Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 125–143
Statistique Canada

Modélisation de l'erreur d'appariement et son effet sur les estimations de l'erreur d'observation du recensement

PAUL P. BIEMER[1]

RÉSUMÉ

L'efficacité des estimateurs de système dual du sousdénombrement du recensement repose en grande partie sur l'hypothèse selon laquelle des personnes dénombrées lors de l'étude d'évaluation peuvent être appariées de façon précise aux mêmes personnes dénombrées lors du recensement. Or, les erreurs d'appariement et les non-appariements erronés, qui sont inévitables, diminuent l'exactitude des estimateurs. De fait, des études ont démontré que l'ampleur de l'erreur résultante peut être suffisamment importante par rapport à l'erreur d'observation du recensement pour que l'estimation devienne inutilisable. Le présent mémoire a pour objet d'exposer un modèle d'analyse de l'effet de l'erreur d'appariement sur les estimateurs du sous-dénombrement et d'illustrer son utilisation possible dans le cadre du programme d'évaluation du sousdénombrement du recensement de 1990. L'erreur quadratique moyenne de l'estimateur de système dual est d'abord dérivée du modèle proposé et les composantes de l'EQM résultant de l'erreur d'appariement sont définies et expliquées. Nous étudierons ensuite, à la lumière du modèle, l'incidence de l'erreur d'appariement sur l'EQM de l'estimateur du sous-dénombrement du recensement. Enfin, nous illustrerons une méthodologie permettant d'utiliser le modèle pour optimaliser la conception des études d'évaluation de l'erreur d'appariement et donnerons la forme des estimateurs.

MOTS CLÉS: Sous-dénombrement; estimation de système dual; saisie-resaisie; erreur non due à l'échantillonnage; erreur de traitement.

1. INTRODUCTION

Sekar et Deming (1949) ont été les premiers à suggérer l'utilisation des méthodes de saisie-resaisie aux fins de l'évaluation du recensement et de l'enregistrement des naissances et des décès. Pour permettre l'estimation de l'erreur d'observation du recensement, la méthode nécessite l'appariement des personnes dénombrées dans le cadre d'une enquête-échantillon de la population et des personnes dénombrées dans le cadre du recensement en vue de déterminer le nombre de personnes dénombrées dans les deux cas. La mise en oeuvre de la méthode de saisie-resaisie peut se buter à nombre de difficultés susceptibles d'introduire un biais considérable dans l'estimation de la taille de la population totale, N [voir, par exemple, Burnham et collaborateurs (1987) et Wolter (1986)]. Un problème qui se pose assez souvent est l'incapacité d'apparier de façon précise les personnes dénombrées dans le cadre de l'enquêteéchantillon et les personnes dénombrées dans le cadre du recensement. Seltzer et Adlakta (1974) ont démontré que l'erreur d'appariement peut entraîner des biais relatifs de l'ordre de 33 % et qu'elle peut être positive ou négative selon que le nombre des faux non-appariements soit supérieur à celui des faux appariements ou l'inverse (voir aussi, Scheuren et Oh, 1985). Wolter (1983) souligne que l'une des raisons justifiant la non-correction de données du recensement des États-Unis de 1980 était la présomption d'erreurs d'appariement dans le cadre du programme de post-dénombrement de 1980.

Le présent mémoire fournit un cadre de base pour l'évaluation de l'erreur d'appariement afférente aux études de saisie-resaisie (en particulier pour leur application à des populations humaines) et pour l'évaluation de l'incidence de ces erreurs sur l'exactitude de l'estimation de

[1] Paul P. Biemer est chef du Department of Experimental Statistics et directeur du University Statistics Center de l'Université du Nouveau-Mexique, Las Cruces, Nouveau-Mexique, États-Unis.

N. En vue d'asseoir l'étude de l'erreur d'appariement sur une base simple et connue, nous adopterons le modèle original de saisie-resaisie de Sekar-Deming. On trouvera d'autres développements de la technique de Sekar-Deming dans les ouvrages de Marks, Seltzer et Krotki (1974), et Wolter (1986).

Considérons une population U dont la taille est de N. Un recensement est réalisé dans le cadre duquel N_c personnes sont dénombrées. Afin de permettre l'estimation de N-N_c (désigné sous le nom d'erreur d'observation du recensement), laquelle équivaut à l'estimation de N, on réalise une enquête postcensitaire (EP) dont la période de référence est identique à celle du recensement. Supposons : a) que ni le recensement ni l'EP ne comprennent de faux événements (par exemple, des doubles comptes, des falsifications, des personnes hors du champ de l'enquête ou non identifiables) ou que le nombre de ces événements peut être estimé avec exactitude et soustrait de N_c; et b) que le fait d'être dénombré dans le cadre du recensement et dans le cadre de l'enquête postcensitaire constituent deux événements indépendants.

Les personnes dénombrées dans l'EP et les personnes dénombrées dans le recensement sont appariées afin de déterminer le nombre de personnes dénombrées dans les deux cas. Soient x_{11} l'estimateur non biaisé de conception du nombre total des personnes comprises dans les populations de l'EP et du recensement, et N_p, l'estimateur non biaisé de conception de la taille de la population de l'EP. L'estimateur Sekar-Deming (désigné depuis peu sous le nom d'estimateur de système dual ou ESD) de N est

$$\hat{N} = \frac{N_p N_c}{x_{11}} \tag{1}$$

Comme nous le verrons plus loin, deux types d'erreurs peuvent avoir une incidence sur $\hat{N}$: l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage. Bien que l'erreur non due à l'échantillonnage puisse provenir de plusieurs sources, la source qui nous intéresse ici est l'erreur d'appariement; c'est-à-dire l'erreur découlant d'un classement incorrect des personnes dénombrées dans le cadre de l'EP comme ayant été recensées (erreurs positives fausses) ou comme n'ayant pas été recensées (erreurs négatives fausses).

En utilisant les développements en séries de Taylor, on peut dériver les formes générales des moments de $\hat{N}$. Il est possible de démontrer que, pour les termes de l'ordre de $1/n$, où n représente la taille de l'échantillon de l'EP,

$$\text{Biais } (\hat{N}) \doteq -N[\text{Biaisrel } (\hat{p}_{11}) - \text{Varrel } (\hat{p}_{11})] \tag{2}$$

$$\text{x } [1 + \text{Biaisrel } (\hat{p}_{11})]^{-1}$$

et

$$\text{Var } (\hat{N}) \doteq N^2 \text{ Varrel } (\hat{p}_{11}) \; [\; 1 + \text{Biaisrel } (\hat{p}_{11}) \;]^{-2} \tag{3}$$

où $\hat{p}_{11} = x_{11}/N_p$ est un estimateur de p_{11}, la proportion réelle de la population de l'EP dénombrée dans le recensement; Biaisrel $(\hat{p}_{11})$ = Biais $(\hat{p}_{11})/p_{11}$; et Varrel $(\hat{p}_{11})$ = Var $(\hat{p}_{11})\, E^{-2}(\hat{p}_{11})$. Ces équations supposent que N_c, le nombre de personnes recensées, a une variance égale à zéro. Il s'agit bien sûr d'une simplification puisque, comme il a déjà été mentionné, une estimation des faux événements du recensement a pu être soustraite du nombre de personnes recensées afin d'obtenir N_c, et que cette correction peut être sujette à des erreurs d'échantillonnage ou à d'autres erreurs. Néanmoins, l'hypothèse retenue est valable compte tenu de l'accent mis dans le présent mémoire sur l'erreur d'appariement et son incidence sur $\hat{N}$. Une application de cette méthodologie permettant la prise en compte de l'erreur dans l'estimateur N_c est présentée dans la dernière section du mémoire.

À partir des équations (2) et (3), on constate que l'erreur quadratique moyenne totale (EQM) de $\hat{N}$ est fonction de l'EQM totale de $\hat{p}_{11}$. La section suivante sera consacrée à l'étude de certains modèles d'évaluation de l'incidence de l'erreur d'appariement sur $\hat{p}_{11}$. En admettant que j $(j=1,\ldots,n)$ constitue l'indice de la $j^{ième}$ personne de l'échantillon EP, α_j peut être défini comme la probabilité qu'une erreur de classement se produise pour une personne j lors du processus d'appariement et nous pouvons étudier deux hypothèses formulées à l'égard des probabilités α_j.

2. MODELES D'ERREUR D'APPARIEMENT

2.1 Erreur d'appariement non corrélée

Supposons que:

1. L'événement {élément j est mal classé} est indépendant de l'événement {élément j' est mal classé} chaque fois que $j \neq j'$.
2. $\alpha_j = \theta$ si l'élément j a vraiment été recensé, cette probabilité est désignée comme étant la probabilité d'une erreur négative fausse, et $\alpha_j = \phi$ si l'élément j n'a pas vraiment été recensé, cette probabilité est désignée comme étant la probabilité d'une erreur positive fausse.

