Techniques d'enquête, décembre 2004
Vol. 30, N° 2, pp. 233-243
Statistique Canada, N° 12-001-XPB au catalogue

Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés

Hui Zheng et Roderick J.A. Little [1]

Résumé

Les échantillonneurs se méfient souvent des approches d'inférence fondées sur un modèle, parce qu'ils craignent que soient commises des erreurs de spécification lorsque les modèles sont appliqués à de grands échantillons provenant de populations complexes. Nous soutenons que le paradigme de l'inférence fondée sur un modèle peut donner de très bons résultats dans les conditions d'enquête, à condition que les modèles soient choisis de façon à tenir compte du plan d'échantillonnage et d'éviter de faire des hypothèses fortes au sujet des paramètres. L'estimateur d'Horvitz-Thompson (HT) est un estimateur simple sans biais par rapport au plan de sondage du total de population finie. Du point de vue de la modélisation, l'estimateur HT fonctionne bien lorsque les ratios des valeurs de la variable de résultat et des probabilités d'inclusion sont interchangeables. Si cette hypothèse n'est pas satisfaite, l'estimateur HT peut être très inefficace. Dans Zheng et Little (2003, 2004), nous avons utilisé des fonctions splines pénalisées (*p*-splines) pour modéliser des relations à variation lisse entre le résultat et les probabilités d'inclusion sous échantillonnage à un degré avec probabilité proportionnelle à la taille (PPT). Nous avons montré que les estimateurs fondés sur un modèle à fonction *p*-spline sont généralement plus efficaces que l'estimateur HT et peuvent fournir des intervalles de confiance plus étroits dont la couverture est proche du niveau de confiance nominal. Dans le présent article, nous étendons cette approche aux plans d'échantillonnage à deux degrés. Nous utilisons un modèle mixte fondé sur une *p*-spline qui est ajusté à une relation non paramétrique entre les moyennes des unités primaires d'échantillonnage (UPE) et une mesure de la taille des UPE, et auquel sont intégrés des effets aléatoires pour modéliser la mise en grappes. Pour l'estimation de la variance, nous considérons les méthodes d'estimation de la variance fondées sur un modèle bayésien empirique, la méthode du jackknife et la méthode des répliques répétées équilibrées (BRR). Des études en simulation portant sur des données simulées et des échantillons tirés des microdonnées à grande diffusion du Recensement de 1990 montrent que l'estimateur fondé sur un modèle à fonction *p*-spline donne de meilleurs résultats que l'estimateur HT et que les estimateurs linéaires assistés par un modèle. Les simulations montrent aussi que les méthodes d'estimation de la variance produisent des intervalles de confiance dont la couverture est satisfaisante. Fait intéressant, ces progrès peuvent être observés pour un plan d'échantillonnage courant à probabilités de sélection égales, où la sélection à la première étape est PPT et où les probabilités de sélection à la deuxième étape sont proportionnelles à l'inverse des probabilités de sélection à la première étape, et où l'estimateur HT mène à la moyenne non pondérée. Dans les situations favorisant le plus l'utilisation de l'estimateur HT, les estimateurs fondés sur un modèle ont une efficacité comparable.

Mots clés : Pondération; maximum de vraisemblance restreint (REML); estimation bayésienne empirique.

1. Introduction

Dans une enquête par sondage, soit y_i la valeur d'un résultat Y pour l'unité i, et soit S l'ensemble d'unités échantillonnées. L'estimateur d'Horvitz-Thompson (HT) (Horvitz et Thompson 1952) $\hat{Y}_{\mathrm{HT}} = \sum_{i \in S} y_i / \pi_i$, où π_i est la probabilité de sélection de l'unité i, est un estimateur sans biais par rapport au plan de sondage du total de population finie (et de la moyenne lorsqu'on le divise par le chiffre de population connu N). Il peut aussi être considéré comme un estimateur fondé sur un modèle projectif (Firth et Bennett 1998) du modèle linéaire suivant reliant y_i à π_i :

$$y_i = \beta \pi_i + \pi_i \varepsilon_i ,$$

où l'on suppose que les ε_i sont i.i.d. et suivent une loi normale de moyenne nulle et de variance σ^2.

Dans Zheng et Little (2003, 2004), nous avons proposé un modèle non paramétrique

$$y_i = f(\pi_i) + \varepsilon_i , \varepsilon_i \sim \text{ind } N\left(0, \pi_i^{2k} \sigma^2\right),$$

utilisant des splines pénalisées pour modéliser la moyenne du résultat y_i sous forme d'une fonction à variation lisse f des probabilités de sélection π_i. Nous avons montré dans Zheng et Little (2003) que les estimateurs fondés sur un modèle non paramétrique sont plus efficaces que l'estimateur HT pour les échantillons généraux à un degré sélectionnés avec probabilité proportionnelle à la taille (PPT) et ne sont pas nettement moins efficaces que l'estimateur HT lorsque les données sont générées à l'aide d'un modèle favorisant l'estimateur HT.

1. Hui Zheng, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115. Courriel : zheng@hcp.med.harvard.edu; Roderick J.A. Little, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109. Courriel : rlittle@umich.edu.

Dans le présent article, nous considérons l'échantillonnage à deux degrés. À la première étape, un sous-ensemble de m unités primaires d'échantillonnage (UPE) est tiré d'une population contenant H UPE, avec probabilités inégales $\pi_{1,h}$, $h = 1, ..., H$. Numérotons les UPE incluses dans l'échantillon de 1 à m. À la deuxième étape, un échantillon aléatoire simple (eas) de n_h parmi N_h unités secondaires d'échantillonnage (USE) est tiré à partir des UPE échantillonnées étiquetées h avec la probabilité de sélection $\pi_{2,h}$. La probabilité globale de sélection de l'unité i dans l'UPE h est $\pi_h = \pi_{1,h}\pi_{2,h}$, et l'estimateur HT de la moyenne d'un résultat Y est $\bar{y}_w = \sum_{h=1}^{m}\sum_{i=1}^{n_h} y_{hi} / (\pi_{1,h}\pi_{2,h}) / N$, où y_{hi} est la valeur de Y pour l'unité i dans l'UPE h et N est le nombre total connu d'unités (USE) dans l'ensemble de la population. Dans un plan d'échantillonnage adopté couramment, la probabilité de sélection au premier degré est proportionnelle à une estimation de la taille de l'UPE et les probabilités de sélection au deuxième degré sont proportionnelles à l'inverse des probabilités de sélection de premier degré, si bien que les probabilités globales de sélection π_h sont égales pour toutes les USE. Dans ce cas, la moyenne pondérée par l'inverse des probabilités est égale à la moyenne simple d'échantillon $\bar{y} = \sum_{h=1}^{m}\sum_{i=1}^{n_h} y_{hi} / \sum_{h=1}^{m} n_h$.

Nous supposons dans tout l'article que les probabilités de sélection $\pi_{1,h}$ sont connues pour toutes les UPE $h = 1, \ldots, H$. Aux sections 2 et 3, nous supposons que les nombres N_h d'USE sont également connus pour toutes les UPE comprises dans la population. À la section 4, nous examinons la situation courante où N_h est connu uniquement pour les UPE échantillonnées, mais où le N_h pour les UPE non échantillonnées peut être estimé au moyen d'un modèle de régression fondé sur des variables auxiliaires connues pour toutes les UPE de la population.

