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

Techniques d’enquête

Le piège de l’inverse de la variance : une correction simple par le ratio pour combiner des données asymétriques

par Anton Grafström, Wilmer Prentius et Åsa Ranlund

Date de diffusion : le 29 juin 2026

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 suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

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.

Techniques d'enquête, juin 2026
Vol. 52, nº 1, pp. 131-143
Statistique Canada, nº 12-001-X au catalogue

Le piège de l'inverse de la variance : une correction simple par le ratio pour combiner des données asymétriques

Anton Grafström, Wilmer Prentius et Åsa Ranlund[1]

Résumé

La combinaison d'estimations provenant d'enquêtes indépendantes au moyen de poids fondés sur l'inverse de la variance peut entraîner un biais négatif lorsque des variances inconnues sont estimées et que la variable cible est non négative et positivement asymétrique. Dans de tels cas, de fortes corrélations positives apparaissent généralement entre les estimateurs et leurs estimateurs de variance correspondants, ce qui fait en sorte que les combinaisons linéaires standard comptant des poids fondés sur l'inverse de la variance présentent un biais négatif. Nous proposons une méthode remarquablement simple pour réduire ce biais : remplacer le poids standard par le ratio entre l'estimateur et l'estimateur de variance. Selon un modèle linéaire reliant les deux, nous montrons que le nouvel estimateur pondéré par le ratio est approximativement sans biais, alors que la combinaison classique fondée sur l'inverse de la variance présente un biais vers le bas. Au moyen de simulations, nous démontrons que la nouvelle méthode rapproche le biais et l'erreur quadratique moyenne de ce qui est optimal pour un large éventail de variables cibles différentes. Comme notre méthode repose uniquement sur des statistiques sommaires couramment publiées, elle peut être adoptée immédiatement pour réduire ce biais largement répandu et améliorer la fiabilité des résultats scientifiques dans divers domaines.

Mots-clés : Combinaison de l'information provenant d'enquêtes; échantillonnage fondé sur le plan de sondage; enquêtes environnementales; estimateur par combinaison linéaire.

1. Introduction

La combinaison de l'information provenant de multiples sources indépendantes est une pratique fondamentale en sciences empiriques. Pour atteindre la plus grande précision possible, les chercheurs synthétisent couramment les résultats de différentes expériences ou enquêtes à l'aide d'une moyenne pondérée. L'approche de référence, enseignée dans les manuels scolaires et intégrée aux logiciels de méta-analyse, est la pondération par l'inverse de la variance, selon laquelle chaque estimation est pondérée par la valeur inverse de sa variance (Cochran, 1954). Les variances réelles des estimateurs étant inconnues pour la plupart (sinon la totalité) des applications pratiques, les estimations de la variance sont souvent utilisées au lieu des variances réelles. Cette méthode est statistiquement optimale selon des hypothèses standard, car elle fournit la combinaison linéaire sans biais la plus précise de l'information disponible. Son utilisation est omniprésente et constitue le fondement de la synthèse des données probantes dans divers domaines.

Toutefois, le caractère optimal de la pondération par l'inverse de la variance repose sur des hypothèses fréquemment non respectées en pratique. Un défi crucial, mais parfois négligé se pose lorsque les variables cibles sont non négatives et présentent une forte asymétrie positive. Cette structure de données n'est pas marginale; elle est au contraire caractéristique d'un vaste éventail de phénomènes. En épidémiologie, les dénombrements de maladies dans une région sont souvent nuls ou faibles, et il y a des foyers ponctuels. En économie, la richesse est fortement asymétrique et habituellement concentrée entre les mains d'une faible proportion de personnes. En écologie, les chercheurs s'intéressent à l'abondance d'espèces rares ou à la

1. Anton Grafström, Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Suède. Courriel : anton.grafstrom@slu.se; Wilmer Prentius, Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Suède. Courriel : wilmer.prentius@slu.se; Åsa Ranlund, Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Suède. Courriel : asa.ranlund@slu.se.

superficie d'habitats, des situations dans lesquelles une grande partie de la base de sondage produit une réponse nulle (Fisher, Corbet et Williams, 1943; Preston, 1948). Dans tous ces cas, les données se caractérisent par une prépondérance de petites valeurs et une queue de grandes valeurs.

Cette caractéristique, comme l'ont fait remarquer Gregoire et Schabenberger (1999) et l'ont présenté plus en détail Grafström, Ekström, Jonsson, Esseen et Ståhl (2019), entraîne une forte corrélation positive entre l'estimateur d'un total et son estimateur de variance correspondant. Une conséquence de cette dépendance est que les estimations plus petites tendent à être associées à des variances estimées plus faibles, ce qui se traduit par des intervalles de confiance plus étroits. Cette asymétrie accroît la probabilité que les intervalles de confiance se situent en dessous de la valeur réelle plus souvent qu'au-dessus (Gregoire et Schabenberger, 1999).

Lorsque des estimations issues de plusieurs enquêtes indépendantes sont combinées à l'aide d'une combinaison linéaire comptant des poids fondés sur l'inverse de la variance, cette corrélation peut introduire un biais systématique. L'enquête pour laquelle le total estimé est plus faible tend à présenter une variance estimée plus petite, elle reçoit donc un poids disproportionnellement grand, ce qui entraîne un estimateur combiné négativement biaisé. Ce problème est particulièrement problématique lorsque la variable d'intérêt est rare ou distribuée de façon inégale, soit précisément les situations où la combinaison des estimations est la plus souhaitable pour améliorer la précision. Ainsi, il existe un besoin manifeste d'avoir un estimateur amélioré qui tient compte de la structure de corrélation et atténue ce biais.

