Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Contributions de J. N. K. Rao aux modèles multiniveaux d'enquêtes complexes et vraisemblance composite

par Mary E. Thompson

Date de diffusion : le 8 septembre 2025

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2024 de Statistique Canada
Le futur des statistiques officielles

Contributions de J. N. K. Rao aux modèles multiniveaux d'enquêtes complexes et vraisemblance composite

Mary E. Thompson[1]

Résumé

Dans le cadre des modèles multiniveaux à estimer à l'aide de données provenant d'enquêtes comportant des plans de sondage complexes, le présent article souligne certaines contributions de l'article de référence de Rao, Verret et Hidiroglou (Techniques d'enquête, 2013) et des travaux connexes ultérieurs.

Mots-clés : Plan de sondage complexe; échantillonnage à plusieurs degrés; utilisation analytique des données d'enquête; composantes de variance; estimation du système de fonction.

1. Introduction : Modèles multiniveaux et composantes de variance

L'une des nombreuses orientations de la recherche du professeur J. N. K. Rao concerne ses travaux sur les modèles multiniveaux et la vraisemblance composite. Le présent article abordera l'article intitulé « Techniques d'enquête » qu'il a corédigé en 2013 avec Verret et Hidiroglou ainsi que ses extensions.

Qu'est-ce qu'un modèle multiniveau? Un exemple classique (voir par exemple Singer, 1998) est le modèle *MATHACH [math achievement ou résultats en mathématique]* dans lequel les unités élémentaires de la population sont les élèves au secondaire d'une certaine classe au sein d'une administration donnée et la variable de réponse Y représente une transformation appropriée de leurs résultats à un test de mathématiques normalisé. Dans (1) ci-dessous, i indexe leur école et j indexe les élèves d'une école. La variable indépendante $\mathbf{x}_{ij}$ pourrait être quelque chose comme le vecteur de lignes $(1, SES_{ij})$, où SES est une variable de statut socio-économique. La spécification du modèle peut être exprimée comme suit :

$$Y_{ij} \mid \mathbf{x}_{ij}, \mathbf{V}_i = \mathbf{v}_i : \underset{ind}{} N(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{x}_{ij}^T\mathbf{v}_i, \sigma_e^2), \tag{1}$$

$$\mathbf{V}_i : \underset{iid}{} N_p(0, \boldsymbol{\Sigma}_v). \tag{2}$$

L'école est le niveau supérieur, parfois appelé unité de « niveau 2 », et l'étudiant est le niveau inférieur, parfois appelé unité de « niveau 1 ». Le vecteur de coefficient $\boldsymbol{\beta}$ comprendrait une ordonnée à l'origine et constituerait ce que l'on appelle un *effet fixe*, alors que $\mathbf{V}_i$ désigne un *effet aléatoire* au niveau de l'école de la même dimension ou d'une dimension inférieure qui comporte une distribution multivariée à distribution normale.

Le ministère de l'Éducation souhaiterait connaître les paramètres $\boldsymbol{\beta}$ et savoir si le système se porte bien en général. S'il souhaitait obtenir des moyennes pour certaines écoles, il pourrait tirer des estimations des paramètres $\boldsymbol{\beta}$ et des effets aléatoires $\mathbf{v}_i$, ce qui lui permettrait d'obtenir des estimations moyennes sur petits domaines.

[1]Mary E. Thompson, Département de statistique et de science actuarielle, Université de Waterloo, Waterloo (Ontario), N2L 3G1

Cependant, la modélisation multiniveau vise principalement à évaluer la variation, par exemple pour estimer la part de la variation des notes qui est due à la variation entre les écoles. Pour simplifier, si nous supposons que $\mathbf{\Sigma}_v$ est diagonale, alors les objets pertinents de l'inférence sont les variances des composantes $\mathbf{V}$, c.-à-d. la « variation au niveau de l'école », et la mesure dans laquelle ils contribuent à la variation globale dans Y – en particulier s'ils étaient très faibles par rapport à la variance au niveau de l'étudiant σ_e^2.
Le problème présente certaines difficultés : notamment, si l'ampleur des variances sur la diagonale de $\mathbf{\Sigma}_v$ *est* très faible, il peut être difficile d'obtenir une bonne estimation.

