Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 93–104
Statistique Canada

Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986

NATHANIEL SCHENKER[1]

RÉSUMÉ

Cet article porte sur les méthodes de traitement des données manquantes dans les enquêtes postcensitaires en vue de l'estimation de l'erreur de couverture dans le recensement; à titre d'illustration, nous analysons le test des opérations de redressement de 1986 (Diffendal 1988). Les méthodes précitées comprennent des méthodes d'imputation fondées sur le hot-deck et des modèles de régression logistique de même que des méthodes de redressement par la pondération. Nous analysons également la sensibilité des estimations de sous-dénombrement tirées du test de 1986 à la variation des modèles d'imputation.

MOTS CLÉS: Imputation; non-réponse; enquête postcensitaire; redressement par pondération; sous-dénombrement.

1. INTRODUCTION

Les données manquantes peuvent être une importante source d'incertitude dans l'estimation de l'erreur de couverture pour les recensements décennaux aux États-Unis (Freedman et Navidi 1986; Fay, Passel et Robinson 1988, Chapitre 6). Pour les recensements décennaux de 1960 et de 1980, diverses méthodes de traitement des données manquantes ont produit plusieurs estimations de l'erreur de couverture.

Le Census Bureau a appliqué de nombreux tests à des méthodes d'estimation de l'erreur de couverture pour être en mesure de traiter le problème des données manquantes et les autres problèmes qui pourraient surgir au cours du recensement décennal de 1990. Un de ces tests a été le Test des opérations de redressement (TOR) de 1986 (Diffendal 1988), qui était fondé sur le recensement de 1986 du Central Los Angeles County. Une modification des méthodes appliquées sur le terrain et du plan de sondage du TOR a permis de réduire la quantité de données manquantes par rapport à 1980 (Hogan et Wolter 1988). Malgré cela, certains problèmes persistent.

Cet article vise à décrire les problèmes qu'ont posés les données manquantes dans le TOR et la façon dont ils ont été traités dans le processus d'estimation. Dans la section 2, nous décrivons brièvement la méthode d'estimation de l'erreur de couverture appliquée dans le TOR. Dans les sections 3 à 6, nous analysons le genre et la quantité de données manquantes de même que les méthodes utilisées pour résoudre ce problème, notamment une méthode de redressement par pondération pour les cas de non-réponse au questionnaire (non-interview), une méthode d'imputation "hot-deck" pour les caractéristiques démographiques et les caractéristiques du logement manquantes ainsi qu'une méthode d'imputation fondée sur des modèles de régression logistique pour certains éléments binaires ayant trait au dénombrement dans le recensement. La section 7 présente des estimations de l'erreur de couverture calculées selon divers modèles d'imputation et diverses méthodes de traitement appliquées à des cas précis. Les estimations minimum et maximum du taux de sous-dénombrement calculées à l'aide de ces modèles ou de ces méthodes sont 8.50 et 10.16% pour les personnes d'origine hispanique,

[1] Nathaniel Schenker, Undercount Research Staff, Statistical Research Division, Bureau of the Census, Washington (DC) 20233, É.-U. Cet article est un compte rendu des recherches qui ont été faites par un membre du personnel du Census Bureau. Les opinions qui y sont exprimées sont celles de l'auteur et ne reflètent pas nécessairement la position du Census Bureau.

5.86 et 7.81% pour les personnes d'origine asiatique et 5.81 et 6.59% pour les autres personnes. Les estimations correspondantes calculées à l'aide du TOR étaient 9.85, 7.32 et 6.21% respectivement. Enfin, la section 8 présente une analyse et les conclusions.

2. ESTIMATION DE L'ERREUR DE COUVERTURE DANS LE RECENSEMENT

Diffendal (1988) expose en détail la méthode d'estimation de l'erreur de couverture du recensement utilisée dans le TOR. La présente section expose brièvement les points qu'il est nécessaire de connaître pour comprendre le reste de cet article. L'erreur de couverture a été estimée à l'aide de données d'une enquête postcensitaire (EP) menée auprès de personnes demeurant dans la région de recensement. On a commencé par prélever un échantillon d'îlots dans la région. Ensuite, chaque unité de logement contenue dans ces îlots a fait l'objet d'une enquête afin d'en connaître les occupants le jour du recensement, les occupants le jour de l'EP (et l'endroit où ces derniers demeuraient le jour du recensement) et de déterminer les caractéristiques de ces occupants.

Deux échantillons ont servi à estimer l'erreur de couverture du recensement. L'échantillon P (pour population) était composé des personnes qui demeuraient dans les îlots échantillonnés de l'EP le jour de cette enquête. On tentait d'apparier chaque membre de l'échantillon P à une personne qui avait été recensée afin de déterminer si ce membre avait lui-même été recensé; le taux d'appariement à l'intérieur de chaque domaine d'analyse servait essentiellement à estimer le taux de saisie du recensement pour ce domaine. L'échantillon D (pour dénombrement) était composé des personnes qui demeuraient dans les îlots échantillonnés de l'EP le jour du recensement; cet échantillon a servi à estimer le nombre d'enregistrements erronés (par ex., enregistrements fictifs ou redoublés) et de personnes non appariables (par ex., personnes pour lesquelles aucun nom n'était indiqué) dans le recensement à l'intérieur de chaque domaine. On tentait d'apparier chaque membre de l'échantillon D à une personne qui avait participé à l'enquête postcensitaire. Chaque cas d'appariement était considéré comme un enregistrement exact puisque l'EP indiquait que la personne devait avoir été recensée. Chaque cas de non-appariement était approfondi pour déterminer s'il s'agissait d'un enregistrement erroné ou d'un enregistrement exact qui avait été oublié dans l'EP (laquelle n'est pas censée être à l'abri des erreurs de couverture).

