Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 69-80
Statistique Canada

Suréchantillonnage géographique dans les enquêtes démographiques aux États-Unis

JOSEPH WAKSBERG, DAVID JUDKINS et JAMES T. MASSEY[1]

RÉSUMÉ

Aux États-Unis, les enquêtes démographiques polyvalentes comptent souvent parmi leurs principaux objectifs la production d'estimations de petites sous-populations définies, par exemple, selon la race, l'origine ethnique et le revenu. Le suréchantillonnage géographique est l'une des techniques fréquemment envisagées pour accroître la fiabilité des statistiques sur ces petites sous-populations (ou domaines), dans la mesure où l'information sur les îlots ou les groupes d'îlots du Bureau of the Census permet d'identifier les régions où se concentrent les sous-populations auxquelles on s'intéresse. Les auteurs passent en revue les questions relatives au suréchantillonnage géographique des régions, parallèlement à la présélection des ménages en vue d'améliorer la précision des estimations sur les petits domaines. Ils présentent les résultats d'une évaluation empirique de la réduction de la variance obtenue par suréchantillonnage géographique et évaluent la robustesse de l'efficacité de l'échantillonnage dans le temps, à mesure que les données servant à la stratification perdent de leur actualité. L'article aborde aussi la question du suréchantillonnage simultané de plusieurs petites sous-populations.

MOTS CLÉS: Échantillonnage; stratification; populations rares.

1. INTRODUCTION

Ceux qui parrainent un grand nombre de vastes enquêtes démographiques polyvalentes souhaitent des analyses distinctes sur des sous-populations définies selon la race, l'origine ethnique et le revenu. En général, les échantillons équiprobables ne sont pas suffisants pour permettre une analyse assez précise de ces domaines, si bien qu'on doit procéder à un suréchantillonnage quelconque. Cette exigence soulève des problèmes méthodologiques intéressants, car aucun registre de la population américaine n'autorise le prélèvement d'échantillons stratifiés en fonction des domaines désirés. Le Bureau of the Census garde bien des listes des logements qui identifient les sous-populations désirées, mais les chercheurs de l'extérieur n'y ont pas accès. Quand l'enquête exige une interview personnelle, les chercheurs externes sont donc contraints à recourir aux techniques d'échantillonnage aréolaire. Le Bureau lui-même procède parfois à un suréchantillonnage géographique, car ses listes ne sont mises à jour que tous les dix ans. Nous décrirons des méthodes de suréchantillonnage efficaces pour les domaines précités, dans le contexte de l'échantillonnage aréolaire.

Les données du recensement décennal américain sur la concentration de diverses caractéristiques démographiques sont largement diffusées pour de petites unités géographiques; ainsi, on possède des données sur la race et l'origine ethnique pour chaque îlot et des données sur le revenu pour chaque groupe d'îlots. (Par «îlot», on entend une zone indivisible, circonscrite des quatre côtés par une route. Les groupes d'îlots sont constitués de plusieurs îlots contigus.) On peut recourir à ces données pour améliorer à peu de frais la précision des statistiques sur des sous-populations rares en suréchantillonnant les îlots ou les groupes d'îlots dans lesquels se concentrent les membres des groupes en question et en abandonnant ou en sous-échantillonnant les personnes qui n'appartiennent pas aux groupes rares recherchés. C'est à Kish (1965, partie 4.5) qu'on doit la théorie générale applicable aux plans d'échantillonnage de ce genre. Waksberg (1973) a fait une présentation personnelle de cette théorie en l'illustrant d'exemples tirés du recensement de 1960. Kalton et Anderson (1986) ont donné d'autres exemples pertinents et analysé les méthodes connexes, ce qu'on retrouve également dans un article rédigé par Kalton dans United Nations (1993). Dans le présent article, nous étendrons les illustrations antérieures à d'autres domaines, actualiserons les résultats à 1990 et évaluerons de façon empirique la robustesse de ces méthodes dans le temps.

Nous commencerons par passer brièvement en revue les aspects associés à l'élimination et au sous-échantillonnage des personnes qui ne font pas partie des groupes visés. Ensuite, nous examinerons la théorie de la répartition optimale en vertu de laquelle les strates sont définies d'après la densité des populations rares, et nous appliquerons cette théorie à plusieurs populations rares. La partie principale de l'article consiste toutefois en une évaluation empirique de la réduction de la variance qui résulte du suréchantillonnage géographique de diverses minorités et populations rares, ainsi que de la robustesse de telles réductions dans le temps. Enfin, nous aborderons les problèmes particuliers que pose l'échantillonnage simultané de plusieurs populations rares, avant d'établir nos conclusions.

2. COÛT DE L'ENQUÊTE ET DÉCISION D'EFFECTUER UNE PRÉSÉLECTION

Soit U, l'univers visé, par exemple les particuliers ou les ménages pour lesquels on dispose d'une base de sondage, et

[1] Joseph Waksberg, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.; David Judkins, Research Triangle Institute, 5901-B Peachtree-Dunwoody Road, Suite 500, Atlanta, GA 30325, U.S.A.; James T. Massey, autrefois de Westat Inc., maintenant décédé.

soit D, une petite sous-population présentant un intérêt particulier, par exemple les personnes de race noire que l'on ne peut distinguer du reste de U lors de l'échantillonnage. Soit $\boldsymbol{Y}$, la valeur vectorielle des caractéristiques auxquelles on s'intéresse, par exemple le revenu annuel, la situation d'emploi et le nombre de consultations médicales de l'année antérieure. Certaines enquêtes n'ont d'autre objectif qu'estimer la distribution de $\boldsymbol{Y}$ dans D. Dans une enquête de ce genre, les éléments U-D identifiés au moment de la présélection des membres de U seront supprimés de l'échantillon. Un questionnaire général peu coûteux permet d'effectuer cette présélection et d'établir qui recevra le questionnaire intégral.

On désire parfois estimer la distribution de $\boldsymbol{Y}$ simultanément dans D et dans U. Dans ce cas, on gardera au moins quelques membres U-D découverts lors de l'interview de présélection afin de leur faire passer l'interview complète. Lorsqu'on recourt au suréchantillonnage géographique, l'échantillon initial comprendra un suréchantillon des membres de U-D habitant dans les régions à forte concentration de population D. Même lorsqu'on s'intéresse à l'univers U-D, on s'efforce habituellement d'éviter le suréchantillonnage de U-D là où la population D est très concentrée car la variation de la probabilité de sélection de U-D entraîne des effets exagérés du plan d'échantillonnage sur les statistiques relatives à U et à U-D. À cause des effets plus marqués du plan d'échantillonnage, la taille de l'échantillon supplémentaire de U-D n'entraîne souvent qu'une réduction négligeable de la variance des statistiques relatives à U-D. On préfère généralement utiliser les fonds supplémentaires destinés aux interviews additionnelles auprès des membres de la population U-D pour accroître la taille de l'échantillon initial.

Il est assez facile d'établir des méthodes de sous-échantillonnage qui aboutiront à l'obtention d'un échantillon équiprobable de U-D. Le sous-échantillonnage peut-être centralisé après la présélection, ou l'intervieweur peut s'en charger avant de laisser le ménage sélectionné, en se renseignant sur la composition du ménage. En effet, on a mis au point des techniques qui facilitent considérablement le sous-échantillonnage par l'intervieweur (Waksberg et Mohadjer 1991). Il n'est pas nécessaire d'apprendre à ce dernier comment effectuer un tirage aléatoire. Lors des sondages papier et crayon, l'intervieweur reçoit des instructions sur les aspects qui devront être abordés avec tel ou tel ménage avant l'interview, d'une maison à l'autre. Ces instructions sont randomisées centralement avant la présélection en vue de donner les taux de sondage souhaités. Avec l'IPAO, il est possible de programmer le sous-échantillonnage pour qu'il s'effectue automatiquement sur l'ordinateur portatif; c'est l'ordinateur qui signale à l'intervieweur quel ménage répondra au questionnaire complet et quel ménage sera rejeté consécutivement au sous-échantillonnage.