L'exemple des résultats en mathématique est formulé en termes de modèle de population hypothétique, sans procéder à un échantillonnage des unités. On retrouve toutefois des cas correspondants dans les populations étudiées. Prenons l'exemple d'une enquête nationale sur la santé comportant un plan d'échantillonnage aréolaire et dans laquelle Y est une mesure appropriée de l'indice de masse corporelle (IMC) d'une certaine sous-population, et la variabilité de Y au niveau du domaine par rapport à la variation au niveau de l'individu constitue une donnée d'intérêt.

Le modèle multiniveau classique est gaussien, mais que se passe-t-il si la variable Y n'est pas continue mais binaire, ou continue mais non transformable en un modèle gaussien? Considérons un échantillon de votes exprimés dans un échantillon probabiliste de bureaux de vote dans un État lors d'un jour d'élection, le principal problème étant de prévoir le résultat final de l'élection dans cet État. On pourrait également considérer la question d'évaluer si la variation des résultats des bureaux de vote reflète des effets locaux sous-jacents. Désignons Y_{ij} comme l'indicateur de vote du candidat A par la personne qui vote *j* au bureau de vote *i*. Un modèle de la variable binaire Y_{ij} pourrait être

$$Prob(Y_{ij} = 1 \mid \mathbf{x}_{ij}, \mathbf{V}_i = \mathbf{v}_i) = expit(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{x}_{ij}^T\mathbf{v}_i), \mathbf{V}_i :_{iid} N_p(0, \mathbf{\Sigma}_v).$$

La modélisation multiniveau a gagné en popularité dans les domaines de l'épidémiologie et des sciences sociales avec la démocratisation du calcul statistique à partir du milieu des années 1990. Cette période a également été marquée par la mise au point d'un nombre croissant de logiciels d'enquête complexes et d'approches d'analyse à plus grande capacité de calcul. La solution que l'on utilisait habituellement pour mener une estimation analytique de points à partir d'un échantillon d'enquête consistait à considérer une équation (ou un système) d'estimation basée sur le recensement pour un estimateur de superpopulation. On suppose en pareil cas que la population d'enquête est générée par échantillonnage aléatoire simple à partir de la superpopulation; puis on estime cette équation (ou ce système) d'estimation à partir de l'échantillon en utilisant des poids d'enquête d'inflation et on résout l'équation d'estimation de l'échantillon qui en résulte. L'objectif est d'obtenir un estimateur conforme au plan de sondage de l'estimateur au niveau de la population d'un paramètre du modèle. La *méthode de pseudo-vraisemblance*, proposée par Binder (1983) et officialisée par Skinner (1989), constituait une extension directe. Cette approche s'est avérée problématique pour l'estimation de la composante de variance dans un problème à deux niveaux. En effet, même s'il semble aller de soi de penser que le nombre d'unités de niveau supérieur dans l'échantillon devient important, il convient, dans de nombreuses applications, d'estimer que le nombre d'unités de niveau inférieur dans chaque unité de niveau supérieur est fixe. Dans ce cadre asymptotique, la méthode de base produit un biais qui ne disparaît pas à mesure que le nombre d'unités de niveau supérieur augmente. Un problème identifié très tôt (voir par exemple Pfeffermann et coll., 1998) consistait donc à formuler un cadre approprié pour assurer la cohérence du plan des estimateurs des composantes de la variance basés sur les enquêtes. Les premières techniques proposées représentaient des moyens de rééchelonner les poids d'enquête, en particulier au niveau le plus élevé.

En septembre 2004, une réunion d'ouverture d'un programme sur les statistiques et les sciences sociales a eu lieu à l'Institut des sciences mathématiques statistiques et appliquées (SAMSI). Les méthodes d'enquête étaient l'un des sous-thèmes abordés, et plusieurs développeurs de logiciels statistiques pour les sciences sociales étaient représentés, présentant des adaptations à des données d'enquête complexes pour Mplus, Stata et d'autres plates-formes. La question qui a reçu le plus d'attention était de savoir comment traiter la modélisation multiniveau, ce qui s'est avéré être une préoccupation majeure pour le groupe d'échantillonnage d'enquête au cours de l'année de projet 2004-2005 du SAMSI.