Särndal, Swensson et Wretman (1992) ont discuté d'estimateurs assistés par modèle pour remplacer l'estimateur HT pour les échantillons à deux degrés avec information auxiliaire disponible au niveau de l'UPE ou de l'USE. Dans le premier cas, soit x_h un vecteur de variables auxiliaires au niveau de l'UPE pour l'UPE h. Nous supposons que les totaux d'UPE $t_h = \sum_{i=1}^{N_h} y_{hi}$ sont reliés à x_h selon un modèle linéaire :

$$E(t_h \mid x_h) = x_h^T \beta, \mathrm{Var}(t_h) = \sigma_h^2, \ h = 1, ..., H$$

(Särndal et coll. 1992). Nous estimons β par la régression pondérée par les probabilités

$$\hat{B} = \left(\sum_{h=1}^{m} x_h x_h^T / \left(\sigma_h^2 \pi_{1,h}\right)\right)^{-1} \sum_{h=1}^{m} x_h t_h^* / \left(\sigma_h^2 \pi_{1,h}\right),$$

où $t_h^* = \sum_{i=1}^{n_h} y_{hi} / \pi_{2,h}$, qui mène aux totaux projetés $\hat{t}_h = x_h^T \hat{B}$, $h = 1, \ldots, H$. En pratique, qu'elles soient obtenues simplement par hypothèse (par exemple, σ_h proportionnel à une mesure de taille de la strate h) ou estimées, les estimations $\hat{\sigma}_h^2$ remplacent σ_h^2 dans la formule qui précède. L'estimateur par la régression généralisée (GR) du total général est

$$\hat{T}_A = \sum_{i=1}^{H} \hat{t}_h + \sum_{h=1}^{m} \frac{(t_h^* - \hat{t}_h)}{\pi_{1,h}},$$

et l'estimation de la moyenne est $\hat{T}_A / N$. Le terme $\sum_{h=1}^{m} (t_h^* - \hat{t}_h) / \pi_{1,h}$ est un terme de calage du biais qui rend l'estimateur convergent par rapport au plan de sondage.

Dans le second cas, où l'information auxiliaire est connue au niveau des USE, soit x_{hi} l'ensemble de variables auxiliaires pour l'USE i dans l'UPE h, $h = 1, \ldots, H; i = 1, \ldots, N_h$. La relation entre le résultat et l'information auxiliaire est modélisée par

$$E(y_{hi} \mid x_{hi}) = x_{hi}^T \beta, \ldots, \mathrm{Var}(y_{hi}) = \sigma_{hi}^2, \ h = 1, ..., H, \ i = 1, ..., N_h.$$

L'estimation de β par la régression pondérée par les probabilités est

$$\hat{B} = \left(\sum_{h=1}^{m}\sum_{i=1}^{n_h} x_{hi} x_{hi}^T / \left(\sigma_{hi}^2 \pi_{hi}\right)\right)^{-1} \sum_{h=1}^{m}\sum_{i=1}^{n_h} x_{hi} \, y_{hi} / \left(\sigma_h^2 \pi_{hi}\right),$$

où π_{hi} est la probabilité que l'unité (h, i) soit incluse dans l'échantillon. L'estimateur GR pour le total général est

$$\hat{T}_B = \sum_{h=1}^{H}\sum_{i=1}^{N_h} \hat{y}_{hi} + \sum_{h=1}^{m}\sum_{i=1}^{n_h} \frac{(y_{hi} - \hat{y}_{hi})}{\pi_{hi}},$$

où $\hat{y}_{hi} = x_{hi}^T \hat{B}$. L'estimateur de la moyenne est $\hat{T}_B / N$.

Ces deux méthodes ne tiennent compte ni l'une ni l'autre des corrélations des résultats à l'intérieur des UPE. Ces corrélations peuvent être modélisées en traitant les moyennes d'UPE comme des effets aléatoires dans un modèle hiérarchique. Dans le cas où l'information au niveau de l'UPE x_h est disponible pour toutes les UPE, un modèle de ce genre est :

$$y_{hi} \mid \mu_h \overset{\text{ind}}{\sim} N(\mu_h, \ \sigma^2)$$

$$\mu \sim N_H(\varphi, \ D) \qquad (1)$$

où $\mu = (\mu_1, ..., \mu_H)^T$, $\varphi = (\varphi_1, ..., \varphi_H)^T$ où μ_h est le résultat moyen dans l'UPE h, $\varphi_h = x_h^T \beta$ et D est la matrice de covariance des moyennes d'UPE. L'estimateur fondé sur un modèle de $\bar{Y}$ est donné par

$$\hat{E}(\bar{Y} \mid \mathbf{y}, x_h) = \frac{1}{N}\left(\sum_{h=1}^{m} [n_h \bar{y}_h + (N_h - n_h)\, \hat{\mu}_h] + \sum_{h=m+1}^{H} N_h \hat{\mu}_h\right),$$

où $\hat{\mu}_h = \hat{E}(y_{hi} \mid \mathbf{y}, x_h)$, et $\mathbf{y}$ est le vecteur des résultats dans l'échantillon.

Diverses hypothèses au sujet de φ et D dans (1) mènent aux modèles suivants :

Effets aléatoires permutables (XRE) : (Holt et Smith 1979; Ghosh et Meeden 1986; Little 1991; Lazzaroni et Little 1998) : $\varphi_h \equiv \mu_o, h = 1, ..., H$ et $D = \tau^2 I_H$;

Autorégressif (AR1) : (Lazzaroni et Little 1998) : $\varphi_h \equiv \mu_o, h = 1, ..., H$ et $D = r^2\{\rho^{|i-j|}\}$;

Linéaire (LIN) : (Lazzaroni et Little 1998) : $\varphi_h = \alpha + \beta x_h, h = 1, ..., H$ et $D = \tau^2 I_H$;

Non paramétrique : (Elliott et Little 2000) : $\varphi_h = f(x_h), h = 1, ..., H$ et $D = 0$.

Les modèles non paramétriques d'Elliott et Little (2000) reposent sur l'hypothèse d'une fonction non paramétrique de moyenne reliant le résultat aux variables de plan de sondage. En supposant que $D = 0$, les moyennes d'UPE sont modélisées de sorte qu'elles soient égales à la fonction de moyenne f au lieu de fluctuer autour de celle-ci. Les modèles non paramétriques mixtes, dans lesquels les hypothèses sur D (par exemple, $D = \tau^2 I_H$) sont relâchées, représentent une extension naturelle du modèle d'Elliott et Little (2000) et des modèles linéaires mixtes avec structure paramétrique de la moyenne.

Il mérite d'être souligné que certains estimateurs appartenant à la famille de modèles susmentionnée correspondent aux estimateurs fondés sur le plan de sondage types. Par exemple, dans un plan de sondage avec probabilités égales où les n_h sont approximativement constants d'une UPE à l'autre, la moyenne non pondérée correspond à l'estimateur fondé sur un modèle spécial supposant que φ_h est constant.

