Techniques d'enquête, juin 1998
Vol. 24, n° 1, pp. 81-91
Statistique Canada

Combinaison de bases multiples pour estimer la taille et les chiffres de la population

DAWN E. HAINES et KENNETH H. POLLOCK[1]

RÉSUMÉ

Le présent article traite de méthodes efficaces d'estimation de la taille et des chiffres de la population, à partir de données extraites de listes multiples et d'une base aréolaire indépendante. Ces travaux constituent un prolongement de la méthode proposée par Hartley (1962), qui porte sur deux bases de sondage générales. Un des principaux inconvénients des listes vient de ce que celles-ci sont habituellement incomplètes. Nous proposons dans cet article plusieurs méthodes pour pallier ces lacunes. Un plan d'échantillonnage mixte alliant l'utilisation d'une liste et d'une base aréolaire permet d'inclure des bases de sondage multiples et de couvrir entièrement la population-cible. Pour chaque combinaison de bases de sondage qui est proposée, nous indiquons les notations qui s'y rapportent, la fonction de vraisemblance et les estimateurs de paramètres. Nous présentons également les résultats d'une étude de simulation qui compare les diverses caractéristiques des estimateurs proposés.

MOTS CLÉS: Base de sondage incomplète; échantillonnage par capture et recapture; estimateur de sélection; méthode à base double; estimation par bases multiples.

1. INTRODUCTION

La théorie classique de l'échantillonnage présume que la base de sondage est complète. En pratique, toutefois, cette hypothèse s'avère souvent non confirmée. En effet, les imperfections dans la base de sondage, dues par exemple à des omissions, des dédoublements et des enregistrements erronés, sont presque inévitables dans tout large exercice de collecte de données (Hansen, Hurwitz et Madow 1953). L'information recueillie à partir de listes et de bases aréolaires est utilisée pour estimer la taille et les chiffres d'une population inconnue. À titre d'exemple, un écologiste ou un biologiste de la faune peuvent utiliser une liste et une base aréolaire pour estimer le nombre de nids d'aigles à tête blanche à l'intérieur d'une région donnée. De même, le U.S. Bureau of the Census utilise une technique d'estimation double pour mesurer le sous-dénombrement du recensement décennal. Pour leur part, Darroch, Fienberg, Glonek et Jonker (1993) décrivent une méthode de saisie multiple avec trois échantillons pour estimer la taille de la population, lorsque les probabilités d'inclusion sont hétérogènes. Dans un même ordre d'idées, des autorités agricoles pourraient être intéressées à estimer par exemple le nombre d'exploitations porcines et le nombre total de porcs en Caroline du nord. Il est courant que des sources multiples soient utilisées pour estimer la taille et les chiffres de la population.

Les listes présentent une liste des unités d'échantillonnage dans la population-cible. Ces listes sont établies au fil des ans, à partir de l'information obtenue des scientifiques, des autorités municipales, des comtés, des États et des organismes fédéraux. Les éléments d'information que l'on retrouve sur une liste d'échantillonnage incluent par exemple le nom, l'adresse, le numéro de téléphone, le numéro de sécurité sociale ou la description physique du lieu. Ces éléments et d'autres variables de stratification de toutes sortes sont utilisés pour identifier des personnes, des animaux, des entreprises ou d'autres établissements. Pour estimer le nombre de nids d'aigles à tête blanche dans une région, la liste d'échantillonnage de l'année courante est construite à partir de celle de l'an dernier. Avec l'ajout des nouveaux nids, la liste de l'an dernier devient rapidement désuète et incomplète; en raison de cette incomplétude, les estimations basées uniquement sur les listes sous-estiment habituellement la taille réelle de la population. L'addition d'information complémentaire tirée d'une base aréolaire peut s'avérer une méthode efficace pour estimer la taille et les chiffres de la population.

Une base aréolaire est un ensemble de régions géographiques définies par des frontières identifiables. L'ensemble de la région dans laquelle les données sont recueillies est divisée en unités d'échantillonnage exhaustives et s'excluant mutuellement, désignées segments. Les segments sont habituellement stratifiés en fonction d'une caractéristique d'intérêt. Lorsqu'un échantillon aléatoire stratifié de segments a été prélevé, les enquêteurs visitent les segments échantillonnés et notent les mesures pour toutes les unités de déclaration qui s'y trouvent.

Le National Agricultural Statistics Service (NASS) utilise actuellement une méthode à bases multiples pour l'échantillonnage et l'estimation d'un grand nombre de denrées agricoles. Fecso, Tortora et Vogel (1986) présentent une révision des bases d'échantillonnage utilisées pour le secteur agricole aux États-Unis, alors que Nealon

[1] Dawn E. Haines, U.S. Bureau of the Census, Washington, DC 20233; Kenneth H. Pollock, North Carolina State University, Department of Statistics, Box 8203, Raleigh, NC 27695-8203, U.S.A.

(1984) décrit en détails les estimateurs par bases multiples et base aréolaire qui sont utilisés par le ministère américain de l'Agriculture. Enfin, Kott et Vogel (1995) présentent une vue d'ensemble des enquêtes par bases multiples.

Nous examinons, à la section 2, les estimations obtenues à partir d'information extraite de deux ou plusieurs listes indépendantes et démontrons le lien qui existe entre ces méthodes et celles par capture et recapture. À la section 3, nous examinons des estimateurs plus efficaces de la taille et des chiffres de la population, lorsque l'information d'une base aréolaire indépendante est disponible. Nous étendons ensuite ces méthodes aux listes dépendantes, à la section 4. Les résultats d'une étude de simulation qui compare différents estimateurs sont résumés à la section 5. Enfin, la section 6 présente un résumé de nos résultats et discute de futures orientations de recherche.

2. LISTES D'ÉCHANTILLONNAGE MULTIPLES

2.1 Estimation de la taille de la population

Les listes utilisées pour estimer la taille de la population sont habituellement incomplètes et ne couvrent pas l'ensemble de la population. Une solution pour pallier ce problème d'incomplétude est de fusionner deux ou plusieurs listes incomplètes. Le fait de combiner ainsi plusieurs listes peut améliorer la couverture de la population-cible et, de ce fait, fournir de meilleurs estimateurs. Dans le cas de listes multiples, on présume habituellement que la probabilité d'inclusion dans une liste donnée est égale pour chaque élément de la population; les éléments de la liste constituent donc eux-mêmes nos «échantillons». À titre d'exemple, il existe une probabilité égale que les gens choisissent ou non d'inscrire leur numéro de téléphone dans l'annuaire. Dans le cas des nids d'aigles à tête blanche, la liste de cette année est construite à partir des nids observés l'année précédente. Si nous présumons que la probabilité qu'un nid soit dénombré est égale pour tous les nids, alors l'hypothèse qui précède est valide. L'hypothèse est également valide dans les expériences par capture et recapture où la première liste est formée de tous les animaux capturés au premier échantillonnage, alors que la deuxième contient tous les animaux capturés au deuxième échantillonnage. Ce scénario correspond au modèle M_t dans les ouvrages traitant de l'échantillonnage par capture et recapture (voir par exemple Otis, Burnham, White et Anderson (1978) pour plus de détails à ce sujet). Le modèle M_t suppose que le risque de capture, à chaque prélèvement, est le même pour tous les animaux dans la population; cette probabilité peut cependant varier d'un prélèvement à un autre.

