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

Article

Inférence bayésienne pour les quantiles de population finie sous échantillonnage avec probabilités inégales

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

Décembre 2012

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Comment accéder à ce produit

Le produit nº 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Ce produit est aussi disponible en version imprimée standard au prix de 30 $CAN l'exemplaire et de 58 $CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	10 $CAN	20 $CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste Statistique Canada
Finances
Immeuble R.-H.-Coats, 6e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
p	provisoire
r	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
E	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 221-233
Statistique Canada, N° 12-001-X au catalogue

Inférence bayésienne pour les quantiles de population finie sous échantillonnage avec probabilités inégales

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

Résumé

Le présent article décrit l'élaboration de deux méthodes bayésiennes d'inférence au sujet des quantiles de variables d'intérêt continues d'une population finie sous échantillonnage avec probabilités inégales. La première de ces méthodes consiste à estimer les fonctions de répartition des variables étudiées continues en ajustant un certain nombre de modèles de régression probit avec splines pénalisées sur les probabilités d'inclusion. Les quantiles de population finie sont alors obtenus par inversion des fonctions de répartition estimées. Cette méthode demande considérablement de calculs. La deuxième méthode consiste à prédire les valeurs pour les unités non échantillonnées en supposant qu'il existe une relation variant de façon lisse entre la variable étudiée continue et la probabilité d'inclusion, en modélisant la fonction moyenne ainsi que de la fonction de variance en se servant de splines. Les deux estimateurs bayésiens fondés sur un modèle avec splines donnent un compromis désirable entre la robustesse et l'efficacité. Des études par simulation montrent que les deux méthodes produisent une racine carrée de l'erreur quadratique moyenne plus faible que l'estimateur pondéré par les poids de sondage et que les estimateurs par le ratio et par différence décrits dans Rao, Kovar et Mantel (RKM 1990), et qu'ils sont plus robustes à la spécification incorrecte du modèle que l'estimateur fondé sur un modèle de régression passant par l'origine décrit dans Chambers et Dunstan (1986). Lorsque la taille de l'échantillon est petite, les intervalles de crédibilité à 95 % des deux nouvelles méthodes ont une couverture plus proche du niveau nominal que l'estimateur pondéré par les poids de sondage.

Mots clés : Analyse bayésienne ; fonction de répartition ; erreurs hétéroscédastiques ; régression avec splines pénalisées ; échantillons.

1. Introduction

Nous considérons l'inférence pour les quantiles d'une variable continue d'une population finie d'après un échantillon sélectionné avec probabilités inégales. Les quantiles de population finie sont habituellement estimés par les quantiles pondérés par les poids de sondage, c'est-à-dire un estimateur de type Horvitz-Thompson. Souvent, dans les sondages, la variable de plan de sondage (ici, la probabilité d'inclusion) ou une variable auxiliaire corrélée est mesurée sur des unités non échantillonnées, et cette information peut être utilisée pour accroître l'efficacité des estimateurs pondérés par les poids de sondage (Zheng et Little 2003 ; Chen, Elliott et Little 2010).

Les méthodes d'utilisation d'information auxiliaire pour estimer les fonctions de répartition en population finie ont fait l'objet d'études approfondies. Chambers et Dunstan (1986) ont proposé une méthode fondée sur un modèle et ont illustré leur approche au moyen d'un modèle de régression linéaire avec ordonnée à l'origine nulle pour une superpopulation. Dans la suite de l'exposé, nous donnons à cet estimateur le nom d'estimateur CD. Dorfman et Hall (1993) ont appliqué l'approche CD en remplaçant le modèle de régression linéaire par un modèle non paramétrique. Lombardía, González-Manteiga et Prada-Sánchez (2003, 2004) ont proposé une approximation par le bootstrap de ces estimateurs fondée sur le rééchantillonnage d'une version lissée de la distribution empirique des résidus. Kuk et Welsh (2001) ont également modifié l'approche CD pour résoudre la question des écarts par rapport au modèle en estimant la distribution conditionnelle des résidus sous forme d'une fonction de la variable auxiliaire. Rao, Kovar et Mantel (RKM 1990) ont démontré les avantages des estimateurs par le ratio et par différence fondés sur le plan de sondage par rapport à l'estimateur CD quand le modèle est mal spécifié. Wang et Dorfman (1996) ont proposé une moyenne pondérée des estimateurs CD et RKM. Kuk (1993) a proposé un estimateur à noyau qui combine la distribution connue de la variable auxiliaire avec une estimation par la méthode du noyau de la distribution conditionnelle de la variable étudiée sachant la valeur de la variable auxiliaire. Chambers, Dorfman et Wehrly (1993) ont proposé un estimateur fondé sur un modèle avec lissage par noyau, et Wu et Sitter (2001), ainsi que Harms et Duchesne (2006) ont proposé des estimateurs par calage.

La recherche sur l'utilisation d'information auxiliaire pour l'inférence au sujet des quantiles de population finie (définis comme étant l'inverse de la fonction de répartition) est plus limitée. Chambers et Dunstan (1986) ont discuté de l'estimation en prenant l'inverse de l'estimateur CD de la fonction de répartition, mais n'ont pas comparé les propriétés de cet estimateur des quantiles à d'autres options. Rao et coll. (1990) ont proposé de simples estimateurs par le ratio et par différence des quantiles qui étaient beaucoup

1. Qixuan Chen, professeur adjoint, Department of Biostatistics, Columbia University Mailman School of Public Health, 722 West 168 Street, New York, NY 10032. Courriel : qc2138@columbia.edu ; Michael R. Elliott et Roderick J.A. Little, professeurs, 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.

plus efficaces que l'estimateur pondéré par les poids de sondage quand la variable d'intérêt résultant de l'enquête était approximativement proportionnelle à la variable auxiliaire.

Nous supposons ici que l'on procède à un échantillonnage avec probabilités inégales où les probabilités d'inclusion sont connues pour toutes les unités de la population. Nous élaborons deux estimateurs bayésiens fondés sur un modèle avec splines des quantiles de population finie dans lequel sont intégrées les probabilités d'inclusion. La première méthode consiste à estimer la fonction de répartition d'un certain nombre de valeurs d'échantillon en utilisant des estimateurs prédictifs bayésiens avec splines pénalisées (Chen et coll. 2010). Nous estimons ensuite les quantiles de population finie en prenant l'inverse de la fonction de répartition prédictive. La deuxième méthode consiste à utiliser un estimateur prédictif bayésien avec splines pénalisées à deux moments, qui prédit les valeurs des unités non échantillonnées en se basant sur un modèle normal, dont la moyenne et la variance sont toutes deux modélisées au moyen de splines pénalisées sur les probabilités d'inclusion. Nous comparons la performance de ces deux nouvelles méthodes à celle de l'estimateur pondéré par les poids de sondage, de l'estimateur CD et des estimateurs par le ratio et par la différence de RKM, en réalisant des études par simulation sur des données générées artificiellement et sur des données d'enquêtes agricoles.

2. Estimateurs des quantiles

Soit s un échantillon aléatoire de taille n tiré avec probabilités inégales de la population finie de N unités identifiables selon les probabilités d'inclusion $\{\pi_i, i = 1, ..., N\}$ que l'on suppose être connues pour toutes les unités avant qu'un échantillon soit tiré. Soit Y une variable étudiée continue, pour laquelle les valeurs $\{y_1, y_2, ..., y_n\}$ sont observées dans l'échantillon aléatoire s. L'α-quantile de Y dans la population finie est défini comme étant :

$$\theta(\alpha) = \inf\left\{t;\ N^{-1}\sum_{i=1}^{N} \Delta(t - y_i) \geq \alpha\right\}, \quad (1)$$

où $\Delta(u) = 1$ quand $u \geq 0$ et $\Delta(u) = 0$ autrement. On estime souvent $\theta(\alpha)$ en utilisant l'α-quantile pondéré par les poids de sondage $\hat{\theta}(\alpha) = \inf\{t, \hat{F}_w(t) \geq \alpha\}$, où $\hat{F}_w(t)$ est la fonction de répartition pondérée par les poids de sondage donnée par

$$\hat{F}_w(t) = \sum_{i\in s} \pi_i^{-1}\Delta(t - y_i) \Big/ \sum_{i\in s} \pi_i^{-1}.$$

