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

COMPARAISON DE DEUX MÉTHODES DE MODÉLISATION DE DONNÉES SUR LA SANTÉ ET L'ENVIRONNEMENT

Sandra A. Ham[1], Michelle M. Yore[1], Peter Mariolis[1]

RÉSUMÉ

Le domaine de la santé publique s'intéresse de plus en plus à l'incidence de l'environnement sur la santé. Idéalement, les études écologiques ou contextuelles consistent à analyser ces relations à l'aide de données sur la santé publique, complétées par des caractéristiques environnementales, à l'intérieur de modèles multiniveaux ou hiérarchiques. Dans ces modèles, les données sur la santé des personnes constituent le premier niveau, et les données concernant les collectivités, le deuxième. La plupart des données sur la santé publique sont le fruit de plans d'enquêtes par sondage complexes qui nécessitent des analyses tenant compte de la mise en grappes, de la non-réponse et de la post-stratification a pour obtenir des estimations représentatives de la prévalence de comportements à risque pour la santé. La présente étude utilise les données de l'enquête intitulée *Behavioral Risk Factor Surveillance System* (BRFSS). Administrée par les États américains en collaboration avec les Centers for Disease Control and Prevention, cette enquête évalue chaque année les facteurs de risque pour la santé chez plus de 200 000 adultes. À l'heure actuelle, les données de la BRFSS sont disponibles au niveau des secteurs statistiques métropolitains (*metropolitan statistical areas* ou MSA); elles fournissent des renseignements de qualité sur la santé en vue d'études sur les effets environnementaux. Les analyses menées au niveau des MSA et combinant des données sur la santé et sur l'environnement se compliquent en raison des exigences communes du plan d'échantillonnage et des analyses multiniveaux. Dans une étude de l'activité physique et de certains facteurs environnementaux, nous comparons deux méthodes de modélisation à partir des données de la BRFSS menée en 2000. Chacune des méthodes que nous décrivons constitue un moyen valide pour analyser les données d'enquêtes par sondage complexes complétées par des renseignements sur l'environnement, mais chacune tient compte de manière différente du plan d'enquête et de la structure multiniveaux des données; elles conviennent donc pour des questions de recherche légèrement différentes.

Mots clés : enquêtes, étude écologique, modèles mixtes

1. INTRODUCTION

Le domaine de la santé publique s'intéresse directement à l'incidence de l'environnement sur la santé (Sallis et coll., 1998). Selon la théorie de l'écologie sociale en matière de relations, les personnes s'inscrivent dans des ensembles de facteurs contextuels ou environnementaux dont elles subissent l'incidence, d'où la nécessité d'établir des modèles de données à structure hiérarchique (Stokols, 1992; Stokols, 1996; Masse et coll., 2002). Idéalement, les études écologiques consistent à analyser ces relations à l'aide de données sur la santé publique, complétées par des effets contextuels, à l'intérieur de modèles multiniveaux ou hiérarchiques. Dans ces modèles statistiques, les données sur la santé de personnes (i) vivant dans des collectivités (j) constituent le premier niveau, et les données concernant les collectivités, le deuxième niveau. On peut évaluer les effets socioéconomiques sur la santé aux niveaux des personnes et des collectivités, comme le montre l'équation $\text{Santé}_{ij} = \text{SSE}_{ij} + \text{Environnement}_j$. Une collectivité peut être un village, une ville, un comté, un secteur métropolitain, un district sanitaire ou toute autre unité géographique de petite taille. On établit des inférences en comparant les collectivités en fonction d'une foule de combinaisons d'effets contextuels. Les ensembles nationaux de données sur la santé publique constituent le plus vaste échantillon de collectivités et sont rapidement et facilement utilisables. Dans le cadre d'études écologiques, ils permettent donc de générer des hypothèses et de cerner des corrélats contextuels des comportements et des résultats en matière de santé.