Stata GLLAMM - pour *generalized linear latent and mixed models* ou modèles linéaires généralisés à effets latents et mixtes - a été développé par M^{me} Sophia Rabe-Hesketh. M. Anders Skrondal et elle ont abordé le problème de l'estimation des paramètres de régression logistique à plusieurs niveaux, notamment en utilisant une méthode de

pseudo-vraisemblance lorsque la variable de réponse était binaire. Dans leur article intitulé *JRSS-B* de 2006, ils ont évalué diverses méthodes de remise à l'échelle des poids. Ils ont finalement conclu qu'avec de petites tailles d'unités de niveau supérieur, même avec une remise à l'échelle des poids, la méthode du pseudo-maximum de vraisemblance doit être utilisée avec prudence.

2. L'article rédigé par Rao, Verret et Hidiroglou (2013)

En 2010, lors de la réunion de la Société statistique du Canada à l'Université Laval, le professeur Rao a dévoilé une nouvelle approche, basée sur l'estimation des fonctions d'estimation, qui a finalement donné lieu à l'article de Rao-Verret-Hidiroglou (2013), ou RVH. Cette méthode et ce cadre théorique allaient considérablement élargir l'ensemble des scénarios dans lesquels une estimation cohérente des composantes de la variance était possible. Ce fut un moment vraiment passionnant.

2.1 Modèle et formule

Voici le modèle général de superpopulation à deux niveaux, au niveau de la population finie, tiré de cet article; il existe des densités générales pour Y étant donné $\mathbf{x}$ et $\mathbf{V}$, et pour $\mathbf{V}$ en soi :

$$Y_{ij} \mid \mathbf{x}_{ij}, \mathbf{V}_i = \mathbf{v}_i : \underset{ind}{} f(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1),\ j = 1, \ldots, M_i$$

et

$$\mathbf{V}_i : \underset{ind}{} f(\mathbf{v}_i \mid \boldsymbol{\theta}_2), i = 1, \ldots N.$$

Les méthodes standard des modèles multiniveaux qui ignorent le plan d'échantillonnage et supposent que ce modèle est valable pour l'échantillon peuvent engendrer des estimateurs asymptotiquement biaisés des paramètres de modèle $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$.

Supposons qu'un échantillon à deux degrés soit prélevé avec un échantillonnage à probabilités inégales, le premier degré se situant au niveau supérieur et le second au niveau inférieur.

Les cas particuliers d'un modèle multiniveau avec les mêmes unités sont le modèle simple à moyenne d'erreurs emboîtées (gaussien) dans lequel les V_i sont des valeurs réelles, et les paramètres sont μ, σ_e^2 and σ_v^2 :

$$Y_{ij} \mid v_i : \underset{ind}{} N(\mu + v_i, \sigma_e^2), V_i : \underset{iid}{} N(0, \sigma_v^2)$$
$$\boldsymbol{\theta} = (\sigma_e^2, \sigma_v^2);$$

et le modèle linéaire à deux niveaux, où l'ordonnée à l'origine et les autres coefficients de régression ont une composante aléatoire avec une matrice de variance-covariance $\boldsymbol{\Sigma}_v$, σ_e^2 est la variance résiduelle, et T dénote la transposition :

$$Y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i : \underset{ind}{} N(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{v}_i, \sigma_e^2), \mathbf{V}_i : \underset{iid}{} N_p(0, \boldsymbol{\Sigma}_v)$$
$$\boldsymbol{\theta} = (\sigma_e^2, \boldsymbol{\Sigma}_v);$$

et la régression logistique à deux niveaux, où le coefficient de régression a également une composante aléatoire avec une matrice de variance-covariance $\boldsymbol{\Sigma}_v$:

$$Prob(Y_{ij}=1|\mathbf{x}_{ij},\mathbf{v}_i): {}_{ind}\ expit(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\mathbf{x}_{ij}^T\mathbf{v}_i), \mathbf{V}_i: {}_{iid}\ N_p(0,\boldsymbol{\Sigma}_v)$$
$$\boldsymbol{\theta}=\boldsymbol{\Sigma}_v.$$

