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

Article

Calage adaptatif pour la prédiction de totaux de population finie

par Robert G. Clark et Raymond L. Chambers

Décembre 2008

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 181-192
Statistique Canada, N° 12-001-X au catalogue

Calage adaptatif pour la prédiction de totaux de population finie

Robert G. Clark et Raymond L. Chambers [1]

Résumé

Les poids d'échantillonnage peuvent être calés de manière à refléter les totaux connus de population d'un ensemble de variables auxiliaires. Le biais des prédicteurs des totaux de population finie calculés en utilisant ces poids est faible si ces variables sont reliées à la variable d'intérêt, mais leur variance peut être élevée si l'on utilise un trop grand nombre de variables auxiliaires. Dans le présent article, nous élaborons une approche de « calage adaptatif » où les variables auxiliaires qu'il convient d'utiliser dans la pondération sont sélectionnées en se servant de données d'échantillon. Nous montrons que, dans de nombreux cas, les estimateurs calés adaptativement ont une erreur quadratique moyenne plus faible et de meilleures propriétés de couverture que les estimateurs non adaptatifs.

Mots clés : Enquête par sondage ; pondération de l'échantillon ; approche de prédiction ; estimation ridge ; sélection du modèle ; méthodes séquentielles (pas à pas).

1. Introduction

Les prédicteurs de totaux de population finie sont habituellement calculés par sommation pondérée de valeurs d'échantillon. Très souvent, on dispose de variables auxiliaires dont les valeurs d'échantillon et les totaux de population sont connus. Les poids peuvent être construits de manière à ce que les sommes pondérées des variables auxiliaires concordent avec les totaux connus de population, processus appelé calage (Deville et Särndal 1992). Les prédicteurs des totaux de population finie fondés sur les poids calés ont généralement un biais de prédiction beaucoup plus faible que ceux calculés sans information auxiliaire.

Les auteurs de la littérature existante sur la prédiction en population finie supposent essentiellement qu'un ensemble de variables auxiliaires utiles est choisi sans référence aux données d'échantillon. Cependant, en pratique, il existe parfois un grand ensemble de variables auxiliaires possibles qui ne devraient pas toutes être utilisées. L'augmentation du nombre de variables auxiliaires réduit généralement le biais des prédicteurs calés, mais accroît la variance, de sorte qu'utiliser trop de variables auxiliaires peut effectivement augmenter l'erreur quadratique moyenne des prédicteurs calés. Souvent, le choix des variables auxiliaires qu'il convient d'utiliser n'est pas évident, et des données d'échantillon peuvent être nécessaires pour déterminer quel ensemble de ces variables est approprié pour les prédicteurs des totaux de variables d'intérêt particulières. Dans le présent article, nous élaborons des méthodes permettant de faire cette détermination. Notre approche peut être appelée calage adaptatif, parce que l'ensemble de variables est choisi adaptativement en partant des données d'échantillon, plutôt que statiquement, sans tenir compte de l'échantillon dont on dispose.

Nous utilisons le cadre de prédiction pour l'estimation en population finie (voir, par exemple, Brewer 1963 ; Royall 1970 ; Valliant, Dorfman et Royall 2000). Suivant cette approche, les valeurs de population des variables d'intérêt sont traitées comme des variables aléatoires. Le but est de prédire le total de population (qui est aussi une variable aléatoire) ou d'autres grandeurs de population finie en utilisant des données d'échantillon sur la variable d'intérêt, et des données de population sur certaines variables auxiliaires. L'échantillon peut avoir été sélectionné par échantillonnage probabiliste ou par une autre méthode, et est traité conditionnellement à la méthode choisie dans l'inférence. L'une des caractéristiques principales de l'approche est l'utilisation d'un modèle stochastique de la variable d'intérêt. L'un des aspects du cadre de prédiction est que la spécification incorrecte du modèle, par exemple par omission de variables auxiliaires importantes, peut causer un biais important.

Un autre cadre est celui de l'approche assistée par modèle (Särndal, Swensson et Wretman 1992). Dans cette approche, on utilise encore un modèle stochastique, mais celui-ci joue un rôle moins crucial. La nature aléatoire de l'échantillonnage est exploitée pour faire en sorte que les estimateurs soient approximativement sans biais, même si le modèle est incorrect. Quand le modèle est correct, les deux approches donnent des estimateurs approximativement sans biais, mais celle fondée sur un modèle produit généralement des estimateurs d'intérêt dont la variance est plus faible. Si le modèle est spécifié incorrectement, les prédicteurs fondés sur un modèle et les estimateurs de leur variance peuvent présenter un biais plus important, mais des méthodes fondées sur un modèle robuste ont été élaborées pour contourner ce problème. Par exemple, Royall et Herson (1973a, 1973b) discutent de la prédiction robuste, et Royall

1. Robert G. Clark et Raymond L. Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australie. Courriel : Robert_Clark@uow.edu.au.

et Cumberland (1978, 1981a, 1981b) élaborent des estimateurs de variance robustes à l'hétéroscédasticité. Pour des comparaisons du cadre de prédiction et du cadre assisté par modèle, voir, par exemple, Smith (1976) et Hansen, Madow et Tepping (1983).

Le problème de la sélection d'un ensemble de variables auxiliaires dans le cadre assisté par modèle a été étudié par Silva et Skinner (1997), ainsi que par Skinner et Silva (1997). Ils ont constaté que l'ajout de variables de calage réduit l'erreur quadratique moyenne (EQM) jusqu'à un certain point, après lequel l'addition d'autres variables accroît l'EQM. Choisir les variables de calage adaptativement, en se basant sur des données d'échantillon, ont donné de meilleures estimations que le calage sur toutes les variables ou sans en utiliser aucune. L'applicabilité de ces travaux à la prédiction fondée sur un modèle n'est pas claire, parce que le rôle joué par le modèle dans les deux cadres est très différent. Des modèles mal spécifiés peuvent donner lieu à des prédicteurs fondés sur le modèle considérablement biaisés, tandis que les estimateurs assistés par modèle sont approximativement sans biais, même si des variables importantes sont omises. Par conséquent, diverses stratégies de sélection du modèle pourraient être appropriées dans les deux cadres de travail. En outre, les différences entre diverses variantes devraient, en principe, être plus prononcées dans le cadre de prédiction que dans le cadre assisté par modèle.

Chambers, Skinner et Wang (1999) ont proposé une approche de sélection des variables de calage dans le cadre de prédiction qui s'appuie sur la sélection ascendante, descendante ou séquentielle. (Dans la suite du texte, cette approche sera appelée CSW.) Leur décision d'omettre (ou d'ajouter) une variable à chaque étape était fondée sur la minimisation de l'erreur quadratique estimée de la prédiction (EQMP) pour le prédicteur d'intérêt. L'approche n'a pas été évaluée par une étude en simulation, et les estimateurs de l'EQMP utilisés n'étaient pas robustes à l'hétéroscédasticité.

Le but du présent article est de développer l'approche de base de CSW afin de l'appliquer à une gamme plus étendue de situations, y compris les populations hétéroscédastiques et les échantillons à plusieurs degrés, et d'évaluer l'approche en utilisant des études par simulation réalistes. Nous utiliserons des estimateurs de l'EQMP qui sont robustes à l'hétéroscédasticité, ainsi qu'à la corrélation dans le cas des sondages à plusieurs degrés. Nous évaluerons les propriétés des estimateurs par simulation en nous servant de deux populations : des données financières sur les fermes provenant d'une enquête sur les fermes et des données sur la population active provenant d'un recensement de population.

Comme dans l'étude CSW, l'approche de base consistera à construire un ensemble de variables auxiliaires par sélection séquentielle de variables, en partant d'un ensemble initial de données. Cet algorithme construit un ensemble de variables auxiliaires grâce à une suite de décisions multiples de choix entre deux ensembles emboîtés de variables. Nous comparons plusieurs critères pour faire un choix entre les deux ensembles emboîtés, dont la signification statistique et un certain nombre d'estimateurs possibles de l'erreur quadratique moyenne de prédiction (EQMP). Nous considérons trois estimateurs différents de l'EQMP, à savoir un estimateur non robuste, un estimateur robuste à l'hétéroscédasticité et un estimateur robuste à l'hétéroscédasticité ainsi qu'aux corrélations entre les unités primaires d'échantillonnage dans l'échantillonnage à plusieurs degrés.