Woodruff (1952) a proposé une méthode de calcul des limites de confiance pour l'α-quantile pondéré par les poids de sondage. En premier lieu, on obtient une pseudo-population en pondérant chaque unité de l'échantillon par son poids de sondage ; on estime l'écart-type du pourcentage d'unités inférieur à l'α-quantile estimé ; on multiplie ensuite l'écart-type estimé par le centile z approprié, puis on l'ajoute et on le soustrait de α pour construire les limites de confiance pour le pourcentage d'unités inférieures à l'α-quantile estimé. Enfin, les valeurs de la variable étudiée correspondant aux limites de confiance du pourcentage d'unités inférieures à l'α-quantile estimé sont lues sur les unités pondérées de la pseudo-population rangées par ordre de taille. L'estimation de la variance du pourcentage d'unités de la pseudo-population dont la valeur est inférieure à l'α-quantile estimé est discutée dans Woodruff (1952). Sitter et Wu (2001) ont montré que les intervalles de Woodruff donnent de bons résultats, mêmes dans les cas modérés à extrêmes des queues de la fonction de répartition. Une autre estimation de la variance a été établie par Francisco et Fuller (1991) en utilisant une version lissée de la version du test de signification en grand échantillon.

2.1 Approche fondée sur un modèle bayésien avec inversion de la fonction de répartition estimée

La fonction quantile de population finie est l'inverse de la fonction de répartition (FR) de population finie, définie comme étant $F(t) = N^{-1}\sum_{i=1}^{N} \Delta(t - y_i)$, où $\Delta(x) = 1$ quand $x \geq 0$ et $\Delta(x) = 0$ ailleurs. Nous pouvons estimer les quantiles de population finie en commençant par construire une estimation prédictive continue et strictement monotone de $F(t)$, en traitant $\Delta(t - y)$ comme une variable de résultat binaire et en appliquant des méthodes d'estimation des proportions en population finie.

En particulier, Chen et coll. (2010) ont proposé un estimateur prédictif bayésien avec splines pénalisées (PBSP) pour les proportions de population finie sous échantillonnage avec probabilités inégales. Ils font la régression de la variable étudiée binaire z sur les probabilités d'inclusion dans l'échantillon, en utilisant le modèle de régression probit avec splines pénalisées (2) avec m nœuds fixes présélectionnés :

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

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

Des unités autoreprésentatives sont incluses en prenant $\pi_i = 1$. En supposant que les lois a priori pour β et τ^2 sont non informatives, ils ont simulé des tirages de z pour les unités non échantillonnées à partir des lois prédictives a posteriori. Un tirage à partir de la loi a posteriori de la proportion de population finie s'obtient alors en calculant la moyenne des unités échantillonnées observées et des tirages d'unités non échantillonnées. Le procédé est répété de nombreuses fois pour simuler la loi a posteriori de la proportion

de population finie. Des études par simulation ont indiqué que l'estimateur PBSP est plus efficace que l'estimateur pondéré par les poids de sondage et que l'estimateur par la régression généralisée de la proportion de population finie, avec une couverture des intervalles de confiance plus proches des niveaux nominaux.

Nous employons l'approche PBSP n fois pour estimer $F(t)$ à chacune des valeurs échantillonnées de y, $t = \{y_1, y_2, ..., y_n\}$. Cet estimateur ne tient pas compte du fait que nous estimons une fonction de répartition complète, et il ne s'agit pas nécessairement d'une fonction monotone. En outre, l'interpolation linéaire des n fonctions de répartition estimées peut mener à une mauvaise estimation de la fonction de répartition de la population finie. Pour contourner ces deux problèmes, nous ajustons une courbe de régression cubique lisse aux n fonctions de répartition estimées en imposant des contraintes de monotonicité (Wood 1994). Nous désignons la fonction de répartition estimée résultante par $\hat{F}(t)$. L'estimateur fondé sur un modèle bayésien de $\theta(\alpha)$, obtenu par inversion de la fonction de répartition (FR), est alors défini comme il suit :

$$\hat{\theta}_{\text{inv-FR}}(\alpha) = \inf\{t;\ \hat{F}(t) \geq \alpha\}. \qquad (3)$$

Nous ajustons également deux autres courbes de régression lisse monotone aux limites supérieures et inférieures des intervalles de crédibilité (IC) à 95 % de ces fonctions de répartition estimées, désignées par $\hat{F}_U(t)$ et $\hat{F}_L(t)$. Afin de réduire le temps de calcul dans nos études par simulation, nous estimons uniquement la fonction de répartition à $k < n$ points présélectionnés de l'échantillon.

L'idée fondamentale qui sous-tend cette approche est illustrée graphiquement à la figure 1. Supposons qu'un échantillon de taille 100 est tiré d'une population finie. Nous choisissons 20 observations dans l'échantillon et estimons les fonctions de répartition respectives et les IC à 95 % associés en utilisant l'estimateur PBSP. À la figure 1(a), nous représentons les estimations PBSP pour ces 20 observations par des points noirs et les limites inférieure et supérieure de l'IC à 95 %, par des signes « - » que nous relions par un trait plein. À la figure 1(b), nous ajoutons trois courbes de prédiction lisses monotones en utilisant un trait plein noir pour l'estimation ponctuelle et des traits pointillés noirs pour les limites supérieure et inférieure des IC à 95 %.

À la figure 1(c), nous traçons à travers le graphique une droite horizontale passant par la valeur α sur l'axe des y. Nous lisons x_A, x et x_B respectivement sur l'axe des x de façon telle que $\hat{F}_L(x_A) = \alpha$, $\hat{F}(x) = \alpha$ et $\hat{F}_U(x_B) = \alpha$. Alors, x est l'estimation bayésienne avec inversion de la fonction de répartition de $\theta(\alpha)$. Si l'IC à 95 % de la fonction de répartition $F(\cdot)$ est construit en divisant en parties égales les queues de la distribution a posteriori, l'intervalle formé par x_A et x_B est un IC à 95 % de $\theta(\alpha)$. La preuve en est la suivante : si α est la limite inférieure de l'IC à 95 % de $F(x_A)$, seulement 2,5 % des tirages de $F(x_A)$ dans la distribution a posteriori sont plus petits que α. C'est-à-dire que

$$\Pr(F^{-1}(\alpha) > F^{-1}(F(x_A))) \equiv \Pr(\theta(\alpha) > x_A) = 0,025.$$

De même si α est la limite supérieure de l'IC à 95 % de $F(x_B)$, $\Pr(\theta(\alpha) < x_B) = 0,975$. Par conséquent, il y a une probabilité de 95 % que $\theta(\alpha)$ soit compris entre x_A et x_B dans la distribution a posteriori, étant donné l'échantillon.

Cette approche fondée sur un modèle bayésien avec inversion de la fonction de répartition permet d'éviter de fortes hypothèses de modélisation et peut être appliquée à des distributions normales ou asymétriques. L'estimation de la fonction de répartition à chacune des n unités échantillonnées permet d'utiliser complètement l'information fournie par l'échantillon, mais requiert d'importants calculs ; l'estimation de la fonction de répartition à $k < n$ valeurs réduit le temps de calcul au prix d'une certaine perte d'efficacité. Dans l'approche classique, les quantiles de population sont estimés par inversion de la fonction de répartition empirique non lissée. Nous recommandons d'ajuster une courbe de régression cubique lisse aux fonctions de répartition estimées avant d'inverser la fonction de répartition estimée résultante. Les estimations résultantes des quantiles sont plus efficaces, parce que la courbe lisse exploite l'information provenant de toutes les données. Des simulations dont les résultats ne sont pas présentés ici donnent à penser que la courbe de la fonction de répartition estimée en se basant sur un sous-ensemble bien choisi de k unités échantillonnées est similaire à celle estimée en se basant sur la totalité des unités échantillonnées, mais le temps de calcul est réduit considérablement.

Nous suggérons de choisir le sous-ensemble de k points de données à intervalles égaux dans le milieu de la distribution, et à intervalles plus fréquents dans les extrémités afin d'améliorer l'estimation de la fonction de répartition dans les queues. Par exemple, dans notre étude par simulation avec un échantillon de taille 100, nous avons estimé les fonctions de répartition à 20 points : les 3 valeurs les plus faibles, les 3 valeurs les plus grandes et 14 autres points uniformément espacés dans le milieu de l'échantillon rangé par ordre de valeur.

