Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 55–64
Statistique Canada

Estimation de la population par échantillonnage multiple et sous-dénombrement lors du recensement en présence d'erreurs d'appariement

YE DING et STEPHEN E. FIENBERG[1]

RÉSUMÉ

Les auteurs examinent le recensement par saisie-resaisie multiples en assouplissant l'hypothèse classique d'un appariement parfait. Ils proposent des modèles avec erreur d'appariement permettant de caractériser les méthodes d'appariement sujettes à des erreurs. Les données observées prennent la forme d'un tableau de contingence 2^k auquel manque une cellule et suivent une distribution multinomiale. Les auteurs proposent une méthode pour estimer la population. Cette approche s'applique à la fois aux modèles log-linéaires habituels pour les tableaux de contingence et aux modèles log-linéaires de l'hétérogénéité du potentiel de saisie. Enfin, les auteurs illustrent leur méthode et procèdent à une estimation en recourant à une répétition générale du recensement de 1990, effectuée en 1988 par le U.S. Bureau of the Census.

MOTS CLÉS: Recensement par saisie-resaisie; estimations de la population totale; modèles log-linéaires; erreurs d'appariement; recensement par resaisie multiple.

1. INTRODUCTION

La technique du recensement par resaisie multiple a été appliquée dans de nombreux domaines pour estimer l'importance d'une population fermée. Ainsi, Cormack (1968) et Seber (1982) offrent une excellente revue de nombreuses techniques utilisées. Nous envisagerons ici une séquence d'échantillons, $s_1, \ldots, s_k$, pour laquelle les sujets du i-ième échantillon sont identifiés de façon unique, par exemple par étiquetage ou marquage, avant d'être remis dans la population (Darroch 1958). Les méthodes de recensement par resaisie multiple habituelles reposent sur les hypothèses que voici:

(1) **Appariement idéal**. Les sujets d'une liste (source d'information, échantillon) peuvent être couplés à ceux d'une autre liste sans erreur. En d'autres termes, aucune erreur de classification ne survient lorsqu'on s'efforce de déterminer si un sujet donné a été enregistré par les deux sources d'information ou par une seule d'entre elles.

(2) **Indépendance**. Les listes sont indépendantes l'une de l'autre, c'est-à-dire que la probabilité qu'un sujet fasse partie d'une liste ne dépend pas de l'inclusion du sujet aux listes antérieures.

(3) **Homogénéité (potentiel de saisie identique)**. Tous les sujets de la population à l'étude ont la même probabilité d'être observés (saisis) dans une liste (échantillon) quelconque.

(4) **Finitude**. La population est «fermée». Elle ne subit aucun changement consécutivement aux naissances, aux décès, à l'émigration ou à l'immigration au cours de la période durant laquelle s'effectue l'échantillonnage.

Darroch (1958) a examiné le recensement par resaisie multiple sous l'angle de ces quatre hypothèses. De son côté, Fienberg (1972) s'est servi d'un modèle log-linéaire pour autoriser une dépendance statistique entre des types spécifiques de l'échantillon. Il a donc abandonné l'hypothèse d'indépendance. Darroch, Fienberg, Glonek et Junker (1993) ont pour leur part mis au point un modèle log-linéaire élargi qui autorise l'hétérogénéité au niveau individuel et la dépendance, mais exige au moins trois échantillons, à savoir $k = 3$. Plusieurs auteurs se sont penchés sur les problèmes de couplage attribuables aux erreurs d'appariement et de non-appariement dans le contexte du recensement à deux échantillons utilisé par le U.S. Bureau of the Census pour évaluer la couverture du recensement. Ding et Fienberg (1994), par exemple, ont envisagé la modélisation des erreurs d'appariement dans le recensement à deux échantillons et mis au point une méthode systématique pour estimer la population totale. L'inclusion d'un troisième échantillon (tiré des dossiers administratifs) au modèle et l'estimation de la couverture du recensement avaient déjà été envisagées par le Bureau of the Census dans le passé. On n'a pas écarté cette possibilité pour étendre et évaluer l'approche du système dual. Dans le présent article, nous examinerons les modèles d'erreur d'appariement relatifs au problème du recensement à échantillon multiple tolérant les principes de dépendance et d'hétérogénéité.

Nous supposons ici que les observations tirées des données d'un recensement par saisie multiple font partie d'une classification croisée 2^k, avec ou sans le i-ième échantillon définissant la catégorie de la i-ième dimension. En vertu d'un telle classification, la cellule qui correspond

[1] Ye Ding, Research Scientist, Bureau of Biometrics, New York State Health Department, Concourse, Room C-144, Empire State Plaza, Albany, New York 12237, U.S.A.; Stephen E. Fienberg, Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

à l'absence des k échantillons n'existe pas. L'objectif consiste à estimer le nombre de sujets de la population qui ne sont pas observés, donc qui correspondent au chiffre de la cellule manquante dans le tableau de contingence 2^k. À la partie 2, nous examinerons les effets des erreurs d'appariement sur le tableau 2^k incomplet observé. La partie 3 propose quelques modèles pour les erreurs d'appariement en mesure de caractériser une méthode d'appariement sujette aux erreurs. À partir de ces modèles et des hypothèses (3) et (4), nous élaborerons une procédure pour estimer la taille de la population à partir d'un modèle log-linéaire. Enfin, à la partie 5, nous nous servirons des méthodes envisagées pour analyser les données de la répétition générale de 1988 du recensement entreprise par le U.S. Bureau of the Census.

2. LES ERREURS D'APPARIEMENT DANS LES RECENSEMENTS À ÉCHANTILLON MULTIPLE

Nous commencerons par répartir les erreurs d'appariement en deux grandes catégories: les faux appariements et les non-appariements erronés. Pour bien comprendre la nature de ces erreurs dans un recensement à échantillon multiple, prenons le cas d'un recensement à trois échantillons. Supposons qu'aucune donnée ne manque ou qu'il n'y ait aucune erreur dans la saisie de l'information sur un sujet de la population et qu'on prélève trois échantillons de la population, s_1, s_2 et s_3. Supposons, par exemple, qu'on retrouve les sujets 1, 3, 4 et 7 dans l'échantillon s_1, les sujets 3, 4 et 8 dans l'échantillon s_2 et les sujets 4, 9 et 10 dans l'échantillon s_3. Sous sa forme vectorielle, cette situation s'exprime comme suit: $s_1 = (1, 3, 4, 7)$, $s_2 = (3, 4, 8)$ et $s_3 = (4, 9, 10)$. Il n'y a pas d'erreurs d'appariement, pourvu que les renseignements soient complets et exacts. On obtient le tableau incomplet 2^3 qui suit au moyen des trois échantillons:

Tableau 1

Tableau original sans erreur d'appariement

	s_1			
	Présent		Absent	
	s_2		s_2	
s_3	Présent	Absent	Présent	Absent
Présent	1	0	0	2
Absent	1	2	1	-

Supposons par ailleurs qu'en raison de données manquantes ou de renseignements incorrects, on observe en réalité ce qui suit:

$$s_1 = (1, 3, 4, 7), \quad s_2 = (3^*, 4^*, 8), \quad s_3 = (4, 9, 10),$$

où 3* et 4* représentent les sujets 3 et 4, mais associés à des renseignements incorrects, ce qui entraîne deux non-appariements erronés lors du couplage des échantillons. Partant de l'hypothèse que tous les appariements sont exacts, on obtient le tableau 2^3 incomplet que voici:

Tableau 2

Observations avec erreurs d'appariement

	s_1			
	Présent		Absent	
	s_2		s_2	
s_3	Présent	Absent	Présent	Absent
Présent	0	1	0	2
Absent	0	3	3	-

Lorsqu'on compare les deux tableaux, il est facile de voir l'effet des erreurs d'appariement:

(i) Le nombre d'observations peut augmenter dans certaines cellules et diminuer dans d'autres, si bien que les totaux marginaux, en particulier le nombre total de sujets différents observés dans les trois échantillons, peuvent varier, sous réserve de la contrainte que le nombre total d'observations dans les échantillons, x_{1++}, x_{+1+}, et x_{++1}, reste le même. Une modification du nombre total de sujets différents dans les échantillons soulèverait un problème distinct du problème habituel posé par une erreur de classification dans l'analyse des données catégoriques, où on examine la possibilité qu'une erreur se glisse dans la répartition des sujets entre les différentes catégories (p. ex., lire Chen 1979).

(ii) Parallèlement, la probabilité dans certaines cellules peut changer pourvu que la probabilité p_{1++}, p_{+1+}, et p_{1++}, d'être saisie dans un échantillon demeure la même.

En raison de la complexité des erreurs d'appariement du cas à trois échantillons, nous avons besoin d'une terminologie particulière pour faciliter la description. Par rapport à l'échantillon s_1, nous dirons donc qu'un sujet se trouve à l'état 1 si on l'observe dans l'échantillon et à l'état 0 dans le cas contraire. Le triplet (i,j,k), $0 \leq i, j, k \leq 1$ indiquera un sujet à l'état i, j et k, pour les échantillons s_1, s_2 et s_3, respectivement. Ainsi, (1,0,0) dénotera un sujet qu'on observe uniquement dans s_1 et (1,1,1), un sujet saisi dans les trois échantillons. Le niveau d'un sujet (i,j,k) sera $i + j + k$, soit le nombre d'échantillons dans lequel se retrouve le sujet. Il existe donc quatre niveaux différents, 0, 1, 2 et 3. Un sujet n'aura le niveau 0 que s'il n'apparaît dans aucun échantillon, alors qu'il sera de niveau 3 quand il est présent dans les trois échantillons. Si le couplage est incorrect selon la règle d'appariement, le sujet (1,1,0) se décomposera en «deux sujets différents», (1,0,0) et (0,1,0), en supposant qu'il n'y ait pas de faux appariement. Par ailleurs, le sujet (1,0,0) apparié par

erreur au sujet (0,1,0) donnera une seule observation (1,1,0). Pour plus de commodité, nous appellerons une telle décomposition ou combinaison une *transition*. Les transitions ne peuvent aller que du niveau 3 ou du niveau 2 au même niveau (s'il n'y a pas d'erreur d'appariement) ou à un niveau inférieur, s'il n'y a pas de faux appariements. Plus précisément, un sujet (1,1,1) peut faire la transition vers l'un des 5 jeux possibles de sujets que voici:

$$\{(1,1,1)\},\quad \{(1,0,0),(0,1,1)\},\quad \{(0,1,0),(1,0,1)\}$$

$$\{(0,0,1),(1,1,0)\},\quad \{(1,0,0),(0,1,0),(0,0,1)\}.$$

Au niveau 2, un sujet (1,1,0) pourra se décomposer en $\{(1,0,0),(0,1,0)\}$ ou demeurer $\{(1,1,0)\}$. Il en ira autant des sujets $\{(0,1,1)\}$ et $\{(1,0,1)\}$. De ce qui précède, on peut résumer l'effet d'une erreur d'appariement au moyen du diagramme suivant:

$$\boxed{\text{Tableau 1}} \rightarrow \{\text{Processus d'appariement}\} \rightarrow \boxed{\text{Tableau 2}}$$

où le tableau 1 représente le tableau 2^k incomplet original, sans erreur d'appariement, et le tableau 2 correspond au tableau 2^k incomplet observé, avec erreur d'appariement. Par conséquent, la probabilité par cellule et le chiffre prévu d'une cellule du tableau 1 seront notés par $\{r_{ijk}\}$ et $\{l_{ijk}\}$, et par $\{p_{ijk}\}$ et $\{m_{ijk}\}$ pour le tableau 2, où $1 \leq i,j,k \leq 2$.

3. QUELQUES MODÈLES POUR LES ERREURS D'APPARIEMENT

Nous présenterons maintenant des modèles qui décrivent les erreurs d'appariement. Chacun d'eux nous permet de formuler la réattribution de la probabilité et du chiffre prévu des cellules associés au tableau 1.

Modèle (1). Outre les hypothèses d'homogénéité et de finitude expliquées en §1, nous supposons: (i) que le couplage n'entraîne pas de faux appariements; (ii) que la probabilité qu'un sujet garde son état original est égale à θ et que la probabilité d'effectuer une transition vers un ensemble possible de sujets est égale à $(1-\theta)/(m-1)$, où m représente le nombre de jeux possibles de sujets pour lesquels la transition est autorisée. Pour le sujet (1,1,1) examiné à la dernière partie, par exemple, $m = 5$.

