Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Un modèle hiérarchique bayésien de non-réponse pour les données catégoriques d'un tableau à double entrée provenant de petits domaines avec incertitude au sujet de l'ignorabilité

par Balgobin Nandram et Myron Katzoff

Juin 2012

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 89-102
Statistique Canada, N° 12-001-X au catalogue

Un modèle hiérarchique bayésien de non-réponse pour les données catégoriques d'un tableau à double entrée provenant de petits domaines avec incertitude au sujet de l'ignorabilité

Balgobin Nandram et Myron Katzoff [1]

Résumé

Nous étudions le problème de la non-réponse non ignorable dans un tableau de contingence bidimensionnel qui peut être créé individuellement pour plusieurs petits domaines en présence de non-réponse partielle ainsi que totale. En général, le fait de prendre en considération les deux types de non-réponse dans les données sur les petits domaines accroît considérablement la complexité de l'estimation des paramètres du modèle. Dans le présent article, nous conceptualisons le tableau complet des données pour chaque domaine comme étant constitué d'un tableau contenant les données complètes et de trois tableaux supplémentaires pour les données de ligne manquantes, les données de colonne manquantes et les données de ligne et de colonne manquantes, respectivement. Dans des conditions de non-réponse non ignorable, les probabilités totales de cellule peuvent varier en fonction du domaine, de la cellule et de ces trois types de « données manquantes ». Les probabilités de cellule sous-jacentes (c'est-à-dire celles qui s'appliqueraient s'il était toujours possible d'obtenir une classification complète) sont produites pour chaque domaine à partir d'une loi commune et leur similarité entre les domaines est quantifiée paramétriquement. Notre approche est une extension de l'approche de sélection sous non-réponse non ignorable étudiée par Nandram et Choi (2002a, b) pour les données binaires ; cette extension crée une complexité supplémentaire qui découle de la nature multivariée des données et de la structure des petits domaines. Comme dans les travaux antérieurs, nous utilisons un modèle d'extension centré sur un modèle de non-réponse ignorable de sorte que la probabilité totale de cellule dépend de la catégorie qui représente la réponse. Notre étude s'appuie sur des modèles hiérarchiques bayésiens et des méthodes Monte Carlo par chaîne de Markov pour l'inférence a posteriori. Nous nous servons de données provenant de la troisième édition de la National Health and Nutrition Examination Survey pour illustrer les modèles et les méthodes.

Mots-clés : Échantillonneur de Metropolis-Hastings ; algorithme SIR ; modèle de non-réponse non ignorable ; modèle d'extension.

1. Introduction

Généralement, les données des enquêtes par sondage sont résumées dans des tableaux de contingence à double entrée. Nous considérons le problème de la non-réponse non ignorable pour un grand nombre de tableaux de contingence de dimensions $r \times c$, pour chacun des domaines spécifiques. Dans nombre de ces enquêtes, des données manquent, si bien que la classification des individus échantillonnés n'est que partielle. Chaque tableau à double entrée présente donc à la fois des cas de non-réponse partielle (données manquantes pour l'une des deux catégories) et des cas de non-réponse totale (données manquantes pour les deux catégories). Comme on ne sait pas nécessairement de quelle façon les données manquent, il peut être souhaitable de privilégier un modèle dans lequel existe une certaine différence entre les données observées et les données manquantes (c'est-à-dire que les données manquantes ne sont pas ignorables). Pour un tableau de contingence $r \times c$ général, nous abordons la question de l'estimation des probabilités de cellule des tableaux à double entrée lorsque la non-réponse est peut-être non ignorable, mais que l'on ne dispose vraiment d'aucune information au sujet de l'ignorabilité. Dans de telles conditions, nous aimerions exprimer le degré d'incertitude au sujet de l'ignorabilité. Nandram et Choi (2002a, b) ont décrit un modèle d'extension approprié pour les données binaires lorsqu'il existe des données provenant de nombreux petits domaines. Nous étendrons ces travaux aux tableaux de contingence $r \times c$.

En désignant par x les covariables et par y la variable réponse, Little et Rubin (2002) décrivent trois types de mécanismes de création de données manquantes. Ils diffèrent selon que la probabilité de réponse a) est indépendante de x et de y ; b) dépend de x, mais non de y, ou c) dépend de y et éventuellement de x. Les données manquent entièrement au hasard (MCAR, *missing completely at random*) sous (a), manquent au hasard (MAR, *missing at random*) sous (b) et ne manquent pas au hasard (MNAR, *missing not at random*) sous (c). Les modèles pour les mécanismes de création de données manquantes MCAR et MAR sont dits ignorables si les paramètres de la variable dépendante du modèle et ceux de la variable réponse sont distincts (Rubin 1976). Les modèles pour les mécanismes de création des données manquantes de type MNAR sont dits

1. Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institue, 100 Institute Road, Worcester, MA 01609. Courriel : balnan@wpi.edu ; Myron Katzoff, Office of research and Methodology, National Center for Health Statisitcs, CDC, 3311 Toledo Road, Hyattsville, MD 20782. Courriel : mjk5@cdc.gov.

non ignorables. La difficulté générale que pose un modèle de non-réponse non ignorable tient au fait que les paramètres ne sont pas identifiables [par exemple, voir Nandram et Choi (2004, 2005, 2008, 2010), et Nandram, Han et Choi (2002)].

Pour un tableau de contingence $r \times c$, soit $I_{ijkl} = 1$ si le l^{e} individu dans le i^{e} domaine se trouve dans la j^{e} ligne et la k^{e} colonne, et 0 autrement. En outre, soit $J_{il} = 1$ si le l^{e} individu dans le i^{e} domaine a fourni des renseignements complets et 0 autrement. Enfin, soit $P(J_{il} = 1 \mid I_{ijkl} = 1, I_{ij'k'l} = 0, j' \neq j, k' \neq k) = \pi_{ijk}$. Pour la non-réponse totale (ou non-réponse d'une unité), si $\pi_{ijk} = \pi_i$, le modèle est ignorable ; pour la non-réponse partielle (non-réponse à certaines questions), si les valeurs de colonne manquent, mais que les valeurs de ligne sont observées et que $\pi_{ijk} = \pi_{ij}$ (ou $\pi_{ijk} = \pi_i$), le modèle est ignorable ; si les valeurs des lignes manquent, mais que les valeurs de colonne sont observées et que $\pi_{ijk} = \pi_{ik}$ (ou $\pi_{ijk} = \pi_i$), le modèle est ignorable. Tous les autres modèles sont non ignorables ; voir Rubin (1976) pour une explication plus détaillée.

Nandram et Choi (2002a, b) se servent d'un modèle d'extension pour étudier les données binaires en présence de non-réponse non ignorable. Le modèle d'extension, qui est un modèle de non-réponse non ignorable, dégénère en un modèle de non-réponse ignorable (dans l'esprit de Draper 1995) quand la valeur d'un paramètre de centrage est fixée à l'unité. Cela permet d'exprimer la certitude au sujet de l'ignorabilité ; voir également Forster et Smith (1998).

Nous discutons du modèle proposé par Nandram et Choi (2002a, b) pour des données binaires provenant de petits domaines. De sorte que J_{il} désigne les indicateurs de réponse et I_{il}, la réponse binaire. Spécifiquement, en introduisant les paramètres de centrage γ_i pour le domaine *i* afin d'intégrer l'incertitude au sujet de l'ignorabilité, le modèle de Nandram et Choi (2002a, b) est

$$
\begin{aligned}
I_{il} \mid p_i &\overset{\text{iid}}{\sim} \text{Bernoulli}\,(p_i), \\
J_{il} \mid \{\pi_i, J_{il} = 0\} &\overset{\text{iid}}{\sim} \text{Bernoulli}\,(\pi_i), l = 1, \ldots, n_i, i = 1, \ldots, l, \\
J_{il} \mid \{\pi_i, \gamma_i, y_{il} = 1\} &\overset{\text{iid}}{\sim} \text{Bernoulli}\,(\gamma_i \pi_i),\ 0 < \gamma_i \pi_i < 1.
\end{aligned}
$$

Quand $\gamma_i = 1$, le modèle de non-réponse non ignorable dégénère en un modèle de non-réponse ignorable. Ici, γ_i est le ratio des chances de succès parmi les répondants aux chances de succès parmi l'ensemble des individus pour le i^{e} domaine. Le paramètre γ_i décrit la portée de la non-ignorabilité du mécanisme de réponse pour le domaine *i*, et c'est donc grâce à ce paramètre γ_i qu'est intégrée l'incertitude au sujet de l'ignorabilité. Nandram et Choi (2002a, b) définissent $\delta_i = \pi_i \{\gamma_i p_i + (1 - p_i)\}$ comme étant la probabilité qu'un individu du domaine *i* réponde dans l'ensemble de la population et, en croyant que tous les domaines sont semblables, ils considèrent que les $(p_i,\ \delta_i, \gamma_i)$ suivent une loi commune. A priori, ils choisissent des lois bêta pour p_i et π_i, respectivement.

Ici, les paramètres ne sont pas identifiables. Cependant, quand $\gamma_i = 1$, ils le sont tous. Autrement dit, le caractère identifiable des paramètres dépend de γ_i. Notons que si $\gamma_i = 1$, nous obtenons un modèle ignorable pour un mécanisme MAR. Comme les paramètres de ce modèle sont identifiables, il est relativement logique de l'utiliser (ou des modèles similaires) comme modèle de référence. Il faut toutefois souligner que ce modèle n'est toujours pas justifié, parce qu'il repose sur l'hypothèse que les données manquantes ressemblent aux données observées. Donc, pour rendre ce modèle de non-réponse ignorable plus souple, nous utilisons le paramètre γ_i.

Soit γ_{iuv} le nombre d'individus pour lesquels $I_{il} = u$, $J_{il} = v (u, v = 0, 1)$ dans le i^{e} domaine. Alors, sous le modèle,

$$
\begin{aligned}
(y_{i00}, y_{i01}, y_{i10}, y_{i11}) \mid \pi_i, p_i, \gamma_i &\overset{\text{ind}}{\sim} \text{Multinomiale}\,\{n_i, \\
&(1 - p_i)(1 - \pi_i), (1 - p_i)\pi_i, (1 - \gamma_i \pi_i) p_i, \gamma_i \pi_i p_i\}
\end{aligned}
$$

avec indépendance sur les domaines. Ici, y_{i01} et y_{i11} seulement sont observés et, par conséquent, tous les paramètres sont non identifiables si les γ_i sont inconnus. Nous obtenons la fonction de vraisemblance de la même manière pour le tableau de contingence $r \times c$ plus complet avec données manquantes.

