Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

L'agrégation bootstrap des estimateurs non différenciables dans les enquêtes complexes

par Jianqiang C. Wang, Jean D. Opsomer et Haonan Wang

Date de diffusion : 19 décembre 2014

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 ».

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

© Ministre de l'Industrie, 2014

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada (http://www.statcan.gc.ca/reference/copyright-droit-auteur-fra.htm).

This publication is also available in English.

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
0^s	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 2014
Vol. 40, n° 2, pp. 211-233
Statistique Canada, n° 12-001-X au catalogue

L'agrégation bootstrap des estimateurs non différenciables dans les enquêtes complexes

Jianqiang C. Wang, Jean D. Opsomer et Haonan Wang[1]

Résumé

L'agrégation bootstrap est une puissante méthode de calcul utilisée pour améliorer la performance des estimateurs inefficaces. Le présent article est le premier à explorer l'utilisation de l'agrégation bootstrap dans l'estimation par sondage. Nous y examinons les effets de l'agrégation bootstrap sur les estimateurs d'enquête non différenciables, y compris les fonctions de répartition de l'échantillon et les quantiles. Les propriétés théoriques des estimateurs d'enquête agrégés par bootstrap sont examinées sous le régime fondé sur le plan de sondage et le régime fondé sur le modèle. En particulier, nous montrons la convergence par rapport au plan des estimateurs agrégés par bootstrap et obtenons la normalité asymptotique des estimateurs dans un contexte fondé sur le modèle. L'article explique comment la mise en œuvre de l'agrégation bootstrap des estimateurs d'enquête peut tirer parti des répliques produites pour l'estimation par sondage de la variance, facilitant l'application de l'agrégation bootstrap dans les enquêtes existantes. Un autre défi important dans la mise en œuvre de l'agrégation bootstrap en contexte d'enquête est l'estimation de la variance pour les estimateurs agrégés par bootstrap eux-mêmes, et nous examinons deux façons possibles d'estimer la variance. Les expériences par simulation révèlent une amélioration de l'estimateur par agrégation bootstrap proposé par rapport à l'estimateur original et comparent les deux approches d'estimation de la variance.

Mots-clés : Bootstrap; fonction de distribution; estimation des quantiles.

1 Introduction

L'agrégation bootstrap, appelée « bagging » en anglais, est une méthode de rééchantillonnage initialement introduite pour améliorer les algorithmes d'apprentissage « faibles ». L'agrégation bootstrap a été proposée par Breiman (1996), qui a démontré de façon heuristique qu'elle améliorait la performance des prédicteurs à structure arborescente. L'agrégation bootstrap a depuis été appliquée dans un large éventail de contextes et analysée par de nombreux auteurs. Bühlmann et Yu (2002) ont démontré les effets de lissage de l'agrégation bootstrap et de ses variations sur les algorithmes de classification des décisions difficiles et formalisé la notion de « prédicteurs instables ». Chen et Hall (2003) ont dérivé des résultats théoriques de l'agrégation bootstrap d'estimateurs définis par des équations d'estimation. Buja et Stuetzle (2006) ont envisagé l'agrégation bootstrap des statistiques U et soutenu que l'agrégation bootstrap [*Traduction*] « réduit souvent, mais pas toujours, la variance, mais augmente toujours le biais ». Friedman et Hall (2007) ont examiné l'incidence de l'agrégation bootstrap sur les estimateurs non linéaires. Plus récemment, Hall et Robinson (2009) ont discuté des effets de l'agrégation bootstrap sur le choix par validation croisée des paramètres de lissage et présenté des résultats intrigants concernant l'amélioration, par agrégation bootstrap, de l'ordre du choix par validation croisée de la bande passante du noyau.

La littérature susmentionnée étudiait les effets de l'agrégation bootstrap sur différents estimateurs, particulièrement les estimateurs non différenciables non linéaires, sous l'hypothèse d'échantillonnage de données *iid* (indépendantes et identiquement distribuées). Pour les données dépendantes, Lee et Yang

1. Jianqiang C. Wang, Hewlett-Packard Labs, Palo Alto, CA 94304. Jean D. Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523. Courriel : jopsomer@stat.colostate.edu; Haonan Wang, Department of Statistics, Colorado State University, Fort Collins, CO 80523.

(2006) ainsi que Inoue et Kilian (2008) ont étudié les effets de l'agrégation bootstrap sur les séries chronologiques économiques. Les premiers ont étudié les effets de l'agrégation bootstrap sur les prédicteurs non différenciables comme les fonctions de signe et les quantiles, tandis que les seconds ont mis l'accent sur l'agrégation bootstrap des prédicteurs de prétest applicables à la prévision de l'inflation des prix à la consommation aux États-Unis.

Comme le montre ce bref examen de la littérature, l'agrégation bootstrap est une méthode prometteuse utilisée pour améliorer l'efficacité des estimateurs. L'agrégation bootstrap pour les estimateurs d'enquête n'a toutefois pas été envisagée jusqu'ici. Le présent article est le premier à examiner l'utilisation de l'agrégation bootstrap en contexte d'enquête; il comprend une évaluation du gain d'efficacité potentiel, avance un certain nombre de résultats théoriques, et explore les questions de mise en œuvre et d'estimation de la variance. Conformément aux pratiques générales d'enquête, nous examinons seulement les estimateurs qui peuvent s'exprimer sous forme de fonctions de Horvitz-Thompson (HT). Plus précisément, nous étudions trois types d'estimateurs. Premièrement, de nombreux estimateurs d'usage courant peuvent s'exprimer en tant que fonctions différenciables des estimateurs de HT. Par exemple, l'estimateur de Hajek, l'estimateur par le ratio et l'estimateur par la régression généralisée peuvent tous être considérés comme des fonctions différenciables des estimateurs de HT. Deuxièmement, il existe d'autres estimateurs d'enquête non différenciables, dont ceux de Dunstan et Chambers (Dunstan et Chambers 1986) et de Rao-Kovar-Mantel (Rao, Kovar et Mantel 1990), l'estimateur de post-stratification endogène (Breidt et Opsomer 2008) et les estimateurs de proportion de personnes à faible revenu (Berger et Skinner 2003). Troisièmement, d'autres estimateurs sont définis seulement comme solutions d'équations d'estimation pondérées. Pour plus de renseignements sur les équations d'estimation en contexte d'enquête, voir Godambe et Thompson (2009), Fuller (2009) et leurs références.

Bien que l'agrégation bootstrap puisse être considérée comme un type de méthode de répliques, elle est très différente de la méthode bootstrap et d'autres méthodes de répliques conçues pour estimer la variance. Contrairement à ces méthodes, l'agrégation bootstrap a pour but d'améliorer l'estimateur même. L'agrégation bootstrap peut être naturellement intégrée aux enquêtes complexes à grande échelle, car nous pouvons tirer parti des poids de réplication facilement disponibles dans de nombreuses enquêtes pratiques. Dans le présent article, nous montrons comment les répliques créées pour l'estimation bootstrap de la variance peuvent être modifiées et utilisées dans l'agrégation bootstrap de l'estimateur original. Malheureusement, une difficulté inhérente à l'application de l'agrégation bootstrap dans les enquêtes est l'absence d'estimateur de variance fondé sur le plan de sondage. Nous examinons un certain nombre de méthodes proposées pour estimer la variance des estimateurs agrégés par bootstrap, mais il reste du travail à faire dans ce domaine.

Le reste de cet article est organisé comme suit. Dans la section 2, nous définissons nos estimateurs cibles et présentons la version agrégée par bootstrap de chaque estimateur. Dans la section 3, nous présentons les propriétés théoriques des estimateurs agrégés par bootstrap. Dans la section 4, nous montrons comment utiliser les répliques pour appliquer les versions agrégées par bootstrap des estimateurs, et nous examinons l'estimation de la variance pour les estimateurs agrégés par bootstrap résultants. Dans la section 5, nous exposons les résultats des simulations et, dans la section 6, nous présentons quelques conclusions et remarques finales.

2 Agrégation bootstrap des estimateurs

2.1 Approche générale

Dans cette section, nous discutons de la mise en œuvre de l'agrégation bootstrap en contexte d'estimation par sondage. Nous commençons par présenter la notation nécessaire. Soit U une population finie de taille N, où chaque élément $i \in U$ est associé à un vecteur de mesures $\mathbf{y}_i$, dans l'espace euclidéen $\mathbb{R}^q$ à q dimensions. Nous utilisons le plan d'échantillonnage $p(\,)$ pour tirer un échantillon aléatoire $A \subseteq U$ de taille n. Soit $\mathcal{Y} = \{\mathbf{y}_i \mid i \in A\}$ l'ensemble des observations de l'échantillon. Ici, le plan d'échantillonnage pourrait être un échantillonnage aléatoire simple sans remise (EASSR), un échantillonnage de Poisson ou un plan de sondage complexe qui prévoit une stratification et/ou un échantillonnage à plusieurs degrés. Dans chaque plan, la probabilité qu'un élément i soit inclus dans l'échantillon est π_i.

La moyenne de population du vecteur de mesure $\mathbf{y}$ est $\boldsymbol{\mu}$. Elle peut être estimée au moyen de l'estimateur de Horvitz-Thompson (HT) défini comme étant

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i \in A}\frac{\mathbf{y}_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{y}_i}{\pi_i} I_i, \tag{2.1}$$

où I_i est l'**indicateur d'appartenance à l'échantillon** pour le i-ième élément. De façon plus générale, soit θ une quantité d'intérêt de la population et $\hat{\theta}(\mathcal{Y})$ l'estimateur de θ selon les observations de l'échantillon $\mathcal{Y}$. L'estimateur $\hat{\theta}(\mathcal{Y})$ est abrégé en $\hat{\theta}$ s'il n'y a aucune confusion. Comme il est noté dans la section qui précède, nous supposons que $\hat{\theta}$ peut s'exprimer en tant que fonction d'estimateurs simples de la forme (2.1).

