs du total (ou de la moyenne) de population fondés sur un plan d'appariement partiel sont meilleurs que les estimateurs habituels du total (ou de la moyenne) de population sans appariement partiel. On peut donc s'attendre que les estimateurs proposés, T_1, T_2 et T_3, donneront de meilleurs résultats que l'estimateur habituel du total de population, T_0, et que l'estimateur fondé sur le même plan d'appariement partiel mais n'utilisant pas d'information auxiliaire, T'.

Dans le calcul de ces estimateurs, nous supposons que:

(i) la taille de l'échantillon est constante à chaque période,
(ii) la mesure normalisée P_i de la taille de la i-ième unité de sondage du premier degré (USPD) est fixée pour chaque période,
(iii) N et M_i, les tailles de population respectives des USPD et des unités de sondage du second degré (USSD) prélevées à partir de la i-ième USPD, sont constantes dans les deux périodes,
(iv) le total (ou la moyenne) de population de la variable auxiliaire est connu.

Les hypothèses (i) à (iii) s'appliquent à T', T_1, T_2 et T_3, tandis que l'hypothèse (iv) s'applique à T_1, T_2 et T_3, mais non à T' et T_0.

À la première période, un échantillon S_1 de n USPD est tiré avec probabilités proportionnelles à la taille et avec remise (PPTAR) à l'aide d'une mesure normalisée de la taille de la i-ième unité ($i = 1, 2, \ldots, N$). Pour choisir les USSD, nous adoptons la méthode de

Cochran (1977, p. 306), qui stipule que si la i-ième USPD de S_1 est choisie θ_i fois ($i = 1, 2, \ldots, n$), on tire θ_i sous-échantillons indépendants de taille m_i à partir des M_i USSD.

À la deuxième période, nous prélevons un échantillon de λn USPD ($0 < \lambda > 1$) à partir de S_1 selon un plan de sondage aléatoire simple et sans remise (SASSR). Les USSD choisies à la première période sont retenues pour chacune de ces λn USPD appariées. Ensuite, un nouvel échantillon de μn ($\mu = 1 - \lambda$) USPD est tiré indépendamment des n USPD par sondage avec PPTSR, avec P_i comme mesure normalisée de la taille de la i-ième USPD. Dans chacune des μn USPD, les USSD sont choisies de la même façon qu'à la première période.

3. NOTATION

Nous définissons y_{ij} (x_{ij}) comme la valeur de la variable à l'étude pour la j-ième USSD dans la i-ième pendant la période donnée (ou la période précédente). De plus, z_{hij} est définie comme la valeur de la variable auxiliaire pour la j-ième USSD dans la i-ième USPD à la h-ième période ($h = 1, 2$). Les moyennes de l'échantillon des USSD dans la i-ième USPD sont

$$\bar{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}, \quad \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \quad \text{et} \quad \bar{z}_{hi} = \frac{1}{m_i} \sum_{j=1}^{m_i} z_{hij}.$$

Le total de population pour la i-ième USPD et le total de population pour l'ensemble des USPD qui correspondent à la variable auxiliaire sont

$$Z_{hi} = \sum_{j=1}^{M_i} z_{hij} \quad \text{et} \quad Z_h = \sum_{j=1}^{N} Z_{hi}.$$

Nous définissons aussi quelques autres notations de la façon suivante:

$$S_b^2(y) = \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y\right)^2 \text{ est la variance entre les USPD;}$$

$$S_w^2(y) = \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{wi}^2(y) \text{ est la variance entre les USSD de l'ensemble des USPD;}$$

$$S_{wi}^2(y) = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_i)^2 \text{ est la variance entre les USSD de la } i\text{-ième USPD;}$$

$$S^2(y) = S_b^2(y) + S_w^2(y);$$

$$C_b(x,y) = \rho_b S_b(x) S_b(y) \text{ est la covariance de } x \text{ et } y \text{ entre les USPD;}$$

$$C_w(x,y) = \rho_w S_w(x) S_w(y) \text{ est la covariance de } x \text{ et } y \text{ entre les USSD de l'ensemble des USPD;}$$

$$C(x,y) = C_b(x,y) + C_w(x,y).$$

Les coefficients de corrélation entre x et y calculés entre les USPD et à l'intérieur des USPD sont respectivement ρ_b et ρ_w.

