

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Comparaison de techniques récentes pour la combinaison d'échantillons probabilistes et non probabilistes

par Julie Gershunskaya, Vladislav Beresovsky et Terrance D. Savitsky

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Comparaison de techniques récentes pour la combinaison d'échantillons probabilistes et non probabilistes

Julie Gershunskaya, Vladislav Beresovsky et Terrance D. Savitsky¹

Résumé

Plusieurs méthodes récentes de quasi-randomisation pour obtenir des inférences à partir d'échantillons non probabilistes sont comparées. Les techniques prises en compte sont élaborées en supposant que la sélection de l'échantillon est régie par un mécanisme aléatoire latent sous-jacent, et qu'elle peut être décelée en combinant des données d'enquête non probabilistes à un échantillon probabiliste de « référence », obtenu à partir de la même population cible. Des processus de rechange sont mis au point pour les raisons suivantes : i) les indicateurs de participation à l'échantillon non probabiliste ne sont disponibles que pour les unités d'échantillonnage observées, et ii) on ne sait généralement pas quelles unités de la population sous-jacente appartiennent à la fois aux échantillons non probabilistes et de référence. La façon dont différents processus permettent de surmonter ces difficultés est considérée, des propriétés théoriques des méthodes sont discutées et on les compare à l'aide de simulations.

Mots clés : Combinaison de données; échantillon non probabiliste; échantillon de référence; probabilités de participation; vraisemblance de l'échantillon; estimation de la variance.

1. Introduction

La prolifération récente d'ordinateurs très puissants et d'Internet s'est traduite par de nouvelles façons de recueillir des renseignements. Il est souvent possible d'obtenir une mine de données rapidement, sans disposer d'un échantillon probabiliste bien conçu. Toutefois, des estimations naïves tirées de telles données peuvent présenter un biais substantiel, même lorsque l'échantillon est grand. À partir d'estimations et d'inférences appropriées fondées sur des données non probabilistes (de commodité), les échantillons nécessitent des hypothèses de modélisation tenant compte du mécanisme de participation à l'échantillon. La méthodologie existante peut être largement regroupée en approches fondées sur un modèle, quasi randomisées et doublement robustes. Elliott et Valliant (2017), Valliant (2020), Beaumont et Rao (2021), Rao (2021), Wu (2022) fournissent un compte rendu complet de ce secteur en pleine croissance.

Dans la présente étude, nous examinons différentes méthodes récentes de quasi-randomisation. Les techniques envisagées sont élaborées en présumant que la sélection de l'échantillon est régie par un mécanisme aléatoire latent sous-jacent et qu'elle peut être décelée en combinant des données d'enquête non probabilistes à un échantillon probabiliste, obtenu à partir de la même population cible. Dans ce contexte, l'échantillon probabiliste est souvent appelé échantillon de « référence ». L'exigence fondamentale est que les échantillons non probabilistes et de référence contiennent un ensemble commun de covariables et l'hypothèse est que les probabilités de participation à l'échantillon non probabiliste sont régies par ces covariables et pourraient être estimées à partir d'un modèle.

Les méthodes récentes de quasi-randomisation disponibles dans la documentation peuvent être largement classées en méthodes reposant sur la pseudo-vraisemblance et sur l'échantillonnage. Beresovsky, Gershunskaya et Savitsky (2024) ont démontré que dans des situations pratiques importantes, où l'échantillon non probabiliste est grand et l'échantillon probabiliste est relativement petit, des méthodes par échantillonnage pourraient être substantiellement plus efficaces que les méthodes fondées sur la pseudo-vraisemblance.