Nous partons d'une loi gamma et, pour permettre le centrage sur le modèle de non-réponse ignorable, nous devons sélectionner chaque γ_i de manière que sa moyenne soit égale à 1. Cependant, nous devons utiliser une loi gamma tronquée, parce que $0 < \pi_i < 1$ et $0 < \gamma_i \leq 1 / \pi_i$. Une idée intéressante de Nandram et Choi (2002a, b) consiste à modéliser le centrage sous forme d'une loi gamma tronquée

$$
\gamma_i \mid \nu \overset{\text{iid}}{\sim} \text{Gamma}(\nu, \nu), 0 < \gamma_i < 1 / \pi_i, 0 < \pi_i < 1.
$$

Le modèle est complet et possède des densités de probabilité a priori non informatives sur tous les hyperparamètres. D'autres distributions peuvent être choisies (par exemple une densité lognormale tronquée) pour les γ_i, mais il ne s'agit pas d'un problème essentiel et cela n'aurait pas beaucoup d'importance.

On peut se servir d'un modèle au niveau du domaine avec effets aléatoires dans lequel, conditionnellement aux données observées, la non-réponse dépend des effets aléatoires au niveau du domaine. Ce modèle peut être formulé en utilisant une fonction de lien logit, mais nous

n'avons pas suivi cette direction pour élaborer nos modèles, en partie parce que nous n'utilisons pas de covariables ici ; voir Nandram et Choi (2010) pour l'utilisation de covariables et d'effets aléatoires.

L'approche décrite dans Nandram et Choi (2002a, b) est intéressante, mais elle ne s'applique pas directement au problème des tableaux de contingence $r \times c$ qui nous occupe. Plus précisément, dans Nandram et Choi (2002a, b), un seul paramètre de centrage est nécessaire par domaine. Dans notre formulation, nous avons besoin de *rc* paramètres de centrage par domaine ; chacun de ces paramètres doit suivre une loi centrée sur l'unité pour permettre la dégénérescence en modèle de non-réponse ignorable. Des contraintes d'inégalité doivent également être incluses dans le modèle de non-réponse non ignorable. En outre, on ne peut écarter la possibilité que ces paramètres soient corrélés. La méthodologie nécessaire pour appliquer les travaux de Nandram et Choi (2002a, b) au tableau de contingence $r \times c$ n'est pas simple. Constatant ces difficultés, Nandram, Liu, Choi et Cox (2005) (avec un seul tableau supplémentaire) et Nandram, Cox et Choi (2005) (avec les trois tableaux supplémentaires) appliquent une idée plus simple, mais moins élégante que dans Nandram et Choi (2002a, b) pour procéder au centrage ; voir aussi Nandram et Choi (2005).

Essentiellement, Nandram, Cox et Choi (2005), ainsi que Nandram, Liu, Cox et Choi (2005) émettent l'hypothèse d'un modèle ignorable, obtiennent des échantillons des probabilités de réponse et se servent de ces probabilités de réponse échantillonnées pour ajuster les probabilités de réponse d'un modèle de non-réponse non ignorable en « contrôlant » ce paramètre. Naturellement, une alternative est possible lorsqu'il existe de l'information sur le degré de non-ignorabilité. Toutefois, l'intégration d'information a priori au sujet d'un écart systématique par rapport à l'ignorabilité est plus complexe dans le cas de notre problème et elle nécessiterait du travail sur le terrain supplémentaire coûteux afin d'obtenir cette information.

Nous discutons maintenant de notre conception du problème de non-réponse non ignorable, fondamentalement un problème de distorsion. En fait, ce problème est extrêmement difficile et nous pensons qu'il n'a vraiment aucune solution, mais que nous devons essayer d'en trouver une. Sans aucune information, il n'est pas possible de dire quelles sont les différences entre les répondants et les non-répondants. Un modèle de non-réponse ignorable est restreint, parce qu'il suppose que les répondants et les non-répondants sont semblables, alors qu'ils peuvent être différents. Les statisticiens doivent non seulement faire face à l'imprécision (erreur d'échantillonnage), mais aussi être suffisamment audacieux pour étudier la subjectivité (l'ignorance découlant de l'information manquante). Malheureusement, il est bien connu que les modèles de non-réponse non ignorable contiennent des paramètres non identifiables. Nous discutons de la façon dont sont identifiés les paramètres clés de non-ignorabilité. Nous savons que si les répondants et les non-répondants sont semblables, les γ_i sont égaux à l'unité, et nous obtenons le modèle de non-réponse ignorable à l'aide de tous les paramètres identifiés. Nous pouvons maintenant étendre le modèle de non-réponse ignorable à un modèle de non-réponse non ignorable en donnant à ces paramètres γ_i une loi centrée à 1, tout en maintenant l'identifiabilité. Un modèle de non-réponse non ignorable peut être formulé pour ajouter de la souplesse au modèle de non-réponse ignorable, comme nous l'avons fait dans nos travaux ; la souplesse est une forme d'analyse de sensibilité, cohérente dans le cas qui nous occupe, et il s'agit en effet d'une évaluation bayésienne de l'incertitude (du risque) (par exemple Greenland 2009). C'est ce que nous avons fait ou essayé de faire dans nos travaux.

Dans le présent article, nous tentons de résoudre le problème difficile de Nandram et Choi (2002a, b) sous sa forme originale pour les tableaux $r \times c$ pour de nombreux domaines. Le plan de l'exposé est le suivant. À la section 2, nous décrivons le modèle hiérarchique bayésien. Précisément, nous décrivons le mécanisme de non-réponse non ignorable et nous construisons une loi a priori propre. À la section 3, nous montrons comment ajuster le modèle en utilisant l'algorithme d'échantillonnage avec rééchantillonnage par importance (SIR, pour *sampling importance resampling*) pour effectuer un sous-échantillonnage à partir d'une densité de probabilité a posteriori approximative après une agrégation innovatrice de la densité a posteriori conjointe complète. À la section 4, nous illustrons notre méthodologie en nous servant de données à grande diffusion recueillies dans 13 États dans le cadre de la troisième édition de la National Health and Nutrition Examination Survey (NHANES III). À la section 5, nous présentons nos conclusions.

2. Le modèle de non-réponse non ignorable

Dans le contexte du problème de non-réponse dans un tableau à double entrée, on peut avoir affaire aussi bien à la non-réponse partielle qu'à la non-réponse totale. Donc, on peut considérer la totalité du tableau de données comme étant constitué de quatre tableaux, à savoir un tableau pour les données complètes et trois tableaux supplémentaires, pour l'information de ligne manquante, pour l'information de colonne manquante et pour l'information de ligne ainsi que de colonne manquante, respectivement. Dans tous l'exposé, nous donnons aux lignes l'indice $j = 1, \ldots, r$, aux colonnes l'indice $k = 1, \ldots, c$, et aux quatre tableaux l'indice $s = 1, 2, 3, 4$. Nous donnons aux domaines l'indice

$i = 1, 2, \ldots, A$ et aux individus dans les domaines, l'indice $l = 1, 2, \ldots, n_i$. Nous allons maintenant décrire le modèle de non-réponse non ignorable (c'est-à-dire le modèle avec extension).

2.1 Processus d'échantillonnage

Nous adaptons la terminologie et les définitions utilisées dans Nandram, Cox et Choi (2005) à la situation considérée ici. Pour l'individu *l* échantillonné dans le domaine *i*, soit

$$I_{ijkl} = \begin{cases} 1, & \text{si la catégorie de résultat est } (j,k) \\ 0, & \text{autrement,} \end{cases}$$

et soit $\boldsymbol{J}_{il}$ l'un des quadruplets $(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)$. Nous supposons que

$$\boldsymbol{I}_{il} \overset{\text{def}}{=} \text{vec}(\{I_{ijkl} \mid j = 1, \ldots, r;\ k = 1, \ldots, c\}) \mid \boldsymbol{p}_i \overset{\text{iid}}{\sim} \text{Mult}\{1, \boldsymbol{p}_i\} \tag{1}$$

et

$$\boldsymbol{J}_{il} \mid \{I_{ijkl} = 1,\ I_{ij'k'l} = 0 \text{ pour tout } j' \neq j \text{ et } k' \neq k \mid \boldsymbol{\pi}_{ijk}\} \overset{\text{iid}}{\sim} \text{Mult}\{1, \boldsymbol{\pi}_{ijk}\}, \tag{2}$$

où $\boldsymbol{p}_i \overset{\text{def}}{=} \text{vec}(\{p_{ijk} \mid j = 1, 2, \ldots, r;\ k = 1, 2, \ldots, c\})$ est un vecteur de probabilités pour le tableau de *rc* catégories pour la variable observée dont la somme doit être égale à l'unité, et pour la cellule (j,k) dans ce tableau bidimensionnel,

$$\boldsymbol{\pi}_{ijk} \overset{\text{def}}{=} \text{vec}(\{\pi_{isjk}\} \text{ pour } s = 1,2,3,4)$$

est un vecteur de probabilités dont la somme doit être égale à l'unité.

Ensuite, définissons les fréquences de cellule y_{isjk}, pour chaque tableau $s = 1, \ldots, 4$ pour le domaine *i*, telles que, pour la cellule (j,k),

$$(y_{i1\,jk}, y_{i2\,jk}, y_{i3\,jk}, y_{i4\,jk}) = \sum_{l=1}^{n_i} I_{ijkl}\, \boldsymbol{J}_{il},$$

où les $y_{i1\,jk}$ sont observées et les y_{isjk}, pour $s = 2, 3, 4$, sont des variables latentes qui satisfont les contraintes observées $\sum_k y_{i2\,jk} = u_{ij}$, $\sum_j y_{i3\,jk} = v_{ik}$ et $\sum_{j,k} y_{i4\,jk} = w_i$. Toutes les inférences seront faites conditionnellement aux quantités observées, u_{ij}, v_{ik} et w_i. Voir Nandram (2009) pour l'analyse d'un seul tableau $r \times c$ sous non-réponse quand les marges sont également aléatoires. Nous désignerons le vecteur des $y_{i1\,jk}$ par $\boldsymbol{y}_1$, le vecteur des y_{isjk}, $s = 2, 3, 4$, par $\boldsymbol{y}_{(1)}$, et le vecteur complet par $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_{(1)})'$.

Les paramètres π_{isjk} ne sont pas identifiables. Si les distributions de ces paramètres sont entièrement connues, la non-identifiabilité disparaît. Donc, le problème essentiel est la façon d'identifier ces paramètres. Nous savons que, si les répondants et les non-répondants sont semblables (c'est-à-dire en ce qui concerne les quatre tableaux, complet et partiellement complets) nous pouvons prendre $\pi_{isjk} = \pi_{is}$, ce qui est le modèle de non-réponse ignorable. Les π_{is} peuvent être estimées par les proportions de cas qui, pour chaque domaine, se retrouvent dans chacun des quatre tableaux. Il s'agit d'un point de départ naturel. Afin d'étendre le modèle de non-réponse ignorable à un modèle de non-réponse non ignorable tout en maintenant l'identifiabilité, nous devons d'abord procéder à une simplification. Nous prenons $\pi_{ijks} = \psi_{ijk}\ \pi_{is}$, qui donne un modèle de non-réponse non ignorable dans lequel les paramètres ψ_{ijk} ne sont pas identifiables.