Examinons d'abord deux listes indépendantes, B_1 et B_2. Supposons que B_1 compte N_{B_1} effectifs et que B_2 en compte N_{B_2}. Supposons également que le domaine $b_1(b_2)$ se compose des éléments $N_{b_1}(N_{b_2})$ qui appartiennent uniquement à la base $B_1(B_2)$ et que le domaine b_1b_2 contient les unités $N_{b_1b_2}$ qui appartiennent aux deux bases de sondage. Le domaine final inclut les éléments de la population-cible existante qui ne figurent dans aucune des deux listes; sa taille correspond à $N - N_{b_1} - N_{b_2} - N_{b_1b_2}$. La notation du domaine pour les listes B_1 et B_2 est présentée au tableau 1. À noter que chaque élément dans chaque base doit être réparti dans un domaine, sans erreur. Les erreurs dans la détermination du domaine sont graves et ne peuvent être corrigées ultérieurement. Ces erreurs ne sont pas examinées durant la phase d'estimation et sont donc considérées comme des erreurs non dues à l'échantillonnage. Selon Nealon (1984), la détermination du domaine constitue la principale source d'erreur non due à l'échantillonnage dans un plan d'échantillonnage à bases multiples (Kott et Vogel 1995).

Tableau 1
Notation du domaine pour les listes B_1 et B_2

Taille du domaine	Probabilité du domaine
N_{b_1}	$p_{b_1} = p_{B_1}(1 - p_{B_2})$
N_{b_2}	$p_{b_2} = (1 - p_{B_1})p_{B_2}$
$N_{b_1b_2}$	$p_{b_1b_2} = p_{B_1}p_{B_2}$
$N - N_{b_1} - N_{b_2} - N_{b_1b_2}$	$1 - p_{b_1} - p_{b_2} - p_{b_1b_2} = (1 - p_{B_1})(1 - p_{B_2})$

Supposons que la probabilité qu'un élément de la population soit inclus dans la liste $B_1(B_2)$ est $p_{B_1}(p_{B_2})$. Comme on présume que les listes B_1 et B_2 sont indépendantes, la probabilité qu'un élément appartienne au domaine b_1 est $p_{b_1} = p_{B_1}(1 - p_{B_2})$. Les probabilités pour les autres domaines sont définies de la même manière. La taille de la population N et les probabilités d'inclusion p_{B_1} et p_{B_2} sont des paramètres inconnus. La fonction de vraisemblance est définie par l'équation

$$\mathcal{L}(p_{B_1}, p_{B_2}, N \mid N_{b_1}, N_{b_2}, N_{b_1b_2}) = \begin{pmatrix} N \\ N_{b_1}, N_{b_2}, N_{b_1b_2} \end{pmatrix} * p_{B_1}^{N_{B_1}} p_{B_2}^{N_{B_2}} (1 - p_{B_1})^{N - N_{B_1}} (1 - p_{B_2})^{N - N_{B_2}}. \quad (1)$$

Les estimateurs du maximum de vraisemblance (EMV) des probabilités d'inclusion dans la base sont définis en maximisant le logarithme de la fonction de vraisemblance (1). Cette opération donne

$$\hat{p}_{B_1} = \frac{N_{B_1}}{\hat{N}} \quad \text{et} \quad \hat{p}_{B_2} = \frac{N_{B_2}}{\hat{N}}, \quad (2)$$

où l'EMV $\hat{N}$ remplace N. Plutôt que de dériver le logarithme de la fonction de vraisemblance pour établir la valeur approximative de N, nous utilisons la «méthode du ratio» pour maximiser la vraisemblance où $\mathcal{L}(N)$ est égal à $\mathcal{L}(N - 1)$ (Darroch 1958). Ce processus tient compte du paramètre discret N et donne l'équation

$$\frac{\mathcal{L}(N)}{\mathcal{L}(N-1)} = \frac{\hat{N}}{(\hat{N} - N_{b_1} - N_{b_2} - N_{b_1 b_2})} *$$

$$(1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) = 1. \qquad (3)$$

Nous présumons ici que N est large, de sorte que

$$\frac{N_{B_1}}{N-1} \approx \frac{N_{B_1}}{N} \quad \text{et} \quad \frac{N_{B_2}}{N-1} \approx \frac{N_{B_2}}{N}.$$

Si l'on remplace les estimateurs de (2) dans (3), on obtient alors

$$\hat{N}_1 = \hat{N} = \frac{N_{B_1} N_{B_2}}{N_{b_1 b_2}}. \qquad (4)$$

Sekar et Deming (1949) ont calculé une estimation de la variance de (4), exprimée par

$$\hat{V}(\hat{N}_1) = \frac{N_{B_1} N_{B_2} N_{b_1} N_{b_2}}{(N_{b_1 b_2})^3}.$$

Le remplacement de (4) dans (2) donne les EMV de p_{B_1} et p_{B_2},

$$\hat{p}_{B_1} = \frac{N_{b_1 b_2}}{N_{B_2}} \quad \text{et} \quad \hat{p}_{B_2} = \frac{N_{b_1 b_2}}{N_{B_1}}.$$

L'estimateur $\hat{N}_1$ de N dans (4) est désigné estimateur de Lincoln-Petersen, dans les modèles par capture et recapture à l'intérieur d'une population fermée. Les éléments de la liste B_1 peuvent être considérés comme les unités saisies lors du premier échantillonnage, alors que les éléments de la liste B_2 seraient les unités saisies au deuxième échantillonnage. Les éléments dans le domaine $b_1 b_2$ correspondent aux éléments saisis à la recapture. Étant donné cette correspondance, on constate facilement que la fonction de vraisemblance pour la taille de la population et les probabilités de capture, pour les deux échantillonnages, sera la même qu'en (1). Par conséquent, les EMV calculés pour deux listes indépendantes seront les mêmes que les EMV correspondants avec le modèle par capture et recapture avec deux échantillonnages.

Si nous poussons plus loin ces hypothèses, nous pouvons prétendre que le fait de combiner k listes indépendantes correspond directement au fait d'avoir k échantillonnages avec le modèle M_t selon des modèles par capture et recapture à l'intérieur d'une population fermée, où $t = k$ (Otis et coll. 1978). La fonction de vraisemblance générale pour k listes indépendantes, $B_1, B_2, ..., B_k$, prend la forme

$$\mathcal{L}(p_{B_1}, ..., p_{B_k}, N \mid N_{b_1}, ..., N_{b_1 ... b_k}) =$$

$$\binom{N}{N_{b_1}, ..., N_{b_1 ... b_k}} \prod_{l=1}^{k} p_{B_l}^{N_{B_l}} (1 - p_{B_l})^{N - N_{B_l}}, \qquad (5)$$

laquelle fonction a exactement la même structure que la fonction de vraisemblance définie par Darroch (1958) et elle est décrite plus en détail par Otis et coll. (1978) et Seber (1982). Les probabilités d'inclusion dans la base de sondage sont définies comme suit

$$\hat{p}_{B_l} = \frac{N_{B_l}}{\hat{N}}, \quad l = 1, ..., k. \qquad (6)$$

Les valeurs de $\hat{N}$ sont établies en calculant de façon numérique le polynôme de $(k - 1)$ degré dans $\hat{N}$ provenant de l'égalité

$$\frac{\mathcal{L}(N)}{\mathcal{L}(N-1)} = \frac{\hat{N}}{(\hat{N} - N_{b_1} - \cdots - N_{b_1 ... b_k})} *$$

$$(1 - \hat{p}_{B_1}) \cdots (1 - \hat{p}_{B_k}) = 1. \qquad (7)$$

Nous choisissons ensuite $\hat{N}$ comme étant la racine qui maximise la valeur de la fonction de vraisemblance (5). L'introduction de cette racine dans (6) donne les EMV des k probabilités d'inclusion dans la base.