4. ESTIMATEURS DE TOTAUX DE POPULATION ET VARIANCES OPTIMUMS

4.1 Cas (i)

Le premier estimateur du total de population Y, de la seconde période est utilisé lorsqu'on n'a pas d'information sur la variable auxiliaire, mais qu'on connaît le total de population de la variable auxiliaire pour les USPD choisies. Il s'exprime ainsi:

$$T_1 = \theta(1)\, T_m(1) + (1 - \theta(1))\, T_u(1) \tag{4.1}$$

où $\theta(1)$ est une constante choisie e telle sorte que la variance de T_1, $V(T_1)$ est minimum, tandis que

$$T_m(1) = \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\} - b(1) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} - \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} \right]$$

est l'estimateur par différence de Y fondé sur l'échantillon apparié, que

$$T_u(1) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\}$$

est l'estimateur de Y fondé sur l'échantillon non apparié et que $k(1)$ et $b(1)$ sont des constantes connues.

Pour cet estimateur, on suppose qu'on connaît le total de population de la variable auxiliaire, Z_i, pour chaque USPD choisie à la première période. On connaît également le total de population pour l'ensemble de USPD, Z, à chaque période. Aucune autre information sur la variable auxiliaire n'est obtenue des répondants ou des unités déclarantes (UD).

Maintenant, en minimisant $V(T_1)$ par rapport à $\theta(1)$ et en résolvant l'équation, on obtient la valeur optimum suivante de $\theta(1)$

$$\theta_0(1) = \lambda A_2(1) / \Delta(1)$$

où

$$A_2(1) = S^2(y) + k^2(1)\, S_b^2(z_2) - 2k(1)\, C_b(z_2, y),$$

$$\Delta(1) = A_2(1) + \mu^2 \{ b^2(1)\, A_1(1) - 2b(1)\beta(1) \}.$$

La valeur optimum de $k(1)$ est obtenue en minimisant $V(T_u(1))$ par rapport à $k(1)$. Cela donne $k_0(1) = C_b(z_2, y) / S_b^2(z_2)$.

On peut montrer qu'en utilisant la méthode proposée par Jessen (1942), la valeur optimum de $V(T_1)$ pour une valeur donnée de λ est

$$V_0(T_1) = \frac{1}{n}\,[A_2(1) + \mu\{b^2(1)A_1(1) - 2b(1)\,\beta(1)\}]A_2(1)/\Delta(1) \tag{4.2}$$

où

$$A_1(1) = S^2(x) + k^2(1)\,S_b^2(z_1) - 2k(1)\,C_b(z_1,x),$$

$$\beta(1) = C(x,y) + k^2(1)C_b(z_1,z_2) - k(1)\{C_b(x,z_2) + C_b(z_1,y)\},$$

$$\Delta(1) = A_2(1) + \mu^2\{b^2(1)\,A_1(1) - 2b(1)\,\beta(1)\}.$$

Lorsqu'on minimise la variance de $T_m(1)$, la valeur optimum de $b(1)$ est

$$b_0(1) = \beta(1)/A_1(1).$$

Si on substitue $b_0(1)$ dans (4.2), la variance optimum devient

$$V_0(T_1) = \frac{1}{n}\left[\frac{A_1(1)\,A_2(1) - \mu\beta^2(1)}{A_1(1)\,A_2(1) - \mu^2\beta^2(1)}\right]A_2(1). \tag{4.3}$$

Lorsqu'on minimise $V_o(T_1)$ dans (4.2) par rapport à μ, la fraction d'appariement optimum se ramène à $\lambda_0 = 1 - \mu_0$ où

$$\mu_0 = A_2(1)\,[A_2(1) + \{A_2^2(1) + A_2(1)\,(b^2(1)A_1(1) - 2b(1)\beta(1))\}^{1/2}]^{-1}. \tag{4.4}$$