Pour centrer le modèle non ignorable sur le modèle ignorable, nous prenons

$$\pi_{isjk} = \begin{cases} \tilde{\psi}_{ijk}\pi_{is}, & \text{pour } s = 1, \\ \psi_{ijk}\pi_{is}, & \text{pour } s = 2,3,4, \end{cases} \tag{3}$$

et exigeons que $\sum_{s=1}^{4} \pi_{is} = 1$. Quelques opérations algébriques donnent alors la relation

$$\begin{aligned} \tilde{\psi}_{ijk}\pi_{i1} &= \left[1 + (1 - \psi_{ijk})\left(\frac{1 - \pi_{i1}}{\pi_{i1}}\right)\right]\pi_{i1} \\ &= a_{ijk}(\pi_{i1}, \psi_{ijk})\,\psi_{ijk}\pi_{i1}, \end{aligned} \tag{4}$$

où $a_{ijk}(\pi_{i1}, \psi_{ijk}) = \{\psi_{ijk}^{-1} + (\psi_{ijk}^{-1} - 1)(\pi_{i1}^{-1} - 1)\}$, dont il découle clairement que $\tilde{\psi}_{ijk} = 1$ si, et seulement si, $\psi_{ijk} = 1$. Soulignons que, puisque $0 \le \pi_{isjk} \le 1$ et $(1 - \pi_{i1})^{-1} \le \min\{\pi_{is}^{-1}; s = 2,3,4\}$, il s'ensuit que $0 < \psi_{ijk} \le (1 - \pi_{i1})^{-1}$.

En combinant (1) et (2), et en notant la définition de π_{isjk} dans (3), comme dans le cas binaire, nous obtenons une distribution multinomiale pour $\boldsymbol{y}$ sachant $\boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{p}$, et nous voyons maintenant que la fonction de vraisemblance pour l'échantillon est

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{p}) &= \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \left\{ \left[\prod_{j,k} (\tilde{\psi}_{ijk}\pi_{i1}p_{ijk})^{y_{i1\,jk}} \right] \right. \\ &\qquad \left. \prod_{s=2}^{4} \prod_{j,k} (\psi_{ijk}\pi_{is}p_{ijk})^{y_{isjk}} \right\} \\ &= \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \left\{ \prod_{s=1}^{4} \prod_{j,k} (\psi_{ijk}\pi_{is}p_{ijk})^{y_{isjk}} \right. \\ &\qquad \left. \prod_{j,k} \left[a_{ijk}(\pi_{i1}, \psi_{ijk}) \right]^{y_{i1\,jk}} \right\}, \end{aligned} \tag{5}$$

où

$$\boldsymbol{y}_{is}^{rc\times 1} = \text{vec}(\{y_{isjk} \mid j=1,\ldots,r; k=1,\ldots,c\}),$$
$$\boldsymbol{y} = (\boldsymbol{y}'_{11}, \boldsymbol{y}'_{12}, \boldsymbol{y}'_{13}, \boldsymbol{y}'_{14}, \boldsymbol{y}'_{21}, \ldots, \boldsymbol{y}'_{24}, \ldots, \boldsymbol{y}'_{A1}, \boldsymbol{y}'_{A2}, \boldsymbol{y}'_{A3}, \boldsymbol{y}'_{A4})',$$
$$\boldsymbol{\pi}^{A4\times 1} = (\pi_{11}, \ldots, \pi_{14}, \pi_{21}, \ldots, \pi_{24}, \ldots, \pi_{A1}, \ldots, \pi_{A4})',$$
$$\boldsymbol{\psi}^{Arc\times 1} = (\psi_{111}, \ldots, \psi_{1rc}, \psi_{211}, \ldots \psi_{2rc}, \ldots, \psi_{A11}, \ldots, \psi_{Arc}),$$

et

$$\boldsymbol{p}^{Arc\times 1} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_A)'.$$

En obtenant des facteurs qui sont des puissances de π_{is}, la fonction de vraisemblance peut également être exprimée sous la forme

$$f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{p}) = \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \left\{ \prod_{s=1}^{4} \pi_{is}^{y_{is..}} \times \prod_{j,k} \{p_{ijk}\ \psi_{ijk}\}^{y_{i.jk}} \left[a_{ijk}(\pi_{i1}, \psi_{ijk}) \right]^{y_{i1jk}} \right\}, \quad (6)$$

où $0 \le \pi_{is} \le 1, \sum_s \pi_{is} = 1$ et $0 \le \psi_{ijk} \le (1-\pi_{i1})^{-1}$. Ici, nous notons que $y_{is..}$ et y_{i1jk} sont des variables observées, mais que les $y_{i.jk}$ sont des variables latentes.

2.2 Construction des lois a priori

Les hypothèses qui suivent décrivent les lois a priori pour le modèle de non-réponse non ignorable :

1. Pour le vecteur de probabilités de cellule $\boldsymbol{p}_i$, nous supposons que

$$\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1 \tau_1),$$

où $\mu_1 = (\mu_{111}, \mu_{112}, \ldots, \mu_{11k}, \mu_{121}, \ldots, \mu_{1rc})'$; $\mu_{1jk} \ge 0$ et $\sum_{j=1}^{r}\sum_{k=1}^{c} \mu_{1jk} = 1$. Le paramètre τ_1 nous informe de la similarité entre les $\boldsymbol{p}_i$: plus la valeur de τ_1 est grande, plus les $\boldsymbol{p}_i$ se ressemblent. Il en est ainsi parce qu'une grande valeur de τ_1 signifie que les variances des $\boldsymbol{p}_i$ sont faibles, et comme elles ont la même moyenne, cela signifie qu'elles sont plus semblables quand τ_1 est grand.

Donc, pour $\boldsymbol{p}$, la densité de probabilité est

$$g_1(\boldsymbol{p} \mid \boldsymbol{\mu}_1, \tau_1) = \prod_{i=1}^{A} g_{1i}(\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1) = \prod_{i=1}^{A} \left\{ \frac{\prod_{j,k} p_{ijk}^{\mu_{1jk}\tau_1 - 1}}{\text{D}(\boldsymbol{\mu}_1 \tau_1)} \right\}, \quad (7)$$

où, pour un k-uplet $\boldsymbol{c}$ et un scalaire t

$$\text{D}(\boldsymbol{ct}) = \frac{\prod_{j=1}^{k} \Gamma(\boldsymbol{c}_j t)}{\Gamma(t)}$$

pour $c_j > 0$ et $\sum_{j=1}^{k} c_j = 1$.

2. Indépendamment des $\boldsymbol{p}_i$, les $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})'$ suivent la spécification

$$\boldsymbol{\pi}_i \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2 \tau_2),$$

avec $\pi_{is} \ge 0$ et $\sum_s \pi_{is} = 1$, où $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22}, \mu_{23}, \mu_{24})', \mu_{2s} \ge 0, \sum_{s=1}^{4} \mu_{2s} = 1$ et τ_2 est une mesure de la similarité entre les $\boldsymbol{\pi}_i$. Donc, pour $\boldsymbol{\pi}_i$, , la densité de probabilité est

$$g_{2i}(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2) = \frac{\prod_{s=1}^{4} \pi_{is}^{\mu_{2s}\tau_2 - 1}}{\text{D}(\boldsymbol{\mu}_2 \tau_2)}. \quad (8)$$

3. Pour chaque i, soit $\boldsymbol{\psi}_i = (\psi_{i11}, \ldots, \psi_{i1c}, \psi_{i21}, \ldots, \psi_{i2c}, \ldots, \psi_{irc})'$ de sorte que $\boldsymbol{\psi} = (\boldsymbol{\psi}'_1, \ldots, \boldsymbol{\psi}'_A)'$. Nous supposons que, pour chaque i, les ψ_{ijk} sont indépendants et identiquement distribués suivant une distribution dérivée de la loi Gamma(β, β), où le support est limité à l'intervalle ouvert $(0, (1-\pi_{i1})^{-1})$; autrement dit, la loi gamma ordinaire est tronquée comme il suit

$$\psi_{ijk} \mid \beta, \boldsymbol{\pi}_i \overset{\text{ind}}{\sim} \text{Gamma}(\beta, \beta)$$
de sorte que $0 < \psi_{ijk} < (1-\pi_{i1})^{-1}$.

Il convient de souligner que ces ψ_{ijk} sont identiquement distribués sur j et sur k. De nouveau, nous pourrions utiliser d'autres distributions, telles que la densité lognormale tronquée, mais cela ne change pas grand-chose. Dans cette formulation, il existe une certaine information au sujet de β parce que nous supposons que les petits domaines partagent un effet commun.

Donc, pour le domaine i, la densité de probabilité pour $\boldsymbol{\psi}_i$ est

$$g_{3i}(\boldsymbol{\psi}_i \mid \beta, \boldsymbol{\pi}_i) = \prod_{j=1}^{r}\prod_{k=1}^{c} \left\{ \frac{\beta^\beta \psi_{ijk}^{\beta-1} e^{-\beta\psi_{ijk}}}{\Gamma(\beta)} \Big/ \int_0^{(1-\pi_{i1})^{-1}} \frac{\beta^\beta \psi_{ijk}^{\beta-1} e^{-\beta\psi_{ijk}}}{\Gamma(\beta)} \text{d}\psi_{ijk} \right\},$$

pour $0 < \psi_{ijk} < (1-\pi_{i1})^{-1}$. En procédant à la transformation $t_{ijk} = \beta\psi_{ijk}$, nous voyons que la constante de normalisation dans le dénominateur de chacun des facteurs qui figurent dans $g_{3i}(\boldsymbol{\psi}_i \mid \beta, \boldsymbol{\pi}_i)$ est $G_\beta[\beta(1-\pi_{i1})^{-1}]$, où $G_\beta(\cdot)$ est la fonction gamma avec paramètre d'échelle β. Pour que l'intervalle d'intégration ne dépende pas de π_{i1}, posons que $\phi_{ijk} = (1-\pi_{i1})\psi_{ijk}$ et que $\boldsymbol{\phi}_i = (\phi_{i11}, \ldots, \phi_{i1c}, \phi_{i21}, \ldots, \phi_{i2c}, \ldots, \phi_{irc})'$. Alors

$$g_{3i}(\boldsymbol{\phi}_i \mid \beta, \boldsymbol{\pi}_i) = \prod_{j=1}^{r}\prod_{k=1}^{c}\left\{ \frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)\, G_{\beta}[\beta\,(1-\pi_{i1})^{-1}]} \right\}, \quad (9)$$

pour $0 < \phi_{ijk} < 1$. La loi priori conjointe de $\boldsymbol{\pi}_i$ et $\boldsymbol{\phi}_i$ est simplement le produit de $g_{3i}(\boldsymbol{\phi}_i \mid \beta, \boldsymbol{\pi}_i)$ et de $g_{2i}(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2)$. Donc, la loi a priori conjointe de $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_A)'$ et de $\boldsymbol{\pi}$ est

$$g^*(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{\mu}_2, \tau_2, \beta) \overset{\text{def}}{=} \prod_{i=1}^{A} \left\{ g_{3i}(\boldsymbol{\phi}_i \mid \boldsymbol{\beta}, \boldsymbol{\pi}_i) \cdot g_{2i}(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2) \right\}.$$

Autrement dit

$$g^*(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{\mu}_2, \tau_2, \beta) = \prod_{i=1}^{A} \left\{ \frac{\prod_{s=1}^{4} \pi_{is}^{\mu_{2s}\tau_2 - 1}}{\mathrm{D}(\boldsymbol{\mu}_2 \tau_2)} \prod_{j=1}^{r}\prod_{k=1}^{c} \frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)\, G_{\beta}[\beta(1-\pi_{i1})^{-1}]} \right\}. \quad (10)$$

Pour achever la description du modèle, nous spécifions les hypothèses concernant les hyperparamètres. Comme il n'existe pas de lois a priori conjuguées, nous utilisons des lois a priori de rétrécissement pour τ_1, τ_2 et β, parce qu'elles sont propres et non informatives. Les lois a priori de la forme $p(\tau_1) \propto 1/\tau_1$, en particulier les lois a priori gamma diffuses propres, sont déconseillées ; voir, par exemple, Gelman (2006). Des demi-densités de la loi de Cauchy et des densités de probabilité de la loi gamma sont d'autres options (pour lesquelles il faudrait spécifier les hyperparamètres). Donc, nous prenons

1. τ_1, τ_2 et β ayant des lois a priori de rétrécissement indépendantes de la forme

   $$f(x) = \frac{a_0}{(a_0 + x)^2}, \text{ pour } x \geq 0,$$

   où a_0 est spécifié ; il est courant en pratique de prendre $a_0 = 1$.

