Techniques d'enquête, décembre 1988
Vol. 14, n° 2, pp. 257–280
Statistique Canada

L'erreur totale dans l'estimateur de système dual: Recensement du Central Los Angeles County de 1986

MARY H. MULRY et BRUCE D. SPENCER[1]

RÉSUMÉ

Le U.S. Bureau of the Census utilise des estimateurs de système dual (ESD) pour évaluer l'erreur de couverture dans le recensement. Ce genre d'estimateur repose sur des données du recensement initial et d'une enquête postcensitaire. Lorsqu'on mesure la précision de l'ESD, il importe de savoir que cet estimateur est soumis à plusieurs composantes de l'erreur d'échantillonnage et de l'erreur non due à l'échantillonnage. Dans cet article, nous décrivons des modèles de l'erreur totale et des composantes d'erreur dans les estimateurs de système dual. Ces modèles établissent un rapport entre des indices observés de la qualité des données, comme le taux d'erreur d'appariement, et les deux premiers moments des composantes d'erreur. Nous analysons également la propagation de l'erreur dans l'ESD et évaluons le biais et la variance de cet estimateur. La méthode proposée est appliquée au recensement du Central Los Angeles County de 1986 dans le cadre du Test des opérations de redressement du U.S. Bureau of the Census. Cette méthode sera aussi utile pour évaluer l'erreur dans l'ESD à l'occasion du recensement de 1990 et pour d'autres applications.

MOTS CLÉS: Erreur non due à l'échantillonnage; enquête postcensitaire; évaluation de la couverture; sous-dénombrement; saisie-resaisie.

1. INTRODUCTION

L'estimateur de système dual (ESD) est utilisé dans plusieurs disciplines pour estimer la taille d'une population. Il peut s'agir de populations d'animaux comme de populations d'êtres humains. Par exemple, le U.S. Bureau of the Census utilise des ESD du nombre de naissances dans son analyse démographique pour estimer la population des États-Unis. Il prévoit utiliser le même genre d'estimateur pour mesurer l'erreur de couverture dans le recensement décennal de 1990. Dans cet article, nous allons insister sur l'application de l'ESD dans le domaine du recensement, où les deux systèmes sont le dénombrement initial et une enquête postcensitaire (EP).

L'estimateur le plus élémentaire fondé sur l'ESD du sous-dénombrement dans le recensement est $\widehat{SD}$ et est défini $\widehat{SD}$ = ESD - REC, où REC désigne le chiffre du recensement. Puisque ESD = REC + $\widehat{SD}$, les ESD produisent aussi d'autres estimateurs de population. Une catégorie plus générale d'estimateurs fondés sur l'ESD (Spencer 1980, 1986) est $(1 - f) \times$ REC $+ f \times$ ESD, qui équivaut à

$$\text{REC} + f \times \widehat{UC}$$

où $0 \leq f \leq 1$.

Les estimations de l'erreur totale dans l'ESD sont indispensables pour déterminer quelle valeur de f produit l'estimateur de population le plus précis. Comme l'intervalle des valeurs

[1] Mary Mulry, Undercount Research Staff, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233. Bruce Spencer, Department of Statistics, Northwestern University, Evanston, IL 60201 et NORC. Les opinions exprimées dans cet article sont celles des auteurs et ne reflètent pas nécessairement la position du U.S. Bureau of the Census.

de f comprend 0 et 1, on peut choisir aussi bien REC que ESD. Les critères permettant de juger de la supériorité d'une série d'estimations démographiques par rapport à une autre peuvent être fondés sur des mesures de la qualité de la distribution de la population (Hogan et Mulry 1987; Spencer 1986). Les estimations de l'erreur totale dans l'ESD jouent aussi un rôle important dans la planification statistique (par ex., combien d'argent devrait être dépensé et quelle devrait être la taille d'un échantillon de l'EP.)

Les ESD sont soumis à plusieurs composantes de l'erreur non due à l'échantillonnage, sans compter l'erreur d'échantillonnage. Nous présentons dans cet article des modèles de l'erreur totale et des composantes d'erreur dans l'ESD. Ces modèles établissent un rapport entre des indicateurs observés de la qualité des données et les deux premiers moments des composantes d'erreur. Nous utilisons ensuite des méthodes de propagation de l'erreur pour estimer le biais et la variance de l'ESD. De cette façon, nous évaluons l'erreur totale ou l'effet combiné des erreurs. Parmi les ouvrages qui abordent les modèles d'erreur pour l'ESD, voir Seltzer et Adlakha (1974).

La méthode exposée est appliquée au recensement du Central Los Angeles County de 1986, connu aussi sous le nom de Test des opérations de redressement (TOR) de 1986 pour la région de Los Angeles (Diffendal 1988). L'échantillon de l'EP réalisée à l'occasion du TOR comprenait environ 6,000 unités de logement et plus de 19,000 personnes. Une analyse de sensibilité montre comment les composantes d'erreur influent l'une sur l'autre, lesquelles s'annulent et lesquelles se complètent. Les méthodes exposées ici pour estimer l'erreur dans l'ESD du TOR peuvent aussi servir à estimer l'erreur dans les ESD du recensement de 1990.

Nous avons cherché à structurer cet article de manière à faciliter la tâche de ceux qui ne veulent pas le lire en entier. Ainsi dans la section 2, nous exposons les fondements de l'ESD utilisé dans le TOR de même que ses principales composantes. Nous décrivons ensuite la méthode utilisée pour évaluer les composantes d'erreur et les combiner dans le but d'estimer l'erreur totale dans l'ESD (Section 3). Une description précise des composantes d'erreur exige une description détaillée de l'ESD, appuyée de symboles (Section 4). Cette description est suivie d'une évaluation des composantes d'erreur (Section 5). Une synthèse de ces composantes aboutit au calcul d'estimations de l'erreur totale dans l'ESD (Section 6). Enfin, nous présentons les principales conclusions (Section 7).

2. ESTIMATEUR DE SYSTÈME DUAL

L'utilisation de l'estimateur de système dual suppose nécessairement qu'il existe deux listes de la population. La première est le résultat du dénombrement initial et la seconde est une liste implicite des personnes incluses dans la base de sondage servant à la formation de l'échantillon P de l'EP, et qui constituent ce que nous appellerons la population de l'échantillon P. La base de sondage proprement dite n'est pas une liste de personnes mais d'îlots de recensement.

L'échantillon P est un des deux échantillons utilisés dans l'EP. Celle-ci comprend l'échantillon D, qui est constitué d'enregistrements du recensement, et l'échantillon P, qui est un

Tableau 1

Probabilité d'inclusion dans une case

	Recensement		
	Inclus	Exclus	Total
Échantillon P inclus	p_{i11}	p_{i12}	p_{i1+}
exclus	p_{i21}	p_{i22}	p_{i2+}
Total	p_{i+1}	p_{i+2}	p_{i++}

Tableau 2

Effectif réel de chaque case

	Recensement		
	Inclus	Exclus	Total
Échantillon P inclus	N_{11}	N_{12}	N_{1+}
exclus	N_{21}	(N_{22})	(N_{2+})
Total	N_{+1}	(N_{+2})	(N_{++})

échantillon de personnes. Le premier sert à estimer le nombre d'enregistrements erronés tandis que le second sert à estimer, par estimation de système dual, le nombre de personnes qui ont été oubliées lors du dénombrement initial.

L'estimateur de système dual repose sur un modèle où la probabilité que le i-ème individu d'une population de taille N ait ou non été recensé et qu'il soit ou non dans l'échantillon P est celle indiquée dans le tableau 2.1 (Wolter 1986a); voir cette source pour une analyse et un renvoi à des ouvrages antérieurs. La taille réelle de la population dans chaque classe est indiquée dans le tableau 2.2.

Dans le tableau ci-dessus, $N_{++} = N$, la population totale. Même si nous connaissions les N_{ij} de la première ligne et de la première colonne, nous ne pourrions observer directement les N_{ij} entre parenthèses et il faudrait donc les estimer à l'aide du modèle. L'ESD de N serait alors sous la forme $N_{1+}N_{+1}/N_{11}$, ce que nous considérons comme l'ESD idéal.

Lorsqu'on estime la taille d'une population en vue d'évaluer l'erreur de couverture dans un recensement, on remplace les N par des estimations tirées du dénombrement initial et de deux échantillons postcensitaires, soit l'échantillon P et l'échantillon D. Les données d'enquête sont pondérées par l'inverse de la probabilité de sélection. Dans les définitions suivantes, le symbole " ˆ " dénote la possibilité d'erreur non due à l'échantillonnage:

N_p = le nombre pondéré de personnes dans la population cible de l'échantillon P

$\hat{N}_p$ = estimation de la population totale d'après l'échantillon P

REC = chiffre du recensement

II_1 = nombre de personnes imputées

II_2 = nombre pondéré d'enregistrements du recensement qui ne renferment pas assez d'information pour être appariés

EE = nombre pondéré d'enregistrements erronés dans le recensement, d'après l'échantillon D

$\widehat{EE}$ = estimation du nombre d'enregistrements erronés dans le recensement

C = REC - II_1 - II_2 - EE = nombre pondéré des personnes recensées, d'après l'échantillon D

$\hat{C}$ = REC - II_1 - II_2 - $\widehat{EE}$ = nombre estimé des personnes recensées, d'après l'échantillon D

M = nombre pondéré des personnes de l'échantillon P qui ont été recensées initialement

$\hat{M}$ = nombre estimé des personnes de l'échantillon P qui ont été recensées initialement.

Selon la notation ci-dessus, $\hat{N}_p$ est un estimateur de N_p, lequel est un estimateur sans biais de N_{1+}. Le *ratio* $\hat{C}/\hat{M}$ sert à estimer le *ratio* N_{+1}/N_{11}. (En soi, $\hat{C}$ et $\hat{M}$ ne sont pas de bons

estimateurs de N_{+1} et N_{11}.) Par conséquent, l'estimateur a la forme suivante $\hat{N}_{++} = \hat{N}_p \hat{C}/\hat{M}$. Le ratio $\hat{C}/\hat{M}$ renferme un facteur de correction pour les enregistrements erronés et les enregistrements qui ne contiennent pas assez d'information pour être appariés II_1 et II_2, de manière que les cas n'ayant aucune chance d'être inclus dans le dénominateur ne figurent pas non plus dans le numérateur.