Si $A_2(1) = A_1(1)$, c'est-à-dire si la variation de la population est la même dans les deux périodes, l'expression (4.3) donne

$$V_0(T_1) = \frac{1}{n}\left[\frac{A^2(1) - \mu\beta^2(1)}{A^2(1) - \mu^2\beta^2(1)}\right]A(1) \tag{4.5}$$

et si on substitue $b_0(1)$ à $b(1)$, la fraction d'appariement optimum donnée dans l'équation (4.4), μ_0, devient

$$\mu_0 = A(1)\,[A(1) + \{A^2(1) - \beta^2(1)\}^{1/2}]^{-1}. \tag{4.6}$$

Quand on substitue μ_0 dans (4.5), la variance se ramène à

$$V_0(T_1) = \frac{1}{2n}\,[A(1) + \{A^2(1) - \beta^2(1)\}^{1/2}]. \tag{4.7}$$

4.2 Cas (ii)

Le deuxième estimateur est l'estimateur habituel dans lequel l'information relative aussi bien à la caractéristique principale qu'à la caractéristique auxiliaire a été obtenue des unités déclarantes et dans lequel le total de population de la caractéristique auxiliaire est connu. Il s'exprime ainsi:

$$T_2 = \theta(2)\, T_m(2) + (1 - \theta(2))\, T_u(2), \tag{4.8}$$

où

$$T_m(2) = \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{2i}}{P_i} - Z_2 \right) \right.$$
$$- b(2) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{1i}}{P_i} - Z_1 \right) \right\} \right.$$
$$\left. \left. - \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{1i}}{P_i} - Z_1 \right) \right\} \right],$$

et

$$T_u(2) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{2i}}{P_i} - Z_2 \right) \right\}.$$

Ici, le total de population global de la variable auxiliaire est connu dans les deux périodes. En outre, l'information sur la variable auxiliaire, z_{ij}, est obtenue pour chaque USSD de l'échantillon. Il s'agit de la façon habituelle d'utiliser l'information auxiliaire dans les plans de sondage décrits dans les ouvrages portant sur le sujet. On peut montrer que la variance optimum de T_2 est

$$V_0(T_2) = \frac{1}{n} [A_2(2) + \mu \{ b^2(2) A_1(2) - 2b(2) \beta(2) \}] A_2(2) / \Delta(2) \tag{4.9}$$

et que le poids optimum est

$$\theta_0(2) = \lambda A_2(2) / \Delta(2)$$

où

$$A_2(2) = S^2(y) + k^2(2)\, S^2(z_2) - 2k(2)\, C(z_2, y),$$

$$A_1(2) = S^2(x) + k^2(2)\, S^2(z_1) - 2k(2)\, C(z_1, x),$$

$$\beta(2) = C(x,y) + k^2(2)\, C(z_1, z_2) - k(2) \{ C(z_1, y) + C(x, z_2) \},$$

$$\Delta(2) = A_2(2) + \mu^2 \{ b^2(2)\, A_1(2) - 2b(2)\, \beta(2) \}.$$

La valeur optimum de $k(2)$ est $k_0(2) = C(z_2,y)/S_2(z_2)$.

En substituant le coefficient optimum de régression, $b_0(2) = \beta(2)/A_1(2)$, obtenue en minimisant la variance de $T_m(2)$, dans (4.9) et en supposant que $A_2(2) = A_1(2) = A(2)$, on obtient

$$V_0(T_2) = \frac{1}{n}\left[\frac{A^2(2) - \mu\beta^2(2)}{A^2(2) - \mu^2\beta^2(2)}\right] A(2). \tag{4.10}$$

Si la valeur optimum de μ est substituée dans (4.10), la variance devient

$$V_0(T_2) = \frac{1}{2n}\left[A(2) + \{A^2(2) - \beta^2(2)\}^{1/2}\right]. \tag{4.11}$$

4.3 Cas (iii)

La troisième façon d'utiliser l'information auxiliaire connue pour améliorer l'estimation du total de population de la période donnée, Y, dans le cadre d'un plan de sondage donné ressemble beaucoup à la deuxième façon. La seule différence est qu'on ne connaît pas le total de population de la caractéristique auxilière; par contre, on connaît la moyenne de population pour les USPD choisies.