À la section 2, nous donnons la notation et les définitions. À la section 3, nous calculons la différence entre les EQMP de deux prédicteurs basés sur des modèles emboîtés et nous élaborons plusieurs options d'estimateur de cette différence. À la section 4, nous présentons les résultats des simulations pour une enquête sur les fermes et une enquête à plusieurs degrés auprès des ménages. La section 5 est consacrée à une discussion. Nous concluons que le calage adaptatif donne généralement de meilleurs résultats que le calage statique, à condition d'utiliser un estimateur non robuste de l'EQMP, ou la signification statistique, comme objectif dans la sélection du modèle.

2. Notation et définitions

Une variable d'intérêt Y_i est observée pour un échantillon s de n unités, qui est un sous-ensemble d'une population finie U contenant N unités. Le but est d'estimer le total de population $T_Y = \sum_{i \in U} Y_i$ et d'autres grandeurs en population finie de Y. Un vecteur de variables auxiliaires $\boldsymbol{x_i}$ est disponible pour $i = 1, ..., n$, dont le total de population $\boldsymbol{T}_x = \sum_{i \in U} \boldsymbol{x_i}$. est connu.

Les estimateurs pondérés de T_Y sont donnés par $\hat{T}_Y = \sum_{i \in s} w_i Y_i$, où w_i peut dépendre des variables auxiliaires, mais non de la variable d'intérêt. Un ensemble de poids est dit calé sur $\boldsymbol{x_i}$ si $\sum_{i \in s} w_i \boldsymbol{x_i} = \boldsymbol{T}_x$.

Le meilleur prédicteur linéaire sans biais (BLUP for *Best Linear Unbiased Predictor*) fondé sur un modèle de régression linéaire est un exemple d'estimateur calé. Le prédicteur BLUP utilisé le plus fréquemment est basé sur le modèle

$$E[Y_i] = \boldsymbol{\beta}^T \boldsymbol{x_i}$$

$$\text{var}[Y_i] = \sigma_i^2 = v_i \sigma^2 \qquad (1)$$

$$\text{cov}[Y_i, Y_j] = 0 (i \neq j)$$

(en supposant que v_i est connu) et est donné par

$$\hat{T}_Y = \sum_{i \in s} Y_i + \sum_{i \in r} \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i \qquad (2)$$

où $r = U - s$ est l'ensemble d'unités non incluses dans l'échantillon et

$$\hat{\boldsymbol{\beta}} = \left\{ \sum_{i \in s} v_i^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^T \right\}^{-1} \sum_{i \in s} v_i^{-1} \boldsymbol{x}_i Y_i \qquad (3)$$

est un estimateur par les moindres carrés pondérés de $\boldsymbol{\beta}$. Le prédicteur BLUP peut aussi s'écrire sous la forme pondérée

$$\hat{T}_Y = \sum_{i \in s} w_i Y_i$$

où les poids w_i sont donnés par

$$w_i = 1 + \boldsymbol{T}_{\boldsymbol{xr}}^T \left\{ \sum_{j \in s} v_j^{-1} \boldsymbol{x}_j \boldsymbol{x}_j^T \right\}^{-1} v_i^{-1} \boldsymbol{x}_i \qquad (4)$$

et $\boldsymbol{T}_{\boldsymbol{xr}} = \sum_{i \in r} \boldsymbol{x}_i$. Il est facile de vérifier que $\sum_{i \in s} w_i \boldsymbol{x}_i = \boldsymbol{T}_{\boldsymbol{x}}$.

Dans le cas de données hétéroscédastiques, il est habituellement difficile de modéliser v_i de manière fiable. Dans ce cas, on dispose d'estimateurs robustes de la variance de prédiction du prédicteur BLUP qui ne requièrent pas que l'on connaisse v_i (Royall et Cumberland 1978, 1981a, 1981b). En ce qui concerne les échantillons à plusieurs degrés, l'hypothèse d'indépendance peut être violée. Le cas échéant, on peut encore utiliser le prédicteur BLUP basé sur (1) et recourir à un estimateur de variance robuste de la variance de prédiction basé sur les grappes finales (par exemple, Valliant et coll. 2000, chapitre 9). Une autre approche, que nous ne considérons pas ici, consisterait à construire un prédicteur BLUP basé sur un modèle dans lequel sont incluses les corrélations intra-grappe (Royall 1976). Nous discuterons plus en détail de l'estimation robuste et non robuste de l'erreur quadratique moyenne de la prédiction du BLUP à la section 3.

Nous devons prendre une décision quant aux éléments à inclure dans $\boldsymbol{x}_i$ dans le prédicteur BLUP. La sélection séquentielle, la sélection ascendante et la sélection descendante sont des algorithmes qui peuvent être utilisés pour décider quel sous-ensemble des variables auxiliaires disponibles il conviendrait d'utiliser. Les trois algorithmes comprennent chacun de nombreux choix entre deux ensembles emboîtés de variables auxiliaires. Supposons que nous devions choisir entre A) l'utilisation d'un prédicteur $\hat{T}_A$ basé sur $\boldsymbol{x}_i$ et B) l'utilisation d'un prédicteur $\hat{T}_B$ basé sur un sous-vecteur $\boldsymbol{x}_{1i}$. Nous pouvons partitionner $\boldsymbol{x}_i$ sous la forme $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}^T, \boldsymbol{x}_{2i}^T)^T$. Les nombres d'éléments de $\boldsymbol{x}_i$, $\boldsymbol{x}_{1i}$ et $\boldsymbol{x}_{2i}$ sont désignés par p, p_1 et p_2, respectivement.

Nous partitionnons de la même manière $\boldsymbol{\beta}$ sous la forme $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$. Le prédicteur $\hat{T}_A$ est sans biais sous le modèle A :

$$E[Y_i] = \boldsymbol{\beta}^T \boldsymbol{x}_i = \boldsymbol{\beta}_1^T \boldsymbol{x}_{1i} + \boldsymbol{\beta}_2^T \boldsymbol{x}_{2i}. \qquad (5)$$

Le prédicteur $\hat{T}_B$ est sans biais sous le modèle B,

$$E[Y_i] = \boldsymbol{\beta}_1^T \boldsymbol{x}_{1i}, \qquad (6)$$

qui est le cas particulier du modèle A où $\boldsymbol{\beta}_2 = \boldsymbol{0}$.

3. Estimation de la différence entre les EQMP

3.1 Comparaison des prédicteurs provenant de modèles hiérarchiques

À l'instar de CSW, notre approche consiste à estimer la différence entre les EQMP de deux estimateurs :

$$\Delta = E[(\hat{T}_A - T_Y)^2] - E[(\hat{T}_B - T_Y)^2]$$

où les espérances sont évaluées par rapport au modèle A, parce que le modèle B est un cas particulier de ce modèle. Habituellement, $\hat{T}_A$ aura un biais plus faible que $\hat{T}_B$, mais une variance plus élevée. L'un ou l'autre prédicteur peut avoir une EQMP plus élevée ou plus faible, selon la population et l'échantillon en question.

Dans le cas de l'échantillonnage à un seul degré, il est habituellement raisonnable de supposer que Y_i et Y_j sont indépendantes pour tout $i \neq j$. À la section 3.2, nous dérivons Δ et un estimateur de cette quantité dans ces conditions. À la section 3.3, nous décrivons le cas particulier instructif où les variances sont égales et où des prédicteurs BLUP sont utilisés ; il s'agit du cas pris en considération par CSW. À la section 3.4, nous généralisons ce cas particulier en décrivant un estimateur de Δ robuste à l'hétéroscédasticité. À la section 3.5, nous étendons encore davantage l'approche en calculant Δ et un estimateur de cette quantité sous échantillonnage à plusieurs degrés, où il peut exister des corrélations entre des valeurs provenant d'une même grappe.