Supposons la classification d'échantillonnage suivante : au premier degré, les unités de niveau supérieur n sont sélectionnées avec des probabilités d'inclusion π_i, et au deuxième degré, les unités de niveau inférieur m_i sont sélectionnées avec des probabilités d'inclusion $\pi_{j|i}$. Supposons que le plan de sondage soit informatif, en ce sens que les probabilités d'échantillonnage ou d'inclusion dépendent de la distribution de Y compte tenu de **x.** Pouvons-nous formuler des équations d'estimation du maximum de vraisemblance au niveau de la population et remplacer les sommes de la population aux deux niveaux par des sommes d'échantillons, pondérées par des poids d'inflation, et obtenir des estimateurs cohérents?

Comme à la section 1, désignons Y_{ij} comme la variable de réponse de l'unité de sondage du second degré *j* dans l'unité de premier degré *i* pour $i=1,...,n$, et $j=1,...,m$. Nous utilisons les lettres minuscules y_{ij} pour représenter les valeurs réalisées de Y_{ij}. Désignons $\mathbf{y}(n)=\{\mathbf{y}_1,\ldots,\mathbf{y}_n\}$ comme les données d'échantillon avec $\mathbf{y}_i=(y_{i1},\ldots,y_{im})^T$ pour $i=1,\ldots,n$.

Comme nous l'avons déjà indiqué, les logiciels statistiques avaient jusqu'alors rééchelonné les poids d'inflation habituels, en particulier au second degré, afin que le logiciel ne soit pas induit en erreur en pensant que les tailles des échantillons du second degré étaient beaucoup plus importantes qu'elles ne l'étaient en réalité. Voir l'étude de Pfeffermann et coll. (1998); Asparouhov (2006); et Grilli et Pratesi (2004) et Rabe-Hesketh et Skrondal (2006) dans le contexte des modèles binaires.

L'idée du professeur Rao était de commencer par des fonctions d'estimation linéaires pour les deux moyennes *et* les paramètres des composantes de variance. Le système d'estimation ponctuelle du modèle simple de moyenne d'erreur emboîtée est donné ci-dessous, où les termes sont les suivants : u_1, u_2 and u_3. Les équations de la fonction d'estimation avec les poids de probabilité d'inclusion inverse et d'inclusion conjointe suivent. Dans des cadres asymptotiques où les tailles d'échantillon au niveau inférieur restent fixes, alors que le nombre d'unités au niveau supérieur tend vers l'infini, les estimateurs sont cohérents avec le modèle de conception pour θ et conformes à la conception pour son estimation basée sur le recensement θ_N.

Examinons l'estimation ponctuelle pour le modèle simple de moyenne d'erreur emboîtée. Les fonctions d'estimation élémentaires sont les suivantes :

- $u_1(y_{ij},\boldsymbol{\theta})=y_{ij}-\mu$
- $u_2(y_{ij},\boldsymbol{\theta})=(y_{ij}-\mu)^2-(\sigma_v^2+\sigma_e^2)$
- $u_3(y_{ij},y_{ik},\boldsymbol{\theta})=[(y_{ij}-\mu)-(y_{ik}-\mu)]^2-2\sigma_e^2,\ j\neq k$.

Les équations d'estimation pondérées (EEP) en vertu du plan des paramètres du modèle sont données par :

$$\sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}u_1(y_{ij},\boldsymbol{\theta})=0$$
$$\sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}u_2(y_{ij},\boldsymbol{\theta})=0$$
$$\sum_{i\in s}w_i\sum_{j<k\in s(i)}w_{jk|i}u_3(y_{ij},y_{ik},\boldsymbol{\theta})=0,$$

où la dernière somme est une somme sur les paires (j,k) d'unités de deuxième phase dans la grappe *i*. Les estimateurs de ces équations d'estimation sont convergents par rapport au modèle pour $\boldsymbol{\theta}$ et au plan pour $\boldsymbol{\theta}_N$.