¹Julie Gershunskaya, U.S. Bureau of Labor Statistics, Washington, DC 20212 (gershunskaya.julie@bls.gov); Vladislav Beresovsky, U.S. Bureau of Labor Statistics, Washington, DC 20212 (beresovsky.vladislav@bls.gov); Terrance D. Savitsky, U.S. Bureau of Labor Statistics, Washington, DC 20212 (savitsky.terrance@bls.gov)

Chu et Beaumont (2019) ainsi que Beaumont et coll. (2024) ont examiné l'utilisation d'arbres de classification dans le contexte de la combinaison d'échantillons probabilistes et non probabilistes. Ils ont utilisé la méthode de pseudo-vraisemblance de Chen, Li et Wu (2020) pour construire une fonction objective permettant de déterminer la répartition optimale et d'estimer les probabilités de participation à l'échantillon non probabiliste. Nous proposons d'utiliser une vraisemblance fondée sur l'échantillon plutôt que la pseudo-vraisemblance comme solution potentiellement plus efficace. Nous appliquons l'algorithme proposé dans une petite étude de simulation pour démontrer l'avantage de la méthode fondée sur l'échantillon par rapport à la pseudo-vraisemblance dans le cas d'un petit échantillon de référence.

Nous commençons par introduire une configuration et formuler des hypothèses de base dans la deuxième partie. Dans la troisième partie, nous examinons brièvement d'autres méthodes et discutons de leurs propriétés asymptotiques. L'application des méthodes dans un algorithme d'arbres de classification est décrite dans la quatrième partie. Les résultats de la simulation sont présentés dans la cinquième partie. La sixième partie propose un résumé.

2. Configuration et notation

Supposons que nous cherchons à estimer la moyenne de population finie $\mu = N^{-1} \sum_{i \in U} y_i$ pour les unités de la population cible U de taille N en fonction de l'échantillon non probabiliste disponible S_c de taille n_c . L'échantillon S_c contient les variables d'intérêt y_i et un ensemble de covariables \mathbf{x}_i , $i = 1, \dots, N$.

Soit I_{ci} , l'indicateur d'inclusion de l'unité de population i , prenant la valeur de 1 si l'unité i est incluse dans S_c , et 0 si elle est dans $U \setminus S_c$. Nous présumons l'existence d'un mécanisme de sélection aléatoire latent sous-jacent et définissons des probabilités *inconnues* $\pi_{ci} = \pi_c(\mathbf{x}_i) = P\{I_{ci} = 1 \mid i \in U, \mathbf{x}_i\}$ pour les unités de population U à inclure dans l'échantillon non probabiliste S_c . Les probabilités d'inclusion π_{ci} sont également souvent appelées probabilités de « participation ».

Supposons que l'échantillon probabiliste S_r de taille n_r est également disponible pour la même population U et que le même ensemble de covariables \mathbf{x}_i observé dans S_c est également disponible pour l'échantillon S_r . L'échantillon probabiliste S_r est sélectionné à l'aide d'un plan d'enquête probabiliste *connu* avec les probabilités d'inclusion $\pi_{ri} = P\{I_{ri} = 1 \mid i \in U\}$, où I_{ri} est l'indicateur d'inclusion de l'échantillon probabiliste de l'unité i .

Nous présumons ce qui suit au sujet du processus de sélection :

H1 : Chaque unité de la population a une probabilité de participation positive : $\pi_{ci} > 0$ pour tout $i \in U$.

H2 : Le mécanisme de sélection de l'échantillon S_c peut être ignoré étant donné que \mathbf{x}_i : $P\{I_{ci} = 1 \mid i \in U, \mathbf{x}_i, y_i\} = P\{I_{ci} = 1 \mid i \in U, \mathbf{x}_i\}$ pour tous $i \in U$.

H3 : Les indicateurs I_{ci} et I_{cj} sont indépendants avec \mathbf{x}_i et \mathbf{x}_j pour $i \neq j$.

H4 : Les indicateurs I_{ci} et I_{ri} sont indépendants avec \mathbf{x}_i pour tout $i \in U$.

L'hypothèse H1 énonce que chaque unité de la population cible a une chance non nulle de faire partie de l'enquête non probabiliste; il s'agit de l'analogie de la couverture de la base de sondage complète dans le cas de l'échantillon probabiliste. L'hypothèse H2 est l'exigence relative aux données manquantes au hasard (MAR) énonçant qu'après

prise en compte des covariables \mathbf{x}_i , le mécanisme de participation à l'échantillon non probabiliste est ignorable pour la variable à l'étude y_i . L'hypothèse d'indépendance mutuelle H3 pour les indicateurs I_c nous permet de formuler une vraisemblance pour les indicateurs en tant que variables indépendantes.