2. Estimation pour le modèle mixte à fonction *P*-spline

La structure linéaire de φ dans le modèle LIN peut donner lieu à des erreurs de spécification si la structure réelle de la moyenne est non linéaire. Le problème de non-linéarité peut être résolu partiellement en ajoutant des termes polynomiaux (par exemple, termes quadratiques ou cubiques) à la partie des effets fixes du modèle LIN. Les modèles non paramétriques mixtes à fonction *p*-spline (Lin et Zhang 1999; Brumback, Ruppert et Wand 1999; Coull, Schwartz et Wand 2001) sont encore plus souples, puisqu'ils remplacent les termes polynomiaux par des fonctions non paramétriques lisses. Nous proposons le modèle non paramétrique mixte à fonction *p*-spline qui suit pour l'inférence au sujet de la moyenne de population :

Modèle non paramétrique mixte à fonction *P*-spline (MMP) :

$$\varphi_h = f(x_h), h = 1, ..., H, \; D = \tau^2 I_H,$$

où f est une fonction spline non paramétrique de degré p :

$$f(x;\beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j + \sum_{l=1}^{K} \beta_{l+p} (x - \kappa_l)_+^p,$$

où $\kappa_1 < ... < \kappa_K$ sont K nœuds fixes, $\beta_0, ..., \beta_{p+K}$ sont les coefficients à estimer et $(x)_+^p = x^p \mathbf{I}(x \geq 0)$.

Un moyen naïf d'estimer $\beta_0,...,\beta_{p+K}$ consiste à les traiter comme s'ils étaient fixes et à les estimer en même temps que les composantes de la variance σ^2 et τ^2 en ajustant un modèle linéaire mixte. Cependant, cette méthode peut produire des estimations de f présentant trop de rugosité et de variabilité. Pour éviter de surajuster le modèle, on peut pénaliser la rugosité de l'estimation $\hat{f}$ en ajoutant un terme de pénalité à la somme des carrés des écarts, de sorte que la solution $\hat{\beta}_0, ..., \hat{\beta}_p$ minimise

$$\sum_{h=1}^{m} (\hat{f}(x_h) - \hat{\mu}_h)^2 + \alpha \sum_{l=1}^{K} \beta_{l+p}^2 .$$

Nous réalisons ceci dans le contexte du modèle en attribuant à $\beta_0, ..., \beta_p$ des lois a priori uniformes, à $(\beta_{p+1}, ..., \beta_{p+K})$ une loi a priori normale $N_m(0, \sigma_\beta^2)$, et en posant que $\alpha = \tau^2 / \sigma_\beta^2$. Nous obtenons ainsi un modèle à spline pénalisée (*p*-spline).

Si $p = 1$, $\hat{f}$ est linéaire par morceau et nous estimons les coefficients $\beta_0, ..., \beta_{K+1}$, σ^2, σ_β^2 et τ^2 en ajustant le modèle linéaire mixte :

$$y = X_1 \beta + X_2 u + \varepsilon, \tag{2}$$

où $y = (y_{11}, y_{12}, ..., y_{mn_m})^T$, $\beta = (\beta_0, \beta_1)^T$, $u = (\beta_2, ..., \beta_{K+1}, u_1, ..., u_m)^T$,

$$X_1 = \begin{bmatrix} 1 & x_1 \\ 1 & x_1 \\ . & . \\ . & x_1 \\ . & x_2 \\ . & . \\ . & . \\ . & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_m \end{bmatrix},$$

$$X_2 = \begin{bmatrix} (x_1-\kappa_1)_+ & \dots & (x_1-\kappa_K)_+ & 1 & 0 & . & 0 \\ . & \dots & . & . & . & . & . \\ . & \dots & . & . & . & . & . \\ (x_1-\kappa_1)_+ & \dots & (x_1-\kappa_K)_+ & 1 & 0 & . & . \\ (x_2-\kappa_1)_+ & \dots & (x_2-\kappa_K)_+ & 0 & 1 & . & . \\ . & \dots & . & . & . & . & . \\ . & \dots & . & . & . & . & . \\ (x_2-\kappa_1)_+ & \dots & (x_2-\kappa_K)_+ & 0 & 1 & . & 0 \\ . & \dots & . & 0 & 0 & . & 1 \\ . & \dots & . & . & . & . & . \\ . & \dots & . & . & . & . & . \\ (x_m-\kappa_1)_+ & \dots & (x_m-\kappa_K)_+ & 0 & 0 & . & 1 \end{bmatrix},$$

où x_h dans X_1 et $(x_h-\kappa_l)_+$ dans X_2 sont tous deux répétés n_h fois. Les termes aléatoires u et ε sont mutuellement indépendants avec

$$u = (\beta_2, \dots, \beta_{K+1}, u_1, \dots, u_m)^T \sim N_{K+m}(0, G),$$

$$G = \begin{bmatrix} \sigma_\beta^2 I_K & 0 \\ 0 & \tau^2 I_m \end{bmatrix}.$$

Nous pouvons estimer les composantes de la variance σ^2, σ_β^2 et τ^2 en ajustant le modèle (2) par la méthode du maximum de vraisemblance restreint (REML).

Les moyennes prévues d'UPE incluses dans l'échantillon sont données par : $\hat{\mu} = X_1\hat{\beta} + X_2\hat{u}$, où $\hat{\beta} = (X_1^T\hat{V}^{-1}X_1)^{-1} \quad X_1^T\hat{V}^{-1}\bar{y}$, $\hat{u} = \hat{G}X_2^T\hat{V}^{-1}(\bar{y} - X_1\hat{\beta})$, où $\hat{V} = X_2\,\hat{G}X_2^T + \hat{\sigma}^2\,\Sigma$, $\Sigma = \mathrm{diag}[\{1/n_h\}_{h=1}^m]$ et $\bar{y} = (\bar{y}_1, \dots, \bar{y}_m)^T$. La moyenne prévue pour une UPE h qui n'est pas sélectionnée au premier degré est $\hat{\mu}_h = x_h^T\hat{\beta}^*$, où

$$x_h = [1 \;\; x_h \;\; (x_h-\kappa_1)_+ \dots (x_h-\kappa_K)_+]^T$$

et

$$\hat{\beta}^* = [\hat{\beta}_0 \;\; \hat{\beta}_1, \;\dots,\; \hat{\beta}_{K+1}]^T.$$

En combinant les prédictions, nous obtenons l'estimateur fondé sur un modèle de la moyenne de population

$$\hat{E}(\bar{Y} \mid y, x_h) = \frac{1}{N}\left(\sum_{h=1}^m [n_h\bar{y}_h + (N_h - n_h)\,\hat{\mu}_h] + \sum_{h=m+1}^H N_h\,\hat{\mu}_h\right).$$

3. Méthodes d'estimation de la variance

3.1 Variance fondée sur un modèle bayésien empirique

Nous pouvons interpréter le modèle (2) comme étant un modèle bayésien dans lequel les paramètres $u = (\beta_2, \dots, \beta_{K+1}, u_1, \dots, u_m)^T$ suivent une loi a priori normale multivariée $N_{K+m}(0, G)$, et $\beta_0, \beta_1, \sigma^2$, σ_β^2 et τ^2 suivent des lois a priori uniformes. Ceci nous mène à la variance bayésienne a posteriori pour le vecteur $(\beta_0, \beta_1, \dots, \beta_{K+1}, u_1, \dots, u_m)^T$ conditionnellement à σ^2, σ_β^2 et τ^2 de la forme

$$\mathrm{Var}((\beta_0, \beta_1, \dots, \beta_{K+1}, u_1, \dots, u_m)^T \mid \sigma^2, \sigma_\beta^2, \tau^2, y) = \sigma^2(X^TX + \Delta)^{-1},$$

où $X = [X_1 \; X_2]$ et

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2/\sigma_\beta^2\, I_K & 0 \\ 0 & 0 & 0 & \sigma^2/\tau^2\, I_m \end{bmatrix},$$

où I_K et I_m sont les matrices identité de dimensions $(K \times K)$ et $(m \times m)$, respectivement.