La décision de garder tous les membres de U-D échantillonnés ou de procéder à un sous-échantillonnage dépend de la taille relative de U et de U-D, de la précision des données qu'on requiert sur les deux populations et du coût relatif de l'interview intégrale et de l'interview de présélection. Soit $c\star$, le coût variable associé à l'échantillonnage d'un membre de U ainsi qu'à la collecte et au traitement de ses données et soit c', le coût variable associé à l'échantillonnage, à la sélection et à l'abandon d'un membre de U. L'équation $c = c\star/c'$ exprime le rapport entre le coût d'une interview complète et celui de l'interview de présélection. Lorsque c est beaucoup plus grand que 1, et que l'enquête porte sur U-D on devrait envisager le sous-échantillonnage même si, ce faisant, l'enquête gagne en complexité. L'interview intégrale étant par définition plus longue que l'interview de présélection, c devrait toujours être légèrement supérieur à 1. Avec les enquêtes par panel et les enquêtes longitudinales, il conviendrait d'inclure le coût des interviews subséquentes à $c\star$, de sorte que le coût de l'interview complète pourrait devenir un très grand multiple du coût de l'interview de présélection c'est-à-dire, $c >> 1$. La même remarque s'applique aux enquêtes qui nécessitent le prélèvement de spécimens, et des travaux de laboratoire onéreux, ainsi qu'aux enquêtes qui font appel à des spécialistes coûteux (un médecin par exemple) pour la collecte des données primaires. Pour ce genre d'enquête, nous recommanderions fortement de ne pas recourir seulement au suréchantillonnage géographique mais de le combiner à la présélection et au sous-échantillonnage. Quand l'enquête se fait de porte à porte et ne comporte qu'une interview donnée par un intervieweur moyen (en mesure de poser des questions et d'enregistrer les réponses mais pas de procéder à une évaluation technique ou anthropologique), c se situe souvent entre 3 et 5, valeur suffisante dans de nombreux cas pour justifier le recours au sous-échantillonnage de U-D dans les régions suréchantillonnées.

3. STRATIFICATION

Même s'il est impossible de distinguer D de U au moment de l'échantillonnage, on suppose qu'on possède certains renseignements sur la distribution de D et de U dans une série d'entités géographiques. Aux États-Unis, les îlots ou groupes d'îlots (GI) constituent des éléments géographiques naturels et le recensement décennal nous renseigne sur eux. (Avant le recensement de 1990, il n'y avait pas d'ilôts dans les régions rurales; pour le suréchantillonnage, on se servait d'unités plus importantes baptisées «secteurs de dénombrement».) Le Bureau of the Census diffuse largement les données sur la composition raciale et ethnique des îlots, ainsi que des données cartographiques qui permettent aux organismes de sondage d'identifier ces îlots lors des années subséquentes. Les données sur le revenu ne sont fournies qu'au niveau des GI.

Il est courant de stratifier les îlots ou les GI en fonction de la concentration locale de D. Ainsi, les îlots où D représente moins de 10% de la population de l'îlot pourraient constituer une strate. D'autres strates pourraient correspondre à 30% et à 60% de la population, ce qui donnerait un total de quatre strates. Le nombre optimal de strates ou de points de découpage n'a guère fait l'objet d'études empiriques. En général, le plan d'échantillonnage est d'autant plus efficace qu'il compte plus de strates, mais il arrive un moment où les complexités opérationnelles soulevées par un grand nombre de strates dépassent les gains d'efficacité. Une certaine sagesse remontant à Kish (1965), veut que la stratification

donne l'essentiel des gains quand on se limite à un nombre assez faible de strates.

4. RÉPARTITION OPTIMALE POUR UN DOMAINE

L'objectif consiste à adapter les formules générales pour la répartition optimale d'un échantillon stratifié, afin de les appliquer à la réduction de la variance attribuable au suréchantillonnage géographique. Le développement mathématique correspond essentiellement à celui de Kish (1965), selon la notation proposée par Kalton dans United Nations (1993). Supposons que la population soit divisée en un certain nombre de strates, comme nous l'avons vu précédemment. Soit N, la taille de la population totale et N_h, la taille de la population de la h-ième strate. Soit P_h, la proportion de la h-ième strate représentant les membres de D et P, la proportion globale de la population que constitue D. On peut se servir du recensement décennal le plus récent pour estimer P_h et P, ou recourir à un autre sondage important plus récent donnant les codes des îlots et/ou des GI pour chaque ménage ou particulier échantillonné, de sorte que le couplage avec le dernier recensement décennal permettra d'affecter le ménage ou le particulier à une strate.

Nous supposerons que c est constant d'une strate à l'autre, même si ce n'est pas toujours le cas. Ainsi, procéder à des interviews dans les îlots où l'on trouve une grande concentration d'Amérindiens, d'Esquimaux ou d'Aléoutes signifie presque toujours se rendre dans des endroits éloignés où le transport s'avère difficile. Toutefois, même l'estimation d'une moyenne nationale pour c soulève des difficultés dans la plupart des enquêtes. En général, on ne pourra donc pas obtenir d'estimations par strates.

On suppose aussi que la distribution de $\boldsymbol{Y}$ dans D est constante d'une strate à l'autre. Plus précisément, on présume que

$$E(\boldsymbol{Y} \mid D \text{ et } h) \equiv E(\boldsymbol{Y} \mid D) \quad \text{et que}$$

$$\text{Var}(\boldsymbol{Y} \mid D \text{ et } h) \equiv \text{Var}(\boldsymbol{Y} \mid D),$$

où la valeur et la variance attendues se rapportent à la population, pas au plan d'échantillonnage. Cette hypothèse laisse habituellement à désirer, mais puisque les éléments du vecteur des caractéristiques auxquelles on s'intéresse se comportent souvent différemment d'une strate à l'autre, il est inutile d'essayer d'être plus précis. Enfin, on suppose que les fractions d'échantillonnage sont assez petites dans toutes les strates pour qu'on néglige les facteurs de correction de la population finie.

Des hypothèses qui précèdent, il ressort que la fraction d'échantillonnage optimale de la h-ième strate de l'enquête dans laquelle on délaisse tous les éléments U-D sélectionnés est

$$f_h = k\sqrt{\frac{P_h}{P_h(c-1)+1}}, \tag{1}$$

où k représente une constante déterminée par les exigences de précision ou les contraintes budgétaires. (Voir une des sources précitées pour une preuve de (1). Cette règle est une application de la répartition de Neyman.) Si $c = 1$ (c'est-à-dire si la présélection coûte aussi cher que l'interview), on peut simplifier l'équation qui précède en $f_h \propto \sqrt{P_h}$, ce qui peut déboucher sur une répartition fort différente d'un échantillon équiprobable pour toutes les strates. Si la présélection est beaucoup moins onéreuse que l'interview (c'est-à-dire si $c >> 1$) et que D n'est pas extrêmement rare (à savoir P_h n'approche pas de zéro), cette relation donne des fraction d'échantillonnage assez uniformes ce qui se traduit par une répartition proportionnelle à la population.

