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

Techniques d'enquête

Estimation de la variance dans le calage à plusieurs phases

par Noam Cohen, Dan Ben-Hur et Luisa Burck

Date de diffusion : le 22 juin 2017

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 STATCAN.infostats-infostats.STATCAN@canada.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

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 « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Signes conventionnels dans les tableaux

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)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2017
Vol. 43, nº 1, pp. 139-154
Statistique Canada, nº 12-001-X au catalogue

Estimation de la variance dans le calage à plusieurs phases

Noam Cohen, Dan Ben-Hur et Luisa Burck[1]

Résumé

L'obtention d'estimateurs dans un processus de calage à plusieurs phases requiert le calcul séquentiel des estimateurs et des poids calés des phases antérieures afin d'obtenir ceux de phases ultérieures. Déjà après deux phases de calage, les estimateurs et leurs variances comprennent des facteurs de calage provenant des deux phases, et les formules deviennent lourdes et non informatives. Par conséquent, les études publiées jusqu'à présent traitent principalement du calage à deux phases, tandis que le calage à trois phases ou plus est rarement envisagé. Dans certains cas, l'analyse s'applique à un plan de sondage particulier et aucune méthodologie complète n'est élaborée pour la construction d'estimateurs calés ni, tâche plus difficile, pour l'estimation de leur variance en trois phases ou plus. Nous fournissons une expression explicite pour calculer la variance d'estimateurs calés en plusieurs phases qui tient pour n'importe quel nombre de phases. En spécifiant une nouvelle représentation des poids calés en plusieurs phases, il est possible de construire des estimateurs calés qui ont la forme d'estimateurs par la régression multivariée, ce qui permet de calculer un estimateur convergent de leur variance. Ce nouvel estimateur de variance est non seulement général pour tout nombre de phases, mais possède aussi certaines caractéristiques favorables. Nous présentons une comparaison à d'autres estimateurs dans le cas particulier du calage à deux phases, ainsi qu'une étude indépendante pour le cas à trois phases.

Mots-clés : Calage; échantillonnage à plusieurs phases; régression généralisée.

1 Introduction

La statistique des sondages fait appel à l'information auxiliaire disponible sur les totaux de population connus pour améliorer les estimations. Un estimateur par calage utilise des poids calés qui, selon une mesure de distance donnée, sont aussi proches que possible des poids de sondage initiaux, tout en satisfaisant un ensemble de contraintes induites par l'information auxiliaire. Des plans d'échantillonnage arbitraires sont permis à toutes les phases de l'échantillonnage, et l'information auxiliaire peut être utilisée à toute phase et est intégrée dans le processus d'estimation.

L'échantillonnage à plusieurs phases assorti du calage sur des données auxiliaires connues est une technique puissante et rentable. Le processus de calage a été étudié abondamment et, parmi les plans à plusieurs phases, le cas particulier de l'échantillonnage à deux phases est une exception qui a fait l'objet de recherches minutieuses. Rao (1973) et Cochran (1977, chapitre 12) ont donné les résultats fondamentaux pour la stratification et la non-réponse sous échantillonnage à deux phases. Un cadre détaillé de l'approche de pondération linéaire sous échantillonnage à deux phases est présenté dans Särndal, Swensson et Wretman (1992, chapitre 9). D'autres procédures d'estimation ont été étudiées pour des plans d'échantillonnage importants, dont le cas où l'échantillon de deuxième phase a été restratifié en utilisant l'information recueillie auprès de l'échantillon de première phase (Binder, Babyak, Brodeur, Hidiroglou et Jocelyn 2000). L'estimation de la variance a été le sujet principal de travaux de recherche dynamiques faisant appel à différentes approches, telles la méthode de linéarisation présentée dans Binder (1996), l'utilisation du jackknife (Kott et Stukel 1997) ou d'autres procédures de rééchantillonnage (Rao et Shao 1992; Fuller 1998; Kim, Navarro et Fuller 2006). Davantage en rapport avec nos travaux, Breidt et Fuller (1993) ont donné des

1. Noam Cohen, Dan Ben-Hur et Luisa Burck, Statistical Methodology Department, The Central Bureau of Statistics, 95464 Jérusalem, Israël. Courriel : avinoam.cohen@mail.huji.ac.il.

procédures d'estimation efficaces pour l'échantillonnage à trois phases en présence d'information auxiliaire, et Hidiroglou et Särndal (1998) ont étudié l'utilisation d'information auxiliaire pour l'échantillonnage à deux phases tout en permettant une légère modification de la fonction de distance qui aboutit à des facteurs de calage additifs (également appelés *facteurs* g) plutôt que multiplicatifs. Une caractéristique commune de ces résultats est la représentation des poids calés de dernière phase au moyen des poids calés des phases antérieures. Il s'agit d'un inconvénient important, car cela requiert le calcul des poids de toutes les phases antérieures pour obtenir ceux des dernières phases, ce qui rend difficile la présentation d'une méthodologie bien établie montrant comment estimer la variance des estimateurs calés sous des plans comptant plus de deux phases.

Afin de résoudre ce problème, nous utilisons la modification de la fonction de distance des moindres carrés généralisée (MCG), introduite par Hidiroglou et Särndal (1998), pour obtenir une représentation du vecteur des poids calés en plusieurs phases qui ne contient que des poids exprimés au moyen des poids de sondage initiaux et n'inclut pas les facteurs g. Partant de cette représentation, nous pouvons construire des estimateurs calés en plusieurs phases possédant la forme d'estimateurs par la régression multivariée, ce qui à son tour permet d'établir une formule générale pour un estimateur convergent de la variance des estimateurs calés en plusieurs phases qui est vérifiée pour tout nombre de phases de calage. Dans le cas relativement simple du calage à deux phases, pour lequel une autre formule d'un estimateur de variance existe dans la littérature, une comparaison montre que les deux estimateurs diffèrent fondamentalement en forme et en interprétation. Il importe de souligner que, dans ce cas particulier, le nouvel estimateur de variance proposé n'apparaît pas supérieur (ni inférieur) en ce qui concerne le biais ou la variance, mais qu'il manifeste certaines autres caractéristiques favorables qui seront discutées à la section 3.2. Cependant, l'objectif principal de l'article n'est pas de prouver la supériorité dans le cas à deux phases, mais de présenter l'approche de rechange sous laquelle la nouvelle représentation des poids calés peut produire une formule explicite pour un estimateur de la variance des estimateurs calés en plusieurs phases qui est vérifiée pour tout nombre de phases.

La présentation de l'article est la suivante. À la section 2, nous donnons la notation, qui est très semblable à celle utilisée par Hidiroglou et Särndal (1998). À la section 3, nous exposons la méthodologie et présentons plus en détail, à la sous-section 3.2, les cas particuliers du calage à deux et à trois phases. À la section 4, nous présentons une étude en simulation pour illustrer certaines caractéristiques de la nouvelle approche. Enfin, nos conclusions sont présentées à la section 5 avec des propositions de domaines à explorer dans des études ultérieures.

2 Notation

La notation que nous utilisons est similaire à celle donnée dans Särndal et coll. (1992) et dans Hidiroglou et Särndal (1998). Considérons une population finie $U = \{1, \ldots, k, \ldots, N\}$. Un échantillon probabiliste de première phase $s_1 \left(s_1 \subseteq U\right)$ est tiré de la population U en utilisant un plan d'échantillonnage qui génère la probabilité de sélection π_{1k} pour la k^{e} unité de la population. Sachant que s_{i-1} a été tiré, l'échantillon de la i^{e} phase $s_i \left(s_i \subseteq s_{i-1}\right)$ est sélectionné à partir de s_{i-1} selon un plan d'échantillonnage ayant les

probabilités de sélection $\pi_{ik|s_{i-1}} \equiv \Pr\left(k \in s_i \,\middle|\, k \in s_{i-1}\right)$. Soulignons la nature conditionnelle des probabilités de sélection de la phase résultante. À partir de ce point, nous travaillons uniquement avec les poids dans le processus d'estimation. Le poids d'échantillonnage de l'unité $k \in s_i$ à la i^{e} phase conditionnée et son poids d'échantillonnage global seront désignés par $w_{ik} = 1/\pi_{ik|s_{i-1}}$ et $w_{ik}^* = \prod_{j=1}^{i} w_{jk}$, respectivement.