Afin de fixer les idées, supposons que l'EP a fait l'objet d'un sondage aléatoire simple et que n est petit comparé à N, ainsi

$$E(\hat{p}_{11}) = p_{11}(1-\theta) + (1-p_{11})\phi, \tag{4}$$

$$\text{Biais}(\hat{p}_{11}) = -p_{11}\theta + (1-p_{11})\phi \tag{5}$$

$$\text{Var}(\hat{p}_{11}) = n^{-1} E(\hat{p}_{11})\,(1-E(\hat{p}_{11})) \tag{6}$$

$$= n^{-1}\,(VE + VAS),$$

où VE, la *variance d'échantillonnage*, est obtenue par

$$VE = p_{11}(1-p_{11})\,(1-\theta-\phi)^2 \tag{7}$$

et où VAS, la *variance d'appariement simple*, est obtenue par

$$VAS = p_{11}\theta(1-\theta) + (1-p_{11})\,\phi(1-\phi) \tag{8}$$

(se reporter à l'appendice pour la preuve).

Les lecteurs qui connaissent le modèle d'erreur de réponse de Hansen, Hurwitz, et Pritzker (1964) reconnaîtront la correspondance entre leur variance de réponse simple et la VAS du présent modèle. Hansen et collaborateurs ont défini une mesure I, désignée sous le nom d'"indice d'incohérence [réponse]", comme le rapport entre la variance de réponse simple et la variance totale d'une seule réponse, c'est-à-dire la proportion de la variance qui constitue une variance de réponse. Dans le cas des réponses d'enquête, I est un indicateur de la fiabilité de réponse des informations recueillies lors de l'enquête. Il est possible d'obtenir une mesure analogue pour l'erreur d'appariement afin de déterminer l'incidence du manque de fiabilité de l'appariement sur la variance de $\hat{p}_{11}$. Cette mesure, notée I_M, est calculée à l'aide de la formule suivante:

$$I_M = \frac{VAS}{VE + VAS}. \tag{9}$$

Il est possible que les hypothèses (1) et (2) se révèlent trop restrictives pour certaines applications. Par exemple, l'hypothèse d'indépendance (1) est infirmée lorsque l'élément B de l'EP et l'élément A du recensement sont appariés de façon erronée, ce qui entraîne le classement erroné de l'appariement correct, l'élément A du recensement et l'élément A de l'EP, comme un non-appariement. Étant donné que cette situation implique que les erreurs relatives aux éléments A et B sont corrélées de façon négative, il s'ensuit que $\text{Var}(\hat{p}_{11})$ aura une valeur plus petite que celle calculée selon la formule (6). Cependant, les erreurs corrélées n'ont aucune incidence sur $E(\hat{p}_{11})$. Un autre type d'erreur d'appariement corrélée se produit lorsque l'appariement est réalisé par des commis faisant preuve d'une tendance variable à commettre des erreurs positives fausses et des erreurs négatives fausses. Un modèle décrivant ce type d'erreurs est présenté à la section suivante.

Il est stipulé dans l'hypothèse 2 que les probabilités d'erreur de classement α_j sont les mêmes pour toute la population de l'EP. Il peut s'agir d'une autre simplification étant donné qu'il est possible de classer certaines personnes, peut-être la majorité, avec un risque d'erreur relativement petit, tandis que l'appariement des autres personnes est plus difficile à réaliser. Fondamentalement, les erreurs d'appariement sont attribuables à l'inexactitude ou à la nature incomplète des informations concernant les caractéristiques de chaque personne dans l'un ou l'autre, ou encore dans les deux systèmes. Par conséquent, si l'échantillon EP peut être poststratifié selon l'exhaustivité des renseignements devant être utilisés aux fins de l'appariement, l'hypothèse peut être plausible (du moins approximativement) pour chaque sous-population. Dans un tel cas, l'erreur d'appariement totale représente une agrégation des taux d'erreur des sous-populations. Ce modèle est étudié dans la dernière sous-section du présent mémoire.

Enfin, l'hypothèse d'un sondage aléatoire simple simplifie de beaucoup l'équation de $\text{Var}(\hat{p}_{11})$. Comme les échantillons EP sont des échantillons complexes, l'hypothèse constitue une simplification qui permet néanmoins d'élaborer des formules utiles pour: a) déterminer quelles sont les composantes de l'erreur d'appariement susceptibles d'avoir l'incidence la plus marquée sur l'EQM totale de $\hat{N}$; b) concevoir des études d'évaluation de l'erreur d'appariement et affecter les ressources nécessaires à la réalisation de ces études. Dans bien des cas, l'ajustement de VE à l'aide d'une constante «d'effet de plan» permettra de compenser pour la majeure partie l'incidence de l'échantillonnage complexe sur $\text{Var}(\hat{p}_{11})$. En outre, pour l'essentiel, les formes d'échantillonnage plus complexes que le sondage aléatoire simple n'ont aucune incidence sur $E(\hat{p}_{11})$, pour autant que l'estimateur $\hat{p}_{11}$ soit pondéré de façon appropriée. En conséquence, la forme de $B(\hat{p}_{11})$ ne dépend pas de cette hypothèse.

2.2 Modélisation des erreurs de commis

Supposons que l'appariement des personnes dénombrées dans l'EP et dans le recensement est effectué par k commis. Soient m_i, le nombre de personnes appartenant à l'EP classées par le commis i, $i=1,\ldots,k$, et l'indice double (i,j), la $j^{ième}$ personne faisant partie de la tâche du $i^{ième}$ commis.

Supposons:

1. L'événement {élément (i,j) est mal classé} et l'événement {élément (i',j') est mal classé} sont indépendants lorsque $i \neq i'$ et conditionnellement indépendants pour tout commis i, $i=i'; j\neq j'; i=l, \ldots, k; j,j' = 1,\ldots,m_i$.

2. $\alpha_{ij} = \theta_i$ si la personne (i,j) a vraiment été recensée, et $= \phi_i$ si la personne (i,j) n'a pas vraiment été recensée.

3. $E(\theta_i) = \theta$; $E(\phi_i) = \phi$; $\text{Var}(\phi_i) = \sigma_\phi^2$; $V(\phi_i) = \sigma_\phi^2$; et $\text{Cov}(\theta_i, \phi_i) = \sigma_{\phi\theta}$.

En ce qui concerne le sous-ensemble de personnes faisant partie de la tâche du $i^{ième}$ commis, les hypothèses 1 et 2 sont analogues aux hypothèses 1 et 2 du modèle exposé à la section précédente. L'hypothèse 3 précise que les probabilités d'erreur d'appariement de commis sont indépendantes et qu'elles constituent des variables aléatoires distribuées de façon identique.

Cette hypothèse est analogue aux hypothèses formulées à l'égard des erreurs d'intervieweur dans le cadre des modèles d'analyse de l'incidence des erreurs d'intervieweur [se reporter, par exemple, à Kish (1962), Hartley et Rao (1978) ainsi que Biemer et Stokes (1985)]. Il convient d'utiliser cette hypothèse si l'objectif visé est d'estimer les paramètres d'un groupe de commis beaucoup plus important dont les k commis de l'EP constituent un échantillon représentatif.

Il est démontré dans l'appendice que $E(\hat{p}_{11})$ est quand même obtenu par l'équation (4) en admettant un sondage aléatoire simple. La formule générale utilisée pour déterminer $\text{Var}(\hat{p}_{11})$ correspond à la formule (A.3) de l'appendice; cependant, une simplification utile est obtenue si on peut supposer que la taille des tâches m_i est approximativement égale à m, la taille moyenne, et que le nombre d'appariements prévus est le même pour chaque tâche de commis (c'est-à-dire que les tâches de commis s'interpénètrent). Alors,

$$\text{Var}(\hat{p}_{11}) = \frac{1}{n}(SV + SMV) + \frac{m-1}{m}\frac{1}{k}CC \tag{11}$$

où CC, représentant la *composante corrélée de la variance d'appariement*, est

$$CC = p_{11}^2\,\sigma_\theta^2 + (1-p_{11})^2\sigma_\phi^2 - 2p_{11}(1-p_{11})\,\sigma_{\phi\theta} \tag{12}$$

et VE, VAS sont obtenues respectivement à l'aide des équations (7) et (8).

Il est à noter que CC résulte de la variabilité des probabilités d'erreur de classement θ_i et ϕ_i. De plus, si on considère que CC constitue la variance de $-p_{11}\,\theta_i + (1-p_{11})\phi_i$ et que les termes de cette équation sont similaires à ceux de l'équation (5), nous constatons que CC représente la variance des biais *nets* parmi les commis. Ce dernier facteur constitue la preuve que la valeur de CC doit être positive. La variance due aux commis a donc pour effet d'accroître la valeur de la variance de $\hat{p}_{11}$.

