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

Article

Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales

par Qixuan Chen, Michael R. Elliott et Roderick J.A. Little

Juin 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2010
Vol. 36, N° 1, pp. 25-37
Statistique Canada, N° 12-001-X au catalogue

Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales

Qixuan Chen, Michael R. Elliott et Roderick J.A. Little [1]

Résumé

Nous proposons un estimateur de prédiction bayésien avec splines pénalisées (PBSP pour *Bayesian Penalized Spline Predictive*) pour une proportion de population finie sous échantillonnage avec probabilités inégales. Cette nouvelle méthode permet d'intégrer directement les probabilités d'inclusion dans l'estimation d'une proportion de population, en effectuant une régression probit du résultat binaire sur la fonction spline pénalisée des probabilités d'inclusion. La loi prédictive a posteriori de la proportion de population est obtenue en utilisant l'échantillonnage de Gibbs. Nous démontrons les avantages de l'estimateur PBSP comparativement à l'estimateur de Hájek (HK), à l'estimateur par la régression généralisée (RG) et aux estimateurs de prédiction fondés sur un modèle paramétrique au moyen d'études en simulation et d'un exemple réel de vérification fiscale. Les études en simulation montrent que l'estimateur PBSP est plus efficace et donne un intervalle de crédibilité à 95 % dont la probabilité de couverture est meilleure et dont la largeur moyenne est plus étroite que les estimateurs HK et RG, surtout quand la proportion de population est proche de zéro ou de un, ou que l'échantillon est petit. Comparativement aux estimateurs de prédiction fondés sur un modèle linéaire, les estimateurs PBSP sont robustes à l'erreur de spécification du modèle et à la présence d'observations influentes dans l'échantillon.

Mots clés : Analyse bayésienne ; données binaires ; régression par splines pénalisées ; probabilité proportionnelle à la taille ; échantillons d'enquête.

1. Introduction

Les organismes scientifiques et les administrations publiques utilisent souvent des plans de sondage avec probabilités inégales pour recueillir leurs données. Le plan de sondage avec probabilités inégales le plus simple est sans doute l'échantillonnage stratifié, dans lequel des unités sont échantillonnées dans diverses strates avec des probabilités d'inclusion différentes. Une autre forme importante d'échantillonnage avec probabilités inégales est l'échantillonnage avec probabilités proportionnelles à la taille (ppt), dans lequel la probabilité d'inclusion est proportionnelle à la valeur d'une variable de taille mesurée pour toutes les unités de la population.

Un plan d'échantillonnage avec probabilités inégales tel que l'échantillonnage ppt est fréquemment utilisé pour obtenir des estimations efficaces des moyennes de population de variables continues, pour lesquelles la variance augmente avec la taille de l'unité. Cependant, dans une enquête polyvalente, les inférences au sujet de variables discrètes présentent souvent un intérêt également (par exemple, Lehtonen et Veijanen 1998, Lehtonen, Särndal et Veijanen 2005). Dans le présent article, nous nous attachons aux méthodes d'inférence pour des proportions de population finie sous échantillonnage avec probabilités inégales fondées sur une variable auxiliaire mesurée pour toutes les unités de la population. Nous utilisons l'échantillonnage ppt comme plan particulier pour illustrer et évaluer nos méthodes.

Les probabilités d'inclusion jouent un rôle important et légèrement différent dans l'inférence fondée sur le plan de sondage et celle fondée sur un modèle en s'appuyant sur des échantillons tirés avec probabilités inégales (Smith 1976, 1994 ; Kish 1995 ; Little 2004). Dans l'inférence fondée sur le plan de sondage, les variables étudiées sont fixes et l'inférence est basée sur la distribution des indicateurs d'inclusion dans l'échantillon ; dans les approches d'estimation classiques fondées sur le plan, telles que l'estimateur de Horvitz-Thompson (HT) (1952) et ses extensions, les unités échantillonnées sont pondérées par l'inverse de leur probabilité d'inclusion. Ces estimateurs sont convergents par rapport au plan (Isaki et Fuller 1982) et fournissent des inférences fiables pour les grands échantillons sans qu'il soit nécessaire de modéliser les hypothèses. Cependant, ces estimateurs peuvent être très inefficaces, comme l'illustre le célèbre exemple de l'éléphant dans Basu (1971). En outre, l'estimation de la variance est fastidieuse, parce qu'elle nécessite les probabilités d'inclusion de deuxième ordre. Les intervalles de confiance correspondants sont fondés sur la théorie asymptotique et peuvent s'écarter des niveaux nominaux pour des tailles d'échantillon moyennes ou faibles.

L'inférence fondée sur un modèle consiste à prédire les valeurs des variables étudiées dans les unités non échantillonnées en introduisant les probabilités d'inclusion comme

1. Qixuan Chen est professeur adjoint, Department of Biostatistics, Columbia University, 722 West 168 Street, New York, NY 10032. Courriel : qc2138@columbia.edu ; Michael R. Elliott est professeur associé et Roderick J.A. Little est professeur au Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109. Courriel : mrelliot@umich.edu et rlittle@umich.edu.

covariable dans le modèle de prédiction (Little 2004). Les estimateurs de prédiction fondés sur un modèle sont convergents et efficaces sous le modèle supposé, mais sont sujets à un biais quand le modèle sous-jacent est mal spécifié. Cette limite motive l'élaboration de modèles statistiques flexibles qui sont plus robustes aux erreurs de spécification du modèle. Pour des données d'enquête continues, Zheng et Little (2003) ont estimé le total de population finie en utilisant une régression non paramétrique sur une fonction spline pénalisée (*p*-spline) des probabilités d'inclusion. Nous proposons ici des estimateurs de prédiction bayésiens avec splines pénalisés (PBSP) qui conviennent pour un résultat binaire, par opposition à un résultat continu. Nous adoptons une approche d'inférence bayésienne pour ce modèle, parce que les méthodes bayésiennes produisent souvent une meilleure inférence pour les problèmes avec des petits échantillons et qu'elles peuvent être mises en œuvre de manière commode pour le modèle que nous proposons au moyen de l'échantillonneur de Gibbs. Dans cette approche, d'autres variables auxiliaires que la probabilité d'inclusion peuvent être incluses dans le modèle, mais nous choisissons la probabilité d'inclusion puisque la modélisation de cette variable est sujette à des erreurs de spécification du modèle.

Nous comparons la performance des estimateurs PBSP à celle des estimateurs de Hájek (HK, type Horvitz-Thompson) et à celle des estimateurs par la régression généralisée (RG) pour un résultat binaire proposé par Lehtonen et Veijanen (1998). L'approche par la régression généralisée est une modification populaire de celle assistée par modèle des estimateurs fondés sur le plan de sondage qui consiste à combiner les prédictions issues d'un modèle avec les résidus du modèle pondéré par les poids de sondage (Montanari 1998) pour produire des estimations qui sont approximativement sans biais sous le plan.

Zheng et Little (2003 ; 2005) ont comparé par simulation les estimations HT, par prédiction avec *p*-splines et RG du total d'une variable observée continue. Ils ont constaté que les estimateurs fondés sur un modèle avec *p*-splines donnaient une meilleure erreur quadratique moyenne que les autres méthodes et que les erreurs-types jackknife fournissaient une meilleure couverture des intervalles de confiance que les inférences HT ou RG. Nous procédons à des comparaisons similaires pour l'inférence au sujet d'une proportion de population dans le cas d'un résultat binaire et montrons que notre estimateur PBSP offre les mêmes avantages par rapport aux estimateurs HK et RG.

2. Estimateur fondé sur le plan de sondage

Supposons que nous avons une population finie constituée de N unités identifiables. Soit Y la variable observée binaire d'intérêt et $p = N^{-1}\sum_{i=1}^{N} Y_i$, la proportion de la population pour laquelle $Y = 1$. Soit π_i la probabilité d'inclusion de l'unité i, que l'on suppose connue pour toutes les unités faisant partie de la population finie avant qu'un échantillon soit tiré. Nous tirons alors de la population finie un échantillon aléatoire avec probabilités inégales s dont les éléments sont $y_1, \ldots y_n$, conformément aux probabilités d'inclusion $\pi_1, \ldots, \pi_N$. L'estimateur HK fondé sur le plan dont il est discuté dans Basu (1971) est défini comme étant

$$\hat{p}_{\mathrm{HK}} = \frac{\sum_{i\in s} y_i/\pi_i}{\sum_{i\in s} 1/\pi_i}. \qquad (1)$$

La variance de $\hat{p}_{\mathrm{HK}}$ peut être estimée par linéarisation de l'estimateur de Yates-Grundy (1953) des totaux,

$$\hat{V}_{\mathrm{YG}}(\hat{p}_{\mathrm{HK}}) = \left(\sum_{k\in s} 1/\pi_k\right)^{-2} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i - \hat{p}_{\mathrm{HK}}}{\pi_i} - \frac{y_j - \hat{p}_{\mathrm{HK}}}{\pi_j}\right)^2. \qquad (2)$$

L'estimateur de variance de Yates-Grundy nécessite les probabilités d'inclusion par paire. Si ces probabilités ne sont pas disponibles, comme cela est le cas dans nos simulations, la formule approximative proposée par Hartley et Rao (1962),

$$\pi_{ij} \approx \frac{n-1}{n}\pi_i\pi_j + \frac{n-1}{n^2}(\pi_i^2\pi_j + \pi_i\pi_j^2) - \frac{n-1}{n^3}\pi_i\pi_j\sum_{k=1}^{N}\pi_k^2,$$

est souvent utilisée. Un intervalle de confiance de niveau $1 - \alpha$ approximatif pour la proportion de population $\hat{p}_{\mathrm{HK}}$ est alors obtenu en se basant sur l'approximation normale.