Une fois les estimations des probabilités de participation à l'échantillon π_{ci} obtenues, nous utilisons leurs valeurs inversées sous la forme habituelle de l'estimateur de Hajek de la moyenne de la population cible :

$$\hat{\mu} = \left(\sum_{i \in S_c} 1 / \hat{\pi}_{ci} \right)^{-1} \sum_{i \in S_c} y_i / \hat{\pi}_{ci},$$

, estimateur dit pondéré selon la propension inverse (IPW).

3. Bref examen des méthodes disponibles

Chen, Li et Wu (2020) ont examiné une vraisemblance de Bernoulli par rapport aux unités de population finies. Ils présentent le logarithme du rapport de vraisemblance comme la somme de deux termes, où le premier terme est un total sur l'échantillon non probabiliste observé et le deuxième, un total de la population finie. Comme la population finie n'est pas disponible, une formulation de *pseudo-vraisemblance* est utilisée, selon laquelle le total de population finie est remplacé par son estimation pondérée pour l'enquête probabiliste :

$$\hat{\ell}^{CLW}(\boldsymbol{\beta}) = \sum_{i \in S_c} \log \left[\frac{\pi_{ci}(\boldsymbol{\beta})}{1 - \pi_{ci}(\boldsymbol{\beta})} \right] + \sum_{i \in S_r} w_{ri} \log [1 - \pi_{ci}(\boldsymbol{\beta})],$$
(1)

où $\boldsymbol{\beta}$ est le vecteur du paramètre dans un modèle de régression logistique $\text{logit}[\pi_{ci}(\boldsymbol{\beta})] = \boldsymbol{\beta}^T \mathbf{x}_i$ et $w_{ri} = \pi_{ri}^{-1}$. Les estimations sont obtenues en résolvant des équations respectives d'estimation fondées sur la pseudo-vraisemblance.

Nous désignons la méthode de pseudo-vraisemblance de Chen, Li et Wu (2020) comme étant « directe », parce que la vraisemblance de Bernoulli formulée pour les indicateurs d'échantillons non probabilistes, I_{ci} , mène directement à l'estimation des probabilités de participation à l'échantillon non probabiliste d'intérêt, π_{ci} . En revanche, plusieurs méthodes sont prises en compte dans la documentation selon lesquelles les vraisemblances de Bernoulli sont formulées pour différentes variables indicatrices liées aux paramètres d'intérêt, en particulier de façons « indirectes ». Ces méthodes comprennent une méthode de pseudo-vraisemblance de Wang, Valliant et Li (2021) et des méthodes fondées sur l'échantillon d'Elliot (2009) et de Savitsky et coll. (2023). Nous appelons ce groupe de techniques les méthodes « indirectes ».

Elliot (2009) propose de tenir compte de la variable indicatrice I_{zi} (dans notre notation) définie sur l'ensemble d'échantillons probabilistes et non probabilistes combinés $S_c \cup S_r$, où $I_{zi} = 1$ pour $i \in S_c$, et $I_{zi} = 0$ pour les unités qui *n'appartiennent pas* à l'échantillon non probabiliste, $i \in (S_c \cup S_r)$, S_c . Il y a un risque non nul que des unités de la population finie puissent être tirées à la fois des échantillons non probabilistes et de référence, $S_c \cap S_r$. Si l'identité des unités appartenant au chevauchement d'échantillonnage est connue, cela est facile à montrer (voir, par exemple, Beresovsky [2019], Beaumont et coll. (2024), Beresovsky, Gershunskaya et Savitsky [2024]) que la relation suivante est valide, reliant la probabilité d'être présent dans la partie de l'échantillon non probabiliste de l'ensemble combiné, $\pi_{zi} = P\{i \in S_c \mid i \in S_c \cup S_r\}$, et les probabilités d'inclusion respectives des échantillons, π_{ci} et π_{ri} :

$$\pi_{zi} = \frac{\pi_{ci}}{\pi_{ci} + \pi_{ri} - \pi_{ci}\pi_{ri}}.$$
(2)