Sous sa forme la plus générale, l'algorithme d'agrégation bootstrap pour l'estimation par sondage se présente comme suit :

1. Pour $b = 1, 2, \ldots, B$:
    a. Tirer le nouvel échantillon A_b de l'échantillon aléatoire A, et désigner les observations du nouvel échantillon par $\mathcal{Y}_b^* = \{\mathbf{y}_i \mid i \in A_b\}$.
    b. Calculer l'estimation du paramètre en se fondant sur le nouvel échantillon A_b, désigné par $\hat{\theta}(\mathcal{Y}_b^*)$.

2. Faire la moyenne sur les estimations répétées $\hat{\theta}(\mathcal{Y}_1^*)$, $\hat{\theta}(\mathcal{Y}_2^*)$, …, $\hat{\theta}(\mathcal{Y}_B^*)$ afin d'obtenir l'estimateur agrégé par bootstrap,

$$\hat{\theta}_{bag} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}(\mathcal{Y}_b^*). \tag{2.2}$$

Dans la littérature sur l'agrégation bootstrap, les nouveaux échantillons A_b sont souvent appelés **échantillons bootstrap** (Breiman 1996) et c'est ce que nous faisons ici, même si nous ne les utilisons pas pour estimer la variance.

Dans l'algorithme, les échantillons bootstrap pouvaient être tirés selon le plan d'échantillonnage plutôt que par répartition empirique des observations de l'échantillon, méthode qui est plus couramment utilisée

dans la littérature sur l'agrégation bootstrap ordinaire (Breiman 1996) et qui équivaut à un échantillonnage aléatoire simple (avec ou sans remise). Par exemple, si l'échantillon A est tiré par échantillonnage stratifié ou en grappes, ce plan de sondage pourrait être pris en considération lors de la sélection des nouveaux échantillons. Plus généralement en contexte d'enquête, l'étape 1 de l'algorithme d'agrégation bootstrap proposé peut être traitée dans le cadre d'un échantillonnage à deux phases, la première phase correspondant à l'échantillon original A et la deuxième, au nouvel échantillon A_b. Ainsi, l'estimateur par extension classique pour plans de sondage à deux phases de Särndal, Swensson et Wretman (1997) est mis en œuvre pour calculer l'estimateur répété $\hat{\theta}\left(\mathcal{Y}_b^*\right)$. Dans le nouvel échantillon A_b, la pseudo probabilité d'inclusion pour le i-ième élément est $\pi_i^* = \pi_i \pi_{i|A}$ où $\pi_{i|A} = \Pr\left(i \in A_b \mid i \in A\right)$ est la probabilité d'inclusion du i-ième élément du nouvel échantillon A_b étant donné son inclusion dans l'échantillon A. Par conséquent, l'estimateur agrégé par bootstrap est une approximation de l'espérance de l'estimateur à deux phases en ce qui concerne la deuxième phase de l'échantillonnage, qui est aussi appelée **espérance bootstrap** dans les méthodes d'agrégation bootstrap ordinaire (Bühlmann et Yu 2002). Bien qu'un plan d'échantillonnage bootstrap général soit possible, nous nous limitons à l'EASSR dans les parties théoriques de cet article. Pour élargir la portée de notre discussion, dans la section sur l'estimation de la variance et dans la section numérique, nous présentons le cas où les échantillons bootstrap sont tirés par EASSR stratifié avec les mêmes strates que l'échantillon original A, ce qui est une extension utile et réaliste.

Prenons, par exemple, l'estimateur de HT défini en (2.1). Le rééchantillonnage bootstrap de l'échantillon réalisé A est tiré sous EASSR de taille k. Selon ce plan de rééchantillonnage, l'estimateur par échantillon répété est défini comme suit :

$$\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right) = \frac{1}{N}\sum_{i \in A_b} \frac{\mathbf{y}_i}{\pi_i^*}, \tag{2.3}$$

où la pseudo probabilité d'inclusion est $\pi_i^* = \pi_i \pi_{i|A} = k\pi_i / n$. La formule (2.2) peut alors être utilisée pour calculer la version agrégée par bootstrap de l'estimateur π^* classique. Un calcul simple montre que l'estimateur agrégé par bootstrap est identique à l'estimateur de HT original si tous les échantillons EASSR de taille k sont dénombrés dans le calcul de (2.2). Il en va de même pour tous les autres estimateurs linéaires. En général, le calcul de l'estimateur agrégé par bootstrap $\hat{\theta}_{bag}$ n'est pas aussi facile. Dans le reste de cette section, nous nous attardons aux calculs de ce genre pour les trois types d'estimateurs non linéaires abordés dans la section 1.

2.2 Agrégation bootstrap des estimateurs différenciables

Pour les estimateurs d'enquête qui sont des fonctions différenciables des estimateurs de HT, la quantité d'intérêt de la population peut aussi s'exprimer sous forme de fonction différenciable des moyennes de population, c'est-à-dire $\theta_d = m\left(\boldsymbol{\mu}\right)$, où $m(\cdot)$ est une fonction différenciable connue. L'indice « d » veut dire **différenciable** par opposition à **non différenciable** $\left(\theta_{nd}\right)$ et à **équation d'estimation** $\left(\theta_{ee}\right)$ à venir plus tard. Un estimateur direct de type « plug-in » de θ_d, fondé sur les observations de l'échantillon $\mathcal{Y}$, peut s'écrire

$$\hat{\theta}_d = m\left(\hat{\boldsymbol{\mu}}\right), \tag{2.4}$$

où $\hat{\boldsymbol{\mu}}$ est défini en (2.1). Ainsi, la version d'échantillon répété de $\hat{\theta}_d$ peut s'écrire

$$\hat{\theta}_d\left(\mathcal{Y}_b^*\right) = m\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)\right),$$

où $\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)$ est défini en (2.3). L'estimateur agrégé par bootstrap de θ_d, dénoté par $\hat{\theta}_{d,bag}$, est alors défini au moyen de la formule en (2.2).

2.3 Agrégation bootstrap des estimateurs non différenciables explicitement définis

Comme exemple de ce type d'estimateurs, prenons la proportion de ménages dont le revenu est inférieur au seuil de pauvreté pour une population donnée. Cette proportion peut s'écrire $\left(1/N\right)\sum_{i=1}^{N} I\left(y_i \le \lambda_N\right)$, où y_i est la valeur du revenu pour le i-ième ménage de la population, et λ_N est le seuil de pauvreté de la population. On peut voir que cette quantité d'intérêt est la moyenne des fonctions noyau de l'indicateur, et que la fonction noyau n'est pas différenciable en λ_N. Ici, nous considérons une classe plus générale où le noyau est une fonction arbitraire non différenciable, mais bornée. Ce type de quantité de population peut s'écrire

$$\theta_{nd} = \frac{1}{N}\sum_{i=1}^{N} h\left(\mathbf{y}_i - \boldsymbol{\lambda}_N\right),$$

où $\boldsymbol{\lambda}_N$ est un paramètre de population inconnu, par exemple la moyenne, un quantile ou une autre quantité de population, et $h\left(\mathbf{y}-\boldsymbol{\lambda}\right): \mathbb{R}^p \to \mathbb{R}$ est une fonction non différenciable de $\boldsymbol{\lambda}$. La quantité de population θ_{nd} généralise la notion de proportion inférieure à un niveau estimé et ressemble à la forme générale d'une statistique U.

Wang et Opsomer (2011) ont étudié une classe d'estimateurs ressemblant à des statistiques U, à savoir les estimateurs d'enquête non différenciables,

$$\hat{\theta}_{nd} = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right), \tag{2.5}$$

où $\hat{\boldsymbol{\lambda}}$ est un estimateur fondé sur le plan de $\boldsymbol{\lambda}_N$. Hors contexte d'une enquête, les estimateurs de ce type sont considérés comme des [*Traduction*] « fonctions non différenciables de la répartition empirique » (Bickel, Götze et van Zwet 1997). Les procédures bootstrap appropriées pour ces estimateurs ont notamment été étudiées par Beran et Srivastava (1985) et par Dümbgen (1993). Nous définissons la version répétée de $\hat{\theta}_{nd}$ fondée sur le nouvel échantillon A_b comme suit

$$\hat{\theta}_{nd}\left(\mathcal{Y}_b^*\right) = \frac{1}{N}\sum_{i\in A_b}\frac{1}{\pi_i^*} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right),$$

où $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ dépend uniquement du nouvel échantillon bootstrap A_b, et l'estimateur agrégé par bootstrap est alors défini comme étant la moyenne des estimateurs répétés. Supposons que le processus de

rééchantillonnage est l'EASSR de taille k, et que chaque sous-échantillon est choisi pour calculer l'estimateur d'agrégation bootstrap; l'estimateur bootstrap prend alors la forme suivante après manipulation :

$$\hat{\theta}_{nd,bag} = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i\binom{n-1}{k-1}}\sum_{A_b \ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right), \tag{2.6}$$

qui remplace $h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right)$ en (2.5) par une quantité « lisse » $\sum_{A_b \ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right) \Big/ \binom{n-1}{k-1}$, en calculant la moyenne des « sauts » de l'estimateur. Dans bien des cas, on peut réduire la variance en effectuant ce remplacement. Le terme $\sum_{A_b \ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right) \Big/ \binom{n-1}{k-1}$ est l'**espérance bootstrap** de $h\left(\mathbf{y}_i - \cdot\right)$ et peut être approximé en utilisant la convolution de $h\left(\mathbf{y}_i - \cdot\right)$ avec la distribution d'échantillonnage de $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$. Les aspects théoriques de $\hat{\theta}_{nd,bag}$ sont examinés dans la section 3.

