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

Modèles mixtes linéaires et non linéaires de niveau agrégé pour l'estimation sur petits domaines d'après des dénombrements binaires provenant d'enquêtes

A.C. Singh[1] et F. Verret[2]

Résumé

Nous proposons un modèle linéaire généralisé avec composantes aléatoires additives (GLMARC pour *generalized linear model with additive random components*) de niveau agrégé applicable aux dénombrements binaires provenant d'enquêtes. Ce modèle comporte une partie linéaire (pour les effets aléatoires) et une partie non linéaire (pour les effets fixes) pour la modélisation de la fonction de moyenne et appartient donc à la classe des modèles mixtes linéaires et non linéaires (MLNL). Il permet d'adopter une approche de type modèle linéaire mixte (LMM) pour l'estimation sur petits domaines (EPD) semblable dans une certaine mesure à la méthode bien connue de Fay-Herriot (1979) et tient donc compte du plan d'échantillonnage. Contrairement à l'approche bayésienne hiérarchique (BH) de You et Rao (2002), la méthode proposée donne lieu à des estimations sur petits domaines et des diagnostics fréquentistes faciles à interpréter, ainsi qu'à un auto-étalonnage reposant sur des estimations directes fiables sur grands domaines. La méthodologie habituelle des LMM ne convient pas dans le cas de dénombrements, à cause de l'absence de contraintes d'intervalle pour la fonction de moyenne et de la possibilité d'obtenir des estimations non raisonnables (p. ex. 0 dans le contexte de l'EPD) des composantes de variance, car le modèle ne permet pas que la partie des effets aléatoires de la fonction de moyenne conditionnelle dépende de la moyenne marginale. La méthode proposée est une amélioration de la méthode élaborée antérieurement par Vonesh et Carter (1992) qui s'appuyait aussi sur des modèles mixtes linéaires et non linéaires, mais qui ne tenait pas compte de la relation entre la variance et la moyenne, quoique cela se fasse habituellement par des contraintes d'intervalle pour les effets aléatoires. En outre, les effets du plan de sondage et l'estimation des effets aléatoires n'étaient pas envisagés. En revanche, dans notre application à l'estimation sur petits domaines, il est important d'obtenir des estimations appropriées des effets fixes ainsi qu'aléatoires. Il convient de souligner que, contrairement au modèle linéaire mixte généralisé (GLMM), le modèle GLMARC se caractérise, comme les LMM, par une grande simplicité d'ajustement. Cette dernière est due au remplacement des effets fixes et aléatoires originaux du GLMM par un nouvel ensemble de paramètres du GLMARC dont l'interprétation est assez différente, car l'effet aléatoire n'est plus intégré dans la fonction prédictrice non linéaire. Toutefois, cela n'a aucune conséquence pour estimation sur petits domaines, parce que les paramètres de petit domaine correspondent aux moyennes conditionnelles globales et non aux paramètres individuels du modèle. Nous proposons pour l'estimation des paramètres une méthode s'appuyant sur l'application itérative du meilleur prédicteur linéaire sans biais (BLUP pour *best linear unbiased predictor*) qui permet l'auto-étalonnage après un agrandissement approprié du modèle. Le problème des petits domaines pour lesquels la taille d'échantillon est faible, voire nulle, ou des estimations directes nulles est résolu en regroupant les domaines pour l'étape de l'estimation des paramètres uniquement. L'application du modèle à l'Enquête sur la santé dans les collectivités canadiennes de 2000-2001 en vue d'estimer la proportion de fumeurs quotidiens dans les sous-populations définies par les régions sociosanitaires provinciales selon le groupe âge-sexe est présentée à titre d'illustration.

MOTS-CLÉS : regroupement des domaines; contraintes d'intervalle; auto-étalonnage; relation variance-moyenne

1. Introduction

Le problème de l'application de modèles mixtes non linéaires de niveau agrégé à l'estimation sur petits domaines examiné dans le présent article s'est posé dans le contexte de la production d'estimations du nombre de fumeurs quotidiens par groupe âge-sexe dans les régions sociosanitaires de la province de l'Île-du-Prince-Édouard d'après les données de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) de 2000-2001. Soit $t_{y,d}$ l'estimation directe du total de petit domaine $T_{y,d}$ de la variable résultat y pour le domaine d représentant la sous-population (r, a, g), où r est la r^{e} région sociosanitaire, a est le a^{e} groupe d'âge (12 à 19 ans, 20 à 29 ans, 30 à 44 ans, 45 à 64 ans et 65 ans et plus) et g représente le sexe (masculin ou féminin). En pratique, dans le cas de données d'enquête, on utilise souvent un modèle linéaire mixte (LMM) de niveau agrégé qui permet de tenir compte des plans d'échantillonnage sous-jacents habituellement non ignorables pour le modèle; voir, par exemple, la méthode bien connue de Fay et Herriot (1979). Ici, l'expression « niveau agrégé » signifie que les covariables du modèle sont

[1] Division de la recherche et de l'innovation en statistique, Statistique Canada, Ottawa K1A 0T6
[2] Division des méthodes d'enquête auprès des ménages, Statistique Canada, Ottawa K1A 0T6

évaluées au niveau du domaine. Par exemple, sous un modèle de moyenne commune de sous-groupes, la moyenne $\mu_{y,d}$ pour le d^e domaine est modélisée par

$$\mu_{y,d} = \nu_{y,d} + \eta_d, \ \nu_{y,d} = A'_{x,d}\beta \qquad (1.1)$$

où $A_{x,d}$ est un vecteur de dimension D de covariables indicatrices $1_{d,ag}, d = 1,...,D;$ $1_{d,ag} = 1$ si d appartient au sous-groupe âge-sexe (a, g) et 0 autrement, et le paramètre fixe β est de dimension q, représentant le nombre total (dix dans notre exemple) de sous-groupes âge-sexe. Le terme d'erreur du modèle η_d est l'effet aléatoire que l'on suppose être $iid \ \ N(0, \sigma_\eta^2)$. On pourrait aussi avoir d'autres covariables, telles que le nombre moyen d'hospitalisations dans le domaine *d* pour lesquelles l'asthme est le diagnostic principal. En dénotant les totaux $N_d A_{x,d}$ par $T_{x,d}$, où N_d est le dénombrement de sous-population du domaine d, nous pouvons écrire le modèle pour $t_{y,d}, d = 1, ..., D$ sous la forme

$$\begin{aligned} t_{y,d} &= N_d (A'_{x,d}\beta + \eta_d) + e_{y,d} \\ &= T'_{x,d}\beta + N_d \eta_d + e_{y,d} \end{aligned} \qquad (1.2)$$

où $e_{y,d}$ est l'erreur d'observation ou d'échantillonnage en supposant que la variance V_d est approximativement connue. Nous supposons que la taille d'échantillon n_d pour le domaine d n'est pas nulle ou qu'elle est très faible, de sorte que $t_{y,d}$ n'est pas nul. En pratique, les domaines peuvent être regroupés en se fondant sur des similarités en ce qui concerne $A_{x,d}$ pour éviter ce problème, ainsi que pour lisser la matrice de covariance $\tilde{V}$ du vecteur regroupé t_y de dimension $\tilde{D}$ afin de le traiter comme s'il était connu; voir Singh (2006) pour plus de détails.