Si une EP était réalisée à la grandeur des États-Unis, les membres de l'échantillon P qui ont quitté le Central Los Angeles County entre le jour du recensement et le jour de l'EP participeraient quand même à cette enquête. On tenterait ensuite d'apparier ces personnes aux enregistrements du Central Los Angeles County faits le jour du recensement et on se servirait des résultats pour estimer l'erreur de couverture dans cette région. De même, les membres de l'échantillon P qui se sont établis dans le Central Los Angeles County entre le jour du recensement et le jour de l'EP contribueraient à l'estimation de l'erreur de couverture à l'extérieur de cette région. Toutefois, comme le recensement et l'EP pour le TOR ne s'appliquaient qu'au Central Los Angeles County (et non à tous les États-Unis), les personnes qui avaient quitté la région d'essai ne participaient pas à l'EP et celles qui s'y étaient établies n'étaient pas concernées. Par conséquent, les données relatives à ces deux catégories de personnes n'ont pas servi à l'estimation. (Notons toutefois que les données relatives aux personnes qui ont déménagé *sans quitter* la région d'essai ont servi à l'estimation). Cette question est analysée plus à fond dans la section 7.2.

L'estimateur "bivalent" de la taille de la population (voir Marks, Seltzer et Krotki 1974; Krotki 1978 et Wolter 1986 pour analyse et bibliographie) est défini par

$$\text{DSE} = N_p\,(\text{CEN-SUB-EE})/\text{M}, \tag{1}$$

où N_p est l'estimateur de la population de l'EP, CEN est le chiffre du recensement non rajusté, SUB est le nombre de substitutions de personnes (pour les cas de non-réponse au questionnaire) dans le recensement, EE est un estimateur du nombre d'enregistrements erronés et de personnes non appariables dans le recensement et M est l'estimateur du nombre de personnes qui font partie à la fois de la population du recensement et de la population de l'EP; les valeurs CEN et SUB sont déduites des données du recensement tandis que N_p, EE et M sont établis à partir des données des échantillons P et D. De fait, le calcul de l'estimateur bivalent équivaut à multiplier le nombre estimé d'enregistrements exacts et appariables du recensement (CEN-SUB-EE) par l'inverse du taux de saisie estimé du recensement (M/N_p).

La théorie de l'estimation "bivalente" suppose que la probabilité de saisie est la même pour tous les membres du domaine auquel s'applique l'estimateur et ce, aussi bien pour le recensement que pour l'enquête postcensitaire (Wolter 1986). Par conséquent, un groupe quelconque de personnes incluses dans le domaine n'aura pas plus de chances ni moins de chances que n'importe quel autre groupe d'être dénombré dans le recensement ou l'EP. Pour rendre cette hypothèse plus réaliste dans le TOR, on a calculé des estimateurs bivalents à l'intérieur de strates formées a posteriori en fonction des caractéristiques des personnes et du logement. Les strates formées a posteriori sont décrites dans Diffendal (1988). Il peut s'agir par exemple des locataires de sexe masculin et d'origine hispanique qui sont âgés de 30 à 44 ans et vivent dans des îlots à majorité hispanique.

En résumé, les données des échantillons P et D qu'il est nécessaire de connaître pour estimer l'erreur de couverture sont le code d'appariement (concordance vs. non-concordance) de chaque membre de l'échantillon P, le code de dénombrement (exact vs. erroné) de chaque membre de l'échantillon D et les caractéristiques des personnes et du logement pour chacun des membres des deux échantillons.

3. MÉNAGES DE L'ÉCHANTILLON P NON INTERVIEWÉS

Il se peut qu'un interviewer n'ait pas été en mesure de réaliser une interview pour une unité de logement occupée à cause, par exemple, d'un refus de répondre de la part des occupants. Sur les 5,935 unités de logement qui étaient considérées comme occupées, 32 (0.5%) ont été classées parmi les cas de non-interview. À cause de ces cas de non-interview, il n'a pas été possible d'obtenir toutes les données voulues concernant le nombre de personnes dans chaque ménage, les caractéristiques des personnes et du logement et les codes d'appariement.

Grâce au plan de sondage de l'EP (fondée sur un échantillon d'îlots), il a été facile de résoudre le problème de la non-interview de ménages de l'échantillon P. Dans chaque îlot échantillonné, les poids de sondage des ménages de non-interview ont été redistribués parmi les ménages interviewés. La repondération pour non-interview suppose essentiellement que la distribution des personnes, des caractéristiques et des codes d'appariement pour les ménages non interviewés à l'intérieur d'un îlot est la même que pour les ménages interviewés. On a utilisé cette hypothèse car les ménages d'un même îlot tendent à avoir plus de caractéristiques en commun que les ménages d'îlots différents, bien qu'il existe probablement des différences notables entre les ménages de non-interview et les ménages interviewés, surtout au point de vue de la taille (voir, par ex., Palmer 1967).