Soit y_k la valeur de la variable cible pour la k^{e} unité de la population à laquelle un vecteur auxiliaire $\mathbf{x}_k = \left(x_{1k},\ldots,x_{jk},\ldots,x_{Jk}\right)$ est associé. Désignons par y le vecteur d'éléments de la variable cible obtenu à la dernière phase d'échantillonnage, p. Comme il est décrit dans Särndal et coll. (1992, chapitre 9), nous partitionnons le vecteur $\mathbf{x}$ comme $\mathbf{x} = \left(\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_p'\right)'$ avec $p \le J$, de sorte que nous pourrions obtenir plus d'une variable auxiliaire à certaines phases. Le total de population de $\mathbf{x}$, $t_{\mathbf{x}} = \sum_U \mathbf{x}_k$ est supposé inconnu. Cependant, certains totaux démographiques peuvent être connus en s'appuyant sur des sources relativement exactes, comme les données de recensement ou d'autres types de fichiers administratifs. Sans perte de généralité, désignons par $\mathbf{x}_1$ le vecteur des variables connues pour toutes les unités dans la population U. Désignons par $\mathbf{x}_2$ le vecteur des variables obtenues dans l'échantillon de première phase s_1, et ainsi de suite. Pour les éléments contenus dans s_r, $r \le p$, l'information complète est alors résumée dans le vecteur $\mathbf{x} = \left(\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_r'\right)'$. Écrivons aussi $t_i = t_{\mathbf{x}_i}$.

Soit X_r la matrice de plan comprenant n_r lignes représentant n_r unités échantillonnées, et un nombre de colonnes correspondant au nombre de variables auxiliaires dans le vecteur $\mathbf{x}_r$. Notons que X_r est obtenue dans l'échantillon s_{r-1} à la $r-1^{\text{e}}$ phase de l'échantillonnage, si bien que nous pouvons concevoir U comme un échantillon s_0. Dans les conditions qui figurent par exemple dans Särndal et coll. (1992) et dans Hidiroglou et Särndal (1998), la matrice de plan X_r englobe toutes les variables auxiliaires $\mathbf{x}_1,\ldots,\mathbf{x}_r$, plutôt que simplement $\mathbf{x}_r$, et est appelée le *vecteur complet*. Néanmoins, l'analyse est la même dans les deux cas.

L'information auxiliaire disponible à chaque phase de l'échantillonnage peut être utilisée pour obtenir des poids améliorés grâce au processus de calage qui produit des facteurs de calage à utiliser dans le processus d'estimation. Nous utilisons l'indice supérieur « * » pour désigner les poids globaux, c'est-à-dire les poids tenant compte de toutes les phases. Le symbole superposé « $\sim$ » désigne les poids calés. Les facteurs g de la i^{e} phase sont désignés par g_{ik}, ce qui donne les poids calés de la i^{e} phase $\tilde{w}_{ik} = \tilde{w}_{i-1,k} w_{ik} g_{ik}$ pour $k \in s_i$, où les $\tilde{w}_{i-1,k}$ sont les poids calés de la $i-1^{\text{e}}$ phase et $\tilde{w}_{0k} = 1$. Pour $k \in s_i$ le calage par rapport à toutes les phases produit des facteurs de calage globaux désignés par g_{ik}^*. Par conséquent, nous aurons les poids calés globaux $\tilde{w}_{ik} = w_{ik}^* g_{ik}^*$, où w_{ik}^* est le poids d'échantillonnage global. Désignons par w_i le vecteur dont les composantes sont $w_{ik}; k = 1,\ldots,n_i$, et par D_i une matrice diagonale de taille n_i avec w_i sur sa diagonale. La même notation sera utilisée avec les vecteurs $w_i^*, \tilde{w}_i$ et g_i.

3 Calage avec la distance MCG

Le calage requiert la spécification d'une fonction de distance mesurant la distance entre les poids initiaux et les nouveaux poids calés. Plusieurs fonctions de distance ont été étudiées, certaines étant résumées dans

Deville et Särndal (1992). Nous nous concentrons sur la mesure de distance par les moindres carrés généralisée (MCG). La forme classique du calage à plusieurs phases sous la fonction de distance MCG consiste à trouver les valeurs $\tilde{w}_{ik}$ pour l'ensemble $k \in s_i$ qui minimisent l'expression

$$\sum_{k \in s_i} \frac{c_{ik}\left(\tilde{w}_{ik} - \tilde{w}_{i-1,k} w_{ik}\right)^2}{\tilde{w}_{i-1,k} w_{ik}} \tag{3.1}$$

sous la contrainte

$$\sum_{k \in s_i} \tilde{w}_{ik} x_{ik} = \sum_{k \in s_{i-1}} \tilde{w}_{i-1,k} x_{ik} \tag{3.2}$$

(autrement, on peut écrire $\tilde{w}_{i-1,k} w_{ik} g_{ik}$ au lieu de $\tilde{w}_{ik}$) où les $\left\{\tilde{w}_{i-1,k} : k \in s_i\right\}$ sont les poids initiaux au début de la phase i, c'est-à-dire les poids calés obtenus à la phase $i-1$; les $\left\{\tilde{w}_{ik} : k \in s_i\right\}$ sont les poids calés de la phase i que nous voulons obtenir; et les $\left\{c_{ik} : k \in s_i\right\}$ sont les facteurs positifs spécifiés utilisés pour contrôler l'importance relative que nous voulons attribuer à chacun des éléments de la somme en fonction de l'information auxiliaire disponible pour $k \in s_{i-1}$. Pour simplifier la notation, supposons à partir de maintenant que $c_{ik} = 1$ pour tout i, k. Les poids résultant de ce scénario de calage sont $\tilde{w}_{ik} = \tilde{w}_{i-1,k} w_{ik} g_{ik}$, où $g_{ik} = 1 + \left(\sum_{l \in s_{i-1}} \tilde{w}_{i-1,l} x_{il} - \sum_{l \in s_i} \tilde{w}_{i-1,l} w_{il} x_{il}\right)' T_i^{-1} x_{ik}$ avec $T_i = \sum_{l \in s_i} w_{il}^* g_{i-1,l}^* x_{il} x_{il}'$. D'où, les facteurs de calage dans ce processus agissent multiplicativement pour donner un facteur de calage global $g_{ik}^* = \prod_{j=1}^{i} g_{jk}$ pour $k \in s_i$ à la fin de la phase i.

La mesure de distance (3.1) peut être critiquée, parce que les facteurs $1/\tilde{w}_{i-1,k} w_{ik}$ pour une phase i pourraient ne pas être forcément tous finis et positifs, car les termes $g_{i-1,k}$ qui figurent dans $\tilde{w}_{i-1,k}$ au dénominateur peuvent être nuls ou négatifs, ce qui contredit la notion de distance. Un autre choix de fonction de distance, et celui que nous utiliserons dans notre analyse, consiste à remplacer (3.1) par

$$\sum_{k \in s_i} \frac{\left(\tilde{w}_{ik} - \tilde{w}_{i-1,k} w_{ik}\right)^2}{w_{i-1,k}^* w_{ik}} \tag{3.3}$$

c'est-à-dire par des poids non calés au dénominateur. Il est facile de vérifier que les poids calés globaux résultant de la minimisation de (3.3) sous la contrainte (3.2) sont (pour $p = 2$, voir Hidiroglou et Särndal 1998)

$$\tilde{w}_{pk} = w_{pk}^* \left(g_{1k} + \ldots + g_{ik} + \ldots + g_{pk} - (p-1)\right) \tag{3.4}$$

où

$$g_{ik} = 1 + \left(\sum_{l \in s_{i-1}} \tilde{w}_{i-1,l} x_{il} - \sum_{l \in s_i} \tilde{w}_{i-1,l} w_{il} x_{il}\right)' T_i^{-1} x_{ik} \tag{3.5}$$

pour $k \in s_p$ avec $T_i = \sum_{l \in s_i} w_{il}^* x_{il} x_{il}'$. Le choix d'une mesure de distance dans la construction des estimateurs calés n'est pas critique, puisque les estimateurs résultants pour une large gamme de mesures de distance sont asymptotiquement équivalents à celui qui utilise la mesure de distance MCG (3.1), Deville et

Särndal (1992). Il en est de même de la mesure de distance (3.3). Puisque l'estimateur de Horvitz-Thompson $X_1'w_1^*$ est sans biais pour t_1 avec un écart-type d'ordre de grandeur $N \cdot O\left(n_1^{-1/2}\right)$, alors $g_{1k} = 1 + O\left(n_1^{-1/2}\right)$ pour tout $k \in s_1$ et donc $\tilde{w}_{1k} = w_{1k}^*\left(1 + O\left(n_1^{-1/2}\right)\right)$. Par induction, $g_{ik} = 1 + O\left(n_i^{-1/2}\right)$ pour tout i et découlant de (3.4), $\tilde{w}_{pk} \big/ w_{pk}^* \to 1$ en probabilité avec n_p. Suggérant de nouvelles techniques en vue d'améliorer l'estimation, Farrell et Singh (2002) ont proposé d'autres types de fonction de distance du khi carré pénalisée.