3. La formulation fondée sur la vraisemblance composite

On s'est rapidement rendu compte que ces fonctions d'estimation au niveau de la population, semblables à la méthode du moment, étaient des exemples d'estimation de la vraisemblance composite, une approche applicable aux modèles linéaires et linéaires généralisés. Les fonctions de vraisemblance composite par paire peuvent prendre plusieurs formes différentes, dont l'une est présentée ci-dessous, où la somme intérieure est une double somme sur des paires d'unités au sein d'une unité de niveau supérieur, et la somme extérieure est prise *sur* les unités de niveau supérieur. La fonction f est la densité marginale conjointe de y_{ij} et y_{ik} dans la même unité de niveau supérieur *i* du modèle à deux niveaux :

$$l_C(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j<k=1}^{M_i} log\, f(y_{ij}, y_{ik} \mid \boldsymbol{\theta})$$

où

$$f(y_{ij}, y_{ik} \mid \boldsymbol{\theta}) = \int f_{y|v}(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i)\, f_{y|v}(y_{ik} \mid \mathbf{x}_{ik}, \mathbf{v}_i)\, f_v(\mathbf{v}_i) d\mathbf{v}_i$$

est la densité marginale conjointe de y_{ij} and y_{ik} .

La version échantillon (pondérée par le plan) du logarithme de la vraisemblance par paire est donnée ici :

$$l_{wc}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \sum_{j<k \in s(i)} w_{jk|i}\, log\, f(y_{ij}, y_{ik} \mid \boldsymbol{\theta}).$$

Pour effectuer une estimation ponctuelle, nous devons résoudre les équations du score composite pondéré

$$\hat{U}_{wc}(\boldsymbol{\theta}) = \partial l_{wc}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \mathbf{0}$$

pour obtenir un estimateur composite pondéré de la vraisemblance $\hat{\boldsymbol{\theta}}_{wc}$.

L'un des avantages de cette formulation est qu'elle permet d'étendre la technique au-delà du modèle linéaire à deux niveaux au modèle linéaire généralisé à deux niveaux. Le professeur Rao a travaillé avec ses collaborateurs Grace Yi et Haocheng Li sur un cadre général, ce qui a donné lieu à un article publié dans la revue *Statistica Sinica* en 2016. Une vraisemblance composite marginale par paire a été formée en multipliant les termes de densité marginale pour les individus de niveau inférieur et les termes de densité marginale conjointe pour les paires, en ignorant la dépendance qui ressort des unités de niveau supérieur *i*. Désignons

$$L_{ij} = \int f_{y|v}(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i)\, f_v(\mathbf{v}_i) d\mathbf{v}_i,$$
$$L_{ijk} = \int f_{y|v}(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i)\, f_{y|v}(y_{ik} \mid \mathbf{x}_{ik}, \mathbf{v}_i)\, f_v(\mathbf{v}_i) d\mathbf{v}_i$$

pour $j \neq k$. La probabilité composite du recensement correspond alors à la formule :

$$C(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j<k} L_{ijk}^{B_{jk}} L_{ij}^{B_j} L_{ik}^{B_k}.$$

Les B_j et B_{jk} peuvent être choisis de différentes manières.

Le théorème de cohérence dans l'article de Yi, Rao et Li (2016) règlent B_j et B_k à 0, et n'utilise que les termes par paire pour les paires d'unités au sein de la même unité de niveau supérieur. La fonction l_{wc} ci-dessous est le logarithme de la vraisemblance composite de l'échantillon :

$$l_{wc}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \sum_{j<k \in s(i)} w_{jk|i} B_{jk} l_{ijk}(\boldsymbol{\theta}),$$

où $w_{jk|i} = \pi^{-1}_{jk|i}$ et $l_{ijk}(\boldsymbol{\theta}) = log(L_{ijk}(\boldsymbol{\theta}))$. Le vecteur de scores de l'échantillon est U_{wc}, le vecteur des dérivées du logarithme de la vraisemblance composite par rapport à chacun des paramètres, et U_{iwc} est le vecteur de la dérivée à l'intérieur de la i e unité de niveau supérieur :

$$U_{wc}(\boldsymbol{\theta}) = \frac{\partial l_w c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i \in s} U_{iwc}(\boldsymbol{\theta}).$$