Pour calculer la variance a posteriori bayésienne empirique de $(\beta_0, \beta_1, \dots, \beta_{K+1}, u_1, \dots, u_m)^T$, nous remplaçons σ^2, σ_β^2 et τ^2 par leurs estimations par le maximum de vraisemblance (ML) ou par le maximum de vraisemblance restreint (REML) $\hat{\sigma}^2$, $\hat{\sigma}_\beta^2$ et $\hat{\tau}^2$, respectivement. La méthode bayésienne empirique sous-estime la variance a posteriori réelle, mais la sous-estimation n'est pas grave pour les tailles d'échantillon utilisées dans nombre d'enquêtes. Il est également possible d'obtenir une solution entièrement bayésienne, mais nous ne la considérons pas ici.

La moyenne prévue de population est $\hat{T}_{\text{prévue}}/N$, où $\hat{T}_{\text{prévue}} = T_1 + \hat{T}_2$, $T_1 = \sum_{h=1}^m n_h\bar{y}_h$ est le total d'échantillon et $\hat{T}_2$ est le total estimé pour les unités non incluses dans l'échantillon,

$$\hat{T}_2 = \sum_{h=1}^m (N_h - n_h)\,\hat{\mu}_h + \sum_{h=m+1}^H N_h\hat{\mu}_h = N_P X_P \left[\hat{\beta}_0 \;\; \hat{\beta}_1 \dots \hat{\beta}_{K+1} \hat{\mu}_1 \dots \hat{\mu}_m\right]^T, \qquad (3)$$

où

$$N_P = \left[(N_1 - n_1)\dots(N_m - n_m)\, N_{m+1} \dots N_H\right],$$

et

$$X_P = \begin{bmatrix} 1 & x_1 & (x_1-\kappa_1)_+ & \cdots & (x_1-\kappa_K)_+ & 1 & 0 & \cdots & 0 \\ . & . & . & \cdots & . & 0 & 1 & 0 & 0 \\ . & . & . & \cdots & . & . & . & 0 & . \\ . & . & . & \cdots & . & 0 & 0 & 1 & 0 \\ 1 & x_m & (x_m-\kappa_1)_+ & \cdots & (x_m-\kappa_K)_+ & 0 & \cdots & 0 & 1 \\ 1 & x_{m+1} & (x_{m+1}-\kappa_1)_+ & \cdots & (x_{m+1}-\kappa_K)_+ & 0 & \cdots & \cdots & 0 \\ . & . & & \cdots & . & . & \cdots & \cdots & . \\ . & . & & \cdots & . & . & \cdots & \cdots & . \\ . & . & & \cdots & . & . & \cdots & \cdots & . \\ 1 & x_H & (x_H-\kappa_1)_+ & \cdots & (x_H-\kappa_K)_+ & 0 & \cdots & \cdots & 0 \end{bmatrix}.$$

La variance a posteriori bayésienne empirique de $\hat{\bar{Y}} = \hat{T}_{\text{prévue}} / N$ est

$$\mathrm{Var}(\hat{\bar{Y}} \mid \sigma^2, \sigma_\beta^2, \tau^2, X, X_P) = \sigma^2 (N_P X_P (X^T X + \Delta)^{-1} X_P^T N_P^T) / N^2 .$$

3.2 Méthode du jackknife

Nous élaborons un estimateur de la variance par le jackknife pour l'estimateur MMP. Nous produisons les répliques jackknife en divisant l'ensemble d'UPE en G sous-groupes égaux et en calculant la g^{e} pseudovaleur selon $\hat{\bar{Y}}_g = G\hat{\bar{Y}} - (G-1)\hat{\bar{Y}}_{(g)}$, où $\hat{\bar{Y}}$ est l'estimateur MMP original et $\hat{\bar{Y}}_{(g)}$ est le même estimateur calculé pour l'échantillon réduit obtenu en excluant les éléments provenant des UPE comprises dans le g^{e} sous-groupe.

L'estimation de la variance de $\hat{\bar{Y}}$ par le jackknife est

$$v(\hat{\bar{Y}}) = \frac{1}{G(G-1)} \sum_{g=1}^{G} (\hat{\bar{Y}}_g - \hat{\bar{\bar{Y}}})^2 ,$$

où $\hat{\bar{\bar{Y}}} = \sum_{g=1}^{G} \hat{\bar{Y}}_g / G$. Afin d'équilibrer la distribution des probabilités de sélection entre sous-groupes, nous stratifions les unités échantillonnées en n/G strates chacune de taille G avec des probabilités d'inclusion de premier degré semblables et nous construisons les G sous-groupes en sélectionnant aléatoirement un élément dans chaque strate. Pour réduire les calculs, nous ne recalculons pas les estimations $\hat{\sigma}^2$, $\hat{\sigma}_\beta^2$ et $\hat{\tau}^2$ pour chaque réplique. Autrement dit, nous calculons des pseudovaleurs de $(\beta_0, \beta_1, ..., \beta_{K+1}, u_1, ..., u_m)^T$ d'après les composantes de la variance estimées à partir de l'échantillon complet.

Miller (1974), ainsi que Shao et Wu (1987, 1989) ont prouvé les propriétés asymptotiques de l'estimateur par le jackknife et de l'estimation de la variance par le jackknife dans le cas de la régression linéaire multiple. Zheng et Little (2004) ont fourni une justification théorique de la méthode du jackknife pour l'estimateur fondé sur un modèle à fonction p-spline dans le cas des plans de sondage à un degré. Les simulations numériques de la section 4 font penser que la méthode du jackknife susmentionnée donne aussi de bons résultats pour les plans d'échantillonnage à deux degrés. Une meilleure performance pourrait être obtenue en utilisant le jackknife pondéré proposé par Hinkley (1977).

3.3 Méthode des répliques répétées équilibrées

La méthode des répliques répétées équilibrées (BRR) peut être appliquée à des plans d'échantillonnage stratifiés avec tirage de deux unités dans chaque strate. Dans le cas des plans d'échantillonnage à une UPE par strate, il est fréquent de regrouper les strates (Kalton 1977) pour l'estimation de la variance par la méthode BRR. Dans notre application, nous supposons que les UPE sont échantillonnées systématiquement à partir d'une liste aléatoirement ordonnée. Nous pouvons considérer ceci approximativement comme un plan d'échantillonnage stratifié à n strates comprenant chacune des UPE pour lesquelles les mesures cumulatives de taille approximative sont $\sum_{i=1}^{H} z_i / n$, où z_i représente les mesures de taille des UPE. Une UPE est échantillonnée dans chacune des n strates. Si nous supposons que n est un nombre pair, nous pouvons approximer le plan d'échantillonnage au moyen d'un plan stratifié à $n/2$ strates dont la mesure de taille est $2\sum_{i=1}^{N} z_i / n$, et où deux unités sont échantillonnées par strate. Nous construisons les demi-échantillons répétés équilibrés en sélectionnant une UPE dans chaque strate, selon un plan fondé sur des matrices de Hadamard (Plackett et Burman 1946). Soit $\hat{\bar{Y}}_b$ l'estimateur à fonction p-spline calculé pour le b^{e} demi-échantillon, au moyen des mêmes nœuds que ceux utilisés pour le calcul fondé sur l'échantillon complet. Le nombre et l'emplacement des nœuds doivent permettre d'ajuster le modèle spline à chaque demi-échantillon. L'estimateur BRR est donné par $v_{\text{BRR}}(\hat{\bar{Y}}) = 1/B \sum_{b=1}^{B} (\hat{\bar{Y}}_b - \hat{\bar{Y}})^2$. Cette estimation de la variance peut contenir un certain biais, parce que le plan d'échantillonnage est traité comme s'il était stratifié avec deux UPE par strate.