3.1 Estimation

L'analyse qui suit est motivée par la nature récursive de $\tilde{w}_{ik}$ dans (3.4), où les poids calés des phases antérieures $1,\ldots,i-1$ sont emboîtés dans chaque facteur g_{ik}, ce qui requiert le calcul séquentiel des poids calés; autrement dit, il faut calculer tous les poids calés des phases antérieures pour obtenir ceux des phases ultérieures. Soient $\hat{B}_{ij}^{+} = \left(\sum_{k\in s_i} w_{ik}^* x_{ik} x_{ik}'\right)^{-1} \sum_{k\in s_j} w_{jk}^* x_{ik} x_{jk}'$ et $\hat{B}_{ij}^{-} = \left(\sum_{k\in s_i} w_{ik}^* x_{ik} x_{ik}'\right)^{-1} \sum_{k\in s_{j-1}} w_{j-1,k}^* x_{ik} x_{jk}'$ les estimateurs de $B_{ij} = \left(\sum_{k\in U} x_{ik} x_{ik}'\right)^{-1} \sum_{k\in U} x_{ik} x_{jk}'$, le coefficient de régression de $\mathbf{x}_j$ sur $\mathbf{x}_i$. La différence entre les deux estimateurs tient au fait que, tandis que $\hat{B}_{ij}^{-}$ utilise l'ensemble complet d'unités connues pour $\mathbf{x}_j$ qui est obtenu dans s_{j-1}, $\hat{B}_{ij}^{+}$ utilise uniquement le sous-ensemble $s_j \subseteq s_{j-1}$ et, donc, plus de variables que $\hat{B}_{ij}^{-}$. Soit $\hat{Z}_{ij} = \hat{B}_{ij}^{+} - \hat{B}_{ij}^{-}$ la différence entre les deux coefficients estimés qui converge vers zéro. Notons aussi $\hat{Z}_{i_1 i_2 \ldots i_k} = \prod_{j=2}^{k} \hat{Z}_{i_{j-1} i_j}$ pour $k \geq 2$ et $\hat{Z}_{i_1} = 1$ pour $k = 1$. Soit $\hat{t}_i^{-} = \sum_{k\in s_{i-1}} w_{i-1k}^* x_{ik}$ et $\hat{t}_i^{+} = \sum_{k\in s_i} w_{ik}^* x_{ik}$ les deux estimateurs de Horvitz-Thompson pour t_i, fondés sur les unités obtenues dans les échantillons s_i et s_{i-1}, respectivement. Notons que tous les estimateurs définis dans le présent paragraphe utilisent les poids de sondage globaux w^* et non les poids calés. Dans le lemme qui suit, nous donnons une représentation de $\tilde{w}_p$, le vecteur de poids calés après p phases de calage, qui dépend uniquement des poids de sondage connus au préalable $\{w_i^*\}_{i=1}^{p}$.

Lemme 3.1 *Considérons un plan d'échantillonnage à plusieurs phases avec un scénario de calage qui produit des facteurs g additifs comme il est défini dans (3.3). Une représentation des poids calés à la phase p fondée entièrement sur les poids de sondage est*

$$\begin{aligned} \tilde{w}_p &= D_p^{*\prime} 1_{n_p} + \sum_{i_1=1}^{p} A_{i_1} - \sum_{i_1<i_2}^{p} A_{i_1 i_2} \\ &\quad + \ldots + (-1)^{k+1} \sum_{i_1<\ldots<i_k}^{p} A_{i_1 i_2 \ldots i_k} + \ldots + (-1)^{p+1} A_{i_1 i_2 \ldots i_p} \end{aligned} \tag{3.6}$$

où $A_{i_1 i_2 \ldots i_k}' = \left(\hat{t}_{i_1}^{-} - \hat{t}_{i_1}^{+}\right)' \hat{Z}_{i_1 i_2 \ldots i_k} \left(X_{i_k}' D_{i_k}^* X_{i_k}\right)^{-1} X_{i_k}' D_p^*$.

Preuve. Voir l'annexe A.

Notons la forme « inclusion-exclusion » de $\tilde{w}_p$ dans le lemme 3.1. La k'^{e} sommation comprend $\binom{p}{k}$ opérandes $A_{i_1 i_2 \ldots i_k}$, pour lesquels chaque $\hat{Z}_{i_1 i_2 \ldots i_k} = \prod_{j=2}^{k} \left(\hat{B}_{i_{j-1} i_j}^{+} - \hat{B}_{i_{j-1} i_j}^{-}\right)$ contient 2^k opérandes. Soit, un total

de $\binom{p}{k}2^k$ opérandes. Le nombre global de termes dans (3.6) est par conséquent 3^p comme il est montré dans la preuve du lemme. Notons aussi que les termes $A_{i_1 i_2 \ldots i_k}$ comprennent le produit des composantes $\hat{t}_{i_1}^- - \hat{t}_{i_1}^+$ et $\hat{Z}_{i_1 i_2 \ldots i_k}$, ayant toutes deux une espérance nulle, de sorte que le poids calé $\tilde{w}_p$ est égal à $D_p^{*\prime} 1_{n_p}$, le poids de sondage global, plus les termes de correction d'ordres de grandeur plus faibles, et maintient la caractéristique bien connue des poids calés. Jusqu'à présent, nous nous sommes limités dans notre discussion à une représentation du vecteur des poids dans un processus de calage à plusieurs phases qui fait intervenir uniquement des paramètres du plan de sondage et n'inclut pas les facteurs g. Or, partant de cette représentation de $\tilde{w}_p$, il est possible de déduire un estimateur novateur pour la variance des estimateurs calés en plusieurs phases. Soit y une variable d'intérêt pour laquelle nous voulons estimer le total de population Y. Soit $\hat{\beta}_j = \left(\sum_{k \in s_j} w_{jk}^* x_{jk} x_{jk}'\right)^{-1} \sum_{k \in s_p} w_{pk}^* x_{jk} y_k$, le coefficient de régression de y sur $\mathbf{x}_j$, et $\hat{Y}_{\mathrm{HT}_p} = 1_{n_p}' D_p^* y$, l'estimateur de Horvitz-Thompson non calé, calculé sur les éléments compris dans s_p. Le réarrangement des termes dans (3.6) produit une représentation plus classique de l'estimateur calé en plusieurs phases $\tilde{w}_p' y$ sous forme d'un estimateur par la régression multivariée

$$\tilde{w}_p' y = \hat{Y}_{\mathrm{HT}_p} + \sum_{i_1=1}^{p} \left(\hat{t}_{i_1}^- - \hat{t}_{i_1}^+\right)' \hat{\gamma}_{i_1} \tag{3.7}$$

où

$$\hat{\gamma}_{i_1} = \hat{\beta}_{i_1} - \sum_{i_1 < i_2}^{p} \hat{Z}_{i_1 i_2} \hat{\beta}_{i_2} +$$

$$\ldots + (-1)^{k+1} \sum_{i_1 < \ldots < i_k}^{p} \hat{Z}_{i_1 i_2 \ldots i_k} \hat{\beta}_{i_k} + \ldots + (-1)^{p-(i_1-1)+1} \hat{Z}_{i_1 \ldots p} \hat{\beta}_p.$$

L'établissement d'un estimateur convergent de la variance des estimateurs calés en plusieurs phases est maintenant simple en ce sens qu'il suit à peu près les étapes utilisées dans le calcul de la variance sous un scénario de calage multivarié à une phase.

