Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

OBTENTION D'ESTIMATIONS RÉGIONALES DE LA PRÉVALENCE DES FACTEURS DE RISQUE LIÉS AU CANCER : COMBINAISON DE DONNÉES PROVENANT DE LA BEHAVIORAL RISK FACTOR SURVEILLANCE SURVEY ET DE LA NATIONAL HEALTH INTERVIEW SURVEY

Michael Elliott[1]

RÉSUMÉ

La National Health Interview Survey (NHIS) et la Behavioral Risk Factors Surveillance Survey (BRFSS) fournissent des renseignements sur les facteurs de risque comportementaux, mais ni l'une ni l'autre ne permet de produire des estimations régionales idéales de leur prévalence. La NHIS est une enquête nationale par interview directe à taux de réponse élevé; cependant, les données à grandes diffusion ne comprennent pas d'identificateurs régionaux. La BRFSS est une enquête téléphonique au niveau de l'État dont sont exclus les ménages n'ayant pas le téléphone et dont le taux de réponse est faible, mais elle fournit des identificateurs au niveau du comté. Nous combinons l'information provenant de la BRFSS et de la NHIS au moyen d'estimateurs par calage, en tablant sur les points forts complémentaires d'une enquête pour compenser les faiblesses de l'autre.

MOTS CLÉS : Estimation par calage, estimation par la méthode itérative du quotient (raking) généralisée, estimation régionale, consommation de cigarettes

1. INTRODUCTION

1.1 Description

Les études de surveillance du cancer nécessitent des estimations exactes des facteurs de risque, y compris les facteurs de risque comportementaux, au niveau régional. Les facteurs de risque comportementaux étudiés incluent les caractéristiques du mode de vie (p. ex., usage du tabac, habitudes alimentaires, activité physique et obésité), le statut économique (p. ex., niveau de scolarité et revenu) et l'utilisation des services de santé (p. ex., caractéristiques de l'assurance-maladie et pratiques en matière de dépistage du cancer). Ces estimations sont utilisées comme données d'entrée dans des équations de régression de l'incidence estimée du cancer au niveau du comté aux États-Unis [(Pickle et coll., (1996) et Nandram et coll., (2000)], ou comme indicateurs avancés des variations des résultats en matière de cancer qui ont des répercussions sur les politiques au niveau tant national que régional. Par exemple, les variations des taux de dépistage du cancer aux États-Unis selon l'âge, la race, le niveau de scolarité et le revenu ont été bien décrites (Potosky et coll., 1998). Bien que l'on ait rapporté récemment l'adoption de stratégies fructueuses visant à intensifier le dépistage du cancer parmi les populations où il est sous-utilisé, le manque d'uniformité de la couverture des études portant sur les interventions a donné lieu à des lacunes. Ces lacune sont manifestes sur le plan géographique, situation qui pourrait justifier un examen plus approfondi de la nécessité de tailler également sur mesure les études sur les interventions (Legler et coll., 2002).

On obtient souvent les estimations de la prévalence des facteurs de risque d'après les données d'enquêtes comme la Current Population Survey (CPS), la National Health Interview Survey (NHIS) ou la Behavioral Risk Factors Surveillance Survey (BRFSS). Malheureusement, aucune de ces enquêtes n'est un instrument idéal pour la

[1] Center for Clinical Epidemiology and Biostatistics, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 612 Blockley Hall, 423 Guardian Drive, Philadelphia, PA USA 19104 (melliott@cceb.upenn.edu)

production d'estimations régionales. La NHIS est une enquête à taux de réponse élevé réalisée par interview sur place auprès d'un échantillon représentatif de la population nationale. Cependant, elles est conçue pour produire des estimations nationales de la prévalence, mais non des estimations au niveau de l'État ou à un niveau infra-État. Par contre, la BRFSS est une enquête téléphonique au niveau de l'État qui fournit des tailles d'échantillons raisonnables pour tous les États et pour de nombreux comtés, mais uniquement pour les ménages qui ont le téléphone, et dont le taux de réponse est nettement plus faible que celui de la NHIS.