L'ESD sert à estimer le sous-dénombrement net en pourcentage, ou *taux de sous-dénombrement net*, dans le recensement,

$$\hat{S} = 100\,(\text{REC} - \hat{N}_{++})/\hat{N}_{++}.$$

Pour la région d'essai en général (notamment le Central Los Angeles County), REC = 355,352, $\hat{N}_p$ = 336,707, $\hat{C}$ = 343,567, $\hat{M}$ = 298,204, et $\hat{N}_{++}$ = 388,040. Suivant ces chiffres, le taux de sous-dénombrement net estimé est 8.42.

3. MÉTHODE D'ÉVALUATION DE L'ERREUR TOTALE

L'ESD est exposé à diverses sources d'erreur, comme des adresses inexactes dans l'échantillon P, des données manquantes (non-réponse partielle ou totale), des erreurs de réponse, des erreurs commises par l'interviewer, le biais de corrélation, l'erreur d'échantillonnage, *etc.* Nous chercherons ici à évaluer les effets de ces diverses sources d'erreur sur l'ESD.

La première étape consiste à exprimer l'ESD comme une fonction des composantes. Celles-ci ont été construites de telle manière que les diverses sources d'erreur agissent d'une façon indépendante ou parfaitement corrélative sur la plupart d'entre elles. En isolant les effets des diverses erreurs, nous pouvons plus facilement définir les principales sources d'erreur.

En second lieu, nous estimons les deux premiers moments de chaque composante d'erreur prise individuellement. Pour cela, nous nous fondons sur les résultats de diverses évaluations du TOR et de programmes de contrôle de la qualité. La manière dont les composantes ont été construites implique que le coefficient de corrélation entre ces composantes est normalement égal à 0 ou à 1.

Nous avons eu recours à des méthodes de simulation par ordinateur pour étudier la propagation des erreurs. Nous avons supposé une distribution multidimensionnelle des composantes d'erreur, par exemple F. La définition de F était compatible avec les deux premiers moments estimés dans la section 5. Nous avons simulé la présence de composantes d'erreur par des tirages pseudo-aléatoires à même F, puis nous avons calculé l'ESD; nous avons répété cette opération 10,000 fois et la distribution empirique de l'ESD ainsi obtenue a servi d'estimation de la distribution réelle. Les deux premiers moments de cette distribution fournissent des estimations numériques de l'erreur totale de l'ESD.

Nous avons réalisé une analyse de sensibilité pour vérifier l'importance d'utiliser une forme de distribution plutôt qu'une autre pour F. Les résultats de cette analyse donnent à penser que la forme de distribution exacte (au-delà des deux premiers moments) est relativement sans importance (voir section 6).

L'analyse de l'erreur dans l'ESD s'est faite selon une approche bayesienne. Nous avons estimé les deux premiers moments des distributions des composantes d'erreur, puis nous avons déterminé la distribution a posteriori du taux de sous-dénombrement en fonction des valeurs observées de $\hat{C}$, $\hat{N}_p$, $\hat{M}$, *etc.*

4. COMPOSANTES DE L'ESD

L'ESD est exposé aux erreurs d'échantillonnage et aux erreurs non dues à l'échantillonnage, y compris la non-vérification des hypothèses qui sous-tendent le modèle de l'ESD. Cet estimateur comporte effectivement un biais mais celui-ci est négligeable en ce qui concerne le

recensement (Wolter 1986a). Les erreurs non dues à l'échantillonnage peuvent influer sur la précision des estimateurs deN_{+1}, de N_{1+}, et de N_{11}. Nous décrivons ci-dessous l'erreur non due à l'échantillonnage.

Dans l'estimation de N_{+1} l'erreur est définie par l'expression $\hat{C} - N_{+1} = (\hat{C} - C) + (C - N_{+1})$. Le premier terme $(\hat{C} - C)$ est l'erreur non due à l'échantillonnage nette, qui entre dans la composition du biais et de la variance, et le second terme $(C - N_{+1})$ est l'erreur d'échantillonnage, qui entre uniquement dans la composition de la variance. Définissons l'erreur non due à l'échantillonnage nette comme c $= \hat{C} - C$.

L'erreur nette c se produit durant le traitement de l'échantillon D, lorsqu'on indique que des répondants ont été enregistrés correctement ou incorrectement, selon le cas, lors du dénombrement initial alors que c'est le contraire. Par conséquent, c a trois composantes : c_e, qui survient durant la collecte et le traitement des données, c_b, qui est le résultat d'un plan de sondage pour l'EP qui ne réussit pas à concilier les estimations des taux de surdénombrement et de sous-dénombrement bruts, et c_i attribuable aux données manquantes, $c = c_e + c_b + c_i$. Ces trois composantes sont étudiées dans les sections 5.5, 5.6 et 5.7 respectivement.

Dans l'estimation de N_{1+} l'erreur est définie par l'expression $\hat{N}_p - N_{1+} = (\hat{N}_p - N_p) + (N_p - N_{1+})$. Le premier terme $(\hat{N}_p - N_p)$ est l'erreur non due à l'échantillonnage, qui entre dans la composition du biais et de la variance, et le second terme $(N_p - N_{11})$ est l'erreur d'échantillonnage, qui entre uniquement dans la composition de la variance. L'erreur non due à l'échantillonnage nette est définie $n_p = \hat{N}_p - N_p$.

L'erreur nette n_p se produit à l'étape de l'interview réservée aux personnes de l'échantillon P, plus précisément lorsque des personnes comprises dans l'échantillon P ne sont pas interviewées. Dans un tel cas, soit que des enregistrements ont été fabriqués ou qu'il y a des données manquantes. Par conséquent, n_p a deux composantes: n_{pf}, l'erreur due à la fabrication, et n_{pi}, l'erreur due aux données manquantes,. $n_p = n_{pf} + n_{pi}$. La première de ces composantes est étudiée dans la section 5.3, tandis que la seconde est l'objet de la section 5.7. En ce qui concerne l'estimation de N_{11} l'erreur est définie par l'expression $\hat{M} - N_{11} = (\hat{M} - M) + (M - N_{11})$. Le premier terme $(\hat{M} - M)$ est l'erreur non due à l'échantillonnage nette, qui entre dans la composition du biais et de la variance, et le second terme $(M - N_{11})$ est l'erreur d'échantillonnage, qui entre uniquement dans la composition de la variance.

Afin de décrire plus facilement l'erreur non due à l'échantillonnage dans l'estimation de N_{11}, considérons les tableaux ci-dessous, qui portent sur les personnes comprises dans la population cible de l'échantillon P et les répondants du même échantillon. Le tableau 3 indique le nombre pondéré de personnes de la population cible de l'échantillon P dans chaque classe tandis que le tableau 4 indique le nombre pondéré de répondants de l'échantillon P dans chaque classe. Quant au tableau 5, on y trouve les estimations de l'effectif de chaque classe établies à l'aide des résultats de l'interview de l'échantillon P et de l'appariement.

Tableau 3

Population cible de l'échantillon P

Échantillon P	Code de dénombrement	
	Recensé	Non recensé
Exclus	D_{11}	D_{12}
Inclus		
Adresse au jour du recensement exacte	D_{21}	D_{22}
Adresse au jour du recensement inexacte	D_{31}	D_{32}

Tableau 4

Code de dénombrement pour les répondants de l'échantillon P

Échantillon P	Code de dénombrement	
	Recensé	Non recensé
Enregistrement fabriqué	A_{11}	A_{12}
Enregistrement non fabriqué		
Adresse au jour du recensement exacte	A_{21}	A_{22}
Adresse au jour du recensement inexacte	A_{31}	A_{32}

Tableau 5

Code d'appariement pour les répondants de l'échantillon P

Échantillon P	Code d'appariement	
	Apparié	Non apparié
Enregistrement fabriqué	B_{11}	B_{12}
Enregistrement non fabriqué		
Adresse au jour du recensement exacte	B_{21}	B_{22}
Adresse au jour du recensement inexacte	B_{31}	B_{32}

Comme le groupe de personnes réputées incluses dans l'échantillon P selon le tableau 3 correspond au groupe de répondants qui sont réputés non fictifs d'après le tableau 4, $D_{21} = A_{21}$ et $D_{31} = A_{31}$. De plus, $A_{11} = 0$ puisqu'un enregistrement fabriqué au cours de l'EP ne pouvait exister lors du recensement. Par conséquent,

$$\mathrm{M} = D_{11} + D_{21} + D_{31} = D_{11} + A_{21} + A_{31}.$$

Comme un enregistrement fabriqué durant l'EP n'a pas d'équivalent dans la liste du recensement, nous supposons que $B_{11} = 0$. Ainsi, $\hat{M} = B_{11} + B_{21} + B_{31} = B_{21} + B_{31}$.

L'erreur non due à l'échantillonnage dans l'estimation de N_{11}, (désignée par m, peut donc être définie comme suit:

$$\begin{aligned} m &= \hat{M} - M \\ &= (B_{11} + B_{21} + B_{31}) - (D_{11} + D_{21} + D_{31}) \\ &= -D_{11} + (B_{21} - A_{21}) + (B_{31} - A_{31}). \end{aligned}$$

L'erreur m a trois composantes: $(B_{21} - A_{21})$, qui est l'erreur introduite au cours de l'appariement (section 5.2), $(B_{31} - A_{31})$, qui est l'erreur introduite par les répondants lorsque ceux-ci indiquent mal l'adresse où ils demeuraient le jour du recensement (section 5.3); et $-D_{11}$. D_{11} a elle-même deux composantes: m_i, l'erreur attribuable à l'absence de codes d'appariement, et m_f, l'erreur due à la fabrication. La première est étudiée dans la section 5.7 tandis que la seconde fait l'objet de la section 5.4.