Le défi est qu'on ne sait généralement pas quelles unités de l'ensemble combiné $S_c \cup S_r$ appartiennent aux deux échantillons S_c et S_r . Elliot (2009) suppose que la probabilité d'appartenance au chevauchement d'échantillonnage est *négligeable* pour toutes les unités de la population finie. Selon cette hypothèse, une relation *approximative* est maintenue :

$$\pi_{zi} = \frac{\pi_{ci}}{\pi_{ci} + \pi_{ri}}. \quad (3)$$

Toutefois, l'exigence de chevauchement négligeable de la méthode d'Elliot (2009) a mené à la croyance répandue selon laquelle la méthode fondée sur des échantillons combinés se limitait aux cas où les fractions de sondage étaient faibles. C'est malheureux, car dans de nombreuses situations pratiques, l'échantillon non probabiliste est grand et le chevauchement de l'échantillonnage pourrait être significatif.

Savitsky et coll. (2023) ont montré que, selon une configuration légèrement différente, la relation (3) *se maintient exactement*, peu importe la taille des échantillons; au lieu de l'union de deux échantillons, les auteurs considèrent une *pile* d'échantillons, $S = S_c + S_r$, et modifient en conséquence la définition de l'indicateur I_{zi} comme prenant la valeur de 1 pour toutes les unités d'un échantillon non probabiliste S_c , et de 0 pour toutes les unités de l'échantillon de référence S_r ; les unités appartenant aux deux échantillons seraient incluses deux fois dans l'échantillon empilé : une fois avec $I_{zi} = 1$ et une fois avec $I_{zi} = 0$ (il n'est pas nécessaire de connaître l'identité de ces unités).

Beresovsky, Gershunskaya et Savitsky (2024) ont prouvé que les indicateurs I_{zi}, I_{zj} pour les unités i et j de l'ensemble empilé étaient *presque* indépendants, avec une corrélation aussi faible que $O(N^{-1})$, quelle que soit l'ampleur du chevauchement; cela motive donc l'utilisation de la vraisemblance de Bernoulli pour les variables aléatoires indépendantes sur l'ensemble empilé.

Wang, Valliant et Li (2021) ont proposé une méthode de pondération de la propension logistique ajustée (ALP) dans le cadre de laquelle ils utilisent une construction imaginaire empilant l'échantillon non probabiliste (partie 1) et la population finie cible (partie 2). Ils ont formulé une vraisemblance de Bernoulli pour une variable indicatrice définie sur l'ensemble empilé proposé sous la forme $\tilde{I}_{zi} = 1$ pour les unités de la partie 1 et $\tilde{I}_{zi} = 0$ pour les unités de la partie 2, et ont utilisé une méthode de pseudo-vraisemblance en remplaçant la partie comprenant les unités de population finies par son estimation pondérée d'enquête fondée sur la probabilité :

$$\hat{\ell}^{ALP}(\gamma) = \sum_{i \in S_c} \log[\tilde{\pi}_{zi}(\gamma)] + \sum_{i \in S_r} w_{ri} \log[1 - \tilde{\pi}_{zi}(\gamma)], \quad (4)$$

où γ est le vecteur du paramètre dans un modèle de régression logistique $\text{logit}[\tilde{\pi}_{zi}(\gamma)] = \gamma^T \mathbf{x}_i$.

La relation entre la probabilité d'être présent dans la partie 1 de la construction empilée et les probabilités de faire partie de l'échantillon non probabiliste est

$$\tilde{\pi}_{zi} = \frac{\pi_{ci}}{\pi_{ci} + 1}, \quad (5)$$

qui est un cas particulier de (3), lorsque l'ensemble de la population finie est disponible au lieu d'un échantillon seulement.

Comme cela a été mentionné, les méthodes reposant sur les identités respectives (2), (3) ou (5) sont « indirectes » en ce sens que les paramètres de Bernoulli respectifs sont des probabilités de présence dans une partie particulière de l'ensemble combiné, plutôt que des paramètres d'intérêt π_{ci} .