2.2 Estimation des chiffres de population

Supposons que les valeurs mesurées y_i sont connues pour toutes les unités dans les k listes indépendantes. La probabilité estimée que le premier élément soit inclus dans au moins une des k listes est égale à

$$\hat{\pi}_1 = \hat{P}\left[\cup_{l=1}^{k} B_l\right] = 1 - (1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) \cdots (1 - \hat{p}_{B_k}),$$

où $\hat{p}_{B_l} = N_{B_l}/\hat{N}$ et $\hat{N}$ est l'EMV de N calculé à partir de l'équation (7), laquelle équation (7)

$$\frac{\hat{N}}{(\hat{N} - N_{b_1} - \cdots - N_{b_1 ... b_k})} (1 - \hat{\pi}_1) = 1$$

devient, sous forme simplifiée,

$$\hat{\pi}_1 = \frac{N_{b_1} + \cdots + N_{b_1 ... b_k}}{\hat{N}}.$$

Un estimateur de Horvitz-Thompson (1952) des chiffres de la population est exprimé par

$$\hat{\hat{Y}}_{\text{H-T}} = \frac{1}{\hat{\pi}_1} \sum_{i \in B_1 \cup ... \cup B_k} y_i$$

$$= \frac{\hat{N}}{N_{b_1} + \cdots + N_{b_1 ... b_k}} \sum_{i \in B_1 \cup ... \cup B_k} y_i = \hat{N} \bar{Y}_L,$$

où $\bar{Y}_L$ est la moyenne des éléments distincts dans les listes. Par conséquent, pour k listes indépendantes, l'estimateur

estimé de Horvitz-Thompson coïncide avec l'estimateur des chiffres de la population proposé par Pollock, Turner et Brown (1994).

Dans certains cas, les valeurs de la variable d'intérêt, y_i, ne sont pas disponibles pour toutes les unités dans les listes. Si les listes sont grandes, des échantillons aléatoires sont alors prélevés de chaque liste et des données sont recueillies sur ces sous-échantillons. S'il y a k listes, il est possible de définir 2^k domaines. Nous examinons un prolongement de l'estimateur proposé par Lund (1968) pour le total de toutes les unités dans les listes

$$\hat{Y}_{L,L} = \sum_{l=1}^{2^k-1} N_l \bar{y}_l,$$

lequel est la somme pondérée de $2^k - 1$ moyennes du domaine, $\bar{y}_l$. Les facteurs de pondération sont déterminés en fonction de la taille du domaine. L'estimateur des chiffres de la population est

$$\hat{Y} = \hat{N} \frac{\hat{Y}_{L,L}}{\sum_{l=1}^{2^k-1} N_l}.$$

3. LISTES MULTIPLES COMBINÉES À UNE BASE ARÉOLAIRE

3.1 Estimation de la taille de la population

Combiner de multiples listes individuelles à une base aréolaire est une des solutions qui s'offrent pour pallier les lacunes des listes d'échantillonnage. Supposons que la région géographique qui nous intéresse est subdivisée en U_A segments. Supposons également qu'un échantillon aléatoire simple formé de u_A segments est sélectionné à partir des U_A segments qui couvrent l'ensemble de la population. Par conséquent, la probabilité qu'un segment soit sélectionné correspond à $p_A = u_A/U_A$. Dans certaines enquêtes, il est possible de subdiviser la région en segments de taille à peu près égale. En pareils cas, la probabilité de sélection d'un segment correspond à peu près à la proportion de la région échantillonnée. L'inclusion d'une base aréolaire ajoute à l'intégralité de la population-cible (Hartley 1962). Nous supposons que chaque unité de déclaration appartient à exactement un segment. Lorsqu'un segment est sélectionné, toutes les unités de déclaration à l'intérieur du segment sont observées. Pour estimer, par exemple, le nombre de nids d'aigles à tête blanche, on suppose que chaque nid n'appartient qu'à un segment et un seul. Cependant, cette hypothèse n'est pas toujours valide. Examinons par exemple le cas d'une exploitation porcine qui s'étendrait au-delà des frontières du segment; dans un tel cas, les éléments de la population peuvent être associés à plus d'un segment. Pour résoudre ce problème, des règles d'association établissant des liens entre les éléments de la population et les segments sont définies durant l'étape de l'estimation. Voir Faulkenberry et Garoui (1991) pour plus de détails à ce sujet. Le National Agricultural Statistics Service utilise trois règles de correspondance pour répartir les éléments de la population entre les segments échantillonnés. Les estimateurs de segment ouvert, fermé et pondéré sont décrits dans Nealon (1984) et aussi dans Sirken (1970).

Examinons le cas où nous avons k listes indépendantes et une base aréolaire. La taille de la population, N, et les probabilités d'inclusion dans la liste, p_{B_i}, $i = 1, ..., k$, sont des paramètres inconnus. Cependant, la probabilité d'inclusion dans la base aréolaire $p_A = u_A/U_A$ est connue. La fonction de vraisemblance est représentée par

$$\mathcal{L}(p_{B_1}, ..., p_{B_k}, N | p_A, n_a, n_{ab_1}, ..., n_{ab_1 \ldots bk}, N_{b_1}, ..., N_{b_1 \ldots b_k})$$

$$= \binom{N}{n_a, n_{ab_1}, ..., n_{ab_1 \ldots b_k}, N_{b_1}, ..., N_{b_1 \ldots b_k}} p_A^{n_A} (1 - p_A)^{N - n_A}$$

$$\prod_{l=1}^{k} p_{B_l}^{N_{B_l}} (1 - p_{B_l})^{N - N_{B_l}},$$

où n_A est le nombre total d'éléments dans les u_A segments de la région échantillonnée et n_a est le nombre d'éléments dans les u_A segments de la région échantillonnée qui n'appartiennent à aucune liste. De même, $n_{ab_1}, ..., n_{ab_1 \ldots b_k}$, $N_{b_1}, ..., N_{b_1 \ldots b_k}$ sont définis comme étant les effectifs des différents domaines. Il est important d'insister sur le fait que l'inclusion d'une base aréolaire peut entraîner la modification de la valeur de N_{b_1}. N_{b_1} correspond maintenant au nombre d'éléments dans la liste B_1 qui ne sont inclus, ni dans les u_A segments de la région sélectionnée, ni dans aucune autre liste.