2.4 Agrégation bootstrap des estimateurs définis par des équations d'estimation non différenciables

Enfin, nous expliquons comment faire l'agrégation bootstrap des estimateurs définis par des équations d'estimation non différenciables. Pour faciliter la présentation, nous considérons un paramètre d'intérêt unidimensionnel. Le paramètre de population θ_{ee} d'intérêt est défini comme suit

$$\theta_{ee} = \inf\left\{\gamma : S(\gamma) \geq 0\right\},$$

où

$$S(\gamma) = \frac{1}{N}\sum_{i=1}^{N}\psi\left(y_i - \gamma\right),$$

et $\psi(\cdot)$ est une fonction non différenciable réelle. Nous pouvons estimer le paramètre de population θ_{ee} au moyen de $\hat{\theta}_{ee}$, où

$$\hat{\theta}_{ee} = \inf\left\{\gamma : \hat{S}(\gamma) \geq 0\right\}$$

avec

$$\hat{S}(\gamma) = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\psi\left(y_i - \gamma\right).$$

Un estimateur de ce type qu'on rencontre souvent est le quantile d'échantillon défini par inversion de la fonction de distribution cumulative de l'échantillon (Francisco et Fuller 1991), où $\psi\left(y_i - \gamma\right) = I_{\left(y_i \leq \gamma\right)} - \alpha$ pour le quantile α.

Sur le plan conceptuel, il existe deux versions de l'agrégation bootstrap $\hat{\theta}_{ee}$: l'une résout l'« équation d'estimation agrégée par bootstrap » définie par agrégation bootstrap de la fonction de score, tandis que

l'autre calcule la moyenne sur les estimations rééchantillonnées de $\hat{\theta}_{ee}$. Comme il est expliqué dans la section 2.1, la première version produit un estimateur équivalant à l'estimateur original, parce que l'espérance bootstrap des échantillons bootstrap de $\hat{S}(\gamma)$ est égale à $\hat{S}(\gamma)$ pour un γ fixe. Nous considérons donc seulement la deuxième version. Pour définir l'estimateur d'équation agrégé par bootstrap, nous commençons par définir la fonction de score répétée $\hat{S}_b(\gamma)$ fondée sur le nouvel échantillon A_b comme suit

$$\hat{S}_b(\gamma) = \frac{1}{N}\sum_{i\in A_b}\frac{1}{\pi_i^*}\psi(y_i - \gamma).$$

L'estimateur répété fondé sur A_b est alors défini comme étant $\hat{\theta}_{ee}(\mathcal{Y}_b^*) = \inf\{\gamma : \hat{S}_b(\gamma) \geq 0\}$ et l'estimateur agrégé par bootstrap devient

$$\hat{\theta}_{ee,bag} = \frac{1}{\binom{n}{k}}\sum\hat{\theta}_{ee}(\mathcal{Y}_b^*), \tag{2.7}$$

où la moyenne est calculée sur tous les échantillons sans remise possibles de taille k choisis à partir de A. Chen et Hall (2003) ont examiné les estimateurs agrégés par bootstrap définis par des équations d'estimation non linéaires sous les conditions *iid* et ils ont conclu que l'agrégation bootstrap n'améliore pas toujours la précision des estimateurs étudiés.

3 Résultats théoriques

Nous commençons par décrire brièvement l'analyse asymptotique des estimateurs agrégés par bootstrap sous échantillonnage général d'une population finie, c.-à-d. dans un contexte fondé sur le plan de sondage. Nous procédons ainsi dans le cadre habituel d'une population croissante, où nous considérons une séquence croissante de populations emboîtées, disons U_N, $N = 1, 2, \ldots$, avec une moyenne de population finie $\boldsymbol{\mu}_N$. Associée à la séquence de populations est une séquence de plans d'échantillonnage utilisés pour tirer un échantillon aléatoire $A_N \subseteq U_N$ de taille n_N, avec probabilités d'inclusion connexes π_{iN}. Comme cela se fait couramment dans la littérature sur les enquêtes, nous supprimons l'indice N dans l'échantillon A, la taille de l'échantillon n et les probabilités d'inclusion π_i. Par souci de concision, nous fournissons seulement les résultats asymptotiques fondés sur le plan pour l'agrégation bootstrap de l'estimateur différenciable $\hat{\theta}_d$ et non différenciable $\hat{\theta}_{nd}$. Les hypothèses formelles qui sous-tendent les résultats et les théorèmes associés aux estimateurs différenciables et non différenciables figurent à l'annexe A.1. La principale conclusion que nous pouvons tirer dans ce contexte fondé sur le plan est que, si nous partons d'un estimateur conforme au plan et que nous laissons le nombre d'échantillons bootstrap k augmenter avec n, les versions agrégées par bootstrap des estimateurs seront elles aussi conformes au plan. Il s'agit clairement d'une propriété clé de ces estimateurs, puisqu'il n'y aurait aucune raison d'en tenir compte s'ils n'étaient pas conformes au plan.

Malheureusement, les résultats fondés sur le plan ci-dessus sont très limités et, surtout, ils ne fournissent pas une distribution asymptotique qui permettrait de faire de l'inférence, autre propriété

hautement souhaitable des estimateurs d'enquête. Nous considérons donc également un contexte fondé sur un modèle, dans lequel nous pouvons obtenir une approximation asymptotique de la variance. En présentant des résultats fondés sur le modèle, nous supposons que le plan d'échantillonnage choisissant l'échantillon original A est un plan d'échantillonnage avec probabilités égales, et les caractéristiques de la population peuvent être considérées comme un échantillon iid d'une répartition de superpopulation. Dans ce contexte, l'estimateur agrégé par bootstrap peut être traité comme une statistique U. Nous pouvons alors appliquer la théorie des statistiques U pour obtenir une expansion asymptotique des estimateurs agrégés par bootstrap. L'analyse est comparable à celles de Bühlmann et Yu (2002) et de Buja et Stuetzle (2006). Aux fins du présent article, nous nous limitons aux échantillons bootstrap de taille k où k est bornée et fixe. Selon cette hypothèse, les estimateurs agrégés par bootstrap peuvent être considérés comme des statistiques U de degré fixe pour lesquelles une théorie asymptotique a été bien élaborée. Un cas plus intéressant survient lorsque la taille k du rééchantillonnage croît avec la taille de l'échantillon n, et que cela aboutit à des statistiques U de degré infini. Ces statistiques ont des applications dans l'étude de l'estimateur de Kaplan-Meier et des estimateurs bootstrap m-sur-n, et les lecteurs sont invités à consulter Frees (1989), Heilig (1997), Heilig et Nolan (2001) et leurs références sur les propriétés statistiques de ces estimateurs. Schick et Wefelmeyer (2004) ont étudié les propriétés des statistiques U de degré infini produites à partir des moyennes mobiles des innovations dans des séries chronologiques. L'étude des estimateurs agrégés par bootstrap considérés comme des statistiques U de degré infini dépasse la portée du présent article; nous nous limitons donc aux échantillons bootstrap de tailles fixes et bornées dans le cas fondé sur un modèle.

Considérons d'abord l'estimateur agrégé par bootstrap en (2.5). Sous l'EASSR, l'estimateur (2.5) peut être simplifié comme suit

$$\hat{\theta}_{nd} = \frac{1}{n}\sum_{i\in A} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right)$$

et la version agrégée par bootstrap de $\hat{\theta}_{nd}$ est définie comme étant

$$\hat{\theta}_{nd,bag} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\binom{n-1}{k-1}} \sum_{A_b \ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right) \tag{3.1}$$

où $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ dépend seulement du nouvel échantillon A_b. Pour faciliter la présentation, nous prenons $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ comme moyenne de l'échantillon. Dans le cas qui nous intéresse, un simple calcul algébrique révèle que

$$\hat{\theta}_{nd,bag} = \frac{1}{\binom{n}{k}} \sum_{A_b \in \mathcal{A}} \left\{ \frac{1}{k}\sum_{i\in A_b} h\left(\frac{k-1}{k}\mathbf{y}_i - \frac{1}{k}\sum_{j\neq i}\mathbf{y}_j \right) \right\},$$

où $\mathcal{A}$ représente les sous-ensembles de taille k de l'ensemble $\{1,2,\ldots,n\}$. L'estimateur $\hat{\theta}_{nd,bag}$ est une statistique U de degré k avec noyau

$$g\left(y_1,\ldots,y_k\right) = \frac{1}{k}\sum_{i=1}^{k} h\left(\frac{k-1}{k}\mathbf{y}_i - \frac{1}{k}\sum_{\substack{j=1 \\ j\neq i}}^{k}\mathbf{y}_j \right)$$

à condition que k reste finie.

On peut voir que l'estimateur agrégé par bootstrap $\hat{\theta}_{nd,bag}$ est une statistique symétrique de $\mathbf{y}_i$, et la théorie standard des statistiques symétriques (Lee 1990) s'applique. Les résultats sont énoncés dans le théorème 1, et les hypothèses et preuves figurent à l'annexe A.2.