3. Estimateur de prédiction bayésien avec splines pénalisées (PBSP)

Royall (1970) a plaidé en faveur de l'utilisation de modèles pour faire des inférences descriptives en population finie en prédisant les valeurs inobservées au moyen de modèles, puisque les inférences fondées sur un modèle devraient être plus efficaces que celles fondées sur le plan de sondage. Afin de modéliser la relation entre le résultat binaire Y et la probabilité d'inclusion continue π, nous devons ajuster une régression binaire de Y en π. Les régressions paramétriques binaires, telles que le modèle logistique ou probit, linéaire ou quadratique, pourraient ne

pas fournir un ajustement adéquat aux données. Un moyen de résoudre ce problème d'inflexibilité consiste à ajuster une régression binaire sur une fonction spline de π en ajoutant certains nœuds. Toutefois, un trop grand nombre de nœuds peut donner de la « rugosité » à l'ajustement du modèle. Un moyen de surmonter ce problème consiste à garder tous les nœuds, mais à restreindre leur influence en ajustant un modèle de régression binaire avec *p*-splines.

Les méthodes courantes de modélisation d'un résultat binaire sont les régressions logistique et probit, qui donnent généralement des résultats semblables. Nous avons opté pour les modèles probit dans notre étude pour des raisons de commodité des calculs. Le modèle de régression probit pour les résultats binaires possède une structure de régression normale tronquée sous-jacente sur des données continues latentes. Si ces données sont connues, les paramètres des modèles de régression binaire avec *p*-splines peuvent être estimés en utilisant les approches classiques pour les modèles de régression normale avec *p*-splines. Dans un contexte bayésien, la loi a posteriori des paramètres dans le modèle probit avec *p*-splines peut être calculée en utilisant l'échantillonnage de Gibbs (Albert et Chib 1993 ; Ruppert, Wand et Carroll 2003, chapitre 16). En revanche, le modèle de régression logistique avec *p*-splines requiert une méthode de calcul plus compliquée, telle que l'algorithme de Metropolis-Hastings. L'avantage en ce qui concerne les calculs rend la fonction de lien probit plus désirable que la fonction de lien logit dans les modèles bayésiens de régression binaire avec *p*-splines.

Les *p*-splines peuvent être de divers types. Quand nous appliquons des *p*-splines, nous devons choisir leur degré et le positionnement des nœuds, ainsi que les fonctions de base utilisées pour présenter le modèle. Nous choisissons d'utiliser des *p*-splines polynomiales tronquées, parce qu'elles sont simples et intuitives. Des estimateurs numériquement plus stables peuvent être obtenus en utilisant des *B*-splines par orthogonalisation des bases de fonctions de puissances tronquées (Eilers et Marx 1996). Le modèle de régression probit avec *p*-splines polynomiales tronquées se représente comme un modèle mixte linéaire généralisé,

$$\Phi^{-1}(E(y_i \mid \beta, b, \pi_i)) = \beta_0 + \sum_{k=1}^{p} \beta_k \pi_i^k + \sum_{l=1}^{m} b_l (\pi_i - k_l)_+^p \quad (3)$$

$$b_l \sim N(0, \tau^2)$$

$$l = 1, \ldots, m\,;\; i = 1, \ldots, n,$$

où $\Phi^{-1}(\cdot)$ désigne l'inverse de la fonction de répartition cumulative d'une loi normale centrée réduite, et les constantes $k_1 < \ldots < k_m$ représentent m nœuds fixes choisis. Une fonction telle que $(\pi_i - k)_+^p$ est appelée une fonction de base spline polynomiale tronquée de puissance p, où $(u)_+^p$ est égal à $\{u \times I(u \geq 0)\}^p$ pour tout nombre réel u. Puisque la fonction de base spline polynomiale tronquée possède $p-1$ dérivées continues, des valeurs plus élevées de p donnent lieu à des splines plus lisses. En spécifiant une loi normale pour b, l'influence des m nœuds est contrainte dans le modèle (3), ce qui équivaut à lisser les splines au moyen de la vraisemblance pénalisée.

Les paramètres du modèle (3) peuvent être estimés en utilisant les méthodes s'appliquant au modèle mixte linéaire généralisé. Une autre approche bayésienne qui simplifie les calculs consiste à émettre l'hypothèse de priors et d'hyperpriors faibles et à utiliser l'échantillonnage de Gibbs pour obtenir des tirages à partir des lois a posteriori des paramètres, de la façon suivante : le modèle de régression probit pour les réponses binaires possède une structure de régression normale sous-jacente sur des données continues latentes ; si ces dernières sont connues, la loi a posteriori des paramètres peut être calculée en utilisant les résultats standard pour les modèles de régression normale ; en outre, sachant la loi a posteriori des paramètres, les données continues latentes peuvent être simulées en partant d'une loi normale tronquée appropriée (Ruppert et coll. 2003, page 290). L'algorithme détaillé de l'échantillonnage de Gibbs est présenté en annexe. En outre, l'inférence bayésienne pour la régression avec *p*-splines peut également être exécutée en utilisant WinBUGS, qui est le logiciel d'analyse bayésienne classique (Crainiceanu, Ruppert et Wand 2005).

La loi a posteriori de la proportion de population est simulée en générant un grand nombre D de tirages et en utilisant l'estimateur de prédiction de la forme $\hat{p}_{\mathrm{PR}}^{(d)} = N^{-1}(\sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j^{(d)})$, où $\hat{y}_j^{(d)}$ est un tirage a partir de la loi prédictive a posteriori de la j^{e} unité non échantillonnée du résultat binaire. La moyenne de ces tirages simule l'estimateur de prédiction bayésien avec splines pénalisées (PBSP) de la proportion de population finie et est désignée par $\hat{p}_{\mathrm{PBSP}}$, où

$$\hat{p}_{\mathrm{PBSP}} = D^{-1} \sum_{d=1}^{D} \hat{p}_{\mathrm{PR}}^{(d)}. \quad (4)$$

L'analogue bayésien d'un intervalle de confiance à $100 \times (1-\alpha)\,\%$ pour la proportion de population est un intervalle de crédibilité à $100 \times (1-\alpha)\,\%$, qui peut être formé de plusieurs façons. Dans les simulations, nous divisons également l'aire α située dans la queue entre les points limites supérieur et inférieur.

Firth et Bennett (1998) ont montré que tout modèle de régression logistique paramétrique contenant un terme d'ordonnée à l'origine et l'inverse des probabilités d'inclusion comme covariable, ajusté par le maximum de vraisemblance ordinaire, non pondéré, présente un « calage interne pour le biais » pour les proportions de population et donne donc lieu à la convergence sous le plan. Cette

propriété est également vérifiée pour les modèles de régression logistique avec *p*-splines polynomiales tronquées sur l'inverse des probabilités d'inclusion, ajustés au moyen de la vraisemblance pénalisée. Si nous utilisons la fonction de lien probit au lieu de la fonction de lien logit et que nous effectuons l'ajustement au moyen de l'algorithme de Monte Carlo par chaînes de Markov au lieu du maximum de vraisemblance pénalisée, l'estimateur PBSP pourrait ne plus avoir la propriété de « calage interne pour le biais ». Cependant, la similarité entre le modèle probit et le modèle logistique implique que l'estimateur de prédiction basé sur le modèle de régression probit avec *p*-splines est approximativement convergent sous le plan. Nous estimons qu'obtenir des estimations efficaces avec couverture des intervalles de confiance proches du taux nominal dans les échantillons finis est plus important que la convergence exacte sous le plan.