Peu d'études (Chou et coll., 2001) ont combiné des données nationales sur la santé avec des facteurs contextuels tels que la sécurité publique, le climat et l'environnement naturel. Les enquêtes nationales sur la santé sont dotées de plans d'enquête complexes, qui nécessitent des analyses tenant compte de la mise en grappes, de la non-réponse et de la post-stratification pour obtenir des estimations représentatives de l'état de santé et des comportements à risque pour la santé. Les analyses menées au niveau des collectivités et combinant des données sur la santé et sur

[1] Centers for Disease Control and Prevention, 4770 Buford Highway NE, Atlanta, Georgia, USA 30341

l'environnement se compliquent en raison des exigences communes du plan d'échantillonnage et des analyses multiniveaux. Malgré que les logiciels statistiques servant à analyser les données d'enquêtes nationales (p. ex. SUDAAN et SAS) tiennent compte des plans d'enquête complexes, ils n'offrent pas les fonctions de modélisation multiniveaux; quant aux logiciels d'analyse multiniveaux (p. ex. MLwiN, HLM et winBUGS), ils ne sont pas adaptés aux plans d'enquête complexes.

Dans la présente étude, nous comparons des méthodes de modélisation pour la recherche écologique en santé publique qui sont axées à la fois sur des plans d'enquête complexes et sur des données concernant des facteurs contextuels. Ces méthodes sont présentées dans le contexte d'un exemple utilisant des données sur l'activité physique tirées de l'enquête américaine intitulée *Behavioral Risk Factor Surveillance System* (BRFSS) menée en 2000. Chacune des méthodes tient compte du plan d'enquête complexe de la BRFSS et permet de générer des hypothèses ou d'évaluer des corrélats liant des facteurs contextuels environnementaux à des comportements et à des résultats en matière de santé. Dans nos observations sur les points forts et les lacunes, nous abordons d'autres décisions en matière de modélisation, dont l'inclusion de corrélats et des questions de recherche pertinentes pour chaque méthode.

2. MÉTHODES

Notre exemple modélise l'activité physique recommandée, en incluant trois mesures du statut socio-économique individuel et trois mesures contextuelles des caractéristiques des collectivités, énumérées dans le tableau 1. Nous avons choisi une mesure du statut socio-économique contextuel, une mesure du climat et une mesure de l'accès aux aires de loisirs. Dans les deux sections suivantes, nous décrivons en détail les données de l'exemple, puis nous présentons les méthodes.

Tableau 1. Sources des données pour l'exemple de l'activité physique*

Niveau	Variable	Source	N
Y	A.P. (activité physique recommandée)	BRFSS 2000	87 050
1	SEXE, ÂGE (18 à 24 ans, 25 à 34 ans, 35 à 44 ans, 45 à 54 ans, 55 à 64 ans, 65 ans et plus), RACE (blancs, non-blancs)	BRFSS 2000	87 050
2	HOMMES (% d'hommes dans le MSA)	U.S. Bureau of the Census	100
2	INDICE-CHALEUR (indice de chaleur le 15 juillet à 14 h, moyenne de 1961 à 1990)	The Zunis Foundation	100
2	PARCS (n^{bre} de parcs d'État et de parcs fédéraux pour 100 000 personnes)	Places Rated Almanac	100

*BRFSS = Behavioral Risk Factor Surveillance System.

2.1 Données sur la santé et données au niveau individuel

La BRFSS est une enquête téléphonique servant à cerner la prévalence des comportements liés aux maladies chroniques et des pratiques de prévention sanitaire au sein de la population civile hors institution âgée de 18 ans et plus dans chaque État américain. Elle est menée conjointement par les Centers for Disease Control and Prevention (Atlanta, Georgie) et par les départements de la santé des 50 États, du district de Columbia, de Guam, de Porto Rico et des îles Vierges américaines. En 2000, plus de 180 000 adultes ont été interviewés (N = 87 050 personnes) dans 100 secteurs statistiques métropolitains (*metropolitan statistical areas* ou MSA) ou MSA primaires, définis pour les besoins du recensement. L'échantillonnage de la BRFSS est fondé sur les États; les MSA retenus étaient donc les secteurs métropolitains les plus populeux de chaque État et du district de Columbia; la taille des échantillons de chaque MSA variait entre 300 et 7 100 personnes. Pour les analyses menées au niveau des MSA, on disposait des poids applicables à ces données. D'autres renseignements concernant la BRFSS se trouvent sur le site http://www.cdc.gov/brfss.