L'ESD idéal peut être exprimé par la formule suivante:

$$N_{1+}\, N_{+1}/N_{11} = (\hat{C} - c)(\hat{N}_p - n_p)(\hat{M} - m).$$

5. COMPOSANTES DE L'ERREUR DANS L'EP

Les estimations des deux premiers moments de la distribution a posteriori du taux de sous-dénombrement découlent des estimations des deux premiers moments des composantes de l'erreur dans l'EP. Ces composantes sont le biais de corrélation, l'erreur d'appariement, l'erreur de déclaration d'adresse, la fabrication dans l'échantillon P, l'erreur dans l'estimation du nombre d'enregistrements erronés, la conciliation des estimations des taux de surdénombrement et de sous-dénombrement bruts, les données manquantes et l'erreur d'échantillonnage. Dans cette section, nous allons décrire la source de chaque composante et définir un modèle pour chacune. Aux fins de la modélisation, nous considérons les composantes d'erreur comme des indicateurs observables de la qualité des données. Nous estimons enfin les deux premiers moments des distributions des erreurs en vue de les utiliser dans le modèle de l'erreur totale défini à la section 6.

5.1 Biais de corrélation

5.1.1 Source d'erreur

Dans l'estimation de système dual, la préoccupation majeure est que l'échantillon P produise une estimation précise de la proportion de la population dénombrée lors du recensement. Le non-respect de l'une ou l'autre des hypothèses d'indépendance qui sous-tendent l'estimation de système dual peut avoir pour effet d'introduire un biais dans l'estimation de la proportion de la population dénombrée lors du recensement et, par voie de conséquence, dans l'estimation de la population.

Trois hypothèses d'indépendance sont posées pour l'estimateur de système dual:

Causalité. Le fait qu'une personne est incluse dans la population recensée est indépendant du fait qu'elle est incluse dans l'échantillon de l'EP. Autrement dit, le rapport des produits croisés satisfait

$$\theta_i = p_{i11}\, p_{i22}\, /p_{i12}\, p_{i21} = 1, \text{ pour } i = 1, \ldots, N.$$

Homogénéité. Les probabilités de saisie satisfont $p_{il+} = p_{1+}$ ou $p_{i+1} = p_{+1}$ pour $i = 1, \ldots, N$, dans chacune des strates formées a posteriori.

Autonomie. Le recensement et l'EP sont le résultat de N essais mutuellement indépendants.

L'hypothèse de l'homogénéité est conforme au modèle composé M_{th} décrit dans Wolter (1986a). Tout l'appareil mathématique exposé dans Wolter (1986a) pour le modèle M_t de Peterson s'applique aussi au modèleM_{th} lorsqu'il y a suffisamment de données pour former des strates a posteriori où s'applique le modèle M_t .

Pour réduire le degré d'hétérogénéité de la population, le U.S. Bureau of the Census exécute une stratification a posteriori des données selon des variables démographiques et géographiques; cette méthode avait été proposée à l'origine par Sekar et Deming (1949). Elle consiste à calculer une estimation de la population dans chaque strate formée a posteriori puis à additionner toutes les estimations pour obtenir la population totale estimée. À moins que l'hypothèse de l'homogénéité ne se vérifie pas du tout, l'estimation se situe normalement entre le chiffre du recensement et le chiffre de la population réelle.

Wolter (1986b) et Cowan et Malec (1986) ont démontré que le non-respect de l'hypothèse de l'autonomie avait un effet négligeable sur le biais de l'ESD mais faisait augmenter sa variance.

La formulation de Wolter prévoit une action individuelle (autonomie) ou collective (absence d'autonomie) des membres des ménages. Pour leur part, Cowan et Malec présentent un modèle qui permet de grouper les cas d'omission dans le recensement (absence d'autonomie). Nous allons maintenant décrire par un modèle l'effet combiné des sources du biais de corrélation sur l'ESD.

5.1.2 Définition

Pour mieux connaître l'effet du biais de corrélation, supposons que $\theta_i = \theta$ pour tous les i et exprimons la taille réelle de la population par la formule

$$N = N_{11} + N_{12} + N_{21} + \theta\ (N_{12}N_{21}/N_{11}),$$

où θ_i est le rapport des produits croisés défini dans la section 5.1.1.

Le biais de corrélation influe uniquement sur le dernier terme de l'expression car les trois autres peuvent être estimés directement. Le paramètre θ, représente l'effet de la non-vérification des hypothèses d'indépendance. Lorsque celles-ci se vérifient, $\theta = 1$.

Le biais de corrélation, qui existe lorsque θ est différent de 1, est la seule composante de t, qui est l'erreur due à la non-applicabilité du modèle. La taille de la population peut être définie par la formule:

$$\begin{aligned} N &= N_{1+}N_{+1}/N_{11} + t \\ &= N_{1+}N_{+1}/N_{11} + (\theta - 1)(N_{12}N_{12}/N_{11}). \end{aligned}$$

En conséquence, le biais de corrélation $t = (\theta - 1)(N_{12}N_{21}/N_{11})$.

5.1.3 Mesure

On peut estimer le paramètre θ au niveau national pour des sous-groupes raciaux et ethniques en utilisant des estimations de la population tirées d'une analyse démographique. L'utilisation de cette méthode suppose toutefois qu'il s'agit là d'estimations précises. Malgré cela, cette formulation permet aussi de faire varier θ afin d'évaluer la sensibilité de l'ESD à l'effet estimé du non-respect des hypothèses d'indépendance.

5.1.4 Estimation

On n'a pas calculé d'estimation de θ pour le TOR de 1986 parce qu'il n'y avait pas d'autre source d'estimations démographiques (par ex., on ne pouvait tirer d'estimations d'une analyse démographique). Cependant, Ericksen et Kadane (1985) ont calculé trois estimations de θ pour la population noire en ce qui regarde le recensement de 1980: 2.1, 2.7 et 3.7. Comme la population visée par le TOR de 1986 était constituée en majeure partie de membres de minorités ethniques (73 pour cent de personnes d'origine hispanique, 12 pour cent de personnes d'origine asiatique et 15 pour cent qui ne sont ni d'origine asiatique ni d'origine hispanique), nous nous servirons des estimations d'Ericksen et Kadane: $E(\theta) = 2.1$, 2.7, ou 3.7, $\text{Var}(\theta) = 0$. Nous considérons ici que θ est fixe mais inconnu. Dans la section 6, nous réalisons une analyse de sensibilité pour illustrer l'effet de diverses valeurs de θ.

Les estimations de θ sont conformes à ce qui est indiqué dans les rapports des observateurs qui ont participé au recensement d'essai de Los Angeles (Childers et coll. 1987). Nous sommes d'avis que le biais de corrélation est plus élevé pour les régions urbaines que pour le pays en général. En conséquence, il peut s'agir ici d'estimations prudentes étant donné le caractère urbain de la région d'essai.

5.1.5 Résumé

Dans le modèle de l'erreur totale, les deux premiers moments de la distribution a posteriori du facteur de correction pour le biais de corrélation sont supposés être $E(\theta) = 2.1$, 2.7 ou 3.7 et $\text{Var}(\theta) = 0$.

5.2 Erreur d'appariement

5.2.1 Source d'erreur

Dans la présente analyse, l'erreur d'appariement désigne les erreurs qui sont commises lorsqu'on compare les enregistrements de l'échantillon P à ceux du recensement. L'erreur d'appariement n'englobe donc pas les erreurs de réponse qui surviennent durant la collecte des données. Bien que d'autres types d'erreur puissent entraîner l'attribution d'un mauvais code de dénombrement à un répondant de l'échantillon P, ces sources sont traitées sous d'autres rubriques.

Une fois l'interview de l'échantillon P terminée, on parcourt les enregistrements du recensement pour vérifier si les répondants en question ont bel et bien été recensés. On attribue alors un code à chacun des répondants selon qu'il a pu être apparié ou non à un enregistrement du recensement. Lorsqu'on attribue le mauvais code de dénombrement à une personne de l'échantillon P durant le traitement des données, on commet ce qu'on appelle une erreur d'appariement. Il y a deux façons de commettre des erreurs d'appariement. Soit qu'une personne est appariée à un enregistrement du recensement alors qu'elle n'a pas été recensée à l'origine (c'est ce qu'on appelle une «fausse concordance»), soit qu'elle est reconnue comme non recensée alors qu'elle l'a effectivement été (c'est ce qu'on appelle une «fausse non-concordance»). L'erreur d'appariement aura pour effet de fausser l'estimation de l'effectif recensé et de l'effectif de l'échantillon P et introduira par conséquent un biais dans l'estimation du nombre de personnes oubliées dans le recensement.

5.2.2 Définition

Le dénominateur N_{11} de l'estimateur de système dual est estimé à l'aide des données de l'échantillon P. Nous avons vu dans la section 4 que:

A_{21} = le nombre pondéré de personnes qui ont été recensées,

B_{21} = le nombre estimé de personnes qui ont pu être appariées.

Par conséquent, l'erreur nette due à l'attribution du mauvais code de dénombrement, m_m, peut être définie $m_m = B_{21} - A_{21}$. L'espérance mathématique et la variance de m_m étant donné la valeur observée $\hat{M}$ sont désignées par $\mathrm{E}(m_m)$ et $\mathrm{Var}(m_m)$.

5.2.3 Mesure

On peut mesurer m_m en traitant une seconde fois une partie de l'échantillon P, c.-à-d. en confiant à du personnel hautement qualifié la tâche d'exécuter un nouvel appariement. On suppose pour cela que des personnes mieux entraînées commettent moins d'erreur même si elles disposent des mêmes documents et des mêmes données qui ont servi à l'appariement initial. Il est possible de rapprocher les nouveaux codes d'appariement et les codes d'appariement initiaux et d'éliminer les différences.

Deux études du processus d'appariement ont été réalisées à l'aide des données du TOR de 1986. Ces études ont servi à évaluer l'appariement pour les personnes ayant déménagé et les personnes n'ayant pas déménagé respectivement.

Dans la seconde étude (Corby et Mulry 1988), on a prélevé un sous-échantillon probabiliste de 35 îlots pour le soumettre à un nouvel appariement; cette opération a été confiée à des spécialistes du bureau central. L'échantillon ainsi obtenu a été stratifié selon le taux d'appariement et des taux de sondage exceptionnellement élevés ont été appliqués aux îlots qui présentaient de faibles taux d'appariement de telle manière que les personnes affectées au contrôle de la qualité pouvaient recueillir le plus de renseignements possible sur les erreurs d'appariement. Comme les îlots adjacents n'ont pas été analysés, il se peut que l'on ait sous-estimé le nombre de cas de fausse non-concordance.