Théorème 3.1 *Soit* $\hat{e}_{rk} = x_{rk}' \hat{\gamma}_r - x_{r+1,k}' \hat{\gamma}_{r+1}$ *pour* $r < p$ *et* $\hat{e}_{pk} = x_{pk}' \hat{\gamma}_p - y_k$. *Un estimateur convergent de la variance de* $\tilde{w}_p' y$ *est*

$$\sum_{1 \le r_1, r_2 \le p} \sum_{k \in s_{r_m},\, l \in s_{r_M}} \frac{w_{r_M l}^*}{w_{r_m l}^*} \left(w_{r_m k}^* w_{r_m l}^* - w_{r_m kl}^*\right) \hat{e}_{r_m k} \hat{e}_{r_M l} \tag{3.8}$$

où $r_m = \min(r_1, r_2)$ *et* $r_M = \max(r_1, r_2)$.

La preuve comprend l'évaluation des ordres de grandeur les plus élevés et l'estimation de leur variance. Une attention particulière est accordée à l'évaluation de la probabilité conjointe des événements $\{k \in s_i, l \in s_j\}$ et à l'estimation de la covariance entre les unités provenant de différentes phases d'échantillonnage.

Preuve. À la première étape, nous allons voir que le remplacement des estimateurs des coefficients $\hat{\gamma}_i; i = 1 \ldots p$ par leurs valeurs réelles γ_i affecte l'estimation de la variance d'un facteur $N^2 o\left(n_p^{-1}\right)$ et, donc, n'affecte pas la convergence de l'estimateur substitué. À cette fin, notons que $\hat{B}_{ij}^+, \hat{B}_{ij}^-$ sont tous deux convergents vers B_{ij}. Écrivons $\hat{B}_{ij}^+ = B_{ij} + \left(\hat{B}_{ij}^+ - B_{ij}\right)$ de sorte que $\hat{B}_{ij}^+ = B_{ij} + O_p\left(n_j^{-1/2}\right)$. Rappelons que $\hat{Z}_{ij} = \hat{B}_{ij}^+ - \hat{B}_{ij}^-$, où $\hat{B}_{ij}^-$ est basé sur s_{j-1}, tandis que $\hat{B}_{ij}^+$ est basé sur son sous-échantillon s_j et, donc, $\hat{Z}_{ij} = O_p\left(n_j^{-1/2}\right)$ et, par conséquent, $\hat{Z}_{i_1 i_2 \ldots i_k}$ est borné par $O_p\left(n_{i_k}^{-1/2}\right)$. De même, $\hat{\beta}_j$ est $\beta_j + O_p\left(n_p^{-1/2}\right)$, parce que y est observé uniquement à la dernière phase d'échantillonnage s_p. Donc, $\hat{\gamma}_i$ est convergent vers γ_i pour tout i, où les $\hat{\beta}_i$ dans $\hat{\gamma}_i$ sont remplacés par β_i dans γ_i. La convergence n'implique pas nécessairement la convergence des moments et, en particulier, pas de la variance. Cependant, pour une population finie, c'est-à-dire un espace de probabilité fini, les concepts coïncident. Il s'ensuit que, pour n_p suffisamment grand, $\mathrm{Var}\left(\hat{Y}_{\mathrm{HT}_p} + \sum_{i_1=1}^{p}\left(\hat{t}_{i_1}^- - \hat{t}_{i_1}^+\right)' \hat{\gamma}_{i_1}\right)$ et $\mathrm{Var}\left(\hat{Y}_{\mathrm{HT}_p} + \sum_{i_1=1}^{p}\left(\hat{t}_{i_1}^- - \hat{t}_{i_1}^+\right)' \gamma_{i_1}\right)$ sont asymptotiquement équivalents et selon la discussion qui précède, la différence peut être quantifiée par

$$\mathrm{Var}\left(\tilde{w}_p' y\right) = \mathrm{Var}\left(\hat{Y}_{\mathrm{HT}_p} + \sum_{r=1}^{p}\left(\hat{t}_r^- - \hat{t}_r^+\right)' \gamma_r\right) + N^2 o\left(n_p^{-1}\right).$$

L'estimateur $\hat{t}_r^+$ est une sommation sur les unités comprises dans s_r, tandis que $\hat{t}_r^-$ est une sommation sur s_{r-1}. En réarrangeant les termes, la variance dans le deuxième membre de l'équation peut s'écrire $\mathrm{Var}\left(\sum_{r=1}^{p}\sum_{i \in S_r} w_{ri}^* e_{ri}\right)$, ce qui est égal à

$$\sum_{1 \le r_1, r_2 \le p}\sum_{k \in U}\sum_{l \in U} w_{r_1 k}^* e_{r_1 k} w_{r_2 l}^* e_{r_2 l} \mathrm{Cov}\left(I_{k \in s_{r_1}}, I_{l \in s_{r_2}}\right)$$

de sorte qu'un estimateur basé sur l'échantillon serait

$$\sum_{1 \le r_1, r_2 \le p}\sum_{k \in s_{r_1},\, l \in s_{r_2}} w_{r_1 k}^* \hat{e}_{r_1 k} w_{r_2 l}^* \hat{e}_{r_2 l}\left[1 - \frac{P\left(k \in s_{r_1}\right) P\left(l \in s_{r_2}\right)}{P\left(k \in s_{r_1}, l \in s_{r_2}\right)}\right]. \tag{3.9}$$

Pour calculer la covariance entre les indicateurs $I_{k \in s_{r_1}}$ et $I_{l \in s_{r_2}}$, nous devons connaître la probabilité conjointe des événements $\{k \in s_i, l \in s_j\}$. Si $s_j \subset s_i$, alors $P\left(k \in s_i, l \in s_j\right)$ est égale à la probabilité conjointe que les deux unités k, l soient dans l'échantillon $s_i = s_{\min(i,j)}$, multipliée par la probabilité conditionnelle que l'unité l soit dans l'échantillon s_j, sachant qu'il appartient à s_i. Formellement, si $s_j \subset s_i$, alors $P\left(k \in s_i, l \in s_j\right) = \frac{w_{il}^*}{w_{jl}^*} w_{i,lk}^{*-1}$, ce qui élimine la dépendance à l'égard de s_{r_2} entre les crochets dans (3.9) et le résultat s'ensuit.

Un autre moyen d'écrire (3.8) est

$$\sum_{1 \le r \le p}\sum_{k,l \in s_r}\left(w_{rk}^* w_{rl}^* - w_{rkl}^*\right)\hat{e}_{rk}\hat{e}_{rl} + 2\sum_{1 \le r_m < r_M \le p}\sum_{k \in s_{r_m}}\sum_{l \in s_{r_M}} w_{r_m k}^* \hat{e}_{r_m k} w_{r_M l}^* \hat{e}_{r_M l}\left(1 - \frac{w_{r_m kl}^*}{w_{r_m k}^* w_{r_m l}^*}\right).$$

Quand $p = 2,$ les termes γ_i coïncident avec les unités de variation obtenues de la décomposition de l'erreur d'échantillonnage de l'*estimateur en deux étapes* de Breidt et Fuller (1993). Des estimations convergentes pour les écarts-types des estimations calées des sous-totaux de population sont calculées de façon ordinaire en multipliant la variable cible par une variable indicatrice pour la sous-population particulière.

