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

Techniques d'enquête

Réponse des auteurs aux commentaires sur l'article « Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada » :

De nouvelles avancées concernant les méthodes de vraisemblance pour l'estimation des probabilités de participation pour des échantillons non probabilistes

par Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois et Kenneth Chu

Date de diffusion : le 25 juin 2024

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 2024
Vol. 50, nº 1, pp. 139-160
Statistique Canada, nº 12-001-X au catalogue

Réponse des auteurs aux commentaires sur l'article « Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada » :

De nouvelles avancées concernant les méthodes de vraisemblance pour l'estimation des probabilités de participation pour des échantillons non probabilistes

Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois et Kenneth Chu[1]

Résumé

Inspirés par les deux excellentes discussions de notre article, nous offrons un regard nouveau et présentons de nouvelles avancées sur le problème de l'estimation des probabilités de participation pour des échantillons non probabilistes. Tout d'abord, nous proposons une amélioration de la méthode de Chen, Li et Wu (2020), fondée sur la théorie de la meilleure estimation linéaire sans biais, qui tire plus efficacement parti des données disponibles des échantillons probabiliste et non probabiliste. De plus, nous élaborons une méthode de vraisemblance de l'échantillon, dont l'idée est semblable à la méthode d'Elliott (2009), qui tient adéquatement compte du chevauchement entre les deux échantillons quand il est possible de l'identifier dans au moins un des échantillons. Nous utilisons la théorie de la meilleure prédiction linéaire sans biais pour traiter le scénario où le chevauchement est inconnu. Il est intéressant de constater que les deux méthodes que nous proposons coïncident quand le chevauchement est inconnu. Ensuite, nous montrons que de nombreuses méthodes existantes peuvent être obtenues comme cas particulier d'une fonction d'estimation sans biais générale. Enfin, nous concluons en formulant quelques commentaires sur l'estimation non paramétrique des probabilités de participation.

Mots-clés : Meilleure estimation linéaire sans biais; meilleure prédiction linéaire sans biais; équation d'estimation; vraisemblance de la population; pseudo-vraisemblance; vraisemblance de l'échantillon.

1. Observations générales

Nous aimerions commencer par remercier les participants à la discussion d'avoir pris le temps de lire notre article et de nous faire part de leurs observations judicieuses et réfléchies sur la pondération par l'inverse de la probabilité pour les échantillons non probabilistes. Dr. Wu nous éclaire sur trois méthodes de pondération fréquemment utilisées pour les échantillons non probabilistes, tandis que Dr. Gershunskaya et Dr. Beresovsky présentent deux nouvelles méthodes : la régression logistique implicite, voir aussi Beresovsky (2019), Savitsky, Williams, Gershunskaya, et coll. (2022), Gershunskaya et Lahiri (2023), et la pseudo-régression logistique implicite. Nous avons grandement apprécié la lecture de ces deux discussions qui nous ont permis d'améliorer nos connaissances dans le domaine et nous ont incités à approfondir nos

1. Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois et Kenneth Chu, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6. Courriels : jean-francois.beaumont@statcan.gc.ca, keven.bosa@statcan.gc.ca, andrew.brennan@statcan.gc.ca, joanne.charlebois@statcan.gc.ca et kenneth.chu@statcan.gc.ca.

réflexions sur le sujet. Dans ce qui suit, nous vous ferons part de ces réflexions ainsi que de nouvelles avancées.

Les sections 2, 3 et 4 sont consacrées aux méthodes de Chen, Li et Wu (2020), Wang, Valliant et Li (2021) et Elliott (2009), voir aussi Elliott et Valliant (2017), respectivement. Nous fournissons des observations supplémentaires sur ces méthodes et nous établissons des liens avec les méthodes de régression logistique implicite et de pseudo-régression logistique implicite. Nous montrons que les trois méthodes sont valides en ce sens qu'elles donnent des fonctions d'estimation sans biais pour les paramètres du modèle de participation, quelle que soit la taille de l'échantillon probabiliste et non probabiliste, ainsi que la taille du chevauchement entre les deux échantillons. Toutefois, seule la méthode Chen-Li-Wu (CLW) possède la propriété d'être équivalente à la méthode du maximum de vraisemblance quand l'échantillon probabiliste est un recensement, ce que nous appelons la propriété de vraisemblance du recensement. Dans la section 2, nous montrons également que la méthode CLW ne tire pas pleinement parti de l'information auxiliaire disponible, ce qui peut entraîner une fonction d'estimation inefficace, particulièrement quand l'échantillon non probabiliste est plus grand que l'échantillon probabiliste. En utilisant la théorie de la meilleure estimation linéaire sans biais, nous proposons une amélioration de la méthode CLW qui répond à ce problème et qui satisfait toujours la propriété de vraisemblance du recensement. Dans la section 5, nous proposons une méthode de vraisemblance de l'échantillon, semblable en esprit à la méthode d'Elliott et de régression logistique implicite, qui tient adéquatement compte du chevauchement entre les deux échantillons, à condition qu'il puisse être identifié dans l'un des deux échantillons. Notre méthode de vraisemblance de l'échantillon satisfait la propriété de vraisemblance du recensement. Au moyen de la théorie de la meilleure prédiction linéaire sans biais, nous obtenons une fonction d'estimation « optimale » applicable quand le chevauchement ne peut être identifié dans aucun des deux échantillons. Il est intéressant de mentionner qu'elle est identique à la fonction d'estimation sous-jacente à notre amélioration de la méthode CLW. À la section 6, nous unifions les méthodes existantes qui n'exigent pas l'identification du chevauchement et nous montrons leur équivalence pour le modèle des groupes homogènes. Une brève synthèse est présentée à la section 7, ainsi que quelques commentaires sur l'estimation non paramétrique des probabilités de participation.

2. Méthode de la vraisemblance de la population et méthode de Chen, Li et Wu (2020)

Nous utilisons la notation de notre article principal : le vecteur des variables auxiliaires pour l'unité k de la population finie U est noté $\mathbf{x}_k$, et l'indicateur de participation (indicateur de participation à l'échantillon non probabiliste $s_{\text{NP}} \subset U$) pour l'unité de population $k \in U$ est désigné par δ_k. La probabilité de participation $p_k = \Pr\left(\delta_k = 1 \middle| \mathbf{x}_k\right) > 0$ est modélisée au moyen d'un modèle paramétrique tel que le modèle logistique $p_k(\boldsymbol{\alpha}) = \left[1 + \exp\left(-\mathbf{x}_k'\boldsymbol{\alpha}\right)\right]^{-1}$, où $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus du modèle. Nous posons l'hypothèse d'indépendance standard suivante :

A1) $\delta_k,\ k \in U$ sont mutuellement indépendants étant donné $\mathbf{x}_k,\ k \in U$.

Idéalement, nous aurions accès à la fois à $\{\mathbf{x}_k; k \in s_{\text{NP}}\}$ et $\{\mathbf{x}_k; k \in U\}$. Ces deux ensembles de données n'auraient pas besoin d'être couplés. Sous ce scénario idéal et l'hypothèse (A1), le logarithme de la fonction de vraisemblance de la population est

$$l(\boldsymbol{\alpha}) = \log\left\{\prod_{k\in U}\left[p_k(\boldsymbol{\alpha})\right]^{\delta_k}\left[1-p_k(\boldsymbol{\alpha})\right]^{(1-\delta_k)}\right\} = \sum\nolimits_{k\in s_{\text{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right] + \sum\nolimits_{k\in U}\log\left[1-p_k(\boldsymbol{\alpha})\right]$$

et la fonction d'estimation de vraisemblance de la population (ou fonction de score) est

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum\nolimits_{k\in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})}\frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1-p_k(\boldsymbol{\alpha})\right]} - \sum\nolimits_{k\in U}\frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1-p_k(\boldsymbol{\alpha})\right]}, \tag{2.1}$$

où $\mathbf{g}_k(\boldsymbol{\alpha}) = \partial p_k(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$. En particulier, $\mathbf{g}_k(\boldsymbol{\alpha}) = p_k(\boldsymbol{\alpha})\left[1-p_k(\boldsymbol{\alpha})\right]\mathbf{x}_k$ pour le modèle logistique.

Dans de nombreux cas réels, le vecteur $\mathbf{x}_k$ n'est pas connu pour l'ensemble de la population, mais il est au moins disponible dans un échantillon probabiliste s_P en plus de l'échantillon non probabiliste s_{NP}. Par conséquent, le terme $\sum_{k\in U}\log\left[1-p_k(\boldsymbol{\alpha})\right]$ dans le logarithme de la fonction de vraisemblance de la population $l(\boldsymbol{\alpha})$ n'est pas calculable, car il dépend de valeurs inconnues de $\mathbf{x}_k$. Chen, Li et Wu (2020) ont proposé de régler ce problème en estimant ce terme au moyen de l'échantillon probabiliste. Cela donne le logarithme de la fonction de pseudo-vraisemblance :

$$\hat{l}(\boldsymbol{\alpha}) = \sum\nolimits_{k\in s_{\text{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right] + \sum\nolimits_{k\in s_P} w_k \log\left[1-p_k(\boldsymbol{\alpha})\right], \tag{2.2}$$

où $w_k = 1/\pi_k$ est un poids d'enquête probabiliste pour l'unité $k \in s_P$ et π_k est sa probabilité de sélection. Nous nous concentrons sur ce poids de base à des fins de simplicité, bien que des méthodes de pondération plus complexes, comportant des ajustements pour la non-réponse et de calage, soient souvent utilisées dans les enquêtes réelles. En prenant la dérivée de (2.2) par rapport à $\boldsymbol{\alpha}$, nous obtenons la fonction d'estimation de pseudo-vraisemblance

$$\hat{\mathbf{U}}_\pi(\boldsymbol{\alpha}) = \sum\nolimits_{k\in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})}\frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1-p_k(\boldsymbol{\alpha})\right]} - \sum\nolimits_{k\in s_P} w_k \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1-p_k(\boldsymbol{\alpha})\right]}. \tag{2.3}$$

On constate facilement que la fonction d'estimation (2.3) est sans biais par rapport à md, conditionnellement à π_k et $\mathbf{x}_k,\ k \in U$, c'est-à-dire $E_{md}\left[\hat{\mathbf{U}}_\pi(\boldsymbol{\alpha})\right] = \mathbf{0}$, à condition que l'hypothèse suivante soit valide :

A2) $E_m\left(\delta_k \middle| \pi_k, \mathbf{x}_k\right) = E_m\left(\delta_k \middle| \mathbf{x}_k\right) = p_k(\boldsymbol{\alpha}),\ k \in U.$

L'indice m fait référence au modèle de participation et l'indice d au plan de sondage probabiliste. Conditionner sur $\pi_k, k \in U$ est naturel quand π_k est disponible dans les deux échantillons puisqu'elle peut être traitée comme une variable auxiliaire potentielle. En effet, l'hypothèse (A2) est automatiquement

satisfaite si π_k est incluse dans le vecteur $\mathbf{x}_k$. À la section 2.1, nous conditionnerons sur π_k, $k \in U$. Ensuite, à la section 2.2, nous examinerons le cas où π_k est traitée comme étant aléatoire et où les inférences sont conditionnelles seulement à $\mathbf{x}_k$, $k \in U$.