2. Nous supposons aussi que $\boldsymbol{\mu}_1 \sim \text{Dirichlet}(1, 1, \ldots, 1)$ et $\boldsymbol{\mu}_2 \sim \text{Dirichlet}(1, 1, 1, 1)$.

Soit $\Omega = (\beta, \boldsymbol{\mu}_1, \tau_1, \boldsymbol{\mu}_2, \tau_2)$. La densité de probabilité pour Ω est alors

$$p(\Omega) = \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2} (rc - 1)!\, 3!$$

pour τ_1, τ_2 et $\beta \geq 0$, $\sum_{j,k} \mu_{1jk} = 1$ et $\sum_{s=1}^{4} \mu_{2s} = 1$.

En vertu du théorème de Bayes, la densité de probabilité a posteriori conjointe est

$$\begin{aligned} h(\Omega, \boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \propto\ & f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{p})\, g_1(\boldsymbol{p} \mid \boldsymbol{\mu}_1, \tau_1)\, g^*(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{\mu}_2, \tau_2, \beta)\, p(\Omega) \\ =\ & \prod_{i=1}^{A} \left[\binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} (1-\pi_{i1})^{-y_{i\cdots}} \right. \\ & \left\{ \prod_{s=1}^{4} \pi_{is}^{y_{is\cdot}} \times \prod_{j,k} \{p_{ijk}\phi_{ijk}\}^{y_{i\cdot jk}} \left[a_{ijk}(\pi_{i1}, \phi_{ijk}) \right]^{y_{i1jk}} \right\} \\ & \times \left\{ \frac{\prod_{j,k} p_{ijk}^{\mu_{1jk}\tau_1 - 1}}{\mathrm{D}(\boldsymbol{\mu}_1 \tau_1)} \right\} \left\{ \frac{\prod_{s=1}^{4} \pi_{is}^{\mu_{2s}\tau_2 - 1}}{\mathrm{D}(\boldsymbol{\mu}_2 \tau_2)} \right\} \\ & \left. \prod_{j=1}^{r}\prod_{k=1}^{c} \frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)\, G_{\beta}[\beta(1-\pi_{i1})^{-1}]} \right] \\ & \times \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2}, \end{aligned} \quad (11)$$

où, en substituant $(1-\pi_{i1})^{-1}\phi_{ijk}$ pour ψ_{ijk},

$$a_{ijk}(\pi_{i1}, \phi_{ijk}) = \left(\frac{1-\pi_{i1}}{\phi_{ijk}} \right) \left[1 + \frac{1}{\pi_{i1}} \left\{ 1 - \pi_{i1} - \phi_{ijk} \right\} \right]. \quad (12)$$

Pour faire des inférences au sujet des p_{ijk}, nous tirerons des échantillons de $h(\Omega, \boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ selon des méthodes Monte Carlo par chaîne de Markov. Cette procédure est décrite à la section 3.

3. Calculs

Nous utilisons l'algorithme SIR pour sous-échantillonner un échantillon aléatoire tiré d'une densité de probabilité a posteriori approximative. L'exécution de cette tâche se fait en trois étapes. Nous agrégeons sur les $\boldsymbol{p}_i, \boldsymbol{\pi}_i$ et $\boldsymbol{\phi}_i$, approximons la densité de probabilité agrégée par une densité plus simple et effectuons l'échantillonnage à partir de cette densité, puis nous sous-échantillonnons ces échantillons pour obtenir des échantillons tirés de la densité de probabilité originale. Nous montrons à la présente section comment exécuter ces trois étapes.

Pour obtenir l'approximation et pour simplifier les calculs, à l'annexe A, nous procédons à l'agrégation sur les $\boldsymbol{p}_i, \boldsymbol{\pi}_i$ et $\boldsymbol{\phi}_i$ pour obtenir

$$h(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \pi_a(\Omega\, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \cdot \prod_{i=1}^{A} I_i,$$

où

$$I_i = \iiint_0^{\frac{\beta b_i}{1-\pi_{i1}}} \frac{G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)}{\left[G_\beta\left(\frac{\beta}{1-\pi_{i1}}\right)\right]^{rc}}$$
$$\prod_{j,k}\left\{\left(\frac{W_i}{\beta}\right)\sum_{s=2}^{4} y_{isjk}\left[\frac{1}{\phi^*_{ijk}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{W_i\phi^*_{ijk}}{\beta}\right\}\right)\right]^{y_{i1jk}}\right\}$$
$$\frac{W_i^{rc\beta-1}\,e^{-W_i}}{\Gamma(rc\beta)G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)}\,\mathrm{d}W_i\left\{\frac{\prod_{j,k}\phi^{*\ y_{i\cdot jk}+\beta-1}_{ijk}}{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\boldsymbol{j})}\right\}$$
$$\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot}+\mu_2 s\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}\right\}\mathrm{d}\boldsymbol{\phi}^*_i\,\mathrm{d}\boldsymbol{\pi}_i, \qquad (13)$$

avec $b_i = \min\{\{1/\phi^*_{ijk}\},\ j=1,\ldots,r;\ k=1,\ldots,c\}$ et

$$\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$$
$$= \frac{a_0}{(a_0+\tau_1)^2}\cdot\frac{b_0}{(b_0+\tau_2)^2}\cdot\frac{c_0}{(c_0+\beta)^2}$$
$$\prod_{i=1}^{A}\frac{\Gamma(rc\beta)}{[\Gamma(\beta)]^{rc}}\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\boldsymbol{j})$$
$$\times\prod_{i=1}^{A}\binom{n_i}{\boldsymbol{y}'_{i1},\boldsymbol{y}'_{i2},\boldsymbol{y}'_{i3},\boldsymbol{y}'_{i4}}\frac{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\boldsymbol{\mu}_1\tau_1)}{\mathrm{D}(\boldsymbol{\mu}_1\tau_1)}\frac{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}{\mathrm{D}(\boldsymbol{\mu}_2\tau_2)}. \qquad (14)$$

Pour évaluer I_i pour chaque $i=1,\ldots,A$, nous procédons comme il suit sachant $(\Omega, \boldsymbol{y}_{i(1)})$:

1. Tirer des échantillons indépendants des vecteurs $\boldsymbol{\pi}_i$ et $\boldsymbol{\phi}^*_i$ des distributions Dirichlet$(\boldsymbol{y}_i^{(2)} + \boldsymbol{\mu}_2\tau_2)$ et Dirichlet$(\boldsymbol{y}_i^{(1)} + \beta\,\boldsymbol{j})$, respectivement. Pour chaque $\boldsymbol{\pi}_i$ et $\boldsymbol{\phi}^*_i$, tirer un échantillon des valeurs de W_i de la distribution gamma tronquée sur l'intervalle $(0, \{\beta b_i/1-\pi_{i1}\})$ avec le paramètre $rc\beta$.
2. Pour chaque $\boldsymbol{\pi}_i, \boldsymbol{\phi}^*_i$ et W_i sélectionné à l'étape (1), calculer R_1R_2, où

$$R_1 = G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)\Bigg/\left[G_\beta\left(\frac{\beta}{1-\pi_{i1}}\right)\right]^{rc} \qquad (15)$$

et

$$R_2 = \prod_{j,k}\left\{\left(\frac{W_i}{\beta}\right)\sum_{s=2}^{4} y_{isjk}\left[\frac{1}{\phi^*_{ijk}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{W_i\phi^*_{ijk}}{\beta}\right\}\right)\right]^{y_{i1jk}}\right\}. \qquad (16)$$

3. Répéter les étapes (1) et (2) 1 000 fois. Puis, calculer la moyenne de R_1R_2 sur ces 1 000 valeurs.

La suite des calculs comporte deux parties. Premièrement, nous utilisons l'échantillonneur de Metropolis-Hastings « griddy » (technique d'approximation par des grilles) pour tirer des échantillons de $\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$. Nous échantillonnons μ_1, μ_2, τ_1 et τ_2 à partir de leur densité de probabilité a posteriori conditionnelle en utilisant des grilles ; cela comporte la transformation de τ_1 et τ_2 pour les faire varier dans l'intervalle unitaire $(0, 1)$. Pour chaque distribution, nous utilisons 100 grilles ; voir Nandram, Cox et Choi (2005) pour une procédure similaire. Ici, le tirage de $\boldsymbol{y}_{(1)}$ est effectué par échantillonnage de la fonction de masse de probabilité conditionnelle par composante. Des échantillons sont tirés de la densité a posteriori conditionnelle de β en utilisant un pas de Metropolis de manière similaire à Nandram et Choi (2002a, b). Nous avons exécuté cet algorithme 11 000 fois en permettant un « rodage » de 1 000 itérations. Nous avons constaté que les autocorrélations entre les itérations étaient faibles, ce qui indiquait que l'échantillonneur produisait un bon mélange. Nous avons également utilisé la méthode des moyennes de lot pour étendre l'évaluation des calculs. Nous avons utilisé des lots de 25 pour calculer les erreurs-types numériques.

Deuxièmement, nous nous servons de l'algorithme SIR pour sous-échantillonner l'échantillon de 10 000 itérations tirées de $\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$. Pour chacune des 10 000 itérations, nous calculons les poids

$$w_m = \frac{h(\Omega^{(m)}, \boldsymbol{y}_{(1)}^{(m)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}{\pi_a(\Omega^{(m)}, \boldsymbol{y}_{(1)}^{(m)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})},$$
$$m = 1, \ldots, M = 10\,000, \qquad (17)$$

et nous rééchantillonnons $\{\Omega^{(m)}, \boldsymbol{y}_{(1)}^{(m)}\}$ avec probabilités proportionnelles aux poids w_m pour $m=1,\ldots,M$, sans remise. Nous utilisons un échantillonnage à 10 % et nous sous-échantillonnons les 10 000 itérations pour en obtenir 1 000 ; l'échantillonnage sans remise est une bonne idée, parce qu'il permet d'éviter les valeurs répétées qui existent déjà parce que l'échantillonneur de Metropolis-Hastings n'est pas vraiment un échantillonneur de type acceptation-rejet et qu'il donne des valeurs répétées. Comme d'habitude sous échantillonnage sans remise, les poids sont calculés chaque fois qu'une valeur est sélectionnée.