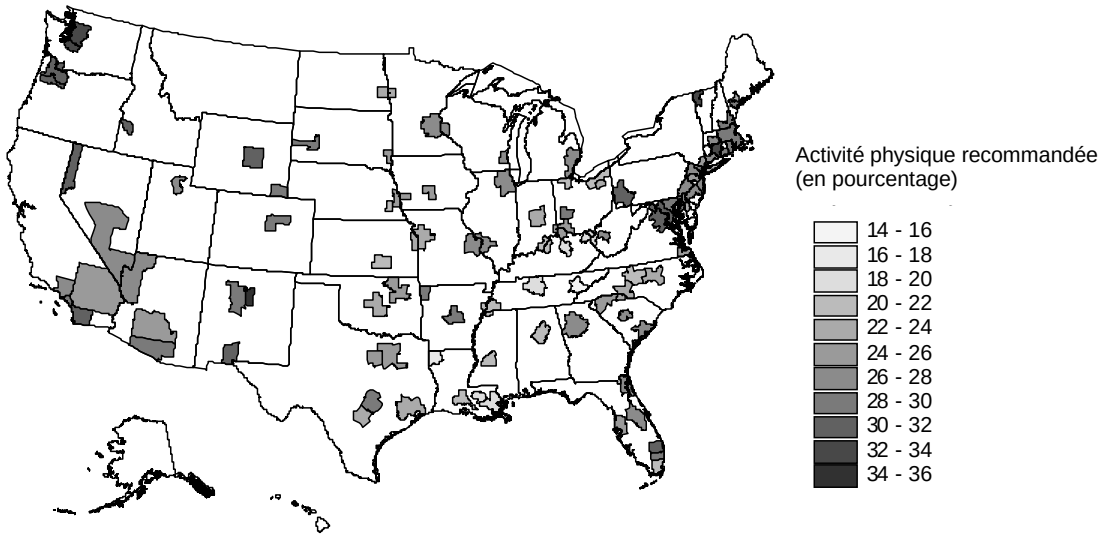
La variable de réponse était une mesure binaire de la conformité au niveau d'activité physique recommandé pendant les heures de loisirs (1 = oui; 0 = non). Selon la classification, les répondants satisfaisaient aux recommandations s'ils déclaraient s'adonner, pendant les heures de loisirs, à une activité physique cinq jours et plus par semaine à raison d'au moins 30 minutes par jour, à une activité physique vigoureuse trois jours et plus par semaine à raison

d'au moins 20 minutes par jour, ou aux deux types d'activité (CDC, 2001). La figure 1 montre la prévalence de l'activité physique recommandée dans chacun des 100 MSA. Dans cet exemple, on a utilisé l'âge, le sexe et la race ou l'ethnie au niveau individuel comme données du premier niveau.

2.2 Données contextuelles

Pour les besoins de ces exemples, les effets contextuels au niveau des MSA étaient le pourcentage d'hommes compris dans le MSA (HOMMES), l'indice de chaleur estivale (INDICE-CHALEUR) et l'accès aux parcs (PARCS) (voir le tableau 1). Toutes les données ont été obtenues auprès de sources pertinentes, reportées sur des tableurs électroniques et centrées autour des moyennes générales respectives des 100 MSA visés par l'analyse. On a obtenu le pourcentage d'hommes vivant dans chaque MSA d'après le Recensement de 2000 du U.S. Bureau of the Census (http://www.census.gov). La fondation Zunis (http://www.zunis.org) a fourni une moyenne sur 30 ans des indices de chaleur enregistrés le 15 juillet à 14 h, d'après les observations météorologiques horaires publiées de 1961 à 1990 par 140 stations météorologiques urbaines des États-Unis, de Porto Rico et de Guam. Les données sur les observations météorologiques ont été rassemblées par le Solar and Meteorological Surface Observation Network et publiées par la U.S. National Oceanic and Atmospheric Administration (http://www.noaa.gov). On a supposé que les indices de chaleur en ville représentaient le climat moyen dans les MSA respectifs. On a obtenu les noms des parcs d'État et des parcs fédéraux de chaque MSA d'après une liste dressée à partir des données d'organismes fédéraux et publiée dans l'almanach intitulé *Places Rated Almanac* (Savageau et D'Agostino, 2000, p. 509 à 568). On a additionné les nombres de tous les parcs d'État et parcs fédéraux énumérés, puis on les a divisés par les chiffres de population des MSA déclarés par le U.S. Bureau of the Census lors du Recensement de 2000 pour calculer le nombre de parcs d'État et de parcs fédéraux pour 100 000 personnes.