En se reportant encore aux ouvrages consacrés à la variance de réponse, on peut définir un paramètre ρ_M analogue au coefficnet de corrélation intra-intervieweur, ρ, defini par Kish (1962). Aux fins du présent mémoire, le symbole ρ_M note la corrélation intra-commis puisqu'il représente la corrélation entre les classements d'appariement de toute paire d'éléments compris dans la même tâche de commis. Selon le modèle,

$$\rho_M = \frac{CC}{SV + SMV}$$

est le rapport de la composante corrélée de la variance à la variance totale associée à un seul classement. Ce rapport peut être interprété comme le degré selon lequel les commis «influent» sur les taux d'appariement à l'intérieur de leurs tâches respectives. Ces considérations nous permettent d'établir une autre équation, équivalente à (11), pour obtenir $\text{Var}(\hat{p}_{11})$, soit

$$\text{Var}(\hat{p}_{11}) = \frac{VE + VAS}{n}[1 + (m-1)\rho_M] \tag{13}$$

2.3 Post-stratification

Le modèle d'erreur non corrélée et le modèle d'erreur de commis tiennent tous deux pour acquis (essentiellement) que le degré de difficulté afférent à la détermination de la classe exacte d'appariement est le même pour toutes les personnes dénombrées dans le cadre de l'EP (hypothèse (2) dans le cas des 2 modèles). Dans le cas du modèle d'erreur de commis, par exemple, le vecteur de la probabilité d'erreur de classement (θ_i, ϕ_i) est le même pour toutes les unités de la tâche du $i^{ième}$ commis. Toutefois, dans la pratique, certaines personnes sont beaucoup plus difficiles à classer que d'autres. Ces différences sont attribuables à des facteurs comme

l'exhaustivité des renseignements d'appariement, le fait que la personne ait déménagé ou non, qu'elle habite une maison unifamiliale ou un appartement, etc.

Une façon simple de modéliser cette situation consiste à stratifier l'échantillon de l'EP en fonction d'une variable, disons Z, corrélée avec les probabilités d'erreur de classement α_j. La variable Z peut être un indicateur de l'exhaustivité des renseignements, du type d'élément, etc.

Supposons que L de ces sous-populations sont identifiées pas l'indice h. Soit (i,h,j) le $j^{ième}$ élément de la $h^{ième}$ sous-population dans la tâche du $i^{ième}$ commis où $i=1,\ldots,k$; $h=1,\ldots,L$, $j=0,\ldots,m_{ih}$; et m_{ih} représente le nombre d'éléments de la tâche du $i^{ième}$ commis compris dans la sous-population h. Comme dans le cas du modèle de l'erreur de commis, l'hypothèse (1) peut encore être retenue. Cependant, supposons en plus que:

2. $\alpha_{ihj} = \theta_{ih}$ si la personne (i,h,j) a vraiment été recensée.

   $= \phi_{ih}$ si la personne (i,h,j) n'a pas vraiment été recensée.

3. $E(\theta_{ih}) = \theta_h$; $E(\phi_{ih}) = \phi_h$

   $\text{Var}(\theta_{ih}) = \sigma^2_{\theta h}$; $Var(\phi_{ih}) = \sigma^2_{\phi h}$;

   $Cov(\theta_{ih}', \theta_{ih}) = \sigma_{\phi\theta h}$ si $h=h'$

   $= 0$ si $h \neq h'$

Selon ces hypothèses, nous obtenons $\text{Biais}(\hat{p}_{11}) = \Sigma\pi_h \text{ Biais } (\hat{p}_{11})$ et $\text{Var } (\hat{p}_{11}) = \Sigma\pi_h^2 \text{ Var } (\hat{p}_{11h}) + \Sigma\pi_h [E(\hat{p}_{11h}) - E(\hat{p}_{11})]^2$ où Biais $(\hat{p}_{11h})$, $E(\hat{p}_{11h})$, et Var $(\hat{p}_{11h})$ sont respectivement obtenus à l'aide des formules (5), (4), et (6), indiquant les paramètres de l'erreur de commis et p_{11} par h et où $\pi_h = E(n_h/n)$, la proportion de la population comprise dans la $h^{ième}$ sous-population.

3. DÉMONSTRATION DE L'INCIDENCE DE L'ERREUR D'APPARIEMENT SUR L'ERREUR TOTALE

Les modèles définis à la section précédente peuvent être utilisés pour démontrer l'incidence de l'erreur d'appariement sur l'erreur quadratique moyenne totale de $\hat{N}$ et $\hat{p}_{11}$. Dans l'exemple qui suit, nous donnerons aux paramètres du modèle des valeurs hypothétiques qui, selon notre expérience, représentent des valeurs types et sont compatibles avec les paramètres de conception de l'EP de 1990.

Dans le cadre de l'EP, des estimations de N seront réalisées pour un certain nombre de sous-populations du recensement. Assumons que le coefficient de variation désiré des estimations est de 1%. Des équipes de commis effectueront l'appariement dans un certain nombre de centres de traitement. (On trouvera des renseignements plus détaillés sur le processus d'appariement dans la section suivante.) Afin d'illustrer l'effet de l'erreur d'appariement sur l'ESD, considérons une sous-population «type» de l'EP. Admettons, pour cette sous-population, que $p_{11} = .85$ et que k, le nombre de commis préposés à l'appariement dans un centre de traitement, est égal à 10. Dans notre analyse, nous avons considéré des valeurs de θ variant entre 0 et .10 ainsi qu'un certain nombre de valeurs types pour le rapport $\gamma = \theta/\phi$, c'est-à-dire, le rapport de la probabilité des erreurs d'appariement négatif faux à la probabilité des erreurs d'appariement positif faux. Étant donné que la mesure de ρ_M pour une erreur d'appariement n'a jamais été effectuée dans le cadre d'aucune étude, il n'existe que peu d'information susceptible d'indiquer l'étendue type de ρ_M. Cependant, si nous assumons que les probabilités d'erreur de commis θ_i et ϕ_i suivent une distribution Bêta unimodale et qu'elles sont corrélées, on peut obtenir une valeur maximale de ρ_M correspondant aux valeurs données des probabilités

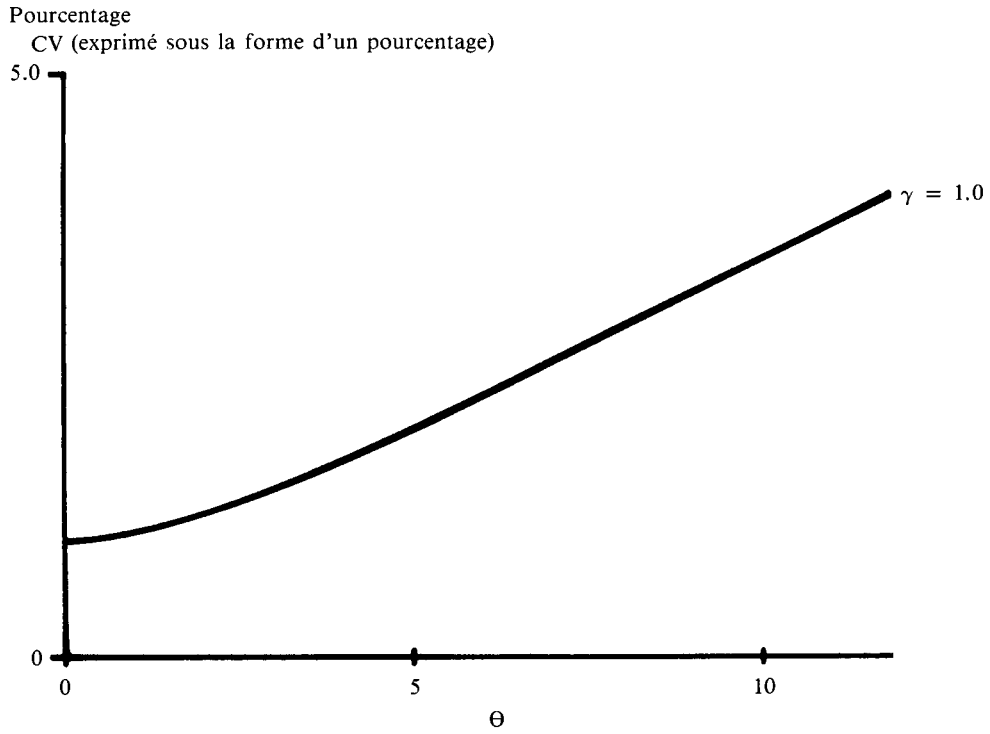

Figure 1. Coefficient de variation de $\hat{p}_{11}$ représenté comme une fonction de Θ pour $\gamma = 1$