Bien que l'utilisation d'un LMM pour résoudre le problème susmentionné ait plusieurs avantages pratiques, dont une approche semi-parmétrique pour le BLUP ne comportant des hypothèses qu'au sujet des deux premiers moments (ce qui permet un effet de surdispersion pour favoriser la robustification), l'auto-étalonnage des estimations sur petits domaines d'après des estimations directes fiables sur de grands domaines afin de réduire le risque de défaillance du modèle et de diagnostics fondés sur des innovations résultant de l'utilisation du filtre de Kalman pour calculer les BLUP (voir Singh, 2006), elle possède quelques limites importantes. Premièrement, le modèle (1.1) pour les moyennes de petit domaine n'impose aucune contrainte d'intervalle pour les paramètres $\mu_{y,d}$, ce qui n'est pas raisonnable dans le cas de données discrètes telles que les dénombrement ou les proportions. Dans notre exemple axé sur les proportions de personnes chez lesquelles, dans une région sociosanitaire particulière, on a posé le diagnostic de diabète, le paramètre $\mu_{y,d}$ est nécessairement compris entre 0 et 1. Pour le modèle de moyenne commune de sous-groupes (1.1) susmentionné, $A'_{x,d}\beta$ se situe dans l'intervalle (0, 1), mais en cas de covariables plus générales, il n'est pas nécessaire qu'il en soit ainsi et le BLUP (qui est compris entre l'estimation directe $N_d^{-1} t_{y,d}$ et l'estimation indirecte ou synthétique $A'_{x,d}\hat{\beta}$, du moins si V est diagonale) ne doit pas nécessairement être compris dans l'intervalle admissible, ce qui rend l'optimalité du BLUP fragile dans cette application du LMM. En outre, l'hypothèse selon laquelle la variance commune σ_η^2 peut remplacer l'erreur de modélisation η_d dans (1.1) peut être une approximation sursimplifiante, car, pour des données discrètes, l'intervalle de valeurs possibles de η_d dépend de la moyenne marginale $\nu_{y,d}$, si bien que sa variance devrait dépendre de la moyenne $\nu_{y,d}$; par exemple, η_d pourrait être de la forme $f(\nu_{y,d})\zeta_d$ pour une certaine fonction lisse $f(\cdot)$ et ζ_d étant $iid \ \ N(0, \sigma_\xi^2)$. Sans ce genre de généralisation du LMM, l'estimation par le maximum de vraisemblance restreint (REML) habituelle de σ_η^2 pourrait être nulle, ce qui est clairement inapproprié, puisque l'on sait que le modèle (1.1) est imparfait, c'est-à-dire que $\sigma_\eta^2 > 0$.

Une solution naturelle du problème susmentionné consiste à considérer un LMM généralisé (GLMM) au lieu du LMM pour l'estimation sur petits domaines. Pour le GLMM, les approches fréquentistes sont généralement axées sur

l'estimation des paramètres fixes de premier et deuxième ordres et non sur les effets aléatoires qui sont évidemment nécessaires pour l'estimation sur petits domaines. Une bonne solution fréquentiste du problème de l'estimation des effets aléatoires pour le GLMM est généralement assez difficile à obtenir; voir Jiang et Lahiri (2006) pour une revue récente. Par ailleurs, les méthodes bayésiennes hiérarchiques conjuguées à des méthodes MCMC peuvent être utilisées pour l'estimation des paramètres fixes et aléatoires; voir Rao (2003, ch. 10). Dans le cas de données d'enquête et de plans de sondage non ignorables, le problème de l'estimation sur petits domaines au moyen d'un GLMM au niveau de l'unité est encore plus compliqué et est le sujet de travaux de recherche courants. Cependant, pour le GLMM de niveau agrégé, You et Rao (2002) ont proposé une extension bayésienne hiérarchique de la méthode de Fay-Herriot. Bien qu'elle soit utile, cette extension ne semble pas permettre d'obtenir la propriété désirable d'auto-étalonnage dans le cadre hiérarchique bayésien. En outre, les estimations sur petits domaines résultantes n'offrent ni l'interprétation pratique simple ni les diagnostics de modélisation. Il serait par conséquent utile de disposer d'une approche de type LMM du GLMM au niveau agrégé pour l'estimation sur petits domaines.

L'objectif du présent article est de proposer une nouvelle méthode fondée sur un GLM avec composantes aléatoires additives (GLMARC) dans lequel la partie aléatoire de la moyenne est linéaire et s'ajoute à la partie non linéaire fixe de la fonction de moyenne. Les modèles de type GLMARC peuvent être considérés comme un cas particulier des modèles mixtes linéaires et non linéaires (MLNL). Les approches antérieures des modèles MLNL s'appuyaient sur le cadre asymptotique « pour petit sigma » (*small-sigma asymptotics*) (c.-à-d. avec σ_η faible) qui n'est peut-être pas raisonnable pour les applications à l'estimation sur petits domaines. En outre, ces approches visaient principalement à estimer les paramètres fixes de la fonction de moyenne et les composantes de la variance, mais non les effets aléatoires.

Dans le présent article, nous proposons un nouveau modèle MLNL dans lequel les paramètres β et η sont tous deux remplacés pour éviter l'hypothèse asymptotique que sigma est petit. Cette approche est raisonnable dans le cas de l'estimation sur petits domaines, parce que les paramètres originaux β et η du modèle ne présentent pas d'intérêt direct. À la section 2, nous passons en revue les approches existantes des modèles MLNL fondés sur l'asymptotique pour petit sigma et exposons les motifs qui nous ont poussés à proposer la méthode. À la section 3, nous décrivons le modèle GLMARC proposé et présentons une méthode BLUP itérative pour l'estimation des paramètres. Nous montrons aussi comment il est possible d'étendre les covariables incluses dans le GLMARC afin de rendre possible un auto-étalonnage analogue au cas du LMM. Nous présentons des diagnostics s'appuyant sur des innovations en supposant que les estimations directes $t_{y,d}$ sont approximativement normales après le regroupement des domaines, au besoin. À la section 4, nous décrivons une application de la méthode proposée aux données de l'ESCC de 2000-2001. Enfin, nous présentons nos conclusions à la section 5.

2. Modèles MLNL alternatifs sans asymptotique pour petit sigma

Les modèles mixtes linéaires et non linéaires (MLNL) ont été considérés auparavant par plusieurs auteurs qui cherchaient uniquement à estimer les paramètres fixes de la fonction de moyenne et les composantes de la variance. L'estimation des effets aléatoires ne présentait pas d'intérêt, contrairement au problème de l'estimation sur petits domaines examiné ici. Par exemple, McCullagh et Nelder (1989, ch. 14) ont utilisé une linéarisation de Taylor de premier ordre sous asymptotique pour petit sigma (c.-à-d. $\sigma_\eta^2 \approx 0$) afin de transformer approximativement la fonction prédictrice non linéaire du GLMM en un modèle MLNL en vue de simplifier les calculs. Cependant, ils ont formulé une mise en garde contre l'hypothèse d'un petit sigma qui n'est vraisemblablement pas défendable; voir aussi Drum et McCullagh (1993). Sutradhar et Rao (1996) ont utilisé une linéarisation de Taylor de deuxième ordre sous asymptotique pour petit sigma afin de saisir les termes d'interaction aléatoires dans les cas où il existe plus d'un facteur aléatoire important. Les principales limites des approches fondées sur l'hypothèse d'un petit sigma susmentionnées tiennent au fait que, dans le contexte de l'estimation sur petits domaines, les composantes de la variance ne sont généralement pas faibles, d'une part, et qu'aucune contrainte d'intervalle n'est imposée aux effets aléatoires, donc aux composantes de la variante, d'autre part, ce qui pourrait entraîner une structure de covariance inadmissible des erreurs. Un autre cadre de modélisation MLNL ne s'appuyant pas sur les petits sigmas est manifestement nécessaire. Par conséquent, Vonesh et Carter (1992) ont proposé l'utilisation directe d'un modèle MLNL qui, dans le cas des données binomiales considérées dans le présent article, peut s'écrire sous la forme :

$$t_{y,d} = N_d[v_{y,d}(\alpha) + \zeta_d^*] + e_{y,d}\,, \qquad e_y \sim (0, V), \quad \zeta_d^* \sim^{iid} N(0, \sigma^2_{\zeta^*}) \tag{2.1}$$