Une stratégie en vue d'obtenir des estimations régionales consiste alors à combiner l'information provenant des deux enquêtes. Bien que diverses méthodes aient été envisagées pour combiner les données provenant d'enquêtes multiples [estimation sur base de sondage double (Hartley, 1962, 1974; Lohr et Rao 2000) et appariement statistique (Rodgers 1984; Moriarity et Scheuren 2001), entre autres], dans le présent contexte, l'exercice pose deux difficultés uniques. Premièrement, la NHIS n'offre pas d'identificateurs régionaux utilisables. Deuxièmement, la BRFSS présente vraisemblablement un biais de sélection et un biais de base de sondage relativement à la NHIS. Donc, nous proposons une nouvelle méthode « assistée par modèle » qui a l'avantage de ne pas nécessiter d'identificateurs régionaux pour l'ensemble de données d'enquête utilisé pour la calibration du biais, mais qui permet néanmoins de tirer parti de ce genre d'identificateurs régionaux s'ils existent. Le concept fondamental consiste à « faire ressembler davantage » les cas de la BRFSS à ceux de la NHIS en procédant à une recalibration des poids de sondage existants des cas de la BRFSS de sorte que les estimations pondérées calculées d'après les données de la BRFSS reflètent plus fidèlement les estimations pondérées basées sur les données de la NHIS. Nous recourrons, pour procéder à cette repondération, à l'estimation par calage. Nous appliquons cette méthode à l'estimation de la prévalence de la consommation de cigarettes en 1997 chez les hommes de 18 ans et plus aux États-Unis, au niveau du comté.

2. ESTIMATION ASSISTÉE PAR MODÈLE : RECALIBRATION DES DONNÉES DE LA BRFSS D'APRÈS CELLES DE LA NHIS

Dans la méthode d'estimation par calage, la NHIS donne un vecteur de « totaux de contrôle » X^s pour un ensemble de variables démographiques clé fournies à la fois par la NHIS et la BRFSS sur une « grande » région *S*. Par exemple, X^s pourrait être un vecteur de sommes de diverses variables démographiques et liées à la santé. Nous considérons alors les estimateurs de la forme

$$\hat{Y}_{CAL} = \sum w_i y_i$$

où les w_i satisfont la contrainte

$$\sum_{i \in S} w_i \mathbf{x}_i = \mathbf{X}^S. \qquad (2)$$

sous réserve de la minimisation d'un critère de distance entre les w_i et les poids de sondage a_i donnés par $D(w_i, a_i)$.

Deville et Sarndal (1992) décrivent une gamme d'options pour $D(w_i, a_i)$. Si $D(w_i, a_i) = (1/2)\sum (w_i - a_i)^2 / a_i$, nous obtenons l'estimateur de régression généralisée ou estimateur GREG (Cassel, Sarndal, et Wretman 1976; Isaki et Fuller 1982). Pour l'estimateur GREG, il est possible d'obtenir des solution analytiques (fermées) de (2) sous la contrainte $D(w_i, a_i)$, soit :

$$w_{i,GREG} = a_i \left\{1 + \left(\mathbf{X} - \hat{\mathbf{X}}\right)\mathbf{T}^{-1}\mathbf{x}_i\right\}$$

où $\mathbf{T} = \sum_{i \in S} a_i \mathbf{x}_i \mathbf{x}_i^T / c_i$ et $\hat{\mathbf{X}} = \sum_{i \in S} a_i \mathbf{x}_i$, où c_i est une constante précisée. S'il existe *p* variables nominales de contrôle chacune de dimension I_k , et que $\mathbf{X}^s$ est donné par les $I_1 \times \cdots \times I_p$ dénombrements de cellule pour un tableaux de contingence complet, alors $\hat{Y}_{GREG}$ est simplement l'estimateur BRFSS stratifié a posteriori sur les estimations des totaux de population de la NHIS.

De façon plus générale, on peut utiliser un algorithme itératif pour résoudre (2) sous la contrainte $D(w_i, a_i)$. Le critère de minimisation de la distance entre les w_i et les poids de sondage a_i de la BRFSS donne

$$d_i(w_i, a_i) - \mathbf{x}_i^T \boldsymbol{\lambda} = 0 \qquad (3)$$

où $d_i(w_i, a_i) = \dfrac{\partial D(w, a)}{w_i}$ et $\boldsymbol{\lambda}$ représente le vecteur des multiplicateurs de Lagrange induits par la contrainte (2). Quand $D(w, a) = \sum \left[w_i \log(w_i / a_i) - w_i + a_i \right]$ (« estimateur par la méthode itérative du quotient (raking) généralisée »), (3) donne

$$\log\left[\frac{w_i}{a_i}\right] - \mathbf{x}_i^T \boldsymbol{\lambda} = 0$$

ou

$$w_{i,RAKE} = a_i e^{\mathbf{x}_i^T \boldsymbol{\lambda}} \qquad (4)$$

où (2) exige que

$$\sum_{i \in s} a_i e^{\sum_j \lambda_j x_{ji}} x_{ji} - X_j = 0 \text{ pour tout } j = 1, \ldots J \ . \qquad (5)$$

Puis, nous résolvons (5) pour λ_j à l'aide d'un algorithme de Newton-Raphson.