Beresovsky (2019) fait remarquer que l'estimation de π_{ci} peut être effectuée en une seule étape si π_{zi} sont considérés comme une fonction composite $\pi_{zi}(\pi_{ci}(\boldsymbol{\beta})) = \pi_{ci}(\boldsymbol{\beta}) / (\pi_{ci}(\boldsymbol{\beta}) + \pi_{ri})$ dans le logarithme du rapport de vraisemblance

$$\ell^{ILR}(\boldsymbol{\beta}) = \sum_{i \in S_c} \log[\pi_{zi}(\pi_{ci}(\boldsymbol{\beta}))] + \sum_{i \in S_r} \log[1 - \pi_{zi}(\pi_{ci}(\boldsymbol{\beta}))] \quad (6)$$

et est appelée la méthode de « régression logistique implicite » (ILR) (voir aussi Savitsky et coll. [2023] et Beresovsky, Gershunskaya et Savitsky [2024]).

En revanche, Elliott (2009) et Wang, Valliant et Li (2021) appliquent un processus en deux étapes : à l'étape 1, ils trouvent les estimations des paramètres du modèle de régression logistique standard; à l'étape 2, ils utilisent les identités respectives pour obtenir des estimations des paramètres d'intérêt, π_{ci} . Toutefois, une telle démarche en deux étapes peut entraîner une perte d'efficacité et des estimations de π_{ci} se retrouvant hors de l'espace des paramètres.

Beresovsky, Gershunskaya et Savitsky (2024) ont comparé les propriétés asymptotiques des estimations du maximum de vraisemblance en se fondant sur : 1) la méthode de pseudo-vraisemblance directe de Chen, Li et Wu (2020) (CLW), 2) la modification en une étape de la pseudo-vraisemblance « indirecte » de Wang, Valliant et Li (2021) (pseudo-ILR, ou PILR) et 3) la méthode « indirecte » fondée sur des échantillons empilés (ILR).

Ils ont constaté que les trois méthodes étaient asymptotiquement équivalentes. Cependant, il semble que la méthode ILR soit moins sensible à la taille de l'échantillon probabiliste, par rapport aux méthodes CLW ou PILR fondées sur une pseudo-vraisemblance.

4. Arbres de classification

Chu et Beaumont (2019) ainsi que Beaumont et coll. (2024) ont adapté l'algorithme des arbres de classification et de régression (CRAC) dans le contexte de la combinaison d'échantillons probabilistes et non probabilistes. Ils utilisent la pseudo-vraisemblance CLW comme fonction objective pour estimer les probabilités dans les branches gauche et droite d'un fractionnement donné et pour trouver des fractionnements optimaux. Nous proposons d'utiliser à cette fin la vraisemblance reposant sur les échantillons empilés.

L'algorithme d'arbres de classification répartit les données en groupes homogènes $g = 1, \dots, G$, de sorte que toutes les unités d'un groupe donné ont les mêmes probabilités π_{cg} . Cela fonctionne de façon récursive en trouvant un fractionnement binaire optimal à une étape donnée pour une covariable donnée, comme suit :

- 1) calculer les probabilités de participation dans les branches gauche et droite, π_{cgL} et π_{cgR} , sur une grille de fractionnements binaires possibles;
- 2) choisir un fractionnement optimal reposant sur une fonction objective.

Selon la pseudo-vraisemblance CLW, la fonction objective est

$$\hat{I}^{CLW} = - \sum_{g=1}^G \frac{\hat{N}_g}{\hat{N}} \left[\hat{\pi}_{cg} \log(\hat{\pi}_{cg}) + (1 - \hat{\pi}_{cg}) \log(1 - \hat{\pi}_{cg}) \right] \quad (7)$$

et l'estimation de la probabilité de participation pour le groupe homogène g est

$$\hat{\pi}_{cg} = \frac{n_{cg}}{\hat{N}_g}, \text{ where } \hat{N}_g = \sum_{i \in g} w_{ri}. \quad (8)$$

Pour la méthode fondée sur l'échantillon (ILR), la fonction objective est

$$\hat{I}^{ILR} = - \sum_{g=1}^G \left[\sum_{i \in S_{cg}} \log \hat{\pi}_{zi,g} + \sum_{i \in S_{rg}} \log(1 - \hat{\pi}_{zi,g}) \right], \quad (9)$$

où $\hat{\pi}_{zi,g} = \hat{\pi}_{cg} / (\pi_{ri} + \hat{\pi}_{cg})$.