L'estimateur s'exprime ainsi:

$$T_3 = \theta(3)\, T_m(3) + (1 - \theta(3))\, T_u(3), \tag{4.12}$$

où

$$T_m(3) = \frac{1}{\lambda n}\sum_{i=1}^{\lambda n}\frac{M_i}{P_i}\{\bar{y}_i - k(3)\,(\bar{z}_{2i} - \bar{Z}_{2i})\}$$

$$- b(3)\left[\frac{1}{\lambda n}\sum_{i=1}^{\lambda n}\frac{M_i}{P_i}\{\bar{x}_i - k(3)\,(\bar{z}_{1i} - \bar{Z}_{1i})\}\right.$$

$$\left. - \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{P_i}\{\bar{x}_i - k(3)\,(\bar{z}_{1i} - \bar{Z}_{1i})\}\right],$$

et

$$T_u(3) = \frac{1}{n\mu}\sum_{i=1}^{n\mu}\frac{M_i}{P_i}\{\bar{y}_i - k(3)\,(\bar{z}_{2i} - \bar{Z}_{2i})\}.$$

Pour cet estimateur, on suppose que les valeurs tant de la variable principale que de la variable auxiliaire sont obtenues pour chaque USSD de l'échantillon dans les deux périodes. On suppose également que la moyenne de population, $\bar{Z}_i$, de la variable auxiliaire est connue pour les USPD choisies.

La variance optimum de T_3 pour une valeur donnée de λ s'exprime ainsi:

$$V_0(T_3) = \frac{1}{n}\left[A_2(3) + \mu\{b^2(3)A_1(3) - 2b(3)\beta(3)\}\right] A_2(3)/\Delta(3) \tag{4.13}$$

tandis que le poids optimum est donné, comme d'habitude, par l'expression suivante:

$$\theta_0(3) = \lambda A_2(3)/\Delta(3),$$

où

$$A_2(3) = S^2(y) + k^2(3)\, S_w^2(z_2) - 2k(3)\, C_w(z_2,y),$$

$$A_1(3) = S^2(x) + k^2(3)\, S_w^2(z_1) - 2k(3)\, C_w(z_1,x),$$

$$\beta(3) = C(x,y) + k^2(3)\, C_w(z_1,z_2) - k(3)\, \{C_w(z_1,y) + C_w(z_2,x)\},$$

$$\Delta(3) = A_2(3) + \mu^2\, \{b^2(3) A_1(3) - 2b(3)\, \beta(3)\}.$$

La valeur optimum de $k(3)$ est $k_0(3) = C_w(z_2,y)/S_w^2(z_2)$.

Si on substitue le coefficient optimum de régression dans (4.13) et si on suppose que la variance de population est la même dans les deux périodes, (4.13) se ramène alors à

$$V_0(T_3) = \frac{1}{n}\left[\frac{A^2(3) - \mu\beta^2(3)}{A^2(3) - \mu^2\beta^2(3)}\right] A(3). \tag{4.14}$$

Quand la valeur optimum de μ est substituée dans (4.14), la variance devient

$$V_0(T_3) = \frac{1}{2n}\,[A(3) + \{A^2(3) - \beta^2(3)\}^{1/2}]. \tag{4.15}$$

4.4 Efficacité des estimateurs proposés

Nous utiliserons les variances données en (4.7), (4.11) et (4.15) pour comparer l'efficacité des trois estimateurs, T_1, T_2 et T_3 par rapport à

$$T_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i \bar{y}_i}{P_i}.$$

T_0 est l'estimateur de y quand il n'y a pas d'appariement partiel des unités et qu'aucune information auxiliaire n'est utilisée. Nous avons aussi comparé l'efficacité de T_0 par rapport à l'estimateur habituel avec appariement partiel, T', qui n'utilise pas d'information auxiliaire, pour mieux faire comprendre la performance des estimateurs proposés.

L'estimation habituelle avec appariement partiel est définie comme suit:

$$T' = \theta' T'_m + (1 - \theta')\, T'_u, \tag{4.16}$$

où

$$T'_m = \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i \bar{y}_i}{P_i} - b' \left\{ \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i \bar{x}_i}{P_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{M_i \bar{x}_i}{P_i} \right\},$$

et

$$T'_u = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \frac{M_i \bar{y}_i}{P_i}.$$

La variance optimum de T', obtenue en utilisant la valeur optimum de b', $b'_0 = C(x,y)/S^2(x)$, et en supposant que $S^2(y) = S^2(x)$ est

$$V_0(T') = \frac{1}{n} \left[\frac{S^2(y) - \mu C(x,y)}{S^2(y) - \mu^2 C(x,y)} \right] S^2(y). \tag{4.17}$$