2.1 Amélioration de la fonction d'estimation de la méthode CLW au moyen de la théorie de la meilleure estimation linéaire sans biais

Le deuxième terme du côté droit de (2.3) est un estimateur du terme correspondant dans (2.1), $\boldsymbol{\Phi}(\boldsymbol{\alpha}) = \sum_{k \in U} \mathbf{g}_k(\boldsymbol{\alpha}) \big/ \left[1 - p_k(\boldsymbol{\alpha})\right]$. Il s'agit d'un estimateur inefficace de $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ parce qu'il utilise uniquement des données de l'échantillon probabiliste et ignore les données auxiliaires pertinentes de l'échantillon non probabiliste. Un estimateur plus efficace utiliserait donc les données auxiliaires des deux échantillons. Cet estimateur pourrait être obtenu en appliquant le principe de l'information manquante. Le principe de l'information manquante a été introduit par Orchard et Woodbury (1972). Voir aussi Chambers (2023) pour une référence récente sur les applications du principe de l'information manquante aux données d'enquête. Le principe de l'information manquante consiste à remplacer la fonction d'estimation de vraisemblance de la population (2.1) par son espérance conditionnelle aux données observées ou, de façon équivalente, à remplacer $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ par son meilleur prédicteur. Cependant, cela nécessiterait de modéliser le vecteur des variables auxiliaires et cette approche serait généralement difficile à mettre en œuvre.

Comme solution de rechange à l'application du principe de l'information manquante, nous proposons d'estimer $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ au moyen de la théorie de la meilleure estimation linéaire sans biais. Nous considérons l'estimateur linéaire sans biais suivant qui utilise les données auxiliaires des deux échantillons :

$$\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \frac{\gamma_k}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]} + \sum_{k \in s_P} w_k (1 - \gamma_k) \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]}, \tag{2.4}$$

où γ_k, $k \in U$, sont des constantes. Il est facile de montrer que $\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})$ est sans biais par rapport à md pour $\boldsymbol{\Phi}(\boldsymbol{\alpha})$, c'est-à-dire $E_{md}\left[\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right] = \boldsymbol{\Phi}(\boldsymbol{\alpha})$, à condition que l'hypothèse (A2) soit valide. En remplaçant le deuxième terme du côté droit de (2.1) par le terme du côté droit de (2.4), nous obtenons la fonction d'estimation sans biais par rapport à md :

$$\hat{\mathbf{U}}_{\pi}^{\gamma}(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{(1 - \gamma_k)}{\left[1 - p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) - \sum_{k \in s_P} w_k \frac{(1 - \gamma_k)}{\left[1 - p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}). \tag{2.5}$$

Il est facile de voir que la fonction d'estimation (2.3) de la méthode CLW est le cas particulier de (2.5) obtenu en posant $\gamma_k = 0$ pour toutes les unités $k \in U$.

On obtient le meilleur estimateur linéaire sans biais de $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ en trouvant γ_k, $k \in U$, qui minimise $\mathrm{var}_{md}\left[\mathbf{c}'\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right]$ pour tout vecteur fixe $\mathbf{c} \neq \mathbf{0}$. Nous posons les hypothèses suivantes :

A3) I_k, $k \in U$, sont mutuellement indépendants étant donné π_k et $\mathbf{x}_k$, $k \in U$, où I_k est l'indicateur d'inclusion dans l'échantillon probabiliste s_P.

A4) $E_d\left(I_k \,|\delta_k, \pi_k, \mathbf{x}_k\right) = E_d\left(I_k \,|\pi_k, \mathbf{x}_k\right) = \pi_k,\ k \in U.$

L'hypothèse (A3) implique que l'échantillon probabiliste est sélectionné au moyen d'un échantillonnage de Poisson. Elle sert à simplifier les dérivations de $\text{var}_{md}\left[\mathbf{c}'\hat{\mathbf{\Phi}}(\boldsymbol{\alpha})\right]$ même si nous reconnaissons que d'autres plans de sondage peuvent être utilisés en pratique. Mentionnons que ni l'hypothèse (A3) ni les hypothèses (A1) et (A4) ne sont nécessaires pour prouver que la fonction d'estimation (2.5) est sans biais par rapport à md. Sous les hypothèses (A1)-(A4), on peut facilement montrer que

$$\text{var}_{md}\left[\mathbf{c}'\hat{\mathbf{\Phi}}(\boldsymbol{\alpha})\right] = \sum_{k\in U} \frac{\left[1-p_k(\boldsymbol{\alpha})\right]}{p_k(\boldsymbol{\alpha})} \gamma_k^2 \left(\frac{\mathbf{c}'\mathbf{g}_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right)^2 + \sum_{k\in U} \frac{(1-\pi_k)}{\pi_k}(1-\gamma_k)^2 \left(\frac{\mathbf{c}'\mathbf{g}_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right)^2. \qquad (2.6)$$

La variance (2.6) est minimisée quand

$$\gamma_k = \gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = \frac{(1-\pi_k)p_k(\boldsymbol{\alpha})}{(1-\pi_k)p_k(\boldsymbol{\alpha}) + \left[1-p_k(\boldsymbol{\alpha})\right]\pi_k} = \frac{w_k - 1}{(w_k-1) + (p_k^{-1}(\boldsymbol{\alpha}) - 1)},\ k \in U. \qquad (2.7)$$

La substitution par $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$ dans (2.4) donne le meilleur estimateur linéaire sans biais de $\mathbf{\Phi}(\boldsymbol{\alpha})$.

Les propriétés suivantes sont associées à $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$:

i) si $\pi_k = p_k(\boldsymbol{\alpha})$ alors $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = 1/2$;

ii) si $\pi_k > p_k(\boldsymbol{\alpha})$ alors $0 \le \gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) < 1/2$;

iii) si $\pi_k < p_k(\boldsymbol{\alpha})$ alors $1/2 < \gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) \le 1$;

iv) si $\pi_k \to 1$ ou $p_k(\boldsymbol{\alpha}) \to 0$ alors $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) \to 0$;

v) si $\pi_k \to 0$ ou $p_k(\boldsymbol{\alpha}) \to 1$ alors $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) \to 1$.

En raison des propriétés (iii) et (v), quand l'échantillon probabiliste est petit comparativement à l'échantillon non probabiliste, on s'attend à ce que $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$ soit grand pour de nombreuses unités de la population, et la méthode CLW $(\gamma_k = 0)$ peut devenir inefficace par rapport à la solution optimale (2.7). L'inefficacité de la méthode CLW dans ce scénario a été démontrée dans l'étude empirique de Savitsky et coll. (2022). L'explication de cette inefficacité est que la méthode CLW ignore le grand échantillon non probabiliste pour l'estimation du total de la population $\mathbf{\Phi}(\boldsymbol{\alpha})$. En vertu des propriétés (ii) et (iv), la méthode CLW devrait être plus performante dans le scénario inverse où l'échantillon probabiliste est beaucoup plus grand que l'échantillon non probabiliste étant donné qu'elle possède la propriété de vraisemblance du recensement, c'est-à-dire que la fonction d'estimation (2.3) devient équivalente à la fonction d'estimation de vraisemblance de la population (2.1) quand l'échantillon probabiliste est un recensement. Ce scénario n'est pas irréaliste en pratique. À titre d'exemple, le questionnaire détaillé du recensement canadien, mené de façon aléatoire auprès de 25 % de la population canadienne, pourrait constituer un échantillon probabiliste efficace pour l'estimation des probabilités de participation d'un échantillon non probabiliste plus petit.

Si nous substituons (2.7) dans la fonction d'estimation (2.5), nous obtenons la fonction d'estimation « optimale » :

$$\begin{aligned}\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha}) &= \sum\nolimits_{k\in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\pi_k}{\left[\pi_k + p_k(\boldsymbol{\alpha}) - 2\pi_k p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) \\ &\quad - \sum\nolimits_{k\in s_P} \frac{1}{\pi_k} \frac{\pi_k}{\left[\pi_k + p_k(\boldsymbol{\alpha}) - 2\pi_k p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}).\end{aligned} \tag{2.8}$$

On peut facilement démontrer que $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$ est le meilleur prédicteur linéaire sans biais de $\mathbf{U}(\boldsymbol{\alpha})$ donné dans (2.1). Comme la fonction d'estimation (2.3) de la méthode CLW, elle possède la propriété de vraisemblance du recensement.