3.2 Estimation de Δ dans un échantillonnage à un degré avec variance connue

En plus du modèle (5), nous supposons ici que Y_i et Y_j sont indépendantes pour $i \neq j$ et que $\text{var}[Y_i] = \sigma_i^2 = \sigma^2 v_i$ où v_i est connu. Dans ces conditions, l'EQMP de tout prédicteur $\hat{T} = \sum_{i \in s} w_i Y_i$ est donnée par

$$\begin{aligned} \text{EQMP}[\hat{T}] &= E[(\hat{T} - T_Y)^2] \\ &= \left\{ E\left[\sum_{i \in s} w_i Y_i - \sum_{i \in U} Y_i \right] \right\}^2 + \text{var}\left[\sum_{i \in s} (w_i - 1) Y_i - \sum_{i \in r} Y_i \right] \\ &= \left\{ \boldsymbol{\beta}^T \left(\sum_{i \in s} w_i \boldsymbol{x}_i - \sum_{i \in U} \boldsymbol{x}_i \right) \right\}^2 + \sum_{i \in s} (w_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2. \end{aligned}$$

En écrivant $\boldsymbol{d} = \sum_{i\in s} w_i \boldsymbol{x_i} - \boldsymbol{T_x}$, nous pouvons réécrire l'EQMP sous la forme

$$\text{EQMP}[\hat{T}] = \boldsymbol{d}^T(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d} + \sum_{i\in s}(w_i - 1)^2\sigma_i^2 + \sum_{i\in r}\sigma_i^2.$$

Soit $\boldsymbol{d_A} = \sum_{i\in s} w_{Ai}\boldsymbol{x_i} - \boldsymbol{T_x}$ et $\boldsymbol{d_B} = \sum_{i\in s} w_{Bi}\boldsymbol{x_i} - \boldsymbol{T_x}$. Alors, Δ est donné par :

$$\begin{aligned}\Delta &= \text{EQMP}[\hat{T}_A] - \text{EQMP}[\hat{T}_B] \\ &= \boldsymbol{d_A^T}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d_A} - \boldsymbol{d_B^T}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d_B} \\ &\quad + \sum_{i\in s}(w_{Ai} - 1)^2\sigma_i^2 - \sum_{i\in s}(w_{Bi} - 1)^2\sigma_i^2. \qquad (7)\end{aligned}$$

Pour estimer Δ, nous considérons d'abord la façon d'estimer $\boldsymbol{\beta}$ et la variance de $\hat{\boldsymbol{\beta}}$. L'estimateur par les moindres carrés habituels est $\hat{\boldsymbol{\beta}} = S_x^{-1}S_{xy}$ où $S_x = \sum_{i\in s} v_i^{-1}\boldsymbol{x_i}\boldsymbol{x_i^T}$ et $S_{xy} = \sum_{i\in s} v_i^{-1}\boldsymbol{x_i}Y_i$. L'estimateur habituel de la variance de $\hat{\boldsymbol{\beta}}$ est $\widehat{\text{var}}[\hat{\boldsymbol{\beta}}] = \hat{\sigma}^2 S_x^{-1}$, où $\hat{\sigma}^2 = \sum_{i\in s}(Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x_i})^2 v_i^{-1}/(n-p)$.

Nous pouvons estimer $(\boldsymbol{\beta}\boldsymbol{\beta}^T)$ sans biais par $(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T - \widehat{\text{var}}[\hat{\boldsymbol{\beta}}])$. Par conséquent, l'expression suivante est un estimateur sans biais de Δ :

$$\begin{aligned}\hat{\Delta} &= \boldsymbol{d_A^T}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T - \widehat{\text{var}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d_A} - \boldsymbol{d_B^T}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T - \widehat{\text{var}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d_B} \\ &\quad + \sum_{i\in s}(w_{Ai} - 1)^2\hat{\sigma}^2 v_i - \sum_{i\in s}(w_{Bi} - 1)^2\hat{\sigma}^2 v_i. \qquad (8)\end{aligned}$$

L'expression (7) est applicable et l'estimateur (8) en est un estimateur sans biais, pour tout prédicteur pondéré $\hat{T}_A$ et $\hat{T}_B$. Nous nous intéressons au cas particulier où $\hat{T}_A$ et $\hat{T}_B$ sont des prédicteurs BLUP. Dans ce cas, $\hat{T}_A$ est calé sur $\boldsymbol{T_x}$ de sorte que $\boldsymbol{d_A} = \boldsymbol{0}$, et (8) se simplifie alors en

$$\begin{aligned}\hat{\Delta} &= -\boldsymbol{d_B^T}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T - \widehat{\text{var}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d_B} \\ &\quad + \sum_{i\in s}(w_{Ai} - 1)^2\hat{\sigma}^2 v_i - \sum_{i\in s}(w_{Bi} - 1)^2\hat{\sigma}^2 v_i. \qquad (9)\end{aligned}$$

3.3 Un cas particulier important

À la présente sous-section, nous faisons les hypothèses énoncées à la section 3.2 et nous supposons en outre que $v_i = 1$ pour tout i. Nous supposons aussi que la dimension de $\boldsymbol{x_{2i}}$ est 1, autrement dit nous considérons s'il faut utiliser ou non dans la prédiction une variable auxiliaire particulière provenant de $\boldsymbol{x_i}$. Les expressions (7) et (9) se simplifient dans ces conditions.

Soit u_i le résidu d'une régression de x_{2i} sur $\boldsymbol{x_{1i}}$:

$$u_i = x_{2i} - C^T\boldsymbol{x_{1i}}$$

$$C = \left(\sum_{i\in s}\boldsymbol{x_{1i}}\boldsymbol{x_{1i}^T}\right)^{-1}\sum_{i\in s}\boldsymbol{x_{1i}}x_{2i}.$$

À l'aide d'opérations algébriques linéaires simples, nous pouvons montrer que :

$$\boldsymbol{\beta}^T\boldsymbol{d_B} = -\beta_2\sum_{i\in r}u_i$$

et que

$$\sum_{i\in s}(w_{Ai} - 1)^2 - \sum_{i\in s}(w_{Bi} - 1)^2 = \left(\sum_{i\in r}u_i\right)^2 S_u^{-1}$$

où $S_u = \sum_{i\in s}u_i^2$.

Donc, (7) devient

$$\Delta = \sigma^2\left(\sum_{i\in r}u_i\right)^2 S_u^{-1} - \left(\sum_{i\in r}u_i\right)^2\beta_2^2.$$

CSW montrent que $d_B^T\,\widehat{\text{var}}\,[\hat{\beta}_2]\,d_B = \hat{\sigma}^2(\sum_{i\in r}u_i)^2 S_u^{-1}$. Par conséquent, (9) devient

$$\hat{\Delta} = \left(\sum_{i\in r}u_i\right)^2(2\hat{\sigma}^2 S_u^{-1} - \hat{\beta}_2^2).$$

Il est proposé d'adopter $\hat{T}_A$ quand $\hat{\Delta} < 0$, et d'utiliser $\hat{T}_B$ autrement. Il s'ensuit que nous adoptons $\hat{T}_A$ quand $\hat{\beta}_2^2 > 2\hat{\sigma}^2 S_u^{-1}$. Comme l'ont mentionné CSW, cela équivaut à adopter $\hat{T}_A$ quand $F = \hat{\beta}_2^2/(\hat{\sigma}^2 S_u^{-1})$ est supérieur à 2. Notons que F est la statistique utilisée habituellement pour tester l'hypothèse nulle selon laquelle $\beta_2 = 0$. Si n est grand, le seuil pour le test F au niveau de signification de 5 % est 3,96, tandis que nous sommes arrivés à un seuil de 2 pour l'adoption de l'ensemble plus grand de variables. Par conséquent, la décision d'utiliser A au lieu de B en s'appuyant sur un test de signification requiert plus de preuve à l'encontre de B que ne le laisserait supposer une simple comparaison des EQMP estimées de $\hat{T}_A$ et $\hat{T}_B$. Autrement dit, utiliser $\hat{\Delta}$ donne lieu à de plus grands modèles que l'utilisation d'un test de signification.

3.4 Estimation de Δ robuste à l'hétéroscédasticité

Les estimateurs de Δ présentés aux sections 3.2 et 3.3 s'appuient sur le fait que $\text{var}[Y_i]$ est connue au moins jusqu'à une constante de proportionnalité. En pratique, les variances sont, au mieux, connues approximativement et les méthodes qui ne dépendent pas de l'hypothèse que la variance est connue pourraient donner de meilleurs résultats. Nous utiliserons un estimateur de σ_i^2 qui, selon l'hypothèse du modèle (5), est approximativement sans biais pour σ_i^2 en général, et exactement sans biais si $\sigma_i^2 = \sigma^2$:

$$\hat{\sigma}_i^2 = \frac{n}{n-p}(Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x_i})^2.$$

(Un estimateur de rechange serait $\hat{\sigma}_i^2 = (Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x_i})^2$, comme dans Royall et Cumberland 1981b.)