Dans la présente étude, nous supposons que chaque enquête fournit un estimateur sans biais du total, ce qui sous-entend une couverture complète de la population cible dans chaque cas. Lorsque plusieurs bases de sondage sont nécessaires pour assurer une couverture complète, d'autres méthodes doivent être envisagées; voir Elliott, Raghunathan et Schenker (2018) pour obtenir un examen de ces approches.

À titre d'exemple incitatif, considérons l'Inventaire national des forêts (INF) de la Suède, un système fondé sur des échantillons qui fournit des estimations nationales pour plus de 100 variables cibles (Fridman, Holm, Nilsson, Nilsson, Ringvall et Ståhl, 2014). L'INF combine un échantillon permanent, optimisé pour surveiller les tendances temporelles, et un échantillon temporaire, conçu pour estimer la situation actuelle. Pour améliorer l'efficacité, les estimations de ces deux composantes doivent être combinées. Traditionnellement, on l'a fait en utilisant des poids inversement proportionnels aux variances estimées, car la taille des échantillons n'était pas bien définie. Toutefois, pour certaines variables, l'INF à lui seul ne fournit pas suffisamment de renseignements, ce qui nécessite l'intégration de données provenant d'autres enquêtes indépendantes, comme l'Inventaire national des paysages en Suède. Cette intégration est particulièrement cruciale dans le cadre de la production de rapports sur les habitats de l'annexe I de la Suède en vertu de l'article 17 de la directive de l'Union européenne sur les habitats, pour laquelle il est essentiel d'utiliser toutes les sources d'information disponibles.

Pour corriger le biais et l'inefficacité introduits par la pondération classique fondée sur l'inverse de la variance dans de tels contextes, nous proposons une nouvelle stratégie de pondération qui tient explicitement compte de la corrélation positive entre les estimateurs et leurs estimateurs de variance. Cette méthode vise à réduire à la fois le biais et l'erreur quadratique moyenne de l'estimateur combiné. D'autres méthodes de pondération, comme la détermination de poids de variance regroupée, ont été proposées pour atténuer ces

biais. Grafström et coll. (2019) ont dérivé des estimateurs généraux de variance regroupée applicables à n'importe quel plan de sondage et ont présenté le cadre théorique permettant de combiner des échantillons indépendants (regroupement des données) à partir de plans généraux. Toutefois, de telles méthodes sont beaucoup plus restrictives, puisqu'elles exigent une connaissance complète des plans de sondage et des échantillons, ainsi que l'utilisation du même type d'unité d'échantillonnage dans toutes les enquêtes. En méta-analyse, le biais introduit par des données asymétriques a été traité de différentes façons, notamment par la transformation logarithmique des variables (Higgins, White et Anzures-Cabrera, 2008). Or, les transformations introduisent elles-mêmes un biais, nécessitent l'accès aux données complètes et ne sont pas applicables aux plans de sondage généraux.

Dans les présents travaux, nous contribuons à la documentation sur la combinaison de l'information provenant d'enquêtes indépendantes en améliorant l'estimateur classique fondé sur l'inverse de la variance (Cochran, 1954) dans des contextes où il y a des variables positivement asymétriques. En basant notre méthode sur la théorie de l'estimation sans biais, nous proposons une solution robuste et interprétable pour combiner les résultats de multiples enquêtes. Le reste de l'article est structuré comme suit : la section 2 présente la méthode de pondération proposée, la section 3 illustre son application à l'aide d'un exemple et la section 4 conclut par des remarques finales.

2. Méthode

L'objectif est d'estimer un total de population, $Y = \sum_{i \in U} y_i$, d'une variable cible prenant la valeur y_i pour l'unité i dans la population $U = \{1,2,3,\ldots,N\}$ de taille N. Nous supposons que la variable cible est non négative et qu'elle présente une distribution positivement asymétrique, c'est-à-dire qu'elle comporte davantage de petites valeurs que de grandes valeurs. Soit $\hat{Y}_j$, pour $j = 1,\ldots,K$, K estimateurs indépendants sans biais par rapport au plan de Y. Leurs variances sont désignées par $V_j = V(\hat{Y}_j)$ et les estimateurs de variance correspondants par $\hat{V}_j = \hat{V}(\hat{Y}_j)$, et $E(\hat{Y}_j) = V_j$, pour $j = 1,\ldots,K$. Il est bien connu que l'estimateur par combinaison linéaire

$$\hat{Y}_\alpha = \sum_{j=1}^{K} \alpha_j \hat{Y}_j, \quad \text{avec} \quad \sum_{j=1}^{K} \alpha_j = 1, \tag{2.1}$$

est sans biais et

$$\alpha_j = \frac{1/V_j}{\sum_{i=1}^{K} 1/V_i}$$

est optimal en ce sens qu'il possède une variance minimale parmi toutes les combinaisons linéaires possibles; voir Shahar (2017) pour obtenir une démonstration formelle. Dans l'estimateur combiné (2.1), chaque estimateur individuel est pondéré par l'inverse de sa variance, et les poids sont normalisés pour sommer à l'unité. La combinaison naïve, pour laquelle on calcule (2.1) à l'aide d'estimations de variance, $\hat{V}_1,\ldots,\hat{V}_K$, au lieu des variances réelles, est désignée par $\hat{Y}_{\hat{\alpha}}$.

Si $E(\hat{Y}_j) = Y$ pour $j = 1,2,\ldots,K$, et les poids $\hat{\alpha}_j \propto \hat{V}_j^{-1}$ somment à l'unité, il s'ensuit que $E(\hat{Y}_{\hat{\alpha}}) = Y + \sum_{j=1}^{K} \mathrm{Cov}(\hat{\alpha}_j, \hat{Y}_j)$. Par conséquent, une corrélation non nulle entre le poids $\hat{\alpha}_j$ et l'estimateur $\hat{Y}_j$ introduit

un biais. Dans notre situation, où la variable cible est non négative et présente une distribution positivement asymétrique, il découle des résultats de Grafström et coll. (2019) que $\hat{Y}_j$ est approximativement proportionnel à $\hat{V}_j$, ce qui rend $\mathrm{Cov}(\hat{\alpha}_j, \hat{Y}_j)$ négative. L'estimateur $\hat{Y}_{\hat{\alpha}}$ tend alors à être négativement biaisé.