4. Estimateur par la régression généralisée

Pour l'estimation des fréquences de classe d'une variable de réponse discrète, Lehtonen et Veijanen (1998) ont proposé un estimateur par la régression généralisée (RG) $\hat{t}_{\mathrm{RG}}$ du total, qui combine les valeurs prédites $\hat{y}_i = \hat{\Pr}(Y_i = 1 \mid \pi_i)$ en se basant sur un modèle approprié et l'estimateur HT pour les résidus $r_i = y_i - \hat{y}_i$ des unités échantillonnées,

$$\hat{t}_{\mathrm{RG}} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i \in s} r_i / \pi_i. \qquad (5)$$

L'estimateur RG donné par l'équation (5) est alors utilisé pour construire un estimateur des proportions de population en divisant par la taille connue de population N (Duchesne 2003),

$$\hat{p}_{\mathrm{RG\_1}} = \frac{1}{N}\left(\sum_{i=1}^{N} \hat{y}_i + \sum_{i \in s} r_i / \pi_i \right). \qquad (6)$$

Nous considérons également ici une autre version de l'estimateur RG pour l'estimation des proportions de population finie, dans lequel le dénominateur du terme de calage du biais pour les résidus r_i est la taille de population estimée $\sum_{i \in s} 1/\pi_i$,

$$\hat{p}_{\mathrm{RG\_2}} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i + \left(\sum_{i \in s} r_i / \pi_i \right)\left(\sum_{i \in s} 1/\pi_i \right)^{-1}. \qquad (7)$$

Pour estimer la variance de (6), nous utilisons l'estimateur de variance du total estimé d'une variable réponse discrète, donné par Lehtonen et Veijanen (1998), divisé par N^2. Pour estimer la variance de (7), nous appliquons la méthode de linéarisation de Taylor (Särndal, Swensson et Wretman 1992, page 182). Ces deux estimateurs de variance prennent la forme des équations (8) et (9), respectivement,

$$\hat{V}(\hat{p}_{\mathrm{RG\_1}}) = \frac{1}{N^2}\sum_{k \in s}\sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{r_k}{\pi_k} \frac{r_l}{\pi_l}, \qquad (8)$$

$$\hat{V}(\hat{p}_{\mathrm{RG\_2}}) = \left(\sum_{i \in s} 1/\pi_i \right)^{-2} \sum_{k \in s}\sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}, \qquad (9)$$

où $e_k = r_k - (\sum_{i \in s} r_i / \pi_i)(\sum_{i \in s} 1/\pi_i)^{-1}$. Ces estimateurs de variance requièrent aussi les probabilités d'inclusion par paire, qui peuvent être approximées par la méthode de Hartley et Rao (1962).

Cependant, l'approximation de Hartley et Rao peut donner lieu à un biais dans l'estimateur de variance. Donc, nous considérons également la méthode du jackknife pour l'estimation de la variance (Shao et Wu 1989). Nous stratifions l'échantillon en n/G strates, chacune de taille G avec des valeurs similaires des probabilités d'inclusion, puis nous construisons les G sous-groupes en sélectionnant un élément à la fois dans chaque strate, sans remise (Zheng et Little 2005). Soit $\hat{p}_{(g)}$ les mêmes estimateurs RG que ceux donnés par les expression (6) et (7) calculés en se basant sur l'échantillon réduit sans les éléments compris dans le g^{e} sous-groupe, et soit $\overline{p}$ la moyenne des G estimateurs basée sur les G échantillons réduits. L'estimateur de variance jackknife de $\hat{p}_{\mathrm{RG}}$ est

$$\hat{V}_{\mathrm{jackknife}}(\hat{p}_{\mathrm{RG}}) = \frac{G-1}{G}\sum_{g=1}^{G} (\hat{p}_{(g)} - \overline{p})^2. \qquad (10)$$

Nous avons utilisé un modèle de régression logistique pondéré par les poids de sondage ajustés sur d'autres covariables comme modèle auxiliaire pour prédire $\hat{y}_i$ dans les estimateurs RG pour les résultats binaires (Lehtonen et Veijanen 1998 ; Lehtonen et coll. 2005). Comme nous souhaitons ici comparer les estimateurs RG avec l'estimateur PBSP, nous appliquons les estimateurs (6) et (7) avec des modèles de régression probit linéaire et des modèles probit avec *p*-splines, comme il est décrit en détail à la section 5. Pour l'estimateur RG s'appuyant sur un modèle probit linéaire comme modèle auxiliaire, nous utilisons la probabilité d'inclusion comme covariable, de même qu'un poids dans nos simulations.

5. Étude en simulation

5.1 Plan de l'étude en simulation

Nous avons réalisé des études en simulation pour étudier la performance de l'estimateur PBSP comparativement à l'estimateur HK, aux estimateurs RG et aux estimateurs de prédiction fondés sur un modèle linéaire pour diverses populations sous échantillonnage ppt. Nous présentons les résultats des simulations pour les six estimateurs suivants :

a) HK, l'estimateur de Hájek défini par l'équation (1) ;

b) RL, estimateur de prédiction de la forme $\hat{p}_{\mathrm{RL}} = N^{-1}(\sum_{i\in s} y_i + \sum_{j\notin s} \hat{y}_j^{\mathrm{RL}})$ avec prédiction $\hat{y}_j^{\mathrm{RL}}$ obtenue en se servant des prédictions du maximum de vraisemblance provenant du modèle de régression logistique linéaire contenant un terme constant et la réciproque de la probabilité d'inclusion comme covariable. RL possède la propriété de « calage interne pour le biais » et est donc convergent sous le plan. RL est exactement le même que son estimateur RG donné par l'équation (6).

c) PR, l'estimateur de prédiction de la forme $\hat{p}_{\mathrm{PR}} = N^{-1}(\sum_{i\in s} y_i + \sum_{j\notin s} \hat{y}_j^{\mathrm{PR}})$ avec la prédiction $\hat{y}_j^{\mathrm{PR}}$ provenant du modèle probit linéaire bayésien contenant un terme d'ordonnée à l'origine et la probabilité d'inclusion comme covariable ;

d) PR_RG, l'estimateur RG donné par l'équation (7), où $\hat{y}_i$ est la prédiction pour l'unité i quand les paramètres inconnus sont remplacés par les estimations du maximum de vraisemblance pondérées provenant du modèle probit avec un terme constant et la probabilité d'inclusion comme covariable ;

e) PBSP, l'estimateur PBSP défini par l'équation (4), avec $p = 1$, une loi a priori gamma inverse pour τ^2 et l'utilisation de 15 nœuds ;

f) PBSP_RG, l'estimateur RG donné par l'équation (7), où $\hat{y}_i$ est la moyenne a posteriori de $\Pr(Y_i = 1 \mid \pi_i)$ provenant du modèle PBSP.

Nous donnons uniquement les résultats des simulations basés sur les splines linéaires pour l'estimateur PBSP, puisque les simulations non présentées ici laissent entendre que les splines linéaires donnent d'aussi bons résultats que les splines quadratiques ou les splines cubiques dans tous les scénarios de simulation. Nous avons choisi deux nombres fixes de nœuds (15 ou 30) et avons positionné les nœuds à des centiles d'échantillon uniformément espacés. Les choix de nœuds donnent de bons résultats et un total de 15 nœuds est suffisant pour saisir les courbures dans nos simulations. En outre, les estimateurs RG donnés par (6) donnent des résultats comparables à ceux des estimateurs donnés par (7) ; certaines différences entre ces estimateurs se dégagent dans l'application des données réelles décrites à la section 6, ce qui nous mène à donner la préférence à (7) par rapport à (6).

Nous avons simulé deux populations artificielles de taille 2 000 en utilisant deux lois différentes, avec des taux d'échantillonnage de 5 % et de 10 %, où la variable de taille rend les valeurs entières consécutives 71, 72, ..., 2 070. Nous avons ensuite calculé les probabilités d'inclusion dans la population comme étant proportionnelles à la variable de taille, la valeur maximale correspondant à environ 30 fois les valeurs minimales.

Nous avons d'abord généré des données continues Z à partir de lois normales ayant une structure de moyenne $f(\pi)$ et une variance du terme d'erreur constante et égale à 0,04. Nous avons simulé deux structures de moyenne $f(\pi)$ distinctes, à savoir une fonction linéairement croissante (LINUP) $f(\pi_i) = k_1\pi_i$ et une fonction exponentielle (EXP) $f(\pi_i) = \exp(-4{,}64 + k_2\pi_i)$. Afin que l'étendue de Z soit la même pour les diverses structures de moyenne, k_1 prend les valeurs de 3 et de 6, et k_2 prend les valeurs de 26 et de 52, quand le taux d'échantillonnage est de 10 % et de 5 %, respectivement. Les deux populations sont représentées graphiquement à la figure 1. Ensuite, nous avons généré la variable de résultat binaire Y_1, qui est égale à 1 si la valeur de Z est inférieure ou égale à son 10e centile de superpopulation, et égale à 0 autrement. De même, nous avons généré les résultats binaires Y_2 et Y_3 en utilisant les 50e et 90e centiles de Z en superpopulation comme valeur seuil. Ici, la cible de l'inférence est la proportion de population pour laquelle Y est égale à 1.