Théorème 1 *Sous les hypothèses M.1 à M.4 concernant la répartition de superpopulation et les plans d'échantillonnage et de rééchantillonnage*

$$\mathrm{AV}\left(\hat{\theta}_{nd,bag}\right)^{-1/2}\left(\hat{\theta}_{nd,bag}-\theta_{nd,\infty}\right)\xrightarrow{p}N(0,1), \tag{3.2}$$

où la valeur limite $\theta_{nd,\infty}=\lim_{n\to\infty}\mathrm{E}\left[h\left(\mathbf{y}_i-\hat{\boldsymbol{\lambda}}\right)\right]$, *la variance asymptotique*

$$\mathrm{AV}\left(\hat{\theta}_{nd,bag}\right)=\frac{1}{n}\mathrm{Var}\left[u(\mathbf{y}_i)\right]+\frac{(k-1)^2}{n}\mathrm{Var}\left[v(\mathbf{y}_i)\right]+\frac{2(k-1)}{n}\mathrm{Cov}\left[u(\mathbf{y}_i),v(\mathbf{y}_i)\right], \tag{3.3}$$

et

$$u(\mathbf{y})=\mathrm{E}\left[h\left(\mathbf{y}-\hat{\boldsymbol{\lambda}}(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y})\right)\right],$$
$$v(\mathbf{y})=\mathrm{E}\left[h\left(\mathbf{y}_1-\hat{\boldsymbol{\lambda}}(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y})\right)\right].$$

Comme il est indiqué en (3.3), la variance asymptotique de l'estimateur agrégé par bootstrap dépend des fonctions inconnues $u(\mathbf{y})$ et $v(\mathbf{y})$, qui sont des espérances de $h(\cdot)$ en ce qui concerne la répartition de superpopulation. En $u(\mathbf{y})$ et $v(\mathbf{y})$, $\hat{\boldsymbol{\lambda}}(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y})$ est calculé à partir de $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1}$ avec un vecteur arbitraire $\mathbf{y}$. L'espérance porte sur la distribution de vecteurs aléatoires *iid* $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1}$. Cette espérance de haute dimension est difficile à calculer et pourrait ne pas avoir une expression explicite en général. On ne peut pas obtenir la forme exacte de $u(\cdot)$ et $v(\cdot)$ mais on peut l'approximer en utilisant une approche fondée sur le rééchantillonnage. Les fonctions inconnues $u(\cdot)$ et $v(\cdot)$ sont définies comme étant des espérances de quantités respectives liées à la répartition de superpopulation, qui peuvent être approximées par l'espérance concernant la distribution empirique.

La variance asymptotique fondée sur un modèle peut être estimée dans le cadre du processus d'agrégation bootstrap. Nous pouvons calculer les intégrandes $h\left(\mathbf{y}-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*,\mathbf{y}\right)\right)$ et $h\left(\mathbf{y}_1-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*,\mathbf{y}\right)\right)$ en fonction de chaque échantillon bootstrap, $\mathbf{y}$ étant où nous voulons évaluer $u(\cdot)$ et $v(\cdot)$, et $\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*$ étant les valeurs rééchantillonnées. Nous pouvons alors calculer la moyenne de chaque quantité pour approximer l'espérance. Enfin, nous pouvons estimer la variance en calculant la variance d'échantillon des espérances évaluées à chacun des points d'échantillonnage. Pour les estimateurs non lisses comme ceux qui nous intéressent, il est souvent recommandé d'utiliser la méthode bootstrap lisse pour approximer la variance (Efron 1979; Davison et Hinkley 1997). Nous appliquons la méthode bootstrap lisse et ajoutons une petite quantité de bruit à chaque valeur rééchantillonnée afin de lisser la fonction sous-jacente. L'algorithme détaillé est expliqué au moyen d'un exemple dans la section 5.

Nous passons maintenant au résultat fondé sur un modèle des estimateurs agrégés par bootstrap définis par les équations d'estimation en (2.7). Un cas spécial dans ce contexte est l'agrégation bootstrap des

quantiles d'échantillon, qui a été étudiée par Knight et Bassett (2002). Knight et Bassett (2002) ont examiné le rééchantillonnage par bootstrap et par EASSR et étudié les effets de l'agrégation bootstrap sur le terme restant de la représentation des quantiles de Bahadur (Bahadur 1966). Nous abordons la question sous un angle légèrement différent et traitons l'estimateur agrégé par bootstrap comme une statistique U. Les hypothèses et les preuves figurent à l'annexe A.2. Il est à noter que, sous l'hypothèse M.5, la fonction d'estimation non différenciable doit avoir une limite lisse. Dans le théorème suivant, nous linéarisons l'estimateur de l'équation d'estimation agrégé par bootstrap et donnons une expression pour la variance asymptotique.

Théorème 2 *Sous les hypothèses M.1 à M.3 et M.5, le résultat asymptotique suivant tient pour l'estimateur de l'équation d'estimation agrégé par bootstrap (2.7),*

$$\mathrm{AV}\left(\hat{\theta}_{ee,bag}\right)^{-1/2}\left(\hat{\theta}_{ee,bag}-\theta_{ee,\infty}\right)\xrightarrow{p}N(0,1), \tag{3.4}$$

où $\theta_{ee,\infty}$ *est la limite asymptotique de la quantité de population* θ_{ee}, *la variance asymptotique de* $\hat{\theta}_{ee,bag}$ *est*

$$\mathrm{AV}\left(\hat{\theta}_{ee,bag}\right)=\frac{k^2}{n}\mathrm{Var}\left[u(y_i)\right], \tag{3.5}$$

et

$$u(y)=\mathrm{E}\ \inf\left\{\gamma:\frac{1}{k}\sum_{i=1}^{k-1}\psi(y_i-\gamma)+\frac{1}{k}\psi(y-\gamma)\geq 0\right\}. \tag{3.6}$$

Comme nous l'avons vu pour l'estimateur agrégé par bootstrap (3.1), les résultats asymptotiques du théorème 2 font intervenir une fonction inconnue. Là encore, cette fonction peut être calculée par rééchantillonnage des échantillons répétés disponibles.

4 Estimation de la variance

Bien que l'approche fondée sur un modèle permette d'obtenir des distributions asymptotiques et donc une inférence asymptotiquement correcte, nous nous intéressons surtout aux applications fondées sur le plan de l'agrégation bootstrap. En contexte fondé sur le plan de sondage, nous pouvons combiner naturellement la construction de l'estimateur agrégé par bootstrap avec l'estimation de la variance de la statistique originale, en tirant parti des échantillons répétés diffusés par les organismes statistiques. Dans le présent article, nous prenons l'exemple précis d'un échantillonnage aléatoire simple stratifié avec plan d'échantillonnage bootstrap d'un échantillon EASSR stratifié.

Nous commençons par appliquer une version de la procédure bootstrap de Rao et Wu (1988) afin d'estimer la variance des estimateurs d'enquête avant l'agrégation bootstrap. Soient N_h, n_h et k_h, la taille de la population, la taille de l'échantillon et la taille du sous-échantillon dans la strate h, $h=1,2,\ldots,H$. Ici, B échantillons bootstrap sont tirés par échantillonnage aléatoire simple stratifié sans remise de taille

k_h pour calculer la variance bootstrap de la statistique originale et l'estimateur agrégé par bootstrap. Pour chaque échantillon bootstrap, nous attribuons un poids de

$$\frac{N_h}{N}\left(1-k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\right)\frac{1}{n_h}+\frac{N_h}{N}k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\frac{1}{k_h}$$

à chaque élément échantillonné de la strate h, et

$$\frac{N_h}{N}\left(1-k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\right)\frac{1}{n_h}$$

aux éléments non échantillonnés. Nous utilisons ensuite la variance ordinaire des estimateurs d'échantillons répétés comme estimateur de variance. Le schéma de pondération susmentionné est algébriquement identique à l'équation 4.1 de Rao et Wu (1988), où la correction pour population finie est intégrée aux poids de rééchantillonnage. L'estimateur de variance par rééchantillonnage dérivé de la méthode de pondération se réduit à un estimateur de variance ordinaire sous l'EASSR stratifié et garantit l'absence de biais sous le plan. Afin de combiner l'agrégation bootstrap et l'estimateur de variance bootstrap, nous utilisons les mêmes échantillons bootstrap afin de construire les estimateurs agrégés par bootstrap pour les quantités de population qui nous intéressent.

Sous le cadre fondé sur le plan de sondage, aucun estimateur de variance analytique n'est disponible pour l'estimateur agrégé par bootstrap en général. Pour le moment, nous suggérons d'appliquer les deux approches suivantes d'estimation de la variance :

(Var. 1) Utiliser la variance estimée de l'estimateur original même si l'estimateur agrégé par bootstrap a une variance plus faible. Cette méthode produit des intervalles de confiance de même largeur, mais son taux de couverture est supérieur à celui de l'intervalle de confiance original.

(Var. 2) Multiplier la variance estimée de l'estimateur original par un facteur de correction tenant compte de l'amélioration probable de l'efficacité. Ce facteur pourrait être le gain d'efficacité si l'on présume que l'échantillon est un échantillon *iid* d'une superpopulation infinie. On peut déterminer le facteur en utilisant les résultats des théorèmes 1 et 2, ou par expérience bootstrap non paramétrique. Une procédure bootstrap possible est le bootstrap double, qu'on met en œuvre par rééchantillonnage bootstrap ordinaire afin d'estimer la variance de l'estimateur original, et un autre niveau de rééchantillonnage EASSR afin de déterminer la variance de l'estimateur agrégé par bootstrap. On peut estimer le ratio de la variance entre l'estimateur agrégé par bootstrap et l'estimateur original en utilisant ces échantillons bootstrap emboîtés, et multiplier la variance sous le plan de l'estimateur original par ce ratio.

Nous examinons les deux approches dans les simulations de la section 5, mais il s'agit clairement d'un domaine qui devrait faire l'objet de recherches plus approfondies.