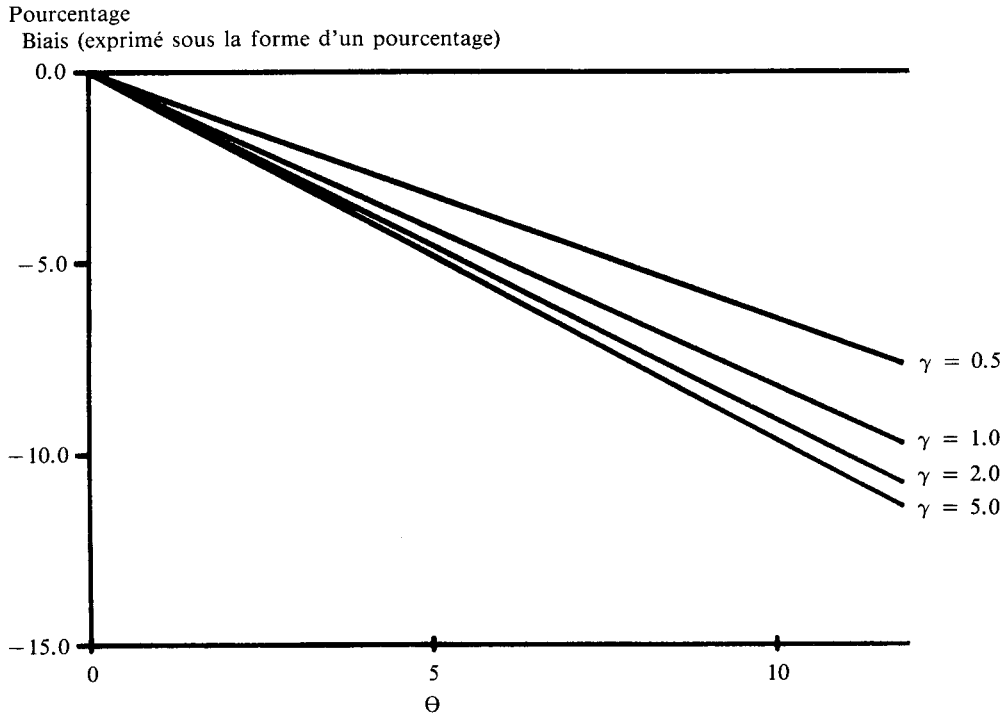

Figure 2. Biais relatif de $\hat{p}_{11}$ représenté comme une fonction de Θ pour des valeurs données de γ

d'erreur prévues θ et ϕ. Algébriquement, la valeur maximale de ρ_M est obtenue à l'aide de la formule

$$\rho_M^* = CC* / (VAS + VE) \tag{14}$$

où $CC* = p_{11}^2\,\theta^2\,(1-\theta)/(1+\theta) + (1-p_{11})^2\,\phi^2(1-\phi)/(1+\phi)$ (se reporter à Johnson et Kotz (1970) pour connaître la théorie sous-jacente). Si θ_i et ϕ_i sont corrélés de façon positive, l'hypothèse admettant une corrélation égale à zéro ne fait qu'accroître l'incidence de CC. En conséquence, les exemples suivants indiquent l'incidence maximale de la variance d'appariement sur les estimations.

Afin d'illustrer l'incidence maximale d'une variance corrélée sur l'exactitude de $\hat{p}_{11}$, on a tracé une courbe de la variation du coefficient de variation de $\hat{p}_{11}$, soit $CV(\hat{p}_{11})$, en fonction de θ pour différentes valeurs de γ. Aux fins de ces calculs, ρ_M^* a été substitué à ρ_M dans l'équation (13). L'étendue de θ était $0 \leq \theta \leq .10$ et celle deγ était $.5 \leq \gamma \leq 5$; c'est-à-dire $\phi = .2\theta$ à $\phi = 2\theta$. Cette étendue des valeurs de γ semble acceptable puisque, typiquement, ϕ est plus petit que θ. La figure 1 illustre la fonction lorsque $\gamma = 1$. Il n'existe pas de différence visible pour les autres valeurs de γ comprises dans l'étendue qui nous intéresse. Il semble donc que la grandeur de ϕ n'a qu'une influence négligeable sur le $CV(\hat{p}_{11})$. De fait, l'expression de $CC*$ démontre que, lorsque $p_{11} = .85$, pas plus de 3% de la variance corrélée est attribuable à la variance de ϕ_i, même lorsque la valeur de ϕ est égale à celle de θ. La figure 1 laisse aussi supposer que la valeur du $CV(\hat{p}_{11})$ peut passer à 2%, c'est-à-dire être doublée, pour les valeurs de θ de l'ordre de 5%.

La figure 2 illustre le biais relatif de $\hat{p}_{11}$, représenté par $BR(\hat{p}_{11})$, pour la même étendue de θ $(0 \leq \theta \leq .1)$ et de $\gamma(.5 \leq \gamma \leq 5)$. Le graphique indique clairement que le biais est moins marqué pour les valeurs plus petites de γ. De fait, le biais est égal à zéro lorsque $\gamma = (1-p_{11})/p_{11}$ ou à .18 en supposant, comme dans le présent exemple, que $p_{11} = .85$. Lorsque θ prend une valeur aussi faible que 5%, le biais relatif se situe entre -2% et -4%, selon la grandeur de γ. Si on compare ce résultat à l'augmentation maximale de $CV(\hat{p}_{11})$ (un point de pourcentage), il appert que le biais peut avoir une incidence beaucoup plus marquée que la variance corrélée.

Afin de déterminer l'incidence possible de l'erreur d'appariement sur $\hat{N}$, on a calculé l'augmentation de l'erreur totale en fonction de θ et pour certaines valeurs de γ. Soit $M(\theta;\gamma)$, $V(\theta;\gamma)$, et $B(\theta;\gamma)$ représentant respectivement l'erreur quadratique moyenne, la variance et le biais de $\hat{N}$ pour des valeurs données de θ et γ. $M(0;\gamma)$ représente l'erreur quadratique moyenne de $\hat{N}$ sans erreur d'appariement (c'est-à-dire $\theta = \phi = 0$) et $M(0;\gamma)^{1/2}$ représente donc de façon approximative l'erreur type de $\hat{N}$. Définissons $RM(\theta;\gamma) = (M(\theta;\gamma)/M(0;\gamma) - 1)^{1/2}$; $RV(\theta;\gamma) = (V(\theta;\gamma)/M(0;\gamma) - 1)^{1/2}$; et $RB(\theta;\gamma) = (B^2(\theta;\gamma)/M(0;\gamma))^{1/2}$.

Ainsi, $RM(\theta;\gamma)$ est égal au rapport de la racine carrée de l'augmentation de l'erreur quadratique moyenne totale de $\hat{N}$ pour un θ et un γ donnés, à la racine carrée de EQM de $\hat{N}$ sans erreur d'appariement. $RV(\theta;\gamma)$ représente la proportion de cet accroissement attribuable à la variance d'appariement, tandis que $RB(\theta;\gamma)$ représente la proportion attribuable au biais d'appariement. Par conséquent, $RM(\theta;\gamma)^2 = RV(\theta;\gamma)^2 + RB(\theta;\gamma)^2$. Les figures 3 et 4 illustrent ces fonctions pour deux valeurs extrêmes de γ, $\gamma = .5$ et 5 et pour $0 \leq \theta \leq .1$. La valeur maximale de la variance corrélée, $CC*$, a une fois de plus été utilisée aux fins du calcul de la variance. Aussi, il est probable que la contribution de la variance d'appariement à l'erreur totale soit exagérée de façon considérable.

Ces figures démontrent que, pour ces valeurs de θ et de γ, la majeure partie de l'erreur est attribuable au biais, bien que la contribution de la variance puisse être non négligeable. De plus, comme il a été mentionné pour les figures 1 et 2, le biais d'appariement compte pour la majeure partie de l'erreur d'appariement totale chaque fois que le pourcentage d'erreurs négatives fausses est supérieur à celui des erreurs positives fausses.