et $$\text{logit}\ v_{y,d}(\alpha) = A'_{x,d}\alpha$$

en adoptant la notation introduite à la section 1 pour le LMM de niveau agrégé. Soulignons que les paramètres de régression fixes β sont remplacés par α, autrement dit que, dans (2.1), nous modélisons la moyenne marginale $v_{y,d}$ au lieu de la moyenne conditionnelle $\mu_{y,d}$, sachant η_d, qui dans le cas du logit est exprimée par

$$\text{logit}\ \mu_{y,d} = A'_{x,d}\beta + \eta_d\,. \tag{2.2}$$

En outre, le nouvel effet aléatoire ζ_d^* se trouve maintenant hors de la fonction prédictive non linéaire sous forme de terme additif. Le modèle (2.1) est fondamentalement un nouveau modèle ne requérant pas l'hypothèse d'un petit sigma dont l'utilisation pour remplacer le modèle (2.2) devrait être valide pour le problème qui nous occupe. De surcroît, aucune équivalence entre les anciens paramètres (β, η) et les nouveaux (α, ζ) n'est stipulée. Cependant, une limite importante du modèle (2.1) tient au fait qu'aucune contrainte d'intervalle n'est imposée à l'effet aléatoire ζ_d^* et que la composante de variance $\sigma^2_{\zeta*}$ ne tient donc pas compte de la relation entre la variance et la moyenne. La moyenne conditionnelle sachant ζ_d^* sous (2.1) est $\mu_{y,d} = (v_{y,d}(\alpha) + \zeta_d^*)$ qui doit être comprise entre 0 et 1. Donc, les contraintes d'intervalle sur ζ_d^* prennent la forme

$$v_{y,d}(\alpha) \le \zeta_d^* \le 1 - v_{y,d}(\alpha) \tag{2.3}$$

ce qui implique que $\sigma^2_{\zeta*}$ devrait dépendre de $v_{y,d}(\alpha)$ au lieu d'être constante pour tout d. Pour le GLMARC proposé à la section suivante, nous considérons que ζ_d^* est de la forme $f(v_{y,d})\zeta_d$, où la forme fonctionnelle de $f(\cdot)$ est motivée par le développement en série de premier ordre de Taylor du GLMM sous (2.2) autour de $\eta_d = 0$, soit

$$\mu_{y,d} := \text{inv logit}\ (A'_{x,d}\beta + \eta_d) = \mu_{y,d}(\beta, \eta_d = 0) + \mu_{y,d}(\beta, \eta_d^*)(1 - \mu_{y,d}(\beta, \eta_d^*))\eta_d \tag{2.4}$$

où η_d^* est compris entre 0 et η_d. Nous devons exprimer (2.4) sous une forme approximativement équivalente faisant intervenir la moyenne marginale $v_d(\alpha)$ et le nouvel effet aléatoire ζ_d, parce que η_d^* n'est pas connu. L'expression (2.4) donne à penser que l'on peut considérer que la forme de travail de la fonction $f(v_{y,d})$ est $v_{y,d}(1 - v_{y,d})$, puis corriger la composante de variance σ^2_η en introduisant le nouvel effet aléatoire ζ_d avec la nouvelle variance σ^2_ζ. Autrement dit, pour le modèle proposé à la section suivante, nous définissons

$$\mu_{y,d} = v_{y,d}(\alpha) + v_{y,d}(\alpha)(1 - v_{y,d}(\alpha))\zeta_d \tag{2.5}$$

où l'effet aléatoire ζ_d peut satisfaire les contraintes d'intervalle

$$-(1 - v_{y,d})^{-1} \le \zeta_d \le v_{y,d}^{-1}\,. \tag{2.6}$$

Les contraintes susmentionnées font dépendre la variance de ζ_d de d par la voie de $v_{y,d}$, ce qui est plutôt incommode, malgré le facteur $f(v_{y,d})$. Une solution consiste à formuler l'hypothèse plus forte que $-1 < \zeta_d < 1$ pour tout d, ce qui implique en fait que l'expression (2.6) est vraie pour tous les $v_{y,d}$ et que $0 \le \sigma^2_\zeta < 1$. (Soulignons que, selon notre expérience, l'hypothèse $|\zeta_d| < 1$ n'est pas raisonnable en pratique.) Compte tenu des contraintes d'intervalle, nous supposerons pour simplifier que les ζ_d sont approximativement i.i.d. $N(0, \sigma^2_\zeta)$, parce qu'ils ne sont pas contraints d'être compris dans (-1, 1).

Nous pouvons justifier l'existence des effets aléatoires susmentionnés de plusieurs façons. Habituellement, dans le cas de données binomiales, la loi bêta binomiale est utilisée pour les dénombrements. Donc, si nous supposons que $\mu_{y,d} \sim \text{Beta}(a_d, b_d)$ dans l'intervalle (0, 1) pour certains $a_d, b_d > 0$, nous avons

$$E(\mu_{y,d}) = \frac{a_d}{a_d + b_d}, \quad V(\mu_{y,d}) = \frac{a_d b_d}{(a_d + b_d)^2(a_d + b_d + 1)}\,. \tag{2.7}$$

Il s'ensuit que, pour la forme générale $f(v_{y,d})$,

$$a_d = v_{y,d}[v_{y,d}(1-v_{y,d})f^{-2}(v_{y,d})\sigma_\zeta^{-2} - 1] \quad (2.8a)$$

et

$$b_d = (1-v_{y,d})[v_{y,d}(1-v_{y,d})f^{-2}(v_{y,d})\sigma_\zeta^{-2} - 1]. \quad (2.8b)$$

Soulignons qu'en choisissant $f(v_{y,d})$ comme $v_{y,d}(1-v_{y,d})$, la condition $\sigma_\zeta^2 < 1$ est suffisante pour s'assurer que a_d ainsi que $b_d > 0$, ce qui justifie la formulation (2.5). Alternativement, si nous supposons que $[v_{y,d}(1-v_{y,d})]^{-1}(\mu_{y,d} - v_{y,d})$ suit une loi normale tronquée sur (-1, 1) de moyenne 0 et de variance σ_ζ^2, alors nous obtenons un ζ_d approximativement $N(0, \sigma_\zeta^2)$. Cette hypothèse est celle que nous faisons dans le modèle proposé, car elle s'avère utile et commode pour les diagnostics de modélisation sous l'hypothèse supplémentaire de normalité approximative de l'erreur d'observation $e_{y,d}$. Cependant, pour les estimations de type BLUP suivies de l'estimation de leur EQM, les hypothèses de normalité susmentionnées ne sont pas nécessaires.

3. Prédiction BLUP itérative pour le GLMARC : la méthode proposée

3.1 Le modèle

La méthode proposée pour l'estimation sur petits domaines à l'aide de données de dénombrement binaires est fondée sur le modèle GLMARC qui suit mentionné à la section précédente. Pour $d = 1,...,D$,

$$t_{y,d} = N_d[v_d(\alpha) + v_d(\alpha)(1-v_d(\alpha 1)\zeta_d] + e_{y,d} \quad (3.1)$$

où $\text{logit}\, v_d(\alpha) = A'_{x,d}\alpha$, $e_y \sim N(0, \sigma_0^2 V)$, $\zeta_d \sim^{iid} N(0, \sigma_0^2\tau_\zeta^2)$, $\sigma_\zeta^2 < 1$ et σ_0^2 est le paramètre de sur/sous-dispersion représentant l'ajustement à la structure de covariance dû à l'omission éventuelle de covariables ou d'effets aléatoires. Il se peut que, pour certains domaines, les tailles réalisées d'échantillon n_d soient nulles ou petites, ou que les produits $n_d N_d^{-1} t_{y,d}$ ne soient pas suffisamment grands pour approximer la normalité de $t_{y,d}$. Pour surmonter ce problème, nous commençons par regrouper les domaines similaires en nous fondant sur la proximité en ce qui a trait aux covariables $A_{x,d}$ afin que le nombre de domaines soit réduit à $\tilde{D}, \tilde{D} < D$, puis nous estimons les paramètres fixes du modèle α, σ_ζ^2 et σ_0^2. Lorsque ces paramètres du modèle sont estimés, nous estimons les effets aléatoires particuliers à d correspondants aux domaines non regroupés, c'est-à-dire en préservant l'identité des paramètres de petit domaine particuliers au domaine; voir, p. ex., Singh (2006). La condition que $|\zeta_d| < 1$ afin de s'assurer que les paramètres de petit domaine $\mu_{y,d}$ satisfont les contraintes d'intervalle n'est pas imposée initialement durant l'estimation par la méthode BLUP, mais une fois que les $\hat{\zeta}_d$ sont obtenus, ils sont tronqués, au besoin, pour satisfaire cette condition. Ce genre de troncature pourrait ne pas être forcément nécessaire en pratique et ne devrait, en principe, pas compromettre gravement l'optimalité de la prédiction BLUP. De surcroît, les estimations habituelles de l'EQM pour le BLUP peuvent encore être utilisées en un sens conventionnel.