En raison de la proportionnalité approximative entre $\hat{Y}_j$ et $\hat{V}_j$, il est possible de construire théoriquement une correction pour $\hat{V}_j$. Si Y était connu, on obtiendrait un estimateur amélioré de la variance de $\hat{Y}_j$ à l'aide de l'estimateur fondé sur le ratio

$$\hat{V}_{jR} = \frac{Y}{\hat{Y}_j}\hat{V}_j,$$

puisque, par exemple, un faible résultat de $\hat{Y}_j$ donnerait lieu à un faible résultat proportionnellement égal de $\hat{V}_j$ lorsque $\hat{Y}_j$ et $\hat{V}_j$ sont proportionnels. Ces estimateurs fondés sur le ratio des variances V_j corrigent ainsi efficacement les résultats trop faibles ou trop élevés de $\hat{V}_j$ par la multiplication avec le ratio $Y / \hat{Y}_j$. Ces estimateurs de variance théorique sont ensuite insérés à la place des variances réelles lors du calcul des poids, ce qui conduit à de nouveaux poids estimés

$$\hat{\alpha}_{jR} = \frac{1/\hat{V}_{jR}}{\sum_{i=1}^{K} 1/\hat{V}_{iR}} = \frac{\hat{Y}_j/\hat{V}_j}{\sum_{i=1}^{K} \hat{Y}_i/\hat{V}_i}. \tag{2.2}$$

Le paramètre inconnu Y s'annule dans la dérivation des poids que nous proposons. En remplaçant α_j par $\hat{\alpha}_{jR}$ dans l'équation (2.1), on obtient notre estimateur par combinaison linéaire, désigné par $\hat{Y}_{\hat{\alpha}_R}$. Dans cet estimateur, chaque estimateur individuel est pondéré par le ratio entre l'estimateur et son estimateur de variance correspondant. Le caractère optimal de cette approche de pondération est atteint lorsque la proportionnalité entre les estimateurs et leurs estimateurs de variance correspondants est parfaite. Si la proportionnalité n'est qu'approximative, on s'attend à ce que la pondération proposée constitue une bonne approximation de la solution optimale. Étant donné l'hypothèse de forte corrélation entre $\hat{V}_j$ et $\hat{Y}_j$, nous pouvons utiliser le modèle

$$\hat{V}_j = \gamma_j \hat{Y}_j + \varepsilon_j, \tag{2.3}$$

où $\gamma_j > 0$ sont des constantes et $E(\varepsilon_j) = 0$ avec une ε_j indépendante de $\hat{Y}_j$ pour $j = 1, \ldots, K$.

Théorème 1 (réduction du biais)**.** *Soit $\hat{Y}_1, \ldots, \hat{Y}_K$, des estimateurs sans biais d'un total de population Y provenant de K enquêtes indépendantes, et les estimateurs de variance sans biais correspondants sont $\hat{V}_1, \ldots, \hat{V}_K$. On suppose que la variable cible est non négative et présente une distribution positivement asymétrique, de sorte que le modèle (2.3) est valide. Alors, le biais approximatif (de premier ordre) de $\hat{Y}_{\hat{\alpha}}$ est*

$$Biais(\hat{Y}_{\hat{\alpha}}) \approx -\frac{\sum_{j=1}^{K} \frac{\gamma_j}{V_j}\left(\sum_{i \neq j} 1/V_i\right)}{\left(\sum_{j=1}^{K} 1/V_j\right)^2},$$

tandis que le biais approximatif (de premier ordre) de $\hat{Y}_{\hat{\alpha}_R}$ est nul.

Une démonstration du théorème 1 est fournie en annexe. Le théorème 1 indique que l'estimateur pondéré par le ratio $\hat{Y}_{\hat{\alpha}_R}$ présente un biais absolu inférieur à la combinaison naïve $\hat{Y}_{\hat{\alpha}}$ lorsque les estimateurs sont approximativement proportionnels à leurs estimateurs de variance. Le message central du théorème 1 est que, lorsque chaque estimateur $\hat{Y}_j$ d'un total est positivement corrélé à son estimateur de variance $\hat{V}_j$, les poids naïfs fondés sur l'inverse de la variance tendent à attribuer un poids trop élevé à l'enquête, qui se trouve à produire une estimation plus faible (et donc une variance estimée plus faible). En pratique, les variables non négatives et positivement asymétriques (par exemple les dénombrements d'espèces rares ou les superficies d'habitats fragmentés) produisent souvent $\rho_j = \text{Corr}\,(\hat{Y}_j, \hat{V}_j)$ au-dessus de 0,8, ce qui rend $\hat{Y}_{\hat{\alpha}_R}$ préférable à $\hat{Y}_{\hat{\alpha}}$.

L'utilisation de cet estimateur amélioré n'exige que la connaissance des estimations et de leurs variances estimées connexes.