Étant donné un budget fixe B, on calcule k grâce à l'équation

$$B = \sum_h N_h f_h c' \left[P_h c + (1 - P_h)\right]. \tag{2}$$

Pour obtenir un échantillon aléatoire simple n du domaine D, il suffit de sélectionner un échantillon de taille n/P, qui coûtera en tout

$$B = ncc' + \left(\frac{n}{P} - n\right)c'. \tag{3}$$

En établissant l'égalité des coûts, on peut calculer la constante de proportionnalité de (1), qui devient:

$$k = \frac{n\left(c - 1 + \frac{1}{P}\right)}{\sum_h N_h P_h \sqrt{c - 1 + \frac{1}{P_h}}}. \tag{4}$$

Pour calculer de façon réaliste les avantages d'une telle répartition, on doit admettre que les estimations de P_h utilisées pour établir la répartition seront quelque peu désuètes au moment où se déroulera l'enquête. Soit A_h, la proportion réelle de D dans la h-ième strate lors de l'échantillonnage et de la collecte des données. On suppose que P n'a pas changé, même si la distribution entre les strates varie selon A_h. En établissant $NP = N_D$ et $N_D A_h = N_{dh}$, on se rend compte que la taille réelle de l'échantillon, n_D, pour D correspond à

$$n_D = \sum_h NPA_h f_h. \tag{5}$$

Selon Kish (1965), la variance de cet échantillon sera plus élevée que celle d'un échantillon aléatoire simple de même taille, pour D. Le facteur d'extension de la variance ou l'effet du plan d'échantillonnage associé aux différents taux de sondage entre les strates est donné par l'équation bien connue

$$deff = \left(\sum_h A_h f_h\right)\left(\sum_h A_h / f_h\right). \tag{6}$$

Par conséquent, la taille *utile* de l'échantillon associé au suréchantillonnage géographique sera

$$\frac{n_D}{deff} = \frac{NP}{\left(\sum_h A_h / f_h\right)}. \quad (7)$$

Si on remplace les formules (1) et (4) dans (7), on obtient

$$\frac{n_D}{deff} = \frac{n\left(c - 1 + \frac{1}{P}\right)}{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP} \sqrt{c - 1 + \frac{1}{P_h}}\right)}. \quad (8)$$

Cette dernière formule nous permet de comparer la variance d'une statistique arbitraire sur le domaine D après suréchantillonnage géographique à la variance de la même statistique résultant d'un échantillonnage aléatoire simple pour le même coût total B. On peut récrire la formule (8) de façon algébrique afin d'établir la part de la variance de l'échantillonnage aléatoire simple supprimée par le suréchantillonnage géographique

$$\frac{\frac{\sigma^2}{n} - \frac{\sigma^2 deff}{n_D}}{\frac{\sigma^2}{n}} =$$

$$1 - \frac{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP} \sqrt{c - 1 + \frac{1}{P_h}}\right)}{\left(c - 1 + \frac{1}{P}\right)}. \quad (9)$$

Il peut sans aucun doute arriver que la réduction soit négative, bref que l'échantillonnage aléatoire simple ait une variance inférieure pour le même coût. Pareille situation survient le plus souvent avec une strate où $NPA_h >> N_h P_h$, bref une strate qu'on croyait contenir une très petite partie de D mais qui en inclut en réalité une forte proportion. Soulignons que si $P_h \equiv P$, il ne faut s'attendre à aucune réduction de la variance consécutivement au suréchantillonnage géographique. De plus, à mesure que c s'approche de l'infini pour une valeur P fixe (ce qui correspond à une présélection de moins en moins chère par rapport à l'interview intégrale), la réduction de la variance approche la valeur nulle. Étant donné les complications supplémentaires qu'entraîne un échantillon stratifié, on en déduit que pour une valeur c élevée et une valeur P moyenne, la personne qui conçoit le plan d'échantillonnage devrait envisager un échantillonnage aléatoire simple au lieu de la stratification. La valeur du suréchantillonnage augmente à mesure que P s'approche de zéro, que c approche de 1, et que D se concentre dans une strate. Lorsque la sous-population D à laquelle on s'intéresse se retrouve dans une seule strate, l'efficacité de l'échantillon s'accroît car il y a de moins en moins de membres de D dans les autres strates, où la différence est plus importante. La partie qui suit évalue de façon empirique la réduction potentielle de la variance attribuable au suréchantillonnage géographique dans diverses conditions, pour plusieurs domaines démographiques.

5. ÉVALUATION EMPIRIQUE

Il est très difficile d'évaluer l'équation (9) pour les domaines auxquels on s'intéresse. On peut extraire des données sur P_h des bandes sommaires du recensement décennal diffusées par le Bureau of the Census au niveau de l'îlot, du groupe d'îlots et du district de dénombrement. Il est donc possible de définir des strates raisonnables et d'évaluer les équations (1) à (4). En supposant que P_h est statique dans le temps, on pourrait aussi évaluer les autres équations. Néanmoins, les Américains ont tendance à déménager souvent, ce qui modifie la composition raciale et ethnique de nombreux îlots (Judkins, Massey et Waksberg 1992). Quand les membres de D déménagent dans une région où ils étaient peu nombreux avant, les avantages du suréchantillonnage géographique s'amenuisent. Puisqu'on ne souhaite pas surestimer les avantages de la méthode, nous en avons cherché une autre afin d'obtenir une estimation raisonnable de A_h à certains points dans le temps après le recensement. Le couplage des données d'îlots ou de GI entre deux recensements consécutifs pourrait constituer une solution, mais il est irréalisable. Jusqu'à présent, les îlots ont été définis et étiquetés indépendamment d'un recensement à l'autre, et on n'a pas essayé de conserver les définitions en prévision d'une analyse longitudinale. C'est pourquoi il faut recourir à d'autres sources pour estimer A_h.

Les microdonnées issues des enquêtes-ménages que poursuit présentement le Bureau of the Census s'avèrent une bonne source d'information sur A_h, pour analyser les avantages de l'échantillonnage géographique à l'égard des populations de race noire et d'origine hispanique. Plus précisément, nous nous sommes servis des données du 1988 National Health Interview Survey (NHIS) effectuée par interview. Des employés du Bureau of the Census nous ont préparé une bande spéciale donnant le code du groupe d'îlots ou du district de dénombrement de 1980 pour tous les ménages ou presque interviewés lors de l'enquête de 1988, pour les habitations construites avant 1980. (Celles construites dans les années 80 auraient été échantillonnées à partir des permis de construction plutôt que par échantillonnage aérolaire. Pour ces cas, les étiquettes des îlots et les groupes d'îlots ne sont pas attribuées aux logements échantillonnés à cause de difficultés techniques.) Les données de l'ENS de 1988 ont ensuite été appariées aux fichiers sommaires du recensement de 1980 par groupe d'îlots ou district de dénombrement, ce qui a permis la répartition des ménages de l'ENS en strates définies selon la concentration de personnes de race noire et d'origine hispanique en 1980. Des facteurs de pondération ont ensuite servis à estimer la distribution de différentes sous-populations dans les strates. (On a supposé que les habitations construites dans les années 80 se trouvaient dans la strate où les sous-populations rares comptaient le moins de représentants.) Des opérations analogues auraient pu être effectuées pour les natifs d'Asie ou

du Pacifique, les Amérindiens, les Esquimaux, les Aléoutes et les personnes à faible revenu, mais on ne l'a pas fait.