Un tel modèle permet d'exprimer les probabilités du tableau avec erreur d'appariement, $\{p_{ijk}\}$, d'après les probabilités du tableau sans erreur d'appariement, $\{r_{ijk}\}$, pour le recensement à trois échantillons de la façon suivante:

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1-\theta}{4} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1-\theta}{4} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1-\theta}{4} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{121} + r_{122},$$

$$p_{212} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{211} + r_{212},$$

$$p_{221} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{211} + (1-\theta) r_{121} + r_{221}.$$

Soit

$$\vec{p} = (p_{111}, p_{112}, p_{121}, p_{211}, p_{122}, p_{212}, p_{221})^T,$$

et

$$\vec{r} = (r_{111}, r_{112}, r_{121}, r_{211}, r_{122}, r_{212}, r_{221})^T,$$

donc

$$\vec{p} = M_1 \times \vec{r}. \tag{1}$$

Dans ce cas, M_1 représente une matrice 7×7 définie par les sept équations qui précèdent, issues du modèle (1). Il est assez facile de vérifier que la probabilité de saisir un sujet dans chaque échantillon est fixe, à savoir $p_{1++} = r_{1++} = p_1, p_{+1+} = r_{+1+} = p_2, p_{++1} = r_{++1} = p_3$. Il doit en être ainsi puisque la probabilité de saisie dans l'échantillon ne dépend pas de la façon dont fonctionne la méthode d'appariement.

On peut aisément généraliser la formule pour k échantillons. Les opérations algébriques deviennent néanmoins fort fastidieuses quand k est élevé. Il est possible de simplifier le modèle en exigeant que les transitions descendent d'au moins un niveau, ce qui nous amène au modèle (2).

Modèle (2). En plus des hypothèses d'homogénéité et de finitude données en §1, on suppose: (i) que le couplage n'entraîne pas de faux appariements; (ii) qu'une transition ne peut se faire que vers le bas, d'au moins un niveau; (iii) que la probabilité qu'un sujet garde son état original est θ et que la probabilité d'une transition vers un ensemble possible de sujets est $(1-\theta)/(m'-1)$, où m' représente le nombre de combinaisons de sujets vers lesquels la transition est possible et autorisée.

Examinons d'abord le cas à trois échantillons. Un sujet (1,1,1) peut se décomposer en trois individus, soit $(1,1,1) \mapsto \{(1,0,0),(0,1,0),(0,0,1)\}$ (le signe « $\mapsto$ » indique qu'il y a décomposition), si les trois appariements présumés ne se concrétisent pas. L'hypothèse (ii) du modèle (2) établit que cette triple erreur a une probabilité négligeable de se reproduire, comparativement à la probabilité d'une transition où un seul appariement n'est pas effectué, si bien que $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$, ou $(1,1,1) \mapsto \{(1,0,1),(0,1,0)\}$, ou $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$.

Le modèle paramétrique qui exprime $\{p_{ijk}\}$ en termes de $\{r_{ijk}\}$ pour le cas à trois échantillons est le suivant:

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1-\theta}{3} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1-\theta}{3} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1-\theta}{3} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{121} + r_{122},$$

$$p_{212} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{211} + r_{212},$$

$$p_{221} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{211} + (1-\theta) r_{121} + r_{221}.$$

Par conséquent

$$\vec{p} = M_2 \times \vec{r}, \tag{2}$$

où M_2 est une matrice 7×7 définie par les sept équations qui précèdent, dérivées du modèle (2). Encore une fois, la probabilité de saisie reste la même, c.-à-d. $p_{1++} = r_{1++} = p_1$, $p_{+1+} = r_{+1+} = p_2$, $p_{++1} = r_{++1} = p_3$.

Pour le problème à k échantillons, soit $p_{\bar{1}}$, la probabilité d'être saisi dans chaque échantillon, à savoir $p_{\bar{1}} = p_{111\ldots1}$, et soit $p_{\bar{1},\bar{2}(h_1,h_2)}$, la probabilité par cellule correspondant à une absence dans le h_1-ième et le h_2-ième échantillons, et d'une présence dans les autres échantillons, *etc.* En vertu du modèle (2), $p_{\bar{1}} = \theta r_{\bar{1}}$. Lorsque $i \leq k-2$, la probabilité de ne pas apparaître dans les h_1-ième, h_2-ième, ..., et h_i-ième échantillons et saisis dans les autres échantillons est égale à

$$p_{\bar{1},\bar{2}(h_1,h_2,\ldots,h_i)} = \theta r_{\bar{1},\bar{2}(h_1,h_2,\ldots,h_i)} + \frac{1-\theta}{k-i+1} \sum_{j=1}^{i} r_{\bar{1},\bar{2}(\{h_1,h_2,\ldots,h_i\}\setminus h_j)}.$$

Quand $i = k-1$, le sujet ne fait partie que d'un échantillon. Par exemple, la probabilité de n'être saisi que dans le premier échantillon est égale à

$$p_{1,\bar{2}} = r_{1,\bar{2}} + (1-\theta) \sum_{h \neq 1} r_{1,1(h),\bar{2}} + \frac{(1-\theta)}{3} \sum_{h_1,h_2 \geq 2} r_{1,1(h_1,h_2),\bar{2}} + \sum_{j=3}^{k-1} \sum_{h_1,h_2,\ldots,h_j \geq 2} \frac{(1-\theta)}{(j+1)} r_{1,1(h_1,h_2,\ldots,h_j),\bar{2}},$$

où $r_{1,1(h_1,h_2,\ldots,h_j),\bar{2}}$ représente la probabilité par cellule du tableau original, correspondant à une présence dans le premier h_1-ième, h_2-ième, ..., h_j-ième échantillon et à une absence dans les autres échantillons. Pour des raisons de symétrie, on peut écrire l'expression pour $p_{1(h),\bar{2}}$, c.-à-d. la probabilité d'être observé dans le h-ième échantillon uniquement et pas dans les autres.