3.2 Estimation des paramètres par la méthode BLUP itérative (IBLUP)

Sachant la surdispersion σ_0^2, la méthode IBLUP proposée peut être utilisée pour estimer α, ζ_d et σ_ζ^2. En fait, il s'avère que, comme σ_0^2 est un facteur multiplicatif pour V et σ_ζ^2, les estimations de α et ζ n'en dépendent pas, mais que leurs variances et l'estimation de σ_ζ^2 en dépendent. Si le modèle était LMM, les ($\sigma_0^2, \sigma_\zeta^2$) pourraient l'une et l'autre être estimées par la méthode REML, qui peut être simplifiée davantage en utilisant la vraisemblance profil pour l'estimation de σ_ζ^2. Il en est ainsi parce que l'on peut obtenir le REML de σ_0^2 sous une forme explicite en tant que moyenne des résidus standardisés qui sont eux-mêmes des fonctions de σ_ζ^2, voir, p. ex., Harvey (1989, ch. 4) pour une approche analogue dans le contexte du filtre de Kalman. Pour le GLMARC, nous pouvons appliquer l'idée

susmentionnée de manière itérative. Dans la suite, pour simplifier, nous posons que $\sigma_0^2 = 1$ et nous décrivons les étapes de la modification du BLUP en IBLUP pour le cas du GLMARC. Ici, les paramètres fixes α, σ_ζ^2 sont estimés pour commencer. Puis, sachant α, σ_ζ^2, les estimations BLUP de $\zeta_d's$ sont calculées.

Étape 0 : Calcul de l'estimation initiale $\alpha^{(0)}$

Poser que $\sigma_\zeta^{2(0)} = 0$ et trouver une estimation convergente $\alpha^{(0)}$ en utilisant les moindres carrés repondérés itérativement (IRLS pour *iteratively reweighted least squares*) standard du GLM. Plus précisément, à l'itération r, la fonction de moyenne $v_d(\alpha)$ est linéarisée en $\alpha = \alpha_r$ sous la forme

$$v_d(\alpha) \approx v_d(\alpha_r) + v_d(\alpha_r)(1 - v_d(\alpha_r))A'_{x,d}(\alpha - \alpha_r) \tag{3.2}$$

qui donne le modèle linéaire approximatif pour $t_{y,d}$ sous la forme

$$t_{y,d} \approx b_{d0(r)}N_d + b_{d1(r)}T'_{x,d}\alpha + e_{y,d} \tag{3.3}$$

où $b_{d0(r)} = v_d(\alpha_r) - b_{d1(r)}A'_{x,d}\alpha_r$, $b_{d1(r)} = v_d(\alpha_r)(1 - v_d(\alpha_r))$.

À $r = 0$, nous n'avons pas besoin de α_0 pour estimer $v_d(\alpha)$. Nous estimons plutôt $v_d(\alpha_0)$ directement d'après la proportion observée en apportant la correction de continuité $(n_d N_d^{-1} t_{y,d} + .5)/(n_d + 1)$ et nous calculons facilement $A'_{x,d}\alpha_0$ en tant que logarithme de $v_d(\alpha_0)/(1 - v_d(\alpha_0))$. Ensuite, avec la variable dépendante ajustée $t_{y,d}$ - $b_{d0(r)}$, nous appliquons la méthode IRLS jusqu'à la convergence afin d'obtenir $\alpha^{(0)}$.

Étape 1 Calcul de $\sigma_\zeta^{2(1)}$

Partant de $\alpha^{(0)}$ calculé à l'étape 0, considérons le modèle linéarisé

$$t_{y,d}^{*(0)} \equiv t_{y,d} - N_d b_{d0}(\alpha^{(0)}) \approx b_{d1}(\alpha^{(0)})T'_{x,d}\alpha + e_{y,d}^{*(0)} \tag{3.4}$$

où $e_{y,d}^{*(0)} = b_{d1}(\alpha^{(0)})N_d\zeta_d + e_{y,d}$, $\quad vecteur\ e_y^{*(0)} \sim N(0, W(\alpha^{(0)}, \sigma_\zeta^2))$

$W(\alpha^{(0)}, \sigma_\zeta^2) = \Gamma(\alpha^{(0)}, \sigma_\zeta^2) + V, \quad \Gamma(\alpha^{(0)}) = \text{diag}\left\{b_{d1}^2(\alpha^{(0)})N_d^2\right\}\sigma_\zeta^2$.

Maintenant, utilisons le REML [voir, p. ex., Rao (2003, ch. 6)] pour obtenir l'estimation $\sigma_\zeta^{2(1)}$. Pour être certain que $0 < \sigma_\zeta^2 < 1$, σ_ζ^2 peut d'abord être transformé par le logit avant l'estimation du REML.

Étape II Calcul de $\alpha^{(1)}$

Partant de $\sigma_\zeta^{2(1)}$ obtenu à l'étape 1, calculer $\alpha^{(1)}$ par les moindres carrés pondérés (WLS) pour le modèle linéarisé susmentionné (3.4) excepté que $W(\alpha^{(0)}, \sigma_\zeta^2)$ est remplacé par $W(\alpha^{(0)}, \sigma_\zeta^{2(1)}) = \Gamma(\alpha^{(0)}, \sigma_\zeta^{2(1)}) + V$, $\Gamma(\alpha^{(0)}, \sigma_\zeta^{2(1)}) = \sigma_\zeta^{2(1)}\text{diag}\ \left\{N_d^2 b_{d1}^2(\alpha^{(0)})\right\}$.

Les étapes I et II qui précèdent sont itérées jusqu'à convergence afin d'obtenir $\hat{\sigma}_\zeta^2$ et $\hat{\alpha}$.

Étape III Estimation de $\zeta_d, d = 1, ..., D$

Sachant $\hat{\sigma}_\zeta^2$ et $\hat{\alpha}$, et le modèle linéarisé (3.4) avec $W = \Gamma(\hat{\alpha}, \hat{\sigma}_\zeta^2) + V = \hat{\sigma}_\zeta^2 \text{diag}\{N_d^2 b_{d1}^2(\hat{\alpha})\} + V$, le EBLUP de ζ_d si V est diagonale prend la forme

$$N_d b_{d1}(\hat{\alpha})\hat{\zeta}_d = \Gamma_d W_d^{-1}(t_{y,d} - N_d v_{y,d}(\hat{\alpha})) \tag{3.5}$$

où Γ_d et W_d sont évalués à $\alpha = \hat{\alpha}$, $\sigma_\zeta^2 = \hat{\sigma}_\zeta^2$. En général, si V est non diagonale, nous avons

$$\text{diag}\{N_d b_{d1}\}\hat{\zeta} = \Gamma W^{-1}(t_y - \text{diag}\{N_d\}v_y(\hat{\alpha})). \tag{3.6}$$