Il se peut que les données recueillies sur un ménage au moyen d'une interview par personne interposée (dans le TOR, cela signifie une interview complète réalisée avec une personne qui n'est pas membre du ménage) soient d'une qualité tellement faible que l'on est porté à classer ce ménage dans les ménages de non-interview. La qualité des données provenant des 189 interviews par personne interposée du TOR est analysée dans la section 4 tandis que la section 7 renferme quelques estimations de l'erreur de couverture établies lorsque les interviews par personne interposée sont considérées comme des non-interviews.

4. CARACTÉRISTIQUES MANQUANTES DANS LES ÉCHANTILLONS P ET D

Même lorsqu'on avait pu réalisé une interview avec un ménage de l'échantillon *P*, les données concernant les caractéristiques des personnes et du logement étaient parfois incomplètes. Il y avait aussi des données incomplètes dans le recensement et, partant, dans l'échantillon *D*.

Les variables utilisées pour la stratification a posteriori dans le TOR (Diffendal 1988) comprenaient la variable de logement Mode d'Occupation (1 = logement occupé par le propriétaire, 2 = logement loué ou occupé sans contrepartie) et les variables personnelles Sexe (1 = homme, 2 = femme), Âge (1 = 0-14, 2 = 15-29, 3 = 30-44, 4 = 45-64, 5 = 65 +) et Origine Ethnique (1 = origine hispanique, 2 = origine asiatique, 3 = autre). En outre, la variable de logement Type de Construction (1 = maison unifamiliale, 2 = immeuble à logements multiples) a servi à traiter les codes d'appariement manquants dans l'échantillon *P* et les codes de dénombrement manquants dans l'échantillon *D* (voir sections 5 et 6).

Le tableau 1 renferme les chiffres concernant les caractéristiques manquantes pour les échantillons *P* et *D* au complet et pour les cas d'interview par personne interposée de l'échantillon *P*. En ce qui concerne les échantillons *P* et *D*, le taux de données manquantes le plus élevé est de 7.0% (pour la variable Origine Ethnique dans l'échantillon *D*), tous les autres taux n'excédant pas 3.5%. Les taux de données manquantes pour les interviews par personne interposée de l'échantillon *P* sont tous plusieurs fois supérieurs aux taux observés pour l'échantillon *P* en général, quoique seule la variable Mode d'Occupation (20.2%) présente un taux de données manquantes supérieur à 10%.

Les caractéristiques manquantes pour chacun des échantillons (*P* et *D*) ont été imputées par une méthode "hot-deck" comportant deux lectures de données après que les données ont été classées selon des critères géographiques. À la première lecture, les valeurs manquantes des variables Mode d'Occupation, Type de Construction et Origine Ethnique ont été imputées à l'aide des observations les plus récentes à cause de la relation étroite qui est censée exister entre ces variables et les caractéristiques géographiques. En outre, on a établi la distribution des variables Sexe et Âge en fonction du genre de ménage (ménage à une personne ou à plusieurs personnes), de l'état matrimonial, du lien avec le chef de ménage et du sexe et de l'âge du chef de ménage en se servant de toutes les observations. À la seconde lecture, les valeurs manquantes des variables Sexe et Âge ont été imputées aléatoirement à partir des distributions établies au cours de la première lecture. Pour plus de détails sur l'imputation des caractéristiques dans le TOR, voir Schenker (1987).

En résumé, le plan de sondage de l'EP (fondée sur un échantillon d'îlots) a servi non seulement à l'élaboration d'un schéma de pondération pour non-interviews (section 3) mais aussi à l'imputation de caractéristiques qui tendent à être réunies dans un même îlot, c'est-à-dire Mode d'Occupation, Type de Construction et Origine Ethnique.

Tableau 1

Chiffres sur les caractéristiques manquantes
(pourcentages entre parenthèses) pour les échantillons *P* et *D*
et les cas d'interview par personne interposée de l'échantillon *P*

Variable	Échantillon *P* (19,552 personnes)	Échantillon *D* (20,976 personnes)	Interviews par personne interposée de l'échantillon *P* (430 personnes)
Mode d'Occupation	690 (3.5)	154 (0.7)	87 (20.2)
Type de Construction	459 (2.3)	343 (1.6)	38 (8.8)
Sexe	418 (2.1)	82 (0.4)	18 (4.2)
Âge	137 (0.7)	432 (2.1)	18 (4.2)
Origine Ethnique	155 (0.8)	1463 (7.0)	17 (4.0)

NOTA: Les 19,552 personnes de l'échantillon *P* comprennent les 430 cas d'interview par personne interposée.

5. CODES D'APPARIEMENT MANQUANTS DANS L'ÉCHANTILLON P

Sur les 19,552 enregistrements de l'échantillon *P* créés à la suite d'une interview complète, 161 (0.8%) n'avaient pas de code d'appariement pour l'estimation "bivalente". Tous ces cas non résolus sauf trois pouvaient être classés dans deux grandes catégories: d'abord, 105 ménages pour lesquels on n'a pas tenté d'appariement à cause de noms incomplets ou de caractéristiques insuffisantes, puis 53 autres ménages qui ont déménagé entre le jour du recensement et le jour de l'EP et pour lesquels il a été difficile de déterminer l'adresse le jour du recensement ou de retracer le questionnaire de recensement.