L'estimateur de $\boldsymbol{\beta}$ sera encore l'estimateur par les moindres carrés pondérés donné par (3). La variance de $\hat{\boldsymbol{\beta}}$ est

$$\begin{aligned}\text{var}[\hat{\boldsymbol{\beta}}] &= \text{var}[S_x^{-1} S_{xy}] \\ &= \text{var}\left[S_x^{-1} \sum_{i \in s} \boldsymbol{x_i} Y_i\right] \\ &= S_x^{-1}\left(\sum_{i \in s} \boldsymbol{x_i}\, \boldsymbol{x_i^T}\, \sigma_i^2\right) S_x^{-1}.\end{aligned}$$

Elle peut être estimée par

$$\widehat{\text{var}}_{\text{robust}}[\hat{\boldsymbol{\beta}}] = S_x^{-1}\left(\sum_{i \in s} \boldsymbol{x_i}\, \boldsymbol{x_i^T}\, \hat{\sigma}_i^2\right) S_x^{-1}.$$

Donc, nous pouvons estimer Δ par

$$\begin{aligned}\hat{\Delta} &= \boldsymbol{d_A^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - \widehat{\text{var}}_{\text{robust}}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d_A} \\ &\quad - \boldsymbol{d_B^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - \widehat{\text{var}}_{\text{robust}}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d_B} \\ &\quad + \sum_{i \in s} (w_{Ai} - 1)^2\, \hat{\sigma}_i^2 - \sum_{i \in s} (w_{Bi} - 1)^2\, \hat{\sigma}_i^2. \qquad (10)\end{aligned}$$

3.5 Estimation de Δ dans l'échantillonnage à plusieurs degrés

Les estimateurs de Δ présentés aux sections 3.2, 3.3 et 3.4 reposent tous sur l'hypothèse que les valeurs de Y sont indépendantes pour différentes unités. Dans l'échantillonnage à plusieurs degrés, on commence par sélectionner un échantillon d'unités primaires d'échantillonnage (UPE), puis on tire un échantillon d'unités dans les UPE sélectionnées. Par exemple, les UPE pourraient être des régions et les unités, des ménages ou des personnes ; ou bien les UPE pourraient être des écoles et les unités, des élèves. Habituellement, les unités provenant d'une même UPE ont tendance à être semblables, de sorte que les valeurs de Y_i et Y_j peuvent être corrélées si i et j appartiennent à la même UPE. À la présente section, nous élaborons un estimateur de Δ qui est approximativement sans biais, même quand il existe des corrélations entre les valeurs de Y dans une même UPE.

Soit s_I l'échantillon d'UPE, tiré de la population U_I. Soit s_g l'échantillon d'unités tirées de l'UPE g, où $g \in s_I$. Soit $r_I = U_I - s_I$ et $r_g = U_g - s_g$. Nous émettons l'hypothèse du modèle (5) et supposons en outre que Y_i et Y_j ne sont pas corrélées pour $i \in g_1$ et $j \in g_2$ si $g_1 \neq g_2$. Les valeurs de Y_i et Y_j peuvent être corrélées si $i \neq j$ avec $i, j \in U_g$.

Soit $\hat{T} = \sum_{i \in s} w_i Y_i$ tout prédicteur et soit $\boldsymbol{d} = \sum_{i \in s} w_i \boldsymbol{x_i} - \boldsymbol{T_x}$. Le biais de $\hat{T}$ est $\boldsymbol{\beta^T d}$, comme à la section 3.2. La variance de $(\hat{T} - T_Y)$ est

$$\begin{aligned}\text{var}[\hat{T} - T_Y] &= \text{var}\left[\sum_{i \in s} (w_i - 1) Y_i - \sum_{i \in r} Y_i\right] \\ &= \text{var}\left[\sum_{g \in s_I}\left(\sum_{i \in s_g} (w_i - 1) Y_i - \sum_{i \in r_g} Y_i\right) - \sum_{g \in r_I} \sum_{i \in U_g} Y_i\right] \\ &= \sum_{g \in s_I} \text{var}\left(\sum_{i \in s_g} (w_i - 1) Y_i - \sum_{i \in r_g} Y_i\right) + \sum_{g \in r_I} \text{var}\left[\sum_{i \in U_g} Y_i\right].\end{aligned}$$

Nous supposons de surcroît que la variance de $\sum_{i \in r_g} Y_i$ et la covariance entre $\sum_{i \in r_g} Y_i$ et $\sum_{i \in s_g} (w_i - 1) Y_i$ sont négligeables comparativement aux autres termes. Il en est ainsi quant on recourt à l'échantillonnage en grappes (parce que, dans ce cas, $s_g = U_g$ and r_g est vide) ou que la fraction d'échantillonnage dans les UPE est faible. La variance devient

$$\text{var}\,[\hat{T} - T_Y] \approx \sum_{g \in s_I} \text{var}\left[\sum_{i \in s_g} (w_i - 1) Y_i\right] + \sum_{g \in r_I} \text{var}\left[\sum_{i \in U_g} Y_i\right].$$

En appliquant cela à Δ, nous obtenons :

$$\begin{aligned}\Delta &= \text{EQMP}[\hat{T}_A] - \text{EQMP}[\hat{T}_B] \\ &= \boldsymbol{d_A^T}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d_A} - \boldsymbol{d_B^T}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d_B} + \sum_{g \in s_I} \text{var}\left[\sum_{i \in s_g} (w_{Ai} - 1) Y_i\right] \\ &\quad - \sum_{g \in s_I} \text{var}\left[\sum_{i \in s_g} (w_{Bi} - 1) Y_i\right]. \qquad (11)\end{aligned}$$

Pour estimer Δ, nous avons besoin des estimateurs de la variance de $\hat{\boldsymbol{\beta}}$ et de $(\sum_{i \in s_g} (w_i - 1) Y_i)$.

En premier lieu, notons que

$$\begin{aligned}\text{var}\,[\hat{\boldsymbol{\beta}}] &= \text{var}\left[S_x^{-1} \sum_{g \in s_I} \sum_{i \in s_g} \boldsymbol{x_i} Y_i\right] \\ &= S_x^{-1} \sum_{g \in s_I} \text{var}\left[\sum_{i \in s_g} \boldsymbol{x_i} Y_i\right] S_x^{-1}.\end{aligned}$$

Nous pouvons estimer cette variance en utilisant la méthode de la variance des grappes finales (*ultimate cluster variance ou UCV*) selon

$$\widehat{\text{var}}_{\text{ucv}}[\hat{\boldsymbol{\beta}}] = S_x^{-1} \sum_{g \in s_I}\left(\sum_{i \in s_g} \boldsymbol{x_i}(Y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{x_i})\right)^2 S_x^{-1}.$$

Cet estimateur bien connu de la variance d'une somme pondérée de données en grappes est équivalent à celui donné par Valliant et coll. (2000, 9.5.5, page 312). Cet estimateur a été appelé « estimateur sandwich de la variance en utilisant les résidus au niveau de la grappe » (Valliant et coll. 2000) et « de la variance des grappes finales » (par exemple, Wolter 1985 décrit essentiellement la même idée dans un cadre de randomisation).

La variance de $(\sum_{i \in s_g} (w_i - 1) Y_i)$ peut aussi être estimée par la méthode d'estimation de la variance des grappes finales :

$$\widehat{\text{var}}\left[\sum_{i\in s_g}(w_i-1)Y_i\right]=\left\{\sum_{i\in s_g}(w_i-1)(Y_i-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)\right\}^2.$$

Donc, nous pouvons estimer Δ par

$$\begin{aligned}\hat{\Delta}_{\text{ucv}} &= \boldsymbol{d}_A^T(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T-\widehat{\text{var}}_{\text{ucv}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d}_A\\ &\quad-\boldsymbol{d}_B^T(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T-\widehat{\text{var}}_{\text{ucv}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d}_B\\ &\quad+\sum_{g\in s_I}\left\{\sum_{i\in s_g}(w_{Ai}-1)(Y_i-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)\right\}^2\\ &\quad-\sum_{g\in s_I}\left\{\sum_{i\in s_g}(w_{Bi}-1)(Y_i-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)\right\}^2. \qquad (12)\end{aligned}$$

4. Étude par simulation

4.1 Simulation d'une enquête sur les fermes

Population et plan d'échantillonnage

La distribution de population des variables auxiliaires, la taille d'échantillon et la taille de population, ainsi que l'hétéroscédasticité et d'autres propriétés de la variable d'intérêt sont des éléments qui, en principe, devraient tous jouer un rôle dans la performance des prédicteurs BLUP adaptatifs. Afin d'évaluer raisonnablement les propriétés de ces estimateurs, il est nécessaire de procéder à une étude par simulation basée sur une grande population réaliste.