Jusqu'à présent dans notre discussion, nous avons donné une représentation du vecteur de poids calés de laquelle nous avons dérivé un nouvel estimateur convergent pour la variance des estimateurs calés en plusieurs phases. Cependant, dans certains cas, les estimateurs peuvent être simplifiés davantage sans perte d'exactitude. Nous discuterons brièvement ici de deux scénarios qui dépendent du fait que n_j est ou non significativement plus petit que $n_{j-1},$ c'est-à-dire du fait que, pour tout $j,$ le sous-échantillon s_j est ou non significativement plus petit que $s_{j-1}.$ Un cas type du premier scénario est celui où l'on possède un ensemble de fichiers administratifs emboîtés dont les tailles diminuent significativement. Le premier ensemble peut être, par exemple, un fichier de registre de population qui contient un nombre limité de variables au sujet de l'ensemble de la population, comme l'âge, le sexe, etc. Le deuxième ensemble peut correspondre à des données d'échantillons provenant d'une enquête de portée nationale dans le cadre de laquelle des données complètes sur les ménages ont été recueillies auprès de toutes les unités échantillonnées, mais en utilisant un questionnaire supplémentaire pour un sous-groupe de ces unités (disons, une unité sur dix). Les données pour ce sous-groupe d'unités peuvent alors être calées sur celles provenant des deux sources d'information précédentes. Un exemple du second scénario est la situation où une ou deux phases de calage sont effectuées sur le même ensemble de données. Autrement dit, contrairement au processus à plusieurs phases habituel, l'élément d'échantillonnage est présent à la première phase seulement, mais non aux phases ultérieures. Un tel scénario peut avoir lieu si nous voulons caler les données d'une enquête sur de nombreuses variables pour lesquelles nous connaissons seulement les totaux de marge, mais ne possédons pas les totaux transversaux. Dans ces conditions, une série de calages sur le même échantillon, mais en utilisant un ensemble différent de variables auxiliaires à chaque phase, en attribuant habituellement aux dernières phases les variables les plus importantes, pourrait être un compromis satisfaisant. Une meilleure façon de caractériser ce scénario serait de le dire *séquentiel*. Sous ces scénarios, $\tilde{w}_p$ et sa variance peuvent être simplifiés considérablement. Ces scénarios peuvent être énoncés comme des corollaires de notre analyse, mais nous choisissons de ne pas les prendre en considération ici afin de nous concentrer sur nos résultats courants.

3.2 Exemples : Calage à deux phases et à trois phases

Calage à deux phases. Nous utiliserons le cas particulier du calage à deux phases $(p = 2)$ pour démontrer la nouvelle méthodologie et ce qui la distingue de l'autre estimateur habituellement utilisé dans la littérature. En notation matricielle, l'estimateur calé est donné, selon (3.7), par

$$\tilde{w}_2' y = \hat{Y}_{\mathrm{HT}_2} + \left(\hat{t}_1^- - \hat{t}_1^+\right)' \hat{\gamma}_1 + \left(\hat{t}_2^- - \hat{t}_2^+\right)' \hat{\gamma}_2$$

où $\hat{\gamma}_1 = \hat{\beta}_1 - \hat{Z}_{12}\hat{\beta}_2$ et $\hat{\gamma}_2 = \hat{\beta}_2.$ Explicitement, sous forme non matricielle,

$$\tilde{w}_2' y = \sum_{k \in s_2} w_{2k}^* y_k + \left(\sum_{k \in U} x_{1k} - \sum_{k \in s_1} w_{1k} x_{1k} \right) \hat{\gamma}_1 + \left(\sum_{k \in s_1} w_{1k} x_{2k} - \sum_{k \in s_2} w_{2k}^* x_{2k} \right) \hat{\gamma}_2$$

où

$$\hat{\gamma}_1 = \left(\sum_{k \in s_1} w_{1k} x_{1k} x_{1k}' \right)^{-1} \left[\sum_{k \in s_2} w_{2k}^* x_{1k} y_k - \left(\sum_{k \in s_2} w_{2k}^* x_{1k} x_{2k}' - \sum_{k \in s_1} w_{1k} x_{1k} x_{2k}' \right) \hat{\gamma}_2 \right]$$

$$\hat{\gamma}_2 = \left(\sum_{k \in s_2} w_{2k}^* x_{2k} x_{2k}' \right)^{-1} \sum_{k \in s_2} w_{2k}^* x_{2k} y_k .$$

Cet estimateur produit des estimations identiques à l'estimateur calé en deux phases utilisé dans Hidiroglou et Särndal (1998) ou dans Särndal et coll. (1992), section 9.7. Cependant, une fois que l'estimateur des paramètres γ_1, γ_2 est calculé, la représentation de $\tilde{w}_2' y$ devient simple et informative, car elle possède la structure d'un simple estimateur par la régression multivariée. Cet estimateur linéaire est fondé sur les coefficients γ qui englobent l'effet total de la variable $\mathbf{x}$ qu'ils multiplient et, donc, diffèrent légèrement des coefficients β. $\hat{\gamma}_i$ englobe l'effet global que le calage sur la variable $\mathbf{x}_i$ a sur l'estimation de Y. Dans le cas général, il tient compte de la projection de y sur $\mathbf{x}_i$, de la projection de y sur $\mathbf{x}_{i+1}$ multipliée par la projection de $\mathbf{x}_{i+1}$ sur $\mathbf{x}_i$, et ainsi de suite. En outre, comme nous allons le montrer, les estimateurs de variance diffèrent significativement en ce qui concerne tant les estimations que la représentation. Étant donné la complexité de l'évaluation de la variance des estimateurs qui comprennent des facteurs g, jusqu'à présent dans la littérature sur le calage à deux phases, il était d'usage pratique de commencer par donner aux facteurs g la valeur approximative de 1, puis d'utiliser la loi de la variation totale pour obtenir deux composantes, une pour chaque phase, conformément à

$$\hat{V}_C \left(\tilde{w}_2' y \right) = \sum_{k,\, l \in s_2} w_{2kl} \left(w_{1k} w_{1l} - w_{1kl} \right) \left(g_{1k} \breve{e}_{1k} \right) \left(g_{1l} \breve{e}_{1l} \right)$$

$$+ \sum_{k,\, l \in s_2} w_{1k} w_{1l} \left(w_{2k} w_{2l} - w_{2kl} \right) \left(g_{2k} \breve{e}_{2k} \right) \left(g_{2l} \breve{e}_{2l} \right) \qquad (3.10)$$

où les termes d'erreur $\breve{e}_{1k} = y_k - x_{1k}' \hat{\gamma}_1$ et $\breve{e}_{2k} = y_k - x_{2k}' \hat{\gamma}_2$ sont tous deux définis pour $k \in s_2$, parce que y est observé uniquement sur s_2, et on notera la représentation simple des termes d'erreur sous la notation faisant appel aux coefficients γ. Les facteurs g sont définis comme dans (3.5). La valeur approximative de 1 donnée aux facteurs g dans le calcul de (3.10) peut indubitablement aboutir à des estimations imprévisibles, car ces facteurs s'écartent de l'unité précisément dans les situations où le calage est essentiel. Par ailleurs, l'estimateur de variance proposé en (3.8) pour un estimateur calé en deux phases est donné par

$$\hat{V}_P \left(\tilde{w}_2' y \right) = \sum_{k,\, l \in s_1} \left(w_{1k} w_{1l} - w_{1kl} \right) \hat{e}_{1k} \hat{e}_{1l} + \sum_{k,\, l \in s_2} \left(w_{2k}^* w_{2l}^* - w_{2kl}^* \right) \hat{e}_{2k} \hat{e}_{2l}$$

$$+ 2 \sum_{k \in s_1,\, l \in s_2} \frac{w_{2l}^*}{w_{1l}} \left(w_{1k} w_{1l} - w_{1kl} \right) \hat{e}_{1k} \hat{e}_{2l} . \qquad (3.11)$$

La différence entre les estimateurs de variance issus des deux méthodes représentées par les équations (3.10) et (3.11) est fondamentale. Elle se manifeste sous divers aspects. Tandis que le terme d'erreur de la deuxième phase est le même dans les deux méthodes, c'est-à-dire $\hat{e}_{2k} = \breve{e}_{2k}$, le terme d'erreur de la première phase diffère. $\breve{e}_{1k}$ est fondé sur la différence entre y_k et le prédicteur de régression $x'_{1k}\hat{\gamma}_1$, tandis que $\hat{e}_{1k}$ est basé sur la différence entre deux prédicteurs de Y provenant des phases un et deux $x'_{1k}\hat{\gamma}_1 - x'_{2k}\hat{\gamma}_2$. Cette modification fait que le premier opérande dans (3.11) est calculé sur s_1 et non sur s_2 où l'échantillon est plus grand. Comme on le voit, l'estimateur (3.11) comprend un troisième opérande qui contient le produit des deux termes d'erreur provenant des deux phases et n'a pas de parallèle dans (3.10). Bien que ce produit soit souvent proche de zéro quand les termes d'erreur ne sont pas fortement corrélés, il peut être non négligeable quand y est fortement corrélé avec $\mathbf{x}_1$. Un avantage évident est l'absence des facteurs g qui rend l'estimateur plus simple à calculer, c'est-à-dire qu'une fois que nous avons calculé les estimations des paramètres $\hat{\gamma}_i; i = 1 \ldots p$, l'estimateur (3.11) peut être calculé en utilisant les paramètres du plan uniquement, sans impliquer les facteurs g provenant de toutes les phases du calage. Enfin, aspect peut-être le plus important du point de vue opérationnel, comme nous le montrerons aussi dans l'étude en simulation, l'avantage de (3.11) est que, pour une grande gamme de plans de sondage, le deuxième opérande représente la majorité absolue de la variance, tandis que dans (3.10), les opérandes sont habituellement du même ordre de grandeur. Cette caractéristique découle du fait que le terme $\left(w^*_{2k}w^*_{2l} - w^*_{2kl}\right)$, qui comprend les poids d'échantillonnage totaux, est très grand comparativement à $w_{2kl}\left(w_{1k}w_{1l} - w_{1kl}\right)$ ou $w_{1k}w_{1l}\left(w_{2k}w_{2l} - w_{2kl}\right)$. Dans l'estimateur de variance, la fonction $f(w) = w_k w_l - w_{kl}$ atteint son maximum sur la diagonale $k = l$, où elle est proportionnelle à w_k^2, et puis elle est multipliée par le carré de son reste $\hat{e}_k$, qui est un terme non négatif. D'où, quand le taux d'échantillonnage de la seconde phase est suffisamment élevé, il accroît fortement les termes qui dépendent des poids totaux w^*_2 de cette phase, comparativement à un terme parallèle provenant de la phase précédente. Donc, le deuxième opérande peut, pratiquement à lui seul, être un bon estimateur de la variance de l'estimateur calé.