L'autre étude a servi à évaluer l'erreur d'appariement pour les personnes ayant déménagé (Childers et coll. 1987). Quatre-vingt-dix personnes ayant déménagé n'ont pu être appariées dans le TOR et ces 90 cas ont été soumis à un nouvel appariement. Cette opération a permis de découvrir onze cas de concordance additionnels, dont deux n'avaient pas été retenus à la suite du contrôle informatique.

5.2.4 Estimation

Nous allons maintenant utiliser les résultats des sous-échantillons qui ont servi aux études précitées pour estimer les moments de la distribution de m_m tirée de l'échantillon de l'EP. Le fait de ne pas effectuer de recherche approfondie dans l'étude portant sur les personnes n'ayant pas déménagé a problablement contribué à sous-estimer le nombre de cas de fausse non-concordance. Les résultats de recherches approfondies antérieures nous indiquent qu'une façon modérée de pallier à l'absence de telles recherches est d'augmenter de 20% l'erreur nette, qui est de 70 en l'occurrence (Hogan et Wolter 1988). D'après les résultats des deux études, l'erreur nette dans l'échantillon de l'EP est de 95. Le taux d'erreur nette est donc $-$ 0.0055. Nous appliquons ce taux uniquement aux cas de l'échantillon P qui ont été résolus car l'erreur d'imputation pour les cas non résolus est traitée dans la section 5.7 - Données manquantes. L'espérance de m_m devient $\mathrm{E}(m_m) = -1831$, lorsque le poids d'échantillonnage global (17) est utilisé.

On n'a pas calculé d'estimation de la variance de l'estimateur de l'erreur d'appariement nette pour les personnes n'ayant pas déménagé. La variance d'échantillon du nombre d'erreurs pour les personnes ayant déménagé est nulle car toutes les personnes de cette catégorie qui n'avaient pu être appariées ont fait l'objet d'un nouvel appariement. Cependant, nous ne croyons pas que la variance réelle est nulle. On pourrait définir une variance en supposant que les erreurs se sont produites suivant une combinaison de processus de Poisson, par exemple les erreurs d'appariement pour les personnes ayant déménagé ayant suivi un processus de Poisson particulier et les erreurs d'appariement pour les personnes n'ayant pas déménagé ayant suivi indépendamment un autre processus de Poisson. Si on considérait que toutes les erreurs suivent un seul et même processus de Poisson, on obtiendrait en définitive une estimation prudente de la variance; en l'occurrence, cette estimation serait le produit de 17 $\times$ 107. Cependant, le modèle de Poisson ne produit pas nécessairement d'estimation prudente si les erreurs se produisent dans des grappes. Dans une tentative pour établir des estimations prudentes de la variance, nous avons multiplié (quelque peu arbitrairement) l'estimation de la variance calculée suivant un seul processus de Poisson par le poids d'échantillonnage global, ce qui a donné

$$\mathrm{Var}(m_m) = (17)^2 \times 107 = 30{,}923.$$

5.2.5 Résumé

Pour le modèle de l'erreur totale, les deux premiers moments de la distribution a posteriori de l'erreur d'appariement nette pour l'échantillon de l'EP sont supposés être $\mathrm{E}(m_m) = -1831$ et $\mathrm{Var}(m_m) = 30{,}923$.

5.3 Erreur de déclaration de l'adresse au jour du recensement

5.3.1 Source d'erreur

Quelques-uns des répondants de l'échantillon P ont déménagé entre le jour du recensement et le jour de l'interview de l'EP. Les réponses fournies à cet égard peuvent être erronées. Par exemple, si le répondant a déménagé, l'adresse antérieure qu'il a indiquée peut être inexacte ou peut avoir été mal géocodée par les préposés du codage. L'une ou l'autre de ces erreurs peut fausser le processus d'appariement en orientant la recherche d'enregistrements dans une région différente de celle où le répondant a été recensé. Ainsi, des répondants qui ont été effectivement recensés pourraient être considérés comme des cas de non-concordance parce que les opérations d'appariement ne permettent pas de repérer les enregistrements correspondants. En classant par erreur des répondants parmi les cas de non-concordance, on introduit un biais par excès dans l'estimation du nombre de personnes oubliées dans le recensement.

Il arrive qu'une erreur de déclaration d'adresse ne se traduise pas par une fausse non-concordance. Si l'adresse au jour du recensement est dans le secteur où doit normalement se trouver l'adresse déclarée et que celle-ci est géocodée correctement, il y aura appariement.

5.3.2 Définition

Le dénominateur N_{11} de l'estimateur de système dual est estimé à l'aide des données de l'échantillon P. Nous avons vu dans la section 4 que:

A_{31} = le nombre pondéré de personnes qui ont été recensées mais dont l'adresse au jour du recensement est inexacte;

B_{31} = le nombre estimé de personnes dont l'adresse au jour du recensement est inexacte mais dont l'enregistrement a été repéré à une autre adresse.

Par conséquent, l'erreur nette due à l'inexactitude de l'adresse au jour du recensement, m_a, peut être définie $m_a = B_{31} - A_{31}$. L'espérance et la variance de m_a , étant donné la valeur observée $\hat{M}$, sont désignées par $\mathrm{E}(m_a)$ et $\mathrm{Var}(m_a)$. respectivement.

5.3.3 Mesure

L'évaluation de m_a repose sur un suivi effectué auprès d'un sous-échantillon de répondants de l'échantillon P qui sont classés parmi les personnes non recensées. Les résultats du suivi servent à estimer l'erreur qui se produit lorsque des personnes qui ont été recensées indiquent mal l'adresse où elles demeuraient le jour du recensement lorsqu'elles répondent à l'EP.

On a évalué l'importance de l'erreur de déclaration d'adresse à la suite du TOR de 1986. Après la production des estimations initiales, on a réalisé une interview de rappel auprès d'un échantillon de 903 personnes qui n'avaient pu être appariées afin de déterminer le nombre de cas de non-concordance attribuables à une erreur de déclaration d'adresse. En ce qui concerne les répondants qui ont dit avoir déménagé sans quitter la région d'essai, on a tenté un appariement à la nouvelle adresse.

5.3.4 Estimation

Les cas échantillonnés pour lesquels on a constaté une erreur de déclaration d'adresse peuvent servir à estimer

L_e = le nombre pondéré de personnes qui, durant l'interview de l'échantillon P, ont mal indiqué l'adresse où elles demeuraient le jour du recensement.

Une recherche d'enregistrements du recensement aux nouvelles adresses produit

r_{am} = l'estimateur du pourcentage de personnes qui ont mal indiqué l'adresse où elles demeuraient le jour du recensement mais qui peuvent être appariées à des enregistrements du recensement.

Par conséquent, l'espérance de l'erreur m_a est estimée par

$$\mathrm{E}(m_a) = -r_{am}L_e.$$

Selon les résultats de l'interview de rappel (Hogan et Wolter 1988), il y aurait eu erreur de déclaration d'adresse dans tout au plus 3.1% des cas de l'échantillon P. On a calculé un taux d'appariement estimé de 33% pour les personnes qui ont mal indiqué l'adresse où elles demeuraient le jour du recensement et qui ont déménagé sans quitter la région d'essai. Si nous supposons un taux d'appariement identique pour les personnes qui ont dit avoir demeuré à l'extérieur de la région d'essai le jour du recensement, l'espérance mathématique est donc $\mathrm{E}(m_a) = -3481$.

On n'a pas estimé la variance de l'erreur due à une mauvaise indication de l'adresse. À notre avis, 900 serait une estimation prudente de la variance pour l'échantillon de l'EP. Par conséquent, la variance pour la région d'essai est

$$\mathrm{Var}(m_a) = (17)^2 \times 900 = 260{,}100.$$

5.3.5 Résumé

Pour le modèle de l'erreur totale, les deux premiers moments de la distribution de l'erreur due à une mauvaise indication de l'adresse au jour du recensement pour l'échantillon de l'EP sont supposés être $E(m_a) = -3481$ and $Var(m_a) = 260{,}100$.

5.4 Fabrication dans l'échantillon P

5.4.1 Source d'erreur

Les intervieweurs peuvent parfois inventer des personnes dans des unités de logement de l'échantillon P. Des études ont montré que la fabrication dans l'interview de l'EP peut introduire un biais appréciable dans les estimations de l'erreur de couverture du recensement fondées sur l'estimateur de système dual.

Essentiellement, la fabrication d'enregistrements peut avoir pour effet de réduire le taux d'appariement pour l'EP, contribuant par le fait même à gonfler l'estimation de l'erreur de couverture.

Au U.S. Bureau of the Census l'expérience montre que c'est la fabrication de ménages entiers qui pose le plus de problèmes dans les enquêtes-ménages. Il est rare que des interviewers ajoutent une personne fictive dans un ménage qui existe réellement.

Le contrôle qualitatif portant sur l'interview de l'échantillon P vise à découvrir les interviews fictives et à interviewer les vrais membres du ménage. Par conséquent, l'établissement des estimations de système dual ne comporte aucune correction statistique pour la fabrication dans l'échantillon P.

5.4.2 Définition

Les éléments N_{11} et N_{1+} de l'estimateur de système dual sont estimés à l'aide des données de l'échantillon P. Nous avons vu dans la section 4 que:

m_f = le nombre pondéré de personnes qui ont été recensées mais auxquelles on a substitué des enregistrements fabriqués;

n_{pf} = l'erreur dans N_{pf} due à la fabrication de ménages dans l'échantillon P.

Les espérances et variances a posteriori de m_f et n_{pf} sont désignées par $E(m_f)$ et $E(n_{pf})$ et $Var(m_f)$ et $Var(n_{pf})$ respectivement.

5.4.3 Mesure

Dans le TOR de 1986, le contrôle qualitatif des interviews a permis d'établir un taux de fabrication d'environ 0.6%. Un suivi exécuté après la production des estimations initiales a permis d'établir un taux de fabrication approximatif de 1.2% (Hogan et Wolter 1988).