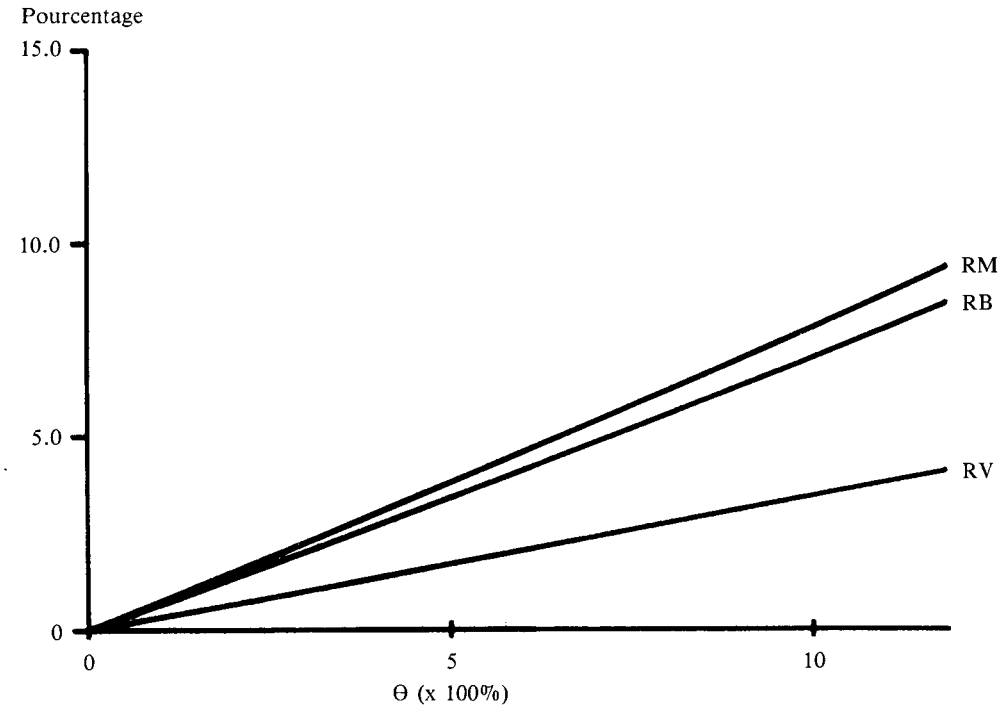

Figure 3. RM(Θ ; γ), RV(Θ ; γ), et RB(Θ ; γ), comme fonctions de Θ pour γ = .5

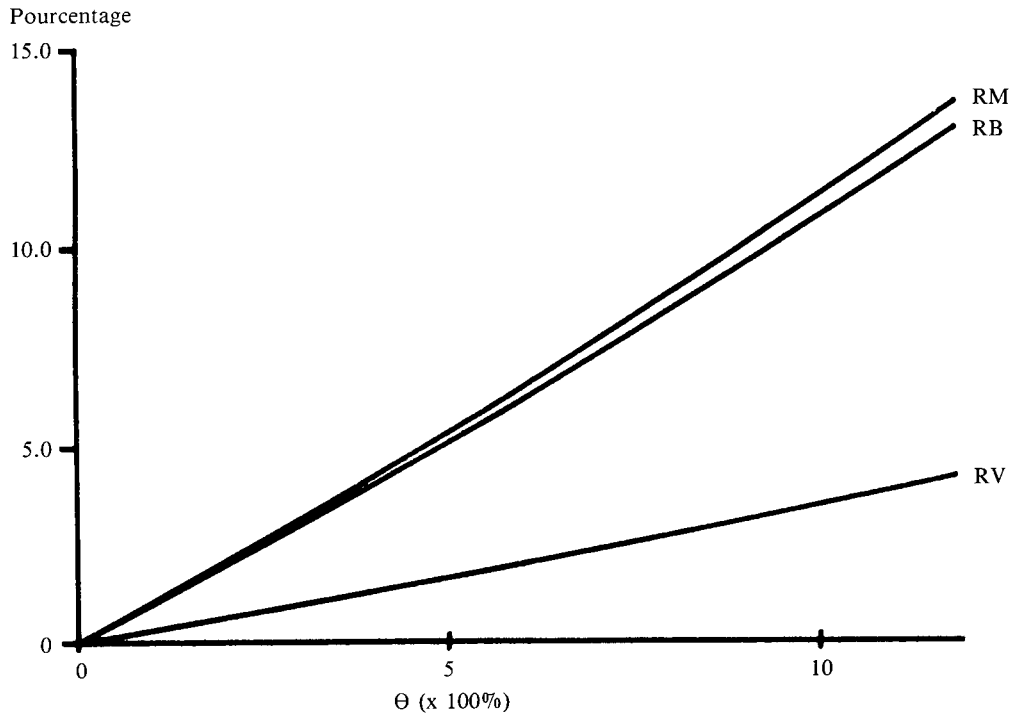

Figure 4. RM(Θ ; γ), RV(Θ ; γ), et RB(Θ ; γ), comme fonctions Θ pour γ = 5

4. ESTIMATION A PARTIR D'ÉTUDES DE RAPPARIEMENT

Plusieurs ouvrages traitent de façon exhaustive des méthodes utilisées pour estimer les composantes de l'erreur de réponse dans des enquêtes-échantillons [se reporter, par exemple, aux ouvrages de Hansen, Hurwitz, Pritzker (1964), Hansen, Hurwitz, Bershad (1961)]. Les méthodes mises en oeuvre pour estimer les composantes de l'erreur d'appariement sont essentiellement les mêmes. Par exemple, pour qu'il soit possible d'estimer la composante corrélée de la variance d'appariement, *CC*, les tâches des commis doivent «s'interpénétrer''. Cette méthode, décrite en détail dans l'ouvrage de Kish (1962), randomise l'attribution des cas d'EP aux commis de sorte que chaque tâche de commis comporte le même nombre prévu de personnes appariées. Ensuite, un estimateur de *CC* est obtenu en faisant la différence entre l'erreur quadratique moyenne des commis entre eux et l'erreur quadratique moyenne intra-commis tirées de l'analyse de la la variance due aux commis. Pour de plus amples renseignements concernant cette méthode, se reporter à l'ouvrage intitulé Bureau of the Census (1985).

La présente section porte surtout sur l'analyse des données relatives aux études de rapparriement qui constitue la méthode d'évaluation de l'erreur d'appariement la plus utilisée. Il existe deux types d'étude de rapparriement. La première tente de reproduire l'appariement initial pour un échantillon de cas en utilisant les mêmes méthodes et règles d'appariement, en donnant la même formation, etc. Ce type de rapparriement a pour objet d'estimer VAS, la variance d'appariement simple ou, de manière équivalente, I_M, l'indice d'incohérence de l'appariement. Le second type de rapparriement est utilisé en vue d'obtenir l'appariement le plus exact possible. En conséquence, sa réalisation nécessite la mise en oeuvre de méthodes plus détaillées de même que le recours à des commis experts hautement qualifiés et à l'arbitrage, c'est-à-dire à la résolution des problèmes de non-correspondance entre les classements originaux et ceux résultant du rapparriement par un tiers, expert en matière d'appariement. Ce type de rapparriement a pour objet d'estimer la valeur du biais d'appariement. De plus, comme il sera démontré plus loin, il est aussi possible d'obtenir une estimation de VAS à partir de ces données.

Les données (non pondérées) recueillies dans le cadre d'une étude de rapparriement peuvent être présentées sous la forme illustrée au tableau 1.

Supposons que l'échantillon utilisé pour le rapparriement constitue un échantillon aléatoire simple de r personnes de l'EP. De plus, il est possible de supposer que le modèle d'erreur non corrélée ou le modèle d'erreur de commis de la section précédente s'applique aussi bien à l'appariement qu'au rapparriement. Soit μ_t $(t = a, b, c, d)$ la proportion moyenne de la case correspondant à t dans le tableau 1.

Tableau 1

Données d'une étude de rapparriement

Classement original	Classement de rapparriement	
	Éléments appariés	Éléments non appariés
Éléments Appariés	a	b
Éléments non appariés	c	d

Alors,

$$\mu_a = p_{11}\,(1-\theta_A)\,(1-\theta_B) + (1-p_{11})\phi_A\phi_B \tag{15}$$

$$\mu_b = p_{11}\,(1-\theta_A)\,\theta_B + (1-p_{11})\phi_A\,(1-\phi_B) \tag{16}$$

$$\mu_c = p_{11}\,\theta_A\,(1-\theta_B) + (1-p_{11})\,(1-\phi_A)\phi_B \tag{17}$$

$$\mu_d = p_{11}\,\theta_A\,\theta_B + (1-p_{11})\,(1-\phi_A)(1-\phi_B) \tag{18}$$

où l'indice A représente l'appariement initial et l'indice B le rappariement.

Définissons,

$$\mu_A = E\left(\frac{a+b}{r}\right) = p_{11}\,(1-\theta_A) + (1-p_{11})\,\phi_A \tag{19}$$

et

$$\mu_B = E\left(\frac{a+c}{r}\right) = p_{11}(1-\theta_B) + (1-p_{11})\,\phi_B. \tag{20}$$

On notera que μ_A et μ_B sont les valeurs prévues des estimations de p_{11} réalisées respectivement à partir des classements originaux et des classements du rappariement. La différence entre ces deux estimations de p_{11}, c'est-à-dire $(b-c)/r$ est appelée *taux de différence net* (TDN). Sa valeur prévue est

$$E(NDR) = \mu_A - \mu_B = -p_{11}(\theta_A - \theta_B) + (1-p_{11})\,(\phi_A - \phi_B). \tag{21}$$

Enfin, la proportion des personnes de l'échantillon r dont le classement du rappariement ne correspond pas au classement du premier appariement est représentée par $(b+c)/r$, quotient désigné sous le nom de *taux de différence brut* (TDB). Sa valeur prévue est

$$E(TDB) = \mu_b + \mu_c$$

$$= p_{11}\,[\theta_A\,(1-\theta_B) + (1-\theta_A)\theta_B] + (1-p_{11})\,[(1-\phi_A)\phi_B + \phi_A\,(1-\phi_B)]. \tag{22}$$

Trois jeux d'hypothèses devront être pris en considération pour l'estimation des composantes de $\mathrm{Var}(\hat{p}_{11})$ et du $\mathrm{Biais}(\hat{p}_{11})$ dans l'étude du rappariement. Dans le premier cas, on suppose que l'étude de rappariement est réalisée dans les mêmes conditions générales que l'étude d'appariement initiale, de sorte que les paramètres d'erreur associés aux deux classements sont sensiblement les mêmes. Par exemple, les commis ont reçu la même formation, possèdent les mêmes aptitudes et utilisent les mêmes méthodes. Selon la deuxième hypothèse, le rapparie-ment est parfait, c'est-àdire que le classement du rappariement peut être considéré comme le vrai classement. La troisième hypothèse se situe quelque part entre la première et la deuxième, elle suppose que des méthodes d'appariement améliorées et plus détaillées sont utilisées pour le rappariement; cependant, nous ne sommes pas prêts à admettre que les classements du rapparie-ment ne comportent pas d'erreur. Nous supposerons plutôt qu'il se produit moins d'erreur lors du rappariement que lors de l'appariement initial.