Une façon classique de pallier l'absence d'un élément binaire comme le code d'appariement est d'imputer l'une des deux valeurs possibles. Par exemple, lorsqu'on a estimé le taux de sous-dénombrement pour le recensement décennal de 1980, on a imputé un code d'appariement pour chaque cas de l'échantillon *P* non résolu en se fondant sur un cas résolu qui présentait les mêmes caractéristiques (Fay, Passel et Robinson 1988, Chapitre 6). Toutefois, une méthode différente a été appliquée dans le TOR. Après avoir imputé toutes les caractéristiques manquantes à l'aide des méthodes décrites dans la section 4, on a imputé une probabilité d'appariement pour chaque code manquant, la probabilité étant estimée au moyen d'un modèle explicite qui sera décrit un peu plus loin. Dans le dénominateur de l'expression qui définit l'estimateur bivalent (équation 1), les cas non résolus sont représentés par la somme pondérée des probabilités imputées.

On a imputé des probabilités plutôt que des valeurs binaires pour deux raisons. La première est que l'imputation de valeurs binaires aléatoires est moins efficiente que l'imputation de probabilités estimées puisqu'elle produit des estimations qui ont une variance plus élevée (voir Rubin 1987, p. 15). En deuxième lieu, comme les probabilités imputées représentent une incertitude à propos des codes d'appariement manquants, ces probabilités devraient pouvoir servir à calculer une variance due à l'imputation. Cependant, comme l'estimateur bivalent (1) est non linéaire en *M*, l'imputation d'une probabilité (ou d'une moyenne) pour chaque élément binaire manquant introduit un biais dans l'estimation (voir Rubin 1987, p. 14). Les statisticiens approfondissent actuellement l'utilisation de probabilités imputées pour des données binaires manquantes.

La méthode de régression logistique suivante a servi à imputer les probabilités d'appariement. Soit X un vecteur de variables explicatives, Y = concordance ou non-concordance, et $p = \Pr(Y = \text{concordance} \mid X)$. Le vecteur de paramètres β du modèle de régression logistique

$$\text{logit}(p) = \log[p/(1-p)] = X'\beta$$

a été estimé sur la base des données relatives aux cas résolus à l'aide des méthodes bayesiennes pour régression logistique catégorique décrites dans Rubin et Schenker (1987); ces méthodes consistent à additionner des observations fractionnaires à chaque case ou cellule dans la régression logistique puis à ajuster le modèle par les méthodes classiques du maximum de vraisemblance. Ainsi, pour le cas non résolu j, étant donné $X = xj$, la probabilité d'appariement imputée a été définie par

$$\hat{p}_j = \text{logit}^{-1}(x_j'\hat{\beta}) = \exp(x_j'\hat{\beta})/[1 + \exp(x_j'\hat{\beta})],$$

où $\hat{\beta}$ désigne l'estimateur de β. Les variables de base qui ont servi à définir X sont le Mode d'Occupation, le Type de Construction, le Sexe, l'Âge et l'Origine Ethnique, de même que les variables qui indiquent s'il s'agit d'une interview ordinaire ou d'une interview par personne interposée ou s'il s'agit d'une personne qui a déménagé ou d'une personne qui n'a pas déménagé entre le jour du recensement et le jour de l'EP.

Le tableau A1 (dans l'annexe) donne les estimations des coefficients de la régression logistique. Les coefficients élevés qui se rattachent au code d'interview et au code "déménagement"

indiquent que les probabilités d'appariement imputées sont beaucoup moins élevées dans le cas des interviews par personne interposée et des personnes ayant déménagé que dans le cas des interviews ordinaires et des personnes n'ayant pas déménagé. Cette situation s'explique peut-être en partie par la difficulté d'apparier les cas d'interview par personne interposée et les cas de déménagement plutôt que par le seul fait que les taux de saisie du recensement sont moins élevés dans ces cas. Si cela est exact, le temps est peut-être venu de considérer d'autres méthodes de traitement des données manquantes; c'est ce à quoi nous nous attachons dans la section 7.

Sur les 19,391 cas résolus de l'échantillon P, 17,018 (87.8%) se sont avérés des cas de concordance. La somme (non pondérée) des 161 probabilités d'appariement imputées était de 124.66 (pour un taux d'appariement imputé de 77.4%). Même si l'on a fait usage d'un échantillon stratifié d'îlots dans le TOR, l'estimation des paramètres de régression logistique a supposé un échantillon aléatoire simple de personnes. Afin de vérifier si l'omission de la stratification n'aurait pas créé de biais, nous avons ajusté une fois de plus le modèle de régression logistique (une fois le TOR achevé) en incluant dans X les variables indicatrices des six strates d'échantillonnage (Diffendal 1988). Nous avons refait la somme des probabilités d'appariement imputées et nous avons obtenu 124.50 (77.3%). La faible incidence de ce changement sur les estimations de l'erreur de couverture dans le recensement est démontrée dans la section 7. Dans la section 8, nous examinons les conséquences des effets de plan qui pourraient découler de la formation de grappes.

6. CODES DE DÉNOMBREMENT MANQUANTS DANS L'ÉCHANTILLON D

Sur les 20,976 cas de l'échantillon D, 3,714 ont fait ou auraient dû faire l'objet d'un suivi. Après le suivi, 979 cas (4.7% de l'effectif, 26.4% des cas devant faire l'objet d'un suivi) n'avaient toujours pas de code de dénombrement. Tous ces cas non résolus sauf neuf pouvaient être classés dans quatre grandes catégories: d'abord, 498 cas qui n'ont pas fait l'objet d'un suivi alors qu'ils auraient dû; ensuite, 257 cas où la personne qui a participé à l'interview de rappel ne connaissait pas la personne en question, puis 137 cas pour lesquels l'interview n'a pas produit assez de renseignements pour déterminer un code de dénombrement, et enfin 78 cas pour lesquels il n'a pas été possible d'obtenir une interview de rappel.