Or, $\hat{\zeta}_d$ provenant de (3.6) peut également être exprimé par

$$N_d b_{d1}(\hat{\alpha})\hat{\zeta}_d = \Gamma_{dd} W_{dd.\bar{d}}^{-1}\left[(t_{y,d} - N_d v_{y,d}(\hat{\alpha})) - W_{d\bar{d}} W_{\bar{d}\bar{d}}^{-1}(t_{y,\bar{d}} - \mathrm{diag}\{N_{d'}, d' \neq d\} v_{\bar{d}}(\hat{\alpha}))\right]$$

$$= \Gamma_{dd} W_{dd}^{-1}\left[W_{dd} W_{dd.\bar{d}}^{-1}(t_{y,d} - N_d v_{y,d}(\hat{\alpha})) - W_{d\bar{d}} W_{\bar{d}\bar{d}}^{-1}(t_{y,\bar{d}} - \mathrm{diag}\{N_{d'}, d' \neq d\} v_{y,\bar{d}}(\hat{\alpha}))\right] \qquad (3.7)$$

où $\bar{d}$ est l'ensemble de tous les domaines sauf le domaine d, et $W_{dd}, W_{d\bar{d}}, W_{\bar{d}\bar{d}}$ sont définis par le partitionnement de la matrice W

$$W = \begin{pmatrix} W_{dd} & W_{d\bar{d}} \\ W_{\bar{d}d} & W_{\bar{d}\bar{d}} \end{pmatrix}, \quad W_{dd.\bar{d}} = W_{dd} - W_{d\bar{d}} W_{\bar{d}\bar{d}}^{-1} w_{\bar{d}d}. \qquad (3.8)$$

Il s'ensuit que l'estimation sur petits domaines de $T_{y,d}$ est donnée par

$$t_{y,d}^{sae} = N_d[v_d(\hat{\alpha}) + v_d(\hat{\alpha})(1 - v_d(\hat{\alpha})\hat{\zeta}_d] \qquad (3.9)$$

parce que $N_d b_{d0}(\hat{\alpha}) + b_{d1}(\hat{\alpha}) T'_{x,d}\hat{\alpha} = N_d v_d(\hat{\alpha})$.

Donc, pour une matrice V diagonale, $N_d^{-1} t_{y,d}^{\mathrm{sae}}$ est nécessairement compris en 0 et 1 et est une combinaison convexe de $N_d^{-1} t_{y,d}$ et $v_d(\hat{\alpha})$. Toutefois, si V est non diagonale, il ne doit pas nécessairement en être ainsi, mais la valeur se situe vraisemblablement dans l'intervalle (0, 1), parce que le premier terme de (3.7) devrait, en principe, être dominant. Comme nous l'avons mentionné plus haut, nous pouvons tronquer $\hat{\zeta}_d$ à ± 1 s'il est situé à l'intérieur de l'intervalle (-1, 1).

Enfin, l'EQM de t_y^{sae} peut être calculée facilement à l'aide d'une approximation de deuxième ordre comme l'a montré Rao (2003, p. 155) parce qu'après la linéarisation, le modèle GLMARC se réduit à un modèle de type Fay-Herriot. Jusqu'à présent, le paramètre de surdispersion σ_0^2 a été fixé à 1 pour simplifier. Afin d'estimer conjointement ($\sigma_0^2, \sigma_\zeta^2$), nous pouvons commencer par estimer σ_ζ^2 par le REML sur la vraisemblance profil après remplacement de σ_0^2 par son EQM, sachant σ_ζ^2 pour la vraisemblance REML. Il s'avère que $\hat{\sigma}_0^2(\sigma_\zeta^2)$ possède une forme simple de moyenne des carrés des résidus standardisés; voir p. ex. Harvey (1989, ch. 4) dans le contexte de la filtration de Kalman pour l'estimation BLUP.

3.3 Étalonnage des estimations sur petits domaines sous le modèle GLMARC

Dans le cas du LMM, il est possible d'obtenir un étalonnage exact en agrandissant le modèle; voir par exemple Singh (2006). Il serait utile d'examiner cet aspect avant d'envisager l'étalonnage pour le GLMARC. Supposons qu'il n'existe qu'une seule valeur globale repère, à savoir que la somme des estimations sur petits domaines pour tous les totaux de domaine devrait être égale à la somme des estimations directes des totaux pour tous les domaines faisant partie du sous-groupe d'étalonnage, B(disons) qui, ici, est constitué de tous les domaines $d, d = 1,...,D$. Considérons maintenant, pour le LMM, le modèle agrandi

$$t_y = 1_B \beta_B + T(x)\beta + \mathrm{diag}\{N_d\}\eta + e_y = T^*(x)\beta^* + T(c)\eta + e_y \qquad (3.10)$$

où 1_B est un vecteur de dimension D dont les éléments sont 1 ou 0 selon que le d^e domaine est ou non dans B, et $T^*(x)$ est la matrice de covariance agrandie, β^* est le vecteur de coefficients β agrandi correspondant et $T(c)$ est simplement $\mathrm{diag}\{N_d\}$.

Dans le cas d'une valeur repère globale, 1_B est simplement un vecteur de valeurs 1. La covariable $V1_B$ faisant intervenir la matrice de covariance des erreurs d'observation est introduite dans le modèle avec un paramètre de régression supplémentaire β_B à titre d'artefact en vue d'induire l'étalonnage.

Pour le montrer, notons que, pour le LMM

$$t_y^{\mathrm{sae}} = T^*(x)\hat{\beta}^* + \Gamma W^{-1}(t_y - T^*(x)\hat{\beta}^*) \qquad (3.11)$$

où $\Gamma = T(c)T'(c)\sigma_\eta^2 = \mathrm{diag}\{N_d^2\}\sigma_\eta^2$, $W = \Gamma + V$ et $\hat{\beta}^*$ est l'estimation WLS de β sous (3.10).

En réécrivant (3.11) sous la forme

$$t_y^{\text{sae}} = t_y - (I - \Gamma W^{-1})(t_y - T^*(x)\hat{\beta}^*)$$
$$= t_y - VW^{-1}(t_y - T^*(x)\hat{\beta}^*) \qquad (3.12)$$

nous obtenons l'étalonnage souhaité, c.-à-d.

$$1'_B(t_y - t_y^{\text{sce}}) = 1'_B VW^{-1}(t_y - T^*(x)\hat{\beta}^*) = 0 \qquad (3.13)$$

parce que (3.13) est l'une des équations WLS correspondant à la nouvelle covariable $V1_B$.

Introduisons maintenant, pour le GLMARC, un nouveau vecteur de covariables $\text{diag}\{N_d^{-1}b_{d(1)}^{-1}(\tilde{\alpha}^*)\}V1_B$ et le paramètre de régression correspondant α_B, où $\tilde{\alpha}^*$ est la valeur réelle inconnue de α^*, et considérons le modèle agrandi

$$t_y = \text{diag}\{N_d\}\, v(\alpha^*) + \text{diag}\left\{N_d v_d(\alpha^*)(1 - v_d(\alpha^*))\right\}\zeta + e_y \qquad (3.14)$$

où $v(\alpha^*)$ =invlogit$(A^*(x)'\alpha^*)$, et $A^*(x)$ désigne la matrice de covariance agrandie des moyennes de domaine, y compris celles pour le nouveau vecteur de covariables. En estimant les paramètres par la méthode IBLUP, pour le modèle linéarisé, nous utilisons la version légèrement modifiée qui suit

$$t_{y,d} \approx b_{d0}(\alpha^{*(r)})N_d + b_{d1}(\alpha^{*(r)})T'^*_{x,d}\alpha^* + b_{d1}(\alpha^{*(r)})N_d\zeta_d + e_d \qquad (3.15a)$$

où $b_{d0}(\alpha^{*(r)}) = v_d(\alpha^{*(r)}) - b_{d1}(\alpha^{*(r)})A'_{x,d}\alpha^{(r)} - (d^e \text{élément de diag}\{N_d^{-1}\}V1_B)$ (3.15b)

et $b_{d1}(\alpha^{*(r)})T'^*_{x,d}\alpha^* = b_{d1}(\alpha^{*(r)})T'_{x,d}\alpha + (d^e \text{ élément de } V1_B)\alpha_B$ (3.15c)

Notons que, dans les derniers termes de (3.15b) et (3.15c), le produit $b_{d1}(\alpha^{*(r)})b_{d1}^{-1}(\tilde{\alpha}^*)$ est remplacé par 1. Cela permet de contourner la spécification inhabituelle de la nouvelle covariable comportant le paramètre α^* inconnu. Cette approche est raisonnable, parce qu'au point de convergence, le produit sera égal à 1 quand $\tilde{\alpha}^*$ est remplacé par $\hat{\alpha}^*$. Cette légère modification des termes b_{d0} et b_{d1} accélère la convergence de l'IBLUP.