Enfin, nous pouvons maintenant faire une inférence exacte (dans les limites des méthodes Monte Carlo par chaîne de Markov) au sujet de $\boldsymbol{p}_i$ a posteriori. Soit $y_{i\cdot jk} = \sum_{s=1}^{4} y_{isjk}$ et $\boldsymbol{y}_i^*$ le vecteur de $y_{i\cdot jk}$. Alors,

$$\boldsymbol{p}_i \mid \boldsymbol{y}_i^*, \boldsymbol{\mu}_1, \tau_1 \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{y}_i^* + \boldsymbol{\mu}_1\tau_1),\ i=1,\ldots,A.$$

Donc, pour chaque valeur de $\boldsymbol{y}_i^*$, $\boldsymbol{\mu}_1$ et τ_i que nous obtenons au moyen de l'algorithme SIR, nous tirons une valeur de $\boldsymbol{p}_i, i=1,\ldots,A$. D'où nous obtenons une densité Rao-blackwellisée pour chaque $\boldsymbol{p}_i$, et l'inférence se poursuit de la manière habituelle.

4. Un exemple

En guise d'illustration, nous choisissons un exemple dans le domaine de la statistique de la santé. À la section 4.1, nous décrivons brièvement les données provenant de la troisième édition de la National Health and Nutrition Examination Survey (NHANES III) que nous utilisons. En particulier, nous étudions la relation entre la densité minérale osseuse et le revenu familial ; voir Nandram, Cox et Choi (2005) pour une discussion de ce problème. À la section 4.2, après une brève discussion de nos calculs, nous présentons l'inférence a posteriori sur les probabilités de cellule. À la section 4.3, en nous servant du facteur de Bayes, nous discutons de la relation entre la densité minérale osseuse et le revenu familial.

4.1 Données de la NHANES III

Le plan de sondage est un plan probabiliste stratifié à plusieurs degrés qui est représentatif de l'ensemble de la population civile ne vivant pas en établissement, âgée de deux mois et plus, des États-Unis. Des renseignements plus détaillés sur le plan de sondage de la NHANES III sont disponibles ailleurs (National Center for Health Statistics 1992, 1994). La collecte des données de la NHANES III comporte deux volets : le premier comprend la sélection de l'échantillon et l'interview des membres des ménages échantillonnés en vue de recueillir des renseignements personnels et le second comprend l'examen physique des personnes interviewées dans un centre d'examen mobile (CEM). L'évaluation de la santé s'appuie sur un examen physique, des tests et des mesures faites par des techniciens, ainsi que des prélèvements pour l'analyse. L'échantillon a été sélectionné auprès des ménages de 81 unités primaires d'échantillonnage à travers les États-Unis continentaux d'octobre 1988 à septembre 1994. Les données finales retenues pour l'étude proviennent des 35 plus grandes unités primaires d'échantillonnage dont la population est égale ou supérieure à 500 000 habitants, et nous considérons 13 régions infranationales.

La non-réponse peut avoir lieu dans les volets interview et examen physique de l'enquête. La non-réponse à l'interview se produit lorsque les personnes échantillonnées ne participent pas à l'interview. Certaines personnes interviewées et incluses dans le sous-échantillon pour l'évaluation de la santé ont manqué l'examen physique à la maison ou au centre d'examen mobile et n'ont donc pas subi la totalité ou une partie des examens.

Les médecins pensent que les personnes obèses ou ayant un excès de poids ne se présentent généralement pas au CEM. Cohen et Duffy (2002) remarquent que les enquêtes sur la santé sont un bon exemple de situation où il paraît plausible qu'il existe un lien entre la propension à répondre et l'état de santé. La NHANES III est en effet un bon exemple.

Les personnes échantillonnées pour participer à la NHANES III peuvent être classées en fonction d'un grand nombre d'attributs, et les chercheurs analysent ces tableaux de contingence afin de déterminer la qualité de l'ajustement des modèles ou l'indépendance. Ici, nous étudions la densité minérale osseuse (DMO) et le revenu familial (RF). Mentionnons ici que, même si le RF est une variable discrète, nous avons classé la DMO en trois catégories (normale, ostéopénie et ostéoporose) et le RF en trois catégories (faible, moyen et élevé). Cependant, nous ne disposons que d'une classification partielle des individus, parce que certains sont classés en fonction d'un seul attribut, tandis que d'autres ne sont pas classés du tout. Parmi les ménages qui ont participé au volet de l'examen physique, environ 62 % ont fourni des données sur le RF et la DMO, 8 % ont fourni des données sur la DMO seulement, 29 % ont fourni des données sur le revenu seulement et 1 % n'ont fourni de données ni sur le revenu ni sur la DMO. Notre problème consiste à estimer les probabilités de cellule et de tester l'association entre la DMO et le RF pour chacune des 13 régions infranationales en utilisant notre modèle d'extension qui regroupe les données de manière adaptative.

Dans le tableau 1, nous présentons les tableaux 3×3 de la DMO et du RF pour les 13 régions susmentionnées. Notons que les données pour les régions 6 et 48 sont suffisantes pour traiter ces régions individuellement. Par contre, les autres régions sont très petites. Les fréquences dans le tableau contenant des totaux de ligne sont généralement faibles, sauf pour la région 17, et les fréquences dans le tableau contenant juste le total sont faibles. Même pour le tableau contenant les données complètes, les fréquences de cellule sont généralement faibles, ce qui nous oblige à utiliser des techniques d'estimation sur petits domaines pour emprunter de l'information.

4.2 Inférence a posteriori des probabilités de cellule

Nous discutons de la performance de nos calculs pour le modèle d'extension, puis de l'inférence a posteriori au sujet des probabilités de cellule. Nous utilisons la moyenne a posteriori (MP), l'écart-type a posteriori (ETP) et l'intervalle de crédibilité à 95 % pour chaque paramètre d'intérêt. Nous présentons aussi les erreurs-types numériques (ETN) pour évaluer la répétabilité de nos calculs.

Au tableau 2, nous présentons des résumés des distributions a posteriori de $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \tau_1, \tau_2$ et β, avant et après l'application de l'algorithme SIR. Ces résumés sont fort semblables, ce qui indique que l'approximation $\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ donnée par l'algorithme SIR n'est pas déraisonnable. Par exemple, les intervalles de crédibilité à 95 % obtenus pour β avant et après l'application de l'algorithme SIR sont (1,081 ; 1,940) et (1,086 ; 1,947),

respectivement, ce qui représente une bonne concordance. Les estimations de τ_1 et τ_2 devraient présenter les écarts les plus importants, mais elles sont également raisonnablement proches [par exemple pour τ_1, l'intervalle de crédibilité à 95 % est (28,282 ; 64,204) avec l'approximation et (27,962 ; 64,425) avec l'algorithme SIR]. Dans les deux cas, les ETN sont faibles, ce qui signifie que les calculs sont répétables.

Au tableau 3, nous comparons notre modèle d'extension (modèle 3) à deux autres. Le modèle 1 (un modèle de non-réponse ignorable) et le modèle 2 (un modèle de non-réponse non ignorable) (pas de centrage) sont décrits à l'annexe B. Pour l'illustration, nous avons choisi trois régions, une grande, une moyenne et une petite. Des différences se dégagent entre les trois modèles. En général, les grandes estimations ont tendance à être plus petites pour le modèle 2 et encore plus petites pour le modèle 1 que pour le modèle 3 (c'est-à-dire que les estimations provenant du modèle 3 sont naturellement les plus proches du modèle 1 et non du modèle 2). Le modèle 2 produit la variabilité la plus importante ; comme prévu, le modèle 3 donne une variabilité légèrement plus grande que le modèle 1. Faute d'espace, nous ne présentons pas les ETM, mais nous constatons qu'elles sont toutes inférieures à 0,005.

Tableau 1
Fréquences des tableaux 3×3 de la DMO et du RF correspondant aux 13 régions infranationales dans la NHANES III

État	Tableau complet									Total de colonne			Total de ligne			Total
4	21	14	9	8	7	3	2	2	0	11	5	6	4	0	1	1
6	257	127	106	92	51	32	32	5	7	178	54	82	65	28	4	20
12	33	18	21	22	4	4	15	5	0	18	11	16	5	6	2	1
17	25	15	13	8	5	3	0	0	1	18	10	16	17	2	2	4
25	9	7	12	6	5	9	2	1	0	9	6	12	1	4	5	1
26	18	11	18	6	5	9	2	1	1	10	5	11	4	3	0	1
29	9	4	10	3	2	4	3	1	2	9	2	9	0	2	4	1
36	42	17	27	32	13	18	9	6	1	43	21	42	9	7	6	1
39	8	6	14	2	5	4	3	0	1	9	7	5	2	3	0	0
42	14	8	11	12	8	4	8	1	2	35	15	24	3	1	0	0
44	12	9	6	8	5	0	5	1	0	19	4	12	7	1	0	1
48	159	44	22	51	11	13	9	6	2	88	12	23	16	8	2	14
53	14	10	15	10	10	14	3	1	1	9	4	8	2	4	1	0

Nota : Dans le tableau 3×3 complet, le premier (deuxième, troisième) ensemble de trois nombres correspond à la première (deuxième, troisième) ligne ; le total de colonne (ligne) renvoie au tableau 3×3 ne contenant que les totaux de colonne (ligne) ; le total renvoie au tableau 3×3 contenant les totaux uniquement.