4. Cas où certains dénombrements d'UPE sont inconnus

Aux sections 2 et 3, nous avons supposé que les nombres d'USE N_h étaient connus pour les UPE échantillonnées et non échantillonnées. À la présente section, nous examinons la situation où N_h n'est connu exactement que pour les UPE échantillonnées (annotées de 1 à m). Nous supposons aussi que les valeurs $M_h, h = 1, ..., H$ d'une variable auxiliaire prédictive de N_h sont connues pour l'ensemble de la population. Par exemple, les M_h pourraient être les nombres d'USE estimés d'après des sources externes, comme un recensement. Nous exécutons une régression de N_h sur M_h en utilisant les UPE échantillonnées et en remplaçant les dénombrements N_h dans (3) pour les UPE non échantillonnées par les prédictions $\hat{N}_h, h = m+1, \ldots, H$ faites d'après la régression. L'estimation résultante du total est

$$\tilde{T} = T_1 + \sum_{h=1}^{m} (N_h - n_h)\hat{\mu}_h + \sum_{h=m+1}^{H} \hat{N}_h \hat{\mu}_h .$$

L'estimation de la variance de $\tilde{T}$ doit être intégrée à la variabilité supplémentaire de $\hat{N}_h$. Plus précisément, une expression de la variance fondée sur un modèle pour $\tilde{T}$ est

$$\mathrm{Var}(\tilde{T} \mid \pi_h, M_h) = \mathrm{Var}(E(\tilde{T} \mid \hat{N}_h, \pi_h, M_h)) + E(\mathrm{Var}(\tilde{T} \mid \hat{N}_h, \pi_h, M_h)),$$

où

$$E(\tilde{T} \mid \hat{N}_h, \pi_h, M_h) = \sum_{h=1}^{m} (N_h - n_h)\,\mu_h + \sum_{h=m+1}^{H} \hat{N}_h\,\mu_h$$

et

$$\mathrm{Var}(\tilde{T} \mid \hat{N}_h, \pi_h, M_h) \approx \sigma^2 (\tilde{N}_P X_P (X^T X + \Delta)^{-1} X_P^T \tilde{N}_P^T),$$

$\tilde{N}_P = [(N_1 - n_1) \ldots (N_m - n_m)\,\hat{N}_{m+1} \ldots \hat{N}_H]$ et X, X_P et Δ sont définis de la même façon que dans (3).

Si les modèles de μ_h et N_h sont tous deux correctement spécifiés, nous pouvons estimer la variance susmentionnée conformément aux modèles correspondants.

5. Simulations

5.1 Plan de simulation

Nous avons réalisé deux simulations pour comparer la méthode de pondération par la probabilité inverse, la méthode assistée par modèle (Särndal et coll. 1992) et la méthode MMP dans le cas de l'échantillonnage à deux degrés.

Dans notre première simulation, nous générons des populations artificielles dont les fonctions de moyenne $f(\pi_{1,h})$ des probabilités de sélection au premier degré sont différentes. Nous simulons quatre fonctions de moyenne différentes, à savoir 1) NULL, une fonction constante; 2) LINDOWN, une fonction linéairement décroissante; 3) EXP, une fonction exponentiellement croissante et 4) SINE, une fonction sinusoïdale.

Nous simulons deux combinaisons de valeurs des composantes de la variance : 1) $\sigma = 0{,}1$ et $\tau = 0{,}2$; 2) $\sigma = 0{,}2$ et $\tau = 0{,}1$. Enfin, nous simulons uniquement des erreurs normales autour des fonctions de moyenne, ainsi que des erreurs normales et lognormales à l'intérieur des UPE.

La population comprend 500 UPE et, au premier degré, 48 UPE sont échantillonnées systématiquement avec probabilité proportionnelle à la taille (PPT) à partir d'une liste aléatoirement ordonnée. Les tailles des UPE sont distribuées uniformément, les valeurs variant de 4 à environ 400. Le nombre d'USE dans chaque UPE est généré à partir d'une loi de moyenne égale à 1,05 fois la mesure de taille et d'erreurs lognormales avec un écart-type de 30.

Nous étudions deux types de plans d'échantillonnage à deux degrés : 1) l'échantillonnage aléatoire simple (eas) à l'intérieur des UPE avec probabilités de sélection proportionnelles à l'inverse des probabilités de sélection au premier degré, aboutissant à une probabilité de sélection égale pour toutes les USE, 2) un échantillonnage aléatoire simple à l'intérieur des UPE avec le même taux d'échantillonnage pour toutes les UPE échantillonnées, si bien que les probabilités de sélection résultantes pour les USE comprises dans l'UPE *h* soient proportionnelles à $\pi_{1,h}$.

Pour chaque échantillon tiré sous les deux plans d'échantillonnage, nous appliquons les méthodes qui suivent.

A. L'estimateur HT.
B. La méthode d'estimation assistée par modèle. Nous utilisons un modèle linéaire de régression de la variable de résultat y_{hi} sur les probabilités de sélection au premier degré qui sont traitées comme des informations de niveau élémentaire. Nous calculons l'estimateur GR au moyen de la formule donnée à la section 1.
C. La méthode MMP, avec les probabilités de sélection au premier degré $\pi_{1,h}$ comme covariable. Nous utilisons 20 centiles égaux de $\pi_{1,h}$ des UPE échantillonnées comme nœuds pour la *p*-spline de régression.
D. La méthode MMP avec les moyennes d'UPE μ_h estimées de la même façon qu'en C, mais en utilisant des nombres estimés d'UPE d'après une régression linéaire simple de N_h sur les mesures de taille, qui sont proportionnelles à $\pi_{1,h}$. Nous réalisons cette partie de la simulation pour étudier la méthode décrite à la section 4.

Les estimations de $\bar{Y}$ obtenues par les méthodes A à D sont calculées pour chacun des 500 échantillons tirés par répétition à partir des populations artificielles (chaque population artificielle n'est générée qu'une seule fois). Pour l'estimateur MMP, nous calculons les estimateurs de la variance par la méthode bayésienne empirique, la méthode du jackknife ($K = 8$) et la méthode BRR pour chaque échantillon répété. Les estimations moyennes de la variance de l'estimateur MMP et du taux de couverture correspondant à un intervalle de confiance à 95 % sont utilisées pour juger de la qualité de l'inférence. Pour la méthode D, nous étudions le biais empirique de l'estimateur de la variance fondé sur un modèle décrit à la section 4, ainsi que les taux de couverture des intervalles de confiance connexes.

Dans la seconde étude en simulation, nous tirons des échantillons de données sur le revenu du ménage provenant de l'échantillon de microdonnées à grande diffusion (EMGD) à 5 % pour l'État du Michigan au recensement des États-Unis de 1990, que nous traitons comme une population finie. Cette simulation est plus réaliste que la précédente en ce sens que les valeurs de la variable de résultat proviennent d'une distribution réelle plutôt que simulée. Les UPE que nous simulons sont fondées sur les grappes géographiques naturelles appelées « secteurs de microdonnées à grande diffusion » (SMGD), qui sont habituellement des comtés et des localités. Il existe 67 SMGD dans l'EMGD à 5 % du Michigan pour lesquels les nombres de familles varient d'environ 1 300 à plus de 10 000. Nous augmentons le nombre d'UPE disponibles en divisant chaque SMGD par cinq, ce qui donne 335 UPE. Le nombre d'USE varie de 134 à 3 058. La figure 1 donne le nuage de points d'un échantillon du revenu moyen du ménage en fonction des tailles d'UPE, ainsi que la courbe de régression $\hat{f}(x)$.