5.4.4 Estimation

Nous allons maintenant estimer les moments des distributions a posteriori de n_{pf} et de m_f tirées de l'échantillon de l'EP. Nous jugeons raisonnable de supposer que la valeur de n_{pf} dans le TOR est négligeable. Par conséquent, l'espérance et la variance de n_{pf} sont définies $E(n_{pf}) = 0$ et $Var(n_{pf}) = 0$.

Les données du contrôle qualitatif peuvent servir à estimer r_f = le taux de fabrication des interviews dans l'échantillon P.

En cherchant les enregistrements du recensement qui se rapportent aux personnes de l'échantillon P qui, selon le contrôle qualitatif, n'ont pas été interviewées alors qu'elles auraient dû l'être, on obtient

r_{fm} = le taux d'appariement pour les personnes qui n'ont pas été interviewées parce que leur ménage a fait l'objet de fabrication dans l'échantillon P.

Comme rien n'a été consigné dans le TOR, on ne peut connaître l'identité des personnes qui n'ont pas été interviewées telles que repérées au cours du contrôle qualitatif. En conséquence, aucun appariement n'a été tenté. Puisqu'il n'existe pas de données permettant d'estimer

directement r_{fm}, nous supposons avec une certaine prudence qu'il existe une similitude entre les personnes qui ont été réellement interviewées et celles qui ne l'ont pas été mais qui auraient dû l'être. C'est pourquoi nous disons que r_{fm} est égal au taux d'appariement global pour l'échantillon P.

Les résultats du suivi exécuté après la production des estimations initiales nous permettent de conclure à un taux de fabrication d'environ 1.2%. Le taux d'appariement pour le TOR est de 88.6% (Diffendal 1988). Par conséquent, l'espérance de l'erreur m_f est $\mathrm{E}(m_f) = -2502$.

On n'a pas calculé d'estimation de la variance de l'estimateur de l'erreur due à la fabrication. À notre avis, il est possible d'obtenir une estimation prudente de la variance en appliquant les définitions de la section 5.4.2. Ainsi, la variance estimée pour la région d'essai est

$$\mathrm{Var}(m_f) = (17)^2 \times 206 = 59{,}534.$$

5.4.5 Résumé

Pour le modèle de l'erreur totale, les deux premiers moments de la distribution de l'erreur nette due à la fabrication d'interviews sont supposés être $\mathrm{E}(m_f) = -2502$ et $\mathrm{Var}(m_f) = 59{,}534$. L'erreur nette due à la fabrication de ménages dans l'échantillon P est supposée négligeable et, par conséquent, $\mathrm{E}(n_{pf}) = 0$ et $\mathrm{Var}(n_{pf}) = 0$.

5.5 Estimation du nombre d'enregistrements erronés

5.5.1 Source d'erreur

Des enregistrements peuvent se trouver parmi les enregistrements du recensement à la suite d'erreurs. C'est ce qu'on appelle des enregistrements erronés. Comme il faut estimer le nombre de personnes réellement recensées pour calculer l'estimateur de système dual, l'estimation de la population totale comporte un facteur de correction pour les enregistrements erronés. En soustrayant des chiffres du recensement le nombre estimé d'enregistrements qui ne se rapportent pas à une seule personne, on obtient une meilleure estimation du nombre de personnes réellement recensées. Cette correction est estimée au moyen de l'échantillon D dans l'EP.

Les enregistrements erronés comprennent les catégories suivantes: (1) personnes décédées avant le jour du recensement, (2) personnes nées après le jour du recensement, (3) enregistrements qui ne se rapportent pas à des personnes réelles, (4) enregistrements répétés, (5) personnes qui ont été recensées dans un autre secteur que celui où est tenté l'appariement. Le secteur de recherche comprend habituellement l'îlot où est située l'adresse en question de même que l'anneau d'îlots avoisinants.

Le type d'erreur étudiée dans la présente section survient lorsqu'on estime l'erreur de couverture dans le recensement. On commet une erreur dans l'estimation du nombre d'enregistrements erronés lorsqu'on considère un enregistrement de l'échantillon D comme erroné alors qu'il est exact ou vice-versa. Il peut donc se produire des erreurs dans les deux sens lorsqu'on estime le nombre d'enregistrements erronés.

Les enregistrements répétés et les enregistrements fabriqués sont les deux catégories d'enregistrements erronés qui ont le plus de chances de ne pas être reconnus comme tels. Nous ne considérons ici que les erreurs d'estimation se rapportant à ces deux catégories car les erreurs qui touchent les autres catégories sont ou bien sans importance ou bien considérées dans une autre section. Les erreurs d'estimation touchant les enregistrements qui concernent des personnes décédées avant le jour du recensement ou des personnes nées après ce jour ont un effet dérisoire. L'erreur d'estimation touchant les personnes qui ont été recensées dans un autre secteur que le secteur de recherche est traitée dans la section 5.6 – Conciliation des estimations des taux de surdénombrement et de sous-dénombrement bruts.

5.5.2 Définition

Le biais de l'ESD attribuable à des codes de dénombrement erronés découle de l'erreur d'estimation de N_{+1}. Lorsqu'on estime le nombre de personnes qui ont été réellement recensées $\hat{C}$, on applique un facteur de correction pour le nombre d'enregistrements erronés $\widehat{EE}$. $\widehat{EE}$ et, par conséquent, $\hat{C}$ sont calculés à l'aide des données de l'échantillon D. On commet une erreur d'estimation pour $\hat{C}$ lorsqu'on attribue le mauvais code de dénombrement à des cas de l'échantillon D. Soit

c_e = la différence entre le nombre pondéré d'enregistrements erronés considérés par erreur comme exacts et le nombre pondéré d'enregistrements exacts considérés par erreur comme erronés.

L'espérance de c_e, étant donné la valeur observée $\hat{C}$, est désignée par $E(c_e)$. La variance de c_e, étant donné la valeur observée $\hat{C}$, est désignée par $\text{Var}(c_e)$.

5.5.3 Mesure

On peut évaluer directement l'erreur administrative en soumettant un échantillon de cas à un nouvel appariement. Les autres types d'erreur, comme la répétition d'enregistrements qui découle de la violation des règles concernant le lieu du domicile le jour du recensement, peuvent être évalués par un examen des distributions de fréquence des enregistrements erronés. Cette méthode est préférable à une méthode d'évaluation directe à cause de la difficulté à obtenir des données précises dans les opérations de suivi subséquentes. Lorsque des tests confirment que l'erreur brute correspondante est à un niveau acceptable, on peut supposer que l'erreur nette est négligeable. Par exemple, si nous répartissons les enregistrements erronés selon le groupe d'âge, nous devrions normalement observer un grand nombre d'enregistrements répétés au sein des groupes de la population qui sont très mobiles et pour lesquels il y a plus de chances de ne pas respecter les règles concernant l'adresse du domicile le jour du recensement.

Dans le TOR de 1986, on a évalué simultanément les opérations de traitement de l'échantillon D et les opérations d'appariement de l'échantillon P dont il a été question dans la section 5.2.3 (Corby et Mulry 1988). Les données pour l'échantillon D, tirées du même sous-échantillon de 35 îlots, ont été traitées à nouveau.

5.5.4 Estimation

Nous allons maintenant estimer les moments de la distribution de c_e d'après l'échantillon de l'EP. Les résultats du second traitement (Hogan et Wolter 1988) indiquent un taux d'erreur nette de 0.0007 en ce qui concerne l'identification des enregistrements exacts. L'espérance de c_e est $E(c_e) = -238$. Cette estimation n'est fondée que sur les cas résolus de l'échantillon D car l'erreur d'imputation pour les cas non résolus est étudiée dans la section 5.7 - Données manquantes.

On n'a pas calculé d'estimation de la variance de l'erreur nette. À notre avis, il est possible d'établir une estimation prudente de la variance en appliquant les définitions de la section 5.2.2. Par conséquent, la variance estimée pour la région d'essai est $\text{Var}(c_e) = (17)^2 \times 14 = 4{,}046$.

5.5.5 Résumé

Pour le modèle de l'erreur totale, les deux premiers moments de la distribution a posteriori de l'erreur nette dans l'estimation du nombre d'enregistrements exacts sont supposés être $E(c_e) = -238$ et $\text{Var}(c_e) = 4{,}046$.

5.6 Conciliation des estimations des taux de surdénombrement et de sous-dénombrement bruts

5.6.1 Source d'erreur

L'échantillon D et l'échantillon P servent tous deux à évaluer les erreurs d'enregistrement dans le recensement. L'échantillon D permet de mesurer le taux de surdénombrement brut par

l'intermédiaire d'enregistrements erronés. L'échantillon P permet de mesurer le taux de sous-dénombrement brut par l'intermédiaire des personnes qui n'ont pas été recensées. Idéalement, il faudrait examiner tous les enregistrements du recensement avant de dire qu'une personne de l'échantillon P n'a pas été recensée. De même, il faudrait parcourir la liste de tous les habitants du pays avant de déterminer si un enregistrement de l'échantillon D est un enregistrement répété ou fictif. Évidemment, de telles recherches sont tout simplement impossibles dans le cadre d'une EP. Il est donc nécessaire de les circonscrire dans des limites raisonnables. Il faut du même coup conserver l'erreur nette bien qu'il faille s'attendre à une hausse des taux de surdénombrement et de sous-dénombrement bruts mesurés à cause de la limitation du secteur de recherche. Les taux de surdénombrement et de sous-dénombrement bruts doivent s'équilibrer pour égaler l'erreur de couverture nette.

Le fait de ne pas réaliser l'équilibre entre le taux de surdénombrement brut estimé et le taux de sous-dénombrement brut estimé peut fausser le nombre d'enregistrements de l'échantillon D qui sont classés par erreur dans les enregistrements erronés. Ce genre d'erreur peut créer soit un biais par excès ou un biais par défaut.

Contrairement à ce que ce fut en 1980, la notion d'équilibrage ne figure pas dans le plan de sondage de l'EP prévue pour 1990 et testée dans le TOR de 1986. Ce plan de sondage prévoit le chevauchement des échantillons P et D. Les mêmes îlots sont contenus dans les deux échantillons. Le secteur de recherche pour l'échantillon P est défini comme le secteur de recherche fondamental. On choisit le secteur de recherche pour l'échantillon D de manière à ce qu'il soit conforme au secteur de recherche pour l'échantillon P.

