No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Recueil du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Dispersion des données en recherche contextuelle sur la santé de la population : Effets de la petite taille des groupes et de l'analyse par grappes sur les modèles multiniveaux linéaires et non linéaires

Philippa Clarke[1], Patricia O'Campo[2] et Blair Wheaton[3]

Résumé

L'usage courant des modèles multiniveaux pour examiner les effets du contexte environnant sur les résultats en matière de santé témoigne de leur valeur en tant que méthode statistique d'analyse de données groupées. Cependant, l'application de la modélisation multiniveaux à des données provenant d'enquêtes à l'échelle de la population est souvent limitée par le petit nombre de cas par unité de deuxième niveau, si bien que l'on relève dans la littérature sur les effets du quartier une tendance récente à appliquer des méthodes d'analyse par grappes, ou classification automatique, pour contourner le problème de la dispersion des données. Dans le présent article, nous utilisons des simulations de Monte Carlo pour étudier les effets des tailles marginales de groupe et des méthodes d'analyse par grappes sur la validité des estimations des paramètres dans les modèles multiniveaux linéaires ainsi que non linéaires.

MOTS-CLÉS : modèles multiniveaux, dispersion des données, analyse par grappes, simulations de Monte Carlo, étude par sondage

1. Introduction

1.1 Dispersion des données dans les modèles multiniveaux

La santé d'une population dépend d'interactions complexes entre les individus et les divers contextes sociaux, culturels et environnementaux dans lesquels ils se trouvent au cours de leur vie. Les environnements physique et social varient selon le quartier, l'école, la région et le pays. Ces contextes diversifiés dans lesquels se trouvent les personnes contribuent aux différences d'état de santé des populations, car la mortalité, la prévalence de la maladie, la fonctionnalité physique et la santé maternelle varient d'un contexte à l'autre (Yen et Syme, 1999; O'Campo et coll., 1997; Yen et Kaplan, 1999; Barr et coll., 2001; Clarke et George, 2005). Le recours de plus en plus fréquent à des modèles multiniveaux pour examiner les associations entre ces contextes au niveau du groupe et une vaste gamme d'indicateurs de la santé individuelle (p. ex., Pickett et Pearl 2001; Ahern, Pickett et Selvin, 2003; Buka et coll., 2003; Merlo, et coll., 2003; Ross et Mirowsky, 2001) témoigne de leur valeur en tant que méthode statistique d'analyse de données groupées ou en grappes. Cependant, l'application de la modélisation multiniveaux à des données provenant d'enquêtes à l'échelle de la population est souvent limitée par la *dispersion des données*, c'est-à-dire un petit nombre de cas par unité de deuxième niveau.

Si les enquêtes à grande échelle permettent d'obtenir assez facilement un nombre élevé de groupes, ces derniers ne contiennent souvent qu'un très petit nombre d'individus. Ainsi, dans le cycle 1.1 de l'Enquête sur la santé dans les collectivités canadiennes, il existe, en moyenne, environ 15 répondants par secteur de recensement (région géographique type utilisée comme approximation des quartiers). Le fait que 5,9 % des secteurs de recensement ne comptent qu'un seul répondant complique encore davantage le problème. Ces groupes « à un élément » sont préoccupants, parce qu'il n'existe aucune variabilité intragroupe, si bien qu'il est impossible de faire la distinction entre les variations au niveau individuel et au niveau du secteur de recensement. La dispersion est également manifeste dans les données d'enquêtes américaines. La National Longitudinal Study of Adolescent Health est une

[1] Philippa Clarke, University of Michigan, 426 Thompson Street, Ann Arbor, Michigan, 48106, USA (pjclarke@umich.edu);
[2] Patricia O'Campo, University of Toronto, 70 Richmond St. E., 4th Floor, Toronto, Canada.
[3] Blair Wheaton, University of Toronto, 725 Spadina Ave., Toronto, Canada.

enquête axée sur l'école réalisée aux États-Unis pour évaluer les comportements ayant une incidence sur la santé des adolescents et leurs résultats au début de l'âge adulte (Bearman, Jones et Udry, 2002). Toutefois, comme les écoles (et non les quartiers) constituent la base de sondage, la dispersion des données est considérable pour les chercheurs qui s'intéressent à l'étude des effets du quartier sur la santé. Au premier cycle de l'enquête, on dénombre, en moyenne, 7,33 sujets par secteur de recensement et près de la moitié de ces secteurs ne comptent qu'un seul sujet. En raison de l'érosion de l'échantillon et de la mobilité résidentielle, la dispersion des données ne fait que croître au cours des enquêtes avec suivi longitudinal.