Les EMV des paramètres sont représentés par $\hat{p}_{B_l} = N_{B_l}/\hat{N}$, où $\hat{N}$ est une solution au polynôme de k-ième degré

$$\hat{N}(1 - p_A)(1 - \hat{p}_{B_1}) \ldots (1 - \hat{p}_{B_k}) =$$

$$(\hat{N} - n_a - n_{ab_1} - \cdots - n_{ab_1 \ldots b_k} - N_{b_1} - \cdots - N_{b_1 \ldots b_k}). \quad (8)$$

Des méthodes numériques sont essentielles pour résoudre l'équation (8) servant à calculer l'EMV $\hat{N}$ de N. Parmi les k racines de (8), nous sélectionnons $\hat{N}$ qui maximise la vraisemblance.

En appliquant cette méthode à une liste et une base aréolaire, nous obtenons

$$\hat{N} = N_{B_1} + \frac{n_a}{p_A}. \quad (9)$$

Cet estimateur est également connu sous le nom d'estimateur de sélection (Kott et Vogel 1995), lequel répartit les éléments en deux groupes distincts. Le premier groupe renferme les éléments qui appartiennent à la fois à la liste et à la base aréolaire et il est désigné domaine de chevauchement. Comme on présume que tous les éléments dans une liste appartiennent à la base aréolaire, la taille du domaine de chevauchement coïncide avec le nombre

d'éléments dans la base B_1 et a la valeur N_{B_1}. Le deuxième groupe renferme les éléments de la base aréolaire qui ne sont pas inclus dans la ou les listes; il est donc désigné domaine sans chevauchement. La taille de ce dernier domaine correspond à une quantité aléatoire non observée, N_a. Le terme n_a désigne le nombre d'éléments que l'on trouve dans les u_A segments qui ne sont pas inclus dans la ou les listes, en vertu d'une règle d'association précise. Une valeur estimée de N_a est n_a/p_A. Par conséquent, $\hat{N}$, à l'équation (9), fournit une estimation de la taille de la population. L'EMV de p_{B_1} qui en résulte est

$$\hat{p}_{B_1} = \frac{N_{B_1}}{N_{B_1} + \dfrac{n_a}{p_A}}.$$

Lorsque des listes multiples sont disponibles, il est possible de les combiner en une seule liste et d'utiliser l'estimateur qui précède pour obtenir une valeur estimée de N. En d'autres mots, supposons que nous avons l'estimateur de sélection

$$\hat{N}_2 = \hat{N} = N_{B_1 \cup \ldots \cup B_k} + \frac{n_a}{p_A} = N_{b_1} + \cdots + N_{b_k} + N_{b_1 b_2} + \cdots + N_{b_1 \ldots b_k} + \frac{n_a}{p_A}. \quad (10)$$

À noter que l'estimateur de sélection $\hat{N}_2$ convient, même lorsque les listes ne sont pas indépendantes les unes des autres. Nous discutons de ce point plus en détails à la section 4.

L'utilisation de cette méthode pour une base aréolaire et deux listes indépendantes donne la fonction de vraisemblance

$$\mathcal{L}(p_{B_1}, p_{B_2}, N \mid p_A, n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2}) = \binom{N}{n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2}} p_A^{n_A} p_{B_1}^{N_{B_1}} p_{B_2}^{N_{B_2}} (1 - p_A)^{N - n_A} (1 - p_{B_1})^{N - N_{B_1}} (1 - p_{B_2})^{N - N_{B_2}}.$$

L'EMV de N est

$$\hat{N}_3 = \hat{N} = (2p_A)^{-1} * \left[(N_{B_1} + N_{B_2})p_A + (n_a - N_{b_1 b_2} - n_{ab_1 b_2})\right] + (2p_A)^{-1} \sqrt{\left[(N_{B_1} + N_{B_2})p_A + (n_a - N_{b_1 b_2} - n_{ab_1 b_2})\right]^2 + 4p_A(1 - p_A)N_{B_1}N_{B_2}}, \quad (11)$$

où $n_{ab_1 b_2}$ représente le nombre d'éléments inclus dans les u_A segments de la région échantillonnée qui appartiennent aux deux listes. On peut obtenir une estimation de la variance de $\hat{N}_3$ en ayant recours à l'approximation par série de Taylor de (11) et à la distribution asymptotique de $(N_{B_1}, N_{B_2}, n_a, N_{b_1 b_2}, n_{ab_1 b_2})$.

3.2 Estimation des chiffres de la population

Lorsque les valeurs de y_i sont connues pour tous les éléments dans les k listes indépendantes ainsi que pour un échantillon de segments d'une base aréolaire, nous utilisons un estimateur de Horvitz-Thompson pour estimer les chiffres de population. Rappelons-nous les hypothèses formulées:

1. La probabilité qu'une unité soit incluse dans la i-ième liste, p_{B_i}, est égale pour toutes les unités.
2. L'inclusion d'une unité dans une base est indépendante de son inclusion dans une autre base.
3. La probabilité qu'une unité soit incluse dans l'échantillon de la base aréolaire formé de u_A segments correspond à $p_A = u_A / U_A$.

Comme nous supposons que les unités de la population n'appartiennent qu'à un segment de la région et que toutes les unités à l'intérieur d'un segment échantillonné sont observées, la troisième hypothèse est valide. Par conséquent, la probabilité que le i-ième élément soit inclus dans au moins une des k listes ou dans l'échantillon de la base aréolaire, ou les deux, est

$$\tilde{\pi}_1 = 1 - (1 - p_A)(1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) \cdots (1 - \hat{p}_{B_k}) = \frac{n_a + n_{ab_1} + \cdots + N_{b_1 \ldots b_k}}{\hat{N}}.$$

L'estimateur de Horvitz-Thompson pour les chiffres de la population est

$$\hat{\hat{Y}}_{\text{H-T}} = \frac{\hat{N}}{n_a + n_{ab_1} + \cdots + N_{b_1 \ldots b_k}} \sum_{i \in \text{ échantillon}} y_i = \hat{N}\,\bar{y}_L,$$

où $\bar{y}_L$ est la moyenne des éléments distincts dans les listes $B_1, \ldots, B_k$ et des éléments dans l'échantillon de la base aréolaire.

Nous pouvons également utiliser l'estimateur de sélection pour estimer les chiffres de population. Le total du domaine de chevauchement connu est combiné à un estimateur du total du domaine sans chevauchement (NOL) pour donner $\hat{Y}_S = Y_L + \sum_{i \in \text{NOL}} y_i / p_A$. Le domaine sans chevauchement est formé des éléments de la base aréolaire qui ne figurent dans aucune liste et $Y_L = Y_{B_1 \cup \ldots \cup B_K}$ est le total des unités distinctes dans les k listes. Dans le cas du sous-échantillonnage, nous pouvons remplacer Y_L dans $\hat{Y}_S$ par l'estimateur de Lund, représenté par

$$\hat{Y}_{L,L} = N_{b_1}\bar{y}_{b_1} + \cdots + N_{b_k}\bar{y}_{b_k} + N_{b_1 b_2}\bar{y}_{b_1 b_2} + \cdots + N_{b_1 \ldots b_k}\bar{y}_{b_1 \ldots b_k}.$$

4. LISTES DÉPENDANTES

Examinons maintenant le cas où il y a dépendance entre les listes, mais où la base aréolaire et les listes demeurent indépendantes. Par exemple, dans les expériences par capture et recapture, la probabilité qu'un animal soit capturé au deuxième échantillonnage peut dépendre de sa capture au premier échantillonnage. Voir Fienberg (1972), Cormack (1989), Wolter (1990), Pollock, Hines et Nichols (1984), Huggins (1989) et Alho (1990) pour obtenir des exemples précis.