En suivant maintenant l'argument LMM sous étalonnage, nous déduisons de (3.13) et (3.15a) que

$$1_B(t_y^* - t_y^{*\text{sae}}) = 0 \qquad (3.16)$$

où $t_y^* = t_y - \text{diag}\{b_{d0}(\hat{\alpha}^*)N_d\}1$ et $t_y^{*sae} = \text{diag}\{b_{d1}(\hat{\alpha}^*)\}T^*(x)\hat{\alpha}^* + \text{diag}\{b_{d1}(\hat{\alpha}^*)N_d\}\hat{\zeta}$.

Cependant, $t_y^* - t_y^{*\text{sae}} = t_y - [v(\hat{\alpha}^*) + \text{diag}\{v_d(\hat{\alpha}^*)(1 - v_d(\hat{\alpha}^*)N_d\}\hat{\zeta}]$

$$= t_y - t_y^{\text{sae}} \qquad (3.17)$$

qui, compte tenu de (3.16), établit l'étalonnage souhaité sous le modèle GLMARC.

3.4 Estimation et étalonnage en présence de domaines regroupés

Il est fréquent en pratique, et cela est effectivement le cas pour l'exemple fondé sur l'ESCC examiné à la section suivante, de devoir regrouper les domaines avant d'estimer les paramètres fixes α et σ_ζ^2. Supposons, à titre d'illustration, que deux domaines seulement d' et d'' sont regroupés. Alors, le nombre total de domaines D est réduit à $\tilde{D} = D - 1$ et, dans le modèle (3.1), d est remplacé par $\tilde{d}$ variant de $1,...,\tilde{D}$, et V est remplacée par $\tilde{V}$. Estimons maintenant α et σ_ζ^2 comme auparavant. Toutefois, dans l'estimation de $\zeta_{d'}, \zeta_{d''}$, le résidu pour le domaine regroupé $[(t_{y,d'} + t_{y,d''}) - (N_{d'}v_{d'}(\hat{\alpha}) + N_{d''}v_{d''}(\hat{\alpha})]$ est réparti proportionnellement entre d' et d'' en fonction des variances relatives de $b_{d'1}(\hat{\alpha})N_{d'}\zeta_{d'}$ et $b_{d''1}(\hat{\alpha})N_{d''}\zeta_{d''}$, de manière à peu près semblable à (3.5); voir aussi Singh (2006).

La propriété d'étalonnage des estimations sur petits domaines pour les domaines regroupés suit les mêmes grandes lignes. Toutefois, les estimations sur petits domaines dans les domaines d' et d'' à l'intérieur du sous-groupe regroupé ne s'obtiennent pas facilement, parce que le vecteur de covariables supplémentaires fait intervenir la

matrice $\tilde{V}$ de dimensions $\tilde{D} \times \tilde{D}$ et non la matrice V. Il ne semble pas exister de moyen optimal de résoudre ce problème, mais une solution simple pourrait consister à répartir le total ($\tilde{d}^e$ligne de $V1_B$, le symbole $\sim$ dénotant la dimension réduite $\tilde{D}$) dans le cas du LMM entre les domaines individuels d' et d'' proportionnellement à $N_{d'}$ et $N_{d''}$. Essentiellement, cela modifie quelque peu la covariable du terme d'effet aléatoire dans le modèle original avant le regroupement. Le cas du GLMARC peut être traité de la même manière.

4. Application du modèle GLMARC aux données de l'ESCC

La méthode proposée a été appliquée aux données du cycle 1.1 de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) qui a été réalisé en 2000-2001. L'objectif était d'estimer le nombre total de fumeurs quotidiens dans 40 sous-populations ou petits domaines de l'Île-du-Prince-Édouard. Les sous-populations ont été définies en fonction des quatre régions sociosanitaires (Queens, East Prince, West Prince et Kings) et de dix sous-groupes âge-sexe (les tranches d'âge étant 12 à 19 ans, 20 à 29 ans, 30 à 44 ans, 45 à 64 ans et 65 ans et plus). Pour que le nombre de degrés de liberté soit suffisant pour la modélisation, il a été décidé de modéliser les données provenant des 4 provinces atlantiques regroupées, ce qui a donné en tout 23 régions sociosanitaires et 230 petits domaines.

Nous avons choisi comme sous-groupes d'étalonnage les dix sous-groupes marginaux âge-sexe et les quatre sous-groupes provinciaux de domaines. Nous avons obtenu ainsi 13 contraintes d'étalonnage (valeurs repères) non redondantes. Le modèle utilisé pour l'illustration a été choisi comme modèle de moyenne commune de sous-groupes dans lequel il est supposé que le nombre moyen de fumeurs quotidiens est le même pour tous les domaines comportant un sous-groupe âge-sexe identique. Le modèle de moyenne commune de sous-groupe est un modèle simple qui représente souvent un bon point de départ faute d'autres covariables prédictives importantes. Dans l'avenir, il est prévu d'étudier l'utilisation de données sur les admissions à l'hôpital et d'autres données administratives pour mieux modéliser les covariables. En bout de ligne, les modèles considérés contenaient 23 paramètres de régression α^*, 10 pour les moyennes communes de sous-groupes âge-sexe et 13 pour les valeurs repères.

Le regroupement de certains domaines a également été nécessaire pour s'assurer que les données d'entrée des modèles possèdent les propriétés requises. Nous avons appliqué pour cela certaines règles empiriques, telles que l'absence d'estimations directes nulles, une taille effective d'échantillon minimale d'au moins 30 et un produit de la taille effective d'échantillon par la probabilité estimée d'occurrence d'au moins 1. Le regroupement des domaines originaux a abouti de cette façon à un total de 217 domaines regroupés pour l'estimation des paramètres fixes du modèle. Les domaines choisis pour constituer un groupe étaient ceux pour lesquels la distance euclidienne était la plus courte par rapport aux courbes de moyenne des covariables du modèle particulières au domaine. Les domaines formant un groupe devaient être compris dans les sous-groupes d'étalonnage, et cette contrainte a dû être assouplie à l'occasion pour satisfaire les trois règles empiriques susmentionnées. Le fait de maintenir les domaines choisis pour former un groupe dans les sous-groupes d'étalonnage donne l'auto-étalonnage des estimations sur petits domaines finales. Toutefois, une légère violation ne devrait pas avoir d'incidence grave sur les contraintes d'étalonnage.

Les modèles LMM et GLMARC ont été appliqués tous deux aux données de l'ESCC. Nous avons procédé à plusieurs tests diagnostiques analogues aux innovations du filtre de Kalman. Nous avons pour cela classé les domaines par ordre décroissant de taille effective d'échantillon, puis nous avons traité le rang comme une pseudovariable temporelle comme il est décrit dans Singh (2006). Les innovations produites par les modèles LMM ainsi que GLMARC passent avec succès le test de normalité de Shapiro-Wilk, les valeurs p étant de 64,28 % et 78,13 %, respectivement. Les figures 4.1 (a, b) pour le GLMARC et 4.2 (a, b) pour le LMM montrent les diagrammes de dispersion et les diagrammes Q-Q des innovations et ne révèlent aucun signe de régularité particulièrement inhabituelle.