Lorsque le niveau de mise en grappes dans les groupes est élevé (c.-à-d. grande taille de groupe), il est bien connu que la désagrégation des données (p. ex., par régression par les moindres carrés ordinaires (MCO)) accroît le risque d'erreurs de type I au cours de l'examen de l'effet du contexte du groupe sur la santé (Kreft, 1996). La supposition que les observations sont indépendantes produit des erreurs-types comportant un biais par défaut qui engendre des intervalles de confiance artificiellement étroits. Les modèles multiniveaux isolent correctement les effets intragroupe et intergroupes, de sorte qu'on tient compte statistiquement d'un haut niveau de mise en grappes dans les groupes. Toutefois, on en sait fort peu sur le seuil *inférieur* auquel la dispersion des données rend les modèles multiniveaux non fiables, voire même inutiles. Plusieurs règles empiriques sont énoncées dans la littérature, la plus fréquente étant le choix de 15 à 30 cas par groupe (Bryk et Raudenbush 1992; Kreft et de Leeuw 1998; Raudenbush et Sampson 1999). Cependant, les règles empiriques pourraient être citées et suivies sans qu'il existe de preuve réelle de ce qu'est vraiment le niveau minimal de dispersion des données. Et la grande étendue de nombres mentionnés dans ces recommandations révèle que fort peu d'études ont été réalisées pour tester explicitement le niveau minimal de mise en grappes nécessaire pour que les modèles multiniveaux produisent des estimations valides et fiables.

Des simulations conçues pour évaluer les problèmes de la dispersion des données commencent à être décrites dans la littérature et les résultats laissent entendre que l'obtention d'estimations sans biais et efficaces dépend plus du nombre de groupes que du nombre d'observations par groupe (Maas et Hox 2002; Maas et Hox 2004; Afshartous 1995; Kreft 1996; Mok, 1995). Cette constatation est rassurante, car l'existence d'un nombre insuffisant de groupes est un problème que posent rarement les données provenant d'enquêtes représentatives de la population. Pourtant, les chercheurs continuent de s'inquiéter de la petite taille des groupes lorsqu'ils étudient les effets contextuels sur la santé et adoptent diverses stratégies pour contourner le problème de dispersion des données. Certains choisissent tout bonnement de laisser de côté la structure hiérarchique des données en utilisant des méthodes de régression par les MCO (p. ex., Robert, 1998; South et Baumer, 2000; Schieman, Pearlin et Meersman, 2006), tandis que d'autres éliminent simplement les quartiers peu peuplés de leurs analyses (Cutrona et coll., 2000). On a également observé une tendance récente à utiliser des méthodes d'analyse par grappes (classification automatique) pour réduire la dispersion des données (Beland, Birch et Stoddart 2002; Buka et coll., 2003; Cutrona, et coll., 2000; Hou et Chen 2003; Wheaton et Clarke, 2003). Ces méthodes comportent habituellement une mise en grappes consistant à regrouper les répondants dans des quartiers « synthétiques » plus grands, afin d'obtenir un nombre plus élevé de cas par unité de deuxième niveau. Cependant, les effets de ce genre de méthodes sur l'exactitude des paramètres des modèles n'ont pas été examinés en détail. Lors de travaux antérieurs (Clarke et Wheaton, 2007), nous avons montré que les stratégies énergiques de mise en grappes ont la conséquence ironique d'effectivement minimiser la variance intergroupe à la suite de l'introduction d'une hétérogénéité intragroupe artificielle. La mise en grappes ciblée des groupes à un seul élément s'est avérée la stratégie la plus efficace en vue de réduire la dispersion des données tout en réduisant au minimum l'accroissement de la variance intragroupe (Clarke et Wheaton, 2007; Wheaton et Clarke, 2003). Ces travaux étaient axés sur des modèles hiérarchiques linéaires, mais l'effet de la mise en grappes sur les modèles non linéaires demeure inconnu.