Figure 1. Prévalence de l'activité physique recommandée* dans 100 secteurs statistiques métropolitains d'après l'enquête *Behavioral Risk Factor Surveillance System* de 2000.



* Activité physique recommandée = personnes ayant déclaré s'adonner à une activité physique au moins 5 fois par semaine x 30 minutes, ou à une activité physique vigoureuse pendant au moins 20 minutes à la fois au moins 3 fois par semaine.

2.3 Première méthode : modèle agrégé

Le modèle agrégé est une régression linéaire à deux degrés de mesures du deuxième niveau, ou niveau des collectivités, ayant comme variable dépendante une prévalence de comportement lié à la santé et, comme variables indépendantes, des mesures contextuelles. Le premier degré utilise un logiciel d'analyse d'enquête pour calculer les

prévalences au sein des collectivités en appliquant des poids individuels pour les besoins des analyses au niveau des collectivités. Le deuxième degré est la régression des prévalences dans les mesures contextuelles à l'aide d'un logiciel statistique standard, après fusion de toutes les données en un seul ensemble de données.

Soit Y_j la prévalence du comportement lié à la santé pour la collectivité *j* avec une distribution normale $N \sim (\mu,\sigma^2)$. L'équation de ce modèle est la suivante : $Y_j = \beta_0 + X\beta + e_j$, où X est un vecteur $X_1, \ldots X_k$ des covariables du deuxième niveau pour la collectivité *j*, β est un vecteur $\beta_1, \ldots \beta_k$ des coefficients du deuxième niveau, et e_j est l'erreur résiduelle qu'on suppose distribuée normalement $N \sim (0, \sigma^2)$. Ce modèle n'utilise aucune covariable de niveau individuel; toutes les variables sont du niveau des collectivités.

Pour les besoins de notre exemple, nous avons d'abord estimé la prévalence de répondants individuels *i* qui satisfaisaient aux niveaux d'activité physique recommandés dans chaque MSA *j* en employant des méthodes standard tenant compte du plan d'enquête complexe, à l'aide du logiciel SUDAAN 8.0. Les données ont été pondérées et post-stratifiées pour chaque MSA. Puis, nous avons fusionné les estimations de prévalence avec une base de données de variables contextuelles au niveau des MSA : pourcentage d'hommes, indice de chaleur et nombre de parcs d'État et de parcs fédéraux pour 100 000 personnes. Le modèle du deuxième niveau était le suivant :

$$Y_j = \beta_0 + \beta_1(\text{HOMMES}) + \beta_2(\text{INDICE-CHALEUR}) + \beta_3(\text{PARCS}) + e_j. \quad (1)$$

Pour l'analyse de régression linéaire du modèle du deuxième niveau, nous avons utilisé le logiciel SAS 8.2.

2.4 Deuxième méthode : modèle multiniveaux avec coordonnées aléatoires à l'origine

Nous présentons maintenant un modèle multiniveaux dans lequel le plan d'échantillonnage complexe a été modélisé (Korn et Graubard, 1999, p. 177 à 181) à l'aide d'un logiciel de modélisation multiniveaux; à chaque collectivité correspondait une seule coordonnée aléatoire à l'origine, définie par les covariables contextuelles. Les poids pour les analyses au niveau des collectivités et les variables de plan d'échantillonnage sont intégrés, sauf lorsqu'il s'agit de facteurs confusionnels, comme nous le verrons plus loin. Nous présentons une version logistique de cette méthode, car la plupart des mesures de la santé tirées d'enquêtes sont catégoriques. Dans notre exemple, les covariables de niveau individuel sont dites du premier niveau, et les covariables contextuelles, du deuxième niveau.