Les tableaux et les graphiques qu'on trouvera dans le reste de l'article présentent les données de diverses sources à divers moments dans le temps. Il convient de se rappeler que les données dont on s'est servi pour créer les strates ne sont pas nécessairement identiques à celles utilisées pour répartir l'échantillon. Les données qui ont permis d'évaluer l'échantillon peuvent venir d'un troisième point dans le temps ou d'une troisième source. Le présent article examine les combinaisons suivantes:

Étiquette	Source des données (stratification)	Source des données (répartition)	Source des données (évaluation)
80/80/80 GI	Recensement de 1980 (GI)	Recensement de 1980	Recensement de 1980
80/80/88 GI	Recensement de 1980 (GI)	Recensement de 1980	ENS 1988
80/88/88 GI	Recensement de 1980 (GI)	ENS 1988	ENS 1988
90/90/90 GI	Recensement de 1990 (GI)	Recensement de 1990	Recensement de 1990
90/90/90 îlot	Recensement de 1990 (îlot)	Recensement de 1990	Recensement de 1990

6. SURÉCHANTILLONNAGE DE LA POPULATION NOIRE

Le tableau 1 présente différents aspects de la ségrégation résidentielle de la population noire aux États-Unis, aspects importants à connaître lorsqu'on entreprend une enquête sur la population. Bien que la proportion de noirs habitant dans des groupes d'îlots à forte concentration de population noire (60% et plus) ait diminué entre 1980 et 1990, la ségrégation demeure très forte au sein de ce groupe. Les colonnes indiquant la population en 1988 revêtent une importance particulière, car elles illustrent la dynamique des données stratifiées dans le temps. En 1988, la proportion de noirs qui vivaient dans un groupe d'îlots où l'on comptait moins de 10% de noirs en 1980 avait doublé, de 9.7% à 20.5%. Ce phénomène a d'importantes répercussions sur l'efficacité du suréchantillonnage géographique, comme on le verra plus loin. Il est aussi intéressant de noter que la population totale des groupes d'îlots à forte concentration de personnes de race noire (c'est-à-dire où l'on recensait plus de 60% de noirs) en 1980 a diminué d'environ 2 millions d'habitants entre 1980 et 1988. La tendance résulte donc en partie de l'abandon de quelques vieux logements ou quartiers. La concentration est plus élevée au niveau de l'îlot qu'à celui du groupe d'îlots, en 1990, ainsi qu'on s'y attendait. (Les données sur les îlots ne sont plus disponibles pour les États-Unis à partir de 1980.) Quoique l'échantillonnage d'îlots coûte légèrement plus cher que l'échantillonnage de groupes d'îlots (à cause du plus grand nombre d'îlots et de la nécessité de prendre certaines dispositions quand l'îlot compte moins d'habitants que la taille idéale de la grappe), elle permet de mieux cibler le domaine auquel on s'intéresse.

La figure 1 résume les effets des données sur la densité de population qui apparaissent au tableau 1 à l'égard du suréchantillonnage de la population noire. On constate l'effet sensible de c sur l'efficacité du suréchantillonnage géographique. Quand c dépasse 20, la meilleure façon d'échantillonner la population noire consiste vraisemblablement à sélectionner un échantillon équiprobable.

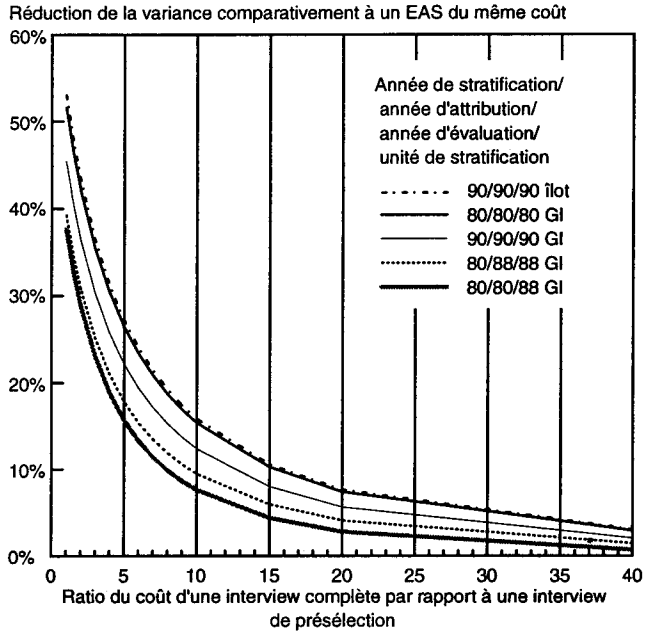

Figure 1. Réduction de la variance par suréchantillonnage géographique de la population noire

La figure montre aussi le risque qu'on court en ne se fiant qu'aux données stratifiées pour évaluer les avantages du suréchantillonnage géographique. La ligne 80/80/80 indique la réduction de la variance qu'on pourrait obtenir si la distribution de la population noire ne variait pas dans le temps selon les strates de densité définies au moyen des données sur les groupes d'îlots de 1980. La ligne 80/80/88 révèle la réduction réelle de la variance obtenue en 1988 grâce aux mêmes strates et à la même répartition. À $c = 5$, la variance diminue de 26% avec une distribution statique, alors qu'elle ne baisse que de 16% si la distribution change. Nous avons essayé de déterminer si la répartition de l'échantillon entre les anciennes strates, en fonction de nouvelles données sur la distribution réduirait la variance davantage en 1988. La réponse est oui, mais pas de beaucoup. La ligne 80/88/88 signale la réduction de la variance réalisable quand on applique la distribution de 1988 aux strates de 1980 et s'en sert pour une enquête en 1988. À $c = 5$, la nouvelle répartition réduit la variance de 18%, modeste amélioration par rapport à la diminution de 16% que permet la répartition selon l'ancienne distribution. Nous en avons conclu que la principale difficulté résidait dans l'ancienne stratification. En 1988, un si grand nombre de noirs avaient quitté les groupes

Tableau 1
Ségrégation résidentielle de la population noire

Strate de densité (pourcentage de Noirs dans l'unité de stratification l'année de la stratification)	Pourcentage de Noirs vivant dans la strate l'année indiquée				Pourcentage de la population totale vivant dans la strate l'année indiquée			
Année de mesure	1980	1988	1990	1990	1980	1988	1990	1990
Année de stratification	1980	1980	1990	1990	1980	1980	1990	1990
Unité de stratification	GI/DD	GI/DD	GI	Îlot	GI/DD	GI/DD	GI	Îlot
< 10%	9.7	20.5	12.0	8.5	78.2	81.4	75.7	77.5
10-30%	13.5	13.2	16.8	13.9	8.9	7.1	11.4	9.6
30-60%	18.9	20.4	20.3	16.2	5.1	5.1	5.7	4.5
60-100%	57.9	45.9	51.0	61.4	7.8	6.4	7.2	8.4
Population totale (milliers)	26,495	29,380	29.986	29,986	226,546	240,876	248,710	248,710
Pourcentage de Noirs aux États-Unis à l'année de la mesure	11.7	12.0	12.1	12.1				

Sources: Recensement décennal de 1980 (totalisation Westat)
Enquête nationale sur la santé de 1988 (totalisation Westat)
Recensement décennal de 1990 (totalisation Westat)

d'îlots où existait une forte concentration de population noire en 1980 pour d'autres groupes d'îlots que la réduction de la variance réalisable par suréchantillonnage avait presque diminué de moitié. La migration des noirs vers des groupes d'îlots où les gens de cette race étaient moins nombreux en 1980 augmente le nombre de noirs de l'échantillon auxquels est associé un facteur de pondération important, ce qui accroît la variabilité des poids, donc la variance. Quoi qu'il en soit, la réduction de la variance qu'indique la ligne 80/80/88 pour les valeurs $c < 10$ suffit indubitablement pour être jugée utile.