5 Simulations

Pour évaluer le comportement pratique de l'agrégation bootstrap en contexte d'enquête, nous générons une population finie de taille $N = 2\ 000$ à trois strates. La taille de chaque strate est N_h où $h = 1,2,3,$ et les proportions des strates sont fixées à $\left(N_1; N_2; N_3\right)/N = \left(0,5;\ 0,3;\ 0,2\right).$ La distribution de la variable cible y_i dans chaque strate est $y_{1i} \sim \left|N\left(-1,1\right)\right|$, $y_{2i} \sim \Gamma\left(1,1\right)$ et $y_{3i} \sim \left|N\left(3,2\right)\right|.$ Une variable auxiliaire x_i est générée par $x_i = A_0 + A_1 y_i + A_2\left(G_i - \alpha/\beta\right)$ où $A_0 = A_1 = 2,\ A_2 = 1,\ \alpha = 2,\ \beta = 1$ et $G_i \overset{iid}{\sim} \Gamma\left(2,1\right).$ Nous tirons de façon répétée des échantillons de taille n par échantillonnage aléatoire simple stratifié de la population d'intérêt et la répartition de la taille de l'échantillon est $\left(n_1; n_2; n_3\right)/n = \left(0,3; 0,3; 0,4\right).$ Dans ce contexte, le plan de sondage est clairement informatif, parce que les observations ne sont pas iid dans la population globale et sont corrélées avec les probabilités d'inclusion.

Nous nous intéressons à trois quantités de population : un quantile α de population, une proportion de la population inférieure à une fraction donnée d'un quantile de population (voir Berger et Skinner 2003, par exemple) et l'estimateur de Rao-Kovar-Mantel (RKM) de la fonction de répartition (Rao et coll. 1990). Le premier est un exemple d'un estimateur fondé sur une équation d'estimation non différenciable, tandis que les deux derniers sont des estimateurs non différenciables explicitement définis. L'estimateur sur échantillon du quantile est obtenu par inversion de la fonction de répartition cumulative estimée. L'estimateur sur échantillon de la proportion inférieure à une fraction donnée d'un quantile de population est l'estimateur HT de la proportion des observations inférieures à la médiane d'échantillon d'une variable d'intérêt multipliée par une constante c,

$$\hat{\theta}_{pr} = \frac{1}{N}\sum_{i \in A}\frac{1}{\pi_i}\mathrm{I}_{\left(y_i \le c\hat{\theta}_{\mathrm{med}}\right)},$$

où $\hat{\theta}_{\mathrm{med}}$ est la médiane d'échantillon de y_i. L'estimateur par la différence de RKM fondé sur le plan de sondage et sur un modèle de ratio est

$$\hat{\theta}_{\mathrm{RKM}} = \frac{1}{N}\left\{\sum_{i \in A}\frac{1}{\pi_i}\mathrm{I}_{\left(y_i \le t\right)} + \sum_{i=1}^{N}\mathrm{I}_{\left(\hat{R}x_i \le t\right)} - \sum_{i \in A}\frac{1}{\pi_i}\mathrm{I}_{\left(\hat{R}x_i \le t\right)}\right\}, \tag{5.1}$$

où $\hat{R}$ est le ratio estimé entre y et x.

La variance sous le plan de ces estimateurs non différenciables est un peu fastidieuse à estimer. Pour le calcul des variances et des intervalles des quantiles d'échantillon, les lecteurs sont invités à consulter Francisco et Fuller (1991), Sitter et Wu (2001) et leurs références. Pour les proportions inférieures à un niveau estimé, voir Shao et Rao (1993) et Berger et Skinner (2003).

Les variances sous le plan des estimateurs originaux $\hat{\theta}_{qr}$, $\hat{\theta}_{pr}$ et $\hat{\theta}_{\mathrm{RKM}}$, sont estimées au moyen de la procédure bootstrap sans remise décrite dans la section qui précède. Nous employons une taille d'échantillon bootstrap de $k_h = n_h/2.$ Les estimateurs agrégés par bootstrap ainsi construits sont souvent

qualifiés de « subagging estimators » (estimateurs sous-agrégés par bootstrap) (Bühlmann et Yu 2002). Il a été établi que les échantillons sans remise de taille $n/2$ produisent des résultats semblables à ceux des échantillons avec remise de taille n en agrégation bootstrap (Buja et Stuetzle 2006; Friedman et Hall 2007). Nous appliquons les deux approches de la variance pour les estimateurs agrégés par bootstrap proposées dans la section qui précède, c.-à-d. une approche identique à celle de l'estimateur non agrégé par bootstrap (Var. 1), et une autre qui multiplie l'estimation de la variance originale par un facteur de correction fondé sur un modèle (Var. 2). Le facteur est déterminé par bootstrap double sur un échantillon particulier. En principe, il faudrait répéter l'exercice pour chaque échantillon, mais ce scénario est exclu par le lourd fardeau de calcul. Les intervalles de confiance des trois estimateurs sont construits par approximation normale. Les intervalles de confiance pour la proportion et l'estimateur de RKM sont construits par approximation normale sur une échelle transformée *logit*, $\log\left[\hat{\theta}\big/\left(1-\hat{\theta}\right)\right]$ ou $\log\left[\hat{\theta}_{bag}\big/\left(1-\hat{\theta}_{bag}\right)\right]$, puis par rétrotransformation (Agresti 2002; Korn et Graubard 1998).

Le tableau 5.1 résume le biais, l'écart-type et le ratio de l'EQM des quantiles d'échantillon originaux et agrégés par bootstrap, tandis que le tableau 5.2 examine les estimateurs de la variance et les intervalles de confiance. Les tailles d'échantillon choisies sont $n = 100$ et 200. Dans le tableau 5.1, nous pouvons voir que l'estimateur de quantile agrégé par bootstrap est plus efficace que l'estimateur original puisque le ratio de l'EQM est inférieur à un dans cette expérience par simulation. En général, plus la taille de l'échantillon diminue, plus les effets de lissage de l'agrégation bootstrap deviennent évidents. Dans le tableau 5.2, nous comparons les deux intervalles de confiance avec estimateur de point d'agrégation bootstrap aux intervalles de confiance originaux. Comme prévu, l'intervalle de confiance construit selon la méthode 1 a la même longueur et une couverture plus étendue que l'original. Dans cet exemple, les intervalles de confiance construits selon la méthode 2 sont plus étroits, mais leur niveau de couverture reste proche du niveau nominal.

Tableau 5.1
Biais, écart-type et ratios de l'EQM des quantiles d'échantillon et des quantiles d'échantillon agrégé par bootstrap; taille de la population $N = 2\,000$, **nombre de bootstraps** $B = 2\,000$ **et résultats de** $2\,000$ **simulations**

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
α	**0,2**	**0,3**	**0,5**	**0,7**	**0,8**	**0,2**	**0,3**	**0,5**	**0,7**	**0,8**
biais$\left(\hat{\theta}_{qt}\right)$	0,002	0,008	0,000	-0,005	-0,035	-0,008	0,005	0,006	0,007	-0,005
biais$\left(\hat{\theta}_{qt,bag}\right)$	0,018	0,019	-0,001	-0,007	-0,043	-0,006	0,009	0,005	0,006	-0,022
é-t$\left(\hat{\theta}_{qt}\right)$	0,093	0,124	0,149	0,181	0,212	0,070	0,076	0,103	0,136	0,148
é-t$\left(\hat{\theta}_{qt,bag}\right)$	0,089	0,112	0,138	0,167	0,197	0,065	0,073	0,099	0,127	0,139
$\dfrac{EQM_p\left(\hat{\theta}_{qt,bag}\right)}{EQM_p\left(\hat{\theta}_{qt}\right)}$	0,946	0,844	0,859	0,854	0,875	0,866	0,924	0,919	0,862	0,912

Tableau 5.2
Biais relatif, probabilité de couverture et largeur de l'intervalle de confiance des estimateurs de variance bootstrap pour les quantiles d'échantillon et des estimateurs de variance non corrigé $\left(\hat{V}_1\right)$ et corrigé $\left(\hat{V}_2\right)$ pour les quantiles d'échantillon agrégé par bootstrap; mêmes conditions de simulation qu'au tableau 5.1

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
α	**0,2**	**0,3**	**0,5**	**0,7**	**0,8**	**0,2**	**0,3**	**0,5**	**0,7**	**0,8**
$\frac{\mathrm{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{qt}\right)\right]}{V\left(\hat{\theta}_{qt}\right)}$	1,208	1,091	1,099	1,135	1,205	1,067	1,117	1,093	1,098	1,180
$\frac{\mathrm{E}\left[\hat{V}_1\left(\hat{\theta}_{qt,bag}\right)\right]}{V\left(\hat{\theta}_{qt,bag}\right)}$	1,327	1,325	1,279	1,331	1,402	1,224	1,217	1,188	1,273	1,326
$\frac{\mathrm{E}\left[\hat{V}_2\left(\hat{\theta}_{qt,bag}\right)\right]}{V\left(\hat{\theta}_{qt,bag}\right)}$	1,307	1,217	1,196	1,184	1,383	1,245	1,249	1,392	1,107	1,104
P.C.(I.C.)	0,944	0,934	0,924	0,928	0,922	0,938	0,951	0,942	0,935	0,950
P.C.$\left(\text{I.C.1}_{\cdot bag}\right)$	0,950	0,946	0,938	0,938	0,939	0,942	0,950	0,946	0,943	0,954
P.C.$\left(\text{I.C.2}_{\cdot bag}\right)$	0,949	0,934	0,932	0,929	0,938	0,944	0,952	0,958	0,927	0,936
Largeur(I.C.)										
Largeur$\left(\text{I.C.1}_{\cdot bag}\right)$	0,386	0,492	0,597	0,729	0,880	0,277	0,309	0,414	0,544	0,612
Largeur$\left(\text{I.C.2}_{\cdot bag}\right)$	0,383	0,472	0,577	0,688	0,874	0,279	0,313	0,448	0,508	0,559

Les tableaux 5.3 et 5.4 résument les résultats fondés sur le plan pour l'estimateur de la proportion de personnes à faible revenu. Le ratio de l'EQM montre que l'estimateur agrégé par bootstrap est uniformément plus efficace que l'estimateur original et que l'EQM de cet estimateur agrégé est inférieur à 50% de l'EQM de l'estimateur original dans certains cas (voir $c = 1,2$). Cela est probablement dû au fait que l'estimateur comporte deux « niveaux » de non-différenciabilité : la médiane d'échantillon est un estimateur non différenciable dont le gain d'efficacité est illustré au tableau 5.1, et la proportion de personnes à faible revenu est une fonction non différenciable de la médiane d'échantillon. Les « sauts » des estimateurs sont lissés par agrégation bootstrap, ce qui donne un estimateur plus stable. La comparaison des intervalles de confiance au tableau 5.4 donne des résultats semblables à ceux obtenus pour les quantiles.