Il est possible d'améliorer le modèle (2) en supposant des taux d'appariement inégaux. Examinons, par exemple, deux décompositions: $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$ et $(1,1,0) \mapsto \{(0,1,0),(1,0,0)\}$. Dans les deux cas, le point commun est qu'un appariement présumé ne s'est pas réalisé. La différence réside dans le fait que, dans le premier cas, il y a couplage de deux sources d'information, comparativement à une seule dans le second. On peut raisonnablement supposer que la probabilité d'une erreur d'appariement diffère dans les deux cas, contrairement à ce que suppose le modèle (2). Ceci nous amène donc au:

Modèle (3). En plus des hypothèses (i) et (iii) du modèle (2), on suppose

$$(1,1,1) \mapsto \begin{cases} (1,1,1) & \text{avec une probabilité } \alpha_1 \\ \{(1,1,0),(0,0,1)\} & \text{avec une probabilité } (1-\alpha_1)/3 \\ \{(0,1,1),(1,0,0)\} & \text{avec une probabilité } (1-\alpha_1)/3 \\ \{(1,0,1),(0,1,0)\} & \text{avec une probabilité } (1-\alpha_1)/3 \end{cases}$$

$$(1,1,0) \mapsto \begin{cases} (1,1,0) & \text{avec une probabilité } \alpha_2 \\ \{(0,1,0),(1,0,0)\} & \text{avec une probabilité } 1-\alpha_2 \end{cases}$$

$$(1,0,1) \mapsto \begin{cases} (1,0,1) & \text{avec une probabilité } \alpha_2 \\ \{(1,0,0),(0,0,1)\} & \text{avec une probabilité } 1-\alpha_2 \end{cases}$$

$$(0,1,1) \mapsto \begin{cases} (0,1,1) & \text{avec une probabilité } \alpha_2 \\ \{(0,1,0),(0,0,1)\} & \text{avec une probabilité } 1-\alpha_2 \end{cases}$$

et (1,0,0), (0,1,0) et (0,0,1) ne varient pas avec une probabilité égale à un.

Selon ce modèle, on peut exprimer la probabilité par cellule $\{p_{ijk}\}$ du tableau 2 en fonction de α_1, α_2 et de la probabilité par cellule du tableau 1, $\{r_{ijk}\}$. Pour cela, il faut tenir compte de toutes les transitions possibles qui donneront un sujet dans la cellule (i,j,k) du tableau 2. Par exemple, prenons le sujet observé (1,0,0). Ce sujet se retrouve dans la cellule (1,2,2) du tableau 2. Soit F, l'observation d'un sujet de statut (1,0,0) et soit E_{ijk}, l'inclusion d'un sujet dans la cellule (i,j,k) du tableau 1. Alors,

$$F = \bigcup_{\{i,j,k\}} (E_{ijk} \cap F).$$

Selon le modèle (3), il n'existe que quatre transitions en vertu desquelles F se réalise, en l'occurrence:

$$(1,1,1) \mapsto \{(1,0,0),(0,1,1)\},$$

$$(1,1,0) \mapsto \{(1,0,0),(0,1,0)\},$$

$$(1,0,1) \mapsto \{(1,0,0),(0,0,1)\},$$

$$(1,0,0) \mapsto \{(1,0,0)\}.$$

Par conséquent,

$$F = (E_{111} \cap F) \cup (E_{112} \cap F) \cup (E_{121} \cap F) \cup (E_{122} \cap F).$$

Étant donné la définition de la probabilité par cellule des deux tableaux, $p(F) = p_{122}$ et $p(E_{ijk}) = r_{ijk}$. En vertu des hypothèses du modèle (3), $p(F \mid E_{111}) = (1 - \alpha_1)/3$, $p(F \mid E_{112}) = p(F \mid E_{121}) = \alpha_2$ et $p(F \mid E_{122}) = 1$.

Puisque $E_{111} \cap F$, $E_{112} \cap F$, $E_{121} \cap F$ et $E_{122} \cap F$ sont quatre possibilités qui s'excluent mutuellement pour que F se réalise,

$$\begin{aligned} p_{122} &= p(E_{111} \cap F) + p(E_{112} \cap F) \\ &\quad + p(E_{121} \cap F) + p(E_{122} \cap F) \\ &= p(F \mid E_{111}) \cdot p(E_{111}) + p(F \mid E_{112}) \cdot p(E_{112}) \\ &\quad + p(F \mid E_{121}) \cdot p(E_{121}) + p(F \mid E_{122}) \cdot p(E_{122}) \\ &= \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122}. \end{aligned}$$

On peut dériver les autres probabilités par cellule du tableau 2 de la même manière, si bien qu'on obtient

$$p_{111} = \alpha_1 r_{111},$$

$$p_{112} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{112},$$

$$p_{121} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{121},$$

$$p_{211} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{211},$$

$$p_{122} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122},$$

$$p_{212} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{211} + r_{212},$$

$$p_{221} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{211} + (1 - \alpha_2) r_{121} + r_{221}.$$

Donc

$$\vec{p} = M_3 \times \vec{r}, \tag{3}$$

où M_3 représente une matrice 7×7 définie par les sept équations qui précèdent, issues du modèle (3).

Lorsque $\alpha_1 = \alpha_2 = \theta$, on obtient la même formule qu'avec le modèle (2). Dans le cas particulier où $\alpha_1 = \alpha_2 = 1$, $p_{ijk} = r_{ijk}$, ce qui nous ramène au problème classique. Encore une fois, les probabilités de saisie demeurent les mêmes, soit $p_{1++} = r_{1++}$, $p_{+1+} = r_{+1+}$, $p_{++1} = r_{++1}$.

4. ESTIMATION DE LA TAILLE DE LA POPULATION

4.1 Formulation du modèle log-linéaire

Aux fins d'exposition, nous nous limiterons au cas du recensement à trois échantillons, bien que l'extension à k échantillons où $k > 3$ se fasse simplement. Comme nous l'avons fait précédemment, soient l_{ijk} et m_{ijk}, le chiffre par cellule prévu pour les tableaux 1 et 2, respectivement. La relation entre la probabilité par cellule et le chiffre prévu par cellule est $l_{ijk} = r_{ijk}N$ et $m_{ijk} = p_{ijk}N$. Soient

$$\vec{m} = (m_{111}, m_{112}, m_{121}, m_{211}, m_{122}, m_{212}, m_{221})^T,$$

et

$$\vec{l} = (l_{111}, l_{112}, l_{121}, l_{211}, l_{122}, l_{212}, l_{221})^T.$$

Puisque, pour chacun des modèles proposés à la dernière partie, il existe une matrice M dont les inscriptions dépendent des paramètres de la probabilité d'appariement pour le modèle retenu, ce qui $\vec{p} = M \times \vec{r}$, multiplié par N donne

$$\vec{m} = M \times \vec{l}. \tag{4}$$

La paramétrisation de m_{ijk} est simple avec tous les modèles log-linéaires indiqués pour le tableau 1. Ainsi, pour chacun des modèles suggérés par Fienberg (1972), on peut exprimer le chiffre prévu au moyen des fonctions des paramètres du terme u:

$$l_{ijk} = g_{ijk}(u, u_1(i), u_2(j), u_3(k), u_{12}(ij), u_{13}(ik), u_{23}(jk)), \tag{5}$$

On peut ensuite dériver les paramètres de $\{m_{ijk}, (ijk) \neq (222)\}$ à partir de (4).