La méthode IRL ne fournit pas d'expression explicite pour l'estimation de π_{cg} . On peut la trouver comme solution à l'équation

$$\sum_{i \in S_g} \frac{\pi_{ri}}{\pi_{ri} + \pi_{cg}} = n_{rg}, \quad (10)$$

Où S_g est la partie de l'échantillon empilé S appartenant au groupe g .

Il convient de souligner que, même si $\hat{\pi}_{cg}$ est une constante dans un groupe homogène donné, les estimations $\hat{\pi}_{zi,g}$ dans un groupe donné peuvent varier.

(Comme cela est mentionné dans Beaumont et coll. [2024], les estimations des probabilités de participation $\hat{\pi}_{cg}$ pour les groupes homogènes fondés sur la méthode PILR sont les mêmes que celles fondées sur CLW fournies par (8), même si les fonctions objectives fondées sur les vraisemblances de CLW et PILR diffèrent. Nous n'incluons pas la méthode PILR dans l'étude de simulation de la cinquième partie.)

5. Simulations

Nous avons effectué une petite étude de simulation, dans le cadre de laquelle nous avons supposé disposer déjà d'un arbre mature, jusqu'à un certain niveau g . Nous nous concentrons sur le fractionnement du nœud g en deux parties.

La configuration de la simulation est la suivante. Soit $N_g = 1,000$, la taille de la population correspondant au nœud g . Nous générons une seule covariable x_{ig} à partir de la distribution normale standard : $x_{ig} \sim N(0,1)$. Soit $y_{ig} = 1 + x_{ig} + \dot{\alpha}_{ig}$, la variable de l'étude avec $\dot{\alpha}_{ig} \sim N(0,1.5^2)$.

Nous avons établi des probabilités réelles de participation à un échantillon non probabiliste à $\pi_{cg,L} = 0.80$ pour les unités i avec $x_{ig} \leq 0$ et $\pi_{cg,R} = 0.20$ pour les unités i avec $x_{ig} > 0$. Cette configuration fournit l'échantillon non probabiliste d'environ $n_{cg} \approx 500$ unités.

Nous utilisons le plan de sondage fondé sur la probabilité proportionnelle à la taille pour générer l'échantillon probabiliste, pour lequel les probabilités d'inclusion $\pi_{rg,i}$ sont proportionnelles à x_{ig} . Nous examinons deux scénarios de simulation : 1) un plus petit échantillon de taille $n_{rg} = 100$ et 2) un plus grand échantillon de taille $n_{rg} = 500$.

Les résultats après 1 000 exécutions de simulation sont présentés aux figures 5-1 et 5-2.

Au tableau 5-1, nous présentons les résultats des estimations correspondantes de la moyenne de population μ .

Figure 5-1

Estimations des probabilités de la « branche gauche », pour des $\pi_{cg,R} = 0.80$ réels, pour des scénarios de taille d'échantillon plus petite ($n_{rg} = 100$) et plus grande ($n_{rg} = 500$)