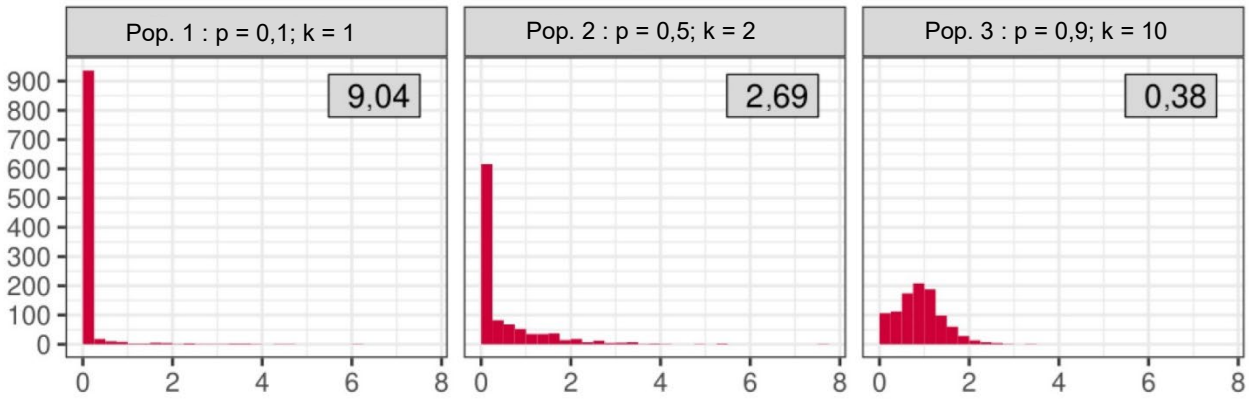
3. Exemple de simulation

Nous fournissons dans la présente section un exemple portant sur le cas de $K = 3$ différents estimateurs pour illustrer le rendement de la nouvelle pondération proposée à la section 2. Trois populations, présentées à la figure 3.1, ont été générées selon

$$Z_i = 1(i \le Np)\frac{X_i}{k}, \quad X_i \sim \chi^2(k),$$

où p désigne la proportion de valeurs non nulles par rapport à la taille totale de la population de $N = 1\,000$ et $k = 1$, 2 et 10.

Figure 3.1 Histogrammes des trois populations différentes utilisées dans l'étude par simulation; p désigne le nombre d'unités non nulles dans la population et k désigne les degrés de liberté de la distribution du khi carré $\chi^2(k)$ (mise à l'échelle pour la moyenne) utilisée pour les unités non nulles



Note : Dans chaque section, le nombre figurant dans le coin supérieur droit indique l'asymétrie de la variable aléatoire $Z^* = YX/k$, où Y désigne une variable aléatoire de Bernoulli indépendante ayant un paramètre p.

Trois plans de sondage différents a, b, c ont été examinés. Les plans de sondage a et b reposaient tous deux sur un échantillonnage aléatoire simple (EAS) sans remise, et les tailles d'échantillon étaient $n_a = 50$

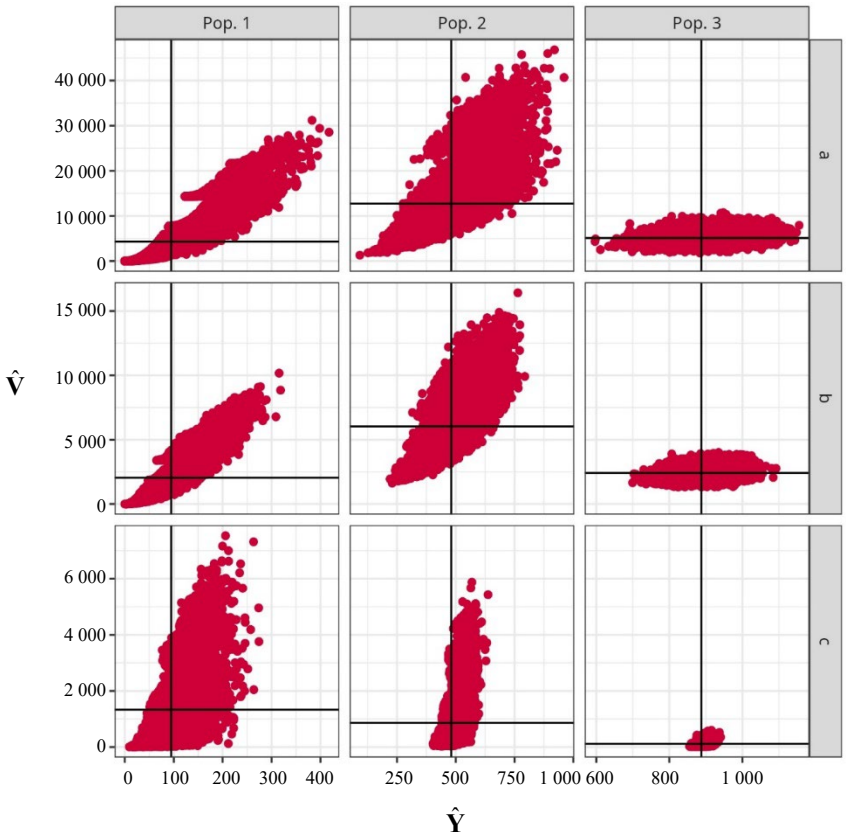
et $n_b = 100$. Pour le plan d'échantillonnage c, la population a d'abord été ordonnée selon la variable cible. La population ordonnée a ensuite été divisée en 10 groupes séquentiels de taille égale (déciles). À partir de chacun de ces déciles, un échantillon aléatoire simple de 10 unités a été tiré. Les variances réelles des estimateurs de Horvitz-Thompson (HT) (Horvitz et Thompson, 1952) des totaux sont présentées au tableau 3.1, et la relation entre les estimations de HT et les estimations de variance correspondantes est illustrée à la figure 3.2.