Nous avons généré une population simulée de 80 000 unités à l'aide de données d'échantillon sur 1 652 fermes provenant de l'Australian Agricultural and Grazing Industry Survey (AAGIS) de 1988 comme point de départ. Nous avons choisi comme variable d'intérêt le total des cultures commerciales et avons considéré comme variables auxiliaires possibles l'EDT (estimation dérivée de la taille), le nombre de moutons, la superficie cultivée, le nombre de bovins, la région (29 régions) et l'industrie (5 industries). L'EDT était une combinaison linéaire des variables nombre de moutons, superficie cultivée et nombre de bovins. L'ensemble de données contenait aussi un poids d'échantillonnage qui était approximativement égal à l'inverse de la probabilité de sélection. Nous avons éliminé 27 cas aberrants pour lesquels la valeur de l'EDT était très grande, car, dans une enquête, ces cas seraient placés dans une strate entièrement dénombrée. Nous avons ensuite construit une population de 80 000 unités selon un plan d'échantillonnage avec probabilité proportionnelle à la taille avec remise, les probabilités étant proportionnelles au poids d'estimation figurant dans le fichier de l'échantillon original.

Ensuite, nous avons sélectionné 250 échantillons sans remise à partir de la population simulée. Nous avons stratifié les échantillons selon la région et l'EDT, en divisant l'EDT en quatre catégories, ce qui a donné 116 strates. Les limites des catégories ont été déterminées de telle façon que les sommes des catégories de l'EDT soient égales. Nous avons simulé des tailles totales d'échantillon de 250, 500, 1 000 et 1 500. Les tailles d'échantillon de strate étaient proportionnelles aux tailles d'échantillon de l'AAGIS originale selon la région et l'EDT.

Variables auxiliaires et méthode de sélection séquentielle

Nous avons inclus les variables auxiliaires correspondant au modèle contenant une constante (ordonnée à l'origine), le nombre de moutons (x1), la superficie cultivée (x2), le nombre de bovins (x3), l'industrie, l'interaction de l'industrie et de x1, x2 et x3, ainsi que la région. Cela donne un total de 52 variables auxiliaires possibles. Certaines de ces variables sont colinéaires, mais sont néanmoins incluses dans l'ensemble de variables possibles, afin de donner pour le processus de sélection du modèle un choix plus étendu de modèles possibles. Nous avons également considéré l'ensemble de 139 variables auxiliaires contenant cet ensemble, ainsi que l'interaction de la région et de x1, x2 et x3. Nous avons construit les modèles par sélection ascendante, en partant du modèle contenant uniquement la constante. Nous avons ajouté les variables en prenant pour critère l'étape qui réduisait le plus l'EQMP estimée, pour divers estimateurs possibles de Δ. Nous avons également essayé la sélection séquentielle, mais son exécution était considérablement plus lente pour le plus grand ensemble de variables et son utilisation n'a pas amélioré fortement l'efficacité des prédicteurs BLUP adaptatifs.

Nous avons également calculé un prédicteur BLUP adaptatif basé sur le test de signification statistique, en prenant $p < 0,05$ comme seuil d'inclusion. Pour chaque modèle consécutif, nous avons déterminé la signification statistique de chacune des variables non incluses dans le modèle à l'aide d'un test t standard. À chaque étape, nous avons inclus dans le modèle la variable dont la valeur p était la plus faible. Quand il n'a plus été possible d'ajouter d'autres variables ayant un effet significatif, nous avons arrêté la procédure et le modèle résultant est celui qui a été choisi.

Un certain nombre de modifications ont été nécessaires pour que l'algorithme de sélection ascendante fonctionne fiablement : les variables auxiliaires n'ont pas été ajoutées au modèle si la corrélation de paire de Pearson était égale ou supérieure à 0,95 (ou égale ou inférieure à -0,95) ; et les variables n'ont pas été ajoutées si leur ajout donnait lieu à des équations de calage impossibles à résoudre.

Estimateurs utilisés

Nous avons calculé plusieurs prédicteurs BLUP, en incluant toutes les variables auxiliaires, en incluant

uniquement la constante et l'EDT, et en incluant les variables auxiliaires choisies par sélection ascendante en utilisant l'estimateur non robuste de Δ (décrit à la section 3.2) ou l'estimateur robuste à l'hétéroscédasticité de Δ (décrit à la section 3.4), en partant de l'ensemble de 52 et de l'ensemble de 139 variables auxiliaires possibles. (Le grand ensemble de 139 variables a été évalué uniquement pour des tailles d'échantillon égales ou supérieures à 500.)

L'utilisation d'estimateurs ridge (par exemple, Bardsley et Chambers 1984) est un autre moyen d'aborder le problème de sélection des variables, si bien que nous les avons inclus dans la simulation pour comparer leurs propriétés à celles des prédicteurs BLUP adaptatifs. Les estimateurs que nous avons pris en considération jusqu'à présent incluent ou excluent chaque variable. Si une variable est incluse, les poids doivent être calés exactement sur cette variable, en ce sens que $\sum_{i\in s} w_i \boldsymbol{x_i} = \boldsymbol{T_x}$. La régression ridge introduit une pénalité pour l'absence de calage, mais ne requiert pas nécessairement que les poids donnent un calage parfait pour toutes les variables. Dans cette régression, le vecteur de poids d'échantillon $\boldsymbol{w}$ est choisi de manière à minimiser

$$\sum_{i\in s} (w_i - 1)^2 v_i^{-1} + \sum_{j=1}^{p} c_j^{-1}\left(\sum_{i\in s} w_i x_{ij} - T_{xj}\right)^2.$$

Les c_j sont des coefficients de coût non négatifs indiquant avec quelle priorité il faut satisfaire la contrainte de calage j. Une valeur de 0 indique que la contrainte doit être satisfaite précisément, tandis que les coefficients de coût plus élevé accordent moins de poids à la contrainte. Donc, l'estimateur ridge permet une réduction lisse de la dimension effective du modèle, en procédant effectivement à une interpolation entre l'inclusion d'une variable de calage $(c_j = 0)$ et son exclusion $(c_j = \infty)$.

Habituellement, les c_j sont fixés à λc_j^*, où c_j^* reflète une évaluation un peu subjective de l'importance relative de chaque contrainte, et λ est choisi de manière que les poids finaux w_i aient des propriétés raisonnables, par exemple, qu'ils soient tous supérieurs ou égaux à 0, ou à 1. Nous fixons c_j^* à 0 pour la constante (reflétant une ordonnée à l'origine dans le modèle), à 1 pour $x1$, $x2$ et $x3$, à 10 pour les indicateurs de région, à 5 pour les indicateurs d'industrie, et à 100 pour les interactions. Nous avons choisi c_j^* en nous fondant sur les variables qui, à notre avis, étaient susceptibles d'être les plus utiles. Pour chaque échantillon, nous avons déterminé numériquement la valeur λ de façon qu'elle soit la valeur la plus faible permettant que tous les poids soient supérieurs ou égaux à 1.

Nous avons fondé toutes les méthodes sur la même procédure de modélisation de $\text{var}_M[Y_i]$. En premier lieu, nous avons ajusté un simple modèle contenant la constante, x1, x2 et x3 aux valeurs d'échantillon de Y en utilisant les moindres carrés ordinaires. Puis, nous avons procédé à la régression du logarithme des carrés des résidus de ce modèle sur le logarithme de l'EDT. Ensuite, nous avons élevé les valeurs ajustées de ce modèle à la puissance e pour obtenir les estimations de σ_i^2 pour chaque $i \in s$. Nous avons ensuite tronqué les valeurs estimées de σ_i^2 de sorte qu'aucune valeur ne soit supérieure ou inférieure d'un facteur 4 à la valeur médiane. Cet ajustement avait pour but d'éviter les valeurs extrêmes de σ_i^{-2} qui pourraient entraîner une instabilité dans le calcul des estimations par les moindres carrés pondérés de $\hat{\boldsymbol{\beta}}$. Les résultats étaient sensibles dans une certaine mesure à la méthode de modélisation de la variance, particulièrement l'ajustement final pour éviter les valeurs extrêmes. Les prédicteurs BLUP basés sur un modèle de variance grossier avec $\sigma_i^2 \propto \text{DSE}_i$ avaient une variance supérieure de 10 à 20 % à celle des prédicteurs BLUP présentés ici.