Le résultat de cohérence est que, sous des conditions de régularité, la solution de $\hat{\boldsymbol{\theta}}_w$ of $\mathbf{U}_{wc}(\boldsymbol{\theta}) = \mathbf{0}$ tend en probabilité vers $\boldsymbol{\theta}$ quand $n \to \infty$, où

$$\mathbf{U}_{wc}(\boldsymbol{\theta}) = \frac{\partial l_w c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i \in s} U_{iwc}(\boldsymbol{\theta}) = \mathbf{0}$$

et

$$\mathbf{U}_{iwc}(\boldsymbol{\theta}) = \sum_{j<k \in s(i)} w_{jk|i} B_{jk} \partial l_{ijk} / \partial \boldsymbol{\theta}.$$

L'article contient également des études de simulation. L'une d'entre elles porte sur un modèle mixte de régression logistique dans lequel Y_{ij} est une réponse binaire avec une moyenne conditionnelle $\mu_{ij} = E(Y_{ij} \mid \mathbf{x}_i, v_i)$ et où en particulier $logit(\mu_{ij}) = \beta_0 + \beta_1 X_{ij} + v_i$; le v_i est i.i.d. $N(0, \sigma^2)$; et les poids d'enquête sont informatifs. On constate que lorsque les tailles d'échantillon au sein des grappes sont faibles, le fait de maximiser la vraisemblance pondérée par paire permet de bien estimer les paramètres β et le paramètre de variance σ, ce qui est mieux que la méthode de maximisation du logarithme du rapport de vraisemblance non pondéré ou du logarithme de la pseudo-vraisemblance.

Dumitrescu, Qian et Rao (2021) se sont efforcés d'étendre la théorie à ce même contexte, où, comme ils l'ont affirmé, « la structure du plan (échantillonnage à deux degrés) correspond à la hiérarchie (niveaux) ». Ils ont étendu la théorie au-delà de la cohérence des estimations ponctuelles aux propriétés des tests d'hypothèses sur les paramètres du modèle multiniveaux, en prouvant un résultat de normalité asymptotique pour les estimateurs composites pondérés de la vraisemblance et en établissant des distributions asymptotiques pour les statistiques du test du score et du test du rapport de vraisemblance. Les résultats ont été étayés par des études de simulation.

Thompson et coll. (2022) (sans justification théorique rigoureuse) ont appliqué la technique de RVH à une étude de la variable continue mesurant le syndrome de stress post-traumatique dans le cadre d'une enquête longitudinale menée dans la région de Galveston Bay, au Texas. Dans ce cas, il y avait une structure de modèle à trois niveaux, où le niveau le plus bas était les vagues d'enquête au sein des individus, le niveau intermédiaire était les individus, et le niveau le plus élevé était les unités primaires d'échantillonnage (UPE) ou « grappes ». On s'attendait à ce qu'il y ait un effet aléatoire au niveau des grappes comportant une composante indépendante et une composante à corrélation spatiale. Les paramètres de la fonction moyenne et de la composante de variance étaient d'intérêt. La technique de RVH a été adaptée à l'estimation conjointe des paramètres de la fonction de la moyenne et des composantes de la variance. Le plan d'échantillonnage a été intégré par la pondération des fonctions d'estimation, et il s'est avéré dans ce cas tout à fait pratique de reconstruire les poids aux niveaux supérieur et moyen (ils étaient égaux à 1 au niveau le plus bas). Les résultats ont confirmé l'existence d'un effet de grappe non négligeable avec autocorrélation spatiale.

On a inclus les renseignements liés au plan de sondage parmi les covariables de régression afin de rapprocher le modèle de ce qui serait supposé dans les analyses basées sur un modèle. Ils ont de plus été intégrés en pondérant les fonctions d'estimation. Une analyse bayésienne utilisant WinBugs et des distributions antérieures non informatives donne des résultats similaires pour les coefficients de régression et pour l'estimation de la composante de variance.