Tableau 2
Données de la NHANES pour 13 régions : Comparaison de la densité de probabilité a posteriori approximative et de la densité de probabilité a posteriori correcte en utilisant les moyennes a posteriori (MP), les écarts-types a posteriori (ETP), les erreurs-types numériques (ETN) et les intervalles de crédibilité à 95 % des hyperparamètres

	Approximation				Corrigée			
	MP	ETP	ETN	Int. à 95 %	MP	ETP	ETN	Int. à 95 %
μ_{21}	0,528	0,031	0,001	(0,463 ; 0,582)	0,525	0,031	0,008	(0,456 ; 0,578)
μ_{22}	0,131	0,021	0,001	(0,096 ; 0,181)	0,133	0,021	0,002	(0,094 ; 0,179)
μ_{23}	0,328	0,028	0,001	(0,274 ; 0,383)	0,328	0,028	0,005	(0,269 ; 0,383)
μ_{24}	0,013	0,006	0,000	(0,004 ; 0,027)	0,014	0,006	0,000	(0,004 ; 0,029)
τ_2	21,638	9,559	0,255	(8,347 ; 46,587)	20,078	8,632	0,303	(8,538 ; 38,625)
μ_{111}	0,280	0,023	0,001	(0,234 ; 0,324)	0,277	0,023	0,004	(0,228 ; 0,319)
μ_{112}	0,133	0,016	0,000	(0,102 ; 0,165)	0,134	0,017	0,002	(0,101 ; 0,165)
μ_{113}	0,200	0,019	0,000	(0,163 ; 0,238)	0,199	0,019	0,003	(0,162 ; 0,236)
μ_{121}	0,105	0,015	0,000	(0,078 ; 0,135)	0,107	0,015	0,002	(0,079 ; 0,135)
μ_{122}	0,065	0,011	0,000	(0,044 ; 0,088)	0,065	0,011	0,001	(0,044 ; 0,087)
μ_{123}	0,072	0,012	0,000	(0,050 ; 0,096)	0,073	0,012	0,001	(0,049 ; 0,097)
μ_{131}	0,061	0,011	0,000	(0,041 ; 0,083)	0,061	0,011	0,001	(0,040 ; 0,083)
μ_{132}	0,037	0,008	0,000	(0,023 ; 0,054)	0,036	0,008	0,001	(0,022 ; 0,054)
μ_{133}	0,048	0,009	0,000	(0,031 ; 0,068)	0,048	0,009	0,001	(0,031 ; 0,068)
τ_1	45,960	10,094	0,153	(28,282 ; 64,204)	45,177	10,562	0,679	(27,962 ; 64,423)
β	1,472	0,218	0,004	(1,081 ; 1,940)	1,449	0,208	0,022	(1,086 ; 1,947)

Nota : Les hyperparamètres sont $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \tau_1, \tau_2$ et β.

Tableau 3
Moyennes a posteriori des probabilités de cellule et des intervalles de crédibilité (IC) à 95 % pour trois régions (grande, moyenne et petite) selon les trois modèles

	Modèle 1			Modèle 2			Modèle 3		
Cellule	**MP**	**ETP**	**IC à 95 %**	**MP**	**ETP**	**IC à 95 %**	**MP**	**ETP**	**IC à 95 %**
a. Grande									
(1,1)	0,239	0,044	(0,157 ; 0,326)	0,196	0,046	(0,117 ; 0,295)	0,259	0,038	(0,189 ; 0,335)
(1,2)	0,140	0,035	(0,078 ; 0,213)	0,127	0,035	(0,068 ; 0,200)	0,132	0,029	(0,082 ; 0,197)
(1,3)	0,240	0,044	(0,159 ; 0,332)	0,198	0,047	(0,118 ; 0,301)	0,248	0,037	(0,175 ; 0,322)
(2,1)	0,092	0,032	(0,039 ; 0,162)	0,098	0,040	(0,037 ; 0,188)	0,077	0,022	(0,039 ; 0,126)
(2,2)	0,074	0,028	(0,029 ; 0,136)	0,077	0,030	(0,030 ; 0,144)	0,056	0,020	(0,024 ; 0,099)
(2,3)	0,133	0,036	(0,070 ; 0,210)	0,121	0,042	(0,056 ; 0,219)	0,110	0,028	(0,058 ; 0,168)
(3,1)	0,036	0,020	(0,008 ; 0,083)	0,069	0,039	(0,013 ; 0,153)	0,047	0,018	(0,018 ; 0,086)
(3,2)	0,023	0,015	(0,003 ; 0,061)	0,043	0,025	(0,007 ; 0,100)	0,032	0,014	(0,009 ; 0,063)
(3,3)	0,025	0,017	(0,003 ; 0,066)	0,071	0,040	(0,010 ; 0,154)	0,042	0,016	(0,016 ; 0,079)
b. Moyenne									
(1,1)	0,233	0,034	(0,169 ; 0,302)	0,213	0,043	(0.141 ; 0.305)	0,254	0,032	(0,194 ; 0,318)
(1,2)	0,143	0,028	(0,093 ; 0,200)	0,127	0,032	(0.072 ; 0.196)	0,146	0,024	(0,102 ; 0,197)
(1,3)	0,190	0,031	(0,132 ; 0,254)	0,140	0,034	(0.084 ; 0.218)	0,208	0,027	(0,156 ; 0,259)
(2,1)	0,174	0,031	(0,118 ; 0,237)	0,160	0,042	(0.092 ; 0.249)	0,154	0,027	(0,106 ; 0,211)
(2,2)	0,043	0,018	(0,015 ; 0,083)	0,060	0,028	(0.017 ; 0.124)	0,032	0,012	(0,012 ; 0,059)
(2,3)	0,049	0,020	(0,017 ; 0,095)	0,065	0,031	(0.018 ; 0.136)	0,042	0,014	(0,020 ; 0,072)
(3,1)	0,112	0,025	(0,068 ; 0,167)	0,120	0,041	(0.059 ; 0.209)	0,092	0,020	(0,056 ; 0,134)
(3,2)	0,047	0,018	(0,018 ; 0,088)	0,059	0,026	(0.019 ; 0.118)	0,040	0,014	(0,018 ; 0,071)
(3,3)	0,010	0,009	(0,000 ; 0,033)	0,056	0,032	(0.006 ; 0.122)	0,032	0,012	(0,013 ; 0,059)
c. Petite									
(1,1)	0,196	0,052	(0,103 ; 0,305)	0,164	0,055	(0.077 ; 0.288)	0,253	0,043	(0,175 ; 0,334)
(1,2)	0,081	0,034	(0,028 ; 0,158)	0,081	0,032	(0.030 ; 0.155)	0,091	0,028	(0,043 ; 0,152)
(1,3)	0,213	0,052	(0,118 ; 0,323)	0,175	0,055	(0.087 ; 0.300)	0,220	0,043	(0,137 ; 0,306)
(2,1)	0,093	0,041	(0,028 ; 0,186)	0,111	0,055	(0.029 ; 0.234)	0,073	0,028	(0,030 ; 0,139)
(2,2)	0,056	0,029	(0,012 ; 0,126)	0,066	0,031	(0.018 ; 0.136)	0,045	0,020	(0,014 ; 0,094)
(2,3)	0,115	0,045	(0,042 ; 0,215)	0,118	0,053	(0.038 ; 0.240)	0,092	0,030	(0,041 ; 0,158)
(3,1)	0,115	0,048	(0,036 ; 0,222)	0,113	0,056	(0.031 ; 0.239)	0,081	0,030	(0,033 ; 0,148)
(3,2)	0,044	0,028	(0,006 ; 0,113)	0,065	0,035	(0.013 ; 0.144)	0,043	0,020	(0,012 ; 0,086)
(3,3)	0,087	0,042	(0,022 ; 0,184)	0,107	0,055	(0.023 ; 0.227)	0,103	0,034	(0,047 ; 0,181)

Nota : Voir l'annexe B pour une description des modèles 1 et 2.

4.3 Facteur de Bayes pour la preuve d'association

Nous avons également considéré l'association entre la densité minérale osseuse et le revenu familial. Bien que l'existence d'une telle association paraisse peu probable, il est intéressant d'examiner cette question ; voir Nandram, Cox et Choi (2005) pour une discussion de ce problème. Nous nous servons du facteur de Bayes (Kass et Raftery 1995) pour mesurer la force de la preuve d'une association comparativement à l'absence d'association dans le tableau de contingence $r \times c$. Nous le faisons pour chacune des 13 régions et pour toutes les régions confondues.

Nous utilisons deux procédures, l'une sans modélisation étendue et l'autre s'appuyant sur notre modèle (étendu) de non-réponse non ignorable. La méthode simple consiste à produire les fréquences de cellule selon une technique de ratissage ordinaire, et nous supposons qu'il n'y a pas d'erreur à le faire. Il s'agit d'une procédure dictée par le bon sens que les praticiens des enquêtes utilisent régulièrement. Au moyen de la deuxième procédure fondée sur notre modèle de non-réponse non ignorable, nous avons obtenu 1 000 tableaux combinés pour chaque région, comme il est décrit à la section 3 sur les calculs. Pour chaque région, nous avons obtenu les fréquences de cellule pour les quatre tableaux, et nous les avons totalisées pour obtenir un seul tableau de toutes les fréquences.

Voici la description de la procédure de ratissage pour obtenir les fréquences de cellule. Soit n_{jk} les fréquences de cellule pour les quatre tableaux combinés. Soit $n_{jk}^{(1)}$ les fréquences de cellule pour le tableau contenant les données complètes, $n_{j,c+1}^{(2)}$ celles pour le tableau des totaux de ligne, $n_{r+1,k}^{(3)}$ celles pour le tableau des totaux de colonne et $n_{r+1,c+1}^{(4)}$ celles pour le tableau des totaux. Les fréquences de cellule pour les quatre tableaux sont estimées par

$$n_{jk} = n_{jk}^{(1)} + \left(\frac{n_{jk}^{(1)}}{n_{j\cdot}^{(1)}}\right) n_{j,c+1}^{(2)} + \left(\frac{n_{jk}^{(1)}}{n_{\cdot k}^{(1)}}\right) n_{r+1,k}^{(3)} + \left(\frac{n_{jk}^{(1)}}{n_{\cdot\cdot}^{(1)}}\right) n_{r+1,c+1}^{(4)},$$

$j = 1, \ldots, r, k = 1, \ldots, c.$

Dans chaque cas, nous désignons la somme des fréquences de cellule pour chaque région par n_{jk}. Pour la procédure de ratissage, nous n'avons qu'un seul tableau pour chaque région, tandis que pour le modèle de non-réponse

non ignorable, nous avons un échantillon de 1 000 tableaux pour chaque région. Nous avons aussi un seul tableau combiné pour toutes les régions sous la procédure de ratissage et 1 000 tableaux pour toutes les régions combinées. Nous obtenons le facteur de Bayes pour chaque tableau sous un modèle multinomial-Dirichlet. Il convient de souligner que notre méthode s'appuie sur le modèle d'extension de sorte que les fréquences de cellule sont produites en empruntant de l'information à d'autres régions contrairement à la procédure de ratissage.

Puis, pour chaque tableau, nous prenons

$$\boldsymbol{n}\,|\,\boldsymbol{\pi} \sim \text{Multinomiale}(n, \boldsymbol{\pi}) \text{ et } \boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{1}).$$

Autrement dit, nous prenons une loi a priori uniforme pour $\boldsymbol{\pi}$ avec $\pi_{jk} > 0$ et $\sum_{j=1}^{r}\sum_{k=1}^{c}\pi_{jk} = 1$. Sous l'hypothèse de l'absence d'association, nous avons $\pi_{jk} = \pi_j \pi_k$, où $\pi_j > 0, \sum_{j=1}^{r}\pi_j = 1$ et $\pi_k > 0, \sum_{k=1}^{c}\pi_k = 1$. Donc, l'hypothèse d'une association est que les π_{jk} ne sont soumises à aucune contrainte (à part le fait qu'elles sont non négatives et que leur somme est égale à l'unité), tandis que, pour l'hypothèse d'absence d'association, $\pi_{jk} = \pi_j \pi_k$.