Tableau 3.1
Variances des estimateurs de HT du total pour les trois plans de sondage a,b,c pour chacune des populations de la figure 3.1

Pop.	a	b	c
1	4 321,05 (69,39 %)	2 046,81 (47,76 %)	1 331,19 (38,52 %)
2	12 730,76 (23,46 %)	6 030,36 (16,15 %)	859,50 (6,10 %)
3	5 101,60 (8,05 %)	2 416,55 (5,54 %)	116,56 (1,22 %)

Notes : - Les écarts-types relatifs sont entre parenthèses.
- Horvitz-Thompson (HT).

Figure 3.2 Relation entre les estimations (axe des x) et les estimations de variance correspondantes (axe des y) pour les trois populations (colonnes) et les plans de sondage (rangées), pour 20 000 échantillons chacune



Note : Les lignes noires présentent les totaux de population et les variances réelles des estimateurs.

Quatre combinaisons différentes ont été prises en considération : la combinaison optimale, reposant sur les variances réelles (V) des estimateurs; une combinaison reposant sur la taille d'échantillon du plan (n); la combinaison naïve, reposant sur des estimations de variance $(\hat{V})$, et l'approche proposée, reposant sur des estimations de variance ajustées par le ratio $(\hat{V}_R)$. Les résultats du tirage de 20 000 échantillons par plan de sondage et par population, ainsi que leur combinaison, sont présentés aux tableaux 3.2, 3.3 et 3.4 pour la racine de l'erreur quadratique moyenne relative, l'écart-type et le biais, respectivement. Il convient de mentionner que pour la population 1, environ 0,4 % des combinaisons linéaires n'ont pas pu être construites en raison d'une estimation de variance nulle obtenue à partir de l'échantillon a. Ces cas ont donc été entièrement exclus des résultats. L'effet probable de l'exclusion des zéros est une légère sous-estimation du biais absolu pour la population 1. La meilleure façon de traiter les estimations nulles lors de l'utilisation de combinaisons linéaires fondées sur des données demeure un problème non résolu qui n'est pas abordé dans la présente étude.

Tableau 3.2
Racine de l'erreur quadratique moyenne relative pour la combinaison d'estimations à l'aide de quatre approches différentes : une combinaison optimale (V), une combinaison reposant sur la taille d'échantillon (n), une combinaison naïve ($\hat{V}$) et une combinaison proposée reposant sur un ratio ($\hat{V}_R$). Résultats par combinaison dans les colonnes pour les plans de sondage (*a*) EAS $n = 50$, (*b*) EAS $n = 100$ et (*c*) stratifié $n = 10 \times 10$

Pop.	Type	ab	ac	bc	abc
1	V	39,13 %	33,81 %	30,18 %	27,59 %
	n	39,14 %	34,66 %	30,85 %	28,19 %
	$\hat{V}$	48,88 %	44,14 %	36,02 %	43,12 %
	$\hat{V}_R$	44,44 %	39,21 %	33,66 %	34,90 %
2	V	13,32 %	5,92 %	5,76 %	5,57 %
	n	13,32 %	8,81 %	8,73 %	8,36 %
	$\hat{V}$	14,10 %	5,91 %	5,77 %	5,65 %
	$\hat{V}_R$	13,47 %	5,85 %	5,72 %	5,52 %
3	V	4,56 %	1,21 %	1,19 %	1,18 %
	n	4,56 %	2,79 %	2,85 %	2,78 %
	$\hat{V}$	4,60 %	1,20 %	1,19 %	1,18 %
	$\hat{V}_R$	4,58 %	1,21 %	1,20 %	1,18 %

Note : Échantillonnage aléatoire simple sans remise (EAS).

Tableau 3.3
Écart-type relatif pour la combinaison d'estimations à l'aide de quatre approches différentes : une combinaison optimale (V), une combinaison reposant sur la taille d'échantillon (n), une combinaison naïve ($\hat{V}$) et une combinaison proposée reposant sur le ratio ($\hat{V}_R$). Résultats par combinaison dans les colonnes pour les plans de sondage (*a*) EAS $n = 50$, (*b*) EAS $n = 100$ et (*c*) stratifié $n = 10 \times 10$

Pop.	Type	ab	ac	bc	abc
1	V	39,13 %	33,81 %	30,18 %	27,59 %
	n	39,14 %	34,66 %	30,85 %	28,19 %
	$\hat{V}$	43,87 %	39,99 %	33,88 %	34,99 %
	$\hat{V}_R$	42,83 %	37,93 %	33,14 %	32,47 %

Note : Échantillonnage aléatoire simple sans remise (EAS).

Tableau 3.3 (suite)
Écart-type relatif pour la combinaison d'estimations à l'aide de quatre approches différentes : une combinaison optimale (V), une combinaison reposant sur la taille d'échantillon (n), une combinaison naïve ($\hat{V}$) et une combinaison proposée reposant sur le ratio ($\hat{V}_R$). Résultats par combinaison dans les colonnes pour les plans de sondage (*a*) EAS $n = 50$, (*b*) EAS $n = 100$ et (*c*) stratifié $n = 10 \times 10$

Pop.	Type	ab	ac	bc	abc
2	V	13,32 %	5,92 %	5,76 %	5,57 %
	n	13,32 %	8,81 %	8,73 %	8,36 %
	$\hat{V}$	13,83 %	5,86 %	5,71 %	5,49 %
	$\hat{V}_R$	13,44 %	5,84 %	5,70 %	5,48 %
3	V	4,56 %	1,21 %	1,19 %	1,18 %
	n	4,56 %	2,79 %	2,85 %	2,78 %
	$\hat{V}$	4,59 %	1,20 %	1,19 %	1,18 %
	$\hat{V}_R$	4,58 %	1,21 %	1,20 %	1,18 %

Note : Échantillonnage aléatoire simple sans remise (EAS).