L'une des difficultés de l'estimation GREG tient au fait que les w_i résultants ne sont pas contraints d'être positifs. Or, les poids négatifs ne sont généralement pas utilisés par les progiciels standards et peuvent causer d'autres problèmes (p. ex., estimation générale directe de la prévalence négative). Les poids obtenus par la méthode itérative du quotient généralisée ont aussi tendance à être plus stables, fait qui concorde avec l'interprétation d'« effet principal uniquement » de la méthode itérative du quotient par opposition à l'interprétation d'« interaction » de la post-stratification. Donc, l'analyse qui suit s'appuie sur l'estimateur par la méthode itérative du quotient généralisée.

2.1 Région de calage *S* et variables de calage $\mathbf{X}^S$

La région *S* sur laquelle sont calculées les « distances » de calage pourrait varier de l'entièreté des États-Unis aux régions les plus petites pour lesquelles existent des cas NHIS. Le choix de *S* repose sur un compromis biais-variance, où la plus petite région fera « ressembler davantage » la BRFSS aux cas NHIS existants, mais où le peu d'information de calage provenant de la NHIS pourraient rendre les estimateurs résultant instables; de la même façon, le choix de plus grandes régions supposera une plus grande interchangeabilité sur un ensemble de petits domaines d'estimation, mais réduira la variabilité connexe.

Pour les données de la NHIS à grande diffusion uniquement, la plus petite région de calage *S* est une catégorisation ayant pour dimension quatre régions par sept MSA. Les régions sont les quatre régions de recensement des États-Unis : Nord-Est : Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York,

New Jersey et Pennsylvanie; Midwest : Ohio, Indiana, Illinois, Michigan, Wisconsin, Minnesota, Iowa, Missouri, Dakota du Nord, Dakota du Sud, Nebraska et Kansas; Sud : Delaware, Maryland, District de Columbia, Virginie, Virginie de l'Ouest, Caroline du Nord, Caroline du Sud, Georgie, Floride, Kentucky, Tennessee, Alabama, Mississippi, Arkansas, Louisiane, Texas et Oklahoma; Ouest : Montana, Idaho, Wyoming, Colorado, Nouveau-Mexique, Arizona, Utah, Nevada, Washington, Oregon, Californie, Alaska et Hawaii. Les MSA correspondent à sept catégories définies par : les régions métropolitaines comptant plus de 5 millions d'habitants, de 2,5 à 5 millions d'habitants, de 1 à 2,5 millions d'habitants, de 500 000 à 1 million d'habitants, de 250 000 à 500 000 habitants et de moins de 250 000 habitants, ainsi que les régions non métropolitaines. Il s'agit de l'ensemble de régions de calage utilisées dans la présente analyse.

Les variables de calage qu'il faut inclure dans $\mathbf{X}^S$ devraient de toute évidence comprendre des variables susceptibles d'être sensibles à toute différence de biais de réponse entre la BRFSS et la NHIS. La longueur *J* du vecteur de calage comporte de nouveau un compromis biais-variance, une grande valeur de *J* induisant le degré le plus important de recalibration et, donc, de réduction du biais, mais créant aussi de plus grandes valeurs de w_i qui pourraient aussi accroître la variabilité des estimations régionales de la prévalence. Comme le nombre de variables de calage est limité et que les estimateurs directs de la prévalence sont instables, nous utiliserons généralement une technique de lissage pour obtenir des estimations régionales de la prévalence (voir la section 2.2 plus loin). Nos analyses préliminaires ont indiqué que le recours à ce lissage permet de maximiser la correction éventuelle du biais grâce à l'inclusion de l'ensemble le plus grand de totaux de contrôle. Cet ensemble comprend le comportement à risque commun étudié (ici le comportement à l'égard du tabac), les variables démographiques communes, c'est-à-dire le niveau de scolarité, le revenu du ménage, l'état matrimonial, la race/groupe ethnique, et le sexe, ainsi que les variables communes de la santé c'est-à-dire la situation d'assurance-maladie, l'âge et l'indice de masse corporelle (IMC).

2.2 Lissage des estimations régionales

Notre petite région primaire d'intérêt *s* correspond aux comtés des États-Unis, dont le nombre est, à l'heure actuelle, de 3 115 (n'incluent pas les régions de recensement de l'Alaska). Nous excluons le comté de Kalawao à Hawaii (population 147), car nous ne disposons que de données démographiques très limitées et nous présentons les résultats pour 3 114 comtés américains.