L'objectif du présent article est d'examiner empiriquement les effets de la dispersion des données sur des modèles multiniveaux afin de pouvoir prendre des décisions analytiques éclairées lorsqu'on utilise des données groupées ou mises en grappes. À l'aide de simulation de Monte Carlo, nous étudions les effets des tailles marginales de groupe sur les estimations des paramètres dans le modèle multiniveaux. Nous examinons des modèles hiérarchiques linéaires et non linéaires afin de déceler les différences éventuelles entre les effets de la taille marginale de groupe sur des résultats continus ainsi que discrets. Puis, nous recourons à des méthodes d'analyse par grappes afin de minimiser la dispersion des données et d'examiner les conséquences sur les deux types de modèle dans les simulations. Nous soutenons que l'on pourrait tolérer une plus grande dispersion qu'il n'est généralement supposé et que toute mesure corrective devrait être appliquée avec prudence.

1.2 **Le modèle généralisé multiniveaux**

En général, le modèle multiniveaux peut être conceptualisé comme un système hiérarchique d'équations de régression dans J groupes contextuels contenant chacun N_j individus (Raudenbush et Bryk 2002). Au niveau individuel (niveau 1), nous avons la variable dépendante Y_{ij} et des équations de régression distinctes pour chaque groupe ou « quartier ». Pour les modèles linéaires, la fonction lien identité correspond à la régression de Y_{ij} sur un ensemble de prédicteurs linéaires comprenant une ou plusieurs variables indépendantes X_{ij}, dont les résidus (e_{ij}) suivent une loi normale de moyenne 0 et de variance σ^2 :

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \qquad (1a).$$

Pour les modèles non linéaires, diverses fonctions liens linéarisent une composante de prédicteurs non linéaires sous-jacents. Dans le cas d'un résultat binaire avec distribution binomiale de l'erreur, on utilise la fonction lien logit pour calculer la régression du logarithme du risque (*log odds*) de Y_{ij} sur un ensemble de variables indépendantes à prédicteurs linéaires :

$$Logit(Y_{ij}) = ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{0j} + \beta_{1j} X_{ij} \qquad (1b).$$

Tant pour les modèles linéaires que non linéaires, ces coefficients de niveau 1 peuvent ensuite être modélisés par des variables explicatives au niveau contextuel 2 (p. ex., pauvreté du quartier) :

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \qquad (2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \qquad (3).$$

En introduisant par substitution les équations (2) et (3) dans les équations (1a) et (1b), et en réarrangeant les termes, nous obtenons le modèle linéaire multiniveaux complet :

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + u_{0j} + u_{1j} X_{ij} + e_{ij} \qquad (4a),$$

et le modèle logistique multiniveaux complet :

$$log\,it(Y_{ij}) = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + u_{0j} + u_{1j} X_{ij} \qquad (4b).$$

Dans les deux modèles, u_{0j} représente la variabilité au niveau du groupe autour de l'ordonnée à l'origine suivant hypothétiquement une loi normale de moyenne 0 et de variance τ_{00}, et u_{1j} représente la variabilité au niveau du groupe autour de la pente de la régression suivant hypothétiquement une loi normale de moyenne 0 et de variance τ_{11}.

S'il n'existe aucune variable explicative au niveau 1 ou 2, les équations (4a) et (4b) se réduisent à :

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \qquad (5a),$$

$$log\,it(Y_{ij}) = \gamma_{00} + u_{0j} \qquad (5b),$$

qui sont les modèles entièrement inconditionnels, ou d'analyse de variance unidimensionnelle, pour les cas linéaire et logistique, respectivement. La distribution de la variance en ses composantes donne une statistique utile, le coefficient de corrélation intraclasse (CCI), qui mesure la proportion de la variance du résultat qui est expliquée au niveau du groupe (Raudenbush et Bryk 2002). Pour le modèle linéaire, le CCI est défini comme étant :