Résultats

Le tableau 1 donne la racine de l'erreur quadratique moyenne relative (RRMSE pour *Relative Root Mean Squared Error*) des divers prédicteurs calés. Les quatre premières lignes du tableau correspondent au premier ensemble de variables auxiliaires (52 variables possibles) et les trois dernières, au deuxième ensemble (139 variables auxiliaires possibles). Les biais ne sont pas présentés, mais constituaient généralement une composante relativement petite de l'erreur quadratique moyenne pour tous les prédicteurs présentés, sauf le prédicteur BLUP basé sur un modèle contenant la constante et l'EDT, qui était relativement biaisé. Ce résultat était quelque peu surprenant, car nous nous attendions à ce qu'un bon compromis entre le biais et la variance implique que les biais soient une composante non négligeable de l'erreur quadratique moyenne. Des détails sur les biais et les variances relatives des prédicteurs peuvent être consultés dans les tableaux A1 et A2 de Clark et Chambers (2008).

Parmi les prédicteurs BLUP adaptatifs, les critères de signification ont donné les meilleurs résultats dans tous les cas, suivis par les critères de non-robustesse, tandis que les critères de robustesse ont donné les pires résultats. Pour le petit ensemble de 52 variables possibles, les prédicteurs BLUP adaptatifs basés sur les critères de non-robustesse et les critères de signification ont donné de meilleurs résultats que les prédicteurs BLUP non adaptatifs pour $n = 250$ et $n = 500$; pour $n = 1\,000$ et $n = 1\,500$, ils ont donné d'un peu moins bons résultats que les prédicteurs BLUP contenant toutes les variables, mais de meilleurs résultats que le prédicteur BLUP contenant la constante et la variable de taille. Pour le grand ensemble de 139 variables possibles, les prédicteurs BLUP adaptatifs basés sur les critères de non-robustesse et de signification ont donné de meilleurs

résultats que les prédicteurs BLUP non adaptatifs pour toutes les tailles d'échantillon, particulièrement pour les faibles valeurs de n.

Tableau 1
RRMSE (%) des prédicteurs du total des cultures commerciales pour l'AAGIS

N^{bre} de var.	n	BLUP		BLUP adaptif			Ridge
		Toutes	C^{ste} + taille	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	Test de sign.	
52	250	3,59	3,02	2,97	3,09	2,87	3,30
	500	2,35	2,54	2,33	2,33	2,30	2,31
	1 000	1,56	2,21	1,58	1,64	1,57	1,54
	1 500	1,36	2,22	1,39	1,41	1,37	1,37
139	500	3,52	2,54	2,99	3,44	2,29	2,27
	1 000	1,77	2,21	1,75	1,92	1,72	1,59
	1 500	1,56	2,22	1,51	1,64	1,42	1,42

En général, l'estimateur ridge a donné des résultats presque aussi bons que le meilleur des prédicteurs BLUP adaptatifs quand nous avons utilisé 52 variables auxiliaires et des résultats un peu meilleurs quand nous nous sommes servis de 139 variables auxiliaires possibles.

Le tableau 2 montre combien de variables auxiliaires ont été sélectionnées pour les deux prédicteurs BLUP adaptatifs. L'estimateur $\hat{\Delta}$ robuste a donné de plus grands ensembles de variables auxiliaires que l'estimateur non robuste, environ 10 variables auxiliaires supplémentaires ayant été sélectionnées. Les critères de signification ont produit des ensembles de variables encore plus petits (6 à 10 variables de moins que dans le cas des critères de non-robustesse).

Tableau 2
Moyenne (intervalle interquartile) du nombre de variables auxiliaires sélectionnées dans l'AAGIS

N^{bre} de var.	n	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	Test de sign.
52	250	16,0 (14,0-18,0)	26,9 (24,0-29,0)	9,6 (8,0-11,0)
	500	18,6 (16,0-21,0)	27,4 (25,0-30,0)	11,5 (10,0-13,0)
	1 000	23,6 (21,0-26,0)	29,6 (26,0-33,0)	14,4 (13,0-16,0)
	1 500	27,3 (25,0-29,0)	32,3 (30,0-35,0)	17,2 (16,0-18,8)
139	500	42,1 (37,0-47,0)	69,4 (62,0-75,0)	23,2 (21,0-26,0)
	1 000	51,5 (47,0-56,0)	74,2 (69,0-79,8)	29,9 (27,0-33,0)
	1 500	59,2 (55,0-64,0)	75,8 (71,0-81,0)	34,9 (32,0-38,0)

Le tableau 3 donne la non-couverture de l'intervalle de confiance (IC) des divers prédicteurs. Nous avons défini l'IC à 90 % comme étant l'estimateur +/- 1,64 erreur-type, où la variance était estimée en utilisant l'estimateur de variance robuste à l'hétéroscédasticité (Royall et Cumberland 1978). Conformément à une pratique courante, nous avons fondé les IC sur la variance estimée et non sur l'erreur quadratique moyenne estimée de prédiction. Les estimations par simulation des taux de non-couverture sont assez approximatives, étant donné que nous n'avons utilisé que 250 simulations. Une plus grande étude par simulation peut être effectuée pour obtenir des estimations plus précises de la couverture, mais nous n'avons pas poursuivi cette option étant donné que le processus de sélection séquentielle est gourmand en ressources informatiques. Le tableau 3 donne à penser que le prédicteur BLUP n'utilisant que la constante et la variable de taille donne une non-couverture importante, de même que le prédicteur BLUP adaptatif basé sur l'estimateur $\hat{\Delta}$ robuste. En général, le taux de non-couverture des autres estimateurs est proche des 10 % nominaux.

Tableau 3
Non-couverture de l'intervalle de confiance dans l'AAGIS

N^{bre} de var.	n	BLUP		BLUP adaptif			Ridge
		Toutes	C^{ste} + taille	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	Test de sign.	
52	250	10,0	6,4	10,4	16,8	11,2	10,0
	500	8,0	13,2	12,0	17,2	10,8	8,0
	1 000	7,6	20,4	9,2	12,0	8,4	8,4
	1 500	8,8	34,8	9,2	13,2	9,6	8,8
139	500	16,8	13,2	18,0	29,2	12,8	8,8
	1 000	12,4	20,4	14,0	20,4	13,2	7,2
	1 500	13,6	34,8	13,6	19,6	12,4	11,2

Le total des cultures commerciales est une variable d'intérêt majeure de l'AAGISS, mais les totaux d'autres variables sont importants également, y compris les capitaux propres des fermes. Pour des raisons pratiques, un seul ensemble de poids est normalement utilisé pour toutes les variables. Le tableau 4 montre dans quelle mesure les poids de calage adaptatifs conçus pour le total des cultures commerciales (TCC) donnent de bons résultats quand on les utilise pour estimer le total des capitaux propres des fermes. Dans le cas où 52 variables auxiliaires possibles sont utilisées, les poids BLUP adaptatifs choisis en se fondant sur la variable TCC (en utilisant l'estimateur $\hat{\Delta}$ non robuste) ont donné d'assez bons résultats, de même que l'estimateur ridge. Cependant, des améliorations pourraient être apportées en choisissant les variables auxiliaires en ne se basant que sur la variable des capitaux propres.