Les estimations régionales directes sont données simplement par $\hat{\theta}_{ws}^{DIR} = \sum_{i\in s} w_i y_i \Big/ \sum_{i\in s} w_i$. Cependant, cet estimateur direct peut être assez instable si le nombre d'habitants du comté est faible. Une autre solution consiste à utiliser un estimateur par régression logistique pondérée pour produire des estimations ponctuelles pour les comtés individuels; autrement dit, nous résolvons l'équation de score (Binder 1983)

$$U_w(\boldsymbol{\beta}) = \sum_i w_i \mathbf{z}_i \left(y_i - \text{expit}\left(\mathbf{z}_i^T \boldsymbol{\beta}\right)\right) \qquad (7)$$

où $\text{expit}(x) = \exp(x)/(1+\exp(x))$ et $\mathbf{z}_i \equiv \mathbf{z}_s$ pour tout $i \in s$ représente un vecteur des covariables au niveau du comté. La résolution de l'équation de score nous donne une estimation qui satisfait $U_w(\hat{\boldsymbol{\beta}}_w) = 0$, de sorte que, pour chaque comté, l'estimateur par régression calé est donné par

$$\hat{\theta}_{ws}^{REG} = \text{expit}\left(\mathbf{z}_s^T \hat{\boldsymbol{\beta}}_w\right) \qquad . \qquad (8)$$

Pour les analyses présentées plus bas, $\mathbf{z}_s$ représente la proportion de la population du comté qui, en 1996, était de race noire, celle qui était de race hispanique (ou de toute autre race), celle âgée de 25 ans et plus ayant obtenu un diplôme d'études secondaires (1990), celle de 25 ans et plus ayant obtenu un diplôme collégial (1990), les impôts fonciers par habitant (1992), le montant des fonds et subventions fédéraux par habitant (1997), les prestations de sécurité sociale par habitant (1996), la proportion de personnes dont le revenu est inférieur au seuil de pauvreté (1993), le nombre de crimes graves dont la police est au courant par habitant (1995), le taux de chômage au sein de

la population active civile (1994), le nombre d'établissements de services sociaux par habitant (1995), la rémunération par habitant selon le comté de résidence (1996), le taux de lecture des quotidiens du lundi au vendredi (1997), la population par mille carré (1990), l'indice médian de pouvoir d'achat (2000), les ventes totales au détail, les ventes d'aliments et les ventes de boissons par habitant (2000), le nombre d'emplois de col bleu par habitant (2000) et la population totale (1997 à 2000).

Pour une comparaison à la BRFSS uniquement, nous pouvons obtenir un estimateur direct $\hat{\theta}_{as}^{DIR} = \sum_{i \in s} a_i y_i \Big/ \sum_{i \in s} a_i$.

Similairement, nous pouvons obtenir l'estimateur par régression « non calé » $\hat{\theta}_{as}^{REG}$ qui ne s'appuie pas sur les données de la NHIS en remplaçant $\hat{\boldsymbol{\beta}}_w$ dans (3) par $\hat{\boldsymbol{\beta}}_a$, où $\hat{\boldsymbol{\beta}}_a$ est obtenu par remplacement des poids de calage w_i par les poids de sondage a_i dans (7).

2.3 Inférence

Comme la plupart des techniques de réduction du biais, les poids de calage proposés ont tendance à augmenter la variance des estimations résultantes de la prévalence. Cette augmentation de la variance est due à la croissance de la variabilité des poids proprement dits, ainsi qu'à l'incertitude concernant les totaux de calage de la NHIS. Plus précisément, dans (8), considérons l'estimateur $\hat{\boldsymbol{\beta}}_w$ comme étant une fonction des poids de calage w_i qui, eux-mêmes, ont une variabilité due à l'incertitude concernant les totaux de calage $\mathbf{X}^S$ calculés d'après les données de la NHIS. Conditionnellement au vecteur de pondération w_i , nous obtenons

$$Var(\hat{\boldsymbol{\beta}}_w) = E\left[Var(\hat{\boldsymbol{\beta}}_w) \mid w\right] + Var\left[E(\hat{\boldsymbol{\beta}}_w) \mid w\right]. \qquad (9)$$

Nous estimons $E\left[Var(\hat{\boldsymbol{\beta}}_w) \mid w\right]$ par une méthode du jackknife (Korn et Graubard, 1999, ch. 2) qui tient compte de la stratification au niveau de l'État et des poids des cas, en ignorant les effets mineurs de mise en grappes dus à toute mise en grappes au niveau de l'UPE liée au plan d'échantillonnage par composition aléatoire de numéros de téléphone de Waksberg. Puis, nous utilisons une méthode bootstrap paramétrique (Davidson et Hickley 1997, ch. 2) pour estimer $Var\left[E(\hat{\boldsymbol{\beta}}_w) \mid w\right]$. Nous estimons la covariance de $\mathbf{X}^S$ par une méthode du jackknife qui tient compte de la stratification, de la mise en grappes et des poids du plan d'échantillonnage de la NHIS. Ensuite, nous procédons à des simulations à partir d'une distribution normale multivariée de moyenne $\mathbf{X}^S$ et $Cov(\mathbf{X}^S)$, et nous recalculons les poids de calage w_i au moyen des tirages répétés de $\mathbf{X}^S$. Puis, nous recalculons $\hat{\boldsymbol{\beta}}_w$ en utilisant (8) et les w_i recalculés/répétés, et la variabilité résultante de $\hat{\boldsymbol{\beta}}_w^{rep}$ utilisée pour évaluer $Var\left[E(\hat{\boldsymbol{\beta}}_w) \mid w\right]$. Le théorème central limite donne à penser que l'hypothèse d'une distribution normale multivariée de $\mathbf{X}^S$ est raisonnable, puisque les composantes de $\mathbf{X}^S$ sont des sommes sur des centaines ou sur des milliers de répondants.