Soit y_{ij} la mesure du comportement lié à la santé, tirée d'une distribution binomiale $y_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij})$, et π_{ij} la probabilité que le i^e répondant de la j^e collectivité déclare le comportement en question. Soit X un vecteur de covariables du premier niveau $X_{1ij},\ldots, X_{pij}$ pour le répondant *ij*, et β un vecteur de coefficients du premier niveau $\beta_{1j, \ldots, \beta_{pj}}$. Soit Z un vecteur $Z_{1j},\ldots, Z_{pj}$ de covariables contextuelles du deuxième niveau pour la collectivité *j*, et γ un vecteur de coefficients du deuxième niveau $\gamma_1, \ldots, \gamma_p$. Compte tenu des covariables, le modèle général de la probabilité de déclarer le comportement lié à la santé se présente comme suit : P(X) = Pr(Y = 1|X, Z).

Le modèle du premier niveau pour la probabilité de déclarer un comportement lié à la santé est $y_{ij} = \pi_{ij} + e_{0ij}$, où la variance des termes aléatoires du premier niveau e_{0ij} est $\pi_{ij}(1 - \pi_{ij})$. La probabilité du comportement lié à la santé pour le répondant *ij* est la fonction logit suivante : $\text{logit}(\pi_{ij}) = \beta_{0j} + X\beta$.

Nous ajoutons ensuite au modèle un deuxième niveau dans lequel la coordonnée à l'origine β_{0j} pour un individu est une fonction de covariables contextuelles ou du deuxième niveau mesurées au niveau des collectivités dans l'équation $\beta_{0j} = \gamma_0 + Z\gamma + :_{0j}$. Les effets aléatoires du deuxième niveau $:_{0j}$ sont distribués normalement avec une moyenne 0 et une variance égale à la variance entre les collectivités dans le comportement lié à la santé, comme dans l'équation $[:_{0j}] \sim N(0, \Omega_:)$, $\Omega_: = [\sigma^2_{:0}]$. L'erreur résiduelle au niveau individuel est distribuée normalement avec une moyenne 0 et une variance Ω_e comme dans l'équation $[e_{0ij}] \sim N(0, \Omega_e)$, $\Omega_e = [\sigma^2_{e0}]$.

On peut centrer toutes les covariables autour des moyennes générales pour simplifier l'interprétation du modèle (Raudenbush et Byrk, 2002, p. 31 à 35). On peut ajouter au modèle multiniveaux les interactions ou effets croisés entre les covariables du premier et du deuxième niveau pour vérifier des hypothèses concernant les effets contextuels sur des sous-groupes de population. Le pourcentage de variance expliqué par le modèle multiniveaux

avec coordonnées aléatoires à l'origine compare la variance du deuxième niveau $\sigma^2_{:0}$ d'un modèle inconditionnel du plan d'enquête sans covariables à la variance du deuxième niveau des modèles qui comprennent des covariables.

Au lieu d'employer les méthodes habituelles d'analyse d'enquête, on peut modéliser un plan d'échantillonnage complexe. Dans ce cas, le modèle doit comprendre les poids et toutes les variables du plan qui ne confondent pas l'analyse. Les questions de recherche portant sur des unités géographiques d'analyse risquent d'être confondues avec la stratification géographique du plan d'échantillonnage complexe si l'unité géographique qui nous intéresse est semblable aux strates géographiques, tout en étant différente de ces dernières. Le problème est manifeste lorsque les variables du plan expliquent une grande partie de la variance entre les collectivités avant qu'on ajoute au modèle les mesures contextuelles qui nous intéressent. Lorsque cette confusion existe, les variables correspondant à la stratification géographique sont exclues du modèle multiniveaux avec coordonnées aléatoires à l'origine.