Le tableau 4.1 donne les estimations des modèles et le R-carré pour les deux modèles. Premièrement, l'estimation de la variance de la composante aléatoire est plus faible pour le modèle LMM, comme prévu parce que, dans le GLMARC, elle apparaît comme un facteur multiplicatif d'une fonction que l'on sait être comprise entre 0 et 1. L'estimation inférieure à 1 de la surdispersion pour les deux modèles donne à penser qu'il existe une certaine

sous-dispersion. Enfin, le R-carré, qui est une mesure descriptive de la signification du modèle, prend une valeur élevée (supérieure à 90 %) pour les deux modèles, ne donnant donc aucune raison de s'inquiéter. La comparaison des CV ou des RREQM (tableau 4.1) révèle, quand à elle, certaines différences intéressantes entre les deux modèles. Sur les 230 petits domaines, 76 des estimations directes étaient non publiables parce que le CV était supérieur à 33,3 %. Toutefois, pour les estimations sur petits domaines fondées sur le LMM, le nombre de résultats non publiables n'était plus que de 10 et pour celles fondées sur le GLMARC, il était nul. Les résultats finaux en ce qui a trait aux estimations et à leur précision sont résumés au tableau 4.3. Les colonnes relatives à la précision (Mod-CV, Mod-RREQM et RREQM) ont été calculées en divisant l'erreur type de la racine de l'erreur quadratique moyenne par l'estimation GLMM-ARC pour faciliter la comparaison des méthodes. Les 13 premières lignes donnent les estimations repères et leur précision, tandis que les 11 dernières donnent les estimations et leur précision pour 11 des 40 domaines cibles (classés de la plus petite taille d'échantillon à la plus grande et correspondant aux déciles de la distribution des tailles d'échantillon). La concordance est exacte avec toutes les valeurs repères pour le Nouveau-Brunswick, parce que certains domaines choisis pour former un groupe provenaient de la province de Terre-Neuve-et-Labrador. Il convient aussi de souligner que, même si les valeurs repères sont satisfaites exactement, leur précision peut varier pour les estimations directes, LMM et GLMM-ARC, à cause de l'utilisation d'une EQM ajustée en raison de la surdispersion. Enfin, nous notons que, même si nous n'avons pas observé ici le problème d'estimations négatives ou de composantes de la variance négligeables dans le cas du LMM, les estimations sous le modèle GLMARC ont généralement de bonnes propriétés comparativement à celles obtenues sous le modèle LMM, particulièrement quand la RREQM sous le modèle LMM a tendance à être élevée. De surcroît, étant donné ses propriétés théoriques souhaitables comparativement au LMM, le GLMARC pourrait être préférable en pratique.

5. Conclusion

Nous proposons dans le présent article, pour la modélisation pour petits domaines au niveau agrégé d'après des données d'enquête, un modèle appelé GLMARC qui utilise les méthodes de type LMM sous la forme IBLUP pour l'estimation des paramètres. La méthode proposée offre une alternative simple aux approches bayésiennes hiérarchiques, ainsi qu'un auto-étalonnage et des diagnostics de modélisation faciles à interpréter. La méthode est illustrée à l'aide d'une application aux données de l'ESCC et le concept de regroupement de domaines (Singh, 2006) est également illustré afin de résoudre le problème des domaines pour lesquels les estimations sont nulles ou les tailles d'échantillon, petites. Dans le cas de l'exemple de l'ESCC, nous avons constaté sous un modèle simple de moyenne commune de sous-groupes que, pour les domaines posant ce genre de problème, il est possible d'obtenir des estimations ponctuelles raisonnables, ainsi que les estimations correspondantes de l'EQM en utilisant la méthode proposée qui a tendance à produire des résultats généralement supérieurs à ceux obtenus sous le modèle LMM. Il s'agit là d'une amélioration par rapport à la pratique courante consistant à écarter ce genre de domaines problématiques du processus de modélisation et à ne publier par après que des estimations synthétiques à leur sujet. Dans l'approche proposée, nous montrons comment la méthode d'auto-étalonnage applicable au modèle LMM peut être généralisée au cas du GLMARC, et ce même en cas de regroupement de domaines. Notons que l'intégration de l'auto-étalonnage dans la modélisation n'est pas possible si les domaines posant problème ne font pas partie du processus de modélisation. Enfin, soulignons que même si, en appliquant le modèle simple de la moyenne commune de sous-groupes, la méthode proposée montre que les c.v.. de nombreuses estimations pourraient être améliorés, il serait utile de trouver de meilleures covariables prédictrices afin de réduire davantage la variance de l'erreur de modélisation σ_ζ^2, ce qui accroîtrait encore les gains d'efficacité.

Références

Drum, M.L., et McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics*, 677-689.

Fay, R.E., et Herriot, R.A. (1979). Estimates of Income for small places: an application of James-Stein procedures to Census Data. ASA., 74, 269-277.

Harvey, A.C. (1989). *Forecasting, Structural Time Series, and the Kalman Filter*. Cambridge University Press.

Jiang, J., et Lahiri, P. (2006). Mixed Model Prediction and small area estimation. Test, Vol. 15, 1-96.

McCullagh, P. et Nelder, J.A. (1989). Generalized Linear Models, 2nd ed., London: Chapman and Hall.

Rao, J.N.K. (2003). Small Area Estimation. New York: John Wiley.

Singh, A.C. (2006). Some problems and proposed solutions in developing a small area estimation product for clients. ASA Proc. Surv. Res. Meth. Sec., 3673-3683.

Sutradhar, B.C., et Rao, R.P. (1996). On joint estimation of regression and overdispersion parameters in generalized linear models for longitudinal data. *Multivariate Analysis*, 56, 90-119.

Vonesh, E.G., et Carter, R.L. (1992). Mixed effects nonlinear regression for unbalanced repeated measures. *Biometrics*, 48, 1-17.

You, Y., et Rao, J.N.K. (2002). Small Area Estimation using unmatched sampling and linking models. La revue canadienne de statistique, 30, 3-15.

Figure 4.1 : Diagramme de dispersion et diagramme Q-Q des innovations standardisées sous le modèle GLMARC

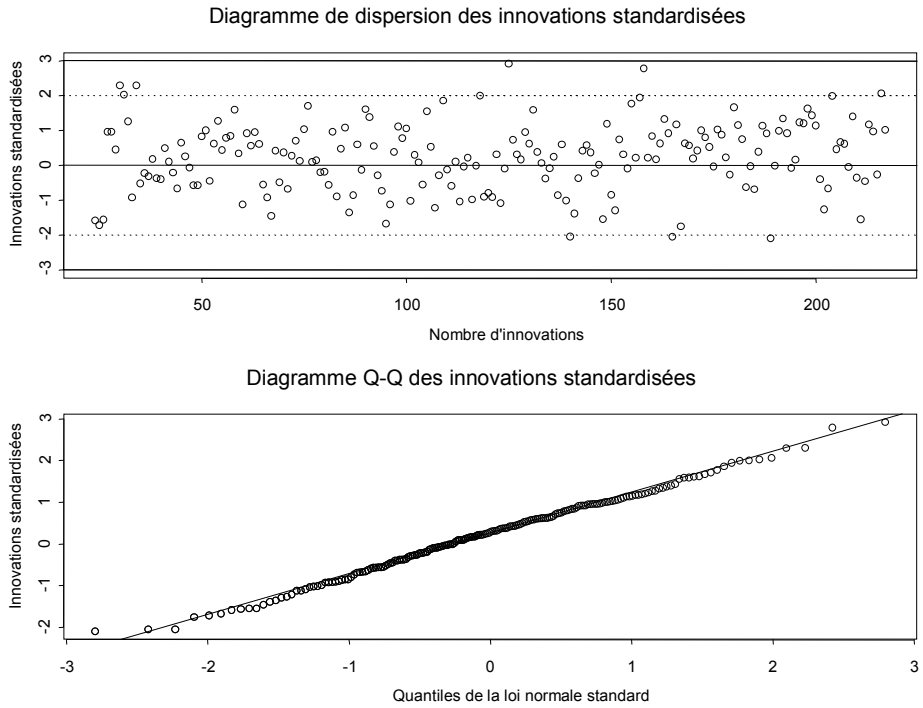

Figure 4.2 : Diagramme de dispersion et diagramme Q-Q des innovations standardisées sous le modèle LMM