Après avoir estimé $Var(\hat{\boldsymbol{\beta}}_w)$, nous estimons la variance de l'estimateur par régression de la prévalence $\hat{\theta}_{ws}^{REG}$ selon la méthode Delta.

La variance de $\hat{\theta}_{ws}^{DIR}$, v_{ws}^{DIR} , est donnée par $E\left[Var(\hat{\theta}_{ws}^{DIR}) \mid w\right] + Var\left[E(\hat{\theta}_{ws}^{DIR}) \mid w\right]$, où le premier élément est estimé par une méthode du jackknife standard (aujourd'hui, typiquement, le jackknife avec suppression d'une unité, puisque la stratification au niveau de l'État n'est plus un problème) et le deuxième élément est estimé par la méthode paramétrique du bootstrap décrite au paragraphe précédent.

Les poids de sondage a_i n'intègrent pas la variabilité d'échantillonnage de la NHIS et, par conséquent, on peut estimer la variance des estimateurs basés sur la BRFSS uniquement $\hat{\theta}_{as}^{REG}$ and $\hat{\theta}_{as}^{DIR}$ par les méthodes types d'approximation par le jackknife ou par série de Taylor.

2.4 Données manquantes

Nous avons réalisé l'analyse qui suit sous une hypothèse de réponses manquant entièrement au hasard (MCAR) (Little et Rubin 1987, ch. 1,5). Sous cette hypothèse, l'absence des données est, en principe, entièrement indépendante des données. Donc, le seul problème que peuvent poser les données manquantes est celui de la recalibration résultant uniquement des différences entre les données manquantes au niveau des grandes régions. Pour éviter que cela se produise, nous avons calculé des poids NHIS distincts pour chaque variable, nous avons fixé leur valeur à 0 si l'élément de donnée en question manquait pour le sujet de la NHIS et nous avons procédé à un rajustement à la hausse de sorte que le total et les poids corrigés pour les cas de données non manquantes soit égaux au total des poids dans la grande région *S* (Korn et Graubard 1999, ch. 4) :

$$a_{ji}^{NHIS^{(2)}} = \begin{cases} 0 \text{ si } x_{ji} \text{ est manquant} \\ \dfrac{\sum_{i \in s} a_{ji}^{NHIS}}{\sum_{i \in s} a_{ji}^{NHIS} I(x_{ji} \text{ observé})} a_i^{NHIS} \text{ si } x_{ji} \text{ est observé} \end{cases}$$

Puis, nous avons utilisé ces poids pour calculer les totaux NHIS de calage

$$\mathbf{X}^S = \left(\sum_{i \in s} a_{1i}^{NHIS^{(2)}} x_{1i}^{NHIS}, \ldots, \sum_{i \in s} a_{Ji}^{NHIS^{(2)}} x_{Ji}^{NHIS} \right)^T .$$

3. RÉSULTATS

À l'aide des méthodes décrites plus haut, nous déterminons les estimations au niveau du comté de la prévalence de l'usage du tabac chez les hommes de 18 ans et plus à l'aide de données provenant de la NHIS et de la BRFSS pour 1997.

Le tableau 1 donne une comparaison des estimations nationales des comportements à l'égard du tabac d'après la NHIS et la BRFSS pour la période allant de 1997 à 2000, ainsi que d'autres estimations au niveau national. Les données de la NHIS produisent des estimations un peu plus élevées de la prévalence de l'usage du tabac. La NHIS semble aussi atteindre une plus grande proportion de personnes appartenant aux catégories de statut socioéconomique le plus faible et le plus élevé, ainsi qu'un nombre un peu plus élevé de minorités, résultat qui concorde avec l'inclusion des ménages n'ayant pas le téléphone et le taux élevé de réponse de la NHIS.

	1997	
	BRFSS	NHIS
Fumeur	23,1 %	24,7 %
N'a jamais fumé	46,9 %	47,6 %
Niveau de scolarité : Pas de diplôme d'études secondaires	13,7 %	19,4 %
Diplôme collégial ou plus élevé	26,2 %	21,8 %
Revenu : <10 000 $	6,3 %	9,5 %
>75 000 $	13,0 %	18,2 %
Afro-américain	9,8 %	11,0 %
Hispanique	9,8 %	9,6 %
Couvert par une assurance-maladie	85,8 %	85,2 %
Homme	48,0 %	48,0 %
Âge moyen (années)	45,0	44,6
IMC moyen	25,9	26,2

Tableau 1 : Estimation nationale de certains indicateurs démographiques et de l'état de santé d'après la BRFSS et la NHIS pour 1997.