On a pallié à l'absence de codes de dénombrement dans l'échantillon D en imputant une probabilité d'enregistrement erroné pour chaque cas non résolu. Dans le terme EE de l'expression définissant l'estimateur bivalent (équation 1), les cas non résolus sont représentés par la somme pondérée des probabilités imputées. La méthode d'imputation différait peu de celle utilisée pour les codes d'appariement. Un seul changement majeur: comme l'absence de codes de dénombrement ne pouvait être constatée qu'à la suite des procédures de suivi, seuls les cas résolus par suivi ont servi à l'estimation du modèle de régression logistique. Les variables de base qui ont servi à définir X pour la régression logistique sont le Mode d'Occupation, le Type de Construction, le Sexe, l'Âge et l'Origine Ethnique, de même que les variables qui indiquent si le questionnaire du recensement concernant le ménage dont fait partie la personne visée a été retourné par la poste ou si le ménage au complet ou une partie de celui-ci n'a pu être apparié avant le suivi. Le tableau 2 (dans l'annexe) donne les estimations des coefficients de la régression logistique.

Sur les 17,262 cas qui ne nécessitaient pas de suivi, 278 (1.6%) ont été classés parmi les enregistrements erronés ou les cas non appariables. Deux mille sept cent trente-cinq cas ont été résolus par suivi et 82 de ces cas (3.0%) ont été classés dans les enregistrements erronés. La somme (non pondérée) des 979 probabilités imputées était de 21.93 (2.2%). Lorsqu'on inclut les variables indicatrices des strates d'échantillonnage dans X, la somme des probabilités imputées devient 23.58 (2.4%). Comme dans le cas de l'échantillon P, ce changement a une très faible incidence sur les estimations de l'erreur de couverture; voir section 7.

7. ESTIMATION DE L'ERREUR DE COUVERTURE SUIVANT D'AUTRES MÉTHODES DE TRAITEMENT DES DONNÉES MANQUANTES ET AUTRES SOURCES DE PROBLÈME

Dans cette section, nous analysons l'incidence d'autres méthodes de traitement des données manquantes et autres sources de problème sur les estimations de l'erreur de couverture pour les trois catégories définies par la variable Origine Ethnique (hispanique, asiatique et autre). Pour une méthode de traitement et une catégorie données, soient $\hat{N}$ la somme des estimations bivalentes pour toutes les strates formées a posteriori qui correspondent à la catégorie donnée et N_c la somme des chiffres de recensement non rajustés pour ces strates. Le taux de sous-dénombrement estimé est alors $100(1 - N_c / \hat{N})\%$.

Nous envisageons tout d'abord d'utiliser les indicateurs des strates d'échantillonnage comme variables explicatives dans les régressions logistiques appliquées aux échantillons P et D pour l'imputation des probabilités d'appariement et d'enregistrement erroné traitées dans les sections 5 et 6. Le TOR a produit des estimations de taux de sous-dénombrement sans faire intervenir ces variables explicatives; ces taux sont 9.85% pour les personnes d'origine hispanique, 7.32% pour les personnes d'origine asiatique et 6.24% pour les autres. Lorsqu'on utilise les indicateurs des strates d'échantillonnage, les taux deviennent 9.82% pour les personnes d'origine hispanique, 7.31% pour les personnes d'origine asiatique et 6.21% pour les autres. L'utilisation des indicateurs de strates d'échantillonnage ne modifie les taux initiaux que de 0.03 point de pourcentage au maximum. Néanmoins, cette utilisation sera considérée dans toutes les méthodes que nous analyserons car elle produit en principe des résultats plus précis; par exemple, on devrait obtenir des erreurs types plus justes.

7.1 Méthodes qui réduisent le taux de sous-dénombrement estimé

Le taux d'appariement pour les 375 cas d'interview par personne interposée de l'échantillon P qui ont été résolus était de 78.9% comparativement à 87.8% pour l'ensemble de l'échantillon P. Même s'il est probable que les cas d'interview par personne interposée aient eu un taux de saisie moins élevé que les cas d'interview ordinaire dans le recensement, l'écart entre les taux d'appariement est peut-être attribuable en partie à l'absence ou à l'inexactitude de données provenant d'interviews par personne interposée (voir section 4). Une façon prudente de procéder consisterait à classer les 189 interviews par personne interposée dans les non-interviews et à appliquer la méthode de pondération décrite dans la section 3; ainsi, les cas d'interview par personne interposée auraient essentiellement le même taux d'appariement que les cas d'interview ordinaire. (Notons que lorsque toutes les interviews par personne interposée sont considérées comme des non-interviews, le code d'interview ne figure plus dans le modèle de régression logistique pour l'imputation des probabilités d'appariement.)