Hypothèse 1. L'appariement et le rappariement ont été réalisés dans les mêmes conditions générales.

Supposons que $\theta_A = \theta_B = \theta$ et $\phi_A = \phi_B = \phi$, c'est-à-dire que le taux d'erreur de classement prévu est le même pour les deux appariements. Alors, selon la formule (21), $E(NDR) = 0$ et aucune estimation du Biais $(\hat{p}_{11})$ ne peut être calculée à l'aide des données. Cependant, selon les formules (22) et (8)

$$\tfrac{1}{2} E(TDB) = VAS \tag{23}$$

De plus, un estimateur de I_M dans la formule (9) est obtenu par

$$\hat{I}_M = TDB / [2\hat{p}_{11}(1-\hat{p}_{11})] \tag{24}$$

où $\hat{p}_{11}$ est l'estimateur EP de p_{11} comme il a été défini pour (2). Comme alternative, un estimateur de $E(\hat{p}_{11})$ peut être obtenu à partir du tableau 1; par exemple, se reporter aux estimateurs des formules (19) et (20).

Hypothèse 2. Rappariement parfait.

Supposons que $\theta_B = \phi_B = 0$; c'est-à-dire que le rappariement est réalisé sans erreur de classement. Alors, selon la formule (21),

$$\begin{aligned} E(TDN) &= -p_{11}\theta_A + (1-p_{11})\phi_A \\ &= \text{Biais } (\hat{p}_{11}). \end{aligned} \tag{25}$$

De plus, la probabilité d'erreur négative fausse, θ_A, est estimée par

$$\hat{\theta} = c/(a+c). \tag{26}$$

et la probabilité d'erreur positive fausse, ϕ_A, est estimée par

$$\hat{\phi} = b/(b+d). \tag{27}$$

Un estimateur de VAS est obtenu à l'aide de la formule

$$\widehat{VAS} = \frac{1}{r}\left(\frac{ac}{a+c} + \frac{bd}{b+d}\right) \tag{28}$$

et on obtient un estimateur de I_M grâce à la formule

$$\hat{I}_M = \widehat{VAS}/\hat{p}_{11}(1-\hat{p}_{11}) \tag{29}$$

où $\hat{p}_{11}$ représente un estimateur de $E(\hat{p}_{11})$ obtenue à partir de l'EP ou du tableau 1.

Hypothèse 3. Le rappariement comprend moins d'erreurs mais n'est pas parfait.

Supposons que $0 < \theta_B < \theta_A$ et $0 < \phi_B < \phi_A$; c'est-à-dire que les probabilités d'erreur de classement sont plus faibles dans le cas du rappariement que dans celui de l'appariement initial, sans être nulles. Il n'existe alors pas d'estimateur non biaisé du Biais $(\hat{p}_{11})$. Cependant, $|E(TDN)|$ sera plus petite que $|\text{Biais } (\hat{p}_{11})|$ si les différences $\mu_A - p_{11}$ et $\mu_B - p_{11}$ ont toutes deux le même signe; c'est-à-dire que les estimateurs définis à partir de l'appariement initial et du rappariement sont biaisés dans la même direction. Ainsi, dans ces conditions, $|TDN|$ constitue l'estimateur minorant de $|\text{Bias } (\hat{p}_{11})|$.

De plus, il n'existe pas d'estimateur non biaisé de VAS. Il est toutefois possible de démontrer à partir de la formule (22) que

$$E(TDB) - 2VAS = p_{11}(\theta_B - \theta_A)(1 - 2\theta_A) + (1 - p_{11})(\phi_B - \phi_A)(1 - 2\phi_A).$$

Ainsi, chaque fois que θ_A et ϕ_A ont tous deux une valeur inférieure à .5, ce qui est vrai pour la plupart des applications pratiques, il s'ensuit que

$$E(TDB) < 2VAS$$

et que $\hat{I}_M$ défini par la formule (24) constitue une sous-estimation de I_M.

5. APPLICATION AU RECENSEMENT DE 1990

Dans le cadre du recensement de 1990, l'échantillon de l'EP sera constitué d'environ 5 000 «îlots» ou groupes d'environ 30 logements contigus et on tentera de procéder à l'appariement de chaque personne comprise dans les îlots avec les personnes recensées. Les variables utilisées aux fins de l'appariement seront le nom, l'adresse, le lien avec le chef de ménage, le sexe, la date de naissance, l'état matrimonial, la race et l'origine hispanique. Le processus d'appariement comportera les quatre étapes suivantes.

Étape 1. *Un appariement sera réalisé par ordinateur selon la technique de Fellegi et Sunter (1969).* Chaque personne faisant partie de l'EP sera classée par ordinateur soit dans la catégorie des personnes appariées à une personne recensée, soit dans celle des personnes non appariées ou encore dans la catégorie des personnes qu'il serait peut être possible d'apparier (c'est-à-dire dont le classement doit être révisé par un commis).

Étape 2. *Des commis effectueront une première révision afin de corriger toutes les erreurs d'appariement ou de non-appariement erroné introduites par l'ordinateur.* De plus, un ensemble normalisé de règles d'appariement sera appliqué à chaque appariement possible. Ainsi, chaque personne dénombrée dans le cadre de l'EP sera classée sous une des rubriques appariement, non-appariement, appariement possible ou cas non résolu.

Étape 3. *Des commis effectueront avec discernement une deuxième révision afin de vérifier de nouveau les classements réalisés aux deux étapes précédentes.* A cette étape-ci, les commis, désignés sous le nom d'équipe d'appariement spéciale (EAS), peuvent aussi décider du bien fondé d'un suivi sur place auprès de certains ménages.

Étape 4. *Une révision «subséquente au suivi sur place».* Les cas sont examinés de nouveau à la lumière de toute information supplémentaire obtenue dans le cadre du suivi sur place. Les codes de classement définitifs sont: apparié (dénombré), non apparié (non dénombré) ou non résolu (statut d'appariement devant être déterminé à l'étape de traitement final).

Les méthodes utilisées pour attribuer les codes «apparié» ou «non apparié» aux cas non résolus sont décrites dans le rapport de Schenker (1987). Ces cas, qui représentent environ 1% de l'échantillon de l'EP, ne sont pas pris en compte dans les tableaux qui suivent puisque les statuts d'appariement n'avaient pas été attribués aux cas non résolus, au moment de ce test. Néanmoins, l'erreur d'imputation peut constituer une source importante d'erreur d'appariement et poser des problèmes d'évaluation particuliers. Il est probable, par exemple, que certains cas d'EP non résolus resteront non résolus lors du rappariement et aucune estimation directe de l'erreur de classement ne peut être calculée pour ces cas. Dans le test décrit plus loin, 83% des cas non résolus dans le cadre de l'EP sont demeurés non résolus au rappariement. Inversement, 41% des cas qui n'ont pas été résolus au rappariement, ont été résolus à l'appariement de l'EP. Si on suppose que les imputations pour les cas qui n'ont pas été résolues lors du rappariement sont erronées, on peut obtenir un majorant de l'erreur d'imputation. De même, il est possible d'obtenir un minorant en supposant que toutes ces imputations sont justes. Cependant, à moins que la proportion représentée par les imputations soit très faible, cette analyse des «meilleurs cas, pires cas» peut produire des bornes trop élevées pour être utiles.