Lorsqu'on examine les données de 1990 à la figure 1, on constate que la ligne 90/90/90 des GI reste toujours quelques points sous la ligne 80/80/80, signe que le suréchantillonnage géographique des groupes d'îlots est un peu moins utile dans les années 90 que dans les années 80. On le doit à la légère diminution de la ségrégation des Américains de race noire en 1990, comparativement à la situation qui prévalait en 1980, ainsi qu'on l'a déjà souligné. D'un autre côté, la ligne 90/90/90 pour les ilôts est presque identique à la ligne 80/80/80. Le suréchantillonnage géographique des îlots devrait donc donner d'aussi bons résultats dans les années 90 que le suréchantillonnage des groupes d'îlots dans les années 80. Bien qu'on n'ait pas encore recueilli de données sur la distribution de la population noire à la fin des années 90 entre les strates de densité de 1990, il est raisonnable de croire que la migration s'est poursuivie, bref que les gains signalés par les lignes de 1990 soient plus faibles (conformément à la tendance générale que révèle la ligne 80/80/88), quand on établit des prévisions sur les économies envisageables pour la fin des années 90 et les premières années qui suivront l'an 2000.

7. SURÉCHANTILLONNAGE DES PERSONNES D'ORIGINE HISPANIQUE

Le tableau 2 présente divers aspects de la ségrégation résidentielle des personnes d'ascendance hispanique aux États-Unis qu'il est important de connaître lorsqu'on élabore une enquête sur la population. Plusieurs points valent la peine

Tableau 2
Ségrégation résidentielle des personnes d'ascendance hispanique

Strate de densité (pourcentage de personnes d'ascendance hispanique dans l'unité de stratification l'année de la stratification)	Pourcentage de personnes d'ascendance hispanique vivant dans la strate l'année indiquée				Pourcentage de la population totale vivant dans la strate l'année indiquée			
Année de mesure	1980	1988	1990	1990	1980	1988	1990	1990
Année de stratification	1980	1980	1990	1990	1980	1980	1990	1990
Unité de stratification	GI/DD	GI/DD	GI	Îlot	GI/DD	GI/DD	GI	Îlot
< 5%	14.8	29.3	10.6	6.6	76.8	79.8	68.4	68.9
5-10%	9.6	9.5	8.7	8.1	8.8	7.7	10.9	10.3
10-30%	22.6	21.2	22.8	22.1	8.5	7.4	11.8	11.5
30-60%	23.1	18.8	24.1	23.3	3.5	3.0	5.1	4.9
60-100%	30.0	21.2	33.9	39.8	2.4	2.0	3.8	4.4
Population totale (milliers)	14,609	19,393	22,354	22,354	226,546	240,876	248,710	248,710
Pourcentage de personnes d'ascendance hispanique aux États-Unis l'année de mesure	6.4	8.1	9.0	9.0				

Sources: Recensement décennal de 1980 (totalisation Westat)
Enquête nationale sur la santé de 1988 (totalisation Westat)
Recensement décennal de 1990 (totalisation Westat)

d'être mentionnés. D'abord il semble qu'à l'inverse des noirs, la ségrégation ait légèrement augmenté pour les membres de ce groupe, entre 1980 et 1990. Néanmoins, il existe des tendances analogues pour les deux populations. Ainsi, en 1980, 30% de la population d'origine hispanique vivait dans des groupes d'îlots comprenant au moins 60% de membres du même groupe. En 1988, ces groupes d'îlots ne comprenaient plus qu'environ 21% de personnes d'ascendance hispanique. En revanche, la proportion de personnes d'origine hispanique habitant un groupe d'îlots qui comptait moins de 5% de personnes de même ascendance en 1980 est passée de 15% à 29% entre 1980 et 1988. Cette évolution traduit un déplacement des personnes d'origine hispanique entre les régions, et un accroissement de la population hispanique aux États-Unis. La restratification de la population hispanique à partir des données de 1990 fait ressortir des tendances similaires à celles révélées par la distribution de 1980.

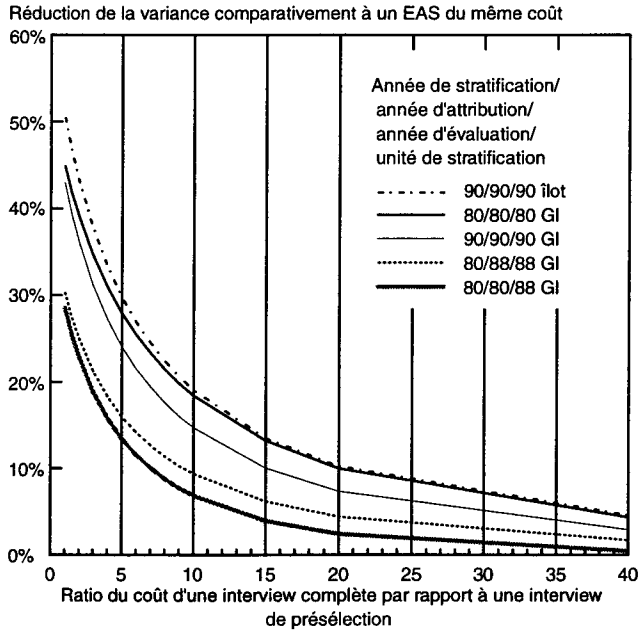

Figure 2. Réduction de la variance par suréchantillonnage géographique des personnes d'ascendance hispanique

La figure 2 donne un aperçu des répercussions de cette ségrégation sur les plans de suréchantillonnage. Les courbes suivent la même tendance générale que celles de la population noire et le suréchantillonnage géographique paraît utile pour les valeurs $c < 10$. Encore une fois cependant, il convient de se rappeler les effets de la migration sur la réduction de la variance. L'écart entre les lignes 80/80/80 et 80/80/88 est plus grand pour la population hispanique que pour la population noire, surtout aux valeurs $c < 5$. Pour l'instant, on ne dispose pas d'une base assez solide pour dire que la situation des années 80 se répétera dans les années 90.

8. SURÉCHANTILLONNAGE DES AUTRES MINORITÉS RACIALES

Les tableaux 3 et 4 présentent les données sur la ségrégation des natifs d'Asie et du Pacifique, d'une part, et des Amérindiens, des Esquimaux et des Aléoutes, d'autre part. Les figures 3 et 4 décrivent les implications correspondantes sur le suréchantillonnage de ces minorités. Les données de 1980 et de 1988 n'ont pas été reportées au tableau pour cette analyse, car les données de 1990 ne laissent pas entrevoir un suréchantillonnage bon marché des populations concernées, même avec une stratification selon la densité de population. La réduction en pourcentage de la variance est fort importante; elle dépasse celle obtenue avec les populations noire et hispanique, car la présélection nécessaire est beaucoup plus importante. Toutefois, le fait qu'il s'agisse de très petites populations aux États-Unis signifie qu'on aura toujours besoin d'une vaste présélection pour disposer d'un échantillon de taille suffisante au moment de l'interview. Ainsi, avec un ratio de coût égal à trois, il faudrait sélectionner 61,000 personnes (soit environ 24,000 ménages) pour obtenir un échantillon d'Amérindiens, d'Esquimaux et d'Aléoutes ayant la même précision qu'un échantillon aléatoire simple (théorique) de 1,000 personnes, de cette population. (Bien sûr, pour présélectionner 24,000 ménages, on devrait choisir un nombre supérieur d'unités de logement, de manière à tenir compte des logements vacants et des non-réponses). Les chiffres correspondant pour les natifs d'Asie et du Pacifique sont de 18,000 personnes, ou environ 7,000 ménages.