Si on substitue la valeur optimum de μ dans (4.17), la variance de T' devient

$$V_0(T') = \frac{1}{2n} [S^2(y) + \{S^4(y) - C^2(x,y)\}^{1/2}]. \tag{4.18}$$

Pour calculer l'efficacité des divers estimateurs, les hypothèses suivantes ont été faites au sujet des coefficients de corrélation et de la constante k:

$$\rho_b(x,z_2) = \rho_b(z_1,y) = \rho_b(z_1,z_2) = \rho_b;$$

$$\rho_w(x,z_2) = \rho_w(z_1,y) = \rho_w(z_1,z_2) = \rho_w;$$

$$k(1) = k(2) = k(3) = 1.$$

Les valeurs de l'efficacité n'ont été présentées que pour des valeurs positives de ρ_b et ρ_w et une série de valeurs de

$$\delta = S_w^2(y)/S_b^2(y),\ R_b = S_b^2(z)/S_b^2(y) \text{ et } R_w = S_w^2(z)/S_b^2(y).$$

Si on regarde le tableau 2, on constate qu'aucune des stratégies T_1, T_2 ou T_3 (plan de sondage et estimateur) n'est uniformément plus efficace que la stratégie T_0. C'est le contraire pour T', qui est toujours plus efficace que T_0, et qui, au pire, n'offre qu'un faible gain par rapport à T_0 (voir tableau 1).

D'après les résultats des tableaux 1 et 2, il faut préférer T_1 à T' seulement lorsque $R_b = 0.05$, $\rho_b = 0.8$ et $R_b = 0.5$.

Tableau 1

Efficacité de T' par rapport à T_0

ρ_b	δ	$\rho_w = 0.2$	$\rho_w = 0.8$
0.2	0.05	1.01	1.01
	0.5	1.01	1.04
	5.0	1.01	1.17
0.8	0.05	1.22	1.25
	0.5	1.11	1.25
	5.0	1.02	1.25

T_2 est meilleur que T' quand

(i) $\rho_w = 0.2$, $R_b = R_w = 0.05$;

(ii) $\rho_b = \rho_w = 0.8$, $R_b = R_w = 0.05, 0.5$;

(iii) $\delta = 0.5, 5.0$, $R_w = R_b = 0.05, \rho 0.5$, $\rho_b = 0.2$ et $\rho_w = 0.8$.

T_3 est généralement plus efficace que T' quand

(i) $\delta = 5.0$, $\rho_w = 0.8$;

(ii) $\delta = 0.5$, $\rho_w = 0.8$ and $R_w = 0.05, 0.5$.

Le gain maximum d'efficacité de T' par rapport à T_0 est de 25% (voir tableau 1). D'après les chiffres du tableau 2, le gain maximum de T_1 par rapport à T_0 est de 155%; il est obtenu quand $\rho_b = \rho_w = 0.8, \delta = 0.05$ et $R_b = 0.5$. Le gain maximum d'efficacité de T_2 par rapport à T_0 est de 172%; il est obtenu quand $\rho_b = \rho_w = 0.8$ et $\delta = R_w = 0.05$. Nous constatons également que lorsque $\rho_b = \rho_w = 0.8$, et $\delta = R_w = 5.0$, le gain maximum de T_3 par rapport à T_0 est de 104%. Il est donc évident que l'utilisation d'une variable auxiliaire a beaucoup amélioré l'efficacité de l'appariement partiel des unités.

Si maintenant nous comparons entre elles les trois stratégies T_1, T_2 et T_3, nous pouvons conclure qu'aucune n'est uniformément meilleure qu'une autre, même si le gain maximum d'efficacité de T_2 est supérieur au gain maximum d'efficacité de T_1, lui-même plus élevé que le gain maximum de T_3 par rapport à T_0. En général, T_1 est meilleur que T_2 quand $\rho_w = 0.2$, tandis que T_2 est meilleur que T_1 quand $\rho_w = 0.8$. T_1 est préférable à T_3 quand $\rho_b = 0.8$, $\rho_w = 0.2$ et $R_b = 0.05, 0.5$ ou quand $\rho_b = \rho_w = 0.8$ et $\delta = R_b = 0.05$. Enfin, T_3 est meilleur que T_2 quand $\rho_w = 0.8$ et $R_b = 5.0$ ou quand $\rho_b = \rho_w = 0.2$ et $R_b = 0.5, 5.0$.