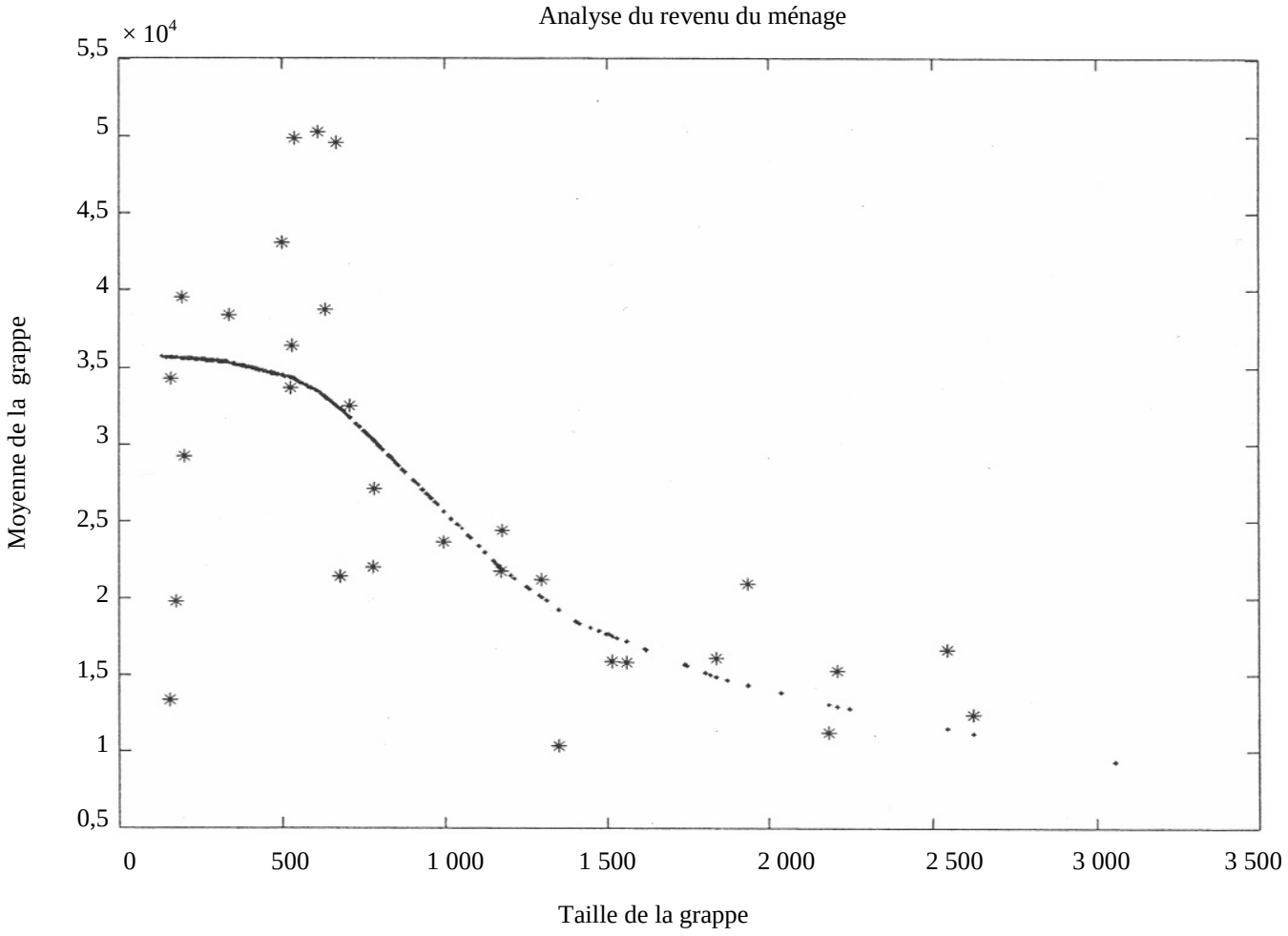

Figure 1. Courbe de régression *p*-spline (trait pointillé) et revenu moyen du ménage (étoiles) dans les UPE échantillonnées.

Nous tirons cinq cents échantillons à deux degrés, chacun comprenant 30 UPE et 20 USE (familles) tirées de chaque UPE sélectionnée. L'échantillonnage de premier degré est un échantillonnage PPT systématique où les mesures de taille sont égales aux dénombrements (nombre d'unités secondaires) des UPE. L'échantillon de deuxième degré est un échantillonnage aléatoire simple avec probabilités de sélection proportionnelles à l'inverse des probabilités de sélection de premier degré. Pour estimer la moyenne, nous utilisons des dénombrements réels d'UPE comme variable x_h, dont les valeurs sont proportionnelles aux probabilités de sélection de premier degré. Nous appliquons le modèle non paramétrique mixte à fonction *p*-spline formulé en (2). Nous utilisons 10 centiles d'échantillon également espacés des dénombrements d'UPE comme nœuds de la *p*-spline.

5.2 Résultats

Le tableau 1 donne le biais empirique et la racine de l'erreur quadratique moyenne (REQM) pour quatre méthodes d'estimation de la moyenne de population finie appliquées à un échantillon sélectionné avec probabilités égales à partir de populations générées avec des erreurs normales ainsi que lognormales à l'intérieur des UPE et deux combinaisons (σ, τ). Le biais empirique et la REQM sont estimés par le biais moyen et l'erreur quadratique moyenne calculés d'après les 500 échantillons répétés.

Le tableau 1 semble indiquer que les méthodes MMP donnent des estimateurs à biais faible. Dans le cas de l'échantillonnage avec probabilités égales, l'estimateur MMP est presque aussi efficace que l'estimateur HT quand la fonction de moyenne f est constante. Dans les cas plus généraux, tels que NULL et LINDOWN, où f est linéaire, mais non constante, la méthode linéaire assistée par modèle et la méthode MMP sont comparables et toutes deux plus efficaces que l'estimateur HT en ce qui concerne la racine de l'erreur quadratique moyenne. Pour les populations EXP et SINE, dont les fonctions de moyenne ne sont pas linéaires, la méthode MMP est supérieure à l'estimateur HT

ainsi qu'à l'estimateur linéaire assisté par modèle. Pour obtenir cette plus grande efficacité, il faut connaître l'information sur le plan de sondage, y compris les probabilités $\pi_{1,h}$ et les dénombrements d'UPE N_h pour l'ensemble de la population. Si l'on utilise des dénombrements d'UPE estimés $\hat{N}_h$ à la place de N_h, l'estimateur résultant est moins efficace qu'avec le N_h connu, mais l'estimateur MMP peut encore surpasser l'estimateur HT quand la fonction moyenne n'est pas constante. Les comparaisons sur les populations dont les erreurs à l'intérieur des UPE sont normales ou lognormales produisent des résultats comparables.