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \qquad (6a).$$

Puisque la loi binomiale pour la fonction lien logistique suppose une variance de niveau 1 de $\pi^2/3$ (Snijders et Bosker, 1999), le CCI pour le modèle non linéaire est défini comme étant :

$$\rho = \frac{\tau_{00}}{\tau_{00} + \pi^2/3}. \qquad (6b)$$

2. Méthodologie

2.1 Méthode de simulation

Une simulation de Monte Carlo est réalisée à l'aide d'un modèle hiérarchique à deux niveaux. Le nombre de groupes est maintenu constant à 200, et la taille de groupe varie, prenant les valeurs de 2, 5, 10 et 20 observations par groupe. Le nombre de groupes est choisi de façon à représenter le nombre le plus grand de groupes habituellement observés dans les données provenant d'enquêtes représentatives de la population, tandis que les tailles de groupes révèlent les cas extrêmes de dispersion de données, ainsi que les tailles de groupes plus grandes habituellement testées dans les simulations (Mok, 1995; Maas et Hox, 2004). Pour chacune de ces quatre conditions, 1 000 ensembles de données simulées ont été générés pour les modèles linéaires ainsi que non linéaires.

Les paramètres de la simulation sont établis d'après les modèles multiniveaux complets (Équations 4a et 4b). L'ordonnée à l'origine (γ_{00}) est fixée à 1,00 et les coefficients d'effet fixe ($\gamma_{10}, \gamma_{01}, \gamma_{11}$) sont fixés à 0,3, ce qui représente un effet de taille moyenne (Cohen, 1988). Un ensemble de valeurs de X et Z est généré aléatoirement à partir d'une loi uniforme standard. Comme l'ont fait Maas et Hox (2004), la variance résiduelle au niveau 1 (σ^2) est fixée à 0,5 dans le modèle linéaire. (Il n'existe pas de niveau 1 dans le modèle non linéaire.)

Les valeurs de population des composantes de la variance de niveau 2 sont calculées à l'aide de la formule pour une valeur de CCI de 0,1 (révélant le seuil inférieur de mise en grappes au niveau du groupe sur la variable de résultat habituellement observé dans les données provenant d'enquêtes représentatives de la population (Gulliford et coll., 1999)).Donc, pour le modèle non linéaire, la valeur de population de la variance de niveau 2 de l'ordonnée à l'origine (τ_{00}) est fixée à 0,366, d'après l'équation 6b. Pour simplifier, τ_{00} et τ_{11} sont contraintes d'être égales (à l'instar de Maas et Hox, 2004) et la covariance de niveau 2 (τ_{01}) est fixée à zéro. Pour le modèle linéaire, l'équation 6a est modifiée afin de tenir compte de l'hétéroscédasticité dans le terme d'erreur aléatoire u_{1j} (où il s'agit d'une fonction de la variable indépendante de niveau 1 (X_{ij})) (Goldstein, 1995; Mok, 1995) :

$$\rho = \frac{\text{var}(niveau\ 2)}{\text{var}(niveau\ 2) + \text{var}(niveau\ 1)} \qquad (7a),$$

où

$$\text{var}(niveau\ 2) = \tau_{00} + 2\tau_{01}x_{ij} + \tau_{11}x_{ij}^2, \qquad (7b)$$

et

$$\text{var}(niveau\ 1) = \sigma^2 \qquad (7c).$$

Donc, pour le modèle multiniveaux linéaire, la variance de population au niveau 2 est fixée à 0,44 (sur la base d'une valeur moyenne de x_{ij} de 0,5 générée aléatoirement à partir d'une loi uniforme). (Aucune formule équivalente du CCI ajusté n'est disponible pour les modèles non linéaires si bien que seule la formule du CCI résiduelle est utilisée.)

Partant de ces paramètres, les valeurs de Y_{ij} sont générées pour chacun des 4 000 ensembles de données simulées (Y est une variable continue pour le modèle linéaire et prend les valeurs de 0 ou 1 pour le modèle non linéaire), et les effets des quatre contraintes distinctes sur les valeurs estimées des paramètres sont examinés pour le modèle linéaire ainsi que non linéaire. Des suites reproductibles de nombres aléatoires ont été générées dans les simulations afin de maintenir la comparabilité entre les modèles. Toutes les simulations ont été réalisées en Mplus version 4.2 (Muthen et Muthen, 1998). Les modèles sont estimés par la méthode du maximum de vraisemblance et les erreurs-types sont calculées à l'aide d'un estimateur intercalé robuste. Les modèles non linéaires sont estimés à l'aide d'un algorithme d'intégration numérique.