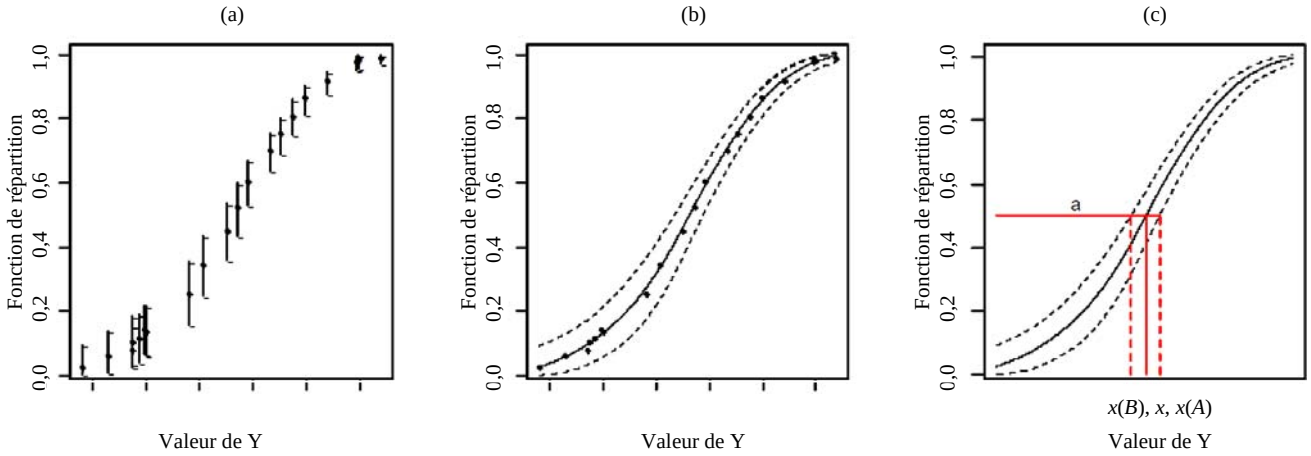

Figure 1 Approche fondée sur un modèle bayésien avec inversion de la fonction de répartition (FR) pour estimer les fonctions de répartition de population finie et les quantiles associés, illustrée en utilisant un échantillon de taille 100 tiré d'une population finie. (a) La méthode PBSP est utilisée pour estimer les fonctions de répartition de la population finie à vingt points de l'échantillon ; les points représentent les estimateurs PBSP et les signes moins représentent les limites supérieure et inférieure des IC à 95 %. (b) Trois modèles de régression cubiques lisses monotones sont ajustés sur les estimateurs PBSP, les limites supérieures et les limites inférieures, respectivement ; la courbe en trait plein représente les fonctions de répartition continues prédictives et les deux courbes en trait interrompu représentent les IC à 95 % des fonctions de répartition. (c) L'estimation ponctuelle et l'IC à 95 % de l'α- quantile de population sont obtenus en inversant la fonction de répartition estimée ; x est l'estimation ponctuelle et x(B) et x(A) sont les limites inférieure et supérieure de l'IC à 95 %

2.2 Approche prédictive bayésienne avec deux moments modélisés par splines pénalisées

Nous considérons d'autres estimateurs des quantiles de population finie de la forme :

$$\tilde{\theta}(\alpha) = \inf\left\{t;\ N^{-1}\left(\sum_{i\in s}\Delta(t-y_i)+\sum_{j\notin s}\Delta(t-\hat{y}_j)\right)\geq\alpha\right\}, \quad (4)$$

où $\hat{y}_j$ est la valeur de la j^{e} unité non échantillonnée prédite par une régression sur les probabilités d'inclusion $\{\pi_i\}$. Un modèle normal de base pour un résultat continu repose sur l'hypothèse d'une fonction moyenne linéaire en $\{\pi_i\}$, c'est-à-dire :

$$Y_i \overset{\text{ind}}{\sim} N(\beta_0+\beta_1\pi_i, c_i\sigma^2), \quad (5)$$

avec des constantes c_i connues pour modéliser la variance non constante. Ce modèle donne une estimation biaisée de $\theta(\alpha)$ si la relation n'est pas linéaire. Pour estimer les totaux de population finie, Zheng et Little (2003, 2005) ont remplacé dans (5) la fonction moyenne linéaire par une spline pénalisée, et ont supposé que $c_i = \pi_i^{2k}$ pour une certaine valeur connue de k. Des simulations ont donné à penser que leur estimateur fondé sur un modèle du total de population finie donne de meilleurs résultats que l'estimateur pondéré par les poids de sondage, même quand la structure de variance est mal spécifiée.

Pour l'estimation des quantiles au lieu du total, il est important de spécifier correctement la structure de la variance afin d'éviter un biais. Par conséquent, nous étendons le modèle avec spline pénalisée de Zheng et Little (2003) en modélisant la moyenne ainsi que la variance en utilisant des splines pénalisées. Le modèle avec deux moments modélisés par splines pénalisées peut s'écrire (Ruppert, Wand et Carroll 2003, page 264) :

$$Y_i \overset{\text{ind}}{\sim} N(\text{SPL}_1(\pi_i,k), \exp(\text{SPL}_2(\pi_i,k'))),$$

$$\text{SPL}_1(\pi_i,k) = \beta_0 + \sum_{k=1}^{p}\beta_k\pi_i^k + \sum_{l=1}^{m_1} b_l(\pi_i-k_l)_+^p,$$

$$b_l \overset{\text{iid}}{\sim} N(0,\tau_b^2),$$

$$\text{SPL}_2(\pi_i,k') = \alpha_0 + \sum_{k=1}^{p}\alpha_k\pi_i^k + \sum_{l=1}^{m_2} \nu_l(\pi_i-k'_l)_+^p,$$

$$\nu_l \overset{\text{iid}}{\sim} N(0,\tau_\nu^2). \quad (6)$$

Dans (6), la moyenne et le logarithme de la variance sont modélisés par des splines pénalisées (SPL_1) et (SPL_2) sur $\{\pi_i\}$. La modélisation du logarithme de la variance fait en sorte que les estimations de la variance soient positives. Nous permettons des nombres (m_1, m_2) et des emplacements (k, k') différents des nœuds pour les deux splines.

Ruppert et coll. (2003) ont proposé une approche itérative pour estimer les paramètres de (6). Ils ont d'abord supposé que SPL_2 était connue et ont ajusté un modèle linéaire mixte pour estimer les paramètres dans SPL_1. Ils ont calculé le carré de la différence entre Y et SPL_1, qui suivait une loi gamma de paramètre de forme ½ et de paramètre d'échelle $2\mathrm{SPL}_2$. Ils ont ensuite ajusté un modèle linéaire mixte généralisé pour les carrés des différences afin d'estimer les paramètres dans SPL_2. Ils ont itéré les procédés susmentionnés jusqu'à ce que les estimations des paramètres convergent. Cette approche itérative est simple à mettre en œuvre. Cependant, ici, notre objectif n'est pas d'estimer les paramètres, mais d'obtenir des prédictions bayésiennes de Y pour les unités non échantillonnées afin de pouvoir utiliser (4) pour estimer les quantiles.

Crainiceanu, Ruppert, Carroll, Joshi et Goodner (2007) ont élaboré une méthodologie inférencielle bayésienne pour (6). Ils ont constaté que la mise en œuvre de la méthode MCMC en utilisant des pas de Metropolis-Hastings multivariés est instable avec de mauvaises propriétés de mélange. Ils ont proposé d'ajouter des termes d'erreur à la deuxième spline pour rendre les calculs plus faisables, en remplaçant l'échantillonnage à partir de lois conditionnelles complètes complexes par de simples pas de Metropolis-Hastings univariés. Cette idée peut s'exprimer comme

$$Y_i \overset{\text{ind}}{\sim} N(\mathrm{SPL}_1(\pi_i, k), \sigma_\varepsilon^2(\pi_i)),$$

$$\log(\sigma_\varepsilon^2(\pi_i)) \overset{\text{iid}}{\sim} N(\mathrm{SPL}_2(\pi_i, k'), \sigma_A^2).$$

Nous avons utilisé une loi a priori $N(0, 10^6)$ pour les paramètres à effets fixes β et α, et une loi a priori gamma inverse propre $\mathrm{IGamma}(10^{-6}, 10^{-6})$ pour les composantes de la variance τ_b^2 et τ_v^2. Nous avons fixé les valeurs de $\sigma_A^2 = 0{,}1$. Les lois conditionnelles a posteriori complètes sont décrites en détail dans Crainiceanu et coll. (2007).