Le taux d'appariement pour les 277 cas de déménagement (entre le jour du recensement et le jour de l'EP) de l'échantillon P qui ont été résolus était de 66.1%. Bien que l'on croit généralement que les personnes ayant déménagé ont un taux de saisie moins élevé dans le recensement que les personnes n'ayant pas déménagé, le faible taux d'appariement observé dans le cas des personnes ayant déménagé est peut-être attribuable en partie aux difficultés inhérentes à l'appariement de cette catégorie de personnes, par exemple la difficulté d'obtenir l'adresse à laquelle demeurait la personne le jour du recensement. Une façon prudente de procéder consisterait à classer tous les cas de déménagement dans les cas non résolus puis à imputer des probabilités d'appariement pour les cas non résolus à l'aide d'un modèle de régression logistique qui n'utilise pas le code "déménagement" comme variable explicative. Ainsi, les personnes ayant déménagé auraient essentiellement le même taux d'appariement que les personnes n'ayant pas déménagé.

Sur les 979 cas non résolus de l'échantillon D, 257 portaient le code d'interview de suivi $W1$, qui signifiait que le répondant ne connaissait pas la personne en question. Ce code pouvait aussi indiquer que la personne en question était un individu fictif. C'est pourquoi, après le TOR,

Tableau 2

Taux de sous-dénombrement estimés (en %) selon l'origine ethnique et suivant diverses méthodes de traitement des cas d'interview par personne interposée de l'échantillon *P*, des cas de déménagement de l'échantillon *P* et des cas de l'échantillon *D* portant le code *W*1

Indicateur de méthode (1 = autre, 0 = TOR)			Origine hispanique	Origine asiatique	Autre
Interview par personne interposée	Personne ayant déménagé	W1			
0	0	0	9.82	7.31	6.21
0	0	1	9.30	6.76	5.83
0	1	0	9.33	7.24	6.19
0	1	1	8.80	6.69	5.81
1	0	0	9.55	6.52	6.24
1	0	1	9.03	5.96	5.86
1	1	0	9.04	6.45	6.22
1	1	1	8.51	5.90	5.84

NOTA: Les indicateurs des strates d'échantillonnage ont servi de variables explicatives dans les régressions logistiques pour l'imputation des probabilités d'appariement et d'enregistrement erroné.

tous les enregistrements portant le code *W*1 ont été vérifiés par des spécialistes de l'appariement. Ceux-ci identifiaient les enregistrements qui pouvaient laisser croire par un signe quelconque (par exemple: une note de l'interviewer) qu'il s'agissait d'un cas fictif; ils en ont relevé 118. Une méthode différente de celle utilisée dans le TOR consisterait à classer les 118 cas dans les enregistrements erronés résolus avant l'imputation. Cela aurait pour effet de hausser les taux observé et imputé d'enregistrement erroné.

Le tableau 2 donne les taux de sous-dénombrement estimés selon l'origine ethnique pour le plan factoriel 2x2x2, où les facteurs indiquent si des méthodes autres que celle utilisée dans le TOR ont été appliquées aux cas d'interview par personne interposée, aux cas de déménagement et aux cas portant le code *W*1. L'écart entre l'estimation la plus faible et l'estimation la plus élevée du taux de sous-dénombrement est de 1.31 point de pourcentage pour les personnes d'origine hispanique, de 1.41 point pour les personnes d'origine asiatique et de 0.43 point pour les autres.

Il convient de souligner qu'il y a peu d'interaction entre les méthodes de traitement des cas d'interview par personne interposée, des cas de déménagement et des cas portant le code *W*1 pour chaque origine ethnique. De fait, on peut utiliser le modèle additif simple ci-dessous pour prévoir les valeurs du tableau 2 pour chaque origine ethnique:

$$\hat{Y} = \hat{\alpha}_0 + I_p\hat{\alpha}_p + I_m\hat{\alpha}_m + I_w\hat{\alpha}_w, \tag{2}$$

où $\hat{Y}$ est l'estimation prévue du taux de sous-dénombrement, I_p, I_m, et I_w sont les indicateurs de méthode (1 = autre, 0 = TOR) pour les cas d'interview par personne interposée, les case de déménagement et les cas portant le code *W*1 respectivement et $\hat{\alpha}_0$, $\hat{\alpha}_p$, $\hat{\alpha}_m$, et $\hat{\alpha}_w$, sont les estimations de paramètres données dans le tableau 3. Le paramètre α_0 est le taux de sous-dénombrement estimé lorsqu'aucune autre méthode que celle du TOR n'est utilisée; α_p, α_m, et α_w sont les effets de l'application d'autres méthodes aux cas d'interview par personne

Tableau 3

Estimations des paramètres du modèle additif (2) servant à prévoir les taux de sous-dénombrement estimés du tableau 2

	Origine hispanique	Origine asiatique	Autre
$\hat{\alpha}_o$	9.82	7.31	6.21
$\hat{\alpha}_p$	−0.28	−0.7925	0.03
$\hat{\alpha}_m$	−0.505	−0.0675	−0.02
$\hat{\alpha}_w$	−0.525	−0.5525	−0.38

interposée, aux cas de déménagement et aux cas portant le code *W*1 respectivement. Lorsqu'on utilise l'équation (2) pour prévoir les valeurs du tableau 2, le résiduel le plus élevé est 0.02%.