2.2 Analyse par grappes

Après avoir généré les données simulées pour chaque contrainte, nous avons effectué une analyse par grappes disjointe pour rassembler les groupes en « quartiers synthétiques » plus grands ayant des caractéristiques semblables. La mise en grappes s'appuyait sur les similarités des groupes déterminées en se basant sur les variables au niveau du groupe générées dans les données simulées initiales. Nous avons utilisé la procédure FASTCLUS de SAS pour affecter les observations à une grappe, et une seule. (Les grappes ne forment pas une structure arborescente comme dans une analyse par grappes hiérarchique (Anderberg, 1973).) Les observations ont été regroupées en grappes en se basant sur les distances euclidiennes entre les valeurs des variables au niveau du

groupe. Nous avons utilisé un critère de distance strict pour nous assurer que le degré de similarité soit élevé avant que les groupes soient rassemblés en quartiers synthétiques.

2.3 Analyse statistique

Les effets de la petite taille du groupe dans la simulation sont examinés en fonction du biais pour toutes les estimations des paramètres et de leurs erreurs-types. Le biais est évalué en déterminant si la moyenne de la distribution d'échantillonnage des estimations sous chaque contrainte est centrée autour de la valeur réelle. Si $\hat{\theta}$ est l'estimation d'échantillon du paramètre de population θ, alors le biais = $\frac{E(\hat{\theta})-\theta}{\theta}$. Pour tout paramètre, un biais supérieur à 10 % est généralement considéré comme significatif (Muthen et Muthen, 2002). La précision des estimations est déterminée par l'examen de la distribution d'échantillonnage des erreurs-types pour chaque paramètre. L'écart-type des estimations des paramètres dans les simulations est un indicateur de l'erreur-type de population lorsque le nombre de répliques est grand (Muthen et Muthen, 2002). Il est comparé à la moyenne des erreurs-types estimée pour chaque estimation de paramètre dans les simulations, et le biais dans les erreurs-types est calculé de la même façon que pour les autres estimations de paramètres, tel qu'il est décrit plus haut.

3. Résultats

Le tableau 1 présente les effets fixes et les erreurs-types connexes pour les modèles multiniveaux linéaire et non linéaire pour les quatre contraintes de simulation. Les valeurs vraies des paramètres sont entre parenthèses dans le tableau 1. Pour chacune des quatre contraintes de simulation, les estimations des paramètres et les erreurs-types sont sans biais pour les variables de résultat continues ainsi que discrètes. Même dans les cas extrêmes de dispersion des données (taille de groupe = 2), le biais dans les estimations des paramètres à effets fixes est sans importance (moins de 2 % pour le modèle linéaire et moins de 6 % pour le modèle non linéaire) et il n'existe aucune preuve d'imprécision dans les estimations (les erreurs-types concordent avec les valeurs des paramètres).

Tableau 1. Résultats des simulations pour les modèles multiniveaux linéaire et non linéaire : effets fixes et erreurs-types

	Modèle multiniveaux linéaire				**Modèle multiniveaux non linéaire**			
	Effets fixes				Effets fixes			
Taille de groupe	$\hat{\gamma}_{00}$ (1,0)	$\hat{\gamma}_{10}$ (0,3)	$\hat{\gamma}_{01}$ (0,3)	$\hat{\gamma}_{11}$ (0,3)	$\hat{\gamma}_{00}$ (1,0)	$\hat{\gamma}_{10}$ (0,3)	$\hat{\gamma}_{01}$ (0,3)	$\hat{\gamma}_{11}$ (0,3)
2	0,999	0,302	0,301	0,302	1,030	0,305	0,320	0,318
5	1,000	0,303	0,299	0,301	1,002	0,299	0,302	0,303
10	1,000	0,301	0,300	0,299	1,000	0,299	0,303	0,305
20	1,000	0,300	0,300	0,299	1,000	0,306	0,301	0,299
	Erreurs-types				Erreurs-types			
Taille de groupe	$\hat{\gamma}_{00}$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{01}$	$\hat{\gamma}_{11}$	$\hat{\gamma}_{00}$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{01}$	$\hat{\gamma}_{11}$
2	0,040 (0,039)	0,042 (0,042)	0,039 (0,042)	0,042 (0,043)	0,176 (0,176)	0,157 (0,165)	0,142 (0,149)	0,162 (1,72)
5	0,027 (0,027)	0,028 (0,028)	0,027 (0,027)	0,028 (0,029)	0,097 (0,099)	0,097 (0,099)	0,091 (0,099)	0,100 (0,102)
10	0,022 (0,021)	0,022 (0,023)	0,022 (0,021)	0,022 (0,022)	0,073 (0,072)	0,074 (0,075)	0,071 (0,073)	0,076 (0,078)
20	0,019 (0,019)	0,019 (0,019)	0,018 (0,019)	0,019 (0,019)	0,059 (0,058)	0,060 (0,060)	0,059 (0,061)	0,061 (0,062)

Note : Les valeurs vraies des paramètres sont entre parenthèses.