Prenons le cas où nous avons deux listes, B_1 et B_2, qui sont dépendantes. Supposons que p_{11} représente la probabilité d'être inclus dans les deux listes. Si B_1 et B_2 sont indépendantes, alors $p_{11} = p_{B_1} p_{B_2}$ où p_{B_1} et p_{B_2} sont les probabilités d'inclusion, respectivement pour B_1 et B_2. Supposons également que $p_{10}(p_{01})$ est la probabilité d'être inclus dans la liste $B_1(B_2)$ mais non dans la liste $B_2(B_1)$. La probabilité d'exclusion des deux listes est représentée par $p_{00} = 1 - p_{B_1} - p_{B_2} + p_{11}$.

La fonction de vraisemblance est définie par l'équation

$$\mathcal{L}(p_{B_1}, p_{B_2}, p_{11}, N \mid p_A, n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1b_2}, n_{ab_1b_2})$$

$$= \binom{N}{n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1b_2}, n_{ab_1b_2}} p_A^{n_A} (1 - p_A)^{N - n_A}$$

$$(p_{B_1} - p_{11})^{N_{b_1} + n_{ab_1}} (p_{B_2} - p_{11})^{N_{b_2} + n_{ab_2}} p_{11}^{N_{b_1b_2} + n_{ab_1b_2}}$$

$$(1 - p_{B_1} - p_{B_2} + p_{11})^{N - N_{b_1} - N_{b_2} - n_{ab_1} - n_{ab_2} - N_{b_1b_2} - n_{ab_1b_2}}. \quad (12)$$

En maximisant la valeur de (12) par rapport à p_{B_1}, p_{B_2}, p_{11} et N, on obtient l'approximation suivante

$$\hat{N} = N_{b_1} + N_{b_2} + n_{ab_1} + n_{ab_2} + N_{b_1b_2} + n_{ab_1b_2} + \frac{n_a}{p_A},$$

qui coïncide avec l'estimateur de sélection $\hat{N}_2$. En d'autres mots, $\hat{N}$ est également l'estimateur qui est obtenu en combinant les deux listes en une seule, où les dédoublements sont éliminés et où la taille du domaine sans chevauchement est estimée à partir de l'échantillon de la base aréolaire. Il peut également être démontré que la méthode du maximum de vraisemblance à deux degrés de Sanathanan (1972) mène à

$$\hat{N} = \frac{n_a + N_{B_1 \cup B_2}}{p_A + (1 - p_A)\dfrac{N_{B_1 \cup B_2}}{\hat{N}_2}}$$

$$= \hat{N}_2.$$

Par conséquent, l'estimateur du maximum de vraisemblance et l'estimateur de Sanathanan coïncident tous deux avec l'estimateur de sélection. Si l'on possède des données provenant de deux listes dépendantes mais que la nature du lien de dépendance est inconnue, alors nous ne pouvons estimer les paramètres individuels. Lorsque l'information d'une base aréolaire indépendante est disponible, tous les paramètres sont estimables. Cependant, pour estimer N, il suffit d'avoir $N_{B_1 \cup B_2}$ aucune information additionnelle ne nous est donnée par N_{B_1}, N_{B_2}, et $N_{b_1b_2}$.

Il existe différentes méthodes pour modéliser la dépendance entre k listes, pour estimer la taille et les chiffres de la population. Des informations additionnelles sur la population ou de l'information provenant d'une base aréolaire indépendante sont nécessaires pour modéliser avec précision la dépendance. Fienberg (1972) et Cormack (1989) proposent des modèles loglinéaires contraints pour modéliser la dépendance. Pour sa part, Wolter (1990) utilise des contraintes externes, comme un rapport de masculinité connu, pour estimer la taille de la population lorsqu'il y a dépendance. Une autre technique consiste à modéliser les probabilités d'inclusion comme étant une fonction des covariables. Alho, Mulry, Wurdeman et Kim (1993) utilisent un modèle de régression logistique conditionnel pour estimer la probabilité d'être dénombré lors d'un recensement et appliquent ce modèle à l'enquête postcensitaire de 1990. Le rôle des variables auxiliaires dans les expériences par capture et recapture avec probabilités inégales est examiné par Pollock et coll. (1984), Huggins (1989) et Alho (1990).

5. ÉTUDE DE SIMULATION

Nous proposons une étude de simulation pour évaluer l'efficacité globale de différents estimateurs de la taille de la population, dans le cas spécial où il y a utilisation de deux listes et d'une base aréolaire. Il s'agit de la combinaison de bases d'échantillonnage la plus facilement réalisable pour résoudre les problèmes réels inhérents aux enquêtes.

5.1 Conception de l'étude

Afin d'étudier à la fois les cas dépendants et indépendants, nous définissons le paramètre θ qui reflète la structure de dépendance entre les listes B_1 et B_2. Ce paramètre a la même forme que le rapport de cotes et il est représenté officiellement par l'équation

$$\theta = \frac{p_{00} p_{11}}{p_{01} p_{10}}.$$

Dans le cas des deux listes, la valeur de θ détermine une solution unique pour p_{11}. Dans notre étude, les facteurs varient comme suit:

Facteur	Niveau	Définition
N	500, 5 000	Taille de la population
p_A	0,05, 0,10, 0,20	Probabilité d'inclusion pour la base aréolaire A
$p_{B_1} (= p_{B_2})$	0,7, 0,9	Probabilité d'inclusion pour la liste $B_1(B_2)$
θ	0,5, 1,0, 1,5, 2,0	Rapport de cotes

Pour chaque combinaison paramétrique, nous produisons les données (n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, $N_{b_1b_2}$, $n_{ab_1b_2}$). Un millier de répétitions de Monte Carlo sont effectuées pour chaque combinaison paramétrique.

5.2 Estimateurs

Nous comparons quatre estimateurs de la taille de la population, $\hat{N}_1, \hat{N}_2, \hat{N}_3$, et $\hat{N}_4$. $\hat{N}_1$ est l'estimateur de Lincoln-Petersen qui n'inclut pas d'information de la base aréolaire. L'estimateur $\hat{N}_1$ convient lorsque les listes sont indépendantes. Comme cet estimateur ne tient pas compte de l'information de l'échantillon de la base aréolaire, on s'attend à ce qu'il soit inefficace lorsque l'information de la base aréolaire est disponible. L'estimateur de sélection, $\hat{N}_2$, fait la somme des estimations par domaines de chevauchement et sans chevauchement et convient tout particulièrement dans les cas de listes dépendantes. Le troisième estimateur, $\hat{N}_3$, est calculé à partir de la fonction de vraisemblance de la base de sondage indépendante intégrale. Cet estimateur s'appuie sur l'information extraite de la base aréolaire et sur le fait que les listes sont indépendantes ($\theta = 1$).