5.6.2 Résumé

L'erreur de géocodage dans le TOR de 1986 est jugée négligeable et n'est donc pas incluse dans le modèle de l'erreur totale. On trouvera en annexe un modèle de l'erreur d'équilibrage.

5.7 Données manquantes

5.7.1 Source d'erreur

Les échantillons P et D ne sont pas à l'abri des données manquantes. Il y a des cas dans l'échantillon D où l'on ne dispose pas des données nécessaires pour déterminer si la personne en question a été enregistrée correctement au cours du recensement. De même, il y a des cas dans l'échantillon P où, à cause d'un manque de données, on ne peut dire si la personne a été réellement recensée. Devant l'incapacité de résoudre le cas, on impute statistiquement la probabilité que la personne soit recensée.

Plusieurs situations peuvent faire qu'un code de dénombrement est indéterminé. L'interviewer peut ne pas être en mesure de réaliser l'interview avec une personne de l'échantillon P ou l'interview de rappel. Un questionnaire de l'échantillon P ou de l'échantillon D peut ne pas contenir toutes les données démographiques et les données sur le logement nécessaires à l'estimation. Même si des questionnaires renferment tous les renseignements voulus, les circonstances peuvent être à ce point obscures qu'il n'est pas possible de déterminer un code de dénombrement.

5.7.2 Mesure

Au lieu de considérer séparément les composantes c_i, m_i et n_{pi} nous évaluons l'erreur de l'ESD attribuable aux données manquantes. Notre méthode consiste à soumettre divers modèles de compensation de la non-réponse jugés acceptables à une analyse de sensibilité. Tout d'abord, avant même l'exécution de l'EP, nous choisissons une méthode jugée préférable d'imputation pour les cas non résolus des échantillons P et D. Selon les problèmes qui surgissent durant la collecte et le traitement des données de l'EP, nous pouvons proposer diverses méthodes de traitement des données manquantes jugées acceptables. On peut alors calculer l'ESD suivant ces divers modèles de compensation de la non-réponse. L'intervalle des estimations possibles

indique dans quelle mesure l'ESD est sensible à la méthode d'imputation. Par exemple si l'intervalle est étroit, on en déduit que les estimations sont robustes et que les données manquantes n'en modifient pas réellement la fiabilité.

5.7.3 Estimation

On a évalué l'effet des données manquantes sur les estimations du TOR de 1986 en étudiant l'intervalle des estimations obtenues lorsqu'on utilise, à la place de la méthode désignée, des méthodes d'imputation fondées sur d'autres hypothèses plausibles. Nous parlons ici plus précisément de diverses méthodes de traitement des cas d'interview par personne interposée, des cas de déménagement et des cas désignés comme enregistrements fictifs (Schenker 1988). Suivant ces méthodes, on a classé les cas d'interview par personne interposée de l'échantillon P dans les cas de non-interview et on a fait un redressement par pondération. Ainsi, les cas d'interview par personne interposée avaient essentiellement le même taux d'appariement que les cas d'interview sans personne interposée. En ce qui a trait aux cas de déménagement de l'échantillon P, on les a tous classés dans les cas non résolus puis on a imputé des probabilités d'appariement pour tous ceux-là au lieu d'en imputer uniquement pour les cas réellement non résolus. Ainsi, les personnes ayant déménagé avaient essentiellement le même taux d'appariement que les personnes n'ayant pas déménagé. En ce qui a trait aux enregistrements fictifs, ils ont été identifiés par suite d'un examen des cas non résolus de l'échantillon D confié à des spécialistes de l'appariement, qui ont reconnu des cas non résolus comme des cas fictifs. Cela a eu pour conséquence de hausser les taux observé et imputé d'enregistrement erroné.

Les modèles 000 et 111 du tableau 4 de l'article de Schenker donnent, respectivement, les limites supérieures et inférieures des taux de sous-dénombrement estimés. Les deux modèles diffèrent du TOR en ce que les migrants internes tiennent la place des migrants externes. Les migrants internes de l'échantillon P sont les répondants de cet échantillon qui ont déménagé entre le jour du recensement et le jour de l'interview de l'EP. Dans le TOR de 1986, les migrants internes de l'échantillon P qui venaient de l'extérieur de la région d'essai n'ont pas été inclus dans l'estimation de l'EP. Le fait d'exclure les migrants externes de l'estimation revient essentiellement à supposer qu'ils avaient le même taux de saisie dans le recensement que les personnes incluses dans l'estimation. Habituellement, les personnes ayant déménagé ont un taux de saisie moindre que celles n'ayant pas déménagé. Le modèle 000 correspond à la méthode utilisée dans le TOR tandis que le modèle 111 correspond à toutes les autres méthodes.

5.7.4 Résumé

On évalue l'effet des données manquantes sur la distribution de l'erreur totale en calculant les paramètres de la distribution du taux de sous-dénombrement selon plusieurs modèles d'imputation acceptables. Les modèles qui déterminent les limites supérieure et inférieure du taux de sous-dénombrement servent à l'analyse de l'erreur totale.

5.8 Erreur d'échantillonnage

5.8.1 Source d'erreur

L'ESD observé est exposé à l'erreur d'échantillonnage car $\hat{N}_p$, $\hat{C}$ et $\hat{M}$ sont déterminés à partir d'échantillons. La taille de l'échantillon de l'EP dépend du degré d'erreur d'échantillonnage toléré et des sommes allouées pour l'enquête. Toutes choses étant égales par ailleurs, plus l'échantillon sera grand, moins il y aura d'erreur d'échantillonnage dans les estimations. L'estimateur et le plan d'échantillonnage influent sur l'erreur d'échantillonnage. Selon le plan de l'EP réalisée dans le cadre du TOR, les données des échantillons P et D sont tirées du même échantillon d'îlots. L'échantillon P comprend toutes les personnes qui habitent dans les unités de logement contenues dans les îlots échantillonnés. L'échantillon D comprend tous les enregistrements du recensement qui se rapportent aux îlots échantillonnés. L'estimation de l'erreur d'échantillonnage tient compte de la corrélation qui tend à exister entre les cas d'ommission et les cas d'enregistrement erroné à l'intérieur des îlots et des unités de logement.

L'expérience montre que de nombreux secteurs où la population est difficile à dénombrer sont caractérisés par des taux d'ommission et d'enregistrement erroné plus élevés.

5.8.2 Mesure

On peut utiliser le modèle de randomisation appliqué couramment aux enquêtes par sondage pour estimer la variance de l'ESD. Le coefficient de variation, qui est le rapport entre la racine carrée de la variance de l'ESD observé et la moyenne de la distribution de l'ESD, renseigne sur l'erreur d'échantillonnage dont est entaché l'ESD.

L'estimateur de la variance de l'estimateur de système dual observé (Moriarity 1987), $v(\hat{N}_{++})$, est défini (sous forme d'une série de Taylor)

$$v(\hat{N}_{++}) = \hat{N}^2_{++}\,(v(\hat{N}_p)/\hat{N}^2_p + v(\hat{M})/\hat{M}^2 - 2c(\hat{N}_p,\hat{M})/\hat{N}_p\hat{M}))$$
$$+ \hat{N}^2_p\, v(\hat{E})/\hat{M}^2 + 2\hat{N}_{++}(\hat{N}_p c(\hat{E},\hat{M})/\hat{M}2 - c(\hat{E},\hat{N}_p)/\hat{M}),$$

où

$\hat{E}$ = $II_2 + \widehat{EE}$,
$v(X)$ = l'estimateur de la variance d'un estimateur X,
$c(X,Y)$ = l'estimateur de la covariance de X et Y.

Les catérogies II_2, données insuffisantes pour l'appariement, et $\widehat{EE}$, enregistrements erronés, sont traitées comme un seul groupe dans l'estimation de la variance. Les estimateurs de la variance et de la covariance reflètent l'échantillonnage d'îlots par grappe et les grappes d'îlots.

5.8.3 Estimation

L'écart type de l'estimation de système dual établie pour la région d'essai (388,040) est 3,100.37. Le coefficient de variation est de 0.008. Par conséquent, l'écart type pour le taux de sous-dénombrement net estimé est 0.7%.

5.8.4 Résumé

L'erreur d'échantillonnage est de 3,100.37 pour l'ESD du TOR et de 0.70% pour l'estimateur du taux de sous-dénombrement net.

6. MODÈLE DE L'ERREUR TOTALE

Nous allons maintenant résumer l'effet combiné des composantes d'erreur à l'aide des distributions *a posteriori* du taux de sous-dénombrement net. On estime le biais du taux de sous-dénombrement net estimé, $B(S)$, en calculant la différence entre et la moyenne de la distribution *a posteriori*. Pour construire une telle distribution, nous avons appliqué une méthode de simulation à 10,000 itérations, qui nous a permis d'obtenir des composantes d'erreur pseudo-aléatoires et de les additionner aux estimations du TOR. En utilisant les formules de la section 5.1.2, nous avons obtenu les équations suivantes:

$$\begin{aligned} N &= (\hat{N}_p - n_p) + (\hat{C} + c - (\hat{M} - m)) \\ &\quad + \theta(\hat{C} - c - (\hat{M} - m))(\hat{N}_p - n_p - (\hat{M} - m))/(\hat{M} - m) \\ &= (\hat{C} - c)(\hat{N}_p - n_p)/(\hat{M} - m) \\ &\quad + (\theta - 1)(\hat{C} - c - (\hat{M} - m))(\hat{N}_p - n_p - (\hat{M} - m))/(\hat{M} - m). \end{aligned}$$

Plusieurs distributions ont été utilisées pour reproduire diverses estimations de l'erreur d'imputation et du biais de corrélation (θ pour paramètre) et diverses distributions marginales pour les composantes - à savoir normale, gamma et uniforme.