La loi a posteriori de l'α-quantile de population finie est simulée en générant un grand nombre D de tirages et en utilisant l'estimateur prédictif de la forme

$$\tilde{\theta}^{(d)}(\alpha) = \inf\left\{t;\ N^{-1}\left(\sum_{i\in s}\Delta(t - y_i) + \sum_{j\notin s}\Delta(t - \hat{y}_j^{(d)})\right) \geq \alpha\right\},$$

où $\hat{y}_j^{(d)}$ est un tirage à partir de la loi prédictive a posteriori de la j^{e} unité non échantillonnée de la variable résultat continue. La moyenne de ces tirages simule l'estimateur prédictif bayésien avec deux moments modélisés par splines pénalisées (PB2SP) de l'α-quantile de population finie,

$$\hat{\theta}_{\mathrm{PB2SP}}(\alpha) = D^{-1}\sum_{d=1}^{D}\tilde{\theta}^{(d)}(\alpha).$$

L'intervalle de crédibilité à 95 % bayésien pour l'α-quantile de population dans les simulations est formé en divisant également la queue de la distribution entre les points finaux supérieur et inférieur.

3. Étude par simulation

3.1 Étude par simulation avec données artificielles

Nous avons d'abord simulé une superpopulation de taille $M = 20\,000$. La variable de taille X dans la superpopulation prend 20 000 valeurs entières consécutives allant de 710 à 20 709. Puis, nous avons tiré une population finie de taille $N = 2\,000$ de cette superpopulation par échantillonnage systématique avec probabilité proportionnelle à la taille (ppt) où la probabilité était proportionnelle à l'inverse de la variable de taille. Par conséquent, dans la population finie, la distribution de la variable de taille est asymétrique avec étalement à droite. La variable résultat étudiée Y a été tirée d'une loi normale de moyenne $f(\pi)$ et de variance d'erreur égale à 0,04 (erreur homoscédastique) ou π (erreur hétéroscédastique). Trois structures de moyenne $f(\pi)$ ont été simulées : pas d'association entre Y et π (NULL) $f(\pi) = 0{,}5$, une association linéaire (LINUP) $f(\pi) = 6\pi$, et une association non linéaire (EXP) $f(\pi) = \exp(-4{,}64 + 52\pi)$. Pour chacune des six conditions de simulation, nous avons généré un millier de répliques de la population finie et nous avons tiré de chaque population un échantillon ppt systématique ($n = 100$) avec x comme variable de taille ; donc $\pi_i = nx_i / \sum_{j=1}^{N} x_j$. Les nuages de points de Y en fonction de π pour ces six populations sont présentés à la figure 2.

Nous avons comparé les propriétés de l'estimateur bayésien avec fonction de répartition inverse et de l'estimateur bayésien PB2SP à cinq autres approches :

a) PS, l'estimateur pondéré par les poids de sondage défini par inversion de $\hat{F}_w$;
b) PS lisse, l'estimateur pondéré par les poids de sondage lisse. Une courbe de régression cubique lisse a été ajustée à $\hat{F}_w$ et désignée par $\tilde{F}_w$. L'estimateur pondéré par les poids de sondage lisse est alors défini comme $\tilde{\theta}_w = \inf\{t;\ \tilde{F}_w \geq \alpha\}$;
c) CD, l'estimateur de Chambers et Dunstan (1986), en supposant le modèle suivant : $Y_i = \beta\pi_i + \sqrt{\pi_i}U_i$, où U_i est une variable aléatoire dont les valeurs sont indépendantes et identiquement distribuées de moyenne nulle ;
d) Ratio, l'estimateur par le ratio de RKM (1990) donné par $\{\hat{\theta}_y(\alpha)/\hat{\theta}_x(\alpha)\}\times\theta_x(\alpha)$, où $\hat{\theta}_y(\alpha)$ et $\hat{\theta}_x(\alpha)$ désignent respectivement les estimations pondérées par les poids de sondage pour Y et la

variable de taille X, et $\theta_x(\alpha)$ est le quantile de population connu de X ;

e) Diff, l'estimateur par la différence de RKM (1990) donné par $\hat{\theta}_y(\alpha) + \hat{R} \times \{\theta_x(\alpha) - \hat{\theta}_x(\alpha)\}$, où $\hat{R}$ est l'estimation pondérée par les poids de sondage de Y/X.

Les sept estimateurs pour les 10^e, 25^e, 50^e, 75^e et 90^e centiles de la population finie ont été comparés pour ce qui est du biais empirique et de la racine carrée de l'erreur quadratique moyenne (REQM). Étant donné la complexité de l'estimation de la variance des estimateurs CD et RKM, nous avons comparé seulement la largeur moyenne et le taux de non-couverture de l'intervalle de confiance/crédibilité (IC) à 95 % pour les deux estimateurs fondés sur un modèle bayésien et l'estimateur pondéré par les poids de sondage. Pour l'IC à 95 %, nous avons utilisé la méthode de Woodruff pour l'estimateur pondéré par les poids de sondage, la méthode illustrée à la figure 1(c) pour l'estimateur bayésien avec fonction de répartition inverse et la probabilité a posteriori de 95 % du quantile avec queues égales pour l'estimateur PB2SP. Nous avons utilisé des splines cubiques avec 15 nœuds également espacés.

Les tableaux 1 et 2 montrent le biais empirique et la REQM pour les trois distributions normales avec erreurs homoscédastiques et erreurs hétéroscédastiques, respectivement. Dans l'ensemble, le biais empirique dans l'estimation des cinq quantiles est semblable lorsque l'on utilise les estimateurs bayésiens, les deux estimateurs pondérés par les poids de sondage et les deux estimateurs fondés sur le plan de sondage de RKM. Par contre, l'estimateur CD produit un grand biais et une grande REQM dans tous les scénarios, sauf LINUP avec erreur hétéroscédastique, où le modèle sous-jacent de l'estimateur est spécifié correctement. Les deux estimateurs fondés sur un modèle bayésien produisent des racines carrées de l'erreur quadratique moyenne plus petites que les autres estimateurs, et cet accroissement de l'efficacité est important dans certains scénarios, en particulier lorsque l'on utilise l'estimateur PB2SP. Par l'application d'une courbe de régression cubique lisse à la fonction de répartition empirique estimée pondérée par les poids de sondage, l'estimateur pondéré par les poids de sondage lisse produit un gain d'efficacité par rapport aux estimateurs pondérés par les poids de sondage classiques, mais la REQM demeure plus grande que pour l'estimateur bayésien avec fonction de répartition inverse. Les comparaisons des trois estimateurs fondés sur le plan de sondage donnent à penser qu'aucun de ces estimateurs ne domine uniformément les deux autres. En particulier, l'estimateur pondéré par les poids de sondage a une plus petite REQM que les estimateurs par différence et par le ratio de RKM pour les cinq quantiles dans la population NULL et pour les quantiles inférieurs dans les populations LINUP et EXP ; par ailleurs, les estimateurs RKM ont une plus petite REQM pour les quantiles supérieurs dans les populations LINUP et EXP.

Le tableau 3 donne la largeur moyenne et le taux de non-couverture de l'IC à 95 % pour les deux estimateurs fondés sur un modèle bayésien et l'estimateur pondéré par les poids de sondage. Dans l'ensemble, les deux estimateurs fondés sur un modèle bayésien donnent de plus courtes largeurs moyennes de l'IC à 95 % que l'estimateur pondéré par les poids de sondage. Le taux de couverture de l'IC à 95 % est comparable pour les trois estimateurs, excepté quand α est égal à 0,1, auquel cas l'IC à 95 % de l'estimateur PB2SP possède la largeur moyenne la plus courte et une très bonne couverture, tandis que l'estimateur pondéré par les poids de sondage présente une sous-couverture importante. Cette situation est due au fait que la méthode de Woodruff pour estimer la variance de l'estimateur pondéré par les poids de sondage est fondée sur une hypothèse de grand échantillon, alors qu'ici l'échantillonnage ppt fait qu'un petit nombre seulement de cas sont échantillonnés dans la queue inférieure de la distribution.