Nous croyons que $\hat{N}_3$ est le meilleur estimateur lorsque les listes B_1 et B_2 sont indépendantes, tandis que $\hat{N}_2$ serait le meilleur lorsqu'il y a dépendance. Nous avons donc examiné également un estimateur d'avant essai pour tester l'indépendance des listes. Selon notre définition, $\hat{N}_4$ est égal à $\hat{N}_2$ lorsque les données portent fortement à croire que les listes B_1 et B_2 ne sont pas indépendantes. Sinon, $\hat{N}_4 = \hat{N}_3$. Officiellement,

$$\hat{N}_4 = \begin{cases} \hat{N}_2 \text{ si VDA} > \chi^2_{1,0,05} = 3{,}84 \\ \hat{N}_3 \text{ autrement,} \end{cases}$$

où VDA est la variable chi-carré du test de validité de l'ajustement pour tester H_0: $\theta = 1$ et est calculé à partir du tableau à double entrée suivant.

	Dans B_1	Pas dans B_1	
Dans B_2	$n_{ab_1b_2}$	n_{ab_2}	$n_{A\cap B_2}$
Pas dans B_2	n_{ab_1}	n_a	$n_{A\cap B_2'}$
	$n_{A\cap B_1}$	$n_{A\cap B_1'}$	n_A

Figure 1. Classification des éléments échantillonnés à partir de la base aérolaire

La figure 1 répartit les n_A éléments selon qu'ils sont présents dans les listes B_1 et B_2 ou qu'ils en sont absents.

5.3 Comparaison des estimateurs

Les tableaux 2 et 3 présentent les pourcentages du biais relatif et de l'erreur quadratique moyenne relative des estimateurs $\hat{N}_1, \hat{N}_2, \hat{N}_3$, et $\hat{N}_4$ pour des populations de tailles correspondant respectivement à 500 et 5 000. Nous réduisons le biais et l'erreur quadratique moyenne par N afin de pouvoir comparer directement des estimateurs basés sur des populations de tailles différentes. Une comparaison entre $\hat{N}_1$ et $\hat{N}_3$ montre l'avantage qu'il y a à prélever un échantillon de la base aréolaire. En pratique, ces avantages dépendent du coût relatif de l'échantillon de la base aréolaire. Cependant, nous ne tenons pas compte, dans la présente étude, des coûts d'échantillonnage. La probabilité d'être inclus dans les deux listes, p_{11}, est indiquée entre parenthèses, sous la colonne θ. Lorsque $p_B = p_C = 0{,}9$, la valeur de p_{11} doit se situer entre 0,8 et 0,9. Cependant, lorsque θ varie de 0,5 à 2, p_{11} se situe uniquement entre 0,806 et 0,817.

L'estimateur $\hat{N}_2$ est sans biais pour N et a le plus faible pourcentage de biais relatif. Les estimateurs $\hat{N}_1$ et $\hat{N}_3$ sont asymptotiquement cohérents pour N et donnent des biais dont la valeur se rapproche de 0, lorsque $\theta = 1$. Par contre, $\hat{N}_1$ et $\hat{N}_3$ ont un large biais lorsque $\theta \neq 1$. Le biais relatif, en pourcentage, de $\hat{N}_4$ est inférieur à celui de $\hat{N}_3$ mais il ne se rapproche pas de zéro. Le biais ne change pas de façon significative à mesure que p_A augmente de 0,05 à 0,10 à 0,20.

Lorsque $N = 500$ et $p_B = p_C = 0{,}9$, $\hat{N}_3$ a le plus faible pourcentage d'erreur quadratique moyenne relative. Ceci s'explique notamment du fait que l'éventail limité des valeurs de p_{11} est similaire à la valeur de p_{11} lorsqu'il y a indépendance (0,810). Le pourcentage de l'erreur quadratique moyenne relative pour $\hat{N}_3$ est de 40 à 50 % inférieur à celui de $\hat{N}_2$. Par contre, il n'est que de 15 à 30 % inférieur à celui de $\hat{N}_1$. Par conséquent, lorsque la probabilité d'inclusion dans les listes est très élevée, alors $\hat{N}_1$ et $\hat{N}_3$ sont tous deux nettement préférables à $\hat{N}_2$. En outre, si les coûts d'échantillonnage dans la base aréolaire sont élevés, alors $\hat{N}_1$ peut s'avérer un estimateur de remplacement acceptable pour $\hat{N}_3$. Lorsque $N = 500$ et $p_B = p_C = 0{,}7$, c'est $\hat{N}_3$ qui a la plus faible erreur quadratique moyenne relative, en pourcentage, lorsqu'il y a indépendance; lorsque $\theta = 2$, c'est $\hat{N}_2$ qui a le plus faible pourcentage d'erreur quadratique moyenne relative. Si $N = 5\,000$ et $p_B = 0{,}7$, alors $\hat{N}_3$ a le plus faible pourcentage d'erreur quadratique moyenne relative, seulement lorsque $\theta = 1$. Pour toutes les autres valeurs de θ, c'est $\hat{N}_2$ qui obtient le plus faible pourcentage. Dans tous les cas, $\hat{N}_3$ présente une très faible variance et l'erreur quadratique moyenne relative, exprimée en pourcentage, est due principalement au biais dans $\hat{N}_3$. Pour $\theta < 1$, $\hat{N}_3$ tend à avoir un biais positif, alors que $\hat{N}_3$ a un biais négatif lorsque $\theta > 1$. Dans le cas où $N = 5\,000$ et $p_B = 0{,}9$, $\hat{N}_3$ le plus faible pourcentage d'erreur quadratique moyenne relative lorsque $\theta = 1$, mais c'est $\hat{N}_2$ qui a le plus faible pourcentage lorsque $\theta = 0{,}5$ et 2. Enfin, lorsque $\theta = 1{,}5$, aucun estimateur n'est supérieur à un autre, en ce qui a trait au pourcentage d'erreur quadratique moyenne relative.

Comme prévu, les pourcentages d'erreur quadratique moyenne relative de $\hat{N}_2$, $\hat{N}_3$, et $\hat{N}_4$ diminuent à mesure que la valeur de p_A augmente. Par conséquent, à mesure que

l'information extraite de la base aréolaire augmente, ces pourcentages diminuent. De même, à mesure que la taille de la population augmente, de 500 à 5 000, les pourcentages d'erreur quadratique moyenne relative diminuent eux aussi. Comme les valeurs de p_A dans notre étude de simulation sont faibles, la variance de $\hat{N}_2$ est élevée. Cependant, même si $\hat{N}_3$ comporte un biais, son erreur-type est très faible et ceci se traduit par un pourcentage d'erreur quadratique moyenne relative plus faible. L'estimateur $\hat{N}_4$ réduit le biais de $\hat{N}_3$ mais comporte une grande erreur-type; $\hat{N}_4$ n'est donc pas un estimateur particulièrement utile. Lorsque les valeurs de θ et p_A sont plus élevées, $\hat{N}_2$ devrait donner de meilleurs résultats que $\hat{N}_3$. Pour leurs valeurs de θ et p_A examinées ici, nous recommandons d'utiliser l'estimateur $\hat{N}_3$, de préférence à tous les autres proposés.