Tableau 5.3
Biais, écart-type et ratio de l'EQM de la proportion estimée inférieure à une constante c multipliée par une médiane estimée et l'estimateur de proportion agrégé par bootstrap; taille de la population $N = 2\ 000$, nombre de bootstraps $B = 2\ 000$ et résultats de $2\ 000$ simulations

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
c	**0,2**	**0,4**	**0,6**	**1,2**	**1,5**	**0,2**	**0,4**	**0,6**	**1,2**	**1,5**
biais$\left(\hat{\theta}_{pr}\right)$	-0,002	-0,002	-0,003	0,011	0,006	0,000	-0,002	-0,005	-0,004	-0,004
biais$\left(\hat{\theta}_{pr,bag}\right)$	-0,004	-0,004	-0,007	0,017	0,009	-0,001	-0,005	-0,009	-0,001	-0,004
é-t$\left(\hat{\theta}_{pr}\right)$	0,034	0,039	0,038	0,034	0,046	0,023	0,027	0,026	0,026	0,036
é-t$\left(\hat{\theta}_{pr,bag}\right)$	0,031	0,035	0,031	0,020	0,034	0,022	0,025	0,022	0,017	0,029
$\frac{EQM_p\left(\hat{\theta}_{pr,bag}\right)}{EQM_p\left(\hat{\theta}_{pr}\right)}$	0,861	0,821	0,709	0,538	0,581	0,883	0,860	0,783	0,434	0,671

Tableau 5.4
Biais relatif, probabilité de couverture et largeur de l'intervalle de confiance des estimateurs de variance bootstrap pour les proportions d'échantillon et des estimateurs de variance non corrigé $\left(\hat{V}_1\right)$ et corrigé $\left(\hat{V}_2\right)$ pour les proportions d'échantillon agrégé par bootstrap; mêmes conditions de simulation qu'au tableau 5.3. Le sigle « I.C.T. » est utilisé pour désigner les intervalles de confiance obtenus par transformation logit.

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
c	**0,2**	**0,4**	**0,6**	**1,2**	**1,5**	**0,2**	**0,4**	**0,6**	**1,2**	**1,5**
$\frac{\mathrm{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{pr}\right)\right]}{V\left(\hat{\theta}_{pr}\right)}$	1,122	1,191	1,325	1,472	1,281	1,140	1,191	1,251	1,350	1,217
$\frac{\mathrm{E}\left[\hat{V}_1\left(\hat{\theta}_{pr,bag}\right)\right]}{V\left(\hat{\theta}_{pr,bag}\right)}$	1,323	1,471	1,959	4,095	2,307	1,293	1,428	1,766	3,064	1,821
$\frac{\mathrm{E}\left[\hat{V}_2\left(\hat{\theta}_{pr,bag}\right)\right]}{V\left(\hat{\theta}_{pr,bag}\right)}$	1,240	0,963	1,190	1,174	1,149	1,145	1,262	1,319	2,039	1,524
P.C.(I.C.T.)	0,969	0,970	0,984	0,991	0,980	0,964	0,974	0,977	0,983	0,946
$\mathrm{P.C.}\left(\mathrm{I.C.T.1.}_{bag}\right)$	0,979	0,983	0,995	0,998	0,995	0,974	0,980	0,988	0,998	0,976
$\mathrm{P.C.}\left(\mathrm{I.C.T.2.}_{bag}\right)$	0,976	0,944	0,973	0,922	0,942	0,962	0,969	0,968	0,993	0,957
Largeur(I.C.T)										
$\mathrm{Largeur}\left(\mathrm{I.C.T.1.}_{bag}\right)$	0,144	0,166	0,168	0,157	0,197	0,098	0,115	0,114	0,113	0,149
$\mathrm{Largeur}\left(\mathrm{I.C.T.2.}_{bag}\right)$	0,139	0,134	0,131	0,085	0,140	0,093	0,108	0,099	0,092	0,136

Les tableaux 5.5 et 5.6 résument les résultats fondés sur le plan de sondage pour l'estimateur de RKM. Là encore, nous observons le gain d'efficacité en appliquant la méthode de l'agrégation bootstrap, et le gain se situe entre 2 % et 12 %. Les deux estimateurs de variance de la quantité agrégée par bootstrap donnent d'assez bons résultats. Les deux versions des intervalles de confiance pour les estimateurs agrégés par bootstrap ont des taux de couverture réels proches de 95 %, et les intervalles de confiance calculés selon l'approche du facteur de correction (Var. 2) sont légèrement plus courts que ceux calculés selon la méthode 1.

Tableau 5.5
Biais, écart-type et ratios de l'EQM de l'estimateur de RKM et de l'estimateur de RKM agrégé par bootstrap (5.1); taille de la population $N = 2\ 000$, nombre de bootstraps $B = 2\ 000$ et résultats de $2\ 000$ simulations

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
t	**0,5**	**1,5**	**2,5**	**3,5**	**4,5**	**0,5**	**1,5**	**2,5**	**3,5**	**4,5**
$\mathrm{biais}\left(\hat{\theta}_{\mathrm{RKM}}\right)$	0,000	0,000	0,000	0,000	0,000	-0,001	0,001	0,000	0,000	0,001
$\mathrm{biais}\left(\hat{\theta}_{\mathrm{RKM},bag}\right)$	-0,001	0,000	-0,001	0,000	0,000	-0,001	0,001	0,000	0,001	0,001
$\text{é-t}\left(\hat{\theta}_{\mathrm{RKM}}\right)$	0,043	0,044	0,030	0,015	0,012	0,030	0,030	0,020	0,011	0,009
$\text{é-t}\left(\hat{\theta}_{\mathrm{RKM},bag}\right)$	0,042	0,042	0,028	0,014	0,012	0,030	0,029	0,019	0,011	0,009
$\frac{EQM_p\left(\hat{\theta}_{\mathrm{RKM},bag}\right)}{EQM_p\left(\hat{\theta}_{\mathrm{RKM}}\right)}$	0,965	0,911	0,877	0,914	0,917	0,976	0,928	0,917	0,918	0,981

Tableau 5.6
Biais relatif, probabilité de couverture et largeur de l'intervalle de confiance des estimateurs de variance bootstrap pour l'estimateur de RKM (5.1) et des estimateurs de variance non corrigé $\left(\hat{V}_1\right)$ et corrigé $\left(\hat{V}_2\right)$ pour l'agrégation bootstrap des estimateurs de RKM; mêmes conditions de simulation qu'au tableau 5.5

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
t	**0,5**	**1,5**	**2,5**	**3,5**	**4,5**	**0,5**	**1,5**	**2,5**	**3,5**	**4,5**
$\frac{\mathrm{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{\mathrm{RKM}}\right)\right]}{V\left(\hat{\theta}_{\mathrm{RKM}}\right)}$	1,081	1,192	1,078	1,082	1,078	1,016	1,045	1,138	1,121	1,016
$\frac{\mathrm{E}\left[\hat{V}_1\left(\hat{\theta}_{\mathrm{RKM},bag}\right)\right]}{V\left(\hat{\theta}_{\mathrm{RKM},bag}\right)}$	1,115	1,324	1,183	1,198	1,156	1,038	1,138	1,223	1,210	1,062
$\frac{\mathrm{E}\left[\hat{V}_2\left(\hat{\theta}_{\mathrm{RKM},bag}\right)\right]}{V\left(\hat{\theta}_{\mathrm{RKM},bag}\right)}$	1,087	1,117	0,962	1,042	1,019	1,009	1,083	1,106	1,118	1,002
P.C.(I.C.)	0,958	0,963	0,955	0,956	0,959	0,954	0,956	0,966	0,964	0,948
$\mathrm{P.C.}\left(\mathrm{I.C.1}_{\cdot bag}\right)$	0,958	0,968	0,958	0,967	0,964	0,958	0,964	0,970	0,970	0,956
$\mathrm{P.C.}\left(\mathrm{I.C.2}_{\cdot bag}\right)$	0,957	0,954	0,937	0,951	0,950	0,955	0,958	0,959	0,960	0,948
Largeur(I.C.)										
$\mathrm{Largeur}\left(\mathrm{I.C.1}_{\cdot bag}\right)$	0,171	0,183	0,116	0,074	0,052	0,122	0,122	0,083	0,049	0,034
$\mathrm{Largeur}\left(\mathrm{I.C.2}_{\cdot bag}\right)$	0,169	0,168	0,105	0,069	0,049	0,120	0,120	0,079	0,047	0,033