Le tableau 2 montre une amélioration comparable pour la méthode MMP dans le cas de l'échantillonnage avec probabilités inégales, ce qui donne à penser que l'élément essentiel de l'accroissement de l'efficacité est la meilleure prédiction donnée par les modèles non paramétriques. Les tableaux 1 et 2 font penser l'un et l'autre que les estimateurs fondés sur un modèle à fonction *p*-spline sont caractérisés par un biais empirique par rapport au plan de sondage très faible. Nous pensons qu'il en est ainsi parce que les fonctions de moyenne souples donnent de bonnes prédictions des moyennes d'UPE.

Au tableau 3, nous comparons les estimations ponctuelles et la couverture des intervalles de confiance à 95 % calculée par trois méthodes d'estimation de la variance pour l'estimateur MMP, à savoir la méthode fondée sur un modèle bayésien empirique, la méthode du jackknife et la méthode BRR. La méthode bayésienne empirique est généralement satisfaisante, mais a tendance à sous-estimer la variance réelle de l'estimateur MMP, ce qui donne lieu à un sous-couverture dans certains cas. Les méthodes jackknife et BRR ont tendance à produire des estimations plus robustes de la variance. En général, l'estimateur MMP donne des estimations plus efficaces que les estimateurs HT et linéaires assistés par modèle habituels et des estimations fondées sur le plan de sondage satisfaisantes.

Tableau 1
Biais et REQM empiriques des estimateurs MMP, HT, GR et MMP avec N_h estimé sous plans d'échantillonnage avec probabilités égales

		MMP		Horvitz-Thompson		Linéaire assisté par modèle		MMP avec N_h estimé	
	($\times 10^{-3}$)	BIAIS	REQM	BIAIS	REQM	BIAIS	REQM	BIAIS	REQM
Erreurs	NULL	1,1	29,7	0,8	30,0	0,8	29,9	1,3	30,1
normales	LINDOWN	3,5	30,7	3,6	36,4	3,7	30,7	2,3	30,4
$\tau = 0{,}2$	EXP	−4,4	29,1	−9,4	53,0	−9,5	36,7	−4,3	29,1
$\sigma = 0{,}1$	SINE	4,8	32,5	2,1	42,0	−0,3	35,9	5,2	34,3
Erreurs	NULL	5,7	22,0	6,6	22,5	6,6	22,1	5,5	22,3
normales	LINDOWN	0,5	20,4	−0,6	27,1	−0,3	20,5	1,6	20,6
$\tau = 0{,}1$	EXP	0,9	23,1	1,9	50,3	−4,2	31,7	0,4	23,4
$\sigma = 0{,}2$	SINE	7,0	22,3	6,5	34,9	3,8	26,4	8,0	26,4
Erreurs	NULL	1,7	32,3	0,9	32,3	0,7	32,3	1,5	32,5
lognormales	LINDOWN	2,9	31,9	3,8	39,4	2,7	32,1	3,2	32,0
$\tau = 0{,}2$	EXP	−0,6	28,4	−5,9	51,5	−6,9	36,4	−0,3	28,5
$\sigma = 0{,}1$	SINE	6,9	33,8	1,5	43,7	−1,9	39,0	−3,1	35,0
Erreurs	NULL	8,5	30,5	9,6	31,3	9,2	31,0	9,1	30,8
lognormales	LINDOWN	3,6	32,3	1,9	37,5	3,6	32,1	6,4	33,1
$\tau = 0{,}1$	EXP	3,9	29,0	6,8	53,8	1,0	34,4	3,7	29,4
$\sigma = 0{,}2$	SINE	−2,9	30,1	−8,9	44,7	−12,0	38,4	−3,8	35,9

Tableau 2
Biais et REQM empiriques pour les modèles MMP, HT, GR et MMP avec N_h estimé sous plans d'échantillonnage avec probabilités inégales

		MMP		Horvitz-Thompson		Linéaire assisté par modèle		MMP avec N_h estimé	
	($\times 10^{-3}$)	BIAIS	REQM	BIAIS	REQM	BIAIS	REQM	BIAIS	REQM
Erreurs	NULL	−4,5	29,3	−3,7	33,6	−3,2	30,5	−4,5	29,3
normales	LINDOWN	−0,9	27,0	3,7	35,5	1,8	27,7	−0,7	26,9
$\tau = 0{,}2$	EXP	5,8	32,0	1,9	56,8	0,4	39,4	14,1	34,4
$\sigma = 0{,}1$	SINE	7,1	30,1	6,1	39,5	3,6	32,8	5,3	30,4
Erreurs	NULL	−7,7	21,3	−7,7	24,9	−6,6	21,1	−7,6	21,2
normales	LINDOWN	1,1	20,7	3,2	30,6	1,2	20,7	3,5	21,1
$\tau = 0{,}1$	EXP	−2,3	20,9	−6,5	53,3	−7,2	30,0	−3,0	20,9
$\sigma = 0{,}2$	SINE	5,6	20,9	6,9	36,2	4,0	28,6	4,3	21,1
Erreurs	NULL	−0,5	28,5	−2,0	30,6	−2,1	29,5	−0,3	28,5
lognormales	LINDOWN	5,4	32,6	5,0	39,0	3,7	34,1	6,0	32,7
$\tau = 0{,}2$	EXP	−1,3	28,6	−7,6	62,6	−7,1	36,8	−9,3	30,3
$\sigma = 0{,}1$	SINE	3,7	31,2	2,3	43,1	0,1	36,1	1,6	31,0
Erreurs	NULL	3,6	22,8	5,7	28,8	5,7	24,2	3,6	22,7
lognormales	LINDOWN	6,0	26,8	9,3	37,5	7,5	27,3	2,5	26,0
$\tau = 0{,}1$	EXP	0,8	26,3	−2,3	50,8	−3,5	33,1	11,5	29,0
$\sigma = 0{,}2$	SINE	3,7	26,9	2,9	37,6	−0,1	30,2	2,2	27,8

Tableau 3
Estimation de la variance et des taux de couverture empiriques des IC à 95 % par les méthodes fondée sur un modèle, jackknife et BRR

		Variance empirique	Modèle bayésien empirique		Jackknife ($K = 8$)		BRR	
			Estimation		Estimation		Estimation	
	Forme	($\times 10^{-5}$)	($\times 10^{-5}$)	%	($\times 10^{-5}$)	%	($\times 10^{-5}$)	%
Erreurs	NULL	88	74	92,8	94	96,4	96	94,4
normales	LINDOWN	94	73	89,6	94	94,6	98	94,2
$\tau = 0{,}2$	EXP	85	70	91,4	88	94,6	85	93,4
$\sigma = 0{,}1$	SINE	83	67	91,6	90	95,8	85	94,4
Erreurs	NULL	48	45	93,8	48	96,0	49	93,8
normales	LINDOWN	42	45	96,8	51	96,2	51	96,8
$\tau = 0{,}1$	EXP	53	54	95,0	61	97,2	59	95,2
$\sigma = 0{,}2$	SINE	44	46	95,8	55	96,6	49	96,0
Erreurs	NULL	104	83	91,8	104	94,8	100	93,6
lognormales	LINDOWN	102	98	93,6	106	95,6	107	95,0
$\tau = 0{,}2$	EXP	81	77	93,4	97	96,4	89	94,8
$\sigma = 0{,}1$	SINE	92	99	94,8	97	95,2	92	93,4
Erreurs	NULL	93	97	94,2	100	96,2	99	95,2
lognormales	LINDOWN	104	101	93,6	106	96,0	102	92,8
$\tau = 0{,}1$	EXP	84	81	94,6	84	95,2	82	95,0
$\sigma = 0{,}2$	SINE	110	96	94,4	98	95,6	92	93,0

Les tableaux 4 et 5 donnent la variance empirique de l'estimateur MMP quand les dénombrements des UPE non échantillonnées N_h sont estimés. Ils donnent aussi la variance estimée moyenne de cet estimateur et les taux de couverture correspondants des IC à 95 %. Les intervalles de confiance sont les intervalles habituels calculés selon la théorie normale d'après nos estimateurs ponctuel et de variance. Ces deux tableaux montrent que la méthode d'inférence discutée à la section 5 a tendance à sous-estimer la variance réelle de l'estimateur MMP utilisant $\hat{N}_h$, ce qui donne à l'occasion une sous-couverture de la moyenne de population. Il faudra déterminer dans l'avenir si les méthodes JRR et BRR produisent aussi des inférences satisfaisantes pour cette méthode.