Dans chaque réplique simulée, nous avons généré une population finie avant de tirer un échantillon, puis nous avons calculé la proportion de population finie réelle pour laquelle Y est égale à 1 et l'avons désignée p. Nous avons alors tiré un échantillon ppt systématiquement d'une liste ordonnée aléatoirement de la population finie. Pour chaque combinaison de taille de population et d'échantillon, nous avons obtenu 1 000 répliques et avons comparé les six estimateurs en ce qui concerne le biais, la racine carrée de l'erreur quadratique moyenne (REQM) et le taux de non-couverture de l'intervalle de confiance/crédibilité à 95 % empiriques. Nous présentons les résultats des simulations aux tableaux 1 à 3. Soit $\hat{p}_i$ une estimation de p_i basée sur le i^{e} échantillon ppt ; le biais et la REQM empiriques sont définis comme il suit :

$$\text{Biais} = \frac{1}{1\,000}\sum_{i=1}^{1\,000}(\hat{p}_i - p_i),$$

$$\text{REQM} = \sqrt{\frac{1}{1\,000}\sum_{i=1}^{1\,000}(\hat{p}_i - p_i)^2}.$$

5.2 Résultats des simulations

La figure 2 donne les moyennes a posteriori de $\Pr(Y_i = 1 \mid \pi_i)$ et les intervalles de crédibilité à 95 % basés sur le modèle probit bayésien avec p-splines linéaires pour un échantillon ppt aléatoire tiré du cas EXP. Le tracé supérieur gauche est le diagramme de dispersion de la variable continue Z dans un échantillon ppt, avec trois droites horizontales parallèles superposées représentant les 10e, 50e et 90e centiles de la superpopulation, respectivement. Dans le tracé supérieur droit, la variable binaire Y, définie comme étant égale à 1 si la valeur de Z est inférieure ou égale au 10e centile de la superpopulation, est représentée par des cercles noirs, et la $\Pr(Y_i = 1 \mid \pi_i)$ de superpopulation est représentée par une courbe en trait plein noire. La

courbe en trait plein grise et les deux courbes en trait interrompu grises représentent les moyennes a posteriori de $\Pr(Y_i = 1 \mid \pi_i)$ et les intervalles de crédibilité à 95 % basés sur le modèle de régression probit bayésien avec *p*-splines linéaires. Les deux autres tracés sont similaires au tracé supérieur gauche, mais en prenant les 50e et 90e centiles de la superpopulation comme valeur seuil pour définir Y. Ces tracés montrent que les probabilités réelles que $Y = 1$ sont comprises dans les intervalles de crédibilité à 95 % et sont proches des moyennes a posteriori de $\Pr(Y_i = 1 \mid \pi_i)$. Nous concluons que le modèle de régression probit bayésien avec *p*-splines donne un bon ajustement pour les résultats binaires dans le cas non linéaire.

Le tableau 1 présente le biais empirique ($\times 10^3$) pour les six estimateurs dans les deux populations générées à partir de LINUP et EXP. Dans l'ensemble, les estimateurs fondés sur le plan de sondage (a, d et f) produisent un biais plus faible que ceux fondés sur le modèle (b, c et e). Dans le cas LINUP, le modèle de régression probit linéaire est spécifié correctement, de sorte que le biais empirique des estimateurs PR est semblable au biais empirique de l'estimateur PBSP ; par contre, dans le cas EXP, les données requièrent l'ajustement d'un modèle de régression probit non linéaire et, donc, le biais de l'estimateur PR est plus grand que celui de l'estimateur PBSP quand les proportions réelles de population sont 0,1 et 0,5. Cependant, le biais de l'estimateur RL est semblable au biais empirique de l'estimateur PBSP à cause de sa propriété de calage interne pour le biais. Comparativement aux estimateurs fondés sur un modèle PR et PBSP, les estimateurs PR_RG et PBSP_RG réduisent le biais grâce à l'ajout du terme de calage du biais. En outre, quelque soit le modèle auxiliaire utilisé, les deux estimateurs RG produisent un biais empirique semblable.

Le tableau 2 donne la racine carrée de l'erreur quadratique moyenne empirique ($\times 10^3$) pour les six estimateurs. La racine carrée de l'erreur quadratique moyenne empirique de l'estimateur PBSP est beaucoup plus petite que celle de l'estimateur HK, sauf quand p est égale à 0,1 dans le cas EXP. Dans l'ensemble, l'estimateur PR donne des résultats comparables à l'estimateur PBSP. Afin d'offrir une protection contre l'erreur de spécification du modèle, les estimateurs RG perdent une certaine efficacité comparativement aux estimateurs de prédiction fondés sur un modèle correspondant. L'estimateur PR_RG a une REQM similaire à l'estimateur PBSP_RG, mais chacun des deux estimateurs RG a une plus petite REQM que l'estimateur HK grâce à l'utilisation de modèles auxiliaires.

Le tableau 3 donne la probabilité de non-couverture ($\times 10^2$) des intervalles de confiance/crédibilité à 95 %, c'est-à-dire la probabilité que la proportion réelle de population finie se situe à l'extérieur de l'IC à 95 % des estimateurs. Pour calculer les variances des estimateurs, nous utilisons l'estimateur de variance de Yates-Grundy défini par l'équation (2) pour l'estimateur HK, la méthode de rééchantillonnage jackknife définie par l'équation (10) pour l'estimateur RL, ainsi que la méthode de linéarisation (V1) définie par l'équation (9) et la méthode de rééchantillonnage jackknife (V2) pour les estimateurs PR_RG et PBSP_RG. Dans l'ensemble, la couverture de l'intervalle de crédibilité est plus proche du taux nominal pour l'estimateur PBSP que pour les cinq autres estimateurs, surtout quand la proportion de population p est proche de 0 ou de 1, ou que peu d'observations sélectionnées dans l'échantillon proviennent des queues de la distribution. En particulier, l'estimateur PBSP donne lieu à une amélioration importante de la couverture quand p est proche de 0 tant dans le cas LINUP que dans le cas EXP, puisque peu de données provenant de la queue inférieure des deux populations sont incluses dans l'échantillon. Notons que la couverture améliorée de l'estimateur PBSP est réalisée avec des intervalles qui sont plus étroits en moyenne que ceux des estimateurs HK, LR, PR_RG et PBSP_RG. Comme dans le cas du biais et de la REQM empiriques, l'estimateur PBSP_RG n'améliore pas la couverture comparativement à l'estimateur PR_RG en utilisant un modèle auxiliaire flexible.

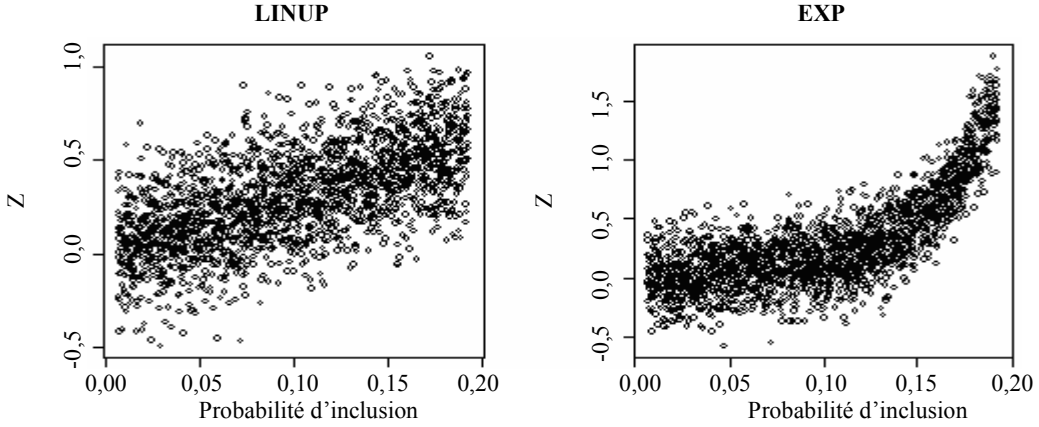

Figure 1 Deux populations artificielles simulées (N = 2 000)