Le facteur de Bayes est le ratio de la vraisemblance marginale sous association par rapport à la vraisemblance marginale en l'absence d'association. Ce ratio mesure la force de la preuve d'association par opposition à l'absence d'association ; voir Kass et Raftery (1995). Soit $p_a(\boldsymbol{n})$ la vraisemblance marginale sous association et $p_0(\boldsymbol{n})$ la vraisemblance marginale en l'absence d'association. Alors, en posant que $n_{j\cdot} = \sum_{k=1}^{c} n_{jk}$ et $n_{\cdot k} = \sum_{j=1}^{r} n_{jk}$, il est facile de montrer que

$$p_a(\boldsymbol{n}) = p_0(\boldsymbol{n})\left\{\prod_{u=0}^{n-1}\frac{u+rc}{(u+r)(u+c)}\right\}\frac{\prod_{j=1}^{r} n_{j\cdot}!\prod_{k=1}^{c} n_{\cdot k}!}{\prod_{j=1}^{r}\prod_{k=1}^{c} n_{jk}!},$$

où $p_0(\boldsymbol{n}) = n!\prod_{u=0}^{n-1}(u+rc)^{-1}$. Observons que $p_0(\boldsymbol{n})$ n'est pas une fonction de $\{n_{jk}\}$. Donc, en tant que mesure d'association, c'est l'écart de $\prod_{j=1}^{r} n_{j\cdot}!\prod_{k=1}^{c} n_{\cdot k}!$ par rapport à $\prod_{j=1}^{r}\prod_{k=1}^{c} n_{jk}!$ qui importe. Par contre, nous notons que, pour la statistique d'indépendance de Pearson classique, ce sont les écarts de n_{jk} par rapport à $n_{j\cdot}n_{\cdot k}$ qui importent. Cependant, soulignons que ce test ne peut pas être appliqué, parce que bon nombre de fréquences de cellule prévues sont inférieures à 5 sous l'hypothèse de l'absence d'association et l'échantillonnage multinomial

Nous présentons nos résultats au tableau 4 et à la figure 1 qui correspond aux données du tableau 1 pour la classification croisée de la densité minérale osseuse et du revenu familial. Nous présentons les logarithmes des vraisemblances marginales (base *e*) et des facteurs de Bayes ; ceux-ci doivent être interprétés en appliquant la règle empirique de Kass et Raftery (1995).

À la figure 1, nous voyons que les boîtes à moustaches se trouvent toutes au-dessus de zéro, sauf celle pour la troisième région qui ne donne aucune preuve d'association ; il n'existe peut-être pas non plus de preuve d'association pour la région 42 (10 dans la figure). Un résumé de ces résultats est présenté au tableau 4. Les facteurs de Bayes indiquent une association dans toutes les régions, sauf la région 12, et leur valeur réelle est nettement plus grande sous le modèle de non-réponse non ignorable. Les valeurs pour la région 6 et pour toutes les régions confondues sont élevées (336,3 c. 5,8 et 3 798,2 c. 0,183).

Tableau 4
Données de la NHANES pour 13 régions. Comparaison des vraisemblances marginales négatives et des facteurs de Bayes ou de l'association de la DMO et du RF d'après la procédure de ratissage et le modèle d'extension, selon la région

	Ratissage			Extension	
Région	$-\ln\{p_0(n)\}$	$-\ln\{p_a(n)\}$	***FB***	$-\ln\{p_a(n)\}$	***FB***
4	26,19	23,07	22,855	$23,5_{0,014}$	$14,78_{0,169}$
6	45,73	43,98	5,766	$40,5_{0,038}$	$336,27_{11,465}$
12	31,14	38,01	0,001	$33,4_{0,054}$	$0,37_{0,027}$
17	29,13	27,03	8,134	$27,0_{0,026}$	$10,27_{0,191}$
25	25,44	26,02	0,558	$23,8_{0,029}$	$9,55_{0,202}$
26	26,89	23,18	40,562	$23,9_{0,018}$	$24,71_{0,370}$
29	23,21	20,87	10,301	$21,3_{0,018}$	$8,40_{0,115}$
36	34,99	36,09	0,330	$33,1_{0,064}$	$21,13_{0,928}$
39	23,77	24,89	0,325	$23,6_{0,044}$	$2,24_{0,68}$
42	29,51	30,21	0,497	$30,3_{0,099}$	$4,33_{0,255}$
44	25,61	30,48	0,008	$24,4_{0,027}$	$5,19_{0,137}$
48	38,83	35,34	32,650	$39,1_{0,060}$	$2,15_{0,081}$
53	27,11	24,82	9,865	$24,2_{0,017}$	$19,40_{0,282}$
All	53,43	55,13	0,183	$46,1_{0,049}$	$3,798,24_{151,82}$

Nota : Dans la colonne des régions, « Toutes » désigne toutes les régions confondues ; la notation a_b signifie que la moyenne est *a* et que l'erreur-type est *b* sur les 1 000 itérations. Le terme $\ln\{p_0(n)\}$ est le même pour les deux procédures.

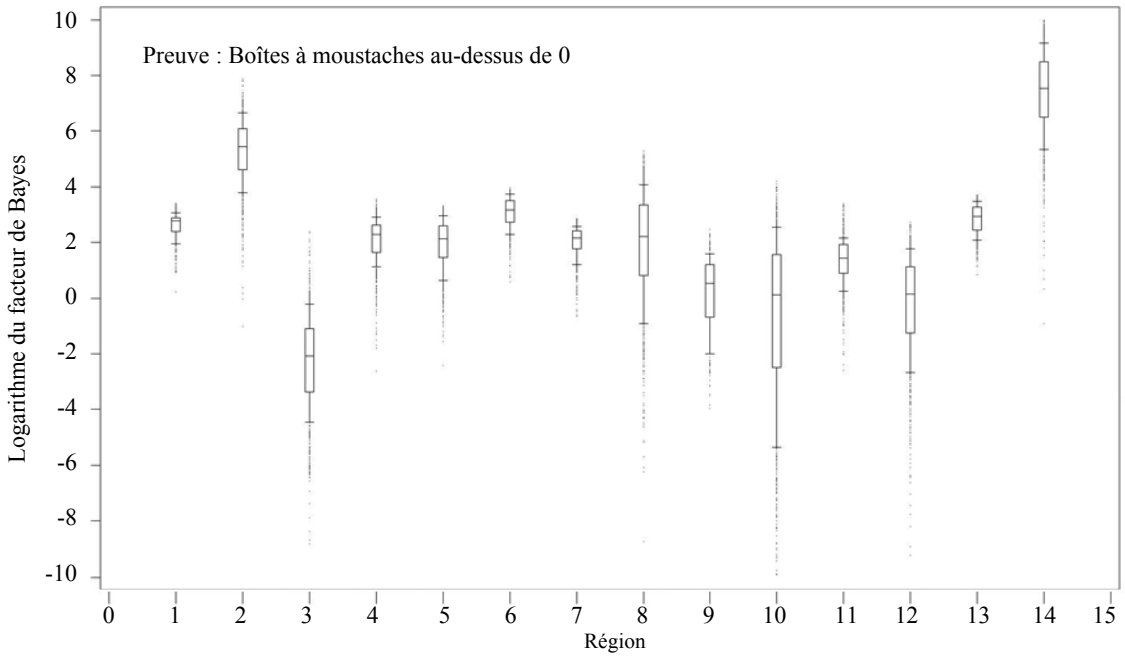

Figure 1 Boîtes à moustaches des logarithmes du facteur de Bayes selon la région pour évaluer la preuve de l'association entre la DMO et le RF

5. Conclusion

L'objectif du présent article était d'élaborer une méthodologie pour analyser les données provenant de tableaux incomplets de contingence à deux entrées, chaque tableau correspondant à un domaine ou région. Nous avons pour cela étendu la méthodologie bayésienne de Nandram et Choi (2002a, b) pour données binaires à des tableaux de contingences $r \times c$ pour petits domaines. Nous avons construit un nouveau modèle de non-réponse non ignorable bayésien (c'est-à-dire le modèle d'extension) qui est centré sur le modèle de non-réponse ignorable. Nous avons utilisé des méthodes Monte Carlo par chaîne de Markov (spécifiquement l'échantillonneur Metropolis-Hastings « griddy ») pour ajuster le modèle. Nous avons comparé notre modèle à un modèle de non-réponse ignorable et un modèle de non-réponse non ignorable. Enfin, nous avons illustré notre méthode en estimant les probabilités de cellule pour le tableau de contingence 3×3 de la densité minérale osseuse et du revenu sur 13 régions infranationales.

Nous avons montré qu'il existe des différences entre les trois modèles. En utilisant les données sur la densité minérale osseuse et le revenu familial, nous avons montré que notre modèle d'extension est un compromis entre le modèle de non-réponse ignorable et le modèle de non-réponse non ignorable. À l'aide du facteur de Bayes, nous avons montré qu'il existe des différences entre les tests de l'association de la densité minérale osseuse et du revenu familial quand les fréquences de cellule sont estimées au moyen de notre modèle et en utilisant une procédure de ratissage. En fait grâce à l'emprunt d'information, nous constatons que la preuve de l'association est nettement plus forte sous notre modèle que sous la procédure de ratissage.

Trois pistes supplémentaires pourraient être explorées. Premièrement, nous pouvons construire un modèle en vue d'intégrer l'écart systématique par rapport à l'ignorabilité. Cette tâche suscitera du travail sur le terrain coûteux supplémentaire pour obtenir l'information nécessaire. Deuxièmement, il serait également intéressant de relâcher l'hypothèse voulant que les marges du tableau de contingence soient fixes ; voir, par exemple, Nandram (2009) qui a procédé à cet examen pour une seule grande région. Troisièmement, le calage pourrait être encore amélioré (c'est-à-dire en intégrant de l'information sur les marges).

Remerciements

Nous exprimons notre reconnaissance à l'égard de deux examinateurs et d'un rédacteur associé pour l'aide apportée dans la présentation du document.