Bien que l'estimateur pondéré par les poids de sondage se comporte de manière comparable aux estimateurs fondés sur un modèle bayésien avec splines pour ce qui est du biais empirique global, le biais conditionnel des estimations varie considérablement à mesure qu'augmente la moyenne d'échantillon des probabilités d'inclusion. À l'exemple de Royall et Cumberland (1981), nous avons classé les estimations provenant des 1 000 échantillons en fonction de la moyenne d'échantillon des probabilités d'inclusion et nous les avons réparties en 20 groupes de 50, puis nous avons calculé le biais empirique pour chaque groupe. La figure 3 donne le biais conditionnel des deux estimateurs bayésiens et de l'estimateur pondéré par les poids de sondage pour le 90^e centile dans le cas « EXP + erreur homoscédastique ». La figure 3 montre une tendance linéaire du biais dans l'estimateur pondéré par les poids de sondage à mesure qu'augmente la moyenne d'échantillon des probabilités d'inclusion, tandis que le biais de groupe des deux estimateurs fondés sur un modèle bayésien avec splines est moins affecté par cette moyenne. Des constatations comparables sont faites pour d'autres scénarios.

3.2 Étude par simulation avec les données de l'enquête sur les exploitations agricoles à grande échelle

L'estimateur PB2SP repose sur l'hypothèse que la variable résultat suit une loi normale, après conditionnement sur les probabilités d'inclusion. Puisque l'approche fondée sur un modèle bayésien avec fonction de répartition inverse ne comporte pas d'hypothèse de normalité, nous pourrions

nous attendre à ce qu'elle donne de meilleurs résultats que l'approche PB2SP lorsque l'hypothèse de normalité est violée. Cela motive une comparaison de l'estimateur pondéré par les poids de sondage et de l'estimateur bayésien avec fonction de répartition inverse pour des données ne suivant pas une loi normale.

La population considérée ici est définie par 398 exploitations agricoles à grande échelle (qui produisent des céréales, des bovins, des moutons et de la laine) ayant une superficie agricole de 6 000 hectares ou moins qui ont participé à l'Australian Agricultural and Grazing Industries Survey de 1982 réalisée par l'Australian Bureau of Agricultural and Resource Economics (ABARE 2003). La variable Y est le total des recettes monétaires agricoles. Nous avons tiré 1 000 échantillons systématiques ppt de taille égale à 100 en prenant la superficie agricole, X, comme variable de taille, de sorte que les grandes exploitations agricoles sont plus susceptibles que les autres d'être sélectionnées dans l'échantillon. La figure 4 donne le nuage de points de Y en fonction de la variable de taille X pour ces exploitations, chaque cercle plein représentant un échantillon ppt sélectionné. Ce graphique montre que la variation de Y augmente à mesure que X augmente. En outre, la distribution de Y est étalée vers la droite étant donné X. Nous avons réalisé une étude par simulation en utilisant ces données sur les exploitations agricoles à grande échelle pour comparer les deux estimateurs fondés sur un modèle bayésien avec splines à l'estimateur pondéré par les poids de sondage.

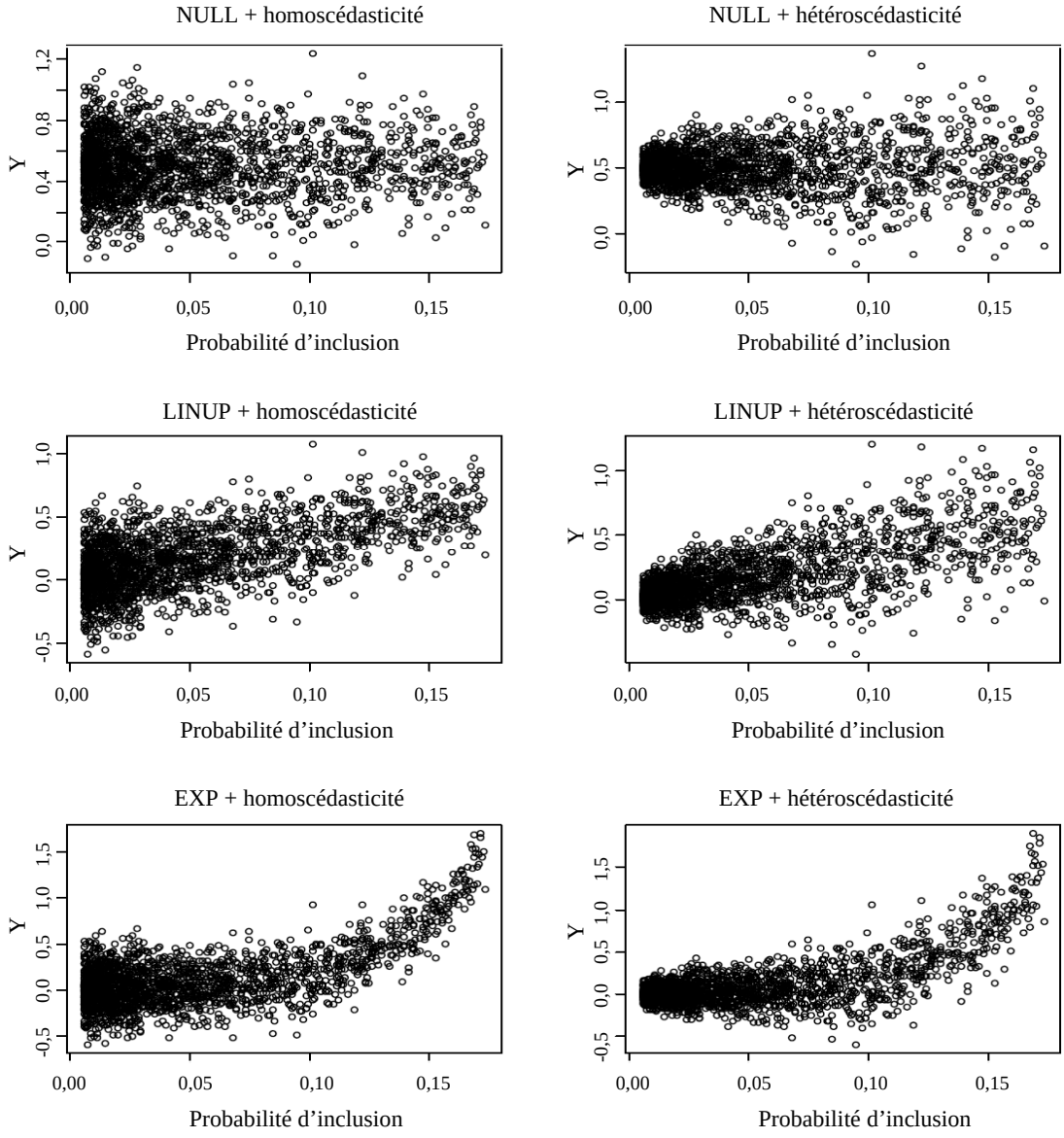

Figure 2 Nuages de points de Y en fonction des probabilités d'inclusion pour les six populations finies artificielles de taille égale à 2 000

Tableau 1
Comparaisons du biais empirique et de la racine carrée de l'erreur quadratique moyenne × 10^3 de $\theta(\alpha)$ pour $\alpha = 0{,}1, 0{,}25, 0{,}5, 0{,}75$ et $0{,}9$: scénarios avec erreurs homoscédastiques