Nous avons vu plus haut que le taux de sous-dénombrement net estimé pour la région d'essai du TOR est 8.42 avec une erreur type de 0.7. Nous avons choisi cette estimation car les composantes de l'erreur non due à l'échantillonnage n'ont été estimées que pour la région d'essai en général. Lorsqu'on construit un ESD pour chaque strate formée a posteriori et que l'on additionne ensuite ces estimateurs pour obtenir une estimation à l'échelle de la région, le taux de sous-dénombrement net estimé (en pourcentage) est 9.02.

Le tableau 6.1 donne les moyennes et les écarts types des composantes d'erreur pour l'échantillon de l'EP. Il convient de rappeler que l'ESD pour la région d'essai est 388,040, $\hat{M} = 298{,}204$, $\hat{C} = 343{,}567$, et $\hat{N}_p = 336{,}707$. Nous avons utilisé le poids d'échantillonnage global (17) dans toutes les simulations de sorte qu'il est possible de comparer les effets de diverses hypothèses comme celles touchant les valeurs du paramètre du biais de corrélation, les distributions d'erreur et les modèles d'imputation. La méthode utilisée s'applique aussi dans les cas où l'on utilise un poids d'échantillonnage différent dans chaque strate.

Le tableau 7 indique les effets de chaque catégorie d'erreur sur la distribution a posteriori du taux de sous-dénombrement lorsqu'on applique la méthode d'imputation du TOR. L'erreur d'appariement nette, l'erreur de déclaration d'adresse et l'erreur due à la fabrication sont toutes des erreurs qui influent sur la valeur de $\hat{M}$. Par conséquent, la présence d'une seule de ces catégories d'erreur suffit à rendre positif le biais de l'estimation du taux de sous-dénombrement net en pourcentage. L'erreur nette de l'échantillon D est une erreur qui influe sur $\hat{C}$. La présence de cette seule catégorie d'erreur suffit à rendre négatif le biais de l'estimation du taux de sous-dénombrement net en pourcentage. Nous avons choisi pour valeur estimée du biais de corrélation la valeur 2.7, qui est la médiane des estimations d'Ericksen et Kadane. La présence du biais de corrélation suffit à rendre négatif le biais de l'estimation du taux de sous-dénombrement net.

Tableau 6

Distributions hypothétiques des estimations d'erreur

	Moyenne	Écart type
Erreur d'appariement nette	-1831	176
Erreur de déclaration d'adresse	-3481	510
Erreur due à la fabrication	-2502	244
Erreur nette de l'échantillon D	-238	64

Tableau 7

Effets de chaque catégorie d'erreurs sur la distribution a posteriori du taux de sous-dénombrement net et biais de l'estimateur du taux de sous-dénombrement

	E(S)	Écart type	B(S)
Erreur d'appariement nette	7.86	0.06	0.56
Erreur de déclaration d'adresse	7.35	0.16	1.07
Erreur due à la fabrication	7.34	0.08	1.08
Erreur nette de l'échantillon D	8.49	0.02	-0.07
Biais de corrélation (2.7)	10.61	0.00	-2.19

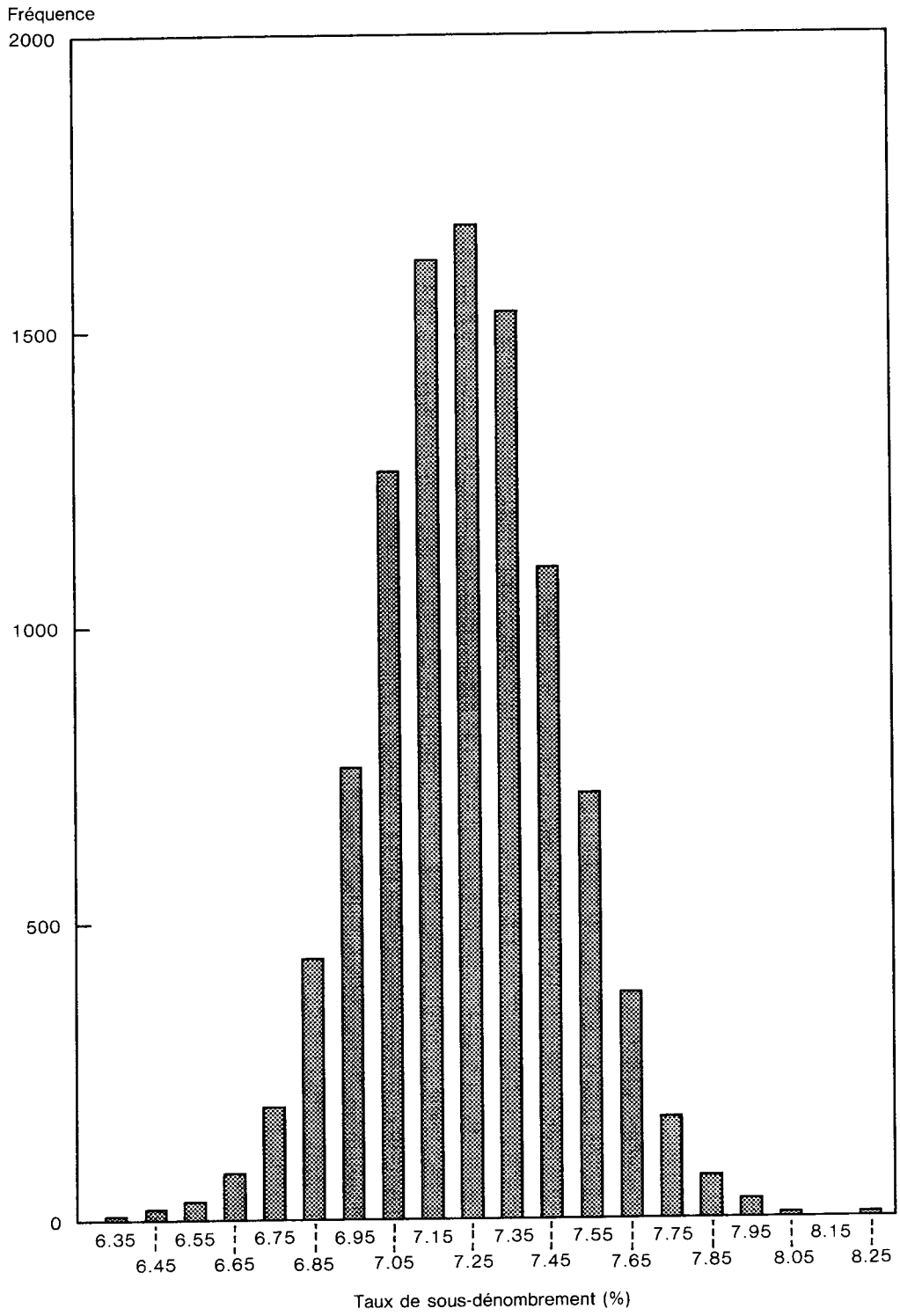

Figure 1. Taux de sous-dénombrement (en pourcentage) lorsque $\theta = 2.7$

Tableau 8

Percentiles de la distribution a posteriori du taux de sous-dénombrement net pour $\theta = 2.7$

	1	5	10	25	50	75	90	95	99
Normale	6.70	6.86	6.94	7.08	7.24	7.40	7.54	7.63	7.79
Uniforme	6.75	6.86	6.93	7.07	7.24	7.42	7.55	7.62	7.73
Gamma	6.67	6.84	6.93	7.08	7.24	7.40	7.53	7.61	7.74

Tableau 9

Distribution a posteriori du taux de sous-dénombrement net pour diverses valeurs de θ

θ	E(S)	Écart type	B(S)
1.0	5.75	0.18	2.67
2.1	6.72	0.22	1.70
2.7	7.24	0.23	1.18
3.7	8.09	0.27	0.33

Nous avons réalisé des simulations où nous tenions constants les deux premiers moments des erreurs n_p, c_e, m_m, m_f, m_a, et θ mais faisions varier les distributions. Nous avons évalué l'erreur totale lorsque les distributions d'erreur sont toutes des distributions normales, puis toutes des distributions gamma et enfin toutes des distributions uniformes. Le changement de distribution avait peu d'effet sur la distribution du taux de sous-dénombrement net. Dans chaque cas, cette distribution se rapprochait sensiblement de la distribution normale. La figure 1 montre la distribution du taux de sous-dénombrement lorsque $\theta = 2.7$, et illustre en même temps les résultats des simulations.

Le tableau 8 donne les percentiles de la distribution du taux de sous-dénombrement net pour divers types de distribution des composantes d'erreur lorsque θ est égal à 2.7 et que la méthode d'imputation du TOR est utilisée. L'écart type de la distribution a posteriori est 0.23. Dans tous les cas, la distribution normale donne une approximation acceptable. L'écart était tout au plus de 0.02 pour les percentiles 5 à 95 et tout au plus de 0.08 pour les percentiles 1 et 99.

Le fait de modifier la valeur estimée du paramètre du biais de corrélation θ influe sur les moments de la distribution a posteriori du taux de sous-dénombrement. Cette influence se manifeste dans la moyenne et l'écart type. Le tableau 9 donne les résultats pertinents pour les diverses valeurs de θ, les erreurs étant distribuées suivant une loi normale. Lorsque $\theta = 1$, il n'y a à peu près pas de biais de corrélation alors que pour les autres valeurs de θ, nous constatons l'existence de sources d'erreur. Du reste, toutes les sources d'erreur sont prises en considération lorsque $\theta = 2.1$, 2.7, et 3.7. La distribution du taux de sous-dénombrement se déplace vers la droite lorsque la valeur estimée du paramètre du biais de corrélation augmente. On observe aussi une augmentation de la variance. Le biais B(S) est positif pour toutes les valeurs de θ considérées mais diminue à mesure que celles-ci augmentent.

Pour les cas non résolus, les simulation ont été réalisées à l'aide de divers modèles d'imputation acceptables. Malgré quelques variations dans les deux premiers moments de la distribution du taux de sous-dénombrement net, l'estimation du taux de sous-dénombrement net dans le TOR présente une certaine robustesse par rapport aux données manquantes. Le tableau 10 donne les résultats des simulations faites à l'aide des modèles 000 et 111 décrits dans la section 5.7.3. Ces modèles ont permis de définir les limites supérieure et inférieure des estimations du taux de sous-dénombrement suivant tous les modèles d'imputation acceptables. Le biais de l'estimateur du taux de sous-dénombrement net varie de 0.93 à 2.79. Autrement dit, le biais représente entre 11 et 33 % du taux de sous-dénombrement net estimé, qui est de 8.42. Le changement de modèle d'imputation n'a presque pas d'effet sur l'écart type.