Dans le contexte des estimateurs non lisses comme ceux considérés ici, il est souvent recommandé d'utiliser un bootstrap lisse plutôt qu'un bootstrap simple pour estimer la variance. Nous avons envisagé de perturber chaque observation rééchantillonnée y^*_{hi} de la strate h pour obtenir

$$\tilde{y}^*_{hi} = \bar{y}_h + \left(1+\sigma_Z^2\right)^{-1/2}\left(y^*_{hi} - \bar{y}_h + s_h Z^*\right), \tag{5.2}$$

où $\bar{y}_h$ et s_h sont la moyenne de l'échantillon et l'écart-type de la strate de l'échantillon original, y^*_{hi} est la valeur rééchantillonnée à l'origine, et Z^* est le bruit aléatoire où $Z^* \overset{iid}{\sim} N\left(0,\sigma_Z^2\right)$. La variance de Z^* contrôle le degré de lissage. Nous avons appliqué cette méthode à l'estimation du quantile et la proportion inférieure à un niveau estimé, mais cela n'a pas semblé améliorer l'efficacité de la procédure d'estimation. Une explication possible est que la contamination par le bruit déstabilise les observations répétées découlant de l'échantillon avec remise et stabilise l'estimateur de variance subséquent dans une certaine mesure. Comme nous avons utilisé la méthode d'échantillonnage sans remise, nous avons évité ce problème en grande partie. Une étude plus approfondie est nécessaire pour comprendre les effets du lissage dans ce contexte.

6 Conclusions

Dans le présent article, nous avons examiné l'utilisation des procédures d'agrégation bootstrap pour les estimateurs d'enquête non linéaires et non différenciables. Nous avons présenté des résultats théoriques de

l'agrégation bootstrap des estimateurs fondés sur le plan de sondage et de ceux fondés sur le modèle. L'estimateur agrégé par bootstrap peut être traité comme l'espérance d'un estimateur à deux phases avec conditionnement sur la première phase, et cette espérance lisse les « sauts » de l'estimateur non différenciable. L'étude empirique révèle le potentiel de l'agrégation bootstrap des estimateurs d'enquête non différenciables. Bien que l'efficacité relative de l'agrégation bootstrap varie d'un scénario à l'autre, les résultats sont prometteurs.

Il reste à déterminer comment estimer la variance des estimateurs agrégés par bootstrap lorsque le plan d'échantillonnage est généralement complexe. Nous avons proposé deux méthodes d'estimation de la variance à des fins pratiques, mais une étude théorique plus approfondie de l'estimation de la variance dans un cadre fondé sur le plan de sondage serait certainement justifiée.

Annexe

A.1 Théorie fondée sur le plan de sondage

Les hypothèses D.1 à D.6 sont utilisées pour illustrer les résultats fondés sur le plan de sondage figurant ci-dessous (théorèmes 3 et 4). L'hypothèse D.1 spécifie les conditions de moments sur la variable étudiée $\mathbf{y}_i$, tandis que l'hypothèse D.2 spécifie les conditions sur la probabilité d'inclusion de second ordre du plan d'échantillonnage. L'hypothèse D.3 garantit que la taille de chaque rééchantillonnage converge à la limite à l'infini. L'hypothèse D.4 spécifie les conditions de lissage sur $m(\cdot)$ dans l'estimateur différenciable. Les hypothèses D.5-D.6 montrent la convergence par rapport au plan de l'agrégation bootstrap des estimateurs d'enquête non différenciables.

(D.1) La variable étudiée $\mathbf{y}_i$ a un moment de population finie $2+\delta$ pour une valeur arbitrairement petite $\delta > 0$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{y}_i^{2+\delta} \| \; < \; \infty,$$

où chaque élément de $\mathbf{y}_i^{2+\delta}$ est l'élément original élevé à la puissance $2+\delta$ et $\|\cdot\|$ est la norme euclidienne.

(D.2) Pour tous les N, $\min_{i\in U_N} \pi_i \geq \pi_N^* > 0$, où $N\pi_N^* \to \infty$, et

$$\lim_{N\to\infty} \sup\, n \cdot \max \left| \pi_{ij} - \pi_i \pi_j \right| < \infty,$$

où π_{ij} est la probabilité d'inclusion conjointe des éléments i, j.

(D.3) Le processus de rééchantillonnage générant A_b est l'EASSR de taille k, où $k = O\left(n^{\kappa}\right)$, $\kappa \in (0,1]$. De plus, chaque nouvel échantillon bootstrap de taille k est utilisé pour calculer l'estimateur agrégé par bootstrap.

(D.4) La fonction $m(\cdot)$ est différenciable et a une dérivée seconde continue non triviale dans un voisinage compact de $\boldsymbol{\mu}_N$.

(D.5) L'estimateur $\hat{\boldsymbol{\lambda}}$ converge vers la cible de population $\boldsymbol{\lambda}_N$ à une vitesse de $\sqrt{n}$, $\lim_{N\to\infty}\boldsymbol{\lambda}_N = \boldsymbol{\lambda}_\infty$ et l'estimateur $\hat{\boldsymbol{\lambda}}$ est une statistique symétrique.

(D.6) La fonction $h(\cdot)$ est bornée et la quantité de population est [*traduction*] « différenciable de manière compacte dans un sens faible » (Dümbgen 1993). Il existe une fonction $g(\cdot)$ telle que

$$\sup_{\mathbf{s}\in C_{\mathbf{s}}}\left|\frac{1}{N}\sum_{i=1}^{N}h\left(\mathbf{y}_i-\boldsymbol{\lambda}_\infty-N^{-\alpha}\mathbf{s}\right)-\frac{1}{N}\sum_{i=1}^{N}h\left(\mathbf{y}_i-\boldsymbol{\lambda}_\infty\right)-g\left(\boldsymbol{\lambda}_\infty\right)N^{-\alpha}\mathbf{s}\right|\to 0,$$

où $C_{\mathbf{s}}$ est un ensemble compact suffisamment important de $\mathbb{R}^p$, $0<\alpha\le 1/2$ et $g\left(\boldsymbol{\lambda}_\infty\right)$ est borné.

Le théorème suivant donne plusieurs approximations asymptotiques de l'estimateur agrégé par bootstrap, selon le taux de convergence de k par rapport à n. Dans les trois cas, l'estimateur agrégé par bootstrap est conforme au plan. Intuitivement, l'estimateur agrégé par bootstrap se comporte comme l'estimateur original lorsque la taille k du rééchantillonnage est importante (tend vers l'infini à une vitesse d'au moins $n^{1/2}$), mais converge à une vitesse différente lorsque le rééchantillonnage est de petite taille.

Théorème 3 *Sous les hypothèses D.1 à D.4, l'estimateur différenciable agrégé par bootstrap* $\hat{\theta}_{d,bag}$ *admet l'expansion de deuxième ordre suivante :*

$$\hat{\theta}_{d,bag}-\theta_d=\begin{cases}\left\{m'\left(\boldsymbol{\mu}_N\right)\right\}^T\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_N\right)+o_p\left(n^{-1/2}\right), & \text{pour } \kappa>1/2\\ \left\{m'\left(\boldsymbol{\mu}_N\right)\right\}^T\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_N\right)+ & \\ \quad\frac{1}{2\binom{n}{k}}\sum\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)^T m''\left(\boldsymbol{\mu}_N\right)\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)+o_p\left(n^{-1/2}\right), & \text{pour } \kappa=1/2\\ \frac{1}{2\binom{n}{k}}\sum\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)^T m''\left(\boldsymbol{\mu}_N\right)\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)+o_p\left(k^{-1}\right), & \text{pour } \kappa<1/2\end{cases}$$

où $\kappa>0$ *est tel que la taille du rééchantillonnage* $k=O\left(n^{\kappa}\right)$.

Preuve du théorème 3 :

La preuve découle facilement d'une expansion de Taylor de chaque estimateur fondé sur un rééchantillonnage $m\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)\right)$ autour de $\boldsymbol{\mu}_N$. Le terme d'expansion linéaire se réduit à

$\{m'(\boldsymbol{\mu}_N)\}^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_N)$ sur la base d'un argument antérieur. Sous D.1 et D.3, le terme quadratique a le même ordre que la variance de l'EASSR de $\hat{\boldsymbol{\mu}}(\mathcal{Y}_b^*)$ et est donc $o_p(1/k)$.

Le théorème 4 donne ensuite la convergence par rapport au plan de l'estimateur agrégé par bootstrap non différenciable.

Théorème 4 *Sous les hypothèses D.1-D.3 et D.5-D.6, l'estimateur agrégé par bootstrap non différenciable* $\hat{\theta}_{nd,bag}$ *est conforme au plan pour sa cible de population* θ_{nd}, *i.e.,* $\hat{\theta}_{nd,bag}-\theta_{nd}=o_p(1)$.

Preuve du théorème 4 :

Nous pouvons établir que $(1/N)\sum_{i\in A}(1/\pi_i)h(\mathbf{y}_i-\boldsymbol{\lambda}_N)$ est conforme au plan pour θ_{nd} en conséquence de D.2 et que $h(\cdot)$ est borné (D.6). Il suffit alors de démontrer que $\hat{\theta}_{nd,bag}-(1/N)\sum_{i\in A}(1/\pi_i)h(\mathbf{y}_i-\boldsymbol{\lambda}_N)=o_p(1)$, ou

$$\frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\left\{\frac{1}{\binom{n-1}{k-1}}\sum_{A_b\ni i}h\left(\mathbf{y}_i-\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)-h(\mathbf{y}_i-\boldsymbol{\lambda}_N)\right\}=o_p(1)$$

suivant (2.6). Nous pouvons établir que l'ensemble d'estimateurs fondés sur le rééchantillonnage $\hat{\boldsymbol{\lambda}}(\mathcal{Y}_b^*)$ est uniformément contenu dans un voisinage de $\boldsymbol{\lambda}_N$, ou, $\sup_{A_b}\left|\hat{\boldsymbol{\lambda}}(\mathcal{Y}_b^*)-\boldsymbol{\lambda}_N\right|=O(N^{-\alpha}\mathbf{s})$ pour certains $\alpha>0$. Nous pouvons alors appliquer D.6 pour conclure à la convergence par rapport au plan de l'estimateur agrégé par bootstrap.