2.2 Lissage des poids

Comme nous l'avons vu plus haut, l'inefficacité possible de (2.3) peut s'expliquer principalement par l'omission de données auxiliaires pertinentes de l'échantillon non probabiliste pour l'estimation de $\boldsymbol{\Phi}(\boldsymbol{\alpha})$. Une autre source possible d'inefficacité peut être attribuable à la variabilité des poids d'enquête $w_k,\ k \in s_P$. En effet, on sait bien que l'estimation de pseudo-vraisemblance peut être inefficace pour l'estimation des paramètres d'un modèle à partir de données d'enquête probabiliste (par exemple voir les travaux récents de Chambers, 2023). On peut utiliser le lissage de poids (Beaumont, 2008) pour régler ce problème. Dans ce contexte, il consiste à remplacer le poids d'enquête w_k dans (2.3) par le poids lissé $\tilde{w}_k = E_{\xi}\left(w_k \,|k \in s_P, \mathbf{x}_k\right)$, où l'indice ξ indique que l'espérance est prise par rapport à un modèle pour π_k (ou w_k). Le poids lissé $\tilde{w}_k$ est souvent inconnu, mais il peut être estimé au moyen de l'échantillon probabiliste et de modèles paramétriques ou non paramétriques. Si π_k est disponible dans l'échantillon non probabiliste et inclus dans le vecteur $\mathbf{x}_k$, $\tilde{w}_k = w_k$ et le lissage de poids n'apporte aucun gain d'efficacité.

En utilisant une relation donnée dans Pfeffermann et Sverchkov (1999), on peut exprimer le poids lissé comme suit :

$$\tilde{w}_k = E_{\xi}\left(w_k \,|k \in s_P, \mathbf{x}_k\right) = \frac{1}{E_{\xi}\left(\pi_k \,|\mathbf{x}_k\right)} = \frac{1}{\tilde{\pi}_k}, \tag{2.9}$$

où $\tilde{\pi}_k = E_{\xi}\left(\pi_k \,|\mathbf{x}_k\right) = \Pr\left(k \in s_P \,|\mathbf{x}_k\right)$. Par cette relation, on peut démontrer facilement que la fonction d'estimation (2.3) est sans biais par rapport à ξmd, quelle que soit la validité de l'hypothèse (A2) et que (2.3) utilise w_k ou $\tilde{w}_k$, c'est-à-dire $E_{\xi md}\left[\hat{\mathbf{U}}_{\pi}(\boldsymbol{\alpha})\right] = \mathbf{0}$ et $E_{\xi md}\left[\hat{\mathbf{U}}_{\tilde{\pi}}(\boldsymbol{\alpha})\right] = \mathbf{0}$, où $\hat{\mathbf{U}}_{\tilde{\pi}}(\boldsymbol{\alpha})$ est la fonction d'estimation (2.3) dans laquelle w_k est remplacé par $\tilde{w}_k$. Mentionnons que les espérances ξmd sont conditionnelles seulement à $\mathbf{x}_k,\ k \in U$, de sorte que π_k est traité comme étant aléatoire. La relation (2.9) peut aussi servir à obtenir un estimateur de $\tilde{\pi}_k$, c'est-à-dire qu'un estimateur convergent de $\tilde{\pi}_k = E_{\xi}\left(\pi_k \,|\mathbf{x}_k\right)$ est $\hat{\tilde{\pi}}_k = 1/\hat{\tilde{w}}_k$, où $\hat{\tilde{w}}_k$ est un estimateur convergent de $\tilde{w}_k = E_{\xi}\left(w_k \,|k \in s_P, \mathbf{x}_k\right)$.

L'utilisation de $\tilde{w}_k$ au lieu de w_k dans la fonction d'estimation (2.3) augmente son efficacité au détriment de nécessiter la validité d'un modèle pour w_k et l'estimation de $\tilde{w}_k$. On peut avancer un argument

semblable pour améliorer l'efficacité de $\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})$ en remplaçant w_k par $\tilde{w}_k$ dans (2.4). L'estimateur qui en résulte est sans biais par rapport à ξmd, c'est-à-dire $E_{\xi md}\left[\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right]=\boldsymbol{\Phi}(\boldsymbol{\alpha})$, et sa variance $\text{var}_{\xi md}\left[\mathbf{c}'\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right]$ prend la même forme que (2.6), où π_k est remplacé par $\tilde{\pi}_k$, à condition que les hypothèses (A1)-(A4) soient valides ainsi que l'hypothèse suivante :

A5) $\pi_k, k\in U$ sont mutuellement indépendantes étant donné $\mathbf{x}_k, k\in U$.

Par conséquent, la valeur optimale de γ_k, désignée par $\gamma_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$, et la fonction d'estimation optimale, désignée par $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$, sont à nouveau données par les expressions (2.7) et (2.8), respectivement, où π_k est remplacé par $\tilde{\pi}_k$.

L'utilisation de π_k dans (2.8) est impossible si elle n'est pas observée dans l'échantillon non probabiliste, un scénario vraisemblable dans la pratique. Dans ce cas, on peut utiliser une estimation de $\tilde{\pi}_k$ dans (2.8) pour remplacer π_k. Si π_k est observé dans l'échantillon non probabiliste, mais non incluse dans $\mathbf{x}_k$, on peut tout de même souhaiter remplacer π_k par une estimation de $\tilde{\pi}_k$ pour améliorer l'efficacité de la fonction d'estimation optimale (2.8).

2.3 Sélection des variables

La fonction d'estimation (2.8) n'est pas la dérivée du logarithme d'une fonction de pseudo-vraisemblance. Par conséquent, la méthodologie que nous avons utilisée dans notre article principal pour dériver un critère d'information d'Akaike (AIC), basé sur Lumley et Scott (2015), n'est pas applicable directement. Pour la sélection des variables, une option consiste à utiliser notre AIC proposé avec la méthode CLW. Une fois les variables auxiliaires sélectionnées, la fonction d'estimation (2.8), $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$, ou sa version lissée $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$, peut servir à estimer $\boldsymbol{\alpha}$ plutôt que la fonction d'estimation CLW. Sinon, on pourrait aussi envisager des méthodes de sélection de variables non fondées sur la vraisemblance.

Mentionnons finalement que la méthodologie élaborée dans la présente section pour obtenir un estimateur optimal (ou meilleur estimateur linéaire sans biais) de $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ pourrait également servir à combiner deux échantillons probabilistes indépendants tirés de la même population. Cette idée pourrait être évaluée dans des travaux futurs.

3. La méthode de Wang, Valliant et Li (2021) et la méthode de pseudo-régression logistique implicite

La méthode de Wang-Valliant-Li (WVL) consiste à créer une population artificielle U_A en empilant l'échantillon non probabiliste s_{NP} par-dessus la population U. Chaque élément de U_A est considéré comme distinct même si les unités de l'échantillon non probabiliste sont présentes deux fois dans U_A. Un indicateur R_i est défini pour chaque élément $i\in U_A$; $R_i=1$ si $i\in s_{\text{NP}}\cap U_A$, et $R_i=0$ si $i\in U\cap U_A$. Nous utilisons l'indice i pour désigner les éléments de la population artificielle U_A afin de les distinguer des unités de la population U. Pour une unité donnée $k\in s_{\text{NP}}$, il y a deux éléments distincts de U_A dont l'étiquette diffère

de celle de cette unité k; $R=0$ pour un élément et $R=1$ pour l'autre. On suppose que l'échantillon probabiliste est sélectionné à partir des éléments dans $U \cap U_A$ pour lesquels $R_i = 0$. Les auteurs ont également supposé que les indicateurs R_i, $i \in U_A$, sont mutuellement indépendants étant donné $\mathbf{x}_i$, $i \in U_A$, et ont obtenu le logarithme d'une fonction de pseudo-vraisemblance semblable à celle de CLW en modélisant $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right)$ au moyen d'un modèle logistique. Ensuite, ils ont établi la relation $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right) = p_i/(1+p_i)$, qui leur a permis d'estimer la probabilité de participation p_i. Parce qu'ils ont utilisé un modèle logistique pour $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right)$, ils ont implicitement modélisé p_i au moyen du modèle exponentiel $p_i(\boldsymbol{\alpha}) = \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)$, qui a la caractéristique indésirable d'admettre des estimations supérieures à 1. Cependant, rien dans leur théorie ne les aurait empêchés d'utiliser un autre modèle pour p_i, comme le modèle logistique, et ainsi d'obtenir implicitement un modèle pour $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right)$. C'est exactement ce que Dr. Gershunskaya et Dr. Beresovsky ont proposé dans leur discussion. Ils appellent leur méthode pseudo-régression logistique implicite, qui est simplement la méthode WVL avec un modèle logistique pour p_i.

Pour un modèle paramétrique arbitraire pour p_i, la fonction d'estimation de pseudo-régression logistique implicite ou de WVL peut être exprimée comme suit :

$$\hat{\mathbf{U}}_{\pi}^{\text{WVL-PRLI}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1+p_k(\boldsymbol{\alpha})\right]} - \sum_{k \in s_P} w_k \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1+p_k(\boldsymbol{\alpha})\right]}. \tag{3.1}$$

Comme pour la méthode CLW, on peut possiblement améliorer la fonction d'estimation (3.1) en remplaçant le poids d'enquête w_k par le poids lissé $\tilde{w}_k$. La fonction d'estimation qui en résulte est désignée par $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{WVL-PRLI}}(\boldsymbol{\alpha})$.

Dr. Gershunskaya et Dr. Beresovsky ont souligné que, contrairement à δ_k, l'indicateur R_i est entièrement observé une fois que les échantillons probabilistes et non probabilistes sont observés. Cependant, cette caractéristique de R_i est trompeuse. Les indicateurs R_i pour les éléments de l'échantillon probabiliste et de l'échantillon non probabiliste n'apportent aucune nouvelle information sur le mécanisme de participation au-delà de ce qui est observable à propos de δ_k et utilisé dans la méthode CLW. Autrement dit, les deux méthodes utilisent les mêmes informations observées : $\left\{w_k, \mathbf{x}_k; k \in s_P\right\}$ et $\left\{\delta_k, \mathbf{x}_k; k \in s_{\text{NP}}\right\}$. De plus, selon nous, la méthode WVL pose deux problèmes principaux, qui sont décrits ci-dessous.