Tableau 10

Distribution a posteriori du taux de sous-dénombrement net selon divers modèles d'imputation acceptables lorsque $\theta = 2.7$

	E(S)	Écart type	B(S)
TOR	7.24	0.23	1.18
Modèle 000	7.49	0.23	0.93
Modèle 111	5.63	0.22	2.79

On peut estimer la variance totale du taux de sous-dénombrement net estimé en faisant la somme de la variance d'échantillonnage et de la variance non due à l'échantillonnage. Lorsque $\theta = 2.7$, l'écart type observé pour les modèles 000 et 111 (tableau 10) est 0.22, ce qui donne une variance non due à l'échantillonnage de 0.0005 lorsque toutes les erreurs sont prises en considération. L'écart type du taux de sous-dénombrement net estimé est 0.70, ce qui donne une variance d'échantillonnage de 0.49. La variance totale est donc de 0.0054 et l'erreur type de 0.73. Le coefficient de variation du taux de sous-dénombrement net est 0.083. La variance non due à l'échantillonnage a très peu de poids dans la variance totale par rapport à la variance d'échantillonnage.

7. CONCLUSIONS

Lorsqu'on applique la stratification a posteriori dans l'estimation, le taux de sous-dénombrement estimé pour le TOR est de 9.02. La stratification a posteriori a pour effet d'accroître le taux de sous-dénombrement net estimé de 0.6, ce qui est à moins d'un écart type (0.73) de la valeur estimée 8.42. Bien que la stratification a posteriori devrait normalement produire des estimations renfermant un moins haut niveau d'erreur, le résultat est conforme à l'analyse d'erreur.

Puisque nous considérons toutes les sources d'erreur dans la distribution a posteriori du taux de sous-dénombrement net, nous ne connaissons pas la distribution du paramètre du biais de corrélation θ. Nous pourrions supposer une distribution a priori pour θ, mais d'autres pourraient critiquer cette hypothèse. Si nous étions sûrs que θ est égal à 2.7, un intervalle de confiance à 95% pour le taux de sous-dénombrement net serait

$$4.77 < S < 9.55.$$

On obtient cet intervalle en prenant le taux de sous-dénombrement net estimé après stratification a posteriori (9.02) et en le corrigeant en fonction des deux biais estimés 2.79 et 0.93 du tableau 10 et de deux écarts types, 2×0.73. Nous croyons qu'il s'agit là d'une estimation prudente puisque nous utilisons deux biais estimés différents, issus respectivement des modèles d'imputation 000 et 111. Si nous considérons maintenant toutes les valeurs que peut prendre θ entre 2.1 et 3.7, un intervalle de confiance à 95% pour S serait (4.43, 10.32); ce serait là une estimation très prudente.

Nous croyons que la méthode qui vient d'être exposée dans cet article peut être appliquée dans le recensement de 1990 si on lui apporte les modifications nécessaires. Quant aux recherches futures, elles devraient porter plus spécialement sur l'estimation de l'erreur non due à

l'échantillonnage pour les strates formées a posteriori, la définition d'une distribution pour le paramètre du biais de corrélation et l'établissement de modèles pour l'erreur de déclaration adresse.

REMERCIEMENTS

Les auteurs tiennent à remercier Aref Dajani pour avoir mis au point le logiciel qui a servi aux simulations de même que Chris Moriarity pour avoir calculé les variances. Nous voudrions aussi exprimer notre reconnaissance à Kirk Wolter et Howard Hogan, qui ont contribué à l'élaboration des modèles par leurs conseils et leurs commentaires judicieux. Bruce Spencer remercie également les membres du Undercount Research Staff du U.S. Bureau of the Census pour avoir collaboré à cette étude en vertu d'une Convention sur la statistique (Joint Statistical Agreement).

ANNEXE

Définition de l'erreur d'équilibre

À cause de la non-linéarité de l'estimateur de système dual, il est difficile d'illustrer par un modèle additif les implications techniques de la conciliation des taux de surdénombrement et de sous-dénombrement bruts estimés. C'est pourquoi nous nous appliquons ici à définir un modèle multiplicatif, qui répondra mieux aux exigences de l'estimateur de système dual.

Le fait de restreindre les secteurs de recherche pour les échantillons D et P touche deux éléments de l'ESD. D'une part, cette restriction introduit un biais dans l'estimateur du nombre d'enregistrements erronés, $\widehat{EE}$. D'autre part, elle introduit un biais dans l'estimateur du nombre de personnes de l'échantillon P qui ont été recensées, $\hat{M}$.

Pour analyser les effets de cette restriction sur l'estimateur de système dual dans le cadre du Test des opérations de redressement, nous devons définir deux variables:

b = la proportion des enregistrements du recensement exacts qui se trouvent dans le secteur de recherche de l'échantillon P,

g = le rapport entre le nombre d'enregistrements du recensement exacts qui se trouvent dans le secteur de recherche de l'échantillon D et le nombre d'enregistrements exacts qui se trouvent dans le secteur de recherche de l'échantillon P.

La proportion g reflète l'erreur commise dans l'exécution de l'enquête lorsque le secteur de recherche de l'échantillon D diffère de celui de l'échantillon P. Selon la méthode utilisée pour le TOR, $g = 1$. Voyons maintenant ce qui arriverait si g n'était pas égal à 1.

À cause de la limitation du secteur de recherche, seul un pourcentage b des personnes de l'échantillon P qui ont été recensées peuvent être appariées à un enregistrement du recensement. Cela a pour effet d'introduire un biais égal à $(1 - b)\, N_{11}$ dans l'estimateur du nombre de personnes de l'échantillon P qui ont été recensées. Par conséquent, le nombre observé de personnes qui figurent à la fois dans l'échantillon P et la liste du recensement est l'estimateur réel de $b\, N_{11}$.

De même, la limitation du secteur de recherche ne permet de vérifier l'exactitude que d'un pourcentage b des enregistrements du recensement et de ce nombre, seul un pourcentage g, c'est-à-dire ceux pour lesquels le secteur de recherche coïncide avec le secteur de recherche de l'échantillon D, sera identifié comme correct. Cela a pour conséquence d'introduire un biais égal à $(1 - bg)\, N_{1+}$ dans l'estimateur du nombre de personnes qui ont été réellement recensées. Ce biais se trouve à l'origine dans l'estimateur du nombre d'enregistrements erronés, $\widehat{EE}$. Par conséquent, le nombre observé de personnes qui ont été réellement recensées est l'estimateur réel de bgN_{1+}.

Si $g = 1$, l'estimateur de système dual ne comporte aucun biais car $bgN_{+1}N_{1+}/(bN_{11}) = N_{1+}N_{+1}/N_{11}$.

En ce qui a trait à l'estimation de N_{+1} on peut définir l'erreur due au non-équilibrage par

c_b = erreur dans l'estimation du nombre d'enregistrements erronés parce que les secteurs de recherche des échantillons P et D ne coïncident

L'erreur c_b sera non nulle si g n'égale pas 1. Le rapport g peut être supérieur ou inférieur à 1. L'erreur est définie $c_b = b(g - 1)N_{+1}$.

Mesure

Dans le TOR, c_b a été mesurée lors d'un test qui visait à confirmer que l'équilibrage ne posait pas de problème et que le plan de sondage était sous contrôle. On pouvait dire que le plan de sondage était sous contrôle lorsque le pourcentage d'enregistrements appariés qui se trouvaient dans l'îlot échantillonné était élevé. Comme le plan de sondage était sous contrôle, on peut supposer que g est approximativement égal à 1 et que c_b est négligeable.

Estimation

Le géocodage semble avoir été très bien exécuté pour la région d'essai du TOR. Toutefois, on ne s'est pas attaché à évaluer de façon formelle les effets d'une erreur de géocodage sur l'estimation de $\widehat{EE}$. Par conséquent, on suppose que g est égal à 1, ce qui implique les relations $E(c_b) = 0$, et $Var(c_b) = 0$.

BIBLIOGRAPHIE

CHILDERS, D., DIFFENDAL, G., HOGAN, H., SCHENKER, N., et WOLTER, K. (1987). The technical feasibility of correcting the 1990 Census. *Proceedings of the Social Statistics Section, American Statistical Association*, 36-45.

CORBY, C., et MULRY, M. (1988). Memorandum à K.M. Wolter, Objet: Matching Error Pilot Study. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

COWAN, C.D., et MALEC, D.J. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.

DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County *Techniques d'enquête,* 14, 75-92.

ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-108, 129-131.

HOGAN, H., et MULRY, M. (1987). Operational standards for determining the accuracy of census results. *Proceedings of the Social Statistics Section, American Statistical Association*, 46-55.

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

MORIARITY, C. (1987). STSD 1986 Test Census Memorandum II-12, Objet: Documentation of the Calculation of the Los Angeles Post-Enumeration Survey Block Weights and Dual System Estimate Variances. Statistical Support Division, U.S. Bureau of the Census, Washington, D.C.

SCHENKER, N. (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986. *Techniques d'enquête,* 14, 93-104.

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 ADLAKHA, A. (1974). On the Effect of Errors in the Application of the Chandrasekaran-Deming Techniques. Reprint 14, University of North Carolina, Laboratory for Population Statistics.

SPENCER, B.D. (1986). Conceptual issues in measuring improvements in population estimates. *Proceedings of the Second Annual Research Conference,* U.S. Bureau of the Census, Washington, D.C., 393-407.

SPENCER, B.D. (1980). Implications of equity and accuracy for undercount adjustment: a decision - theoretic approach. *Proceedings of the 1980 Conference on Census Undercount*, U.S. Bureau of the Census, Washington, D.C., 204-216.

WOLTER, K.M. (1986a). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

WOLTER, K.M. (1986b). A Combined Coverage Error Model for Individuals and Housing Units. SRD Research Report Number Census/ SRD/RR-86/27, Statistical Research Division Report Series, U.S. Bureau of the Census, Washington, D.C.