A.2 Théorie fondée sur le modèle

Les hypothèses M.1 à M.4 sont utilisées pour montrer les résultats fondés sur le modèle (théorèmes 1 et 2). L'hypothèse M.1 spécifie la répartition de superpopulation des caractéristiques de population $\mathbf{y}_i$. Les hypothèses M.2 et M.3 supposent un échantillonnage aléatoire simple sans remise pour le plan de sondage et le processus de rééchantillonnage. L'hypothèse M.5 est requise pour montrer les résultats asymptotiques fondés sur un modèle pour l'estimateur agrégé par bootstrap défini par les équations d'estimation.

(M.1) La séquence de caractéristiques de population $\mathbf{y}_i$ constitue un échantillon *iid* d'une distribution de probabilités de densité $f_Y(\mathbf{y})$.

(M.2) Le plan d'échantillonnage est ignorable ou, ce qui revient au même, les observations échantillonnées et non échantillonnées sont distribuées de la même façon.

(M.3) Le processus de rééchantillonnage générant A_b est l'EASSR de taille k, où la taille de l'échantillon bootstrap k est bornée. De plus, chaque nouvel échantillon bootstrap de taille k est utilisé pour calculer l'estimateur agrégé par bootstrap.

(M.4) La fonction $h(\cdot)$ est bornée.

(M.5) Soit $S_\infty(\gamma) = \mathrm{E}\,\psi(y_i - \gamma)$ une fonction continue de γ, et $\theta_{ee,\infty}$ la plus petite racine de $S_\infty(\gamma) = 0$; pour un y arbitraire à l'appui de la variable aléatoire y_i, la quantité

$$\inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k-1}\psi(y_i - \gamma) + \frac{1}{k}\psi(y - \gamma) \geq 0\right\}$$

appartient à un ensemble compact avec probabilité 1.

Preuve du théorème 1 :

L'estimateur agrégé par bootstrap $\hat{\theta}_{nd,bag}$ est une statistique symétrique, à condition que $\hat{\boldsymbol{\lambda}}$ soit symétrique (Lee 1990). Nous pouvons le projeter sur une seule dimension, disons $\mathbf{y}_1$, mais les projections sur d'autres observations sont équivalentes en raison de la symétrie,

$$\begin{aligned}
&\mathrm{E}\left\{\hat{\theta}_{nd,bag}\,\middle|\,\mathbf{y}_1 = \mathbf{y}\right\} \\
&= \mathrm{E}\left\{\frac{1}{n}\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni 1} h\left(\mathbf{y}_1 - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\,\middle|\,\mathbf{y}_1 = \mathbf{y}\right\} + \mathrm{E}\left\{\frac{n-1}{n}\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni \{i,1\}, i \neq 1} h\left(\mathbf{y}_1 - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\,\middle|\,\mathbf{y}_1 = \mathbf{y}\right\} \\
&= \frac{1}{n}u(\mathbf{y}) + \frac{k-1}{n}v(\mathbf{y}).
\end{aligned}$$

Nous pouvons alors dériver la linéarisation suivante de l'estimateur agrégé par bootstrap en appliquant la théorie des statistiques symétriques,

$$\hat{\theta}_{nd,bag} - \theta_{nd,\infty} = \frac{1}{n}\sum_{i=1}^{n}\left\{u(\mathbf{y}_i) - \theta_{nd,\infty}\right\} + \frac{k-1}{n}\sum_{i=1}^{n}\left\{v(\mathbf{y}_i) - \theta_{nd,\infty}\right\} + o_p\left(n^{-1/2}\right),$$

où $u(\cdot)$, $v(\cdot)$ et $\theta_{nd,\infty}$ sont définies dans le théorème 1. Il est facile de calculer la variance asymptotique (3.3) étant donné l'hypothèse d'échantillonnage *iid* .

Preuve du théorème 2 :

L'estimateur agrégé par bootstrap défini en (2.7) peut être traité comme une statistique U d'ordre k d'un échantillon, avec fonction noyau

$$h(y_1, y_2, \ldots, y_k) = \inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k}\psi(y_i - \gamma) \geq 0\right\}.$$

Nous pouvons appliquer directement une formule bien connue pour linéariser la statistique U (Serfling 1980; van der Vaart 1998, p. 161) afin d'obtenir la linéarisation

$$\hat{\theta}_{ee,bag} - \theta_{ee,\infty} = \frac{k}{n}\sum_{i=1}^{n}\left\{u\left(y_i\right) - \theta_{ee,\infty}\right\} + o_p\left(n^{-1/2}\right),$$

où

$$\begin{aligned} u(y) &= \mathrm{E}\ h\left(y, y_1, y_2, \ldots, y_{k-1}\right) \\ &= \mathrm{E}\ \inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k-1}\psi\left(y_i - \gamma\right) + \frac{1}{k}\psi\left(y - \gamma\right) \geq 0\right\}. \end{aligned}$$

L'estimateur d'équation d'estimation agrégé par bootstrap (2.7) peut être linéarisé comme suit :

$$\hat{\theta}_{ee,bag} - \theta_{ee,\infty} = \frac{k}{n}\sum_{i=1}^{n}\left\{u\left(y_i\right) - \theta_{ee,\infty}\right\} + o_p\left(n^{-1/2}\right). \qquad \text{(A.1)}$$

La variance asymptotique de $\hat{\theta}_{ee,bag}$ peut être obtenue directement de la linéarisation (A.1).

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition, New York: John Wiley and Sons.

Bahadur, R.R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics,* 37, 577-580.

Beran, R. et Srivastava, M. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, 13, 95-115.

Berger, Y.G. et Skinner, C.J. (2003). Variance estimation for a low income proportion. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 52 (4), 457-468.

Bickel, P., Götze, F. et van Zwet, W. (1997). Resampling fewer than n observations: gains, losses and remedies for losses. *Statistica Sinica,* 7, 1-31.

Breidt, F. et Opsomer, J. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Annals of Statistics*, 36, 403-427.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Bühlmann, P. et Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30 (4), 927-961.

Buja, A. et Stuetzle, W. (2006). Observations on bagging. *Statistica Sinica*, 16 (2), 323-351.

Chen, S.X. et Hall, P. (2003). Effects of bagging and bias correction on estimators defined by estimating equations. *Statistica Sinica*, 13 (1), 97-109.

Davison, A. et Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95, 125-140.

Dunstan, R. et Chambers, R.L. (1986). Model-based confidence intervals in multipurpose surveys. *Applied Statistics*, 35, 276-280.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1-26.

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

Frees, E.W. (1989). Infinite order U-statistics. *Scandinavian Journal of Statistics*, 16, 29-45.

Friedman, J.H. et Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137 (3), 669-683.

Fuller, W. (2009). *Sampling Statistics*. John Wiley and Sons.

Godambe, V. et Thompson, M. (2009). Estimating functions and survey sampling. Dans C. Rao et D. P. (éditeurs) (Eds.), *Handbook of Statistics, vol. 29: Sample Surveys: Inference and Analysis*, 669-687. Elsevier/North-Holland.

Hall, P. et Robinson, A. (2009). Reducing variability of crossvalidation for smoothing parameter choice. *Biometrika*, 96, 175-186.

Heilig, C.M. (1997). *An Empirical Process Approach to U-processes of Increasing Degree*. Thèse de doctorat, University of California, Berkeley.

Heilig, C.M. et Nolan, D. (2001). Limit theorems for the infinite-degree U-process. *Statistica Sinica*, 11, 289-302.

Inoue, A. et Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association*, 103 (482), 511-522.

Knight, K. et Bassett, J.G. (2002). Second order improvements of sample quantiles using subsamples. Unpublished manuscript.

Korn, E.L. et Graubard, B.I. (1998). Intervalles de confiance pour les proportions à petit nombre d'évènements positifs prévus estimées au moyen des données d'enquête. *Techniques d'enquête*, 24 (2), 209-218.

Lee, A.J. (1990). *U-statistics: Theory and Practice*. Marcel Dekker Inc.

Lee, T.-H. et Yang, Y. (2006). Bagging binary and quantile predictors for time series. *Journal of Econometrics*, 135 (1-2), 465-497.

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

Rao, J. et Wu, C. (1988). Resampling inference with complex surveys. *Journal of the American Statistical Association*, 83 (401), 231-241.

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1997). *Model Assisted Survey Sampling*. Springer-Verlag Inc (Berlin; New York).

Schick, A. et Wefelmeyer, W. (2004). Estimating invariant laws of linear processes by U-statistics. *The Annals of Statistics*, 32, 603-632.

Sering, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons.

Shao, J. et Rao, J. (1993). Standard errors for low income proportions estimated from stratified multistage samples. *Sankhya B,* 55, 393-414.

Sitter, R.R. et Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics & Probability Letters*, 52 (4), 353-358.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wang, J.C. et Opsomer, J.D. (2011). On the asymptotic normality and variance estimation of nondifferentiable survey estimators. *Biometrika*, 98, 91-106.