Les figures 1 et 2 montrent les estimateurs par régression non calé et calé de la prévalence de l'usage du tabac chez les hommes en 1997 donnés par $\hat{\theta}_{as}^{REG}$ et $\hat{\theta}_{ws}^{REG}$ pour les 3 114 comtés des États-Unis. Le profil de base est le même pour les deux cartes, les taux les plus élevés figurant dans les deux bandes qui partent du Sud et s'étendent à travers l'Est de l'Ohio et à travers les montagnes Rocheuses. Les taux ont tendance à être plus faibles dans les centres urbains que dans les régions rurales environnantes.

Comme le laissent entendre les figures 1 et 2, la corrélation entre les deux estimateurs par régression est assez forte (r=0,76). La figure 3 donne le tracé de l'écart entre les estimateurs par régression calé (sur les données de la NHIS) et non calé (données de la BFRSS uniquement) pour les hommes et pour les femmes comparativement à la moyenne des deux estimateurs [tracés « Bland et Altman » (Bland et Altman, 1986)]. Ces tracés montrent qu'il existe des écarts non négligeables entre les deux estimateurs et, plus particulièrement, qu'en moyenne, l'estimateur rajusté d'après les données de la NHIS produit des estimations plus élevées du taux d'usage du tabac et que l'augmentation de la prévalence due à la recalibration des poids a tendance à être la plus importante pour les comtés où la prévalence est la plus élevée, tandis qu'elle a tendance à être réduite quelque peu pour les comtés où la prévalence est la plus faible. En général, les taux augmentent dans le cas de l'estimation calée sur les données de la NHIS, résultat qui concorde avec les estimations plus fortes de la prévalence obtenues d'après les données de la NHIS : les 2,5, 50 et 97,5 percentiles de la distribution des taux de prévalence de l'usage du tabac chez les hommes adultes au niveau du comté sont 21,2 %, 27,9 % et 36,2 % pour l'estimateur par régression non calé et 21,7 %, 30,7 % et 40,2 % pour l'estimateur par régression calé. Les estimations calées sur les données de la NHIS ont tendance à être plus fortes pour les comtés ruraux, mais égales ou plus faibles pour les comtés urbains. La régression de l'écart relatif entre la prévalence de l'usage du tabac chez les hommes produite au moyen des estimateurs de régression non calé et calé sur les covariables de comté, ainsi que l'examen de la somme des carrés de type III résultante montrent que la variable de prestations de sécurité sociale par habitant, qui est principalement une approximation de l'âge, a le coefficient de régression partielle le plus important (0,89); viennent ensuite la variable de race hispanique (0,87), le taux de pauvreté (0,81) et le niveau de scolarité (0,77). Ces variations concordent avec la correction du biais pour les ménages non raccordés au téléphone et du biais de « classe moyenne » dans les taux de réponse (Goyer et coll., 2002). (Cette dernière hypothèse est contestée : voir Groves et Couper, 1998.)

La figure 4 donne les erreurs-types des estimations de la prévalence de l'usage du tabac chez les hommes pour les 3 114 comtés des États-Unis produites par l'estimateur de régression calé. La méthode de lissage employée donne généralement de forte variances pour les comtés pour lesquels les valeurs de certaines covariables sont extrêmes. Les erreurs-types des estimations de la prévalence de l'usage du tabac au niveau du comté en 1997 calées sur les données de la NHIS étaient, en moyenne, 1,56 fois plus grandes pour les hommes et 2,29 fois plus grandes pour les femmes, à cause de l'incertitude concernant les poids recalibrés attribuable à la variabilité d'échantillonnage de la NHIS, ainsi qu'à l'augmentation de la variabilité des poids.

4. DISCUSSION

Nous examinons dans le présent rapport la question de savoir si l'on peut combiner les données à grande diffusion provenant de la National Health Interview Survey et de la Behavioral Risk Factor Surveillance Survey pour produire de meilleures estimations régionales de la prévalence des comportements associés à un risque de cancer, tout spécialement l'usage du tabac et le dépistage du cancer. Comme la base de sondage de la NHIS comprend les ménages n'ayant pas le téléphone et que le taux de réponse de cette enquête est plus élevé que celui de la BRFSS, on pourrait supposer que les estimations fondées sur les données de la NHIS sont moins biaisées que les estimations équivalentes obtenues d'après la BRFSS. Comme les données à grande diffusion de la NHIS ne contiennent pas d'identificateurs de petites régions, nous envisageons l'utilisation de méthodes de calage pour rajuster les poids d'échantillonnage de la BRFSS de sorte que les estimations pondérées fondées sur la BRFSS concordent mieux avec les estimations pondérées fondées sur la NHIS.