En 1986, un test d'essai de ces méthodes d'appariement d'EP a été effectué à Los Angeles. Un échantillon d'environ 4 000 personnes a été apparié avec le recensement d'essai de Los Angeles, puis apparié de nouveau par des spécialistes du recensement en vue d'évaluer le biais d'appariement. Des méthodes spéciales on été utilisées lors du rappariement afin d'assurer un classement d'appariement très exact. Le tableau 2 illustre les taux de désaccord entre les quatre étapes d'appariement et de rappariement. Ce tableau met en lumière l'amélioration des classements à chaque étape supérieure, amélioration indiquée par un taux de désaccord décroissant dans la colonne réservée au rappariement. Les données indiquent également que peu de classements sont modifiés lors de l'étape subséquente au suivi sur place (.68% en désaccord avec l'étape 3). De plus, le TDB de l'étape finale (par rapport au rappariement) est très bas, soit de moins de 1%.

En supposant que le processus de rappariement conduit à l'attribution de vrais statuts d'appariement, le tableau 3 indique les estimations de θ, la probabilité d'erreur d'appariement négatif faux, et de ϕ, la probabilité d'erreur d'appariement positif faux, pour chaque étape d'appariement. Il ressort de ce tableau que le taux de non-appariement faux est plus élevé dans le cas de l'appariement par ordinateur et dans le cas du premier appariement effectué par des commis. Cependant, le contraire est vrai pour les deux dernières étapes d'appariement.

Tableau 2

Comparaison des taux de désaccord entre les étapes d'appariement (%)

	Étape 2	Étape 3	Étape 4	Rappariement
Étape 1	2.9	4.4	4.7	5.5
Étape 2	0	3.3	4.0	4.8
Étape 3	3.3	0	.68	1.6
Étape 4	4.0	.68	0	.87

Tableau 3

Estimations de θ et ϕ pour les étapes d'appariement

Étape d'appariement	Estimation de θ (x100%) (taux de non-appariement faux)	Estimation de ϕ (x100%) (taux d'appariement faux)
1	6.2	2.3
2	5.1	3.3
3	1.5	2.1
4	.1	.3

Tableau 4

Resultats de l'étude de rappariement (pondérés).

Classement de l'appariement initial	Classement du rappariement	
	Appariés	Non appariés
Appariés	16 690	9
Non appariés	85	2 178

Tableau 5a

Resultats du rappariement pour les cas qui ont conservé leur statut jusqu'à la fin du procédé de rappariement.

Classement de l'appariement initial	Classement du rappariement	
	Appariés	Non appariés
Appariés	14 458	0
Non appariés	640	1 775

Tableau 5b

Resultats du rappariement pour les cas dont le classement a été modifié au cours d'au moins une des quatre étapes.

Classement de l'appariement initial	Classement du rappariement	
	Appariés	Non appariés
Appariés	2 223	9
Non appariés	21	403

En utilisant la méthodologie de la section précédente, il est possible d'estimer Biaisrel$(\hat{p}_{11})$, Biaisrel$(\hat{N})$ et I_M, l'indice d'incohérence de l'appariement. Le tableau 4 révèle les résultats de l'étude du rappariement, pondérés en fonction des probabilités de sélection de l'échantillon du rappariement. Pour ce tableau, le Biaisrel$(\hat{p}_{11})$ est estimé à $-.4$ %, et Biaisrel$(\hat{N})$ est donc estimé à .4 %. Ces calculs sont effectués à l'aide de la formule (2) en supposant que le coefficient de variation de $\hat{p}_{11}$ est de 1 % et en remplaçant le Biaisrel$(\hat{p}_{11})$ par son estimation. L'indice d'incohérence de l'appariement est estimé à .49 %, ce qui représente un très faible pourcentage. Le taux d'appariement positif faux, $\hat{\phi}$, est égal à .004 et le taux d'appariement négatif faux, $\hat{\theta}$, est égal à .005.

Comme il a été mentionné à la deuxième section, la probabilité d'erreur d'appariement peut, entre autres choses, être fonction de l'exhaustivité des données de l'EP ou du recensement. L'échantillon du rappariement à été divisé en deux sous-échantillons afin d'illustrer dans quelle mesure les taux d'erreur d'appariement peuvent varier. Le premier sous-échantillon était composé des cas qui avaient été classés comme «appariés» ou «non appariés» à toutes les étapes d'appariement sans exception. Le second sous-échantillon était constitué du reste de l'échantillon, c'est-à-dire des cas dont le classement a été modifié au cours d'au moins une des quatre étapes. Cette division de l'échantillon correspond à peu près à une division reposant sur l'exhaustivité des données d'appariement étant donné que la plupart des cas classés dans la même

catégorie au cours des quatre étapes sont ceux pour lesquels les données recueillies sont les plus complètes. Les tableaux 5a (cas pour lesquels les données sont complètes) et 5b (cas pour lesquels les données sont incomplètes) illustrent les résultats pondérés.

Dans le premier cas, le taux d'appariement négatif faux est de .44% tandis que le taux d'appariement positif faux est nul. Donc, aucun de ces cas n'a fait l'objet d'un appariement erroné, bien qu'un certain nombre d'entre eux aient été faussement désignés comme nonappariements. Ces données peuvent attester de la plus grande habileté du personnel responsable du rapparierment à apparier les cas de l'EP. I_M est estimé à .39%, ce qui représente un indice très bas. En ce qui concerne les cas pour lesquels les données sont incomplètes, le taux d'appariement négatif faux est de .93% tandis que le taux d'appariement positif faux est de 2.18%. I_M est estimé à 1.1%, ce qui représente encore un indice assez bas. Cependant, ces données indiquent que le risque de produire de faux appariements est plus grand en ce qui concerne les cas pour lesquels l'information est incomplète.

Les données recueillies dans le cadre de cette étude indiquent que l'erreur d'appariement induit un faible biais négatif ($-.4\%$) dans l'estimateur $\hat{N}$, qui correspond à une sous-estimation de la valeur de N d'environ un million de personnes (en supposant que $N = 250$ millions de personnes). Le biais est seulement de $-.7\%$ même pour les cas les plus difficiles. Il serait intéressant d'analyser certains sous-groupes démographiques de la population (personnes ayant déménagé, enquêtés-substituts et personnes habitant un appartement) afin de déterminer l'importance de l'erreur d'appariement pour ces sous-groupes.

6. RÉSUMÉ

Les modèles et les formules de l'EQM qui ont été développés dans le présent mémoire peuvent être utilisés pour évaluer l'incidence de l'erreur d'appariement sur des estimations de l'erreur d'observation du recensement. Dans le cadre du recensement des États-Unis de 1990, le biais de l'erreur d'appariement semble être l'élément le plus important de l'EQM($\hat{N}$). Des études préliminaires ont démontré que le biais de l'erreur d'appariement pour le recensement de 1990 est de peu d'importance, sa valeur étant inférieure à .5%. Cette estimation ne tient pas compte de l'erreur d'imputation qui s'applique à environ 1% des cas de l'EP. De plus, l'exactitude des estimations de biais repose dans une large mesure sur la véracité de l'hypothèse selon laquelle le rapparierment permet d'obtenir le vrai classement d'appariement. D'autres études seront nécessaires pour déterminer la validité de cette hypothèse.

Lors de l'élaboration des formules de l'erreur quadratique moyenne totale de $\hat{N}$, il a été supposé qu'il était peu probable que N_c comporte une erreur. Cependant, dans la pratique, il convient de soustraire de N_c une estimation du nombre d'événements faux du recensement (ou des dénombrements erronés), représentée par EE. Comme cet estimateur est obtenu à partir d'un appariement des éléments d'un échantillon de la population du recensement aux éléments de la population de l'EP, EE est aussi sujet à une erreur d'échantillonnage et à une erreur d'appariement. Par exemple, une personne peut être classée comme dénombrée de façon erronée alors qu'elle l'avait été correctement (erreur de dénombrement positif faux), ou peut être classée comme dénombrée correctement alors qu'elle l'avait été de façon erronée (erreur de dénombrement négatif faux). La méthodologie et le modèle présentés pour évaluer l'incidence des erreurs d'appariement positif faux et d'appariement négatif faux pour x_{11} peuvent facilement s'appliquer à l'estimateur des dénombrements erronés. On notera que les termes des formules d'approximation de Taylor, (2) et (3), utilisées pour calculer le biais et la variance de $\hat{N}$ seront remplacés par les termes correspondant au biais et à la variance de EE.

En ce qui concerne les recherches à venir, des études portant sur la variance corrélée de l'erreur d'appariement seront nécessaires pour déterminer dans quelle mesure la variance due aux commis contribue à l'erreur totale de $\hat{N}$. Il est possible que $CC*$, l'incidence maximale de l'erreur corrélée, se traduise par une surestimation considérable de l'incidence de l'erreur de commis.

REMERCIEMENTS

La rédaction du présent mémoire a été rendue possible grâce à un Joint Statistical Agreement avec le Bureau of the Census des États-Unis. L'auteur souhaite remercier Aref Dejani du Bureau of the Census pour avoir mis à sa disposition le soutien informatique nécessaire. Il remercie également Bernice Garrett qui a dactylographié le mémoire et s'est chargée de sa relecture.