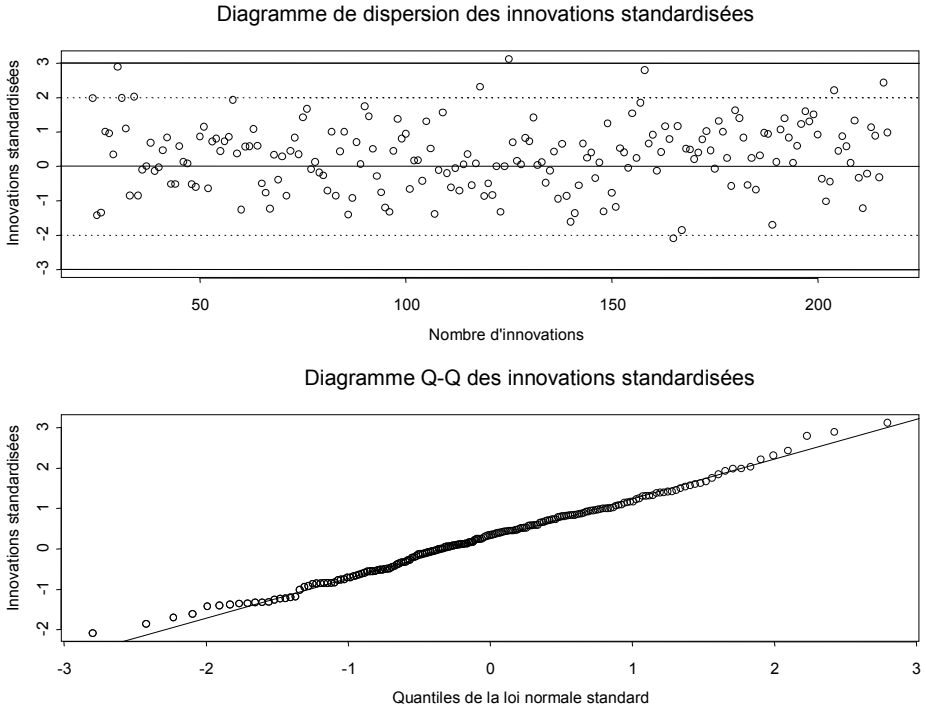

Tableau 4.1 : Estimations des paramètres et R-carré des modèles GLMARC et LMM

Paramètre					LMM	GLMARC
	Repère	*Âge*	*Sexe*	*Géographie*		
$\hat{\beta}$	•	20-29	Masculin		-0,0030	-0,0240
	•	30-44	Masculin		-0,0015	-0,0164
	•	45-64	Masculin		-0,0023	-0,0203
	•	65+	Masculin		-0,0009	-0,0076
	•	12-19	Féminin		0,0001	0,0018
	•	20-29	Féminin		-0,0013	-0,0125
	•	30-44	Féminin		-0,0025	-0,0228
	•	45-64	Féminin		-0,0017	-0,0154
	•	65+	Féminin		0,0001	0,0063
	•			Î.-P.-É.	0,0011	0,0067
	•			Nouvelle-Écosse	-0,0006	-0,0031
	•			Nouveau-Brunswick	-0,0004	-0,0021
	•			Atlantique	0,0029	0,0235
		12-19	Masculin		0,0853	-2,3091
		20-29	Masculin		0,3776	-0,6017
		30-44	Masculin		0,2865	-0,9313
		45-64	Masculin		0,2471	-1,1227
		65+	Masculin		0,1017	-2,1653
		12-19	Féminin		0,0804	-2,3570
		20-29	Féminin		0,2122	-1,3821
		30-44	Féminin		0,2771	-0,9438
		45-64	Féminin		0,1997	-1,4150
		65+	Féminin		0,0709	-2,5236
$\hat{\sigma}_{\eta}^2$					0,00123	0,06053
$\hat{\sigma}_{0}^2$					0,95139	0,91428
R^2					0,9619	0,9865

Tableau 4.2 : Catégories de c.v. des estimations directes et des estimations sur petits domaines sous les modèles GLMARC et LMM

CLASSE	Mod-CV – Directe	Mod-RREQM – LMM	RREQM – GLMARC
[0]	2	0	0
(0-16,5 %)	32	130	119
[16,5 %-33,3 %)	120	90	111
[33,3 %-50 %)	58	10	0
[50 %-100 %)	14	0	0
[100 %-)	4	0	0

Tableau 4.3 : Estimations sur petits domaines et leur précision sous les modèles GLMARC et LMM

Géographie	Âge	Sexe	Taille de l'échantillon	Estimation directe	Estimation LMM	Estimation GLMARC	Mod-c.v. Directe	Mod-RREQM LMM	RREQM GLMARC
---	20-29	MASCULIN	881	52 496,97	52 496,97	52 496,97	5,85	5,71	5,63
---	30-44	MASCULIN	2 153	90 252,64	90 252,64	90 252,64	3,70	3,61	3,56
---	45-64	MASCULIN	2 520	74 592,38	74 592,38	74 592,38	4,10	4,00	3,95
---	65+	MASCULIN	1 247	16 720,60	16 720,60	16 720,60	8,99	8,77	8,65
---	12-19	FÉMININ	1 256	15 952,30	15 952,30	15 952,30	8,94	8,72	8,61
---	20-29	FÉMININ	1 141	39 406,69	39 406,69	39 406,69	5,92	5,77	5,69
---	30-44	FÉMININ	2 605	82 692,72	82 692,72	82 692,72	3,95	3,85	3,80
---	45-64	FÉMININ	2 830	67 787,47	67 787,47	67 787,47	4,52	4,41	4,35
---	65+	FÉMININ	1 996	17 095,44	17 095,44	17 095,44	8,42	8,21	8,10
Î.-P.-É.	---	---	3 651	27 491,75	27 491,75	27 491,75	4,70	4,58	4,52
N.-É.	---	---	5 319	184 684,46	184 684,46	184 684,46	3,33	3,24	3,20
N.-B.	---	---	4 996	147 603,97	147 733,11	147 735,09	3,17	3,09	3,05
ATLANT.	---	---	17 836	474 686,16	474 686,16	474 686,16	1,88	1,83	1,81
1140	12-19	MASCULIN	25	67,41	86,71	87,61	46,21	35,34	24,51
1140	20-29	FÉMININ	30	100,36	211,42	205,95	24,57	17,58	19,63
1110	20-29	MASCULIN	51	625,73	556,08	573,79	22,44	10,74	13,52
1120	12-19	MASCULIN	59	159,32	172,29	178,17	42,19	28,75	22,11
1110	20-29	FÉMININ	66	388,55	303,10	295,83	28,91	16,46	18,04
1140	65+	FÉMININ	79	123,87	102,39	88,01	38,85	28,15	19,40
1120	30-44	MASCULIN	92	1 288,21	1 213,87	1 230,43	20,24	10,11	12,76
1120	65+	FÉMININ	115	230,57	222,06	208,74	36,78	27,08	20,49
1110	65+	FÉMININ	123	67,61	80,80	95,85	28,78	25,28	21,32
1110	45-64	FÉMININ	159	518,33	525,00	516,05	19,55	13,35	13,85
1130	45-64	FÉMININ	218	1 592,19	1 764,76	1 734,63	15,48	11,83	12,39