Annexe A

Densité a posteriori conjointe du modèle d'extension

Premièrement, en intégrant la fonction de densité a posteriori conjointe sur $\boldsymbol{p}$, nous obtenons

$$h(\Omega,\boldsymbol{\pi},\boldsymbol{\phi},\boldsymbol{y}_{(1)}|\boldsymbol{y}_1,\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) \propto \prod_{i=1}^{4}\frac{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\boldsymbol{\mu}_1\tau_1)}{\mathrm{D}(\boldsymbol{\mu}_1\tau_1)}$$
$$\left[\binom{n_i}{\boldsymbol{y}'_{i1},\,\boldsymbol{y}'_{i2},\,\boldsymbol{y}'_{i3},\boldsymbol{y}'_{i4}}(1-\pi_{i1})^{-y_{i\ldots}}\right.$$
$$\times\left\{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot\cdot}}\prod_{j,k}\phi_{ijk}^{y_{i\cdot jk}}[a_{ijk}(\pi_{i1},\phi_{ijk})]^{y_{i1jk}}\right\}$$
$$\times\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{\mu_{2s}\tau_2-1}}{\mathrm{D}(\boldsymbol{\mu}_2\tau_2)}\prod_{j=1}^{r}\prod_{k=1}^{c}\frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}}\right.$$
$$\left.\left.\frac{\phi_{ijk}^{\beta-1}e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)G_{\beta}[\beta(1-\pi_{i1})^{-1}]}\right\}\right]$$
$$\times\frac{a_0}{(a_0+\tau_1)^2}\cdot\frac{b_0}{(b_0+\tau_2)^2}\cdot\frac{c_0}{(c_0+\beta)^2}, \qquad \text{(A.1)}$$

où le vecteur de dimensions $rc\times 1$ est

$$\boldsymbol{y}_i^{(1)} \overset{\text{def}}{=} (y_{i.11}, y_{i.12}, \ldots, y_{i.1c}, y_{i.21}, \ldots, y_{i.2c}, y_{i.r1}, \ldots, y_{i.rc})'.$$

Maintenant, désignons par $\boldsymbol{j}$ un vecteur de dimensions $rc\times 1$ de valeurs 1 et soit

$$\boldsymbol{y}_i^{(2)} \overset{\text{def}}{=} (y_{i1..}, y_{i2..}, y_{i3..}, y_{i4..})'.$$

Alors, en agrégeant sur $\boldsymbol{\pi}$ et $\boldsymbol{\phi}$, nous obtenons

$$h(\Omega,\boldsymbol{y}_{(1)}|\boldsymbol{y}_1,\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) \propto \frac{a_0}{(a_0+\tau_1)^2}\cdot\frac{b_0}{(b_0+\tau_2)^2}\cdot\frac{c_0}{(c_0+\beta)^2}$$
$$\times\prod_{i=1}^{A}\binom{n_i}{\boldsymbol{y}'_{i1},\,\boldsymbol{y}'_{i2},\,\boldsymbol{y}'_{i3},\boldsymbol{y}'_{i4}}\frac{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\boldsymbol{\mu}_1\tau_1)}{\mathrm{D}(\boldsymbol{\mu}_1\tau_1)}\,\frac{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}{\mathrm{D}(\boldsymbol{\mu}_2\tau_2)}$$
$$\times\prod_{i=1}^{A}\left[\frac{\beta^{\beta}}{\Gamma(\beta)}\right]^{rc}\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\,\boldsymbol{j})\,I_i, \qquad \text{(A.2)}$$

où

$$I_i=\iiint\prod_{j,k}\left\{\frac{\left[\left(\frac{1-\pi_{i1}}{\phi_{ijk}}\right)\left(1+\frac{1}{\pi_{i1}}\{1-\pi_{i1}-\phi_{ijk}\}\right)\right]^{y_{i1jk}}}{(1-\pi_{i1})^{y_{i\cdot jk}+\beta}\,G_{\beta}\left(\frac{\beta}{1-\pi_{i1}}\right)}\right\}$$
$$\times\left\{\prod_{j,k}e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}\right\}\left\{\frac{\prod_{j,k}\phi_{ijk}^{y_{i\cdot jk}+\beta-1}}{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\,\boldsymbol{j})}\right\}$$
$$\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot\cdot}+\mu_{2s}\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}\right\}\mathrm{d}\boldsymbol{\phi}_i\,\mathrm{d}\,\boldsymbol{\pi}_i. \qquad \text{(A.3)}$$

Notons que $0\le\pi_{is}\le 1, \sum_{s=1}^{4}\pi_{is}=1$ et $0\le\phi_{ijk}\le 1$. Nous simplifions le calcul pour I_i dans (A.3) en deux étapes.

Premièrement, dans (A.3), nous effectuons la transformation

$$\phi_{ijk}=T_i\,\phi_{ijk}^{*} \qquad \sum_{j=1}^{r}\sum_{k=1}^{c}\phi_{ijk}=T_i.$$

Les nouvelles variables ϕ_{ijk}^{*} satisfont les relations $0\le\phi_{ijk}^{*}\le 1, \sum_{j=1}^{r}\sum_{k=1}^{c}\phi_{ijk}^{*}=1$ et les T_i sont soumis aux contraintes $0\le T_i\le 1/\phi_{ijk}^{*}$, pour $j=1,\ldots,r, k=1,\ldots,c$ et $0\le T_i\le rc$. Suite à cette transformation, nous avons

$$I_i=\iiint_0^{b_i}\prod_{j,k}\left\{\left[\frac{T_i}{1-\pi_{i1}}\right]^{\sum_{s=2}^{4}y_{isjk}}\right.$$
$$\left.\left[\frac{1}{\phi_{ijk}^{*}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{T_i}{1-\pi_{i1}}\phi_{ijk}^{*}\right\}\right)\right]^{y_{i1jk}}\right\}$$
$$\times\left\{\frac{\left(\frac{T_i}{1-\pi_{i1}}\right)^{rc\beta-1}e^{-\frac{\beta T_i}{1-\pi_{i1}}}}{(1-\pi_{i1})\left[G_{\beta}\left(\frac{\beta}{1-\pi_{i1}}\right)\right]^{rc}}\right\}\left\{\frac{\prod_{j,k}\phi_{ijk}^{*\,y_{i\cdot jk}+\beta-1}}{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\,\boldsymbol{j})}\right\}$$
$$\times\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot\cdot}+\mu_{2s}\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}\right\}\mathrm{d}T_i\,\mathrm{d}\boldsymbol{\phi}_i^{*}\mathrm{d}\boldsymbol{\pi}_i,$$

où $b_i=\min\{\{1/\phi_{ijk}^{*}\}\,j=1,\ldots,r\,;\,k=1,\ldots,c\}$.

Deuxièmement, en posant que $W_i=\{\beta T_i/1-\pi_{i1}\}$ et en absorbant le facteur $\beta^{rc\beta}/\Gamma(rc\beta)$ dans I_i, avec certaines opérations algébriques supplémentaires, nous obtenons

$$I_i=\iiint_0^{\frac{\beta b_i}{1-\boldsymbol{\pi}_{i1}}}\frac{G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)}{\left[G_{\beta}\left(\frac{\beta}{1-\pi_{i1}}\right)\right]^{rc}}$$
$$\prod_{j,k}\left\{\left(\frac{W_i}{\beta}\right)^{\sum_{s=2}^{4}y_{isjk}}\left[\frac{1}{\phi_{ijk}^{*}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{W_i\phi_{ijk}^{*}}{\beta}\right\}\right)\right]^{y_{i1jk}}\right\}$$
$$\times\frac{W_i^{rc\beta-1}e^{-W_i}}{\Gamma(rc\beta)G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)}$$
$$\mathrm{d}W_i\left\{\frac{\prod_{j,k}\phi_{ijk}^{*\,y_{i\cdot jk}+\beta-1}}{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\,\boldsymbol{j})}\right\}\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{y_{is\cdot\cdot}+\mu_2 s\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}\right\}\mathrm{d}\boldsymbol{\phi}_i^{*}\,\mathrm{d}\boldsymbol{\pi}_i. \qquad \text{(A.4)}$$

Annexe B

Modèles de non-réponse ignorable et non ignorable

Posons que $\psi_{ijk} \equiv 1$ dans le modèle d'extension pour former le modèle de non-réponse ignorable. Pour $i = 1, \ldots, A$, nous prenons alors

$$\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2 \tau_2)$$

et indépendamment

$$\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1 \tau_1).$$

En outre, $p(\tau_2) = \{1/(1+\tau_1)^2\}, \tau_1 \geq 0, \mu_1 \sim \text{Dirichlet}(\mathbf{1})$, $p(\tau_2) = \{1/(1+\tau_2)^2\}, \tau_1 \geq 0$ et $\mu_2 \sim \text{Dirichlet}(\mathbf{1})$. Ici, nous avons l'indépendance à tous les niveaux et les vecteurs **1** sont de dimension appropriée, chaque coordonnée étant égale à l'unité. Notons que tous les paramètres du modèle ignorable doivent être identifiés et estimés.

Soit $\pi_{isjk} = \pi_{is}\psi_{ijk}$ dans le modèle d'extension pour former le modèle de non-réponse non ignorable. Dans ce cas, pour $i = 1, \ldots, A$,

$$\boldsymbol{\pi}_{ijk} \mid \boldsymbol{\mu}_2, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2 \tau_2)$$

et indépendamment

$$\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1 \tau_1).$$

Dans ce modèle, les paramètres $\boldsymbol{\pi}_{ijk}$ ne sont pas identifiables et nous prenons $\tau_2 \sim \text{Gamma}(\alpha_0, \beta_0)$, où α_0 et β_0 doivent être spécifiés. La spécification du modèle est alors achevée en attribuant à $\tau_1, \boldsymbol{\mu}_1$ et à $\boldsymbol{\mu}_2$ les mêmes propriétés distributionnelles qu'au paragraphe précédent.

Comme dans Nandram, Cox et Choi (2005), α_0 et β_0 sont spécifiés comme il suit. Le modèle de non-réponse ignorable et ajusté de manière à obtenir un échantillon de la densité de probabilité a posteriori de τ_2. Puis, α_0 et β_0 sont obtenus en utilisant la méthode des moments. Nandram, Cox et Choi (2005) ont constaté que l'inférence au sujet de $\boldsymbol{p}_i$ n'est pas très sensible au choix de ces paramètres.

Bibliographie

Cohen, G., et Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics*, 18, 13-23.

Draper, D. (1995). Assessment and propagation of model uncertainty (avec discussion). *Journal of the Royal Statistical Society*, Séries B, 57, 45-97.

Forster, J.J., et Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society*, Séries B, 60, 57-70.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Greenland, S. (2009). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Sciences*, 24, 195-210.

Kass, R.E., et Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, deuxième édition. New York : John Wiley & Sons, Inc.

Nandram, B. (2009). Bayesian inference of the cell probabilities of a two-way categorical table under non-ignorability. *Communications in Statistics - Theory and Methods*, 38, 3015-3030.

Nandram, B., et Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., et Choi, J.W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., et Choi, J.W. (2004). A nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse. *Journal of Nonparametric Statistics*, 16, 821-839.

Nandram, B., et Choi, J.W. (2005). Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines : une application aux données de la NHANES. *Techniques d'enquête*, 31, 79-92.

Nandram, B., et Choi, J.W. (2008). Une répartition bayésienne des électeurs indécis. *Techniques d'enquête*, 34, 41-54.

Nandram, B., et Choi, J.W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, 105, 120-135.

Nandram, B., Cox, L.H. et Choi, J.W. (2005). Analyse bayésienne des données catégoriques manquantes non ignorables : une application à la densité minérale osseuse et au revenu familial. *Techniques d'enquête*, 31, 233-247.

Nandram, B., Han, G. et Choi, J.W. (2002). Un modèle bayesien hiérarchique de non-réponse non-ignorable pour les données multinomiales des petites régions. *Techniques d'enquête*, 28, 157-170.

Nandram, B., Liu, N., Choi, J.W. et Cox, L. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.

National Center for Health Statistics (1992). Third national health and nutrition examination survey. *Vital and Health Statistics*, Series 2, 113.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Smith, A.F.M., et Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46, 84-88.