Pour les besoins de notre exemple axé sur l'activité physique, nous avons utilisé un modèle pondéré au moyen de poids uniformisés avec une moyenne de 1,0 pour chaque MSA. Le plan d'échantillonnage complexe du BRFSS comprend l'échantillonnage stratifié, la post-stratification et la pondération. Nous avons exclu les variables correspondant aux strates géographiques, car elles confondaient l'interprétation des résultats lorsqu'on utilisait le MSA comme unité d'analyse du deuxième niveau. Notre modèle pour l'activité physique est le suivant :

$$\text{logit}(\pi_{ij}) = \beta_{0j} + \beta_1(\text{SEXE})_{ij} + \beta_2(\text{ÂGE})_{ij} + \beta_3(\text{RACE})_{ij} \quad (2)$$

$$\beta_{0j} = \gamma_0 + \gamma_1(\text{HOMMES})_j + \gamma_2\,(\text{INDICE-CHALEUR})_j + \gamma_3\,(\text{PARCS})_j + :_{0j}. \quad (3)$$

Nous avons créé des variables fictives pour les catégories SEXE, ÂGE et RACE. À l'aide du logiciel MlwiN 1.10, nous avons effectué des analyses en utilisant une estimation marginale de quasi-vraisemblance du premier ordre pour assurer la convergence du modèle, puis une estimation pénalisée de quasi-vraisemblance du deuxième ordre pour obtenir les estimateurs les moins biaisés. Nous aurions pu utiliser une distribution extra-binomiale, mais nous ne l'avons pas fait, après avoir vérifié le modèle inconditionnel sans constater de surdispersion ni de sous-dispersion dans cet ensemble de données du BRFSS. Si l'on appliquait des poids bruts, la taille de population pondérée du MSA deviendrait un facteur dans le modèle et les données présenteraient une distribution extra-binomiale. Nous avons exécuté trois modèles : un modèle inconditionnel avec plan mais sans covariables, un modèle avec effets individuels et, enfin, un modèle complet avec effets individuels et contextuels.

3. RÉSULTATS

Le tableau 2 montre les estimations de paramètres, les erreurs-types, les valeurs *P* et le pourcentage de variance entre les MSA qui est expliqué par le modèle agrégé et le modèle multiniveaux avec coordonnées aléatoires à l'origine. S'il est impossible de comparer directement les estimations de paramètres, on peut toutefois illustrer plusieurs points à l'aide de ce tableau. Dans le cas du modèle agrégé, on ne disposait pas d'estimations de paramètres pour les covariables de niveau individuel. L'orientation des estimations des effets contextuels était la même dans les deux modèles : positive pour les HOMMES et les PARCS, négative pour l'INDICE-CHALEUR. Les valeurs *P* des effets contextuels diffèrent d'un modèle à l'autre, ce qui nous porte à conclure que les PARCS constituent un corrélat significatif dans le cas du modèle agrégé, mais non significatif dans celui du modèle multiniveaux. Dans le modèle multiniveaux, l'importance des effets individuels est supérieure à celle des effets contextuels, ce qui montre l'importance relative des effets individuels par rapport aux effets contextuels en ce qui concerne l'activité physique recommandée. La variance du deuxième niveau expliquée par le modèle agrégé était de 31 %; elle était de 22 % dans le cas du modèle multiniveaux par rapport au modèle inconditionnel, et de 29 % par rapport au modèle comprenant les effets individuels.

Tableau 2. Estimations de paramètres pour le modèle agrégé et le modèle multiniveaux avec coordonnées aléatoires à l'origine[†]

	Modèle agrégé (linéaire)		Modèle multiniveaux (logistique)	
	β (E.-T.)	valeur *P*	β (E.-T.)	valeur *P*
Effets fixes				
SEXE	S.O.	S.O.	-0,10 (0,02)	<0,0001
ÂGE (35 à 44 ans*)	S.O.	S.O.	-0,23 (0,03)	<0,0001
RACE	S.O.	S.O.	-0,32 (0,02)	<0,0001
HOMMES	1,54 (0,47)	0,0014	0,09 (0,02)	0,0007
INDICE-CHALEUR	-0,19 (0,06)	0,0028	-0,01 (0,00)	0,012
PARCS	1,03 (0,37)	0,0057	0,04 (0,02)	0,066
Composantes de la variance				
Variance expliquée entre MSA	31 %		22 % (modèle complet par rapport au modèle inconditionnel) 29 % (modèle complet par rapport au modèle des effets individuels)	