Le tableau 2 donne les composantes de la variance et les erreurs-types pour les modèles multiniveaux linéaire ainsi que logistique pour les quatre contraintes de simulation. Lorsque la taille de groupe est très faible (taille de

groupe = 2), les composantes de la variance au niveau du groupe sont surestimées par le modèle tant linéaire que non linéaire. Dans le cas du modèle linéaire, la variance de l'ordonnée à l'origine aléatoire est surestimée de 14 % pour les tailles de groupe marginales, tandis que, dans le modèle non linéaire, la variance de l'ordonnée à l'origine et de la pente est surestimée de plus de 30 %. Les erreurs-types de ces estimations sont elles aussi surestimées dans les situations où les tailles de groupe sont marginales. Lorsque la taille de groupe est égale à 2, les erreurs-types présentent un biais par excès pouvant aller jusqu'à 80 % pour la variance de l'ordonnée à l'origine au niveau du groupe dans le modèle linéaire et jusqu'à 32 % pour la variance de la pente au niveau du groupe dans le modèle non linéaire. Par conséquent, la puissance de détection d'une variance intergroupe significative dans ces modèles est inférieure à 0,3.

Tableau 2. Résultats des simulations pour les modèles multiniveaux linéaire et non linéaire : effets aléatoires et erreurs-types

	Modèle multiniveaux linéaire			**Modèle multiniveaux non linéaire**	
	Effets aléatoires			Effets aléatoires	
Taille de groupe	$\hat{\sigma}^2$ (0,5)	$\hat{\tau}_{00}$ (0,044)	$\hat{\tau}_{11}$ (0,044)	$\hat{\tau}_{00}$ (0,366)	$\hat{\tau}_{11}$ (0,366)
2	0,487	0,050	0,048	0,485	0,520
5	0,501	0,042	0,042	0,358	0,370
10	0,500	0,043	0,043	0,358	0,366
20	0,500	0,043	0,043	0,361	0,359
	Erreurs-types			Erreurs-types	
Taille de groupe	$\hat{\sigma}^2$	$\hat{\tau}_{00}$	$\hat{\tau}_{11}$	$\hat{\tau}_{00}$	$\hat{\tau}_{11}$
2	0,057 (0,053)	0,058 (0,033)	0,031 (0,029)	0,556 (0,520)	0,855 (0,649)
5	0,027 (0,028)	0,016 (0,016)	0,015 (0,016)	0,181 (0,184)	0,213 (0,218)
10	0,018 (0,017)	0,010 (0,009)	0,010 (0,010)	0,105 (0,106)	0,122 (0,122)
20	0,012 (0,012)	0,007 (0,007)	0,007 (0,007)	0,070 (0,071)	0,078 (0,076)

Note : Les valeurs vraies des paramètres sont entre parenthèses.

Le tableau 3 donne les résultats pour les estimations des paramètres après une analyse par grappes des groupes. Pour simplifier, nous ne présentons les résultats que pour la contrainte où la taille de groupe originale était 2, puisque elle est la seule qui produit un biais dans les estimations des paramètres. Après l'analyse par grappes, la taille moyenne de groupe passe à 6,3 observations par groupe, avec un total final de 63 745 groupes. Cet accroissement de la taille de groupe n'introduit aucun biais dans les estimations des paramètres fixes ou de leurs erreurs-types pour les modèles à variable de résultat continue. Cependant, dans le modèle logistique, les effets fixes sont sous-estimés lorsqu'on utilise des données mises en grappe, particulièrement les effets au niveau du groupe (estimation avec un biais par défaut de 12 % pour l'ordonnée à l'origine et de 20 % pour γ_{01}). Les erreurs-types de ces estimations des effets fixes sont également systématiquement trop faibles dans le modèle non linéaire (biais par défaut de 26 % à 33 %). De surcroît, pour les deux types de modèles, la mise en grappes donne lieu à une sous-estimation importante des effets aléatoires au niveau du groupe et de leurs erreurs-types. Les variances de l'ordonnée à l'origine aléatoire et de la pente présentent toutes deux un biais par défaut de 72 % à 88 %. Dans le cas du modèle non linéaire, la mise en grappes a également introduit une covariance positive considérable dans les données. Cette sous-estimation des composantes de la variance au niveau du groupe est assortie d'une surestimation de la variance intragroupe dans les modèles linéaires (par excès de 14 %), ce qui laisse entendre que la méthode de mise en grappes a introduit une hétérogénéité intragroupe artificielle dans les données en regroupant des observations provenant de groupes différents.