Dans la plupart des cas, la valeur de l'erreur quadratique moyenne relative, en pourcentage, de $\hat{N}_4$ se situe entre celle de $\hat{N}_2$ et $\hat{N}_3$. Nous écrivons l'estimateur $\hat{N}_4$ comme étant $\hat{N}_4 = \delta\hat{N}_2 + (1 - \delta)\hat{N}_3$, où $\delta = 0$ ou 1, selon les résultats du test de validité de l'ajustement. Il n'est pas nécessaire que les pourcentages de l'erreur quadratique moyenne relative et du biais relatif de $\hat{N}_4$ se situent entre ceux de $\hat{N}_2$ et $\hat{N}_3$, parce que δ n'est pas indépendant de $\hat{N}_2$ et $\hat{N}_3$.

5.4 Limitations de l'étude

Notre étude avait pour but de comparer le biais, l'erreur-type et l'erreur quadratique moyenne de quatre estimateurs de la taille de la population, en présumant de probabilités d'inclusion égales dans les deux listes. De futures études pourraient être menées en incluant des probabilités d'inclusion inégales et des valeurs de θ plus élevées. De toute évidence, l'avantage de $\hat{N}_3$ sur $\hat{N}_1$ dépend du coût d'échantillonnage d'une base aréolaire. Pour notre étude, nous avons retenu uniquement les cas où les valeurs de p_A étaient faibles. Or de faibles valeurs de p_A sont associées à des coûts élevés d'échantillonnage par base aréolaire. Cependant, même dans ce cas, nous observons une réduction significative des pourcentages de l'erreur quadratique moyenne relative et du biais relatif, ce qui justifie l'utilisation de $\hat{N}_3$ sur $\hat{N}_1$. Nous n'examinons pas ici de fonction objective qui tiendrait compte à la fois des coûts d'échantillonnage et des pourcentages de l'erreur quadratique moyenne relative et du biais relatif.

Tout au long de cet article, nous avons présumé que la probabilité d'inclusion, à l'intérieur d'une liste donnée, était la même pour toutes les unités. Haines (1997) examine le cas où les probabilités d'inclusion sont représentées comme étant une fonction d'une covariable. Lorsque les probabilités d'inclusion sont hétérogènes, il se peut que la probabilité d'inclusion dans une liste soit alors plus élevée pour les grandes unités que pour les petites. Les probabilités d'inclusion hétérogènes jouent un rôle important dans l'estimation des chiffres de la population, lorsque la variable de réponse présente une distribution fortement asymétrique ou des valeurs rares. Haines (1997) propose également deux méthodes de stratification qui sont utiles lorsque la stratification d'une base aréolaire et des listes se fait en fonction de la même variable. Ces résultats feront l'objet de futures publications.

6. DISCUSSION

Cette étude porte principalement sur l'estimation de la taille de la population, à partir de plusieurs bases de sondage. L'information provenant d'une base aréolaire et/ou d'une ou de plusieurs listes est recueillie et combinée, pour obtenir divers estimateurs. Nous calculons les estimateurs de la taille de la population, lorsque l'information est disponible uniquement pour k listes indépendantes et aussi lorsque l'information est disponible pour un échantillon de la base aréolaire, en plus des listes. Nous présentons ensuite une étude de simulation, qui vise à comparer la performance des estimateurs dans le cas spécial où l'on utilise deux listes et une base aréolaire. À la lumière des résultats de cette simulation, nous recommandons l'estimateur calculé à partir de la fonction de vraisemblance indépendante intégrale, $\hat{N}_3$, lorsque les listes sont indépendantes ou presque indépendantes. Dans les cas de dépendance de modérée à forte, nous recommandons plutôt l'estimateur de sélection $\hat{N}_2$.

Nous examinons aussi l'estimation des chiffres de la population, en regard de deux scénarios. Dans le premier cas, nous supposons que des observations sont disponibles pour toutes les unités qui forment les listes. Dans le deuxième cas, par contre, nous supposons que l'information n'est disponible que pour des sous-échantillons de chaque liste. Nous utilisons un estimateur de type Horvitz-Thompson lorsque les listes sont indépendantes et un estimateur de sélection, lorsque les listes sont dépendantes.

Dans cette étude, nous nous sommes intéressés principalement à l'estimation de la taille de la population. En pratique, toutefois, il se pourrait que l'on veuille estimer les chiffres de la population en regard de plusieurs caractéristiques, selon un plan d'échantillonnage à plusieurs degrés avec probabilités d'inclusion inégales. Au nombre des études qui traitent de ce dernier sujet, mentionnons celles de Bankier (1986), Skinner (1991) et Skinner, Holmes et Holt (1994).

7. REMERCIEMENTS

Les auteurs aimeraient remercier le rédacteur et les deux examinateurs pour leurs commentaires utiles au sujet d'une version antérieure du présent article. Cette recherche a été financée partiellement par le U.S. Geological Survey, Biological Resources Division. Christine Bunck est directrice du programme BEST. Les vues qui y sont exprimées appartiennent aux auteurs et ne sont pas nécessairement partagées par le Census Bureau.