	Biais empirique					REQM empirique				
	0,1	**0,25**	**0,5**	**0,75**	**0,9**	**0,1**	**0,25**	**0,5**	**0,75**	**0,9**
NULL										
FR inverse	-6	-3	-1	-1	-5	46	37	36	37	45
PB2SP	-5	-1	1	2	6	41	33	31	34	42
PS	-5	-3	-1	-4	-6	54	41	39	41	50
PS lisse	-7	-4	-1	-2	-5	50	39	37	38	47
CD	-197	-272	-265	-108	168	203	274	266	115	189
Ratio de RKM	3	25	33	16	6	77	125	159	112	79
Diff de RKM	-5	-1	6	14	14	58	58	94	122	113
LINUP										
FR inverse	-15	-3	-2	-1	-2	70	49	39	34	33
PB2SP	-3	-1	1	4	7	56	43	35	31	29
PS	-15	-3	-3	-2	-6	77	57	48	44	42
PS lisse	-14	-5	-2	-1	-4	72	53	45	42	41
CD	101	35	-37	-49	1	104	38	39	53	31
Ratio de RKM	-23	-9	2	5	-0.2	95	67	53	51	40
Diff de RKM	-15	-4	-4	-0.2	-2	77	55	45	43	38
EXP										
FR inverse	-8	0.4	4	7	4	60	45	41	43	49
PB2SP	-10	-6	-3	0.3	13	52	40	35	36	36
PS	-9	-3	-2	-2	-8	65	49	46	50	72
PS lisse	-12	-5	-2	-1	-2	62	47	43	46	68
CD	92	54	14	19	61	96	57	21	31	75
Ratio de RKM	-17	-11	1	3	-5	87	65	50	53	55
Diff de RKM	-9	-4	-2	-2	-7	65	49	47	47	59

Tableau 2
Comparaisons du biais empirique et de la racine carrée de l'erreur quadratique moyenne × 10^3 de $\theta(\alpha)$ pour $\alpha = 0{,}1, 0{,}25, 0{,}5, 0{,}75$ et $0{,}9$: scénarios avec erreurs hétéroscédastiques

	Biais empirique					REQM empirique				
	0,1	**0,25**	**0,5**	**0,75**	**0,9**	**0,1**	**0,25**	**0,5**	**0,75**	**0,9**
NULL										
FR inverse	-9	-8	-2	4	1	30	24	22	24	31
PB2SP	-6	-6	1	7	7	25	21	19	23	27
PS	-4	-3	-2	-1	-5	34	26	23	26	35
PS lisse	-4	-5	-2	1	-4	34	26	23	26	35
CD	-298	-325	-253	-46	270	302	327	255	60	288
Ratio de RKM	8	31	32	16	5	81	143	154	94	57
Diff de RKM	-5	-1	6	17	16	44	54	87	113	97
LINUP										
FR inverse	-11	-1	5	2	-3	32	24	24	29	35
PB2SP	-10	-1	7	3	1	29	22	22	24	30
PS	-5	-1	-0.1	-1	-4	31	28	33	45	51
PS lisse	-11	-3	2	-0.4	-5	32	26	30	44	50
CD	10	7	6	7	11	20	13	13	20	32
Ratio de RKM	-7	-3	2	3	1	36	29	30	35	41
Diff de RKM	-5	-2	-1	1	-0.2	32	27	28	33	41
EXP										
FR inverse	-8	-3	5	7	-3	30	23	23	30	48
PB2SP	-11	-7	2	6	7	28	23	20	25	36
PS	-3	-3	-2	1	-2	30	26	26	41	84
PS lisse	-8	-5	1	2	-5	30	23	24	39	86
CD	18	16	35	84	68	27	21	38	88	81
Ratio de RKM	-5	-6	-1	2	-0.1	36	31	27	32	62
Diff de RKM	-3	-3	-2	1	-0.1	32	28	28	31	67

Tableau 3
Comparaisons de la largeur moyenne et du taux de non-couverture de l'IC à 95 % × 10^3 de $\theta(\alpha)$ pour α = 0,1, 0,25, 0,5, 0,75 et 0,9

	Largeur moyenne de l'IC à 95 %					Taux de non-couverture de l'IC à 95 %				
	0,1	**0,25**	**0,5**	**0,75**	**0,9**	**0,1**	**0,25**	**0,5**	**0,75**	**0,9**
				Erreurs homoscédastiques						
NULL										
FR inverse	199	156	141	152	184	46	35	44	38	67
PB2SP	178	134	118	134	177	52	55	61	59	50
PS	195	164	151	167	237	112	65	46	40	38
LINUP										
FR inverse	257	207	157	139	141	61	45	37	46	52
PB2SP	230	167	134	123	121	58	54	44	57	59
PS	248	231	188	179	187	119	60	42	41	39
EXP										
FR inverse	234	184	163	177	234	59	44	47	40	42
PB2SP	217	157	132	144	156	54	59	55	53	60
PS	231	199	175	210	402	106	64	47	40	40
				Erreurs hétéroscédastiques						
NULL										
FR inverse	146	104	90	101	137	42	43	38	38	47
PB2SP	107	89	79	89	107	38	49	37	68	65
PS	146	101	91	113	169	80	60	51	37	42
LINUP										
FR inverse	131	107	104	124	154	70	31	36	42	40
PB2SP	125	97	87	93	116	47	35	50	58	52
PS	141	110	133	184	219	138	69	41	50	42
EXP										
FR inverse	131	99	99	134	242	63	49	34	40	41
PB2SP	116	92	84	98	139	57	55	40	63	59
PS	135	100	106	186	378	111	65	46	45	34

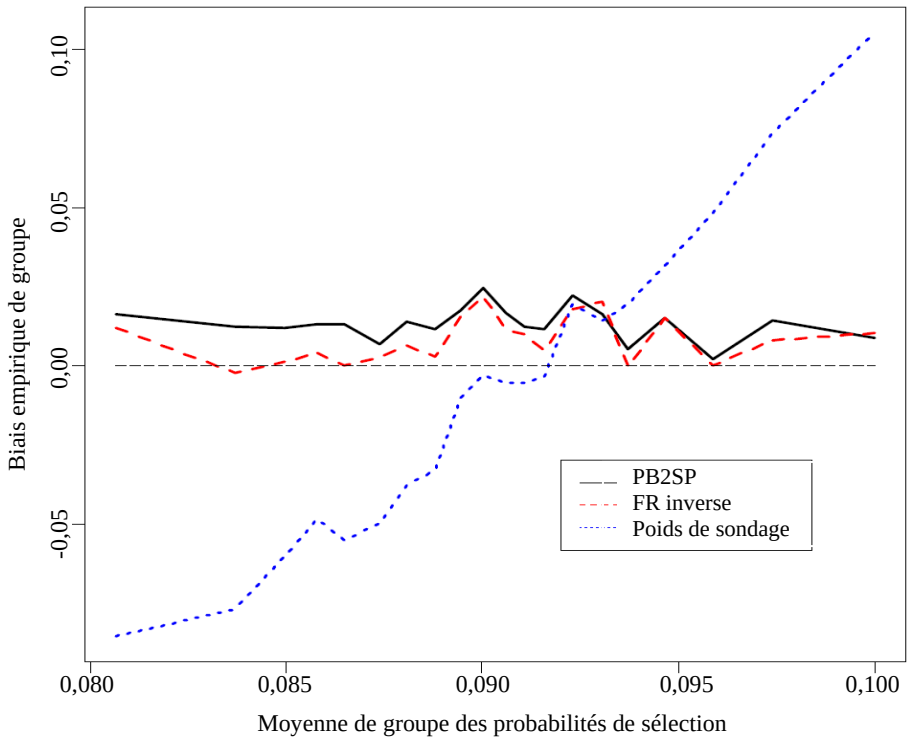

Figure 3 Variation du biais empirique des trois estimateurs pour le 90e centile dans le cas « EXP + homoscédasticité »

Le tableau 4 donne les résultats des simulations. L'approche bayésienne avec fonction de répartition inverse donne en général un biais empirique et une REQM plus petits, et une plus courte largeur moyenne de l'IC à 95 % que l'estimateur pondéré par les poids de sondage. L'IC à 95 % de l'approche bayésienne avec fonction de répartition inverse possède aussi une couverture plus proche du niveau de confiance nominal que l'estimateur pondéré par les poids de sondage quand α est égal à 0,1 et à 0,25. Cependant, dans la queue supérieure avec $\alpha = 0{,}90$, le taux de non-couverture pour l'approche bayésienne avec fonction de répartition inverse est plus élevé que le niveau nominal de 0,05, tandis que l'IC de Woodruff de l'estimateur pondéré par les poids de sondage a de bonnes propriétés. Ces résultats sont en harmonie avec ceux de Sitter et Wu (2001) selon lesquels les intervalles de Woodruff ont de bonnes propriétés même dans les parties moyennes à extrêmes des queues de la distribution. Puisque l'hypothèse de normalité conditionnelle n'est pas raisonnable ici, l'estimateur PB2SP est biaisé et l'IC à 95 % a une mauvaise couverture.