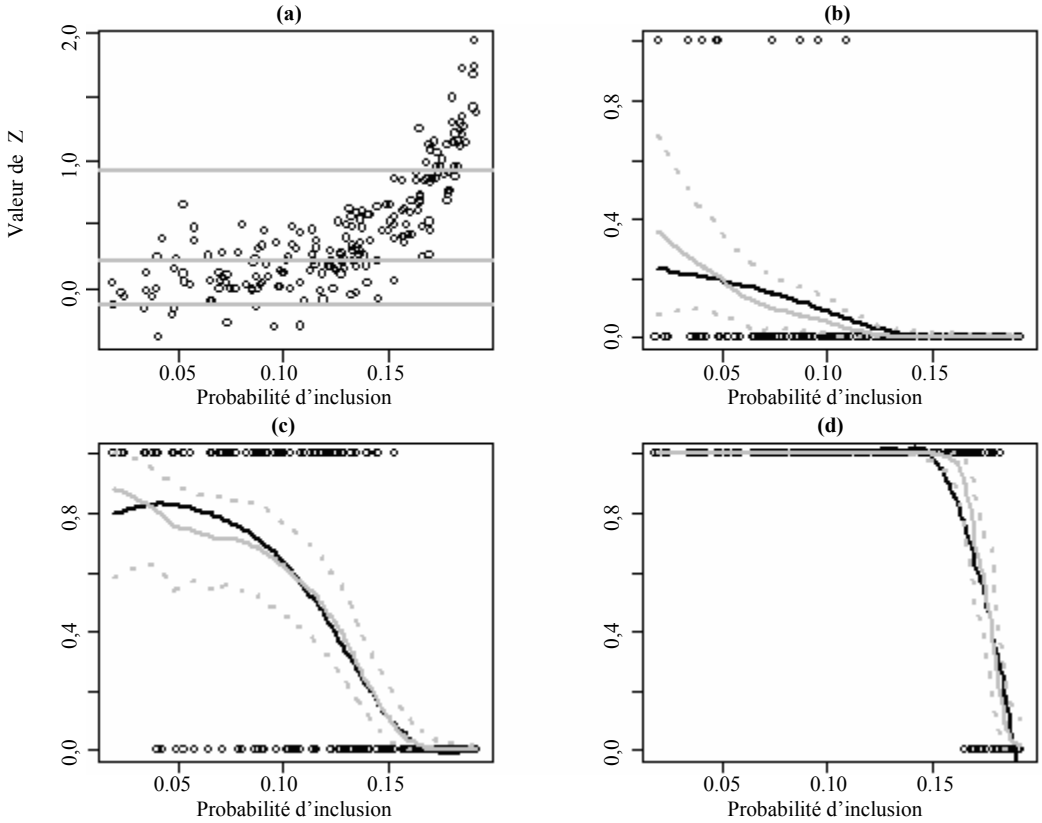

Figure 2 **Un échantillon ppt aléatoire tiré du cas EXP (n = 200, N = 2 000) : (a) diagramme de dispersion de *Z* ; les trois droites grises représentent les 10^e^, 50^e^ et 90^e^ centiles de la superpopulation, respectivement. (b) Les cercles noirs représentent les unités observées de la variable binaire étudiée *Y* dans l'échantillon, définie comme étant *Y* = *I* (*Z* ≤ 10^e^ centile) ; les courbes en trait plein et en trait interrompu grises sont les moyennes a posteriori de Pr(Y_i = 1|π_i) et les intervalles de crédibilité à 95 %, respectivement, simulés en se basant sur un modèle de régression probit avec *p*-splines sur π ; la courbe noire est la probabilité Pr(Y_i = 1|π_i) de superpopulation. (c) Similaire à (b), mais avec *Y* = *I* (*Z* ≤ 50^e^ centile). (d) Similaire à (b), mais avec *Y* = *I* (*Z* ≤ 90^e^ centile)**

Tableau 1
Biais empirique × 1 000 des six estimateurs (le biais absolu minimum dans une ligne est en caractères italique)

Population	*n*	Prop. réelle	HK	RL	PR	PR_RG	PBSP	PBSP_RG
LINUP	100	0,10	*-0,01*	13,0	10,3	1,6	8,0	1,2
		0,50	-4,0	*-2,9*	-4,3	-3,0	-5,2	-3,3
		0,90	-0,4	0,3	-2,5	0,3	-2,9	*0,08*
	200	0,10	2,5	7,9	5,8	1,5	5,1	*1,4*
		0,50	3,3	-0,1	-1,3	*-0,06*	-1,7	-0,2
		0,90	1,6	0,4	-1,0	*0,3*	-1,2	*0,3*
EXP	100	0,10	*1,2*	18,1	25,8	4,7	17,0	3,9
		0,50	-4,0	-3,5	12,5	-1,6	*-1,4*	-3,4
		0,90	-1,3	-0,2	-1,0	*-0,1*	-1,0	-0,2
	200	0,10	3,1	11,0	22,1	3,5	13,4	*2,7*
		0,50	3,8	-0,6	14,0	0,4	*0,01*	-0,7
		0,90	2,3	0,1	-0,7	0,1	-0,7	*0,02*

Le choix des priors et hyperpriors dans les modèles mixtes peut avoir une incidence énorme sur les inférences. Nous avons utilisé une loi a priori $N(0,10^6)$ pour les paramètres à effets fixes, β_i. Dans nos simulations, nous présentons les résultats basés sur une loi a priori gamma inverse appropriée pour τ^2, à savoir $\tau^2 \propto \mathrm{GI}(0,1,0,1)$. Pour évaluer la sensibilité du choix des lois a priori, nous avons également calculé les résultats en utilisant $\tau^2 \propto \mathrm{GI}(0,01,0,01)$ et $\tau^2 \propto \mathrm{GI}(0,001,0,001)$, ainsi qu'une loi a priori uniforme impropre sur τ (Gelman 2006). Ces divers priors ont peu d'effet sur l'inférence a posteriori de la proportion d'intérêt.

6. Exemple de la vérification fiscale

Nous comparons maintenant l'estimateur PBSP à d'autres méthodes sur une population réelle comportant des données de vérification fiscale du revenu (Compumine 2007). L'ensemble de données comprend 3 119 déclarations de revenus produites en Suède par des personnes qui, durant l'année, ont vendu des fonds communs de placement gérés dans un pays étranger. Le résultat d'intérêt Y est la question de savoir si la déclaration de revenus est incorrecte (code égal à 1 pour incorrecte et égal à 0 pour correcte) ; il est mesuré pour toutes les observations figurant dans cet ensemble de données. Ici, nous avons traité les 3 119 déclarations de revenus comme une population finie, de sorte que la proportion de population réelle de déclarations de revenus incorrectes est 0,517. Puisque le montant des bénéfices positifs réalisés est une caractéristique importante de la détermination du montant que le contribuable a caché pour le soustraire à l'imposition dans sa déclaration de revenus provenant de la vente d'un fonds étranger, nous l'avons choisi comme variable de taille pour tirer l'échantillon ppt. La principale mesure d'intérêt étant le montant total non déclaré aux fins de l'impôt par le contribuable, il est raisonnable d'attribuer une valeur de une couronne suédoise aux bénéfices négatifs, c'est-à-dire le montant minimum de bénéfice positif, les valeurs négatives n'étant pas permises dans la variable de taille.

Nous avons tiré sans remise à partir de listes de population classées aléatoirement un millier d'échantillons ppt systématiques répétés de taille 300 et 600. Les déclarations contenant les bénéfices les plus importants ont été incluses avec certitude dans les échantillons de taille 300 et 600 : il existait 78 et 241 de ces déclarations, respectivement. La figure 3 montre que la distribution de la probabilité d'inclusion est étalée vers la droite pour la population, même après avoir exclu les observations dont la probabilité d'inclusion est égale à 1.

Nous avons appliqué aux échantillons ppt les 6 mêmes estimateurs que dans l'étude en simulation avec 30 nœuds et comparé leurs propriétés en ce qui concerne le biais, la REQM, ainsi que la largeur moyenne et le taux de non-couverture des intervalles de confiance/crédibilité à 95 % empiriques. Pour l'estimateur PBSP, un nombre fixe de 30 nœuds sont positionnés à des centiles d'échantillon uniformément espacés des probabilités d'inclusion. Pour les estimateurs RG, ni l'estimateur de variance par linéarisation ni celui par le jackknife ne possédant des propriétés essentiellement meilleures que l'autre, nous présentons l'inférence basée sur l'estimateur de variance par linéarisation pour simplifier les calculs. Nous donnons les résultats pour les estimateurs RG basés sur l'équation (6), ainsi que (7) au tableau 4.