En général, les méthodes d'estimation par régression logistique où l'on utilise les poids calibrés d'après la NHIS produisent des estimations plus élevées de la prévalence de l'usage du tabac que celles calculées en utilisant uniquement les poids de la BRFSS. La recalibration des poids a tendance à augmenter le plus les taux de prévalence de l'usage du tabac des comtés où ils sont les plus élevés, et à réduire dans une certaine mesure ceux des comtés où ils sont les plus faibles. L'âge, la race ou le groupe ethnique, le taux de pauvreté et le niveau de scolarité sont les prédicteurs les plus importants des différences entre l'estimateur non calé et l'estimateur calé sur les données de la NHIS, observation qui concorde avec le « biais de classe moyenne » qui serait, suppose-t-on, créé par le biais de sélection et la non-couverture par la base de sondage dans les enquêtes téléphoniques à faible taux de réponse.

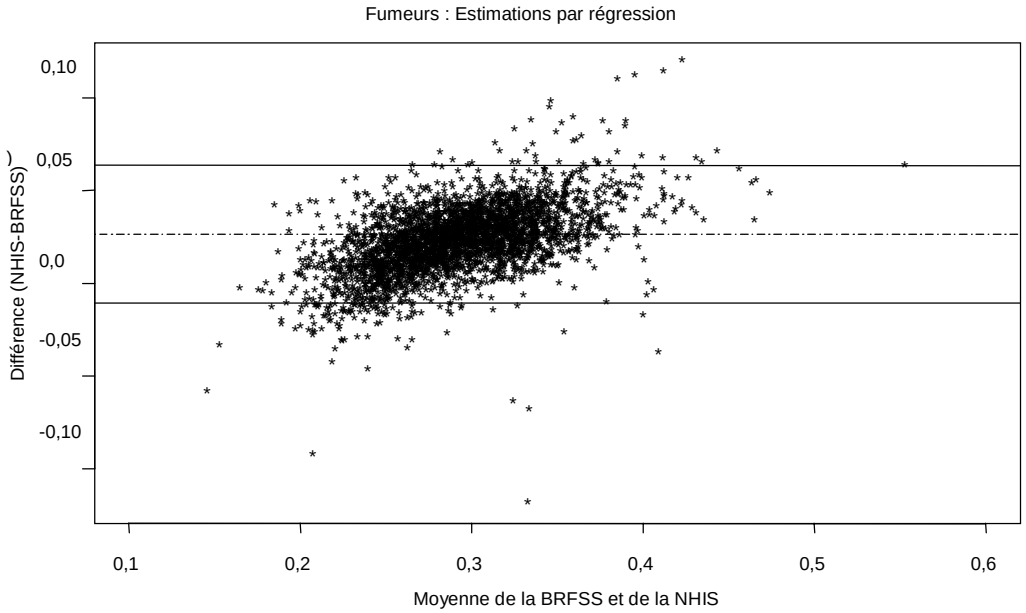

Figure 3 : Tracé de la différence entre les estimations par régression calées sur les données de la NHIS et non calées (BRFSS uniquement) de la prévalence de l'usage du tabac chez les hommes de 18 ans et plus en 1997 en fonction de la moyenne des estimations calées et non calées, pour les 3 114 comtés des États-Unis. (—·—·—) différence moyenne; (—) moyenne +/- 2 écarts-types de la différence