Tableau 2
Résultats des simulations pour $N = 500$

			p_A					
			0,05		0,10		0,20	
p_B	θ		% Biais relatif	% PEQMR	% Biais relatif	% PEQMR	% Biais relatif	% PEQMR
0,7	0,5 (0,462)	$\hat{N}_1$	62,30	66,01	60,64	64,04	63,26	66,81
		$\hat{N}_2$	0,30	49,07	-0,75	32,37	0,85	22,58
		$\hat{N}_3$	55,52	58,95	48,15	51,15	40,53	43,32
		$\hat{N}_4$	48,15	58,88	37,88	49,25	24,95	38,80
	1 (0,490)	$\hat{N}_1$	0,47	19,26	1,01	19,08	-0,11	19,45
		$\hat{N}_2$	0,45	57,34	0,34	39,61	0,88	27,25
		$\hat{N}_3$	0,43	18,21	0,83	16,93	0,14	15,75
		$\hat{N}_4$	2,40	27,57	1,39	22,94	0,29	17,96
	1,5 (0,508)	$\hat{N}_1$	-35,60	40,06	-36,48	40,58	-35,69	40,26
		$\hat{N}_2$	3,11	66,43	-5,08	41,96	0,30	28,79
		$\hat{N}_3$	-32,07	36,79	-31,01	35,28	-24,04	28,88
		$\hat{N}_4$	-22,74	47,62	-26,21	37,57	-17,06	30,38
	2 (0,522)	$\hat{N}_1$	-60,07	62,91	-61,31	64,06	-60,41	63,28
		$\hat{N}_2$	-6,12	66,59	-1,15	46,68	1,67	30,99
		$\hat{N}_3$	-55,36	58,35	-51,21	54,19	-40,89	43,99
		$\hat{N}_4$	-41,39	63,79	-34,79	55,45	-18,60	41,35
0,9	0,5 (0,806)	$\hat{N}_1$	5,37	6,79	5,27	6,63	5,59	6,97
		$\hat{N}_2$	0,08	14,78	-0,06	10,17	-0,06	6,55
		$\hat{N}_3$	5,04	6,44	4,62	5,93	4,24	5,53
		$\hat{N}_4$	5,94	9,48	5,03	7,05	4,34	5,72
	1 (0,810)	$\hat{N}_1$	0,30	5,01	0,17	5,01	0,25	4,94
		$\hat{N}_2$	0,78	20,72	0,41	14,06	-0,06	9,03
		$\hat{N}_3$	0,33	4,83	0,20	4,68	0,17	4,24
		$\hat{N}_4$	3,23	13,79	1,88	9,35	1,00	5,98
	1,5 (0,814)	$\hat{N}_1$	-4,29	7,07	-4,39	7,32	-4,55	7,37
		$\hat{N}_2$	-0,65	21,52	0,35	15,88	0,02	10,27
		$\hat{N}_3$	-4,07	6,78	-3,83	6,73	-3,49	6,15
		$\hat{N}_4$	-0,43	13,77	-1,18	10,92	-1,43	8,20
	2 (0,817)	$\hat{N}_1$	-8,28	10,27	-8,40	10,36	-8,33	10,32
		$\hat{N}_2$	-0,29	25,59	0,39	17,66	0,35	11,41
		$\hat{N}_3$	-7,80	9,82	-7,35	9,38	-6,30	8,20
		$\hat{N}_4$	-2,52	17,96	-3,10	14,02	-2,73	10,33

Tableau 3
Résultats des simulations pour $N = 5000$

			p_A					
			0,5		0,10		0,20	
p_B	θ		% Biais relatif	% PEQMR	% Biais relatif	% PEQMR	% Biais relatif	% PEQMR
0,7	0,5 (0,462)	$\hat{N}_1$	61,47	61,82	61,39	61,76	61,69	62,04
		$\hat{N}_2$	-0,18	15,78	0,26	10,65	-0,15	6,72
		$\hat{N}_3$	54,84	55,17	49,06	49,38	39,38	39,65
		$\hat{N}_4$	19,73	38,12	4,77	19,52	-0,01	7,21
	1 (0,490)	$\hat{N}_1$	-0,28	6,14	-0,13	5,99	0,35	6,15
		$\hat{N}_2$	0,43	18,14	0,47	12,85	-0,20	8,34
		$\hat{N}_3$	-0,22	5,82	-0,03	5,35	0,16	4,88
		$\hat{N}_4$	0,26	9,82	-0,04	7,44	0,11	5,95
	1,5 (0,508)	$\hat{N}_1$	-36,21	36,68	-36,29	36,78	-35,90	36,38
		$\hat{N}_2$	0,41	20,39	-0,16	14,21	0,39	9,55
		$\hat{N}_3$	-32,87	33,37	-29,97	30,49	-24,13	24,66
		$\hat{N}_4$	-19,11	31,15	-11,51	23,92	-3,12	14,03
	2 (0,522)	$\hat{N}_1$	-61,04	61,30	-60,53	60,81	-60,64	60,92
		$\hat{N}_2$	0,40	20,09	0,60	15,43	0,31	9,67
		$\hat{N}_3$	-55,69	55,96	-50,24	50,55	-41,46	41,76
		$\hat{N}_4$	-14,10	36,31	-2,34	20,96	0,26	9,84
0,9	0,5 (0,806)	$\hat{N}_1$	5,56	5,70	5,52	5,67	5,54	5,68
		$\hat{N}_2$	-0,12	4,55	0,11	3,19	-0,03	2,08
		$\hat{N}_3$	5,21	5,35	4,86	5,01	4,22	4,35
		$\hat{N}_4$	4,97	5,41	3,64	4,88	2,26	3,79
	1 (0,810)	$\hat{N}_1$	-0,02	1,58	0,08	1,55	0,01	1,57
		$\hat{N}_2$	-0,09	6,16	-0,17	4,08	-0,14	2,79
		$\hat{N}_3$	-0,03	1,53	0,05	1,48	-0,02	1,35
		$\hat{N}_4$	0,37	3,19	0,11	2,18	0,09	1,89
	1,5 (0,814)	$\hat{N}_1$	-4,66	5,00	-4,52	4,85	-4,61	4,90
		$\hat{N}_2$	-0,25	7,54	0,11	4,95	-0,09	3,14
		$\hat{N}_3$	-4,39	4,73	-3,96	4,32	-3,55	3,85
		$\hat{N}_4$	-2,50	6,31	-2,26	5,02	-1,84	3,82
	2 (0,817)	$\hat{N}_1$	-8,45	8,68	-8,38	8,60	-8,46	8,69
		$\hat{N}_2$	-0,21	7,86	-0,06	5,29	0,01	3,73
		$\hat{N}_3$	-7,95	8,18	-7,39	7,61	-6,49	6,73
		$\hat{N}_4$	-3,76	8,80	-2,77	6,99	-1,25	4,97

BIBLIOGRAPHIE

ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.

ALHO, J.M., MULRY, M.H., WURDEMAN, K., et KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.

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.

CORMACK, R.M. (1989). Log-linear models for capture-recapture. *Biometrics*, 45, 395-413.

DARROCH, J.N. (1958). The multiple-recapture census I: estimation of a closed population. *Biometrika*, 45, 343-359.

DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., et JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88, 1137-1148.

FAULKENBERRY, G.D., et GAROUI, A. (1991). Estimating a population total using an area frame. *Journal of the American Statistical Association*, 86, 445-449.

FECSO, R., TORTORA, R.D., et VOGEL, F.A. (1986). Sampling frames for agriculture in the United States. *Journal of Official Statistics*, 2, 279-292.

FIENBERG, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59, 591-603.

HAINES, D.E. (1997). Estimating Population Parameters Using Multiple Frame and Capture-Recapture Methodology. Thèse de doctorat, North Carolina State University.

HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons.

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

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.

HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.

KOTT, P.S., et VOGEL, F.A. (1995). Multiple-frame business surveys. *Business Survey Methods* (Éd., B.G. Cox). New York: John Wiley & Sons, 185-203.

LUND, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.

NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators. Staff Report 80, U.S. Department of Agriculture, Statistical Reporting Service, Washington, DC.

OTIS, D.L., BURNHAM, K.P., WHITE, G.C., et ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-135.

POLLOCK, K.H., HINES, J.E., et NICHOLS, J.D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40, 329-340.

POLLOCK, K.H., TURNER, S.C., et BROWN, C.A. (1994). Techniques de saisie-ressaisie pour l'estimation de la taille de la population et de totaux de population lorsqu'on ne dispose pas d'une base de sondage complète. *Techniques d'enquête*, 20, 121-128.

SANATHANAN, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 43, 1, 142-152.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*, (2-ième Édition). New York: Macmillan.

SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

SIRKEN, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

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.

WOLTER, K.M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.