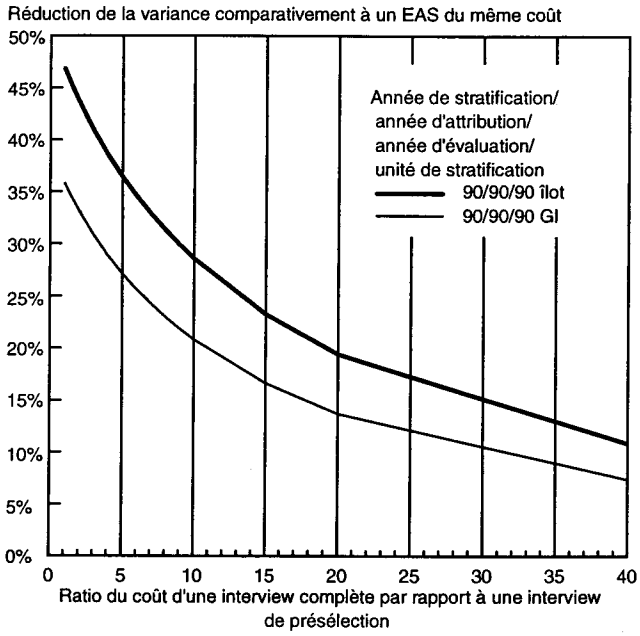

Figure 3. Réduction de la variance par suréchantillonnage géographique des natifs d'Asie et du Pacifique

Tableau 3
Ségrégation résidentielle des natifs d'Asie et du Pacifique

Strate de densité (pourcentage de natifs d'Asie et du Pacifique dans l'îlot ou le groupe d'îlots en 1990)	Pourcentage de natifs d'Asie et du Pacifique vivant dans la strate en 1990		Pourcentage de la population totale vivant dans la strate en 1990	
Unité de stratification:	GI	Îlot	GI	Îlot
< 5%	30.5	19.4	86.4	85.2
5-10%	17.2	17.7	7.2	7.4
10-30%	27.8	32.1	5.0	5.7
30-60%	14.6	18.0	1.0	1.3
60-100%	9.8	13.0	0.4	0.5
Population totale (milliers)	6,968	6,968	248,710	248,710
Pourcentage de natifs d'Asie et du Pacifique aux États-Unis l'année de mesure	2.8	2.8		

Source: Recensement décennal de 1990 (totalisation Westat)

Tableau 4
Ségrégation résidentielle des Amérindiens, des Esquimaux et des Aléoutes

Strate de densité (pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes dans l'unité de stratification en 1990)	Pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes vivant dans la strate en 1990		Pourcentage de la population totale vivant dans la strate en 1990	
Unité stratification:	GI	Îlot	GI	Îlot
< 5%	50.3	34.6	98.3	97.4
5-10%	7.4	12.1	0.8	1.4
10-30%	12.4	15.9	0.6	0.8
30-60%	6.0	7.7	0.1	0.1
60-100%	23.8	29.6	0.2	0.2
Population totale (milliers)	1,793	1,793	248,710	248,710
Pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes aux États-Unis l'année de mesure	0.7	0.7		

Source: Recensement décennal de 1990 (totalisation Westat)

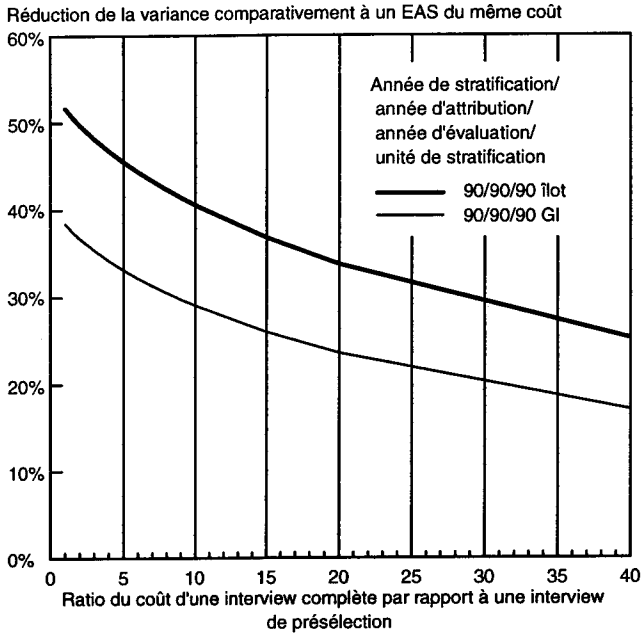

Figure 4. Réduction de la variance par suréchantillonnage géographique des Amérindiens, des Esquimaux et des Aléoutes

9. SURÉCHANTILLONNAGE DES DÉMUNIS

Le tableau 5 montre la distribution de la population défavorisée par groupe d'îlots, classée selon la population de personnes à faible revenu dans les GI, en 1990. Le nombre de GI dans chaque classe varie avec la définition de «faible revenu». Les chiffres qui apparaissent dans le tableau correspondent au pourcentage de personnes défavorisées dans chaque classe. Le tableau 5 illustre une distribution relativement uniforme de la pauvreté entre les différentes classes, pour les trois définitions retenues en 1990. Les données (non indiquées) du recensement de 1970 et du Current Population Survey révèlent que la ségrégation des personnes sous le seuil de pauvreté s'est accrue entre 1970 et 1990 (Waksberg 1995), mais elle reste beaucoup moins importante que la ségrégation des groupes ethniques et raciaux. La concentration est légèrement plus élevée pour le groupe de personnes sous la marque de 150% que pour les groupes nés des deux autres définitions mais, même pour ce groupe, elle demeure considérablement inférieure à celle observée pour les groupes raciaux et ethniques. Comme on peut le constater, avec cette définition, 25% seulement environ des démunis vivent dans des GI comptant 50% ou plus de démunis. Les pourcentages correspondants sont de 19% pour les personnes sous la marque de 125% de pauvreté et de 13% seulement pour celles sous la marque de 100% de pauvreté. Pareille distribution

signifie que le suréchantillonnage des ménages de la strate où on note un pourcentage relativement élevé de personnes à faible revenu ne donnera pas de résultats vraiment supérieurs au suréchantillonnage et à la présélection de la base de sondage complète, à moins que le coût de l'interview intégrale ne dépasse que légèrement celui de l'interview de présélection.

La figure 5 indique le ratio de la variance entre l'échantillon optimal et un échantillon aléatoire simple du même coût, pour les statistiques se rapportant aux défavorisés. Fait intéressant, en dépit de la plus forte concentration qu'entraîne la définition la plus large de «faible revenu», la réduction de la variance attribuable au suréchantillonnage géographique est plus importante avec la définition la plus étroite, car celle-ci exige une présélection accrue, si bien qu'on a plus à gagner d'une stratégie d'échantillonnage qui atténue la présélection. Dans les trois cas, le suréchantillonnage semble déboucher sur des avantages modérés quand c est inférieur à 3 ou 4, soit une réduction de 10% ou de 15% de la variance. Lorsque c atteint 10, les gains sont minimes et le suréchantillonnage des GI n'offre virtuellement aucun avantage avec une concentration élevée de démunis, pour les valeurs de c égales ou supérieures à 20. Bien sûr, il faut aussi tenir compte de la migration, mais nous n'avons pu obtenir les données voulues. La migration devrait se traduire par une réduction de la variance presque certainement plus faible que celle indiquée au tableau. En outre, les données sur le revenu du recensement de 1990 reposent sur un échantillon d'un sixième. La taille de l'échantillon dans un groupe d'îlots typique est légèrement inférieure à 100 ménages. La classification des îlots d'après la proportion de personnes à faible revenu entraîne donc un degré de confusion appréciable et beaucoup de groupes d'îlots ne se retrouveront pas dans la catégorie que leur attribuent les données du recensement, mais dans une classe voisine, ce qui atténuera encore plus la réduction de la variance que l'on peut espérer avec le suréchantillonnage géographique. Pour ces raisons, il est peu probable que le suréchantillonnage géographique débouche sur des gains d'efficacité. En réalité, la variance pourrait augmenter vers le milieu de la décennie ou plus tard. Une étude connexe, inédite, effectuée par Waksberg en 1989 aboutit à des résultats similaires quand on envisage de fusionner les données sommaires sur le revenu selon le code zip aux banques de numéros de téléphone utilisées lors de l'échantillonnage par composition aléatoire. Les gains réalisables grâce à la stratification semblent fort limités.