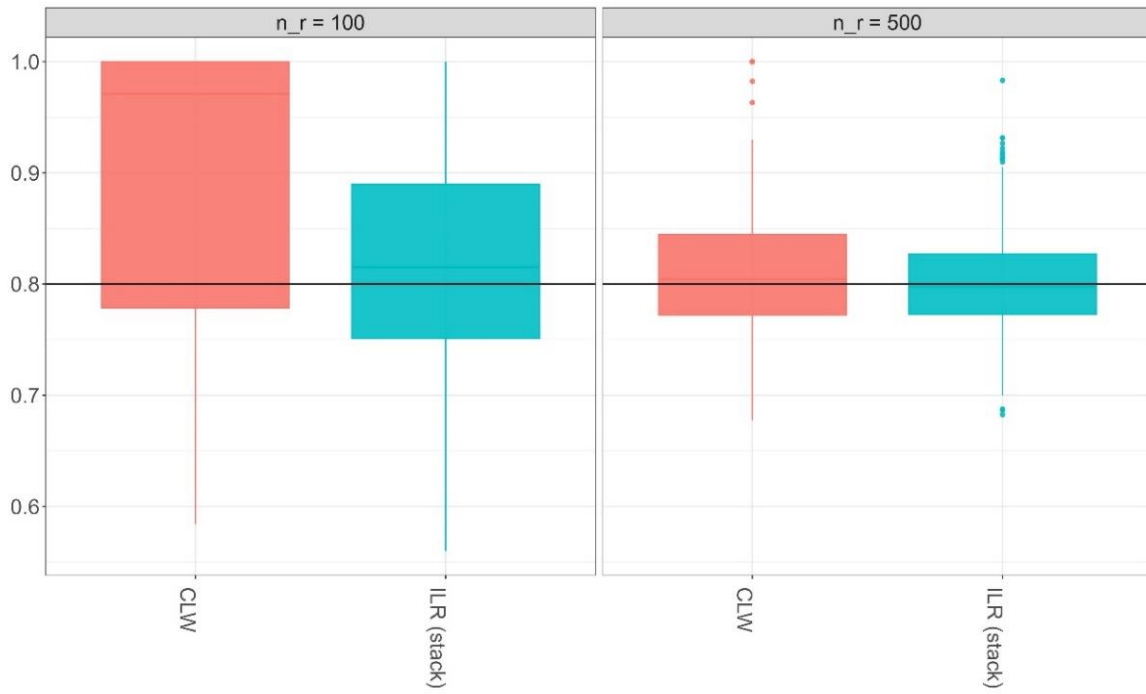


Figure 5-2

Estimations des probabilités de la « branche droite », pour des $\pi_{cg,L} = 0.20$ réels, pour des scénarios de taille d'échantillon plus petite ($n_{rg} = 100$) et plus grande ($n_{rg} = 500$)