Problème 1 : L'hypothèse selon laquelle R_i, $i \in U_A$, sont mutuellement indépendants étant donné $\mathbf{x}_i$, $i \in U_A$, n'est pas valide, car chaque unité de l'échantillon non probabiliste est présente deux fois dans U_A : $R_i = 0$ pour un élément et $R_i = 1$ pour l'autre (voir davantage de précision ci-dessous).

Problème 2 : La fonction d'estimation (3.1) n'a pas la propriété de vraisemblance du recensement, car elle ne se réduit pas à la fonction d'estimation (2.1) de vraisemblance de la population quand l'échantillon probabiliste est un recensement. Quand $s_P = U$, la méthode CLW utilise la même information que si $\left\{\delta_k, \mathbf{x}_k; k \in U\right\}$ étaient connus, mais le logarithme de la fonction de vraisemblance de WVL ne reconnaît pas cette information.

Malgré les deux problèmes décrits ci-dessus, il est facile de montrer que la fonction d'estimation (3.1) est sans biais par rapport à md, à condition que l'hypothèse (A2) soit valide. Les deux fonctions d'estimation $\hat{\mathbf{U}}_{\pi}^{\text{WVL-PRLI}}(\boldsymbol{\alpha})$ et $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{WVL-PRLI}}(\boldsymbol{\alpha})$ sont également sans biais par rapport à ξmd, indépendamment de la validité de l'hypothèse (A2); la fonction d'estimation (3.1) est donc valide. Elle peut être écrite sous la forme (2.5) avec $\gamma_k = \gamma_k^{\text{WVL-PRLI}}(\boldsymbol{\alpha}) = 2\,p_k(\boldsymbol{\alpha}) \big/ \left[1 + p_k(\boldsymbol{\alpha})\right]$. On constate facilement que $0 < \gamma_k^{\text{WVL-PRLI}}(\boldsymbol{\alpha}) \le 1$; elle utilise donc dans une certaine mesure les données auxiliaires de l'échantillon non probabiliste pour l'estimation de $\boldsymbol{\Phi}(\boldsymbol{\alpha})$.

La variance (2.6) est une fonction quadratique de γ_k qui est minimisée quand $\gamma_k = \gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$, $k \in U$. Par conséquent, si $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$ est plus proche de $\gamma_k^{\text{WVL-PRLI}}(\boldsymbol{\alpha})$ que de 0 pour la plupart des unités de la population, c'est-à-dire que $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) > 0{,}5\gamma_k^{\text{WVL-PRLI}}(\boldsymbol{\alpha})$, la fonction d'estimation (3.1) de WVL devrait être plus efficace que la fonction d'estimation CLW ($\gamma_k = 0$), mais resterait moins efficace que la fonction d'estimation optimale (2.8). On peut montrer facilement que $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) > 0{,}5\gamma_k^{\text{WVL-PRLI}}(\boldsymbol{\alpha})$ est satisfait quand $\pi_k < \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}$. Cependant, si $\pi_k > \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}$ pour la plupart des unités de la population, la fonction d'estimation (2.3) de CLW devrait devenir plus efficace que (3.1), en particulier quand l'échantillon probabiliste est un recensement ($\pi_k = 1, k \in U$). Puisque $\left[2 - p_k(\boldsymbol{\alpha})\right]^{-1} > 0{,}5$, la condition $\pi_k < \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}$ est généralement plus courante dans les enquêtes sociales que $\pi_k > \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}$, du moins pour la plupart des unités de la population. On peut aussi démontrer facilement que $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = \gamma_k^{\text{WVL-PRLI}}(\boldsymbol{\alpha})$ quand $\pi_k = 1/3$ alors que $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = 0$ quand $\pi_k = 1$. La fonction d'estimation WVL devrait donc être proche de la fonction d'estimation optimale quand la taille de l'échantillon probabiliste est d'environ un tiers de la taille de la population et que les probabilités de sélection π_k ne sont pas trop variables.

Comme Dr. Wu l'a indiqué dans sa discussion, nous avons eu de la difficulté à comprendre le cadre probabiliste sous-jacent à la relation $\Pr\left(R_i = 1 \,\middle|\, i \in U_A, \mathbf{x}_i\right) = p_i / (1 + p_i)$. Cependant, le cadre ingénieux proposé par Savitsky et coll. (2022) semble correctement justifier cette relation. Ces auteurs ont imaginé une population augmentée fixe U^* obtenue en empilant la population U_1 par-dessus la population U_2, où U_1 et U_2 sont deux populations identiques à U de taille N, mais étiquetées de façon unique de sorte que chacun des $2N$ éléments de U^* sont considérés comme distincts. Premièrement, une des deux populations est choisie au hasard avec une probabilité de ½. Nous désignons cette population sélectionnée au hasard par U_{NP}, qui est soit U_1 soit U_2. L'autre population est désignée par U_P. L'échantillon non probabiliste s_{NP} est observé à partir de U_{NP}, et l'échantillon probabiliste s_P est sélectionné aléatoirement à partir de U_P. En utilisant ce cadre, on peut facilement montrer que

$$\Pr\left(R_i = 1 \,\middle|\, i \in s_{\text{NP}} \cup U_P, \mathbf{x}_i\right) = \frac{\Pr\left(i \in s_{\text{NP}} \,\middle|\, \mathbf{x}_i\right)}{\Pr\left(i \in s_{\text{NP}} \cup U_P \,\middle|\, \mathbf{x}_i\right)} = \frac{1/2\,p_i}{1/2\,p_i + 1/2} = \frac{p_i}{(1 + p_i)}.$$

Mentionnons que le fractionnement aléatoire de U^* en U_{NP} et U_P n'est pas explicitement énoncé dans Savitsky et coll. (2022), mais qu'il est nécessaire pour obtenir l'équation ci-dessus.

Ce cadre probabiliste semble corriger la situation, mais les deux problèmes mentionnés plus haut demeurent. En particulier, il est facile de démontrer que l'hypothèse d'indépendance n'est pas satisfaite puisque, pour toute paire d'éléments $i \in U_1$ et $j \in U_2$,

$$\Pr\left(R_i = 1, R_j = 1 \middle| i, j \in s_{\text{NP}} \cup U_P, \mathbf{x}_i, \mathbf{x}_j\right) = 0 \neq \frac{p_i}{(1+p_i)} \frac{p_j}{(1+p_j)}.$$

Pour deux éléments différents i et j dans la même population, soit U_1 ou U_2, nous avons

$$\Pr\left(R_i = 1, R_j = 1 \middle| i, j \in s_{\text{NP}} \cup U_P, \mathbf{x}_i, \mathbf{x}_j\right) = \frac{p_i p_j}{1+p_i p_j} \neq \frac{p_i}{(1+p_i)} \frac{p_j}{(1+p_j)}.$$

à condition que l'hypothèse (A1) soit valide pour les éléments $i \in U_{\text{NP}}$. Par conséquent, l'hypothèse d'indépendance est raisonnable seulement quand toutes les probabilités de participation (ou au moins un grand nombre d'entre elles) sont petites, et donc, que le chevauchement représente une petite partie de l'échantillon probabiliste. Dans cette situation, les fonctions d'estimation WVL et CLW devraient être à peu près équivalentes. Si la plupart des probabilités de participation sont grandes, le logarithme de la fonction de pseudo-vraisemblance proposée par WVL repose sur une hypothèse d'indépendance incorrecte. En principe, un critère d'information d'Akaike qui repose sur le logarithme d'une fonction de (pseudo) vraisemblance incorrecte n'est pas valide. À quel point les probabilités de participation doivent-elles être petites pour que cette hypothèse d'indépendance soit raisonnable ? L'étude par simulations de Dr. Gershunskaya et Dr. Beresovsky est un premier pas dans cette direction, mais il faudrait d'autres études sur cette question. Mentionnons que le logarithme de la fonction de pseudo-vraisemblance de CLW est valide, quelle que soit l'ampleur des probabilités de participation, pourvu que l'hypothèse (A1) soit valide.

4. La méthode d'Elliott (2009) et la régression logistique implicite

Dans la méthode d'Elliott (2009), voir aussi Elliott et Valliant (2017), on obtient un échantillon combiné s^* en empilant l'échantillon non probabiliste s_{NP} sur l'échantillon probabiliste s_P tout en ignorant le chevauchement (inconnu) possible. Une unité de la population $k \in U$ qui est sélectionnée dans s_P et observée dans s_{NP} est donc présente deux fois dans s^*. Elliott (2009) a supposé implicitement que le chevauchement entre les deux échantillons était négligeable. Comme dans Wang, Valliant et Li (2021), un indicateur z_i, $i \in s^*$, est créé de telle sorte que $z_i = 1$ si $i \in s_{\text{NP}} \cap s^*$, et $z_i = 0$ si $i \in s_P \cap s^*$. Elliott (2009) a proposé de modéliser $\rho_i = \Pr\left(z_i = 1 \middle| i \in s^*, \mathbf{x}_i\right)$ au moyen d'un modèle logistique et, en supposant que les fractions d'échantillonnage sont petites (Elliott et Valliant, 2017), il a établi la relation $p_i = K\tilde{\pi}_i \rho_i / (1-\rho_i)$ utilisée pour estimer p_i, où K est une constante de proportionnalité inconnue. Cela implique que $\rho_i = p_i / (K\tilde{\pi}_i + p_i)$. En pratique, $\tilde{\pi}_i$ est généralement inconnu, mais elle peut être estimée, comme nous l'avons vu à la section 2.