Calage à trois phases. Le calage à plusieurs phases peut être mis en œuvre quand, dans une série d'échantillons de taille décroissante (non croissante), chaque paire de phases consécutives présente certaines variables communes. Il peut être effectué que les échantillons soient emboîtés, c'est-à-dire si s_i est un sous-échantillon de s_{i-1}, ou non. En pratique, le cas le plus simple et le plus fréquent est évidemment le calage à deux phases où un plus petit échantillon (emboîté ou non) est calé sur un échantillon beaucoup plus grand, comme celui d'une Enquête sur la population active, qui est à son tour fréquemment calé sur un fichier administratif contenant des variables démographiques. Cependant, étant donné la faisabilité des calculs et les progrès méthodologiques, les plans comportant un plus grand nombre de phases de calage demeurent répandus et les plans à trois phases occupent le second rang quant à la simplicité et à la mise en œuvre. Par conséquent, cela vaut la peine de s'étendre un peu plus sur l'estimateur pour ce cas.

L'approximation (3.8) contient six termes différents, trois pour les trois phases d'échantillonnage et trois autres pour la covariance entre les phases. Nous désignons ces termes par V_1, V_2, V_3 et C_{12}, C_{13}, C_{23}, respectivement. Chacun correspond à la multiplication d'un terme qui comprend les poids d'échantillonnage par les restes pour les phases pertinentes. Les formules pour le calage à trois phases sont présentées à l'annexe B. Comme nous l'avons exposé pour le cas à deux phases, quand $w_i > 1$, les V_i suivent

vraisemblablement un ordre clair $V_1 < V_2 < V_3$ et V_3 deviendra d'autant plus dominant que les taux d'échantillonnage de la troisième phase seront grands. Cette situation est représentée par le cas 3 dans le tableau 3.1, et dans notre simulation, cela se manifeste aux lignes 2 et 6 du tableau 4.2, où w_{3k} est égal à 10 et à 5, respectivement. Ce n'est manifestement pas très souvent le cas en réalité, car l'approximation dépend aussi des tailles des termes de reste, qui dépendent du choix des variables de calage et de leurs corrélations particulières qui sont parfois très fortes. Le cas échéant, les restes seront très petits et il serait préférable d'utiliser tous les termes de (3.8). Comme pour les termes de covariance, même si C_{13} comprend les poids globaux $\{w^*_{3k}\}$, il est peu probable qu'il ajoute une valeur importante à la variance totale en raison de la corrélation généralement faible entre les restes des phases 1 et 3. Par ailleurs, le terme C_{23}, même s'il est pondéré par les poids globaux de 2^e phase seulement, peut être significatif en raison de la forte corrélation entre les restes des phases 2 et 3, car ils contiennent tous deux le terme $x'_{3k}\hat{\gamma}_3$ pour $k \in s_3$. L'importance relative des termes pour certains plans généraux est spécifiée dans le tableau 3.1. Les coefficients γ, qui englobent l'effet total des variables $\mathbf{x}$ qu'ils multiplient, prennent maintenant une forme plus intéressante et compliquée. Par exemple, $\hat{\gamma}_1$ tient compte des projections de $\mathbf{x}_1$ sur $\mathbf{x}_2$ et de $\mathbf{x}_1$ sur $\mathbf{x}_3$, mais avec déduction de la projection de $\mathbf{x}_1$ sur la projection de $\mathbf{x}_2$ sur $\mathbf{x}_3$.

Tableau 3.1
Une représentation générale de l'importance relative de chacun des termes dans (3.8) pour certains scénarios. Les points noirs indiquent une forte dominance, les points gris foncé, une dominance modérée et les points gris clair, une non-dominance

Cas	Description	V_1	V_2	V_3	C_{12}	C_{13}	C_{23}
1	Pratiquement aucun échantillonnage supplémentaire aux deuxième et troisième phases : $w_2 \approx w_3 \approx 1$.	●	●	●	○	●	○
2	Les poids w_1, w_2, w_3 sont de taille modérée.	○	◉	●	○	●	○
3	n_3 nettement plus petit que n_2, indépendamment des tailles de w_1, w_2.	○	○	●	○	○	○

4 Une étude en simulation

L'objectif principal de l'analyse exposée dans le présent article est de fournir un estimateur convergent de la variance des estimateurs calés en plusieurs phases qui est vérifié pour tout nombre de phases de calage. Une étude en simulation pourrait donc être exécutée pour comparer le nouvel estimateur à d'autres décrits dans la littérature. Comme on ne trouve généralement aucun estimateur de rechange dans la littérature pour des plans de calage à trois phases ou plus $(p \geq 3)$, notre comparaison porte principalement sur le cas à deux phases qui est celui le plus étudié. Nous avons également exécuté une étude pour $p = 3$ afin d'évaluer l'écart de l'estimateur proposé par rapport à la valeur simulée réelle. Les études sont décrites ici en termes généraux. Elles visent essentiellement à démontrer la pertinence de l'estimateur proposé, sa concordance avec la « condition limite » du cas à deux phases, et son potentiel en ce qui concerne les plans comportant plus de deux phases. Une étude approfondie en vue de caractériser l'efficacité de l'estimateur proposé en tant que fonction des paramètres du plan, tels que les taux d'échantillonnage, le choix des variables de calage et leur corrélation avec y, etc., est réservée à de futurs travaux de recherche.