Tableau 2
REQM empirique × 1 000 des six estimateurs (la REQM minimale dans une ligne est en caractères italique)

Population	*n*	Prop. réelle	HK	RL	PR	PR_RG	PBSP	PBSP_RG
LINUP	100	0,10	55,1	57,1	*46,3*	51,3	47,2	51,7
		0,50	65,2	50,8	*47,1*	49,7	47,7	50,0
		0,90	26,3	*22,6*	23,3	22,7	23,5	22,9
	200	0,10	39,3	40,9	*31,8*	36,1	*32,0*	36,2
		0,50	45,7	35,9	*32,8*	34,3	*32,8*	34,6
		0,90	17,8	15,4	15,5	15,4	15,5	*15,3*
EXP	100	0,10	*51,2*	60,1	54,4	51,6	51,8	52,4
		0,50	66,1	56,0	*43,0*	53,2	47,0	51,7
		0,90	24,2	12,4	*12,3*	12,4	*12,3*	*12,3*
	200	0,10	35,9	42,4	39,6	*35,6*	36,0	36,2
		0,50	45,1	38,9	*31,3*	36,1	32,1	35,1
		0,90	15,8	*8,0*	8,1	*8,0*	*8,0*	*8,0*

Tableau 3
Taux de non-couverture de l'IC à 95 % × 100 des six estimateurs (le taux de non-couverture le plus proche de 5 dans une ligne est en caractères italique)

Population	*n*	Prop. réelle	HK	RL	PR	PR_RG		PBSP	PBSP_RG	
						V1	V2		V1	V2
LINUP	100	0,10	16,2	18,0	*8,4*	20,9	16,1	9,0	18,4	14,2
		0,50	7,5	9,4	*5,0*	7,2	7,6	4,4	7,3	7,1
		0,90	7,4	11,4	5,7	8,0	9,4	*5,4*	8,4	7,1
	200	0,10	10,8	12,6	6,4	13,9	10,9	*6,2*	12,6	9,4
		0,50	5,5	8,3	5,5	6,2	5,9	*5,1*	6,0	5,5
		0,90	6,0	8,4	4,4	6,1	4,4	*4,7*	6,3	5,5
EXP	100	0,10	15,0	18,1	10,5	19,4	14,8	*9,2*	18,4	14,4
		0,50	*7,4*	13,5	12,2	9,0	11,4	8,9	10,2	8,4
		0,90	*6,1*	10,5	7,9	9,9	7,6	7,0	9,8	7,2
	200	0,10	10,8	13,3	9,9	12,5	11,7	*7,5*	12,4	9,4
		0,50	*6,0*	11,5	14,3	7,2	8,5	6,2	7,5	6,9
		0,90	*5,5*	8,8	5,5	6,8	4,6	*5,5*	6,6	3,7

* V1 : Estimateur de variance par linéarisation ; V2 : estimateur de variance jackknife.

Le tableau 4 montre que l'estimateur PBSP possède un biais légèrement plus grand, mais une REQM plus petite, et un intervalle de crédibilité dont la largeur moyenne est plus étroite et dont la couverture est plus proche du niveau nominal que les estimateurs fondés sur le plan (a), (d) et (f). Des résultats non présentés ici indiquent que l'estimateur PBSP avec une loi a priori uniforme donne d'un peu meilleurs résultats que celui avec une loi a priori gamma inverse en ce qui concerne le biais, la REQM et le taux de couverture empiriques, parce qu'il existe plus de fluctuations dans les données et que la loi a priori uniforme donne plus de flexibilité à la fonction ajustée. L'estimateur PBSP_RG produit un biais plus faible, mais est moins efficace et a un moins bon taux de couverture que l'estimateur PBSP. L'estimateur de prédiction basé sur le modèle de régression probit linéaire comme modèle de prédiction donne de médiocres résultats ici puisque le modèle est mal spécifié, mais son estimateur RG réduit le biais et la REQM, et améliore le taux de couverture. L'estimateur PBSP_RG basé sur l'équation (6) donne de très mauvais résultats en ce qui concerne la REQM comparativement à l'estimateur donné par l'équation (7), à cause d'une situation similaire à l'exemple de l'éléphant de cirque de Basu (1971), où une ou plusieurs observations ayant une très faible probabilité d'inclusion sont sélectionnées dans l'échantillon et, donc, reçoivent des poids très grands. Cependant, l'estimateur PR_RG donné par l'équation (6) a d'aussi bonnes propriétés que celui donné par l'équation (7) avec les prédictions obtenues d'après les estimations pondérées du maximum de vraisemblance, où la probabilité d'inclusion est utilisée comme covariable, ainsi que les poids de sondage. Dans l'ensemble, l'estimateur RG donné par l'équation (7) est préférable à celui donné par l'équation (6). À mesure que la taille d'échantillon augmente, passant de 300 à 600, la probabilité de non-couverture de l'intervalle de crédibilité à 95 % de l'estimateur PBSP s'approche du niveau nominal de 5 % rapidement de 14 % à 5 %, mais les couvertures sont systématiquement inférieures au niveau nominal pour les autres estimateurs.

Comparativement aux estimateurs de prédiction fondés sur un modèle linéaire, l'estimateur PBSP est robuste non seulement à l'erreur de spécification du modèle, mais aussi aux observations influentes présentes dans l'échantillon. Pour démontrer la robustesse aux observations influentes, nous comparons les variations de l'adéquation du modèle en utilisant des modèles probit avec *p*-splines, un modèle probit linéaire et un modèle probit quadratique basés sur l'échantillon ppt uniquement à la figure 4, et basés sur l'échantillon ppt ainsi que sur les observations avec probabilités d'inclusion égales à 1 à la figure 5. Dans chaque figure, la population est stratifiée selon les 100 quantiles des probabilités d'inclusion, et les probabilités réelles que $Y = 1$ sont calculées et représentées par un point noir pour chaque strate. Les courbes grises représentent les moyennes a posteriori de $\Pr(Y_i = 1 \mid \pi_i)$ sur 10 échantillons ppt aléatoires en utilisant 3 000 itérations de l'échantillonneur de Gibbs et des splines linéaires dans le graphique de gauche, en utilisant une régression probit linéaire dans le graphique du milieu et en utilisant une régression probit quadratique dans le graphique de droite. La figure 4 montre que le modèle de régression probit avec *p*-splines est plus souple que les modèles paramétriques pour ce qui est de dégager la courbe parmi les observations. De la figure 4 à la figure 5, les moyennes a posteriori de $\Pr(Y_i = 1 \mid \pi_i)$ ne varient pas, sauf dans les cas où les probabilités d'inclusion sont très grandes en utilisant le modèle avec *p*-splines. Cependant, les courbes des moyennes a posteriori changent considérablement lorsqu'on utilise la régression probit quadratique. Ces comparaisons indiquent que le modèle de régression probit avec *p*-splines est moins susceptible d'être affecté par les observations influentes et, donc, est un bon choix de modèle de prédiction pour l'inférence fondée sur un modèle.

7. Discussion

Les inférences bayésiennes basées sur le modèle avec *p*-splines surpassent celles obtenues avec l'estimateur HK, les estimateurs RG et les estimateurs de prédiction fondés sur un modèle linéaire dans nos simulations. Les estimateurs PBSP sont plus efficaces que les estimateurs HK et RG, et, malgré un biais empirique un peu plus grand, la couverture de leurs intervalles de crédibilité à 95 % est meilleure et la largeur moyenne de l'intervalle est plus étroite, surtout quand la proportion de population est proche de 0 ou de 1 et que peu de données provenant des queues de la distribution sont sélectionnées dans l'échantillon. Ces résultats donnent à penser que les travaux de recherche courants sur l'estimation de la prévalence d'événements rares en population finie sont importants.

L'estimateur PBSP est une extension naturelle des estimateurs de proportions de population finie fondés sur un modèle de régression linéaire ordinaire. Comparativement aux estimateurs de prédiction fondés sur un modèle linéaire, l'estimateur PBSP est robuste à l'erreur de spécification du modèle et à la présence d'observations influentes dans l'échantillon grâce à l'utilisation d'un modèle avec *p*-splines flexible, sans grande perte d'efficacité pour les tailles d'échantillon étudiées. Par conséquent, l'estimateur PBSP est facile à comprendre, mais il requiert des calculs complexes. Toutefois, grâce à WinBUGS, le logiciel statistique bayésien, il peut être implémenté facilement par les praticiens des sondages.