4. Conclusion

La méthode de RVH est une méthode fréquentiste, qui présente de bonnes propriétés en termes de plan et en convergence par rapport au modèle et au plan. On peut s'attendre à ce qu'elle fonctionne bien dans des situations où

les ensembles de données d'enquête sont plus importants et plus structurés que celui examiné par Thompson et coll. (2022). Lumley et Huang (2023 et 2024) ont récemment publié deux articles dans lesquels ils ont mis en œuvre la méthode en utilisant R et étudié ses propriétés. L'une des conclusions du premier article est que l'estimateur de la vraisemblance par paire peut être « étendu à des contextes où les groupes du modèle et les grappes du plan de sondage sont indépendants ».

La deuxième suit cette suggestion, avec une mise en œuvre où, dans le modèle linéaire à deux niveaux, la hiérarchie ne doit pas nécessairement correspondre au plan de sondage. Il existe une application au cas où les UPE sont géographiques et les groupements de niveau supérieur sont des relations génétiques.

Les résultats sont mitigés. Toutefois, parmi les conclusions de ce deuxième article figure la confirmation que « nous recommandons l'estimation pondérée de la vraisemblance par paire lorsque l'échantillonnage est censé être informatif et que les composantes de la variance présentent un intérêt substantiel, et que les variables du plan de sondage ne sont pas disponibles ou ne sont pas appropriées pour être incluses dans le modèle ».

Les travaux de RVH et l'approche de la vraisemblance composite, ainsi que leurs extensions et ramifications, constituent une contribution très importante non seulement pour les méthodes, mais aussi pour les fondements : une méthode fréquentiste d'inférence analytique complexe et des principes fondamentaux appliqués à un problème difficile d'analyse de données, où les conclusions sont importantes et souvent insaisissables.

Remerciements

Cette recherche est en partie soutenue par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG).

Bibliographie

Asparouhov, T. (2006), "Generalized multi-level modeling with sampling weights", *Communications in Statistics - Theory and Methods,* 35, p. 439-460.

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

Dumitrescu, L., Qian, W., et Rao, J.N.K. (2021), "A weighted composite likelihood approach to inference from clustered survey data under a two-level model", *Sankhya,* 83, p. 814-843.

Grilli, L. et Pratesi, M. (2004), "Estimation pondérée dans le cadre de modèles multiniveaux ordinaux et binaires sous un plan d'échantillonnage informatif", *Techniques d'enquête,* 30(1), p. 103-114.

Lumley, T. et Huang, X. (2023), "Weighted composite likelihood for linear mixed models in complex samples", arXiv:2311.13048.

Lumley, T. et Huang, X. (2024), "Linear mixed models for complex survey data: Implementing and evaluating pairwise likelihood", *Stat* 13(1). DOI:10.1002/sta4.657.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1998), "Weighting for unequal selection probabilities in multi-level models", *Journal of the Royal Statistical Society, Series B,* 60, p. 23-56.

Rabe-Hesketh, S. et Skrondal, A. (2006), "Multilevel modeling of complex survey data", *Journal of the Royal Statistical Society, Series A,* 169, p. 805-827.

Rao, J.N.K., Verret, F. et Hidiroglou, M.A. (2013), “Une approche d’inférence fondée sur la vraisemblance composite pondérée pour des modèles à deux niveaux issus de données d’enquête », *Techniques d’enquête,* 39(2), p. 291-311.

Singer, J.D. (1998), “Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models”, *Journal of Educational and Behavioral Statistics*, 24(4), p. 323-355.

Skinner, C.J. (1989), “Domain means, regression and multivariate analysis”, in *Analysis of Complex Surveys*, Eds. Skinner, C.J., Holt, D. and Smith, T.M.F. Chichester: John Wiley & Sons, Inc., p. 59-87.

Thompson, M.E., Meng, G., Sedransk, J., Chen, Q. et Anthopolos, R. (2022), “Spatial multilevel modelling in the Galveston Bay Recovery Study Survey”, in *Advances and Innovations in Statistics and Data Science*, Eds. He, W., Wang, L., Chen, J. and Lin, C.D. Springer Cham, p. 275-293.

Yi, G. Y., Rao, J. N. K., et Li, H. (2016), “A weighted composite likelihood approach for analysis of survey data under two-level models” *Statistica Sinica*, 26, p. 569-587.