Un processus d'estimation sous calage à deux phases a été appliqué aux données d'une enquête récente sur la carrière et la mobilité des titulaires d'un doctorat (TD). Comme il n'existe pas de base de sondage des TD, les données sur les études supérieures ont été extraites d'un recensement de population récent. Cependant, seul un échantillon S_1 qui représente un cinquième des ménages dénombrés au recensement a reçu un questionnaire détaillé contenant des questions sur les études supérieures. Pour l'enquête sur les TD, on a tiré de S_1 un sous-échantillon S_2 dans lequel les personnes qui étaient en fait TD ont reçu un questionnaire encore plus détaillé. Donc, un scénario de calage à deux phases pour estimer les caractéristiques des TD était de mise. La première phase comprenait le calage des variables conjointes de S_1 et S_2 sur les totaux estimés calculés d'après S_1. À la deuxième phase, les données démographiques de S_1 ont été calées sur les totaux connus provenant du registre de la population complète U. Nous avons réalisé une étude en simulation sur ces données, dans laquelle les données d'enquête ont servi de population réelle. Mille échantillons (réalisations) $\{u, s_1, s_2\}$ de tailles $N = 1\,000$, $n_1 = 200$, $n_2 = 50$ ont été tirés aléatoirement de l'ensemble de données S_2 de TD. À chaque échantillon, nous avons appliqué le même processus de calage à deux phases en utilisant l'estimateur donné par (3.7) avec l'équation (3.6) comme représentation des poids calés $\tilde{w}_2$, et son estimateur de variance donné par (3.11) comme un cas particulier de (3.8). Comme nous l'avons déjà mentionné, quand $p = 2$, les estimations $\hat{Y} = \tilde{w}_2' y$ sont identiques sous la nouvelle représentation ou sous la représentation classique utilisée jusqu'à présent dans la littérature, Särndal et coll. (1992). Donc, nous nous sommes concentrés sur les estimateurs de variance (3.10) et (3.11) calculés selon les deux méthodes. Un profil type de la comparaison entre les deux estimateurs de variance dans ce cas particulier du calage à deux phases est présenté à la figure 4.1. On voit que, malgré la différence fondamentale entre les deux estimateurs de variance, dans la plupart des réalisations, la différence entre leurs estimations est assez faible. Néanmoins, pour l'une des réalisations, elle peut aller jusqu'à 20 %. Pour la variable particulière présentée dans la figure, les valeurs moyennes des deux estimateurs de la variance étaient très semblables, à savoir $54{,}17^2$ et $54{,}65^2$, tandis que la valeur réelle dans les données de simulation était de $54{,}46^2$. Même les variances de leur estimateur de l'écart-type, à savoir $5{,}73^2$ contre $5{,}93^2$, étaient presque les mêmes pour cette variable. Ces résultats sont présentés au tableau 4.1. La caractéristique favorable de l'estimateur proposé ressort dans la 5e colonne. Contrairement à l'estimateur classique dans lequel les deux termes de l'estimateur de variance sont du même ordre de grandeur, le 2e terme de (3.11) représente plus de 99 % de la variance, avec une variation de moins de 2 % sur l'ensemble des 1 000 réalisations. Nous avons donné l'explication de ce phénomène à la section 3.2. Les résultats présentés ici se sont répétés pour toutes les variables étudiées et nous avons jugé non pertinent à ce stade de présenter d'autres variables ou d'étudier plus en profondeur ces données particulières ou le cas particulier du calage à deux phases.

Tableau 4.1
Estimateur proposé (P) c. classique (C) pour l'écart-type d'un estimateur calé en deux phases

Variable	Valeur moyenne	É.-T.	Couverture de l'IC	2e terme en pourcentage de $\widehat{\text{É.-T.}}\left(\tilde{w}_2' y\right)$
$\tilde{w}_2' y$	200,43	54,46		
$\widehat{\text{É.-T.}}_C\left(\tilde{w}_2' y\right)$	54,65	5,93	95,2 %	77 % ± 7 %
$\widehat{\text{É.-T.}}_P\left(\tilde{w}_2' y\right)$	54,17	5,73	95,1 %	99 % ± 2 %

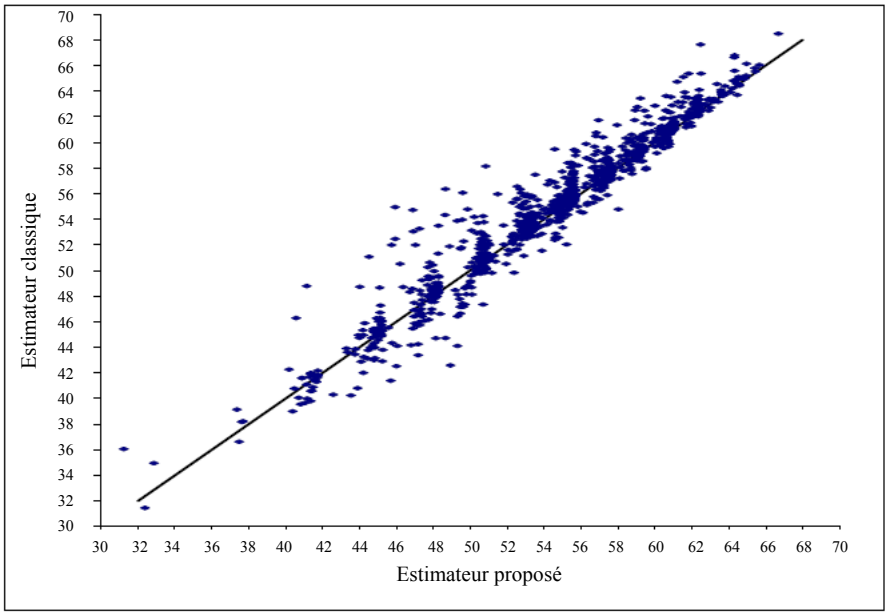

Figure 4.1 Estimations de la variance dans le calage à deux phases. Un profil type de 1 000 réalisations de l'estimateur proposé (équation 3.11) en fonction de l'estimateur classique (équation 3.10) pour la variance d'un estimateur calé de Y. La droite en trait plein est la diagonale principale.

La similarité des estimations des deux estimateurs de variance dans le cas du calage à deux phases est rassurante, mais il n'a pas été possible d'effectuer la comparaison dans le cas du calage à trois phases ou plus, parce qu'il n'existe pas d'alternative à l'estimateur proposé. Une méthode par rééchantillonnage pour l'échantillonnage à deux phases stratifié a été proposée par Kim et coll. (2006), et nous exposons brièvement une ébauche de généralisation pour un cas à trois phases, mais sans formulation explicite ni résultats de simulation. Nous avons ajouté une troisième phase de calage dans notre simulation en utilisant certaines variables en commun avec l'échantillon de deuxième phase des TD, choisies en fonction de l'expérience sur le terrain, et avons procédé de la même façon que dans le cas à deux phases. L'étude en simulation a de nouveau révélé une excellente estimation pour la variance d'un estimateur calé en trois phases pour toutes les variables Y examinées et chacun des différents ensembles de variables de calage à toutes les phases. Les taux de convergence de l'estimateur de variance sont rapides, même pour de très petites tailles d'échantillon, telles que $n = 25$ ou moins à la troisième phase. Certains résultats pour divers paramètres de plan de sondage sont présentés au tableau 4.2. Comme indiqué plus haut, la simulation a été exécutée sur une taille de population de 1 000 de manière que les trois premiers plans aient un poids global de $w^* = 40$, et les trois suivants, de $w^* = 20$. Donc, comme prévu, la variance de l'estimateur calé pour les trois premiers plans est généralement plus élevée, bien qu'elle dépende aussi des tailles d'échantillon des première et deuxième phases, comme le montre, par exemple, le cas artificiel numéro 4 qui dépeint un scénario généralement impossible en pratique. Les biais relatifs $\frac{E\left(\widehat{\text{É.-T.}}_P\right)}{\text{É.-T.}} - 1$ sont proches de zéro pour tous les plans étudiés et les couvertures des intervalles de confiance (IC) à 95 %, estimées également, se sont avérées principalement raisonnables et proches des niveaux nominaux. L'écart-type de $\widehat{\text{É.-T.}}_P$ vaut approximativement 5 % à 10 % de la valeur de l'estimateur, comme le montre la colonne 7.

Tableau 4.2
Valeurs vraie et estimée de l'écart-type d'un estimateur calé en trois phases $\tilde{w}_3' y$ pour divers paramètres de plan de sondage

Cas	n1	n2	n3	Valeur vraie	$\widehat{\text{É.-T.}}_P$	É.-T. de $\widehat{\text{É.-T.}}_P$ en %	Couverture de l'IC à 95 %
1	100	50	25	882,6	866,9	7,1 %	94,9 %
2	500	250	25	781,5	774,1	10,8 %	95,2 %
3	500	100	25	733,9	731,5	10,2 %	96,0 %
4	50	50	50	902,8	892,1	4,8 %	95,6 %
5	200	100	50	598,1	591,4	5,4 %	94,4 %
6	500	250	50	543,0	542,2	8,3 %	96,3 %
7	333	100	33	650,8	654,4	8,6 %	95,3 %

5 Conclusion

Le présent article décrit la construction d'une nouvelle représentation des poids calés en plusieurs phases qui permet de représenter un estimateur calé en plusieurs phases sous la forme d'un estimateur multivarié calé en une phase. Cette représentation rend possible le calcul d'une approximation sous forme explicite de la variance des estimateurs calés en plusieurs phases pour tout nombre de phases. Une comparaison avec une autre approximation connue dans la littérature pour le cas à deux phases montre que, même si les deux approximations sont convergentes, elles diffèrent en ce qui concerne leurs estimations, leur forme et leur interprétation. Nous avons discuté de certains avantages de la nouvelle approximation dans le cas du calage à deux phases et avons aussi montré sa convergence au moyen d'une étude en simulation du calage à trois phases où elle a donné de très bons résultats pour tous les plans étudiés. L'examen de l'efficacité de l'estimateur proposé en fonction des taux d'échantillonnage et d'autres paramètres du plan fera l'objet de futurs travaux de recherche.