Tableau 4
RRMSE (%) des prédicteurs du total des capitaux propres dans l'AAGIS

N^{bre} de var.	n	BLUP		BLUP adaptatif ($\hat{\Delta}$ non robuste)		Ridge
		Toutes	C^{ste} + taille	Basé sur TCC	Basé sur capitaux propres	
52	250	6,85	6,45	6,51	6,13	6,78
	500	4,44	4,44	4,61	4,40	4,28
	1 000	3,09	3,12	3,42	3,14	3,10
	1 500	2,54	2,58	2,90	2,58	2,54
139	500	5,53	4,93	4,98	4,74	4,20
	1 000	3,68	4,03	3,23	3,15	3,08
	1 500	3,04	3,63	2,66	2,60	2,57

4.2 Simulation de l'enquête sur la population active

Population et plan d'échantillonnage

Nous avons construit une population simulée en tirant un échantillon aléatoire simple sans remise de 30 000 personnes de 15 à 64 ans du fichier de l'échantillon à 1 % de l'Australian Census of Population and Housing de 1991. La variable d'intérêt était l'emploi (1 pour les personnes occupées, 0 pour les autres). Nous avons divisé la population simulée en unités primaires d'échantillonnage (UPE) simulées contenant 75 personnes chacune, de telle sorte que la corrélation intra-grappe soit égale à 0,05. (Il s'agit d'une valeur assez typique de la corrélation intra-grappe pour la variable d'emploi dans les unités primaires d'échantillonnage d'une enquête-ménage (voir, par exemple, Clark et Steel 2002). L'algorithme de définition des grappes devait trier les données en fonction d'une variable généré aléatoirement de loi $N(0, \gamma^2)$ plus la variable d'emploi, puis définir les grappes comme des ensembles séquentiels de 75 personnes, où γ était choisi de manière à obtenir la corrélation intra-grappe souhaitée.

La simulation consistait en 250 échantillons à deux degrés répétés. Le premier degré était un échantillon aléatoire simple sans remise de m UPE et le deuxième degré était un échantillon aléatoire simple sans remise de 20 personnes provenant de chaque UPE sélectionnée. La taille totale d'échantillon a été fixée à $n = 200, 400$ et 1 000 personnes. La plupart des enquêtes-ménages nationales sont réalisée auprès d'échantillons de taille beaucoup plus grande que cela, mais il est courant, pour l'estimation, de construire des poststrates dans les strates ou les provinces, et les tailles d'échantillon pour ces régions sont souvent de l'ordre de 200 à 1 000.

Les variables auxiliaires possibles étaient l'âge selon le sexe, où l'âge était enregistré en année simple pour le groupe des 16 à 24 ans, puis en tranche de cinq ans pour les groupes des 25 à 29 ans, 30 à 34 ans, ..., 55 à 59 ans et 60 ans et plus.

Non-réponse

L'une des raisons principales pour lesquelles on se sert de l'âge et du sexe comme variables auxiliaires dans les enquêtes-ménages est que l'on sait que la non-réponse dépend de ces variables. Par exemple, les jeunes hommes représentent habituellement le groupe dont le taux de réponse est le plus faible. Nous avons simulé la non-réponse en supposant que le logit de la probabilité de réponse était égal à $1{,}8 - ((\hat{a}ge - 50)/25)^2$ pour les hommes, et à $2 - 0{,}7((\hat{a}ge - 50)/25)^2$ pour les femmes. Ce modèle donne un taux de réponse de 75 %. Nous avons augmenté la taille de l'échantillon initial de manière que la taille de l'échantillon final de répondants soit égale à $n = 200, 400$ ou 1 000.

Variables auxiliaires et méthode de sélection séquentielle

Nous avons choisi les variables auxiliaires possibles en nous basant sur les cellules déterminées par l'âge selon le sexe. La définition des variables x est donnée au tableau 5. Nous avons choisi cette paramétrisation afin que les variables auxiliaires correspondant à des âges ou groupes d'âge particuliers puissent être abandonnées tout en obtenant encore un modèle raisonnable. Par exemple, si toutes les variables auxiliaires étaient incluses, sauf x_{4i}, la valeur espérée selon le modèle pour les personnes de 17 ans serait la même que celle pour les personnes de 16 ans, au lieu d'être égale à la constante. D'encore meilleurs résultats pourraient être obtenus en utilisant des paramétrisations plus perfectionnées, comme des modèles spline, ce que nous explorerons dans une étude à venir.

Tableau 5
Variables auxiliaires possibles dans la simulation de l'enquête sur la population active

Variable	Définition
x_{1i}	1 (correspond à la constante dans le modèle pour Y)
x_{2i}	1 si la personne i est un homme, -1 si elle est une femme
x_{3i}	1 si la personne i a 16 ans ou plus
x_{4i}	1 si la personne i a 17 ans ou plus
⋮	⋮
$x_{12,i}$	1 si la personne i a 25 ans ou plus
$x_{13,i}$	1 si la personne i a 30 ans ou plus
⋮	⋮
$x_{19,i}$	1 si la personne i a 60 ans ou plus
$x_{20,i}$	x_{3i} si la personne i est un homme $-x_{3i}$ si elle est une femme
⋮	⋮
$x_{36,i}$	$x_{19,i}$ si la personne i est un homme $-x_{19,i}$ si elle est une femme

Nous avons utilisé la sélection séquentielle pour choisir les variables, en partant du modèle ne contenant que le terme constant. À chaque étape, des variables pouvaient être ajoutées ou supprimées, en choisissant celles donnant la meilleure réduction du critère. Quand la sélection séquentielle commençait à tourner en rond (par exemple, ajout de x1, puis ajout de x2, puis suppression de x1, puis suppressions de x2, puis ajout de x1, *etc.*), le processus de construction du modèle était arrêté et le modèle résultant était utilisé comme modèle final. Les estimateurs de Δ pris en considération étaient l'estimateur non robuste, l'estimateur robuste (à l'hétéroscédasticité) et l'estimateur de la variance des grappes finales (UCV) qui est robuste à l'hétéroscédasticité et aux corrélations dans les UPE. Nous n'avons pas utilisé les tests de signification, car ils auraient nécessité l'intégration des corrélations dans les UPE pour être réalistes. Nous ne présentons pas les résultats pour l'estimateur ridge, parce que des points négatifs ont

rarement été produits par cette simulation, de sorte que les résultats étaient fort semblables à ceux des prédicteurs BLUP utilisant toutes les variables auxiliaires.

Résultats

Le tableau 6 donne la RRMSE des divers prédicteurs BLUP adaptatifs et non adaptatifs. L'écart entre les RRMSE du prédicteur BLUP ne contenant que le terme constant et du prédicteur BLUP contenant toutes les variables auxiliaires était assez faible. Il n'est par conséquent pas étonnant que les meilleurs gains mineurs aient été réalisés en utilisant les prédicteurs BLUP adaptatifs plutôt que le prédicteur BLUP contenant toutes les variables. Le prédicteur BLUP adaptatif utilisant l'estimateur $\hat{\Delta}$ non robuste est celui qui a donné la RRMSE la plus faible dans tous les cas.

Tableau 6
RRMSE des prédicteurs de l'emploi dans l'enquête sur la population active

n	BLUP		BLUP adaptatif		
	Toutes	C^{ste}	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	$\hat{\Delta}$ UCV
200	6,54	6,77	6,44	7,06	6,96
400	4,72	4,76	4,61	4,72	4,65
1 000	2,45	2,70	2,43	2,45	2,49

Le tableau 7 donne le nombre moyen de variables sélectionnées pour chacun des prédicteurs BLUP adaptatifs. Des 36 variables auxiliaires possibles, de cinq à neuf environ ont été sélectionnées en se basant sur l'estimateur $\hat{\Delta}$ non robuste. Le nombre de variables sélectionnées augmente parallèlement à la taille d'échantillon. Le critère de robustesse à l'hétéroscédasticité a produit de plus grands ensembles de variables auxiliaires et le critère de variance des grappes finales (UCV) a donné d'encore plus grands ensembles.

Tableau 7
Moyenne (intervalle interquartile) du nombre de variables auxiliaires sélectionnées dans la simulation de la population active

n	Méthode de sélection des variables		
	Non robuste	Robuste	UCV
200	6,5 (5,0- 8,0)	13,4 (10,0-16,0)	16,1 (13,0-19,0)
400	7,4 (6,0- 8,0)	12,1 (9,0-15,0)	14,5 (12,0-17,0)
1 000	8,6 (7,0-10,0)	11,6 (10,0-13,0)	14,2 (12,0-17,0)

Le tableau 8 donne la non-couverture de l'intervalle de confiance (IC) de divers prédicteurs. Les IC à 90 % ont été définis comme étant l'estimateur +/-1,64 erreur-type où la variance a été estimée à l'aide de la méthode de la variance des grappes finales (UVC). L'examen du tableau révèle que le prédicteur BLUP contenant toutes les variables auxiliaires donne lieu à une non-couverture importante pour $n = 200$ et 400. Le prédicteur BLUP adaptatif utilisant l'estimateur $\hat{\Delta}$ non robuste produit une couverture raisonnablement proche de la couverture nominale, tandis que les autres prédicteurs BLUP adaptatifs donnent une non-couverture importante.