[†]E.-T. = erreur-type; S.O. = sans objet; MSA = secteur statistique métropolitain (*metropolitan statistical area*).
*Estimation pour le groupe d'âge de 35 à 44 ans par rapport à celui de 18 à 24 ans. Les estimations concernant les autres groupes d'âge étaient moins significatives.

4. OBSERVATIONS

Nous avons présenté deux méthodes consistant à utiliser les données d'enquêtes par sondage complexes dans la recherche écologique en santé publique. La première méthode est un modèle agrégé visant à lier la prévalence aux mesures contextuelles en utilisant toutes les composantes d'un plan d'enquête complexe. La deuxième est un modèle multiniveaux avec coordonnées aléatoires à l'origine qui utilise la plupart des composantes du plan d'enquête complexe et toutes les données des niveaux individuel et contextuel dont on dispose. Chacune des méthodes que nous avons décrites constitue un moyen valide pour analyser les données d'enquêtes par sondage complexes complétées par des renseignements sur l'environnement, mais chacune tient compte de manière différente du plan d'enquête et de la structure multiniveaux des données; elles conviennent donc pour des questions de recherche légèrement différentes. Aucune méthode, aucun progiciel ne s'impose comme étant idéal pour tous les types d'hypothèse et de questions de recherche. Pour mener nos analyses, nous avons utilisé trois progiciels : SAS, SUDAAN et MLwiN.

Le modèle agrégé présente deux points forts : il s'agit d'une technique simple qui modélise la prévalence des effets sur la santé et il utilise tous les renseignements concernant le plan d'une enquête par sondage. Pour ce qui est de la génération d'hypothèses, les estimations des effets se comparent favorablement à celles du modèle multiniveaux, mais il faut les utiliser avec prudence. Le modèle présente aussi certaines lacunes : il risque de donner lieu à des estimations biaisées et ne tient compte ni de l'autocorrélation spatiale ni de la variabilité des estimations de la prévalence. En outre, les questions de recherche auxquelles on peut répondre se limitent à celles qui sont liées à la prévalence, car le modèle ne permet pas d'inclure les variables de niveau individuel.

Le modèle multiniveaux avec coordonnées aléatoires à l'origine est une méthode prudente qui peut servir à générer des hypothèses et à dégager des corrélats de la santé individuelle. Les interactions croisées permettent d'effectuer des analyses des effets contextuels sur des sous-groupes de population. Dans notre exemple de régression logistique, les données ont été traitées de manière hiérarchique selon un modèle de la probabilité des effets sur la santé individuelle. Cette méthode a pour inconvénient de ne pas tenir compte de la stratification géographique dans le plan d'échantillonnage complexe ni de l'autocorrélation spatiale.

Après avoir choisi une méthode de modélisation, on doit aussi tenir compte de la confusion que risque d'entraîner l'autocorrélation spatiale inhérente aux données. L'autocorrélation spatiale, ou corrélation entre les aires géographiques en raison de leur étroite proximité, peut confondre les estimations générées à partir des modèles qui lient la santé à l'environnement. Selon les spécialistes de la statistique spatiale, les données d'enquête concernant des collectivités ont une structure en treillis, car les unités d'analyse sont des aires qu'on peut dessiner sur une carte

au moyen de polygones. Les données d'enquête sur la santé peuvent comprendre des unités géographiques contiguës ou non contiguës. Lorsqu'il faut effectuer une analyse spatiale, on intègre les méthodes au modèle principal ou on les ajoute à titre d'étape supplémentaire au processus de modélisation. Les logiciels SAS, MLwiN et winBUGS offrent des possibilités d'analyse spatiale qui sont décrites dans leur mode d'emploi; il n'en sera pas question ici, mais ils feront l'objet d'une étude ultérieure.