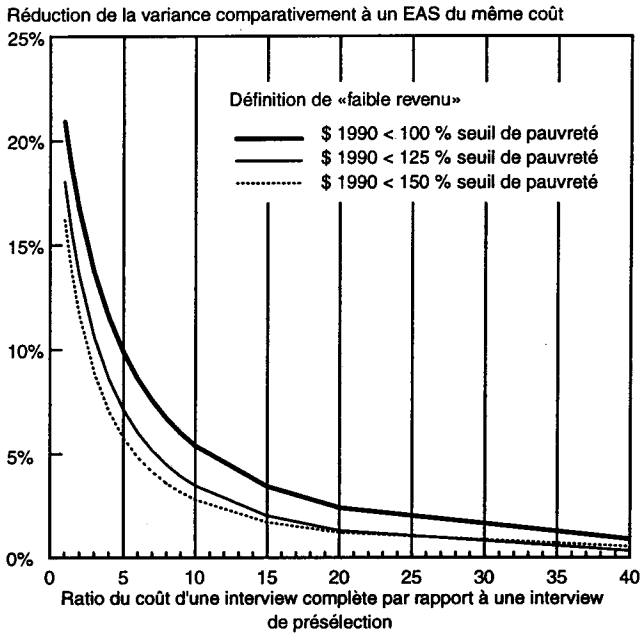

Figure 5. Réduction de la variance par suréchantillonnage géographique des démunis

Tableau 5
Ségrégation résidentielle des démunis

Strate de densité (pourcentage de démunis dans le groupe d'îlots en 1990 selon différentes définitions de «faible revenu»)	Pourcentage des démunis vivant dans la strate en 1990			Pourcentage de la population totale vivant dans la strate en 1990		
Définition de «faible revenu»:	$ < pauvreté	$ < 125% pauvreté	$ < 150% pauvreté	$ < pauvreté	$ < 125% pauvreté	$ < 150% pauvreté
< 5%	5.8	3.2	1.8	33.3	22.4	15.4
5-10%	12.3	8.3	5.7	22.1	19.7	16.7
10-20%	24.8	21 .0	16.8	22.8	25.2	24.8
20-30%	19.8	20.2	19.2	10.7	14.4	16.8
30-40%	14.3	15.9	17.0	5.4	8.1	10.7
40-50%	10.0	12.2	13.7	2.9	4.8	6.7
50-100%	13.0	19.3	25.7	2.8	5.4	8.8
Population totale (milliers)	31,797	42,316	52,521	248,710	248,710	248,710
Pourcentage de démunis aux États-Unis l'année de mesure	12.8	17.0	21.1			

Source: Recensement décennal de 1990 (totalisation Westat de STF-3)

Tableau 6
Ségrégation résidentielle des démunis selon la race et l'ascendance

Strate de densité (taux de pauvreté des personnes de la race ou de l'ascendance indiquée dans le groupe d'îlots en 1990)	Pourcentage de personnes de la race ou de l'ascendance indiquée dont le revenu était inférieur au seuil de pauvreté et qui vivaient dans la strate en 1990			Strate de densité (pourcentage de la minorité indiquée dans l'îlot en 1990)	Pourcentage de personnes de la race ou de l'ascendance indiquée dont le revenu était inférieur au seuil de pauvreté et qui vivaient dans la strate en 1990		
	Domaine				Domaine		
	Noirs	Hispaniques	Autres		Noirs	Hispaniques	Autres
< 5%	0.6	0.6	10.4	< 5%	4.0	4.6	n/a
5-10%	2.2	2.4	19.6	5-10%	3.7	5.1	n/a
10-20%	8.8	11.0	32.6	10-30%	13.2	19.9	n/a
20-30%	13.8	17.0	18.1	30-60%	19.0	25.5	n/a
30-40%	17.0	19.3	9.0	60-100%	60.0	44.8	n/a
40-50%	17.3	17.7	4.6				
50-100%	40.4	32.0	5.6				
Population totale (milliers)	8,557	5,536	17,975	Population totale (milliers)	8,557	5,536	17,975

Source: Recensement décennal de 1990 (totalisation Westat de STF-3)

L'analyse de tableaux plus détaillés (non présentés) révèle que l'efficacité est à peu près la même pour différentes ventilations géographiques, c'est-à-dire par État, par petite ou grande MSA, par noyau urbain, par banlieue ou par région non métropolitaine. Les conclusions tirées de cette analyse s'appliqueront donc approximativement aux enquêtes infranationales.

Le suréchantillonnage géographique demeure toutefois un instrument d'une très grande efficacité pour les populations noires et hispaniques à faible revenu. Comme on peut le constater au tableau 6, les personnes de race noire et celles d'ascendance hispanique qui vivent dans la pauvreté se regroupent beaucoup, à l'inverse des démunis des autres races. Le côté gauche du tableau 6 donne la distribution des défavorisés de race noire, d'origine hispanique et des autres groupes entre les strates de densité définies selon l'indice de pauvreté par rapport à la sous-population sur laquelle on se penche. En prenant un exemple du côté gauche, on se rend compte qu'en 1990, 32% des défavorisés d'origine hispanique vivaient dans un groupe d'îlots où l'indice de pauvreté pour les membres de ce groupe dépassait 50%. Le côté droit présente la distribution des défavorisés de race noire et d'ascendance hispanique pour les strates de densité définies d'après la concentration locale de noirs ou de personnes d'origine hispanique, sans égard au revenu. En prenant un exemple du côté droit, on note qu'en 1990, 44.8% des démunis d'origine hispanique habitaient un groupe d'îlots où les gens de la même ascendance représentaient plus de 60% de la population. Ces chiffres laissent entendre que plus de 90% des défavorisés de race noire et d'origine hispanique vivaient à un endroit où la concentration du groupe racial ou ethnique pertinent dépassait la moyenne. Bref, une stratégie prévoyant le suréchantillonnage des îlots à forte population de personnes de race noire ou d'ascendance hispanique débouchera automatiquement sur un nombre disproportionné de défavorisés. Par ailleurs, on ne retrouve presque aucun démuni de race noire ou d'origine hispanique dans les régions où ces deux groupes indiquent un faible indice de pauvreté. Il s'agit d'un contraste marqué avec les tendances relevées pour les défavorisés qui ne sont ni de race noire, ni d'ascendance hispanique. Apparemment, beaucoup de défavorisés de race blanche mais d'origine non hispanique vivent en voisinage avec d'autres personnes de race blanche mais nanties, peut-être parce que pour elles, la pauvreté a tendance à être un phénomène transitoire ou parce qu'il s'agit de personnes à la retraite qui avaient acheté leur maison à un moment où elles se trouvaient dans une meilleure situation.