7.2 Méthode qui accroît le taux de sous-dénombrement estimé

Comme le TOR était limité à une très petite région des États-Unis, l'enquête postcensitaire n'a pas permis de recueillir des données sur les personnes qui avaient quitté la région d'essai après le recensement. Le fait d'exclure ces personnes de l'estimation revenait à supposer qu'elles avaient le même taux de saisie dans le recensement que les personnes incluses dans l'estimation. C'était là une hypothèse modérée puisque l'on sait que les personnes ayant déménagé ont habituellement un taux de saisie moindre que les personnes n'ayant pas déménagé.

Quatre cent neuf personnes sont venues s'établir dans la région d'essai entre le jour du recensement et le jour de l'EP. Ces nouveaux venus n'ont pas été inclus dans l'estimation puisque le jour du recensement, ils demeuraient à l'extérieur de la région d'essai et que, par conséquent, les données recueillies ce jour-là s'appliquent à d'autres régions. De plus, il n'était pas possible d'apparier ces nouveaux venus à des enregistrements du recensement puisqu'ils vivaient à l'extérieur de la région d'essai le jour du recensement.

Pour avoir une idée des conséquences qu'aurait l'inclusion des sortants dans l'estimation, on pourrait considérer à la place les 409 entrants et imputer des probabilités d'appariement pour ces personnes (puisque leurs codes d'appariement sont inconnus). Les combinaisons qui ont produit les estimations les plus élevées et les plus faibles dans le tableau 2 ont été appliquées aux données du TOR en tenant compte des entrants; les résultats pertinents figurent dans le tableau 4. On remarquera que les estimations obtenues par l'application de méthodes autres

Tableau 4

Taux de sous-dénombrement estimés (en %) selon l'origine ethnique lorsque les entrants sont considérés dans l'estimation avec probabilités d'appariement imputées

Indicateur de méthode (1 = autre, 0 = TOR)			Origine hispanique	Origine asiatique	Autre
Interview par personne interposée	Personne ayant déménagé	W1			
0	0	0	10.16	7.81	6.59
1	1	1	8.50	5.86	5.81

NOTA: Les indicateurs des strates d'échantillonnage ont servi de variables explicatives dans les régressions logistiques pour l'imputation des probabilités d'appariement et d'enregistrement erroné.

que celle utilisée dans le TOR pour les cas d'interview par personne interposée, les cas de déménagement et les cas portant le code *W*1 sont toutes inférieures aux estimations correspondantes du tableau 2 et l'écart maximum est de 0.04 point de pourcentage. Ces résultats étaient prévisibles puisque l'addition de cas ayant un taux d'appariement imputé comparable au taux d'appariement global n'est pas censée avoir beaucoup d'effet sur les estimations. Les estimations obtenues par l'application de la méthode utilisée dans le TOR sont toutes supérieures aux estimations correspondantes du tableau 2, les écarts étant de 0.34 point de pourcentage pour les personnes d'origine hispanique, de 0.50 point de pourcentage pour les personnes d'origine asiatique et de 0.38 point de poucentage pour les autres.

8. RÉSUMÉ ET ANALYSE

Une combinaison de méthodes de pondération et d'imputation (aléatoire et non aléatoire) a servi à traiter les données manquantes dans le TOR. Les cas de non-interview dans l'échantillon *P* ont été résolus grâce à une méthode de redressement par pondération au niveau de l'îlot. Une méthode hot-deck a servi à imputer les caractéristiques manquantes dans les deux échantillons. Pour pallier l'absence de codes d'appariement dans l'échantillon *P* et de codes de dénombrement dans l'échantillon *D*, on a imputé des probabilités estimées par des méthodes de régression logistique.

Comme nous l'avons indiqué dans les sections 5 et 6, l'utilisation de probabilités imputées pour les codes d'appariement et de dénombrement manquants devrait faciliter le calcul de la variance due à l'imputation de ces codes. Pour évaluer entièrement cette variance, il est nécessaire de mesurer la variance due à l'estimation des paramètres de la régression logistique de même que la variance due à l'imputation étant donné β (Rubin et Schenker 1986). Il nous faut donc une matrice estimée des variances-covariances pour $\hat{\beta}$. Comme on a procédé à un échantillonnage par grappes dans le TOR, l'estimation par régression logistique (section 5), qui suppose un échantillonnage aléatoire simple, ne produit pas une estimation juste de la matrice des variances-covariances. Cela ne posait pas vraiment de problèmes dans le TOR puisque celui-ci ne visait pas principalement à mesurer la variance due à l'imputation. En outre, pour ce qui a trait aux taux de non-réponse enregistrés dans le test, la variance due à l'incertitude entourant l'estimation de β sera vraisemblablement peu élevée par rapport à la variance due à l'imputation étant donné β (Rubin et Schenker 1986).

Bien qu'il soit possible en principe d'évaluer la variance due à l'imputation de codes d'appariement et de dénombrement au moyen de la méthode du TOR, on ne peut encore évaluer la variance due à l'imputation de caractéristiques manquantes (section 4). En fait, il serait possible d'évaluer cette variabilité en multipliant les caractéristiques d'imputation dans les échantillons *P* et *D* (Rubin 1987). Il faudrait alors calculer plusieurs estimateurs bivalents – un pour chaque série d'imputations.