4.2 Estimation de la taille de la population

Examinons maintenant le taux d'appariement des divers modèles. Pour estimer la population, on procède de la façon suivante. Tout d'abord, à l'instar de Sanathanan (1972), on estime le maximum de vraisemblance des paramètres du terme u au moyen de l_c, la vraisemblance conditionnelle associée au tableau 2 étant donné n,

$$l_c = n! \prod_{\{(ijk) \neq (222)\}} \frac{(q_{ijk})^{x_{ijk}}}{x_{ijk}!},$$

où $n = \sum_{\{(ijk) \neq (222)\}} x_{ijk}$ et $q_{ijk} = m_{ijk}/n$. Sanathanan (1972) montre que dans des conditions de régularité convenables, les estimations conditionnelles et inconditionnelles du maximum de vraisemblance sont à la fois cohérentes et présentent la même distribution normale asymptotique. Quand on supprime les paramètres redondants du terme u au moyen des contraintes liées au modèle log-linéaire du tableau 1, la difficulté consiste à établir le maximum de l_c, sous réserve de la seule contrainte que voici:

$$\sum_{\{(ijk) \neq (222)\}} m_{ijk} = n.$$

Sur le plan numérique, il s'agit d'un problème d'optimisation sans limite linéaire. Rao (1957) a examiné les conditions de régularité qui débouchent sur une estimation unique du maximum de vraisemblance pour les paramètres, avec une distribution multinomiale. Les conditions qu'il a établies sont satisfaites par la paramétrisation de $\{q_{ijk}\}$. Une fois que les estimations conditionnelles du maximum de vraisemblance des paramètres du terme u sont connues, on se sert du modèle log-linéaire établi pour le tableau 1 afin d'obtenir les estimations conditionnelles du maximum de vraisemblance pour $\{l_{ijk}\}$, soit le chiffre prévu par cellule du tableau 1, y compris le chiffre prévu pour la cellule manquante. L'estimation de N est donc

$$\hat{N} = \sum_{\{ijk\}} \hat{l}_{ijk}.$$

S'il n'y a pas d'erreur d'appariement, puisque $\alpha_1 = \alpha_2 = 1$ au modèle (3), $m_{ijk} = l_{ijk}$. Par conséquent,

$$\hat{N} = n + \hat{m}_{222},$$

ce qui revient à la méthode d'estimation du problème classique du recensement à resaisie multiple énoncé par Fienberg (1972) avec les modèles log-linéaires proposés par cet auteur.

Comme on l'a indiqué précédemment, on a spécifié un modèle log-linéaire pour le tableau 1 et on considère que les observations s'inscrivent au tableau 2, dont le modèle paramétrique pour le chiffre prévu par cellule est défini par le modèle log-linéaire et par le modèle retenu pour les erreurs d'appariement. Pour savoir si un modèle log-linéaire spécifié pour le tableau 1 convient, on peut appliquer les tests habituels d'adéquation de Pearson et du ratio de vraisemblance, X^2 et G^2, décrits dans Fienberg (1972), au tableau 2. Chaque statistique se caractérise par une distribution χ^2 asymptotique selon l'hypothèse nulle que le modèle convient avec $2^k - 1 -$ (nombre de paramètres indépendants du modèle) degrés de libertés.

5. ANALYSE DES DONNÉES DE LA RÉPÉTITION GÉNÉRALE DU RECENSEMENT DE 1988 ST. LOUIS

Le U.S. Bureau of the Census se sert de l'estimation à système dual (ESD) articulée sur le recensement type à deux échantillons pour évaluer la couverture du recensement depuis 1950. En 1988, le Bureau a procédé à une répétition générale en prévision du recensement décennal de 1990 à trois endroits: St. Louis (Missouri), Columbia (Missouri) et l'ouest de l'État de Washington. Zaslavsky et Wolfgang (1993) présentent les données d'un sous-groupe de la population extrait de l'enquête postcensitaire (EPC) de la répétition générale de St. Louis. Ce sous-groupe est essentiellement composé d'adultes de race noire et de sexe masculin vivant en milieu urbain, dont on croit la population sous-estimée par les méthodes à système dual. Les données résultantes viennent de trois sources: l'échantillon C est celui du recensement proprement dit; l'échantillon P dérive de l'EPC et on a recouru à une troisième source d'information, en l'occurrence la liste administrative complémentaire (LAC) tirée des dossiers administratifs de divers services d'État et services fédéraux (y compris sécurité de l'emploi, permis de conduire, impôt, service sélectif et administration des anciens combattants) avant le recensement. L'échantillon C et l'échantillon P procurent les données nécessaires à l'ESD habituelle ou à l'approche par saisie et resaisie. On peut combiner les données de la LAC à celles du recensement et de l'échantillon P pour effectuer une analyse sous l'angle d'un triple échantillon, même si le but original de l'exercice était d'améliorer la couverture de l'échantillon P. Le tableau 3 présente les données à trois échantillons relatives à la strate d'échantillonnage 11 de l'EPC pour St. Louis, obtenues par groupement des données originales du tableau 1 de Zaslavsky et Wolfgang (1993) pour quatre strates a posteriori définies par les paramètres propriétaire/locataire × 20-29, 30-44 ans.

Tableau 3

Données à trois échantillons pour la strate 11 de St. Louis

LAC	Recensement			
	Présent échantillon P		Absent échantillon P	
	Présent	Absent	Présent	Absent
Present	300	51	53	180
Absent	187	166	76	-

On peut analyser les données d'un tel système triple au moyen du modèle (2) des erreurs d'appariement et des données d'une étude sur l'erreur d'appariement (EEP ou étude de réappariement) distincte, associées aux mêmes strates échantillonnées a posteriori. L'EEP est l'une des opérations effectuées par le Bureau of the Census pour évaluer l'EPC. On l'applique typiquement à un échantillon de cas en recourant à des procédures plus développées et à du personnel hautement qualifié, et en procédant à de nouvelles interviews afin d'estimer le biais associé au processus de couplage antérieur. Dans leur analyse de l'étude sur l'erreur d'appariement entreprise lors d'un recensement expérimental, en 1986, à Los Angeles, Hogan et Wolter (1988) signalent que le réappariement avait été effectué indépendamment de l'appariement original et qu'on avait établi l'écart entre les résultats de l'appariement et du réappariement. À cause de l'approche intensive utilisée, les auteurs estiment que les résultats du réappariement illustrent un couplage véritable, et que les différences entre les données des deux exercices représentent le biais qui fausse les résultats de l'appariement original. Par conséquent, les données de l'EEP permettent d'estimer le taux d'erreur lors de l'appariement initial. Mulry, Dajani et Biemer (1989) ont rédigé un rapport sur l'EEA pour la répétition générale de 1988, dans lequel ils fournissent les résultats du réappariement pour les trois sites d'essai. Le tableau 4 reproduit les données du rapport qui nous intéressent.