5. APPLICATION

Les estimateurs proposés ont été appliqués à une enquête sur la hauteur des arbres dans des plantations de teck. L'objectif était d'estimer la hauteur moyenne des arbres en utilisant la circonférence des troncs comme information auxiliaire.

Tableau 2

Efficacité de T_1, T_2, et T_3 par rapport à T_0

		$\rho_w = 0.2$								
		$R_b = 0.05$			$R_b = 0.5$			$R_b = 5.0$		
ρ_b	δ	0.05	R_w 0.5	5.0	0.05	R_w 0.5	5.0	0.05	R_w 0.5	Stratégie
0.2	0.05	1.04	1.04	1.04	0.83	0.83	0.83	0.20	0.20	T_1
		1.01	0.73	0.18	0.81	0.62	0.17	0.20	0.19	T_2
		0.98	0.71	0.18	0.98	0.71	0.18	0.98	0.71	T_3
	0.5	1.03	1.03	1.03	0.87	0.87	0.87	0.27	0.27	T_1
		1.04	0.85	0.26	0.88	0.74	0.25	0.27	0.25	T_2
		1.02	0.84	0.26	1.02	0.84	0.26	1.02	0.84	T_3
	5.0	1.02	1.02	1.02	0.97	0.97	0.97	0.60	0.60	T_1
		1.04	1.03	0.67	0.99	0.99	0.65	0.60	0.60	T_2
		1.03	1.03	0.67	1.03	1.03	0.67	1.03	1.03	T_3
0.8	0.05	1.62	1.62	1.62	2.53	2.53	2.53	0.45	0.45	T_1
		1.53	0.94	0.19	2.35	1.23	0.20	0.45	0.38	T_2
		1.16	0.77	0.18	1.16	0.77	0.18	1.16	0.77	T_3
	0.5	1.34	1.34	1.34	1.74	1.74	1.74	0.45	0.45	T_1
		1.34	1.03	0.27	1.76	1.28	0.29	0.54	0.48	T_2
		1.11	0.88	0.26	1.11	0.88	0.26	1.11	0.88	T_3
	5.0	1.07	1.07	1.07	1.13	1.13	1.13	0.83	0.83	T_1
		1.10	1.09	0.69	1.16	1.15	0.72	0.84	0.83	T_2
		1.05	1.03	0.67	1.05	1.03	0.67	1.05	1.03	T_3

		$\rho_w = 0.8$									
		$R_b = 0.05$				$R_b = 0.5$			$R_b = 5.0$			
ρ_b	δ	5.0	0.05	R_w 0.5	5.0	0.05	R_w 0.5	5.0	0.05	R_w 0.5	5.0	Stratégie
0.2	0.05	0.20	1.05	1.05	1.05	0.83	0.83	0.83	0.20	0.20	0.20	T_1
		0.11	1.07	0.85	0.23	0.85	0.70	0.21	0.19	0.19	0.12	T_2
		0.18	1.04	0.83	0.23	1.04	0.83	0.23	1.04	0.83	0.23	T_3
	0.5	0.27	1.06	1.06	0.89	0.89	0.89	0.89	0.27	0.27	0.27	T_1
		0.15	1.21	1.30	0.41	1.00	1.06	0.38	0.28	0.28	0.19	T_2
		0.26	1.18	1.26	0.41	1.18	1.26	0.41	1.18	1.26	0.41	T_3
	5.0	0.60	1.17	1.17	1.17	1.09	1.09	1.09	0.62	0.62	0.62	T_1
		0.46	1.31	1.64	2.03	1.22	1.51	1.87	0.67	0.76	0.84	T_2
		0.67	1.30	1.63	2.00	1.30	1.63	2.00	1.30	1.63	2.00	T_3
0.8	0.05	0.45	1.65	1.65	1.65	2.55	2.55	2.55	0.46	0.46	0.46	T_1
		0.15	1.70	1.22	0.25	2.72	1.64	0.27	0.46	0.42	0.18	T_2
		0.18	1.27	0.98	0.24	1.26	0.98	0.24	1.27	0.98	0.24	T_3
	0.5	0.45	1.50	1.50	1.50	1.88	1.88	1.88	0.56	0.56	0.56	T_1
		0.21	1.75	1.83	0.46	2.34	2.65	0.50	0.59	0.61	0.31	T_2
		0.26	1.40	1.43	0.43	1.40	1.43	0.43	1.40	1.43	0.43	T_3
	5.0	0.83	1.30	1.30	1.30	1.35	1.35	1.35	0.95	0.95	0.95	T_1
		0.85	1.46	1.85	2.25	1.53	1.98	2.53	1.03	1.22	1.38	T_2
		0.67	1.39	1.74	2.04	1.39	1.74	2.04	1.39	1.74	2.04	T_3