Tableau 3. Résultats des simulations pour les modèles multiniveaux linéaire et non linéaire : après l'analyse par grappes des groupes de taille = 2

	Modèle multiniveaux linéaire		**Modèle multiniveaux non linéaire**	
Effets fixes	Estimation du paramètre	Erreur-type	Estimation du paramètre	Erreur-type
$\hat{\gamma}_{00}$	1,008 (1,000)	0,039 (0,039)	0,877 (1,000)	0,121 (0,176)
$\hat{\gamma}_{10}$	0,279 (0,300)	0,041 (0,042)	0,301 (0,300)	0,122 (0,165)
$\hat{\gamma}_{01}$	0,301 (0,300)	0,039 (0,042)	0,240 (0,300)	0,108 (0,149)
$\hat{\gamma}_{11}$	0,291 (0,300)	0,040 (0,043)	0,267 (0,300)	0,115 (0,172)
	Modèle multiniveaux linéaire		**Modèle multiniveaux non linéaire**	
Effets aléatoires	Estimation du paramètre	Erreur-type	Estimation du paramètre	Erreur-type
$\hat{\sigma}^2$	0,569 (0,500)	0,043 (0,053)	---	---
$\hat{\tau}_{00}$	0,005 (0,044)	0,018 (0,033)	0,095 (0,366)	0,103 (0,520)
$\hat{\tau}_{11}$	0,012 (0,044)	0,020 (0,030)	0,096 (0,366)	0,101 (0,649)

Note : Les valeurs vraies des paramètres sont entre parenthèses.

4. Discussion

Bien que la dispersion des données préoccupe de nombreux chercheurs, fort peu d'études ont visé à tester clairement le seuil général auquel cette dispersion pose un problème lorsqu'on veut obtenir des estimations sans biais et efficaces au moyen de modèles multiniveaux. Parallèlement, les chercheurs ne semblent pas avoir de réserves quant à l'utilisation de stratégies d'analyse par grappes (ou classification automatique) pour manipuler artificiellement les groupes contextuels dans les données. Au moyen de simulations de Monte Carlo, nous examinons empiriquement les effets de la dispersion des données dans les modèles multiniveaux et nous efforçons de comprendre les effets qu'ont sur l'exactitude des paramètres des modèles les stratégies d'analyse par grappes conçues en vue de contourner le problème de la dispersion des données.

Corroborant les résultats des travaux de recherche déjà réalisés dans ce domaine (Mok, 1995; Afshartous, 1995; Maas et Hox, 2002; Maas et Hox, 2004), notre étude indique que les modèles multiniveaux produisent des estimations sans biais des effets fixes et de leurs erreurs-types, même dans les cas extrêmes de dispersion des données, et ce aussi bien pour les variables de résultat continues que discrètes. Toutefois, nous dégageons des preuves que les effets aléatoires au niveau du groupe et leurs erreurs-types connexes sont surestimés si les tailles de groupe sont marginales. Par conséquent, la perte de précision de ces estimations réduit la puissance de détection d'une variance intragroupe significative lorsque les tailles de groupe sont très faibles. Quand le niveau de mise en grappes atteint au moins cinq observations par groupe, il n'existe aucune preuve d'un biais dans les estimations des effets fixes ou aléatoires et de leurs erreurs-types.