Tableau 4
Comparaison du biais, de la racine carrée de l'erreur quadratique moyenne, ainsi que de la largeur moyenne et du taux de non-couverture des IC à 95 % empiriques de divers estimateurs dans l'exemple des déclarations de revenus

Méthodes	biais*100		REQM*100		largeur moyenne*100		non-couverture*100	
HK	-2,4	-1,8	12,4	10,2	36	29	14,1	10,2
RL	6,7	5,5	11,9	9,2	27	21	43,5	45,6
PR	-11,6	-10,1	12,4	10,6	18	14	69,8	83,4
PR_RG1	-1,2	-0,4	11,5	8,7	31	25	22,4	16,8
PR_RG2	-1,2	-0,3	11,5	8,8	33	26	16,1	11,4
PBSP	-6,8	-2,7	9,3	5,2	27	19	14,2	5,0
PBSP_RG1	-3,0	-0,5	102,6	56,9	77	57	14,4	9,2
PBSP_RG2	-0,7	0,2	12,0	10,1	34	26	15,9	12,8

* RG_1 : estimateurs RG en utilisant l'équation (6);
RG_2 : estimateurs RG en utilisant l'équation (7).

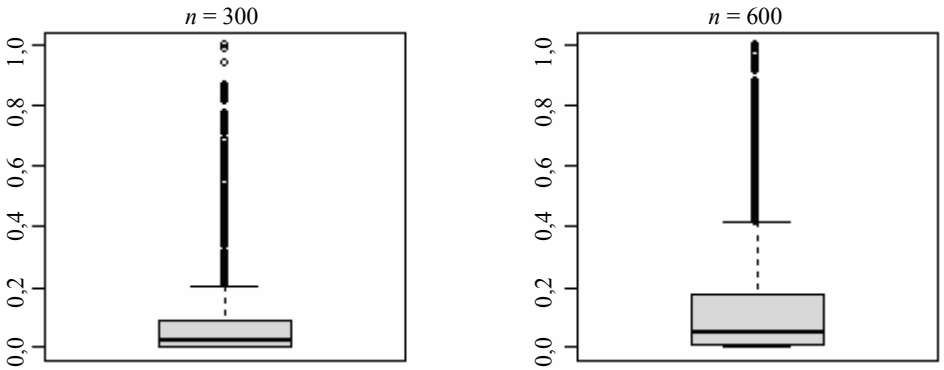

Figure 3 Boîtes à moustache pour les probabilités d'inclusion pour deux tailles d'échantillon dans l'exemple de la vérification fiscale

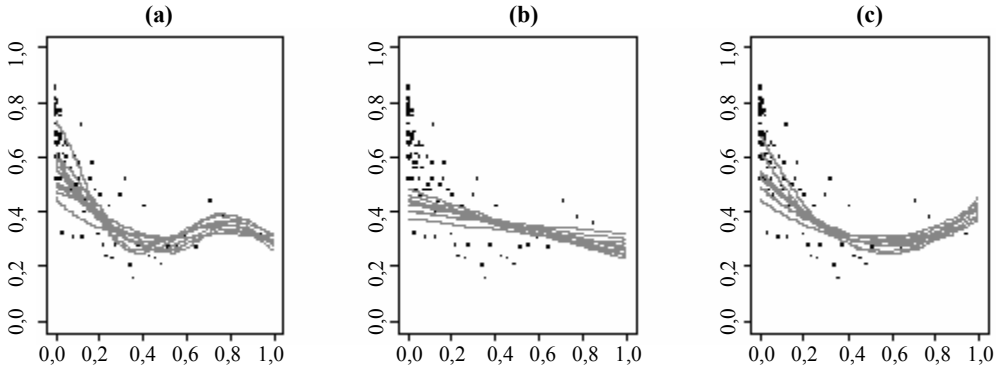

Figure 4 Prédictions basées sur les échantillons ppt uniquement dans l'exemple de la vérification fiscale, axe des X : probabilités d'inclusion π, axe des Y : $P(Y = 1|\pi)$; les points noirs représentent les probabilités réelles $P(Y = 1|\pi)$ dans chaque centile de π ; les courbes grises représentent dix réalisations des moyennes a posteriori de $P(Y = 1|\pi)$. Les modèles de prédiction sont (a) la régression probit linéaire avec *p*-splines, (b) la régression probit linéaire, (c) la régression probit quadratique

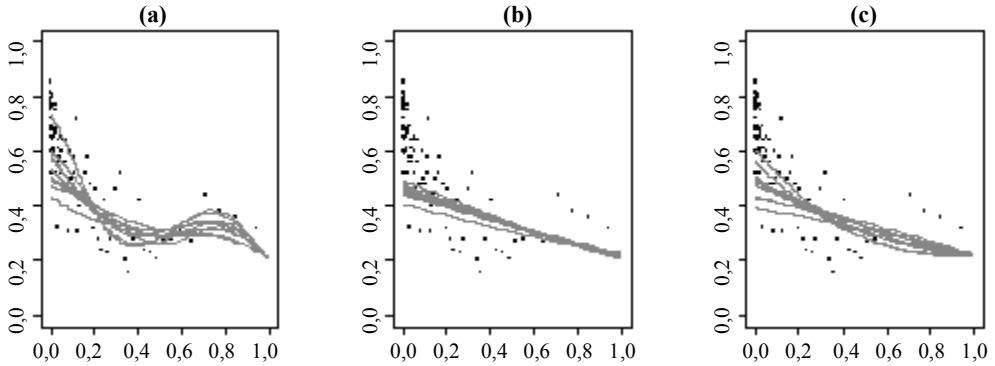

Figure 5 Prédictions fondées sur les données combinées des échantillons ppt et des observations échantillonnées avec certitude dans l'exemple de la vérification fiscale, axe des X : probabilités d'inclusion π, axe des Y : $P(Y = 1|\pi)$; les points noirs représentent les probabilités réelles $P(Y = 1|\pi)$ dans chaque centile de π ; les courbes grises représentent dix réalisations des moyennes a posteriori de $P(Y = 1|\pi)$. Les modèles de prédiction sont (a) la régression probit linéaire avec p-splines, (b) la régression probit linéaire, (c) la régression probit quadratique

Les estimateurs PBSP ne sont pas sensibles aux deux choix de loi a priori pour τ^2 considérés ici, mais l'exemple de la vérification fiscale semble indiquer que la loi a priori uniforme produit un biais et une REQM un peu plus petits, des intervalles de crédibilité à 95 % plus étroits et une meilleure couverture quant un modèle de prédiction non linéaire est nécessaire. L'exemple de la vérification fiscale montre aussi que, dans l'estimateur RG, une taille de population estimée s'appuyant sur la somme des inverses des probabilités d'inclusion est préférable à la taille réelle de population quand une ou plusieurs observations dont la probabilité d'inclusion est très faible sont incluses dans l'échantillon, puisque l'estimateur RG avec le dénominateur N possède une variance élevée et une faible efficacité dans ce cas.

Les estimateurs fondés sur le plan et leurs intervalles de confiance à 95 % peuvent fournir des inférences valides pour les proportions de population quand l'échantillon est grand. Cependant, ces propriétés asymptotiques ne semblent pas tenir quand la taille de l'échantillon est moyenne ou faible. L'approche PBSP peut donner des inférences plus valides pour les petits échantillons, surtout quand la proportion de population réelle que l'on veut estimer est proche de 0 ou de 1, quoique la couverture de l'intervalle de confiance semble être inférieure au taux nominal quand la taille d'échantillon diminue et le manque de parcimonie du modèle est un problème. Lors de l'estimation de proportions en dehors des queues de la distribution, l'estimateur PBSP produit une REQM un peu plus faible et une couverture des intervalles de confiance plus proche du taux nominal que les estimateurs HK et RG, mais l'amélioration n'est pas aussi significative que dans les queues. Sous ce scénario, pour éviter les calculs compliqués de l'estimateur PBSP, l'estimateur PR_RG basé sur l'équation (7) est une alternative pour les praticiens des sondages.

Le choix de l'estimateur de variance pose un problème pour certains plans de sondage avec probabilités inégales pour les estimateurs fondés sur le plan, mais l'approche de prédiction bayésienne avec p-splines fournit une approximation par simulation de la loi a posteriori complète de la proportion de population. Un effort supplémentaire en vue d'estimer la variance ou l'intervalle de crédibilité à 95 % de l'estimateur PBSP n'est pas nécessaire, car ceux-ci peuvent être obtenus en même temps que les estimateurs ponctuels. Zheng et Little (2005) comparent trois estimateurs de variance de l'estimateur fondé sur un modèle avec p-splines pour un total de population finie dans un échantillon ppt, y compris l'estimateur de variance bayésien empirique fondé sur un modèle, l'estimateur de variance jackknife et l'estimateur de variance par la méthode des répliques répétées équilibrées (BRR). Les études en simulation montrent que la méthode du jackknife donne de bons résultats, tandis que la méthode BRR a tendance à produire des erreurs-types prudentes et que l'estimateur bayésien empirique fondé sur un modèle est vulnérable à l'erreur de spécification de la structure de variance. Dans les travaux présentés ici, l'intervalle de crédibilité au niveau $1 - \alpha$ pour l'estimateur PBSP de la proportion de population est construit en divisant α également entre les points limites supérieur et inférieur de la loi a posteriori de p. Cette approche purement bayésienne basée sur des tirages à partir des lois a posteriori semble donner de bons résultats sous les conditions que nous avons établies et évite les calculs lourds associés aux méthodes jackknife et BRR.