L'analyse spatiale neutralise la corrélation qui existe entre les unités géographiques en raison de leur étroite proximité; on l'utilise en combinaison avec le modèle agrégé ou avec le modèle multiniveaux. Les profils spatiaux peuvent donner lieu à plus d'une interprétation et à différentes solutions analytiques. Selon les interprétations, on peut conclure que la corrélation avec les données du BRFSS est simplement due au fait qu'en raison d'une étroite proximité spatiale, il y a une plus grande similitude entre certains MSA voisins qu'entre des MSA éloignés, ou encore que la mise en grappes est due au fait que certaines régions possèdent une topographie, un climat et une culture semblables. La première interprétation peut nous inciter à neutraliser la corrélation spatiale; la deuxième peut constituer une inférence précieuse tirée de l'analyse et ne pas nécessiter la neutralisation de cette corrélation. Les données d'enquêtes sur la santé au niveau des collectivités peuvent présenter une certaine corrélation spatiale; les chercheurs doivent donc envisager l'opportunité d'effectuer une analyse spatiale fondée sur la question de recherche et l'interprétation des effets contextuels avec et sans neutralisation de la corrélation spatiale.

Dans le cadre de la recherche en santé publique, ces méthodes sont utiles lorsqu'il s'agit de mener des études écologiques qui utilisent les données d'enquêtes complexes avec des mesures contextuelles. Les méthodes présentées ici trouvent des applications dans la recherche en santé publique qui utilise à la fois des données d'enquête et des mesures contextuelles des facteurs environnementaux. Le modèle agrégé permet de lier les effets contextuels à la prévalence d'une collectivité à l'autre au moyen de méthodes et de logiciels traditionnels d'analyse d'enquête. Quant à l'analyse multiniveaux, elle permet de modéliser le plan d'échantillonnage complexe des enquêtes nationales sur la santé pour examiner l'incidence de l'environnement sur les comportements ou les résultats individuels en matière de santé. En outre, l'analyse spatiale neutralise la corrélation spatiale inhérente aux données géographiques.

REMERCIEMENTS

Nous tenons à remercier les personnes suivantes pour leur collaboration : Tom Schmid, William Kalsbeek, Fuzhong Li, Alastair Leyland, Michael Daniels, Constantine Gatsonis et Danny Pfefferman.

RÉFÉRENCES

Centers for Disease Control and Prevention (2001), "Physical Activity Trends--United States, 1990-1998", *Morbidity and Mortality Weekly Report*, 50, pp. 166-169.

Chou, S.Y., Grossman, M. et Saffer, H. (2001), "An Economic Analysis of Adult Obesity: Results from the Behavioral Risk Factor Surveillance System." Paper presented at the Third International Health Economics Association Conference, York, U.K.

Korn, E.L. et Graubard, B.I. (1999), *Analysis of Health Ssurveys*. New York, NY: John Wiley & Sons.

Masse L.C., Dassa, C., Gauvin, L., Giles-Corti, B. et Motl, R. (2002), "Emerging Measurement and Statistical Methods in Physical Activity Research", *American Journal of Preventive Medicine*, 23(2 Suppl.), pp. 44-55.

Raudenbush S.W. et Bryk, A.S. (2002), *The logic of hierarchical linear models. Hierarchical linear models: applications and data analysis methods*, Thousand Oaks, CA: Sage Publications.

Sallis, J.F., Bauman, A. et Pratt, M. (1998), "Environmental and Policy Interventions to Promote Physical Activity", *American Journal of Preventive Medicine*, 15, pp. 379-397.

Savageau, D. et D'Agostino, R. (2000), *Places Rated Almanac: Millenium Edition*, Foster City, CA: Macmillan General Reference USA.

Stokols D. (1992), "Establishing and Maintaining Healthy Environments: Toward a Social Ecology of Health Promotion", *American Psychology*, 47, pp. 6-22.

Stokols D. (1996), "Translating Social Ecologic Theory into Guidelines for Community Health Promotion", *American Journal of Health Promotion*, 10, pp. 282-298.