4. Discussion

L'usage des estimateurs des quantiles de population finie pondérés par les poids de sondage est très répandu chez les praticiens des sondages. Bien qu'ils soient faciles à calculer et puissent fournir des inférences valides en grand échantillon, les estimateurs pondérés avec intervalle de confiance de Woodruff peuvent être inefficaces et donner une mauvaise couverture des intervalles de confiance pour les échantillons de taille petite à modérée. Les estimateurs fondés sur un modèle peuvent améliorer l'efficacité des estimations quand le modèle est spécifié correctement, mais produisent des estimations biaisées s'il est mal spécifié. Pour trouver un compromis entre la robustesse et l'efficacité, nous avons considéré des estimateurs fondés sur des modèles avec splines. Pour l'estimation des quantiles d'une variable étudiée continue, nous pouvons estimer des fonctions de répartition fondées sur le modèle puis inverser ces fonctions pour obtenir les quantiles, ou modéliser directement la variable résultat étudié sur les probabilités d'inclusion. Dans le présent article, nous proposons deux estimateurs des quantiles fondés sur un modèle bayésien avec splines. La première méthode est celle de l'estimateur bayésien avec fonction de répartition (FR) inverse, obtenue en inversant les estimations fondées sur un modèle avec splines des fonctions de répartition. La deuxième méthode est celle de l'estimateur PB2SP, estimé en supposant que la variable résultat étudié continue suit une loi normale dont la fonction moyenne et la fonction variance sont toutes deux modélisées au moyen de splines.

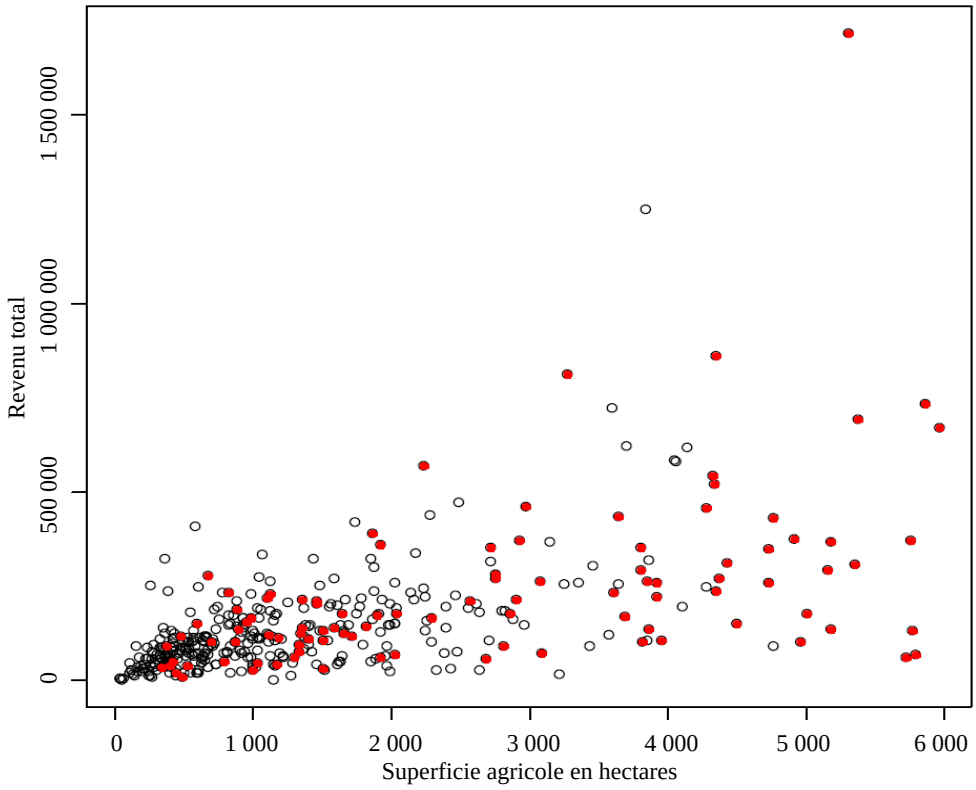

Figure 4 Nuage de points des données sur les exploitations agricoles à grande échelle avec les cercles pleins représentant chacun un échantillon ppt

Tableau 4
Biais empirique × 10^{-2}, racine carrée de l'erreur quadratique moyenne × 10^{-2}, largeur moyenne de l'IC à 95 % × 10^{-2}, et taux de non-couverture de l'IC à 95 % × 10^{3} de $\theta(\alpha)$ pour α = 0,1, 0,25, 0,5, 0,75 et 0,9 : données sur les exploitations agricoles à grande échelle

	0,1	0,25	0,5	0,75	0,9
			Biais empirique		
FR inverse	8	14	10	-22	-60
PB2SP	-110	-125	-63	-12	88
PS	20	-19	-17	-21	-61
			REQM empirique		
FR inverse	117	117	108	164	256
PB2SP	113	141	124	140	206
PS	132	173	167	226	350
			Largeur moyenne de l'IC à 95 %		
FR inverse	402	443	501	697	906
PB2SP	170	327	539	726	964
PS	285	468	615	864	1 589
			Taux de non-couverture de l'IC à 95 %		
FR inverse	96	53	26	52	90
PB2SP	670	258	42	8	17
PS	220	121	68	42	44

Les simulations donnent à penser que les deux estimateurs fondés sur un modèle bayésien avec splines donnent de meilleurs résultats que l'estimateur pondéré par les poids de sondage, les estimateurs par le ratio et par différence fondés sur le plan de sondage, ainsi que l'estimateur CD fondé sur un modèle lorsque le modèle supposé est incorrect. Les nouvelles méthodes donnent toutes deux des racines de l'erreur quadratique moyenne plus petites qu'il n'y ait pas d'association ou qu'il y ait une association linéaire ou une association non linéaire entre le résultat de l'enquête et la probabilité d'inclusion. Dans certains scénarios, l'accroissement de l'efficacité obtenu en utilisant les deux méthodes bayésiennes est considérable. Lorsque l'hypothèse de normalité du résultat étudié sachant les probabilités d'inclusion est vérifiée, l'estimateur PB2SP produit une REQM plus petite et un intervalle de crédibilité plus court que l'approche avec fonction de répartition inverse. En outre, les deux estimateurs fondés sur un modèle bayésien sont robustes à l'erreur de spécification tant de la fonction moyenne que de la fonction variance. En revanche, l'estimateur fondé sur un modèle CD est biaisé et inefficace quand la fonction moyenne ou la fonction variance est mal spécifiée. Enfin, les méthodes fondées sur un modèle bayésien ont l'avantage de permettre de calculer plus facilement l'IC à 95 % et l'inférence fondée sur les lois a posteriori des paramètres. Cette caractéristique est intéressante, parce que l'estimation de la variance pour les autres estimateurs fondés sur le plan de sondage peut être compliquée. La méthode d'estimation de la variance de Woodruff pour l'estimateur pondéré par les poids de sondage donne de bons résultats quand une fraction importante des données est sélectionnée à partir de la population finie, même dans les parties moyenne à extrême des queues de la fonction de répartition. Cependant, lorsque les données provenant de la population sont peu nombreuses, la méthode de Woodruff a tendance à sous-estimer la couverture de l'intervalle de confiance, alors que les deux méthodes bayésiennes donnent une couverture de ces intervalles plus proche du niveau nominal.

Les trois estimateurs fondés sur le plan de sondage ont un biais empirique global comparable à celui des deux estimateurs fondés sur un modèle bayésien avec splines. Toutefois, la variation du biais de l'estimateur pondéré par les poids de sondage présente une tendance linéaire lorsqu'augmente la moyenne d'échantillon des probabilités d'inclusion. En l'absence d'association entre le résultat étudié et la probabilité d'inclusion, les estimateurs par le ratio et par différence donnent un biais et une REQM relativement plus grands que l'estimateur pondéré par les poids de sondage. Cependant, dans certains scénarios de simulation, les estimateurs par le ratio et par différence produisent une REQM plus petite que l'estimateur pondéré par les poids de sondage. La comparaison entre l'estimateur pondéré par les poids de sondage classique et l'estimateur pondéré par les poids de sondage lisse laisse entendre que l'ajustement d'une courbe cubique lisse pour la fonction de répartition pondérée par les poids de sondage peut améliorer l'efficacité, mais que l'estimateur pondéré par les poids de sondage lisse continuera d'avoir une REQM plus grande que l'estimateur bayésien avec fonction de répartition inverse.