L'estimateur PBSP que nous proposons ici peut être étendu en vue d'inclure des covariables auxiliaires supplémentaires en ajoutant des termes linéaires pour ces variables. Pour l'estimation par domaine, un terme d'interaction entre la fonction spline des probabilités d'inclusion et l'indicateur de domaine devrait également être modélisé. Tant les effets additifs des variables auxiliaires que l'interaction entre l'indicateur de domaine et les probabilités d'inclusion peuvent être représentés dans un modèle mixte (Ruppert et coll. 2003, page 231) et estimés en se servant de l'échantillonnage de Gibbs ou de WinBUGS (Crainiceanu et coll. 2005). L'estimateur PBSP pour les proportions de population finie peut également être étendu à un cas plus général de réponse polychotomique. L'approche de l'échantillonnage de Gibbs pour le cas binaire peut être généralisée au cas des catégories ordonnées et appliquée aux catégories non ordonnées suivant une loi multinomiale latente (Albert et Chib 1993). L'estimateur PBSP peut aussi être étendu à l'estimation sur petits domaines en combinant des effets aléatoires de petit domaine avec la fonction spline lisse sur les probabilités d'inclusion (Opsomer, Claeskens, Ranalli, Kauermann et Breidt 2008). Cette extension sera le sujet de futurs travaux de recherche.

Enfin, un examinateur a demandé si l'approche proposée peut être appliquée à une enquête polyvalente comportant de nombreux résultats, puisque la procédure de modélisation ne fournit pas un ensemble unique de poids et doit être répétée pour toutes les variables d'intérêt. Il est vrai que nos méthodes requièrent plus de calculs que les approches existantes, mais la méthode PBSP peut être mise en œuvre facilement avec un algorithme d'échantillonnage de Gibbs ou en utilisant le logiciel WinBUGS, si bien que les calculs ne constituent pas un obstacle majeur. Nous avons mentionné que les simulations décrites dans le présent article comportaient la répétition de l'analyse itérative de Gibbs 6 000 fois, si bien qu'un niveau équivalent de calculs pour une enquête unique de taille comparable permettrait la mise en œuvre de la méthode BPSP pour 6 000 résultats. Ces calculs ont été exécutés sur un PC portable ordinaire. Bien que nous ne défendions l'utilisation automatique d'aucune méthode analytique, fondée sur le plan ou sur un modèle, le fait est que la complexité des calculs n'est plus un obstacle majeur à l'application de ces méthodes. À notre avis, les propriétés statistiques d'une méthode sont plus importantes que le temps de calcul, étant donné les ressources informatiques contemporaines.

Remerciements

Les présents travaux ont été financés en partie par la société Dow Chemical par la voie d'une subvention sans restrictions accordée pour l'étude sur l'exposition à la dioxine réalisée par l'Université du Michigan. Les auteurs remercient les examinateurs et un rédacteur associé de leurs commentaires constructifs concernant la version originale du présent article.

Annexe

Algorithme de l'échantillonnage de Gibbs

Le modèle (3) peut s'écrire sous forme matricielle,

$$\Phi^{-1}(E(y_i \mid \beta, b, X, Z)) = (X\beta + Zb)_i,\ i = 1,...,n$$

$$\beta = (\beta_0, \beta_1, ..., \beta_p)^T,\ b = (b_1,...,b_m) \sim N_m(0, \tau^2 I_m)$$

$$X = \begin{pmatrix} 1 & \pi_1 & ... & \pi_1^p \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & \pi_n & ... & \pi_n^p \end{pmatrix},\ Z = \begin{pmatrix} (\pi_1 - k_1)_+^p & ... & (\pi_1 - k_m)_+^p \\ . & ... & . \\ . & ... & . \\ . & ... & . \\ (\pi_n - k_1)_+^p & ... & (\pi_n - k_m)_+^p \end{pmatrix}.$$

L'algorithme de l'échantillonnage de Gibbs pour l'estimation des paramètres du modèle (3) est le suivant :

a) le modèle de régression probit pour le résultat binaire $y = [y_1, ..., y_n]^T$ correspond à un modèle de régression normal pour des données continues latentes $y^* = [y_1^*, ..., y_n^*]^T$, qui suit une loi normale multivariée tronquée de moyenne $(X\beta + Zb)$ et de matrice de covariance identité (Albert et Chib 1993), et y_i est l'indicateur que $y_i^* > 0$. Partant de certaines valeurs initiales de (β, b), les valeurs des données continues latentes y_i^* peuvent être simulées.

b) En spécifiant une loi a priori normale aplatie appropriée $N(0,10^6)$ sur β et une loi gamma inverse $\mathrm{GI}(0{,}1, 0{,}1)$ sur τ^2, la loi a posteriori de (β, b, τ^2) sachant les données continues latentes simulées y^* est

$$\begin{aligned}(\beta, b) \mid \tau^2, y^* &\sim \mathrm{MVN}_{m+p+1}((C^T C + D/\tau^2)^{-1} C^T y^*, \\ &\qquad (C^T C + D/\tau^2)^{-1}) \\ \tau^2 \mid \beta, b &\sim \mathrm{GI}(0{,}1 + m/2,\ 0{,}1 + ||b||^2/2), \quad (11)\end{aligned}$$

où $C = [X, Z]$ et D est une matrice diagonale avec $p+1$ valeurs de 10^{-6} suivies par m de ces valeurs sur la diagonale. Gelman (2006) a recommandé d'utiliser une loi a priori uniforme sur τ, qui résulte en la loi a posteriori pour τ^2 de la forme

$$\tau^2 \mid \beta, b \sim \mathrm{GI}((m-1)/2, ||b||^2/2). \quad (12)$$

c) À l'itération t, des tirages de $(\beta^{(t)}, b^{(t)}, \tau^{2(t)})$ à partir de la loi a posteriori donnée par l'équation (11) ou

(12) sont utilisés pour générer de nouvelles données latentes $\hat{y}^{*(t)}$, sachant la variable binaire observée y pour l'échantillon, et pour obtenir les valeurs prédites a posteriori $\hat{y}^{(t)}$ pour les unités non échantillonnées. Nous pouvons alors obtenir des tirages à partir de la loi a posteriori de la proportion de population finie à l'itération t sous la forme

$$\hat{p}_{\mathrm{PR}}^{(t)} = N^{-1}\left(\sum_{i\in s} y_i + \sum_{j\notin s} \hat{y}_j^{(t)}\right).$$

Bibliographie

Albert, J.H., et Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of American Statistical Association*, 88, 669-679.

Basu, D. (1971). An essay on the logical foundations of survey sampling. Partie 1, dans *Foundations of Statistical Inference*, (Éds., V.P. Godambe et D.A. Sprott), Toronto : Holt, Rinehart and Winston, 203-242.

Compumine (2007). Re: analysis – Tax audit data mining. Février 2007. http://www.compumine.com/web/public/newsletter/20071/tax-audit-data-mining.

Crainiceanu, C.M., Ruppert, D. et Wand, M. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, 14, 2005, 14.

Duchesne, P. (2003). Estimation of a proportion with survey data. *Journal of Statistics Education*, 11, 3.

Eilers, P.H.C., et Marx, B.D. (1996). Flexible smoothing with *B*-splines and penalties (avec discussion). *Statistical Science*, 11, 89-121.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3, 515-533.

Hartley, H.O., et Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Horvitz, D.G., et Thompson, M.E. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830.

Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-673.

Lehtonen, R., et Veijanen, A. (1998). Estimateurs de régression généralisés logistiques. *Techniques d'enquête*, 24, 53-58.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

Montanari, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 71-79.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. et Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society*, Séries B, 70, 265-286.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Ruppert, D., Wand, M.P. et Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge, UK : Cambridge University Press.

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

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

Smith, T.M.F. (1976). The foundations of survey sampling: A review (avec discussion). *Journal of the Royal Statistical Society*, Séries A, 139, 183-204.

Smith, T.M.F. (1994). Sample surveys 1975–1990: An age of reconciliation? (avec discussion). *Revue Internationale de Statistique*, 62, 5-34.

Yates, F., et Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Séries B, 15, 235-261.

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

Zheng, H., et Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.