Tableau 3

Efficacité estimée des estimateurs proposés par rapport à T_0 dans l'estimation de la hauteur moyenne des arbres dans des plantations de teck

Estimateurs	Hauteur moyenne (m)	Variance (m^2)	Efficacité estimée en %
T_0 (sans appariement)	20.04	6.3118	100
T' (appariement partiel)	18.06	4.0680	155
T_1	17.86	0.0718	8791
T_2	17.31	0.0651	9635
T_3	17.99	4.0183	157

Les arbres utilisés dans l'enquête ont été plantés en 1965 suivant différents espacements, ce qui a produit des plantations ayant le nombre d'arbres suivant par hectare: 2,000, 800, 400 et 250. Pour mesurer la hauteur des arbres, un périmètre de 40 mètres sur 40 a été tracé dans chacune des 8 plantations (USPD) prélevées parmi 16 plantations à l'aide d'un plan de sondage avec PPTAR. Le nombre d'arbres dans chaque plantation a été utilisé comme mesure de la taille. Tous les arbres à l'intérieur du périmètre de 40 m sur 40 formaient les unités de sondage du second degré et leur circonférence à hauteur de poitrine a été mesurée. Pour le calcul de la hauteur, un sous-échantillon d'arbres a été sélectionné des arbres de périmètre 40m sur 40 dans chaque USPD choisie. Une première série de calculs a été faite en 1981 et une seconde en 1983. Le plan de sondage utilisé était le même que celui qui a été décrit dans la section 2 et comportait un appariement partiel des USPD dans une proportion de 50 %.

Les valeurs estimées de l'efficacité sont présentées dans le tableau 3. Les estimations de la variance et de la covariance de l'échantillon ont été utilisées pour calculer les variances optimums de T', T_1, T_2 et T_3 parce que les valeurs de ces variances et covariances pour l'ensemble de la population n'étaient pas connues. Par conséquent, le fait que les valeurs obtenues pour les variances estimées optimums de T_1 et T_2 sont faibles est attribuable d'une part, à l'utilisation de données d'échantillon et, d'autre part, à la nature même des estimateurs.

Nous constatons que l'estimateur T_2 est plus efficace que T_1 et que T_3, tandis que T_1 est plus efficace que T_3 dans l'estimation de la hauteur moyenne des arbres à l'aide de la circonférence comme information auxiliaire.

REMERCIEMENTS

Je tiens à remercier l'arbitre et le rédacteur associé, dont les précieux commentaires m'ont aidé à améliorer le présent document. Je remercie également Dr. O. Abe du Département de statistique de l'Université d'Ibadan, pour les retouches apportées à la version préliminaire du document révisé.

BIBLIOGRAPHIE

ABRAHAM, T.P., KHOSLA, R.K., et KATHURIA, O.P. (1969). Some investigations of the use of successive sampling in pest and disease surveys. *Journal of the Indian Society of Agricultural* Statistics, 21, 43-57.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3e éd). New York: John Wiley.

JESSEN, R.J. (1942). Statistical investigations of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, 304, 54-59.

KATHURIA, O.P. (1975). Some estimators in two-stage sampling on successive occasions with partial matching at both stages. *Sankhyā*, Sér. C, 37, 147-162.

KATHURIA, O.P. (1978). Double sampling on successive occasions using a two-stage design. *Journal of the Indian Society of Agricultural Statistics*, 30, 49-64.

KATHURIA, O.P. et SINGH, D. (1971). Relative efficiencies of some alternative procedures in two-stage sampling on successive occasions. *Journal of the Indian Society of Agricultural Statistics*, 23, 101-114.

SINGH, S., et SRIVASTAVA, A.K. (1973). Use of auxiliary information in two-stage successive sampling. *Journal of the Indian Society of Agricultural Statistics*, 25, 101-114.

SINGH, D. (1968). Estimates in successive sampling using a multistage design. *Journal of the American Statistical Association*, 63, 99-112.