Tableau 8
Non-couverture de l'intervalle de confiance (%) des prédicteurs de l'emploi

n	BLUP		BLUP adaptatif		
	Toutes	C^{ste}	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	$\hat{\Delta}$ UCV
200	17,6	12,0	12,0	20,0	24,0
400	17,2	12,0	14,8	16,8	17,6
1 000	6,4	11,6	7,6	6,8	9,6

Le tableau 9 montre dans quelle mesure les divers poids ont donné de bons résultats quand ils ont été utilisés pour estimer une variable différente, à savoir le chômage (égale à 1 pour les chômeurs et 0 autrement). Nous avons calculé les prédicteurs BLUP adaptatifs en utilisant l'estimateur $\hat{\Delta}$ non robuste et en prenant comme variable d'intérêt l'emploi, ainsi que le chômage. Le prédicteur BLUP adaptatif contenant des variables choisies pour l'emploi avait une RRMSE comprise entre celle du prédicteur BLUP non adaptatif contenant toutes les variables et celle du prédicteur BLUP non adaptatif ne contenant que la constante. Cela donne à penser que ce prédicteur BLUP adaptatif donne des résultats raisonnables, même s'il est appliqué à d'autres variables que l'emploi. En fait, le prédicteur BLUP adaptatif basé sur le chômage avait une RRMSE plus élevée, peut-être parce que les variables auxiliaires n'avaient que peu de pouvoir prédictif, voire aucun, pour le chômage, de sorte qu'essayer d'adapter le choix des variables auxiliaires à cette variable d'intérêt n'a pas donné de bons résultats.

Tableau 9
RRMSE des prédicteurs du chômage dans l'enquête sur la population active

n	BLUP		BLUP adaptatif	
	Toutes	C^{ste}	Basé sur l'emploi	Basé sur le chômage
200	36,3	32,6	34,5	36,0
400	24,1	21,7	22,8	23,7
1 000	14,5	14,2	14,1	14,2

5. Discussion

Les études par simulation décrites ici montrent que les prédicteurs BLUP adaptatifs peuvent donner des améliorations utiles comparativement à de simples options non adaptatives. En ce qui concerne tant les simulations de l'enquête sur les fermes que celles de l'enquête sur la population active, les prédicteurs BLUP adaptatifs basés sur

un estimateur non robuste de Δ et ceux basés sur un test de signification ont les uns et les autres produit une EQMP plus faible que les estimateurs non adaptatifs dans la plupart des cas. Pour ce qui est de l'enquête sur les fermes, les améliorations étaient parfois considérables comparativement à l'option consistant à systématiquement utiliser le modèle complet ou celle consistant à systématiquement utiliser le modèle contenant la constante plus la variable de taille. Dans le cas de l'enquête sur la population active, les améliorations étaient mineures. Les prédicteurs BLUP adaptatifs ont également donné une couverture raisonnable des intervalles de confiance.

Les prédicteurs BLUP adaptatifs basés sur les critères d'estimateur robuste et par la méthode de la variance des grappes finales (UCV) ont donné de nettement moins bons résultats que les autres. Cette constatation est étonnante, car il est connu que les données de l'AAGIS sont hétéroscédastiques et que celles de l'enquête sur la population active sont en grappes, ce qui laisse entendre que l'estimation de la variance par la méthode UCV aurait dû donner de bons résultats. Une analyse plus approfondie de la simulation de l'enquête sur les fermes a montré que, dans la grande majorité des cas, $\hat{\Delta}_{\text{robust}}$ avait une variance plus élevée que $\hat{\Delta}_{\text{nonrobust}}$, particulièrement pour les variables auxiliaires dont le pouvoir prédicteur était faible – voir l'annexe de Clark et Chambers 2008 pour des renseignements plus détaillés. Il semble donc que la méthode robuste ait tendance à aboutir plus fréquemment à la sélection de variables auxiliaires contre-productives, ce qui pourrait expliquer sa performance médiocre.

Nous avons constaté une tendance générale de toutes les méthodes adaptatives à choisir un trop grand nombre de variables auxiliaires, mais malgré cela, les estimateurs adaptatifs ont généralement donné des résultats supérieurs ou comparables à ceux des estimateurs simples non adaptatifs. Selon nous, en pratique, il faudrait conjuguer une recherche automatique de modèle (en utilisant un estimateur $\hat{\Delta}$ non robuste ou un critère de signification statistique) et une certaine forme de jugement subjectif. Par exemple, des modèles pourraient être sélectionnés parmi plusieurs ensembles de variables auxiliaires possibles de tailles différentes. Si les grands ensembles ne donnent que des améliorations apparemment faibles, le statisticien pourrait décider de se limiter à un ensemble plus petit, même s'il paraît légèrement sous-optimal.

Les estimateurs ridge ont aussi donné d'assez bons résultats en ce qui concerne la RRMSE et la couverture de l'intervalle de confiance. Dans l'ensemble, ils ont produit des résultats comparables à ceux des prédicteurs BLUP adaptatifs pour l'estimation du total de la variable d'intérêt quand le choix des variables auxiliaires était fondé sur cette dernière. Cependant, quand nous avons appliqué les poids BLUP adaptatifs à des variables différentes, les estimateurs ridge ont donné des résultats légèrement supérieurs. Une approche encore meilleure pourrait consister à choisir adaptativement les variables auxiliaires qu'il convient d'inclure dans le modèle, ainsi que la façon d'appliquer la régression ridge, en se fondant sur un critère calculé d'après l'échantillon. Cette approche sera le sujet de travaux de recherche à venir.

L'une des réserves qui a été formulée concernant l'approche par prédiction de l'échantillonnage en population finie tient à sa non-robustesse à l'omission de variables auxiliaires importantes. Dans nos simulations s'appuyant sur des données économiques sur les fermes et sur des données sociales, les prédicteurs adaptatifs avaient un biais faible et une erreur quadratique moyenne plus faible que les estimateurs non adaptatifs dans la plupart des nombreux cas examinés dans notre étude par simulation et n'étaient jamais considérablement plus mauvais. À condition que toutes les variables de plan soient considérées comme des variables auxiliaires possibles, le calage adaptatif offre une stratégie robuste et efficace de prédiction en population finie.

Remerciements

Les travaux ont été financés conjointement par l'Australian Research Council et l'Australian Bureau of Statistics. Un rédacteur associé et deux examinateurs ont formulé des commentaires détaillés et judicieux qui nous ont permis d'améliorer considérablement l'article.

Bibliographie

Bardsley, P., et Chambers, R. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33(3), 290-299.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

Chambers, R., Skinner, C. et Wang, S. (1999). Intelligent calibration. *Bulletin of the International Statistical Institute*, 58(2), 321-324.

Clark, R.G., et Chambers, R.L. (2008). *Adaptive calibration for prediction of finite population totals*. University of Wollongong Centre for Statistical and Survey Methodology. Consultable à http://www.cssm.uow.edu.au.

Clark, R.G., et Steel, D.G. (2002). The effect of using household as a sampling unit. *Revue Internationale de Statistique*, 70(2), 289-314.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Hansen, M., Madow, W. et Tepping, B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2), 377-387.

Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71(355), 657-664.

Royall, R.M., et Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73(362), 351-358.

Royall, R.M., et Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76(373), 66-80.

Royall, R.M., et Cumberland, W.G. (1981b). The finite-population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76(376), 924-930.

Royall, R.M., et Herson, J. (1973a). Robust estimation in finite populations 1. *Journal of the American Statistical Association*, 68(344), 880-889.

Royall, R.M., et Herson, J. (1973b). Robust estimation in finite populations 2: Stratification on a size variable. *Journal of the American Statistical Association*, 68(344), 890-893.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model assisted survey sampling*. New York : Springer-Verlag.

Silva, P.L.N., et Skinner, C. (1997). Sélection des variables pour l'estimation par régression dans le cas des populations finies. *Techniques d'enquête*, 23, 25-35.

Skinner, C., et Silva, P.L.N. (1997). Variable selection for regression estimation in the presence of nonresponse. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 76-81.

Smith, T.M.F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society*, Series A, 139(2), 183-202.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*: A Prediction Approach. New York : John Wiley & Sons, Inc.

Wolter, K.M. (1985). *Introduction to variance estimation*. New York : Springer-Verlag.