Dans la présente section et la suivante, nous conditionnons sur $\mathbf{x}_i$ et nous traitons π_i comme étant aléatoire. La théorie reste valide si nous conditionnons à la fois sur $\mathbf{x}_i$ et π_i à condition que $\tilde{\pi}_i$ soit remplacé par π_i dans les développements ci-dessous et que l'hypothèse (A2) soit valide. Le conditionnement sur π_i n'a de sens que s'il est observé dans les deux échantillons, afin qu'il puisse être traité comme une variable auxiliaire potentielle et inclus dans $\mathbf{x}_i$. Si π_i est inclus dans $\mathbf{x}_i$, $\tilde{\pi}_i = \pi_i$ et l'hypothèse (A2) est satisfaite. Dans le cas d'enquêtes probabilistes complexes, il est peu probable que π_i soit observé dans l'échantillon non probabiliste. Dans ce cas, il faut le traiter comme étant aléatoire.

En utilisant le cadre probabiliste introduit par Savitsky et coll. (2022), également décrit à la section 3, on peut facilement démontrer que

$$\rho_i = \Pr\left(z_i = 1 \middle| i \in s^*, \mathbf{x}_i\right) = \frac{\Pr\left(i \in s_{\text{NP}} \middle| \mathbf{x}_i\right)}{\Pr\left(i \in s^* \middle| \mathbf{x}_i\right)} = \frac{p_i}{(\tilde{\pi}_i + p_i)}.$$

Mentionnons que la relation n'exige pas de constante de proportionnalité $(K = 1)$ et qu'elle est valide, quelle que soit la taille des fractions de sondage. Quand le modèle logistique $\rho_i(\boldsymbol{\alpha}) = \left[1 + \exp\left(-\mathbf{x}_i'\boldsymbol{\alpha}\right)\right]^{-1}$ est utilisé, il est facile de voir que le modèle implicite qui en résulte pour p_i est $p_i(\boldsymbol{\alpha}) = \tilde{\pi}_i \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)$, qui admet des estimations supérieures à 1. D'autres modèles pour p_i peuvent être envisagés, comme le modèle logistique. Beresovsky (2019), voir aussi Gershunskaya et Lahiri (2023), a appelé cette méthode régression logistique implicite, qui est essentiellement la méthode d'Elliott avec un modèle logistique pour p_i et donne un modèle implicite pour ρ_i.

On dérive le logarithme d'une fonction de vraisemblance en supposant que z_i, $i \in s^*$, sont mutuellement indépendants étant donné $\mathbf{x}_i$, $i \in s^*$. Pour un modèle paramétrique arbitraire pour p_i, la fonction d'estimation d'Elliott, ou de la méthode de régression logistique implicite, qui en résulte est

$$\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-RLI}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\tilde{\pi}_k}{\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) - \sum\nolimits_{k \in s_P} \frac{1}{\tilde{\pi}_k} \frac{\tilde{\pi}_k}{\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}). \tag{4.1}$$

La fonction d'estimation (4.1) est sans biais par rapport à ξmd. Si $\tilde{\pi}_k$ est remplacé par π_k dans (4.1), la fonction d'estimation qui en résulte est désignée par $\hat{\mathbf{U}}_{\pi}^{E\text{-RLI}}(\boldsymbol{\alpha})$. Elle est sans biais par rapport à md et sans biais par rapport à ξmd, à condition que l'hypothèse (A2) soit satisfaite. La fonction d'estimation (4.1) a une forme semblable à la fonction d'estimation optimale $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ donnée par (2.8), où π_k est remplacée par $\tilde{\pi}_k$. On s'attend à ce que $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ et $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-RLI}}(\boldsymbol{\alpha})$ soient à peu près équivalentes en général, sauf dans les scénarios où la fraction de sondage dans les deux échantillons est grande et où le chevauchement n'est pas petit. Il n'est donc pas surprenant que la fonction d'estimation (4.1) ait obtenu de meilleurs résultats que les fonctions d'estimation CLW et WVL dans l'étude par simulations de Savitsky et coll. (2022).

Les deux problèmes indiqués dans la section 3 à propos de la méthode WVL (ou de pseudo-régression logistique implicite) s'appliquent également à la méthode d'Elliott (ou de régression logistique implicite). La fonction d'estimation (4.1) n'a pas la propriété de vraisemblance du recensement, puisqu'elle ne se réduit pas à la fonction d'estimation (2.1) de vraisemblance de la population quand l'échantillon probabiliste est

un recensement. En effet, elle se réduit à la fonction d'estimation (3.1) de WVL, ou de la méthode de pseudo-régression logistique implicite.

De plus, l'hypothèse que $z_i, i \in s^*$, sont mutuellement indépendants étant donné $\mathbf{x}_i, i \in s^*$, n'est pas valide puisque si l'on utilise le cadre de Savitsky et coll. (2022) ainsi que le fractionnement aléatoire de U^* décrit dans la section 3,

$$\Pr\left(z_i = 1, z_j = 1 \middle| i, j \in s^*, \mathbf{x}_i, \mathbf{x}_j\right) = 0 \neq \frac{p_i}{(\tilde{\pi}_i + p_i)} \frac{p_j}{(\tilde{\pi}_j + p_j)},$$

pour toute paire d'éléments $i \in U_1$ et $j \in U_2$. Pour deux éléments différents i et j dans la même population, soit U_1 ou U_2, nous avons

$$\Pr\left(z_i = 1, z_j = 1 \middle| i, j \in s^*, \mathbf{x}_i, \mathbf{x}_j\right) = \frac{p_i p_j}{\tilde{\pi}_i \tilde{\pi}_j + p_i p_j} \neq \frac{p_i}{(\tilde{\pi}_i + p_i)} \frac{p_j}{(\tilde{\pi}_j + p_j)},$$

quand (A1) est satisfaite pour les éléments $i \in U_{\text{NP}}$ ainsi qu'(A3) et (A5) pour les éléments $i \in U_P$. Même sous ces hypothèses, l'indépendance mutuelle de $z_i, i \in s^*$, n'est pas soutenable sauf si bon nombre des p_i sont petits, et que par conséquent le chevauchement est une portion négligeable de l'échantillon probabiliste. Cette condition semble raisonnablement satisfaite dans l'étude par simulations de Dr. Gershunskaya et Dr. Beresovsky et peut expliquer les bonnes performances du AIC pour la méthode de régression logistique implicite. En principe, un AIC qui repose sur le logarithme d'une fonction de vraisemblance incorrecte n'est pas valide et peut ne pas être efficace pour la sélection de variables.

5. Méthode de vraisemblance de l'échantillon

5.1 Chevauchement connu entre les deux échantillons

Une fonction de vraisemblance de l'échantillon (par exemple Pfeffermann, Krieger et Rinott, 1998) dans le scénario d'intégration de données étudié dans notre article est une fonction de vraisemblance fondée sur les observations des unités de l'échantillon $k \in s_{\text{NP}} \cup s_P$. Supposons d'abord que nous avons accès à $\mathbf{X}_s = \{\mathbf{x}_k; k \in s_{\text{NP}} \cup s_P\}$ en plus de $\{\mathbf{x}_k; k \in s_{\text{NP}}\}$. Ces deux ensembles de données n'ont pas besoin d'être couplés, mais on doit connaître le chevauchement entre les échantillons probabiliste et non probabiliste pour créer $\mathbf{X}_s$ à partir des données auxiliaires des deux échantillons. Nous supposons donc que soit $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ ou $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$ est connu. Cette hypothèse sera assouplie dans la section 5.2.

Sous les hypothèses (A2) et (A4), il est facile de montrer que la probabilité de participation étant donné $k \in s_{\text{NP}} \cup s_P$ est

$$p_{s,k} = \Pr\left(\delta_k = 1 \middle| k \in s_{\text{NP}} \cup s_P, \mathbf{x}_k\right) = \frac{\Pr\left(k \in s_{\text{NP}} \middle| \mathbf{x}_k\right)}{\Pr\left(k \in s_{\text{NP}} \cup s_P \middle| \mathbf{x}_k\right)} = \frac{p_k}{p_k + \tilde{\pi}_k - \tilde{\pi}_k p_k}. \tag{5.1}$$

Cette probabilité de participation conditionnelle se réduit à $p_{s,k} \approx p_k / (p_k + \tilde{\pi}_k)$ quand le chevauchement est négligeable, ce qui est l'hypothèse implicite posée par Elliott (2009). Mentionnons que nos hypothèses (A2) et (A4) n'impliquent pas nécessairement un chevauchement négligeable, en particulier quand les fractions de sondage sont grandes.

Sous les hypothèses d'indépendance (A1), (A3) et (A5), $\left(I_k, \delta_k\right),\ k \in s_{\text{NP}} \cup s_P$ sont mutuellement indépendants étant donné $\mathbf{x}_k,\ k \in s_{\text{NP}} \cup s_P.$ En supposant un modèle paramétrique pour $p_k,$ on peut écrire la fonction de vraisemblance de l'échantillon comme suit :

$$L_s(\boldsymbol{\alpha}) = \prod_{k \in s_{\text{NP}} \cup s_P} \left[p_{s,k}(\boldsymbol{\alpha})\right]^{\delta_k} \left[1 - p_{s,k}(\boldsymbol{\alpha})\right]^{(1-\delta_k)}, \tag{5.2}$$

où $p_{s,k}(\boldsymbol{\alpha})$ est donné par (5.1) avec $p_k = p_k(\boldsymbol{\alpha}).$ Si l'échantillonnage de Poisson n'est pas utilisé pour la sélection de l'échantillon probabiliste, l'hypothèse (A3) n'est pas satisfaite. Il reste à vérifier si la fonction de vraisemblance de l'échantillon (5.2) serait approximativement valide pour les plans de sondage utilisés dans la pratique au-delà de l'échantillonnage de Poisson. Dans un contexte où seules les données de l'échantillon probabiliste sont modélisées, Pfeffermann, Krieger et Rinott (1998) ont montré l'indépendance asymptotique des observations de l'échantillon pour les plans de sondage courants à condition que les observations de la population soient indépendantes. Il est possible qu'un résultat semblable existe également dans le contexte de l'intégration de données d'un échantillon probabiliste et non probabiliste.