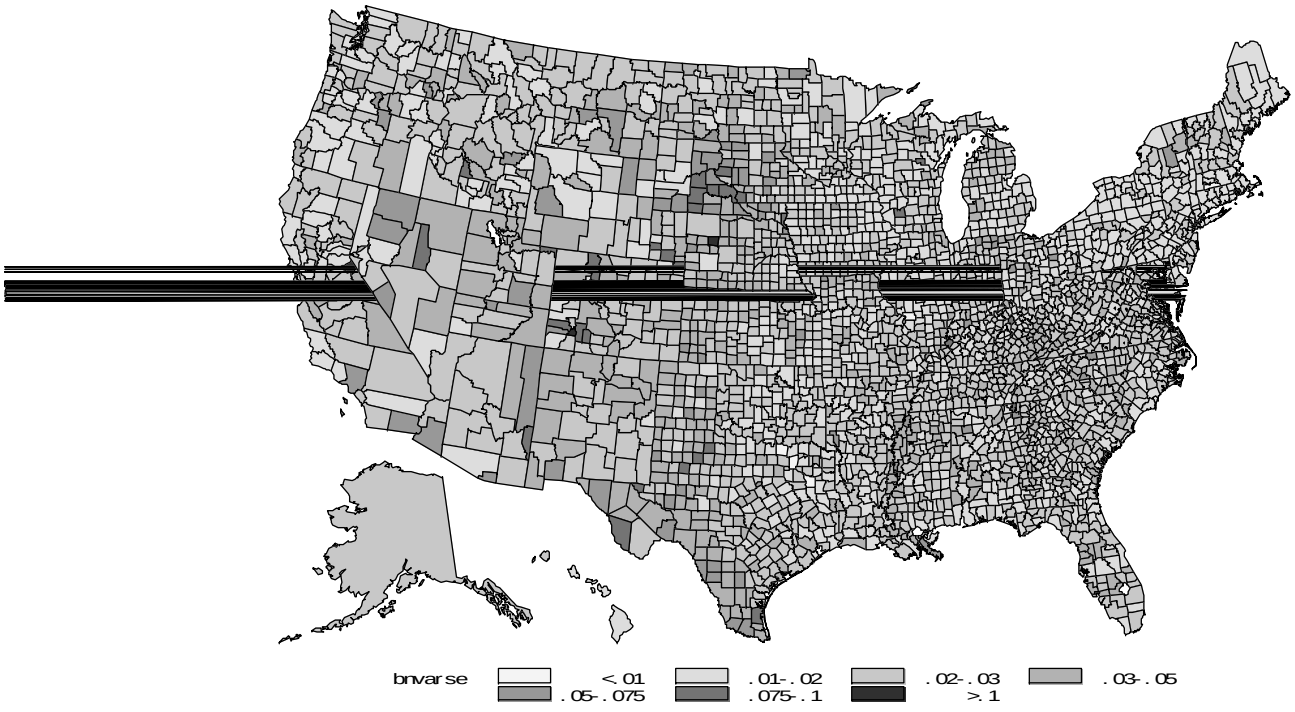

Figure 4 : Erreurs-types des estimations au niveau du comté de la prévalence de l'usage du tabac en 1997 chez les hommes de 18 ans et plus obtenues au moyen de l'estimateur de régression calé sur les données de la NHIS

RÉFÉRENCES

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Bland, J.M. Altman, D.G. (1986). Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet*, 327, 307-310.

Cassel CM, Sarndal CE, Wretman JH. (1976), "Some results on generalized difference estimation and generalized regression estimation for finite populations," *Biometrika* 63, pp. 615-620.

Deville JC, Sarndal C-E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association* 87, pp. 376-382.

Davidson, AC, Hinckley, DV (1997), *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.

Goyer, J., Warriner, K., Miller, S. (2002), "Evaluating Socio-economic Status (SES) Bias in Survey Nonresponse," *Journal of Official Statistics*, 18, pp. 1-12.

Groves, R.M., Couper, M.P. (1998), *Nonresponse in Household Interview Surveys*. New York: John Wiley and Sons.

Hartley, H.O. (1962), "Multiple Frame Surveys," *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.

——— (1974), "Multiple Frame Methodology and Selected Applications," *Sankhya*, C36, pp. 99-118.

Isaki C, Fuller W. (1982), "Survey Design under the Regression Superpopulation Model," *Journal of the American Statistical Association* 77, pp. 89-96.

Korn, E.L., Graubard, B.I. (1999), *Analysis of Health Surveys*. New York: John Wiley and Sons.

Legler J., Meissner H., Breen N., Malec D. (2002), "Mammography and the underserved: a geographical perspective on the congruency of national needs and intervention research," draft manuscript, Washington, DC: National Cancer Institute.

Little R.J.A., Rubin D.B. (1987), *Statistical Analysis with Missing Data*. New York: John Wiley and Sons.

Lohr, S.L., Rao, J.N.K. (2000), "Inference from Dual Frame Surveys," *Journal of the American Statistical Association*, 95, pp. 271-280.

Moriarity, C., Scheuren, F. (2001), "Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure," *Journal of Official Statistics*, 17, pp. 407-422.

Nandram, B., Sedransk, J., Pickle, L.W. (2000), "Bayesian Analysis and Mapping of Mortality Rates for Chronic Obstructive Pulmonary Disease," *Journal of the American Statistical Association*, 95, pp. 1110-1118.

Pickle, L.W., Mungiole, M., Jones, G.K., White, A.A. (1996), *Atlas of United States Mortality*. Hyattsville, MD: National Center for Health Statistics.

Potosky A.L., Breen N., Graubard B.I., Parsons P.E. (1998), "The association between health care coverage and the use of cancer screening tests: Results from the 1992 National Health Interview Survey." *Medical Care*, 36, pp. 257-70.

Rodgers, W.E. (1984), "An Evaluation of Statistical Matching," *Journal of Business and Economic Statistics*, 2, pp. 91-102.