Tableau 4
Estimation de la variance et des taux empiriques de couverture des IC à 95 % en utilisant une fonction *p*-spline et des dénombrements d'UPE estimés, population simulée avec des erreurs normales

	$\sigma = 0{,}1$ et $\tau = 0{,}2$			$\sigma = 0{,}2$ et $\tau = 0{,}1$		
	Variance empirique ($\times 10^{-5}$)	Variance estimée ($\times 10^{-5}$)	Taux de couverture	Variance empirique ($\times 10^{-5}$)	Variance estimée ($\times 10^{-5}$)	Taux de couverture
NULL	90	76	91,8	50	46	93,2
LINDOWN	93	74	90,4	43	46	95,6
EXP	85	72	93,0	55	56	96,2
SINE	110	98	94,8	50	55	97,6

Tableau 5
Estimation de la variance et des taux de couverture empiriques des IC à 95 % en utilisant une fonction *p*-spline et des dénombrements d'UPE estimés, population simulée avec des erreurs lognormales

	$\sigma = 0{,}1$ et $\tau = 0{,}2$			$\sigma = 0{,}2$ et $\tau = 0{,}1$		
	Variance empirique ($\times 10^{-5}$)	Variance estimée ($\times 10^{-5}$)	Taux de couverture	Variance empirique ($\times 10^{-5}$)	Variance estimée ($\times 10^{-5}$)	Taux de couverture
NULL	105	84	91,8	95	99	94,8
LINDOWN	103	98	94,4	110	102	94,4
EXP	81	79	94,6	87	83	94,2
SINE	110	150	96,4	91	130	95,8

Pour l'étude en simulation au moyen des données de l'EMGD à 5 %, la moyenne simple a un biais = –50,9 et une REQM = 2 600, et le modèle non paramétrique mixte à fonction *p*-spline (MMP) a un biais égal à = –41,9 et une REQM = 2 153. Donc, les deux méthodes produisent un biais faible et l'estimateur fondé sur le modèle a une REQM inférieure de 17 % à celle de l'estimateur de la moyenne simple. Cet accroissement de l'efficacité est dû au fait que le revenu moyen des ménages diminue à mesure que le nombre de familles augmente dans les UPE (figure 1). La méthode MMP exploite cette relation dans ses prédictions.

6. Discussion

Les méthodes d'inférence fondées sur un modèle paramétrique utilisées antérieurement ont été critiquées principalement parce qu'un biais par rapport au plan de sondage important peut être introduit si le modèle est spécifié incorrectement. Dans nos modèles non paramétriques, l'hypothèse de linéarité est remplacée par une hypothèse beaucoup plus faible de relation à variation lisse. Par conséquent, les estimateurs fondés sur un modèle sont plus robustes, leurs biais étant faibles pour diverses formes de population.

L'information sur le plan de sondage, comme les probabilités de sélection, joue un rôle important dans l'inférence fondée sur un modèle. Les méthodes avec pondération inverse aux probabilités impliquent des hypothèses simples quant à la relation entre les variables de résultat et les variables de plan de sondage. La méthode que nous proposons permet d'augmenter l'efficacité en appliquant des modèles non paramétriques dans lesquels ces hypothèses sont relâchées.

L'un des résultats intéressants de notre étude est que les estimateurs fondés sur un modèle peuvent être plus efficaces que la moyenne simple dans le cas d'un plan de sondage avec probabilités égales. Dans d'autres études, nous constatons aussi un accroissement de l'efficacité lors de l'utilisation de modèles mixtes non paramétriques à fonction *p*-spline pour estimer les moyennes de strate a posteriori dans les échantillons stratifiés a posteriori.

La méthode bayésienne empirique, ainsi que les méthodes du jackknife et BRR donnent toutes une bonne couverture des intervalles de confiance, lesquels sont plus étroits que ceux produits par les méthodes traditionnelles. Cependant, nous nous attendons à ce que la méthode bayésienne empirique soit sensible aux hypothèses du modèle concernant les composantes de la variance (par exemple, variances à l'intérieur des UPE constantes). Si

les dénombrements d'UPE sont connus pour l'échantillon, mais non pour la population dans son ensemble, les estimations fondées sur un modèle des dénombrements inconnus peuvent encore fournir des estimations valables de la moyenne de population si le modèle suit les dénombrements réels d'UPE de façon suffisamment précise. Nous avons traité le modèle reliant ces dénombrements à la variable auxiliaire de façon paramétrique, mais il pourrait être spécifié de façon non paramétrique sans trop de difficulté.

Nous sommes convaincus que les modèles non paramétriques mixtes à fonction *p*-spline peuvent être appliqués à des plans de sondage plus complexes, tels que les plans stratifiés et les plans à plusieurs degrés. Nous pensons aussi que notre méthode peut être généralisée sans trop d'efforts à des résultats binaires ou ordinaux.

Remerciements

La présente étude a été financée par la bourse DMS 0106914 de la National Science Foundation.

Bibliographie

Brumback, B.A., Ruppert, D. et Wand, M.P. (1999). Comment to variable selection and function estimation in additive nonparametric regression using data-based prior. *Journal of the American Statistical Association*, 94, 794-797.

Coull, B.A., Schwartz, J. et Wand, M.P. (2001) Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics*, 2(3), 337-349.

Elliott, M.R., et Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

Firth, D., et Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, B, 60, 3-21.

Ghosh, M., et Meeden, G. (1986). Empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 81, 1058-1062.

Hinkley, D.V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19, 285-292.

Holt, D., et Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, A, 142, 33-46.

HORVITZ, D.G., et THOMPSON, D.J. (1952). A Generalization Of Sampling Without Replacement From A finite universe. *Journal of American Statistical Association*, 47, 663-685.

Kalton, G. (1977). Practical methods for estimating survey sampling errors. *Bulletin of the International Statistical Institute*, 47, 495-514.

Lazzaroni, L.C., et Little, R.J.A. (1998). Random effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.

Lin, X., et Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society*, B, 61, 381-400.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Miller, R.G. (1974). An unbalanced jackknife. *Annals of Statistics*, 2, 880-891.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag.

Shao, J., et Wu, C.F.J. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models. *Annals of Statistics*, 15, 1563-1579.

Shao, J., et Wu, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

U.S. Census (1990). Dept. of Commerce. Census of Population and Housing, [United States]: fichier de l'échantillon de microdonnées à grande diffusion à 5 % . 3e éditions. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 1995. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributeur], 1996.

Zheng, H., et Little, R.J.A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zheng, H., et Little, R.J.A. (2004). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. À paraître dans *Journal of Official Statistics*.