En utilisant (5.2) et en réorganisant les termes, nous obtenons le logarithme de la fonction de vraisemblance de l'échantillon

$$l_s(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\left[p_{s,k}(\boldsymbol{\alpha})\right] + \sum_{k \in s_P} \log\left[1 - p_{s,k}(\boldsymbol{\alpha})\right] - \sum_{k \in s_{\text{NP}} \cap s_P} \log\left[1 - p_{s,k}(\boldsymbol{\alpha})\right]. \tag{5.3}$$

En prenant la dérivée de $l_s(\boldsymbol{\alpha})$ par rapport à $\boldsymbol{\alpha},$ nous obtenons par des calculs algébriques simples, la fonction d'estimation de vraisemblance de l'échantillon

$$\mathbf{U}_s(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})} \mathbf{g}_k(\boldsymbol{\alpha}) - \sum_{k \in s_P} \frac{1}{\tilde{\pi}_k} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) + \boldsymbol{\Psi}(\boldsymbol{\alpha}), \tag{5.4}$$

où

$$\boldsymbol{\Psi}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}} \cup s_P} I_k \delta_k \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} = \sum_{k \in s_{\text{NP}}} I_k \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} = \sum_{k \in s_P} \delta_k \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]}. \tag{5.5}$$

La fonction d'estimation (5.4) satisfait la propriété de vraisemblance du recensement et sous les hypothèses (A2) et (A4), est sans biais par rapport à ξmd conditionnellement à $\mathbf{X}_s.$ À partir de (5.5), nous constatons que l'utilisation de la fonction d'estimation (5.4) nécessite de connaître le chevauchement seulement dans un des deux échantillons, c'est-à-dire qu'il suffit d'observer soit $\left\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\right\}$ ou $\left\{\delta_k, \mathbf{x}_k; k \in s_P\right\}.$ Cette information peut être obtenue au moyen de questions supplémentaires dans l'enquête probabiliste ou non probabiliste, ou par couplage d'enregistrements, les variables auxiliaires étant des

variables d'appariement possibles. Par exemple, si le vecteur $\mathbf{x}_k$ est distinct pour chaque unité de la population $k \in U$ (par exemple s'il y a au moins une variable auxiliaire continue), on peut connaître $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ et $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$ en appariant chaque unité d'un échantillon à toutes les unités de l'autre échantillon. Autrement dit, si le vecteur $\mathbf{x}_k$ pour une unité $k \in s_P$ est identique au vecteur $\mathbf{x}_l$ d'une unité $l \in s_{\text{NP}}$, nous savons alors que $\delta_k = 1$ (et $I_l = 1$). Sinon, en l'absence d'appariement avec $\mathbf{x}_k$, alors $\delta_k = 0$. Cet appariement peut être répété pour chaque unité $k \in s_P$ afin de déterminer le chevauchement complet $\{\delta_k, \mathbf{x}_k; k \in s_P\}$. Une procédure similaire peut servir à identifier $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$. Si l'on dispose de suffisamment d'informations pour mettre en œuvre la fonction d'estimation (5.4), on peut alors utiliser le AIC classique pour la sélection de variables en utilisant le logarithme de la fonction de vraisemblance de l'échantillon (5.3), c'est-à-dire $\text{AIC} = -2l_s(\hat{\boldsymbol{\alpha}}_s) + 2q$, où $\hat{\boldsymbol{\alpha}}_s$ est la solution de $\mathbf{U}_s(\boldsymbol{\alpha}) = \mathbf{0}$ et q est le nombre de paramètres du modèle. Cette solution est idéale si les unités qui se chevauchent peuvent être déterminées avec précision. Kim et Kwon (2024) ont proposé de façon indépendante une approche non conditionnelle par modèle du score de propension qui semble être très semblable à notre approche de vraisemblance de l'échantillon.

5.2 Chevauchement inconnu entre les deux échantillons

En pratique, il se peut que nous n'observions ni $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ ni $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$. Une façon de régler ce problème consiste à appliquer directement le principe de l'information manquante. Pour cela, il faut remplacer la fonction d'estimation non observée (5.4) par son espérance conditionnelle aux données observées, $\mathbf{X}_{\text{obs}} = \{\{\mathbf{x}_k; k \in s_{\text{NP}}\}, \{\mathbf{x}_k; k \in s_P\}\} \neq \mathbf{X}_s$. Cela donne la fonction d'estimation

$$\begin{aligned} E_{\xi md}\left(\mathbf{U}_s(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right) &= \sum_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})} \\ &\quad - \sum_{k \in s_P} \frac{1}{\tilde{\pi}_k} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})[1 - p_k(\boldsymbol{\alpha})]} + E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right). \end{aligned} \tag{5.6}$$

En utilisant le dernier terme du côté droit de (5.5), on peut réécrire l'espérance $E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right)$ dans (5.6), qui est le meilleur prédicteur de $\boldsymbol{\Psi}(\boldsymbol{\alpha})$, comme suit :

$$E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right) = \sum_{k \in s_P} p_k^{\text{obs}} \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})[1 - p_k(\boldsymbol{\alpha})]},$$

où $p_k^{\text{obs}} = E_{\xi md}\left(\delta_k \middle| \mathbf{X}_{\text{obs}}\right)$, $k \in s_P$. Sous nos hypothèses, on peut montrer que $p_k^{\text{obs}} = E_{\xi md}\left(\delta_k \middle| n_k^{\text{NP}}\right) = n_k^{\text{NP}} / N_k$, où n_k^{NP}, $k \in s_P$, est le nombre d'unités $l \in s_{\text{NP}}$ pour lesquelles $\mathbf{x}_l = \mathbf{x}_k$, et N_k, $k \in s_P$, est le nombre d'unités $l \in U$ pour lesquelles $\mathbf{x}_l = \mathbf{x}_k$. Ce résultat est obtenu en notant que n_k^{NP} obéit à une loi binomiale avec le nombre d'essais N_k et la probabilité $p_k(\boldsymbol{\alpha})$. L'application du principe de l'information manquante dans ce contexte nécessite de connaître $N_k, k \in s_P$. Cette information est généralement inconnue, mais si nous pouvons supposer que les vecteurs de population $\mathbf{x}_k$ sont tous distincts (c'est-à-dire $N_k = 1, k \in U$), nous pouvons entièrement déterminer le chevauchement, comme cela est expliqué plus

haut, et donc $p_k^{\text{obs}} = \delta_k,\ k \in s_P,$ et $E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right) = \boldsymbol{\Psi}(\boldsymbol{\alpha}).$ Dans d'autres cas moins triviaux, N_k pourrait être modélisé au moyen, par exemple, de la distribution de Poisson.

Comme solution de rechange simple à la modélisation de $N_k,$ dans des situations où le chevauchement ne peut être déterminé dans aucun des échantillons, nous proposons de remplacer $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ dans (5.4) par son meilleur prédicteur linéaire sans biais. Cette approche peut entraîner une légère réduction de l'efficacité par rapport au meilleur prédicteur $E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right),$ mais au moins la solution ne dépend pas des valeurs inconnues $N_k,\ k \in s_P,$ comme cela est illustré ci-dessous.

Nous considérons le prédicteur linéaire sans biais suivant de $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ qui utilise les données auxiliaires disponibles des deux échantillons :

$$\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \lambda_k \tilde{\pi}_k \frac{p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} + \sum_{k \in s_P} (1 - \lambda_k) p_k(\boldsymbol{\alpha}) \frac{p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]}, \tag{5.7}$$

où $\lambda_k,\ k \in s_{\text{NP}} \cup s_P,$ sont des constantes. L'estimateur (5.7) est conditionnellement sans biais en ce sens que $E_{\xi md}\left(\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) - \boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_s\right) = \mathbf{0},$ à condition que les hypothèses (A2) et (A4) soient satisfaites. En remplaçant $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ dans (5.4) par le terme du côté droit de (5.7), nous obtenons la fonction d'estimation

$$\begin{aligned} \hat{\mathbf{U}}_s^{\lambda}(\boldsymbol{\alpha}) \quad &= \sum_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})}\left(1 + \lambda_k \frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right) \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})} \\ &\quad - \sum_{k \in s_P} \frac{1}{\tilde{\pi}_k}\left(1 + \lambda_k \frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right) \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})}. \end{aligned} \tag{5.8}$$

Elle est sans biais par rapport à ξmd conditionnellement à $\mathbf{X}_s.$

On obtient le meilleur prédicteur linéaire sans biais de $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ en déterminant $\lambda_k,\ k \in s_{\text{NP}} \cup s_P,$ qui minimise la variance de prédiction $\text{var}_{\xi md}\left(\mathbf{c}'\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) - \mathbf{c}'\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_s\right).$ Sous nos trois hypothèses d'indépendance, cette variance de prédiction est minimisée quand $\text{var}_{\xi md}(\lambda_k \tilde{\pi}_k \delta_k + (1 - \lambda_k) p_k(\boldsymbol{\alpha}) I_k - I_k \delta_k \,|k \in\ s_{\text{NP}} \cup s_P, \mathbf{x}_k)$ est minimisé pour chaque $k \in s_{\text{NP}} \cup s_P.$ La constante λ_k est donc déterminée de façon à ce que $\lambda_k \tilde{\pi}_k \delta_k + (1 - \lambda_k) p_k(\boldsymbol{\alpha}) I_k$ prédise le plus précisément possible $I_k \delta_k,$ c'est-à-dire si l'unité k se trouve dans l'intersection des deux échantillons ou pas. En ajoutant les hypothèses (A2) et (A4), on peut montrer, après des calculs simples, que la valeur de λ_k qui minimise la variance de prédiction est $\lambda_k = \lambda_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha}) = 1 - \gamma_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha}),$ $k \in s_{\text{NP}} \cup s_P,$ où $\gamma_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$ est donné par (2.7) après remplacement de π_k par $\tilde{\pi}_k.$ Si l'on utilise $\lambda_k = \lambda_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$ dans (5.8), on constate que la fonction d'estimation (5.8) se réduit exactement à la fonction d'estimation optimale $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha}),$ qui est donnée par (2.8) avec π_k remplacé par $\tilde{\pi}_k.$ Il est donc intéressant de voir que l'utilisation de la théorie de la meilleure estimation linéaire sans biais dans une méthode de vraisemblance de la population (voir la section 2) équivaut à l'utilisation de la théorie de la meilleure prédiction linéaire sans biais dans une méthode de vraisemblance de l'échantillon.