L'analyse par grappes est une stratégie efficace d'accroissement de la taille de groupe et de correction de la surestimation des effets aléatoires dans les modèles multiniveaux linéaires. Toutefois, il convient de recourir à ces stratégies avec prudence, afin d'éviter toute sous-estimation de la variance aléatoire due à l'introduction d'une hétérogénéité intragroupe artificielle. Les résultats discrets estimés à l'aide de modèles multiniveaux non linéaires ne sont pas des candidats robustes à l'application de méthodes d'analyse par grappes. Ces modèles sont extrêmement vulnérables à la sous-estimation des composantes tant fixes qu'aléatoires et de leurs erreurs-types. Le regroupement des données dispersées lors de l'examen de résultats discrets produirait des résultats incorrects et des conclusions invalides au sujet des effets des contextes sociaux sur la santé.

Remerciements

Cette étude a été financée par l'Initiative stratégique Méthodes et outils de recherche sur la santé publique et la santé des populations – Subventions de projets pilotes des Instituts de recherche en santé du Canada.

Références

Afshartous, David. (1995). "Determination of Sample Size for Multilevel Model Design". Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (Available at http://www.bus.miami.edu/~dafshart/)

Ahern, Jennifer, Pickett Kate E, Selvin, S., Abrams, B. (2003). "Preterm Birth among African American and White Women: A Multilevel Analysis of Socioeconomic Characteristics and Cigarette Smoking." *Journal of Epidemiology and Community Health*, 57:606-611.

Anderberg, Michael R. (1973). *Cluster Analysis for Applications.* New York: Academic Press, Inc.

Barr RG, Diez Roux AV, Knirsch, CA, Pablos-Mendez A. (2001). "Neighborhood poverty and the resurgence of tuberculosis in New York City, 1984-1992". *American Journal of Public Health* 19(9):1487-1493.

Bearman Peter S, Jones Jo, et Udry J. Richard. (2002). "The National Longitudinal Study of Adolescent Health: Research Design." University of North Carolina Population Center, URL: http//www.cpc.unc.edu/addhealth,.

Beland Francois, Birch Stephen, Stoddart Greg. (2002). "Unemployment and Health: Contextual-Level Influences on the Production of Health in Populations." *Social Science and Medicine*, 55:2033-2052.

Bryk Anthony S. et Raudenbush Stephen W. (1992). *Hierarchical Linear Models: Applications And Data Analysis Methods*. Newbury Park, CA: Sage.

Buka Stephen L, Brennan Robert T, Rich-Edwards Janet W, Raudenbush Stephen W, Earls Felton. (2003). "Neighborhood Support and the Birth Weight of Urban Infants." *American Journal of Epidemiology*, 157:1-8.

Cohen, Jacob. (1988). *Statistical Power Analysis*. Hillsdale, New Jersey: Erlbaum.

Clarke, Philippa and George, Linda K. (2005). "The role of the built environment in the Disablement Process." *American Journal of Public Health*; 95(11):1933-1939.

Clarke, Philippa, et Wheaton, Blair. 2007. "Addressing data sparseness in contextual population research: Using cluster analysis to create synthetic neighborhoods." *Sociological Methods and Research*, 35(3), In Press.

Cutrona, Carolyn E, Russell Daniel W, Hessling Robert M, Brown P Adama, Murry Velma. (2000). "Direct and Moderating Effects of Community Context on the Psychological Well-Being of African American Women." *Journal of Personality and Social Psychology*, 79(6):1088-1101.

Goldstein Harvey. (1995). *Multilevel Statistical Models*. London: Edward Arnold, New York: Halsted.

Gulliford, Martin C., Ukoumunne, Obioha C., Chinn Susan. (1999). "Components of Variance and Intraclass Correlations for the Design of Community-based Surveys and Intervention Studies." *American Journal of Epidemiology*, 149:876-883.

Hou Feng, Chen Jaijian. (2003). "Neighborhood Low Income, Income Inequality and Health in Toronto." *Health Reports*, 14(2):21-34.

Wheaton, Blair et Philippa Clarke. (2003). “Space meets time: Integrating temporal and contextual influences on mental health in early adulthood.” *American Sociological Review*;68:680-706.

Yen IH, Kaplan GA. (1999). “Neighborhood social environment and risk of death: multilevel evidence from the Alameda County Study.” *American Journal of Epidemiology*, 149:898-907

Yen IH, Syme SL. (1999). “The social environment and health: A discussion of the epidemiologic literature.” *Annual Review of Public Health*; 20:287-308.