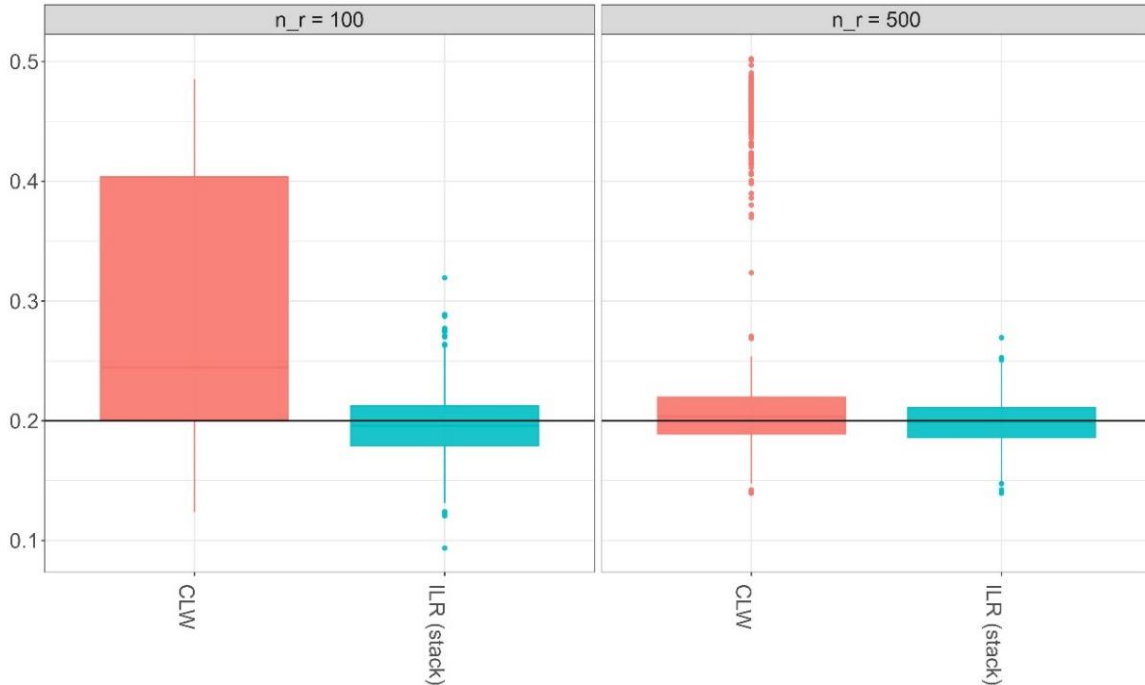


Tableau 5-1

Biais et EQM des estimations de la moyenne de la population, à l'aide d'un estimateur non pondéré et de facteurs de pondération fondés sur les méthodes CLW et ILR

	$n_r = 100$		$n_r = 500$	
	Biais	EQM	Biais	EQM
Non pondéré	-0,481	0,234	-0,481	0,234
CLW	-0,128	0,042	-0,047	0,023
ILR	0,012	0,011	0,001	0,007

6. Résumé

Des recherches antérieures montrent que la méthode fondée sur des échantillons empilés donne des estimations plus efficaces des probabilités de participation comparativement aux méthodes fondées sur la pseudo-vraisemblance dans de nombreux scénarios pratiques, surtout lorsque l'échantillon fondé sur la probabilité est petit et que le chevauchement des domaines définis par les covariables est faible entre les deux échantillons. Par conséquent, nous nous attendions à ce que l'utilisation de cette méthode conjointement avec l'algorithme d'arbres de régression puisse être particulièrement bénéfique. Nos simulations semblent confirmer cette hypothèse.

Bibliographie

Beaumont, J.-F. et J.N.K. Rao. (2021), « Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies », *The Survey Statistician*, 83, p. 11-22.

- Beaumont, J.-F., K. Bosa, A. Brennan, J. Charlebois et K. Chu. (2024), « Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada », *Techniques d'enquête*, 50, p. 77-106.
- Beresovsky, V. (2019), « On application of a response propensity model to estimation from web samples », dans ResearchGate.
- Beresovsky, V., J. Gershunskaya et T.D. Savitsky. (2024), « Review of Quasi-Randomization Approaches for Estimation from Non-probability Samples », <https://arxiv.org/abs/2312.05383>.
- Chen, Y., P. Li et C. Wu. (2020), « Doubly Robust Inference With Nonprobability Survey Samples », *Journal of the American Statistical Association*, 115, p. 2011-2021.
- Chu, K. et J.F. Beaumont. (2019), « The use of classification trees to reduce selection bias for a nonprobability sample with help from a probability sample », dans *Proceedings of the Survey Methods Section, Statistical Society of Canada*.
- Elliott, M.R. (2009), « Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights », *Survey Practice*, 2, p. 813-845.
- Elliott, M.R. et R. Valliant. (2017), « Inference for Nonprobability Samples », *Statistical Science*, 32, p. 249-264.
- Rao, J.N.K. (2021), « On Making Valid Inferences by Integrating Data from Surveys and Other Sources », *Sankhya*, 83, p. 242-272.
- Savitsky, T.D., M.R. Williams, J. Gershunskaya et V. Beresovsky. (2023), « Methods for combining probability and nonprobability samples under unknown overlapps », *Statistics in Transition, New Series*, 24, p. 1-34.
- Valliant, R. (2020), « Comparing Alternatives for Estimation from Nonprobability Samples », *Journal of Survey Statistics and Methodology*, 8, p. 231-263.
- Wang, L., R. Valliant et Y. Li. (2021), « Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts », *Statistics in Medicine*, 40, p. 5237-5250.
- Wu, C. (2022), « A new criterion for confounder selection », *Techniques d'enquête*, 48, p. 283-311.