Annexe A

Pour abréger la notation, nous effectuerons notre analyse sous forme matricielle. Nous utiliserons la convention selon laquelle, pour $j > i,$ la sommation dans les produits scalaires $X_i' w_j$ et $X_i' D_j$ (ou avec w_j^* ou $\tilde{w}_j$) est faite sur les unités $k \in s_j$ (et non sur s_i), c'est-à-dire sur l'échantillon indiqué par le dernier ensemble de poids dans le produit scalaire. D'où, $\hat{Z}_{ij} = \left(X_i' D_i^* X_i\right)^{-1} X_i' \left(D_j^* - D_{j-1}^*\right) X_j$ sous cette notation.

Preuve du lemme 3.1. Les poids qui satisfont l'équation de calage à la j^{e} phase avec les poids initiaux $\tilde{w}_{j-1}$ sont donnés par l'équation (3.4). Sous notre notation matricielle

$$\tilde{w}_j = D_j^* \left[g_1 + \ldots + g_j - (j-1)\right]$$

où $g_j = 1 + X_j T_j^{-1} \left(X_j' \tilde{w}_{j-1} - X_j' D_j \tilde{w}_{j-1}\right)$ (voir l'équation (3.5)). Donc

$$\begin{aligned} \tilde{w}_j &= D_{j-1}^* D_j \left[g_1 + \ldots + g_{j-1} - (j-2) + g_j - 1\right] \\ &= D_j \left[\tilde{w}_{j-1} + D_{j-1}^* \left(g_j - 1\right)\right]. \end{aligned} \tag{A.1}$$

L'insertion de g_j donne $\tilde{w}_j = D_j\left[\tilde{w}_{j-1} + D^*_{j-1}X_jT_j^{-1}\left(X_j'\tilde{w}_{j-1} - X_j'D_j\tilde{w}_{j-1}\right)\right]$ qui fait intervenir le poids $\tilde{w}_{j-1}$ provenant de la phase de calage précédente et son produit scalaire avec X_j' et $X_j'D_j$, tandis que les autres multiplicateurs sont des paramètres du plan. L'expression entre crochets contient trois opérandes et donc, après j phases de calage, nous aurions 3^j opérandes qui contiendraient uniquement des paramètres du plan. L'introduction de $\tilde{w}_{j-1}$ provenant de (A.1) par substitution dans $X_j'D_j\tilde{w}_{j-1}$ donne

$$\begin{aligned} X_j'D_j\tilde{w}_{j-1} &= X_j'D_j\left\{D_{j-1}\tilde{w}_{j-2} + D^*_{j-1}\left(g_{j-1} - 1\right)\right\} \\ &= X_j'D_jD_{j-1}\tilde{w}_{j-2} + X_j'D_jD^*_{j-1}X_{j-1}T_{j-1}^{-1}\left(X_{j-1}'\tilde{w}_{j-2} - X_{j-1}'D_{j-1}\tilde{w}_{j-2}\right) \quad \text{(A.2)} \end{aligned}$$

et donc aussi

$$X_j'\tilde{w}_{j-1} = X_j'D_{j-1}\tilde{w}_{j-2} + X_j'D^*_{j-1}X_{j-1}T_{j-1}^{-1}\left(X_{j-1}'\tilde{w}_{j-2} - X_{j-1}'D_{j-1}\tilde{w}_{j-2}\right). \quad \text{(A.3)}$$

La combinaison des termes donne une expression pour $\tilde{w}_j$ qui fait intervenir les poids calés provenant de la phase $j-2$ uniquement

$$\begin{aligned} \tilde{w}_j &= D_jD_{j-1}\tilde{w}_{j-2} \\ &\quad + D^*_jX_{j-1}T_{j-1}^{-1}\left(X_{j-1}'\tilde{w}_{j-2} - X_{j-1}'D_{j-1}\tilde{w}_{j-2}\right) \\ &\quad + D^*_jX_jT_j^{-1}\left(X_j'D_{j-1}\tilde{w}_{j-2} - X_j'D_jD_{j-1}\tilde{w}_{j-2}\right) \\ &\quad - D^*_jX_jT_j^{-1}\hat{Z}'_{j-1,\,j}\left(X_{j-1}'\tilde{w}_{j-2} - X_{j-1}'D_{j-1}\tilde{w}_{j-2}\right). \quad \text{(A.4)} \end{aligned}$$

L'insertion de (A.2) et (A.3) avec $j = p$ dans (A.1) et la récurrence $p-1$ fois sur les groupes de calage respectifs produisent le résultat souhaité.

Annexe B

Un estimateur convergent du total de population dans le calage à trois phases peut être représenté par $\hat{w}_3'y = \hat{Y}_{\mathrm{HT}_3} + \sum_{i=1}^{3}(\hat{t}_1^- - \hat{t}_1^+)'\,\hat{\gamma}_i$, où

$$\begin{aligned} \hat{\gamma}_1 &= \hat{\beta}_1 - \hat{Z}_{12}\hat{\beta}_2 - \hat{Z}_{13}\hat{\beta}_3 + \hat{Z}_{12}\hat{Z}_{23}\hat{\beta}_3 \\ \hat{\gamma}_2 &= \hat{\beta}_2 - \hat{Z}_{23}\hat{\beta}_3 \\ \hat{\gamma}_3 &= \hat{\beta}_3. \end{aligned}$$

Un estimateur convergent de la variance est donné par

$$\begin{aligned} \hat{V}_P\left(\tilde{w}_3'y\right) &= \sum_{k,\,l\in s_1}\left(w^*_{1k}w^*_{1l} - w^*_{1kl}\right)\hat{e}_{1k}\hat{e}_{1l} + \ldots + \sum_{k,\,l\in s_3}\left(w^*_{3k}w^*_{3l} - w^*_{3kl}\right)\hat{e}_{3k}\hat{e}_{3l} \\ &\quad + 2\sum_{k\in s_1,\,l\in s_2} w_{2l}\left(w_{1k}w_{1l} - w_{1kl}\right)\hat{e}_{1k}\hat{e}_{2l} + 2\sum_{k\in s_2,\,l\in s_3} w_{3l}\left(w^*_{2k}w^*_{2l} - w^*_{2kl}\right)\hat{e}_{2k}\hat{e}_{3l} \\ &\quad + 2\sum_{k\in s_1,\,l\in s_3} w_{2l}w_{3l}\left(w^*_{3k}w^*_{3l} - w^*_{3kl}\right)\hat{e}_{1k}\hat{e}_{3l}. \end{aligned}$$

où $\hat{e}_{1k} = x'_{1k}\hat{\gamma}_1 - x'_{2k}\hat{\gamma}_2$, $\hat{e}_{2k} = x'_{2k}\hat{\gamma}_2 - x'_{3k}\hat{\gamma}_3$ et $\hat{e}_{3k} = x'_{3k}\hat{\gamma}_3 - y_k$ sont définis au théorème 3.1.

Bibliographie

Binder, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases : Une approche de type « recette ». *Techniques d'enquête,* 22, 1, 17-22. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1996001/article/14389-fra.pdf.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M. et Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics,* 28, 751-764.

Breidt, J., et Fuller, W.A. (1993). Regression weighting for multiphase samples. *Sankhyā,* 55, 297-309.

Cochran, W.G. (1977). *Sampling Techniques, 3rd Edition*. New-York: John Wiley & Sons, Inc.

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

Farell, P.J., et Singh, S. (2002). Penalized chi-square distance function in survey sampling. *Proceedings of Joint Statistical Meeting,* NY, États-Unis.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica,* 8, 1153-1164.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête,* 24, 1, 11-20. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1998001/article/3905-fra.pdf.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replicate variance estimation after multi-phase stratified sampling. *Journal of American Statistical Association,* 101, 312-320.

Kott, P.S., et Stukel, D.M. (1997). La méthode du jackknife convient-elle à un échantillon à deux phases ? *Techniques d'enquête,* 23, 2, 89-98. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1997002/article/3621-fra.pdf.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika,* 6, 125-133.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika,* 79, 811-822.

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