Pour les données dont la distribution est normale, nous recommandons d'utiliser l'estimateur PB2SP de préférence aux autres, en raison du biais plus petit, de la REQM plus petite, et de la meilleure couverture et de la plus courte largeur de l'intervalle de confiance. L'estimateur PB2SP et son intervalle de probabilité a posteriori de 95 % sont faciles à obtenir en utilisant l'algorithme proposé par Crainiceanu et coll. (2007), qui offre aussi l'avantage d'un temps de calcul relativement court.

L'estimateur PB2SP peut être biaisé quand l'hypothèse de normalité conditionnelle ne tient pas. Une option dans ce cas consiste à transformer le résultat étudié afin que l'hypothèse de normalité conditionnelle devienne plus raisonnable. L'estimateur PB2SP peut être appliqué aux données transformées et les tirages à partir des lois a posteriori des unités non échantillonnées sont de nouveau transformés pour revenir à l'échelle originale avant d'estimer les quantiles d'intérêt.

Dans nos simulations avec des données non normales, l'approche bayésienne avec fonction de répartition inverse demeurait plus efficace que l'estimateur pondéré par les poids de sondage. L'amélioration de la couverture de l'intervalle de confiance était limitée aux situations où la taille de l'échantillon est petite, avec une méthode de détermination de l'IC de Woodruff donnant de bons résultats quand l'hypothèse de grand échantillon est vérifiée. Donc, pour les données ne suivant pas une loi normale pour lesquelles il n'existe aucune transformation évidente en vue d'améliorer la normalité, nous ne recommandons pas l'approche bayésienne avec fonction de répartition inverse quand l'échantillon est de grande taille. Étant donné les bonnes propriétés de l'estimateur PB2SP dans les conditions de normalité, l'extension à examiner lors de futurs travaux consisterait à relâcher l'hypothèse de normalité dans nos approches proposées.

Nous utilisons la probabilité d'inclusion comme variable auxiliaire ici. Lorsqu'il n'existe qu'une seule variable auxiliaire pertinente, peu importe que l'on modélise la probabilité d'inclusion ou la variable auxiliaire. Par contre, s'il existe plus d'une variable auxiliaire pertinente, la probabilité d'inclusion est la variable auxiliaire principale qui doit être modélisée correctement, puisque la spécification incorrecte du modèle reliant le résultat étudié à la probabilité d'inclusion entraîne un biais. Lorsque d'autres variables auxiliaires sont observées pour toutes les unités de la population finie, nos estimateurs bayésiens peuvent tous deux être étendus facilement afin d'inclure les covariables auxiliaires supplémentaires en ajoutant des termes linéaires pour ces variables dans le modèle avec splines pénalisées correspondant.

Un examinateur a proposé une approche pondérée de rechange fondée sur la loi de Dirichlet, qui est facile à calculer, mais n'utilise pas les variables auxiliaires connues dans les unités non échantillonnées. Une autre possibilité consiste à redéfinir l'estimateur CD au moyen du modèle avec splines que nous avons utilisé pour définir l'estimateur PB2SP. Plus précisément, au lieu de supposer que le modèle de régression passe par l'origine, un modèle avec splines est ajusté aux moments d'ordres un et deux de la loi conditionnelle de la variable résultat étudiée sachant la probabilité d'inclusion. L'estimateur CD fondé sur les splines devrait donner des résultats comparables à ceux de l'estimateur PB2SP, et sa variance peut être estimée en utilisant des méthodes de rééchantillonnage.

Dans le contexte de la statistique officielle, les méthodes décrites dans le présent article illustrent les avantages éventuels d'un changement de paradigme pour passer de méthodes fondées sur le plan de sondage à la modélisation bayésienne en vue de produire des inférences ayant de bonnes propriétés fréquentistes. Nos collègues spécialistes de la statistique fondée sur l'échantillonnage probabiliste ont deux grandes objections à ce point de vue.

Premièrement, l'idée d'une approche exagérément fondée sur un modèle – pire encore, bayésienne – des enquêtes probabilistes est mal acceptée, quoique nous mettions ici l'accent sur des méthodes bayésiennes ayant de bonnes propriétés de randomisation. Selon nous, les méthodes probabilistes classiques fondées sur le plan ne fournissent pas l'approche globale nécessaire pour traiter les problèmes complexes qui se posent de plus en plus souvent en statistique officielle. Des choix judicieux de modèles bien calés sont nécessaires pour s'y attaquer. En accordant de l'attention aux caractéristiques du plan de sondage et en choisissant des lois a priori objectives, on peut obtenir des inférences bayésiennes exemptes de subjectivité, et comme les hypothèses de modélisation sont explicites, elles peuvent être critiquées et perfectionnées. Voir Little (2004, 2012) pour une discussion plus approfondie de ces points.

La deuxième objection est que les méthodes bayésiennes requièrent des calculs trop compliqués pour le secteur de la statistique officielle qui doit calculer correctement et produire rapidement un grand nombre de statistiques régulières. Il est vrai qu'à l'heure actuelle, le calcul bayésien peut sembler rébarbatif aux statisticiens habitués à de simples statistiques pondérées et à des méthodes d'estimation de la variance par rééchantillonnage. Dans un article défendant vigoureusement les approches bayésiennes, Sedransk (2008) mentionne que les difficultés pratiques de calcul sont un inhibiteur. Nous convenons que du travail reste à faire pour répondre à cette objection, mais nous ne pensons pas que le problème soit insurmontable. La recherche sur les méthodes de calcul bayésien a connu une véritable explosion ces dernières décennies, tout comme la capacité de calcul. Des modèles bayésiens ont été ajustés pour résoudre des

problèmes de très grande portée et très complexes, dans certains cas nettement plus complexes que ceux qui se posent habituellement dans le secteur de la statistique officielle.

Remerciements

Nous remercions M. Philip Kokic de la Commonwealth Scientific and Industrial Research Organisation, de nous avoir fourni les données sur les exploitations agricoles à grande échelle (*broadacre farms*). Nous remercions aussi un rédacteur associé et les examinateurs de leurs commentaires constructifs au sujet de la version originale du présent article.

Bibliographie

ABARE (2003). Australian farm surveys report 2003. Canberra.

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of American Statistical Association*, 88, 268-277.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2010). Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 36, 1, 25-37.

Crainiceanu, C.M., Ruppert, D., Carroll, R.J., Joshi, A. et Goodner, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic error. *Journal of Computational and Graphical Statistics*, 16, 265-288.

Dorfman, H., et Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1474.

Francisco, C.A., et Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.

Harms, T., et Duchesne, P. (2006). De l'estimation des quantiles par calage. *Techniques d'enquête*, 32, 1, 41-57.

Kuk, A.Y.C. (1993). A kernel method for estimating finite population functions using auxiliary information. *Biometrika*, 80, 385-392.

Kuk, A.Y.C., et Welsh, A.H. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society*, Séries B, 63, 277-292.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, DOI: 10.1198/016214504000000467. {70}, 99, 546-556.

Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (avec discussion et réplique). *Journal of Official Statistics*, 28, 309-334.

Lombardía, M.J., González-Manteiga, W. et Prada-Sánchez, J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference*, 116, 367-388.

Lombardía, M.J., González-Manteiga, W. et Prada-Sánchez, J.M. (2004). Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimate of a finite population distribution function. *Journal of Nonparametric Statistics*, 16, 63-90.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution function and quantile from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Royall, R.M., et Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

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

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24, 495-506.

Sitter, R.R., et Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters*, 52, 353-358.

Wang, S., et Dorfman, A.H. (1996). A new estimator for the finite population distribution function. *Biometrika*, 83, 639-652.

Wood, S.N. (1994). Monotonic smoothing splines fitted by cross validation SIAM. *Journal on Scientific Computing*, 15, 1126-1133.

Woodruff, R. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complex auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

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.