Tableau 3.4
Biais relatif pour la combinaison d'estimations à l'aide de quatre approches différentes : une combinaison optimale (V), une combinaison reposant sur la taille d'échantillon (n), une combinaison naïve ($\hat{V}$) et une combinaison proposée reposant sur le ratio ($\hat{V}_R$). Résultats par combinaison dans les colonnes pour les plans de sondage (*a*) EAS $n = 50$, (*b*) EAS $n = 100$ et (*c*) stratifié $n = 10 \times 10$

Pop.	Type	ab	ac	bc	abc
1	V	0,16 %	0,05 %	-0,04 %	0,04 %
	n	0,16 %	0,10 %	-0,03 %	0,06 %
	$\hat{V}$	-21,57 %	-18,70 %	-12,23 %	-25,19 %
	$\hat{V}_R$	-11,87 %	-9,92 %	-5,88 %	-12,78 %
2	V	0,02 %	0,02 %	0,01 %	0,01 %
	n	0,02 %	0,03 %	0,01 %	0,02 %
	$\hat{V}$	-2,75 %	-0,79 %	-0,79 %	-1,36 %
	$\hat{V}_R$	-0,99 %	-0,38 %	-0,48 %	-0,73 %
3	V	0,01 %	0,01 %	0,00 %	0,00 %
	n	0,01 %	0,02 %	0,00 %	0,01 %
	$\hat{V}$	-0,15 %	-0,02 %	-0,03 %	-0,04 %
	$\hat{V}_R$	0,06 %	0,00 %	-0,01 %	-0,02 %

Note : Échantillonnage aléatoire simple sans remise (EAS).

Comme prévu théoriquement, il est possible de réduire le biais en utilisant l'estimateur ajusté par le ratio. Même pour les populations qui présentent très peu d'asymétrie, l'estimateur pondéré par le ratio permet de réduire le biais ou, à tout le moins, de ne pas avoir d'effet négatif sur le biais, l'erreur type et l'erreur quadratique moyenne, comparativement à l'approche naïve. Toutefois, l'hypothèse de linéarité sous-jacente à l'estimateur ajusté par le ratio n'étant pas parfaitement respectée, comme l'indique la figure 3.1, la méthode proposée ne permet pas d'éliminer complètement le biais.

Dans la population 1, la pondération par la taille de l'échantillon donne de meilleurs résultats que la pondération par le ratio proposée, laquelle est supérieure à la pondération naïve. Pour les populations 2 et 3, la pondération par le ratio proposée a un meilleur rendement que la pondération par la taille de l'échantillon. Dans de nombreuses applications, la pondération par la taille de l'échantillon n'est pas possible, car les tailles d'échantillon ne sont pas disponibles ou ne peuvent être déterminées. Les résultats du tableau 3.3

mettent également en évidence le problème lié à l'utilisation de la taille de l'échantillon comme poids pour les combinaisons linéaires, indépendamment du fait que la taille de l'échantillon n'est pas toujours disponible dans certains types d'enquêtes, principalement des enquêtes environnementales où les unités d'observation sont échantillonnées indirectement au moyen de parcelles d'échantillonnage. La pondération par la taille de l'échantillon est fortement pénalisée lorsque les efficacités des enquêtes varient considérablement. Lors de la combinaison de plans de sondage présentant des efficacités différentes (effets de plan), illustrés dans ce cas-ci par l'EAS relativement peu efficace b $(n=100)$ et le plan stratifié ordonné très efficace c $(n=100)$, la variance réelle de l'estimateur n'est plus proportionnelle à la taille de l'échantillon et peut donc s'écarter sensiblement de la solution optimale. Cela explique la raison pour laquelle l'approche de pondération par le ratio proposée présente un meilleur rendement pour les populations 2 et 3, où le biais est faible.

4. Remarques finales

Nous avons présenté une nouvelle approche pour déterminer les poids d'une combinaison linéaire de plusieurs estimateurs lorsqu'il faut traiter des variables cibles non négatives et positivement asymétriques. La méthodologie repose sur le fait que, dans de telles conditions, les estimateurs affichent une relation quasi proportionnelle avec leurs estimateurs de variance. Nous avons ensuite démontré que, dans ces conditions, l'estimateur pondéré par le ratio proposé présentait un biais absolu approximatif plus petit que celui de l'estimateur standard fondé sur l'inverse de la variance. Notre exemple de simulation appuie clairement nos conclusions théoriques. De plus, l'estimateur pondéré par le ratio semble très robuste aux écarts par rapport à la quasi-proportionnalité. Dans tous les cas évalués, il a donné de meilleurs résultats que la combinaison naïve. Ainsi, pour un large éventail de variables cibles différentes, des poids fondés sur le ratio peuvent être utilisés pour réduire de façon considérable le biais d'un estimateur par combinaison linéaire.

L'un des principaux avantages de l'approche proposée est qu'elle ne nécessite aucune information supplémentaire et qu'elle s'applique à toute variable cible non négative caractérisée par une distribution positivement asymétrique.

Dans certaines situations, en particulier lors de la combinaison d'estimations issues de plans similaires, le choix des poids en fonction des tailles d'échantillon constitue une option viable pouvant mener à une pondération optimale ou quasi optimale. Toutefois, en présence de plans très différents, de tels poids peuvent être loin d'être optimaux. Non seulement les plans peuvent différer considérablement, mais il en va de même des bases de sondage à partir desquelles on procède à l'échantillonnage et de la distribution de la variable cible. La population sous-jacente peut être définie par une région géographique, qui peut être traitée comme une base de points continue. Dans ce cas, la variable cible peut être définie comme une réponse ponctuelle ou une moyenne locale sur une parcelle de superficie fixe. Des tailles ou des formes de parcelles différentes produisent des réponses différentes, mais mènent aux mêmes totaux de population. Une zone peut également être divisée en différentes populations discrètes (finies) pouvant être échantillonnées à l'aide de méthodes d'échantillonnage de population finie, ce qui produit ainsi des bases de sondage, des unités d'échantillonnage et des distributions de variable cible différentes. Dans le cadre de la surveillance environnementale,