Si la probabilité de sélection π_k est observée pour toutes les unités des échantillons probabiliste et non probabiliste, elle peut et doit être considérée comme une variable auxiliaire potentielle à inclure dans $\mathbf{x}_k.$

Si la probabilité π_k est incluse dans $\mathbf{x}_k$, $\tilde{\pi}_k = \pi_k$ et on peut donc utiliser indifféremment π_k ou $\tilde{\pi}_k$ dans (2.8). Si π_k n'est pas incluse dans $\mathbf{x}_k$, car elle ne semble pas expliquer δ_k après avoir conditionné sur $\mathbf{x}_k$, alors la théorie ci-dessus reste valide et $\tilde{\pi}_k$ peut être utilisée dans (2.8) comme si π_k était inconnue. Il serait également possible de conditionner à la fois sur π_k et $\mathbf{x}_k$, ce qui entraînerait le remplacement de $\tilde{\pi}_k$ par π_k dans les développements ci-dessus. La fonction d'estimation optimale (2.8) serait sans biais par rapport à md conditionnellement à $\mathbf{X}_s$ (et inconditionnellement) si l'hypothèse (A2) est satisfaite.

5.3 Sélection des variables

La fonction d'estimation (2.8) n'est pas la dérivée du logarithme d'une fonction de vraisemblance de l'échantillon. Par conséquent, le AIC classique n'est pas applicable. Il faudrait approfondir la recherche sur la sélection de variables quand on utilise $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$ ou $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ pour l'estimation des probabilités de participation. Cependant, si un grand nombre de p_k sont petits, le chevauchement est une portion négligeable de l'échantillon probabiliste et la fonction d'estimation de vraisemblance de l'échantillon (5.4) devient approximativement équivalente à la fonction d'estimation $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$. Par conséquent, le logarithme de la fonction de vraisemblance de l'échantillon (5.3), si l'on omet le terme d'intersection négligeable, de même que $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ peuvent être utilisés pour calculer le AIC classique et sélectionner les variables auxiliaires pertinentes. Il semble que ce soit semblable au AIC utilisé par Dr. Gershunskaya et Dr. Beresovsky dans leur étude par simulations pour la méthode de régression logistique implicite, si ce n'est l'utilisation de la fonction d'estimation $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-RLI}}(\boldsymbol{\alpha})$ donnée dans (4.1). Les deux fonctions $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-RLI}}(\boldsymbol{\alpha})$ et $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ devraient être semblables quand le chevauchement est négligeable. Il est rassurant d'observer que leur AIC s'est montré performant dans leur étude par simulations. Nous nous attendons à ce que cette performance se détériore à mesure que la taille de l'échantillon non probabiliste augmente et que le chevauchement devient non négligeable.

6. Fonction d'estimation unifiée

Continuons avec le scénario réaliste dans lequel ni $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ ni $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$ n'est connu. Nous avons décrit plusieurs méthodes pour ce scénario dans les sections précédentes, qui ont donné différentes fonctions d'estimation. En supposant qu'on utilise $\tilde{\pi}_k$ plutôt que π_k, elles sont toutes des cas particuliers de la fonction d'estimation générale

$$\hat{\mathbf{U}}_{\tilde{\pi}}^{h}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right] \mathbf{g}_k(\boldsymbol{\alpha}) - \sum_{k \in s_P} \tilde{w}_k h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right] \mathbf{g}_k(\boldsymbol{\alpha}), \tag{6.1}$$

où $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ est une fonction qui dépend de la méthode. Le tableau 6.1 fournit l'expression de $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ pour les méthodes décrites dans les sections précédentes.

Tableau 6.1
Expression de $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ pour différentes méthodes.

Méthode	Fonction d'estimation	$h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$
CLW	$\hat{\mathbf{U}}_{\tilde{\pi}}(\boldsymbol{\alpha})$: (2.3)*	$\left[1-p_k(\boldsymbol{\alpha})\right]^{-1}$
WVL/pseudo-régression logistique implicite	$\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{WVL-PRLI}}(\boldsymbol{\alpha})$: (3.1)*	$\left[1+p_k(\boldsymbol{\alpha})\right]^{-1}$
Elliott/régression logistique implicite	$\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-RLI}}(\boldsymbol{\alpha})$: (4.1)	$\tilde{\pi}_k\left[\tilde{\pi}_k+p_k(\boldsymbol{\alpha})\right]^{-1}$
Meilleure estimation/prédiction linéaire sans biais	$\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$: (2.8)**	$\tilde{\pi}_k\left[\tilde{\pi}_k+p_k(\boldsymbol{\alpha})-2\tilde{\pi}_k p_k(\boldsymbol{\alpha})\right]^{-1}$

* w_k est remplacé par $\tilde{w}_k$ dans (2.3) et (3.1).
** π_k est remplacé par $\tilde{\pi}_k$ dans (2.8).

Certains auteurs (par exemple Beaumont, 2020; Chen, Li et Wu, 2020; et Rao, 2021) ont étudié la fonction d'estimation par calage

$$\hat{\mathbf{U}}_{\pi}^{\text{cal}}(\boldsymbol{\alpha})=\sum_{k\in s_{\text{NP}}}\frac{1}{p_k(\boldsymbol{\alpha})}\mathbf{x}_k-\sum_{k\in s_P}w_k\mathbf{x}_k.$$

Sa version lissée $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{cal}}(\boldsymbol{\alpha})$, obtenue au moyen du remplacement de w_k par $\tilde{w}_k$ dans l'équation ci-dessus, est également un cas particulier de (6.1). À titre d'exemple, si un modèle logistique pour $p_k(\boldsymbol{\alpha})$ est utilisé, la fonction d'estimation (6.1) se réduit à $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{cal}}(\boldsymbol{\alpha})$ quand $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]=\left\{p_k(\boldsymbol{\alpha})\left[1-p_k(\boldsymbol{\alpha})\right]\right\}^{-1}$. La fonction d'estimation par calage n'a pas la propriété de vraisemblance du recensement, mais a une propriété implicite de double robustesse quand un modèle linéaire entre les variables d'enquête et les variables auxiliaires est valide. Elle pourrait également être facilement généralisée au scénario où différentes variables auxiliaires sont disponibles dans différents échantillons probabilistes, à condition que toutes les variables auxiliaires soient observées dans l'échantillon non probabiliste. La fonction d'estimation par calage $\hat{\mathbf{U}}_{\pi}^{\text{cal}}(\boldsymbol{\alpha})$ est le cas particulier de (2.5) avec $\gamma_k=\gamma_k^{\text{CAL}}(\boldsymbol{\alpha})=-\left[1-p_k(\boldsymbol{\alpha})\right]/p_k(\boldsymbol{\alpha})<0$. On s'attend par conséquent à ce qu'elle soit inefficace pour l'estimation de $p_k(\boldsymbol{\alpha})$.

La fonction d'estimation (6.1) est sans biais par rapport à ξmd, à la fois inconditionnellement et conditionnellement à $\mathbf{X}_s$. La probabilité $\tilde{\pi}_k$ dans (6.1) peut aussi être remplacée par π_k si elle est disponible dans l'échantillon non probabiliste. La fonction d'estimation (6.1) reste (conditionnellement) sans biais par rapport à ξmd si l'hypothèse (A2) est satisfaite (par exemple si π_k est incluse dans $\mathbf{x}_k$). Si π_k, $k\in U$ sont traitées comme étant fixes, (6.1) est également (conditionnellement) sans biais par rapport à md sous l'hypothèse (A2). Une fonction d'estimation hybride qui n'exige pas la disponibilité de π_k dans l'échantillon non probabiliste est :

$$\hat{\mathbf{U}}_{\pi,\tilde{\pi}}^{h}(\boldsymbol{\alpha})=\sum_{k\in s_{\text{NP}}}\frac{1}{p_k(\boldsymbol{\alpha})}h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]\mathbf{g}_k(\boldsymbol{\alpha})-\sum_{k\in s_P}w_k h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]\mathbf{g}_k(\boldsymbol{\alpha}). \tag{6.2}$$

Elle est (conditionnellement) sans biais par rapport à ξmd sans nécessiter la validité d'un modèle pour π_k, mais il se peut qu'elle soit moins efficace que (6.1) en raison de la variabilité des poids de sondage probabiliste w_k.

En pratique, que (6.1) ou (6.2) soit utilisé, $\tilde{\pi}_k$ est inconnu et il faut l'estimer. Comme souligné dans la section 2, l'échantillon probabiliste peut servir à estimer $\tilde{w}_k = E_\xi\left(w_k \,|k \in s_P, \mathbf{x}_k\right)$ par $\hat{\tilde{w}}_k$, peut-être au moyen de méthodes non paramétriques, comme des méthodes d'apprentissage automatique. En utilisant la relation (2.9), $\tilde{\pi}_k$ est estimé par $\hat{\tilde{\pi}}_k = 1/\hat{\tilde{w}}_k$. Mentionnons que $\tilde{\pi}_k = E_\xi\left(\pi_k \,|\mathbf{x}_k\right)$ ne peut pas être estimé en modélisant $E_\xi\left(\pi_k | k \in s_P, \mathbf{x}_k\right)$ et en ignorant le plan de sondage probabiliste, comme cela est parfois proposé dans la littérature (par exemple Elliott, 2009; Elliott et Valliant, 2017). Cela s'explique par le fait que le plan de sondage probabiliste est (fortement) informatif par rapport à la distribution de π_k étant donné $\mathbf{x}_k$.