Tableau 4

Étude de réappariement de St. Louis: échantillon P
Source: Mulry, Dajani et Biemer (1989)

Classification à l'appariement initial	Classification au réappariement			
	Apparié	Non apparié	Indéterminé	Total
Apparié	2,667	7	8	2,682
Non apparié	9	427	30	466
Indéterminé	0	7	20	27
Total	2,676	441	58	3,175

Soit α, le taux d'appariement entre les échantillons C et P et $\gamma = 1 - \alpha$, le taux d'erreur du non-appariement. Nous présumons que le réappariement ne comporte aucune erreur. À partir des données du tableau 4, on peut estimer α avec $\hat{\alpha} = 2667/(2667 + 9) = 99.6637\%$ et γ avec $\hat{\gamma} = 1 - \hat{\alpha} = .3363\%$. Le paramètre θ représente le taux d'appariement à trois échantillons pour l'échantillon C, l'échantillon P et la LAC. Deux appariements sont nécessaires à une classification correcte (1,1,1) des trois échantillons, par exemple le premier entre l'échantillon C et l'échantillon P et le second entre l'échantillon P et la LAC. Puisque l'appariement entre le recensement et la LAC n'a pas été évalué, nous supposons que les deux couplages sont indépendants l'un de l'autre et que le taux d'appariement entre l'échantillon P et la LAC est identique à celui entre l'échantillon C et l'échantillon P. On peut donc utiliser $\theta = \alpha^2$ et $\hat{\theta} = \hat{\alpha}^2 = 99.3285\%$. D'après d'autres renseignements qualitatifs, le taux d'appariement paraît anormalement élevé. En outre, le taux d'erreur de l'appariement entre le recensement et la LAC est sans doute plus élevé que le taux d'erreur de l'appariement entre le recensement et l'échantillon P. Faute de meilleurs renseignements quantitatifs, nous avons néanmoins décidé de les utiliser dans les calculs qui suivent.

Tableau 5

Estimations selon divers modèles

Modèle log-linéaire	EP ordinaire		EP avec modèle à erreur d'appariement (2)	
	$\hat{N}$ (E.-T.)	Ajustement (d.l.)	$\hat{N}$ (E.-T.)	Ajustement (d.l.)
[C] [P] [A]	1091.48 (11.24)	248.31 (3)	1083.58 (10.93)	244.56 (3)
[CP] [A]	1204.14 (23.31)	90.60 (2)	1194.73 (22.86)	87.30 (2)
[PA] [C]	1108.34 (13.77)	247.93 (2)	1100.03 (13.40)	244.53 (2)
[CA] [P]	1068.87 (10.47)	230.66 (2)	1061.09 (10.10)	226.42 (2)
[CP] [CA]	1271.11 (52.55)	87.16 (1)	1256.77 (50.97)	84.37 (1)
[CP] [PA]	1598.88 (106.26)	17.55 (1)	1585.03 (104.93)	15.88 (1)
[CA] [PA]	1080.47 (13.38)	230.43 (1)	1072.19 (12.88)	226.44 (1)
[CP] [CA] [PA]	2360.82 (363.25)	- (0)	2309.55 (352.36)	- (0)

Le tableau 5 donne l'estimation de la taille de la population pour divers modèles log-linéaires, avec l'estimation de l'erreur-type et des statistiques d'adéquation. On a calculé l'erreur-type selon la méthode delta envisagée par Fienberg (1972). L'hypothèse d'indépendance entre le recensement et l'échantillon P a été contestée dans le cadre de l'ESD. La méthode du système dual ne nous permet de vérifier cette hypothèse et d'ajuster la dépendance potentielle que de façon limitée, alors que ces deux aspects peuvent être réglés au moyen de modèles log-linéaires pour trois échantillons ou plus. Quatre modèles énumérés au tableau 5 supposent l'indépendance entre le recensement et l'échantillon P: le modèle de totale indépendance [C] [P] [A], [PA] [C], [CA] [P], et [CA] [PA]. Aucun d'eux ne permet un bon ajustement des données. Les trois modèles avec terme d'interaction entre le recensement et l'échantillon P, [CP] [A], [CP] [CA] et [CP] [PA], assurent un meilleur ajustement. L'addition d'un terme d'interaction reliant le recensement à la LAC n'améliore que légèrement l'ajustement du modèle [CP] [CA] par rapport à [CP] [A], signe que le recensement et l'échantillon P sont presque indépendants de la LAC. Le modèle [CP] [PA] est celui qui procure le meilleur ajustement, ce qui semble indiquer que l'hypothèse habituelle d'indépendance de l'ESD n'est pas valable et qu'il existe un lien entre l'échantillon P et la LAC. L'ajustement obtenu avec le modèle (2) de l'erreur d'appariement est légèrement meilleur pour les sept modèles log-linéaires insaturés, mais à peine, en raison du fort taux d'appariement des données de la répétition générale du recensement américain de 1988. Pour le modèle [CP] [PA],

on observe un écart de 0.8738% dans l'estimation de N combiné à un taux de non-appariement de 0.3363%. En supposant que la différence dans l'estimation de N varie de façon à peu près linéaire avec le taux de non-appariement, si le taux de non-appariement avait été de 10%, à savoir s'il y avait eu appariement à 90%, l'écart entre l'estimation habituelle du maximum de vraisemblance de N et notre propre estimation aurait été de 26%.