APPENDICE

Dérivation des formules de l'EQM

Soit U la population de taille N à dénombrer, U_c le sous-ensemble tiré de la population U S l'échantillon tiré de l'EP, et S_c l'ensemble des personnes de l'EP faisant partie du recensement, représentant $S \cap U_c$. Enfin, soit les n éléments de l'échantillon S $u_1, \ldots, u_n$. Définissons les variables

$$\eta_i = 1 \text{ si } u_i \in S_c$$

$$= 0 \text{ si } u_i \notin S_c$$

et

$$y_i = 1 \text{ si } u_i \text{ a été classé (par appariement) dans } S_c.$$

$$= 0 \text{ si } u_i \text{ n'a pas été classé dans } S_c.$$

Modèle d'évaluation de l'erreur de commis

Supposons que: 1) y_i est une variable aléatoire, que les probabilités que $P(y_i = 1 \mid \eta_i = 0) = \phi$ et $P(y_i = 0 \mid \eta_i = 1) = \theta$, et 2) que y_i et y_j sont des variables indépendantes étant donné η_i et η_j pour $i \neq j$. Soit $E(\cdot \mid S)$ l'espérance conditionnelle et $V(\cdot \mid S)$, la variance conditionnelle étant donné S. Alors, $\hat{p}_{11} = \Sigma y_i / n$ et $E(\hat{p}_{11} \mid S) = (1-\theta)\tilde{p}_{11} + \phi(1-\tilde{p}_{11})$ où $\tilde{p}_{11} = \Sigma \eta_i / n$. Le résultat de l'expression (5) est obtenu en calculant l'espérance par rapport S.

De plus, $V(y_i \mid \eta_i = 0) = \phi(1-\phi)$ et $V(y_i \mid \eta_i = 1) = \theta(1-\theta)$. Par conséquent, $V(\hat{p}_{11} \mid S) = \phi(1-\phi)(1-\tilde{p}_{11})/n + \theta(1-\theta)\tilde{p}_{11}/n$.

VAS (8) est obtenu en calculant l'espérance par rapport à S.

Enfin, le résultat de l'expression (6) est obtenu en combinant $VE(\hat{p}_{11} \mid S)$ et $EV(\hat{p}_{11} \mid S)$.

Modèle d'évaluation de l'erreur de commis

Soit (i,j) la $j^{ième}$ personne de la tâche du $i^{ième}$ commis et soit y_{ij} *et* η_{ij} définis en fonction de y_i et η_i. Supposons (1) — (3) pour le modèle d'erreur de commis. E_2, V_2, et C_2 respectivement, l'espérance, la variance et la covariance conditionnelles par rapport aux distributions de l'erreur de commis maintenant l'échantillon de commis fixé. Les valeurs correspondantes E_1 l'espérance, V_1 la variance, C_1 la covariance par rapport à la sélection aléatoire des vecteurs des paramètres des commis k, comme pour l'hypothèse (3), maintenant l'échantillon S fixé. Donc,

$$E_1 E_2 (\hat{p}_{11}) = E_1 \left\{ \sum_i \left[(1-\theta_i) \frac{n_{1i}}{n} + \phi_i \frac{n_{0i}}{n} \right] \right\}$$

$$= (1 - \theta)\tilde{p}_{11} + \phi(1 - \tilde{p}_{11})$$

où $n_{1i} = \sum_j \eta_{ij}$ et $n_{0i} = \sum_j (1 - \eta_{ij})$. Donc, (4) est obtenu en calculant l'espérence de (A.1) par rapport à S.

Considérons la variance de $\hat{p}_{11}$. On a $\text{Var}(\hat{p}_{11}) = \text{VE}(\hat{p}_{11}|\, S) + \text{EV}(\hat{p}_{11}|\, S)$ où $\text{E}(\hat{p}_{11}|\, S)$ provient de (A.1). De plus $n^2 V(\hat{p}_{11}|\, S) = \sum_i \sum_j V(y_{ij}|\, S) + \sum_i \sum_{j \neq j'} \text{Cov}(y_{ij}, y_{ij'}|\, S)$ où $V(y_{ij}|\, S) = V_2(y_{ij}) + V_1E_2(y_{ij})$ et $\text{Cov}(y_{ij}, y_{ij'}|\, S) = C_1\,[E_2(y_{ij}), E_2(y_{ij'})]\,, E_1C_2(y_{ij}, y_{ij'})$ étant égal à zéro. Étant donné que $E_2(y_{ij}) = \phi_i$, pour $\eta_{ij} = 0$, et que $E(y_{ij}) = 1 - \theta_i$ pour $\eta_{ij} = 1$, $V_1E_2(y_{ij}) = \sigma_\phi^2$, si $\eta_{ij} = 0$, et $V_1E_2(y_{ij}) = \sigma_\theta^2$ si $\eta_{ij} = 1$. De plus, $V_2(y_{ij}) = \phi_i(1-\phi_i)$ pour $\eta_{ij} = 0$ et $V_2(y_{ij}) = \theta_i(1-\theta_i)$ pour $\eta_{ij} = 1$. Ainsi

$$E_1V_2(y_{ij}) = \phi(1-\phi) - \sigma_\phi^2 \text{ si } \eta_{ij} = 0$$

$$= \theta(1-\theta) - \sigma_\theta^2 \text{ si } \eta_{ij} = 1$$

De la même façon, on peut démontrer que, pour $j \neq j'$,

$$C_1\,\{E_2(y_{ij}),\, E_2(y_{ij'})\} = \sigma_\theta^2 \text{ si } (\eta_{ij}, \eta_{ij'}) = (1,1)$$

$$= -\sigma_{\phi\theta} \text{ if } (\eta_{ij}, \eta_{ij'}) = (1,0)$$

$$= \sigma_\phi^2 \text{ if } (\eta_{ij}, \eta_{ij'}) = (0,0).$$

Par conséquent,

$$V(\hat{p}_{11}|\, S) = (\Sigma m_i^2 - n)\,/\,n^2\, CC + VAS/n. \tag{A.2}$$

Finalement, en combinant (A.1) et (A.2) dans l'expression

$$V(\hat{p}_{11}) = VE(\hat{p}_{11}|\, S) + EV(\hat{p}_{11}|\, S), \text{ nous obtenons}$$

$$V(\hat{p}_{11}) = 1/n(VE + VAS) + (\Sigma m_i^2 - n)\,/\,n^2\, CC. \tag{A.3}$$

Si de plus, nous supposons que $m_i = m$ pour tout i, nous obtenons la forme de l'expression (11).

BIBLIOGRAPHIE

BIEMER, P. P., et STOKES, S. L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of American Statistical Association* 80, 158-166.

BUREAU OF THE CENSUS (1985). *Evaluating censuses of population et housing*, Statistical Training Document. United States Bureau of the Census, Washington, D.C.

BURNHAM, K. P., ANDERSON, D. R., WHITE, G. C., BROWNIE, C., et POLLOCK, K. H. (1987). *Design et Analysis Methods for Fish Survival Experiments Based on Release — Recapture*. American Fisheries Society Monograph 5.

FELLEGI, I. P., et SUNTER, A. B. (1969). A Theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

HANSEN, M. H., HURWITZ, W. N., et BERSHAD, M. A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.

HANSEN, M. H., HURWITZ, W. N., et PRITZKER, L. (1964). The estimation and interpretation of cross differences and the simple response variance. Dans *Contributions to Statistics* (éd. C. R. Rao), Oxford: Pergamon Press, 111-136.

HARTLEY, H. O., et RAO, J. N. K. Estimation of nonsampling variance components in sample surveys. Dans *Survey Sampling and Measurement*, (éd. N.K. Namboodiri), New York: Academic Press.

JOHNSON, N. L., et KOTZ S. (1969). *Continuous Univariate Distributions II*. Boston: Houghton Mifflen.

KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

MARKS, E. S., SELTZER, W., et KROTKI, K. J. (1974). *Population Growth Estimation*. New York: Population Council.

RICKER, W. E. (1958). *Handbook of Computations for Biological Statistics of Fish Populations*. Fisheries Research Board of Canada. Ottawa: Queen's Printer and Controller of Stationery.

SEBER, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*, (3^{e} éd.). New York: MacMillian.

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.

SELTZER, W., et ADLAKTA A. (1974). On the effect of errors in the application of the Chandrasekaran — Deming techniques, Reprint 14. Laboratory for Population Statistics, University of North Carolina.

SCHENKER, N. (1987). Report on missing data in the 1986 test of adjustment related operations. Survey Research Division Report Series, Bureau of the Census, RR-87/09.

SCHEUREN, F., et OH, H. L. (1985). Fiddling around with nonmatches and mismatches. *Proceedings of the Workshop on Exact Matching Methodologies*. Arlington, Virginia.

WOLTER, K. M. (1983). Affidavit, Mario Cuomo et coll. *vs.* Malcolm Baldrige et coll. U. S. District Court, Southern District of New York, 80 Civ. 4550.