par exemple, il est très courant de rencontrer des plans reposant sur des approches très différentes pour construire la base de sondage à partir de laquelle l'échantillonnage est effectué. Ces plans distincts servent néanmoins à estimer les mêmes quantités. Dans de telles situations, les tailles d'échantillon ne sont même pas comparables et ne devraient pas être utilisées pour dériver des poids. Pour ces raisons, la nouvelle méthode de pondération pourrait être particulièrement avantageuse pour les enquêtes environnementales, dans lesquelles des défis comme la diversité des plans, des bases de sondage, des tailles et des formes de parcelles compliquent la détermination des poids optimaux. Compte tenu de la présence répandue de variables cibles non négatives et positivement asymétriques dans de nombreuses enquêtes et méta-analyses, nous prévoyons que notre stratégie de pondération trouvera une application étendue et contribuera, en fin de compte, à l'obtention d'estimations combinées améliorées et plus précises à partir de multiples enquêtes.

Remerciements

Les auteurs remercient le rédacteur en chef adjoint et les deux examinateurs anonymes pour leurs commentaires et suggestions utiles.

Annexe

A. Démonstration du théorème 1

Biais approximatif de $\hat{Y}_{\hat{\alpha}}$

Le biais de l'estimateur combiné est obtenu au moyen de $\text{Biais}(\hat{Y}_{\hat{\alpha}}) = E(\hat{Y}_{\hat{\alpha}}) - Y$. Étant donné que chaque estimateur est sans biais pour Y et que les poids somment à l'unité, on obtient

$$\begin{aligned}\text{Biais}(\hat{Y}_{\hat{\alpha}}) &= E\left(\sum_{j=1}^{K}\hat{\alpha}_j\hat{Y}_j\right) - Y\sum_{j=1}^{K}E(\hat{\alpha}_j)\\ &= \sum_{j=1}^{K}E(\hat{\alpha}_j\hat{Y}_j) - \sum_{j=1}^{K}E(\hat{\alpha}_j)\,E(\hat{Y}_j) = \sum_{j=1}^{K}\text{Cov}(\hat{\alpha}_j,\hat{Y}_j).\end{aligned}$$

Pour dériver le biais approximatif, nous utilisons une approximation du premier degré de Taylor de $\hat{\alpha}_j$ autour de la référence $(V_1,\ldots,V_K)$. L'approximation est

$$\hat{\alpha}_j \approx \alpha_j + \sum_{i=1}^{K}\frac{\partial\hat{\alpha}_j}{\partial\hat{V}_i}\Big|_{\text{référence}}(\hat{V}_i - V_i).$$

La dérivée partielle de $\alpha_j = (1/V_j)\Big/\left(\sum_{i=1}^{K}1/V_i\right)$ par rapport à V_i est

$$\frac{\partial\alpha_j}{\partial V_i} = \begin{cases} -\dfrac{(1/V_j^2)\left(\sum_{l\neq j}1/V_l\right)}{\left(\sum_{l=1}^{K}1/V_l\right)^2} & \text{si } i = j \\ \dfrac{(1/V_j)\,(1/V_i^2)}{\left(\sum_{l=1}^{K}1/V_l\right)^2} & \text{si } i \neq j \end{cases}$$

Le terme de covariance devient

$$\mathrm{Cov}(\hat{\alpha}_j, \hat{Y}_j) \approx \mathrm{Cov}\left(\sum_{i=1}^{K} \frac{\partial \alpha_j}{\partial V_i} (\hat{V}_i - V_i), \hat{Y}_j \right) = \sum_{i=1}^{K} \frac{\partial \alpha_j}{\partial V_i} \mathrm{Cov}(\hat{V}_i, \hat{Y}_i).$$

Si l'on a K enquêtes indépendantes, les covariances croisées $\mathrm{Cov}(\hat{V}_i, \hat{Y}_j)$ pour $i \neq j$ sont nulles. Ainsi,

$$\mathrm{Cov}(\hat{\alpha}_j, \hat{Y}_j) \approx \frac{\partial \alpha_j}{\partial V_j} \mathrm{Cov}(\hat{V}_j, \hat{Y}_j) = -\frac{(1/V_j^2)\left(\sum_{i \neq j} 1/V_i\right)}{\left(\sum_{i=1}^{K} 1/V_i\right)^2} \mathrm{Cov}(\hat{V}_j, \hat{Y}_j).$$

Cette expression montre que les termes contribuent négativement au biais lorsque $\mathrm{Cov}(\hat{V}_j, \hat{Y}_j) > 0$. Si l'on utilise le modèle (2.3), qui donne $\mathrm{Cov}(\hat{V}_j, \hat{Y}_j) = \gamma_j V(\hat{Y}_j) = \gamma_j V_j$, le biais approximatif total est

$$\mathrm{Biais}(\hat{Y}_{\hat{\alpha}}) \approx \sum_{j=1}^{K} \left(-\frac{(1/V_j^2)\left(\sum_{i \neq j} 1/V_i\right)}{\left(\sum_{i=1}^{K} 1/V_i\right)^2} \gamma_j V_j \right) = -\frac{\sum_{j=1}^{K} \frac{\gamma_j}{V_j} \left(\sum_{i \neq j} 1/V_i\right)}{\left(\sum_{i=1}^{K} 1/V_i\right)^2}.$$

Biais approximatif de $\hat{Y}_{\hat{\alpha}_R}$

On définit $W_j = \hat{Y}_j / \hat{V}_j$ pour $j = 1, \ldots, K$. Les poids sont $\hat{\alpha}_{jR} = W_j \Big/ \sum_{i=1}^{K} W_i$. Le biais est $\mathrm{Biais}(\hat{Y}_{\hat{\alpha}_R}) = \sum_{j=1}^{K} \mathrm{Cov}(\hat{\alpha}_{jR}, \hat{Y}_j)$.