Tableau 6

Données du système dual pour la strate 11 de St. Louis

Échantillon P	Recensement		
	Présent	Absent	Total
Présent	487	129	616
Absent	217	-	
Total	704		

Le tableau 6 présente les données habituelles du système dual pour la strate 11 de St. Louis. Le nombre de gens dénombrés lors du recensement et dans l'échantillon P est $y_{11} = 300$, le nombre de personnes dans le recensement seulement est $y_{12} = 217$ et le nombre de sujets dans l'échantillon P seulement est $y_{21} = 129$. Le dénombrement total s'élève à $y_{1+} = y_{11} + y_{12} = 704$ pour le recensement et à $y_{+1} = y_{11} + y_{21} = 616$ pour l'échantillon P. L'estimation du système dual est $\widehat{\mathrm{DSE}} = y_{1+}y_{+1}/y_{11} = 893$ (p. 232, Bishop, Fienberg et Holland 1975) et la variance estimée de $\widehat{\mathrm{DSE}}$ est $\mathrm{Var}(\widehat{\mathrm{DSE}}) = y_{1+}y_{+1}y_{12}y_{21}/y_{11}^3 = 105.4$ (p. 233, Bishop et coll. 1975). L'erreur-type s'établit à $\mathrm{SE}(\widehat{\mathrm{DSE}}) = 10.27$.

Le sous-dénombrement pour l'estimation de la population $\widehat{\mathrm{DSE}}$ est $(\widehat{\mathrm{DSE}} - y_{1+})/\widehat{\mathrm{DSE}} \times 100\% = 21.17\%$. Avec le modèle permettant le meilleur ajustement, le sous-dénombrement s'élève à $(\hat{N} - y_{1+})/\hat{N} = 55.97\%$ pour l'estimation $\hat{N} = 1599$, en supposant l'absence d'erreur d'appariement, et à 55.58% pour $\hat{N} = 1585$, selon le modèle (2) de l'erreur d'appariement. Il existe donc un biais à la hausse de 55.97% − 55.58% = 0.39% quand on ne tient pas compte des erreurs d'appariement. Ce résultat s'approche beaucoup des 0.37% calculés par Ding et Fienberg (1994) à partir des données du recensement expérimental de 1986 à Los Angeles au moyen d'un taux d'appariement pour deux échantillons de 99.4734%. Le taux d'appariement utilisé ici avec les données de St. Louis est de 99.6637%. Nos estimations révèlent que la population d'adultes de sexe masculin et de race noire vivant en milieu urbain, choisie pour la répétition générale de St. Louis, a été l'objet d'un sous-dénombrement important lors du recensement et que l'estimateur habituel à système dual ou à saisie et resaisie sous-estime gravement le sous-dénombrement. Un troisième échantillon, qualitativement différent, pourrait donner de bons résultats pour ce groupe démographique.

Des probabilités de saisie homogènes font partie des hypothèses de l'approche type à l'estimation d'une population fermée. Darroch et ses collaborateurs (1993) ont mis au point un modèle quasi-symétrique et un modèle à quasi-symétrie partielle permettant de faire varier le potentiel de saisie des sujets. Le premier modèle suppose que les trois échantillons présentent le même genre d'hétérogénéité, tandis que le modèle à quasi-symétrie partielle présume que deux échantillons ont la même hétérogénéité mais pas le troisième. Il s'agit d'un modèle raisonnable puisque le troisième échantillon est qualitativement très différent du recensement et de l'EPC. Ce modèle équivaut à une combinaison de la dépendance et de l'hétérogénéité. Pour les probabilités par cellule multinomiale, y compris la cellule manquante $R = (r_{111}, r_{112}, \ldots, r_{222})$, il s'agit de deux modèles log-linéaires de type $R = A\beta$ assortis d'une matrice de plan d'expérience A appropriée et d'un vecteur des paramètres β. Darroch et ses collaborateurs (1993) donnent les matrices de plan d'expérience pour les deux modèles.

Tableau 7

Modèles à potentiel de saisie hétérogène

Modèle log-linéaire	EP de Darroch et coll. (1993)		EP du modèle à erreur d'appariement (2)	
	$\hat{N}$ (E.-T.)	Ajustement (d.l.)	$\hat{N}$ (E.-T.)	Ajustement (d.l.)
Quasi-symétrie totale	1923.63 (216.84)	133.54 (2)	1906.61 (213.47)	133.50 (2)
Quasi-symétrie partielle	2576.54 (413.28)	11.70 (1)	2557.08 (409.39)	11.72 (1)

La méthode que nous proposons intègre aisément le potentiel de saisie hétérogène, ce qui permet d'estimer la taille de la population, en supposant un modèle d'hétérogénéité pour le tableau 1 et en adoptant l'estimation conditionnelle de vraisemblance (Sanathanan 1972). Le tableau 7 présente les estimations obtenues après ajustement du modèle quasi-symétrique et du modèle à quasi-symétrie partielle aux données de la strate 11. Encore une fois, les erreurs d'appariement de l'analyse n'ont pas d'effet sensible, en raison du taux d'appariement élevé. Le modèle à quasi-symétrie partielle permet un meilleur ajustement que le modèle quasi-symétrique, signe que l'hétérogénéité est plausible et que la LAC paraît avoir une hétérogénéité différente. Le fait que le modèle à indépendance totale s'ajuste mal pourrait aussi s'expliquer en partie par la dépendance entre les échantillons (en particulier entre le recensement et l'échantillon P), et en partie par le potentiel de saisie hétérogène.

Le modèle à quasi-symétrie partielle intègre la dépendance [CP], si bien qu'il constitue une solution de rechange au modèle [CP][PA] du tableau 5. Les deux modèles permettent un ajustement analogue des données, mais aboutissent à des estimations radicalement différentes de N, le modèle qui incorpore l'hétérogénéité produisant une estimation nettement plus élevée, assortie d'une estimation de l'erreur-type estimative beaucoup plus importante. Il se pourrait donc que les paramètres d'hétérogénéité soient très instables et, bien que les deux modèles ne soient pas

hiérarchiques, donc ne puissent se comparer directement, il semble raisonnable de retenir l'estimation la plus faible et la plus stable qui n'intègre pas l'hétérogénéité.

Darroch et coll.(1993) ont analysé quatre sous-strates de la strate 11. Les deux variables de classification croisée des quatre sous-strates O2, R2, O3 et R3 étaient: propriétaire ou locataire d'une maison, et 20 à 29 ans ou 30 à 44 ans. Les données des quatre sous-strates apparaissent au tableau 8, où 1 indique la présence et 0 l'absence dans l'échantillon. Nous avons réanalysés ces données aux fins de comparaison. Le tableau 9 et le tableau 10 donnent une estimation des deux modèles d'hétérogénéité. Comme on l'a indiqué plus haut, le taux d'appariement élevé débouche sur des estimations et un ajustement analogues pour les modèles qui incorporent l'erreur d'appariement. Le modèle à quasi-symétrie partielle donne un ajustement sensiblement plus précis que le modèle quasi-symétrique. Les meilleurs ajustements concernent les sous-strates R2 et R3. Si on additionne les estimations de N pour les quatre sous-strates, le total pour la version à quasi-symétrie partielle du modèle à erreur d'appariement donne $\hat{N} = 2980.8$, soit 16% de plus que l'estimation obtenue avec le modèle groupé du tableau 7. Bien sûr, l'erreur-type de l'estimation augmente du même ordre de grandeur.