10. SURÉCHANTILLONNAGE SIMULTANÉ DE PLUSIEURS DOMAINES RACIAUX-ETHNIQUES

En général, le suréchantillonnage géographique s'avère aussi facile et aussi efficace pour plusieurs domaines raciaux-ethniques que pour un seul. De fait, les taux d'échantillonnage optimaux pour les strates où se concentre chaque sous-population à laquelle on s'intéresse sont à peu près les mêmes que si on se penchait sur un seul domaine. Le taux de présélection global augmente néanmoins, car le nombre de régions à taux d'échantillonnage élevé augmentera avec le nombre de domaines. Ces deux observations résultent du chevauchement restreint entre les régions où il y a forte ségrégation des minorités raciales et ethniques à l'étude.

Le tableau 7 fournit quelques données sur la question, du recensement décennal de 1990. Les seules sous-populations pour lesquelles on remarque un chevauchement important dans les régions à forte concentration sont les personnes d'ascendance hispanique et les natifs d'Asie et du Pacifique. Toutefois, même dans ce cas, le chevauchement ne s'effectue que dans un sens. Comme on compte beaucoup plus de personnes d'origine hispanique que de ressortissants d'Asie et des îles du Pacifique aux États-Unis, la proportion de personnes d'ascendance hispanique habitant les îlots où les

Tableau 7
Mélange résidentiel des minorités

Strate de densité (pourcentage de la minorité indiquée dans l'îlot en 1990)	Pourcentage de Noirs vivant dans la strate en 1990			Pourcentage de personnes d'ascendance hispanique vivant dans la strate en 1990			Pourcentage de natifs d'Asie et du Pacifique vivant dans la strate en 1990			Pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes vivant dans la strate en 1990		
	Domaine de stratification			Domaine de stratification			Domaine de stratification			Domaine de stratification		
	Hispaniques	Natifs d'Asie et du Pacifique	Amérindiens, Esquimaux, et Aléoutes	Noirs	Natifs d'Asie et du Pacifique	Amérindiens, Esquimaux et Aléoutes	Noirs	Hispaniques	Amérindiens, Esquimaux et Aléoutes	Noirs	Hispaniques	Natifs d'Asie et du Pacifique
< 10%	79.2	95.4	99.6	73.4	86.3	99.1	78.9	59.2	99.6	85.9	81.4	95.1
10-30%	12.7	3.8	0.3	15.5	10.7	0.8	15.2	26.9	0.4	8.2	12.3	3.9
30-60%	5.8	0.7	0.0	7.4	2.5	0.1	4.2	10.8	0.0	3.3	4.5	0.8
60-100%	2.2	0.1	0.0	3.6	0.5	0.1	1.6	3.2	0.0	2.5	1.8	0.2

Sources: Recensement décennal de 1990 (totalisation Westat)

natifs d'Asie et du Pacifique représentent plus de 10% de la population n'est que de 13.7%, alors que le pourcentage de natifs d'Asie et du Pacifique dans les îlots où la population de personnes d'origine hispanique dépasse 10% de la population s'élève à 40.8%. Sur le plan pratique, ce chevauchement présente sans doute peu d'importance, on aurait besoin d'un échantillon de présélection très élevé (dans les régions très concentrées et à l'extérieur de celles-ci) pour trouver un nombre suffisant de natifs d'Asie et du Pacifique en vue de respecter les besoins de précision modérés, pareil échantillon ferait ressortir un nombre suffisant de personnes d'origine hispanique sans qu'on doive recourir à une répartition disproportionnée de l'échantillon entre les îlots où se concentrent davantage les personnes d'ascendance hispanique.

11. CONCLUSIONS

Le suréchantillonnage géographique reposant sur les données issues du recensement décennal le plus récent est une stratégie utile lorsqu'on veut accroître la précision des statistiques sur la population de race noire et d'origine hispanique dans les enquêtes-ménages entreprises aux États-Unis, pourvu que le coût de l'interview complète soit de 5 à 10 fois moindre que celui de l'interview de présélection. La même stratégie donne aussi de bons résultats lorsqu'il s'agit d'améliorer la précision des statistiques sur les natifs d'Asie et du Pacifique ou sur les Amérindiens, les Esquimaux et les Aléoutes, même lorsque le rapport entre le coût de l'interview intégrale et le coût de l'interview de présélection est très élevé.

Quoi qu'il en soit, ces constatations ne signifient pas qu'on puisse réaliser une enquête à un coût raisonnable en vue d'obtenir simultanément des statistiques très précises sur toutes les sous-populations désirées et de garder le degré de précision souhaité pour la population globale. La majorité des enquêtes démographiques réclament une précision raisonnable pour les domaines spécialisés et la population, en général. En déplaçant une portion de l'interview intégrale de la population blanche d'ascendance non hispanique vers les autres sous-populations, on devrait réduire la précision des statistiques sur la population totale. Il vaut habituellement la peine de parvenir à un compromis entre la précision des données sur les sous-populations et sur la population totale. Par cela, nous voulons simplement faire ressortir que le suréchantillonnage géographique ne met pas fin à la nécessité de sélectionner de très gros échantillons et d'entreprendre de nombreuses interviews de présélection quand on s'efforce de recueillir des statistiques précises sur des domaines rares, au coût le plus bas. Les statistiques précises sur des sous-populations rares continueront d'être dispendieuses qu'on recoure ou non au suréchantillonnage géographique.

En ce qui concerne les enquêtes sur les défavorisés, le suréchantillonnage géographique ne donne lieu qu'à de légers gains, et cela uniquement quand le coût de l'interview complète dépasse de plusieurs fois celui de l'interview de présélection et de l'abandon d'un ménage. Les gains devraient pour la plupart disparaître avec la détérioration due au passage du temps. En réalité, vers le milieu de la décennie ou un peu plus tard, une fois que les données du recensement seront sérieusement périmées, il se peut fort bien que le suréchantillonnage géographique entraîne une perte d'efficacité au lieu d'une amélioration, à cause de la migration des démunis et de l'erreur d'échantillonnage qui résulte de la quantification de la pauvreté au niveau des groupes d'îlots. Néanmoins, le suréchantillonnage géographique reste un instrument utile quand on s'intéresse surtout aux personnes défavorisées de race noire ou d'origine hispanique.

REMERCIEMENTS

Ce projet de recherche a été entrepris par Westat Inc. dans le cadre de l'entente 200-89-7021 parrainée par le National Center for Health Statistics et les Centers for Disease Control and Prevention. David Judkins et James Massey ont prêté leur concours au projet pendant la période où ils travaillaient respectivement à Westat et au NCHS. Les auteurs tiennent à remercier John Edmonds et Robert Dymowski, de Westat, pour leur participation à la programmation, ainsi que les arbitres pour leurs commentaires et observations sur une version antérieure de cet article.

BIBLIOGRAPHIE

JUDKINS, D., MASSEY, J., et WAKSBERG, J. (1992). Patterns of residential concentration by race and Hispanic origin. *Proceedings of the Social Statistics Section, American Statistical Association,* 51-60.

KALTON, G., et ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A,* 149, 1, 65-82.

KISH, L. (1965). *Survey Sampling.* New York: Wiley.

MASSEY, J., JUDKINS, D., et WAKSBERG. J. (1993). Collecting health data on minority populations in a national survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 75-84.

UNITED NATIONS (1993). *Sampling Rare and Elusive Populations.* Department for Economic and Social Information and Policy Analysis, Statistical Division, National Household Survey Capability Programme. New York.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association*, 429-434.

WAKSBERG, J. (1995). Distribution of poverty in Census block groups (BG's) and implications for sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 497-502.

WAKSBERG, J., et MOHADJER, L. (1991). Automation of within-household sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 350-355.