On calcule l'approximation du premier degré de Taylor de $\hat{\alpha}_{jR}$ autour de la référence ($\hat{Y}_i = Y, \hat{V}_i = V_i = \gamma_i Y$) pour toutes les i. Les dérivées partielles de W_i évaluées à la référence sont

$$\frac{\partial W_i}{\partial \hat{Y}_i}\Big|_{\text{référence}} = \frac{1}{V_i}, \qquad \frac{\partial W_i}{\partial \hat{V}_i}\Big|_{\text{référence}} = -\frac{Y}{V_i^2} = -\frac{1}{\gamma_i V_i}.$$

Selon la règle de la dérivation en chaîne, le développement en série de Taylor de $\hat{\alpha}_{k,R}$ peut s'écrire

$$\hat{\alpha}_{jR} \approx \alpha_{jR} + \sum_{i=1}^{K} \left(\frac{\partial \hat{\alpha}_{jR}}{\partial \hat{Y}_i}\Big|_{\text{référence}} \Delta Y_i + \frac{\partial \hat{\alpha}_{jR}}{\partial \hat{V}_i}\Big|_{\text{référence}} \Delta V_i \right),$$

où $\Delta Y_i = \hat{Y}_i - Y$ et $\Delta V_i = \hat{V}_i - V_i$. Les dérivées partielles sont

$$\frac{\partial \hat{\alpha}_{jR}}{\partial \hat{Y}_i}\Big|_{\text{référence}} = \frac{\partial \hat{\alpha}_{jR}}{\partial W_i}\Big|_{\text{référence}} \frac{1}{V_i}, \qquad \frac{\partial \hat{\alpha}_{jR}}{\partial \hat{V}_i}\Big|_{\text{référence}} = \frac{\partial \hat{\alpha}_{jR}}{\partial W_i}\Big|_{\text{référence}} \frac{-1}{\gamma_i V_i}.$$

Selon l'hypothèse du modèle (2.3), on a $\Delta V_i = \gamma_i \Delta Y_i + \varepsilon_i$. En substituant cela dans le développement, les termes concernant ΔY_i s'annulent et, après quelques manipulations algébriques, on obtient

$$\frac{\partial \hat{\alpha}_{jR}}{\partial \hat{Y}_i}\Big|_{\text{référence}} \Delta Y_i + \frac{\partial \hat{\alpha}_{jR}}{\partial \hat{V}_i}\Big|_{\text{référence}} \Delta V_i = \frac{\partial \hat{\alpha}_{jR}}{\partial \hat{V}_i}\Big|_{\text{référence}} \varepsilon_i.$$

L'expression pour $\hat{\alpha}_{jR}$ se simplifie pour devenir

$$\hat{\alpha}_{jR} \approx \alpha_{jR} + \sum_{i=1}^{K} \frac{\partial \hat{\alpha}_{jR}}{\partial \hat{V}_i}\Big|_{\text{référence}} \varepsilon_i.$$

Comme ε_i pour $i = 1,\ldots,K$ sont (censées être) indépendantes de $\hat{Y}_j$, le terme de covariance approximative est

$$\mathrm{Cov}(\hat{\alpha}_{jR}, \hat{Y}_j) \approx \mathrm{Cov}\left(\sum_{i=1}^{K} \frac{\partial \hat{\alpha}_{jR}}{\partial \hat{V}_i} \Big|_{\text{référence}} \varepsilon_i, \hat{Y}_j \right) = \sum_{i=1}^{K} \frac{\partial \hat{\alpha}_{jR}}{\partial \hat{V}_i} \Big|_{\text{référence}} \mathrm{Cov}(\varepsilon_i, \hat{Y}_j) = 0.$$

Ainsi, le biais approximatif (de premier ordre) est nul.

$$\mathrm{Biais}(\hat{Y}_{\hat{\alpha}_R}) = \sum_{j=1}^{K} \mathrm{Cov}(\hat{\alpha}_{jR}, \hat{Y}_j) \approx 0.$$

Bibliographie

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10(1), 101-129.

Elliott, M.R., Raghunathan, T.E. et Schenker, N. (2018). Combining estimates from multiple surveys. Wiley StatsRef: Statistics Reference Online, 1-10.

Fisher, R.A., Corbet, A.S. et Williams, C.B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 42-58.

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A.H. et Ståhl, G. (2014). Adapting National Forest Inventories to changing requirements – The case of the Swedish National Forest Inventory at the turn of the 20th century. *Silva Fennica*, 48(3).

Grafström, A., Ekström, M., Jonsson, B.G., Esseen, P.-A. et Ståhl, G. (2019). Combinaison d'échantillons probabilistes indépendants. *Techniques d'enquête*, 45(2), 371-387. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019002/article/00003-fra.pdf.

Gregoire, T.G. et Schabenberger, O. (1999). Sampling skewed biological populations: Behavior of confidence intervals for the population total. *Ecology*, 80(3), 1056-1065.

Higgins, J.P., White, I.R. et Anzures-Cabrera, J. (2008). Meta-analysis of skewed data: Combining results reported on log-transformed or raw scales. *Statistics in Medicine*, 27(29), 6072-6092.

Horvitz, D.G. et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.

Preston, F.W. (1948). The commonness, and rarity, of species. *Ecology*, 29(3), 254-283.

Shahar, D.J. (2017). Minimizing the variance of a weighted average. *Open Journal of Statistics*, 7(2), 216-224.