Tableau 8

Données à trois échantillons pour quatre sous-strates de la strate 11
Source: tableau 2, Darroch et coll. (1993)

Échantillon			Sous-strate			
C	*P*	*A*	O2	R2	O3	R3
0	0	1	59	43	35	43
0	1	0	8	34	10	24
0	1	1	19	11	10	13
1	0	0	31	41	62	32
1	0	1	19	12	13	7
1	1	0	13	69	36	69
1	1	1	79	58	91	72

Tableau 9

Estimations pour une quasi-symétrie totale

Sous-strate	EP de Darroch et coll. (1993)		EP avec modèle à erreur d'appariement (2)	
	$\hat{N}$ (E.-T.)	Ajustement (d.l.)	$\hat{N}$ (E.-T.)	Ajustement (d.l.)
O2	780.83 (294.81)	11.70 (2)	777.98 (293.99)	11.69 (2)
R2	394.34 (56.45)	41.09 (2)	391.14 (55.29)	41.02 (2)
O3	765.45 (254.57)	25.99 (2)	759.97 (252.44)	25.98 (2)
R3	361.83 (47.33)	59.31 (2)	358.71 (46.20)	59.22 (2)

Tableau 10

Estimations pour une quasi-symétrie partielle

Sous-strate	EP de Darroch et coll. (1993)		EP du modèle à erreur d'appariement (2)	
	$\hat{N}$ (E.-T.)	Ajustement (d.l.)	$\hat{N}$ (E.-T.)	Ajustement (d.l.)
O2	605.66 (212.63)	7.51 (1)	601.44 (210.93)	7.52 (1)
R2	652.34 (205.12)	0.04 (1)	646.59 (202.58)	0.04 (1)
O3	1124.00 (473.26)	8.27 (1)	1126.90 (476.54)	8.22 (1)
R3	611.78 (200.82)	2.92 (1)	605.91 (198.26)	2.92 (1)

6. RÉCAPITULATION

Dans le présent article, nous proposons des modèles pour les erreurs d'appariement et des modèles pour l'estimation de la population totale et du sous-dénombrement du recensement dans un recensement à échantillon multiple. Nous avons illustré les méthodes proposées en procédant à une nouvelle analyse des données sur la couverture du recensement obtenues lors de la répétition générale de 1988 à St. Louis. Nous envisageons pour cela deux sources d'information, soit les données d'une étude sur l'erreur d'appariement (EEA) et celles d'un système triple où chaque sujet a fait l'objet d'une classification croisée selon sa présence ou son absence dans un des trois échantillons suivants: recensement, enquête postcensitaire (échantillon *P*) et liste administrative complémentaire. Nous avons intégré la formule type du modèle log-linéaire de Fienberg (1972) à notre méthode d'estimation pour tenir compte de la dépendance statistique et des erreurs d'appariement, et afin de disposer d'un test d'adéquation formelle pour les divers modèles. Notre méthode est valable pour n'importe quel modèle de type log-linéaire et nous avons montré comment les modèles d'hétérogénéité peuvent être incorporés à cette approche afin de tenir compte à la fois des erreurs d'appariement et d'un potentiel de saisie hétérogène.

Nos modèles de l'erreur d'appariement supposent des faux appariements négligeables. L'analyse de sensibilité effectuée par Ding (1990) montre que lorsque le taux de non-appariements erronés et le taux de faux appariements ont le même ordre de grandeur, le biais d'appariement résulte principalement du taux de faux non-appariements (lire aussi Fay, Passel, Robinson et Cowan 1988, p. 53). On le doit à une forte probabilité de saisie pour le recensement et l'enquête postcensitaire, si bien qu'une variation analogue des taux de non-appariements erronés et de faux appariements exerce sensiblement plus d'influence sur les non-appariements erronés que sur les faux appariements. Dans le cas des données censitaires de l'essai effectué en 1986 à Los Angeles, Ding et Fienberg (1994) estiment le taux de non-appariements erronés et le taux de faux appariements à environ 0.5% et 0.8%, respectivement. Compte tenu de ces résultats empiriques, il y a lieu de croire que

nos modèles pour les erreurs de non-appariements erronés donnent une approximation raisonnable de la réalité, du moins dans l'application censitaire dont il est question ici.

Nous avons analysé les données du système triple de St. Louis en estimant le taux d'appariement de l'EEA. Les taux d'appariement pourraient ne pas être homogènes avec différentes strates de la population. C'est pourquoi nous suggérons l'utilisation des données de l'EEA associées à la même strate d'échantillonnage. En §3, nous avons élaboré une formule pour un recensement à k échantillons et cette approche peut s'appliquer à un recensement à k échantillons où $k \geq 4$.

BIBLIOGRAPHIE

BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: M.I.T. Press.

CHEN, T.T. (1979). Log-linear models for categorical data with misclassification and double sampling. *Journal of American Statistical Association*, 74, 481-488.

CORMACK, R.M. (1968). The statistics of capture-recapture methods. *Oceanography and Marine Biology, Annals Review*, 6, 455-506.

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 American Statistical Association*, 88, 1137-1148.

DING, Y. (1990). Capture-recapture census with uncertain matching. Thèse de doctorat, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.

DING, Y., et FIENBERG, S.E. (1994). Estimation de système dual du sous-dénombrement dans le recensement lorsqu'il y a erreur d'appariement. *Techniques d'enquête*, 20, 155-165.

FAY, R.E., PASSEL, J.S., ROBINSON, J.G., et COWAN, C.D. (1988). The coverage of population in the 1980 census. Bureau of the Census, U.S. Department of Commerce.

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

HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14, 105-124.

MULRY, M.H., DAJANI, A., et BIEMER, P. (1989). The Matching Error Study for the 1988 Dress Rehearsal. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 704-709.

RAO, C.R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139-148.

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

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters.* New York: MacMillan.

ZASLAVSKY, A.M., et WOLFGANG, G.S. (1993). Triple System Modeling of Census, Post-Enumeration Survey and Administrative List Data. *Journal of Business and Economic Statistics*, 11, 279-288.