Les modèles sur lesquels reposent les méthodes de pondération et d'imputation utilisées dans le TOR supposent que la probabilité qu'une variable soit manquante ne dépend pas de la valeur de celle-ci, étant donné les données observées. Une autre question ayant trait à l'imputation est de savoir quelle est la meilleure façon d'imputer simultanément des caractéristiques et des codes d'appariement (ou de dénombrement). La méthode utilisée dans le TOR (qui consiste à imputer tout d'abord les caractéristiques et ensuite les codes en fonction des caractéristiques imputées) suppose que les codes ne sont pas des variables explicatives utiles pour l'imputation de caractéristiques. Il serait peut-être plus indiqué d'adopter des modèles qui assouplissent les hypothèse du TOR. Rubin, Schafer et Schenker (1988) approfondissent la question.

Les données manquantes constituent une seule source d'erreur dans l'estimation de l'erreur de couverture. D'autres sources, comme l'erreur d'appariement et le non-respect de l'hypothèse des probabilités de saisie constantes (section 2), sont analysées dans Hogan et Wolter (1988). Après avoir évalué toutes ces sources d'erreur pour le TOR, Hogan et Wolter concluent à la supériorité du TOR par rapport au dénombrement initial.

REMERCIEMENTS

L'auteur tient à remercier Robert O'Brien, qui a calculé les estimations bivalentes utilisées dans la section 7. Il veut aussi exprimer sa reconnaissance aux arbitres pour leurs commentaires utiles et aux membres du comité d'étude sur le sous-dénombrement du Département de statistique de l'Université Harvard pour leur participation enrichissante aux discussions sur le traitement des données manquantes dans l'estimation du sous-dénombrement.

ANNEXE

RÉSULTATS DE LA RÉGRESSION LOGISTIQUE

Tableau A1

Résultats de la régression logistique pour l'échantillon *P*

Variable explicative	Codes	Coefficient estimé
Ordonnée à l'origine		1.47
Code d'interview	1 si interview ordinaire, −1 si interview par personne interposée	.36
Code "déménagement"	1 si la personne n'a pas déménagé, −1 si la personne a déménagé	.60
Mode d'occupation	1 si logement occupé par le propriétaire, −1 dans les autres cas	.46
Type de construction	1 si maison unifamiliale, −1 si immeuble à logements multiples	−.16
Sexe	1 pour homme, −1 pour femme	−.09
Âge 1	1 si de 0 à 14 ans, −1 si 65 et plus, 0 dans les autres cas	−.06
Âge 2	1 si de 15 à 29 ans, −1 si 65 ans et plus, 0 dans les autres cas	−.46
Âge 3	1 si de 30 à 44 ans, −1 si 65 ans et plus, 0 dans les autres cas	−.02
Âge 4	1 si de 45 à 59 ans, −1 si 65 ans et plus, 0 dans les autres cas	.13
Origine ethnique 1	1 si hispanique, −1 si autre, 0 si asiatique	−.14
Origine ethnique 2	1 si asiatique, −1 si autre, 0 si hispanique	.11

Tableau A2

Résultats de la régression logistique pour l'échantillon *D*

Variable explicative	Codes	Coefficient estimé
Ordonnée à l'origine		−3.45
Code de questionnaire	1 si retourné par la poste, −1 dans les autres cas	.01
Code de pré-suivi	1 si non-concordance partielle du ménage, −1 si non-concordance totale	−.20

Tableau A2 - Suite

Résultats de la régression logistique pour l'échantillon *D*

Variable explicativ	Codes	Coefficient estimé
Mode d'occupation	1 si logement occupé par le propriétaire, − 1 dans les autres cas	.36
Type de construction	1 si maison unifamiliale, − 1 si immeuble à logements multiples	.17
Sexe	1 pour homme, − 1 pour femme	.08
Âge 1	1 si de 0 à 14 ans, − 1 si 65 et plus, 0 dans les autres cas	− .30
Âge 2	1 si de 15 à 29 ans, − 1 si 65 ans et plus, 0 dans les autres cas	− .04
Âge 3	1 si de 30 à 44 ans, − 1 si 65 ans et plus, 0 dans les autres cas	− .34
Âge 4	1 si de 45 à 59 ans, − 1 si 65 ans et plus, 0 dans les autres cas	.10
Origine ethnique 1	1 si hispanique, − 1 si autre, 0 si asiatique	− .02
Origine ethnique 2	1 si asiatique, − 1 si autre, 0 si hispanique	− .38

BIBLIOGRAPHIE

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

FAY, R.E., PASSEL, J.S., et ROBINSON, J.G. (1988). *The Coverage of Population in the 1980 Census*, 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.

FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression Models for Adjusting the 1980 Census, *Statistical Science*, 1, 3-39.

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

KROTKI, K. (1978). *Developments in Dual System Estimation of Population Size and Growth*, Edmonton, The University of Alberta Press.

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

PALMER, S. (1967). On the Character and Influence of Nonresponse in the Current Population Survey, *Proceedings of the Social Statistics Section*, American Statistical Association, 73-80.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

RUBIN, D.B., SCHAFER, J.L., et SCHENKER, N. (1988). Imputation Strategies for Estimating the Undercount, *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census.

RUBIN, D.B., et SCHENKER, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse, *Journal of the American Statistical Association*, 81, 366-374.

RUBIN, D.B., et SCHENKER, N. (1987). Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior, *Sociological Methodology*, 17, 131- 144.

SCHENKER, N. (1987). Handling Missing Data in the 1986 Test of Adjustment Related Operations, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

WOLTER, K.M. (1986). Some Coverage Error Models for Census Data, *Journal of the American Statistical Association*, 81, 338-346.