Examinons maintenant le modèle des groupes homogènes pour lequel les variables auxiliaires partitionnent la population en G groupes et $p_k(\boldsymbol{\alpha}) = p_g$ pour une unité k dans le groupe g. Le poids lissé pour une unité k dans le groupe g est $\tilde{w}_g = E_\xi\left(w_k | k \in s_{P,g}\right)$, où $s_{P,g}$ est l'ensemble des unités de l'échantillon probabiliste qui se trouvent dans le groupe g. On peut l'estimer simplement par la moyenne des poids dans le groupe g, c'est-à-dire $\hat{\tilde{w}}_g = \hat{N}_g / n_g^P$, où $\hat{N}_g = \sum_{k \in s_{P,g}} w_k$ et n_g^P est la taille de l'échantillon probabiliste dans le groupe g. Pour une unité k dans le groupe g, $\hat{\tilde{\pi}}_k = \hat{\tilde{\pi}}_g = n_g^P / \hat{N}_g$. En remplaçant $\tilde{\pi}_k$ par $\hat{\tilde{\pi}}_k$ dans (6.1) ou (6.2) et en résolvant les équations d'estimation (soit $\hat{\mathbf{U}}^h_{\tilde{\pi}}(\boldsymbol{\alpha}) = \mathbf{0}$ ou $\hat{\mathbf{U}}^h_{\pi,\tilde{\pi}}(\boldsymbol{\alpha}) = \mathbf{0}$) pour n'importe quel choix de $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$, on obtient $\hat{p}_g = n_g^{\mathrm{NP}} / \hat{N}_g$, l'estimation de p_g, où n_g^{NP} est la taille de l'échantillon non probabiliste dans le groupe g. Si on remplace $\tilde{\pi}_k$ par π_k dans (6.1) ou (6.2) et que $h\left[\pi_k, p_k(\boldsymbol{\alpha})\right]$ dépend à la fois de π_k et $p_k(\boldsymbol{\alpha})$, la probabilité de participation estimée dans le groupe g n'est plus $\hat{p}_g = n_g^{\mathrm{NP}} / \hat{N}_g$ et n'a pas de forme analytique.

Mentionnons que $\hat{p}_g = n_g^{\mathrm{NP}} / \hat{N}_g$ peut être supérieur à 1. Cela est plus susceptible de se produire pour les grands échantillons non probabilistes et les petits échantillons probabilistes. Pour un vecteur général $\mathbf{x}_k$, cela peut laisser supposer qu'une solution existerait moins fréquemment avec le modèle logistique ($p_k(\hat{\boldsymbol{\alpha}})$ borné par 1) qu'avec le modèle exponentiel ($p_k(\hat{\boldsymbol{\alpha}})$ non borné). Toutefois, le modèle exponentiel peut ne pas être précis pour les grands échantillons non probabilistes.

Pour le modèle des groupes homogènes, la solution de l'équation d'estimation de vraisemblance de l'échantillon $\mathbf{U}_s(\boldsymbol{\alpha}) = \mathbf{0}$, où $\mathbf{U}_s(\boldsymbol{\alpha})$ est donné dans (5.4), donne

$$\hat{p}_g^{\mathrm{VE}} = \frac{\hat{p}_g}{\hat{p}_g + \left(1 - \dfrac{n_g^I}{n_g^P}\right)}$$

comme étant l'estimation de p_g, où n_g^I est le nombre d'unités dans l'intersection $s_{\mathrm{NP}} \cap s_P$ qui se trouvent dans le groupe g et $\hat{p}_g = n_g^{\mathrm{NP}} / \hat{N}_g$. Comme on s'y attend, $\hat{p}_g^{\mathrm{VE}}$ est proche de $\hat{p}_g$ quand $\hat{p}_g$ et le taux de chevauchement n_g^I / n_g^P sont petits. L'estimation de vraisemblance de l'échantillon $\hat{p}_g^{\mathrm{VE}}$ ne peut pas être supérieure à 1, contrairement à $\hat{p}_g$. Il s'agit d'une propriété souhaitable de la fonction d'estimation de vraisemblance de l'échantillon (5.4), qui résulte de l'exploitation de l'information sur le chevauchement entre les deux échantillons.

7. Conclusion

Dans les sections qui précèdent, nous avons décrit trois méthodes de vraisemblance pour l'estimation des probabilités de participation et la sélection de variables auxiliaires pertinentes qui sont valides quelles que soient la taille des échantillons probabiliste et non probabiliste et celle du chevauchement entre les deux échantillons. Il s'agit des méthodes de vraisemblance de la population et de pseudo-vraisemblance, décrites à la section 2, et de la méthode de vraisemblance de l'échantillon, décrite à la section 5. Si l'échantillon probabiliste est un recensement, la méthode de vraisemblance de la population est la plus efficace et doit être privilégiée. Si l'échantillon probabiliste n'est pas un recensement, mais que le chevauchement est connu dans au moins un des deux échantillons, la méthode de vraisemblance de l'échantillon doit être préférée à la méthode de pseudo-vraisemblance en raison de sa plus grande efficacité. Si le chevauchement est inconnu, la méthode de pseudo-vraisemblance de Chen, Li et Wu (2020) peut être utilisée à la fois pour l'estimation des probabilités de participation et le calcul d'un AIC aux fins de sélection des variables. Cependant, la fonction d'estimation CLW (2.3) peut ne pas être efficace, surtout quand l'échantillon non probabiliste est plus grand que l'échantillon probabiliste, parce qu'elle ne tire pas pleinement parti des données auxiliaires disponibles. Notre fonction d'estimation optimale (2.8), $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$, ou sa version lissée $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$, devrait être plus efficace que les solutions existantes, bien qu'il nous reste à le démontrer par une étude empirique. Il faudrait également approfondir la recherche sur la sélection de variables en cas de chevauchement inconnu, quand $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$ ou $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ est utilisé, sauf dans le cas où un grand nombre des probabilités de participation sont petites et que le chevauchement peut être négligé. Dans ce cas, on peut utiliser le logarithme de la fonction de vraisemblance de l'échantillon (5.3), en ignorant le terme de chevauchement, ainsi que $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ pour calculer le AIC classique.

En pratique, il est rare que les probabilités de participation estimées à partir d'un modèle paramétrique servent directement à calculer les estimations des paramètres de population finie. Souvent, on crée des groupes homogènes par rapport à ces probabilités estimées pour protéger contre les spécifications erronées du modèle et les poids extrêmes. Il est possible que le choix d'une fonction d'estimation n'ait pas d'effet majeur sur les estimations des paramètres de population finie si l'on utilise des groupes homogènes avant de calculer ces estimations. Néanmoins, il semble raisonnable de choisir la fonction d'estimation la plus efficace pour l'estimation des probabilités de participation avant de créer les groupes homogènes.

L'estimation non paramétrique des probabilités de participation au moyen, par exemple, de méthodes d'apprentissage automatique pourrait être utile pour protéger contre d'éventuelles spécifications erronées du modèle. Les méthodes d'apprentissage automatique existantes peuvent servir directement à modéliser p_k si $\{\delta_k, \mathbf{x}_k; k \in U\}$ est connu. Autrement, si $\{\delta_k, \mathbf{x}_k; k \in s_{\text{NP}} \cup s_P\}$ est connu, la probabilité conditionnelle $p_{s,k}$ peut être modélisée au moyen de méthodes d'apprentissage automatique existantes et p_k peut ensuite être estimée au moyen de la relation (5.1). Le cas le plus difficile se présente quand le chevauchement est inconnu. Dans notre article principal, nous avons proposé l'algorithme nppCART comme moyen de créer des groupes homogènes et d'obtenir une protection contre les spécifications erronées du modèle. Notre procédure s'inspire de la méthode de pseudo-vraisemblance de Chen, Li et Wu (2020) et ne nécessite pas la

présence d'un chevauchement négligeable entre les deux échantillons. Une solution de rechange consisterait à envisager des méthodes d'apprentissage automatique ainsi que la méthode d'Elliott (ou la méthode de la régression logistique implicite), comme le proposent Elliott et Valliant (2017) et Elliott (2022). L'idée reviendrait à modéliser $\rho_i = \Pr\left(z_i = 1 \middle| i \in s^*, \mathbf{x}_i\right)$ au moyen d'une méthode d'apprentissage automatique, en ignorant le chevauchement et, par conséquent, l'absence d'indépendance entre les observations. Ensuite, p_i serait estimé pour les unités de l'échantillon non probabiliste au moyen de la relation $\rho_i = p_i / (\tilde{\pi}_i + p_i)$, dont on a montré la validité dans la section 4 en utilisant le cadre de Savitsky et coll. (2022). Si la plupart des probabilités de participation sont petites, le chevauchement est une portion négligeable de l'échantillon probabiliste et peut donc être ignoré. Par conséquent, cette méthode est équivalente à la proposition que nous avons faite ci-dessus de modéliser $p_{s,k}$ au moyen d'une méthode d'apprentissage automatique, puis d'utiliser la relation (5.1) pour estimer p_k. Il reste à évaluer les performances de cette version d'apprentissage automatique de la méthode d'Elliott (ou de la méthode de régression logistique implicite) en cas de chevauchement non négligeable.

Dans notre article principal et dans la présente réponse, nous nous sommes concentrés sur l'estimation de la probabilité de participation p_k pour les unités de l'échantillon non probabiliste. Une fois les estimations $\hat{p}_k$, $k \in s_{\text{NP}}$ calculées, elles peuvent servir à estimer des paramètres de population finie, comme des totaux ou des moyennes de population. L'estimateur pondéré par l'inverse de la probabilité des paramètres de population finie consiste à pondérer les unités de l'échantillon non probabiliste par $1/\hat{p}_k$. Bien entendu, certains estimateurs utilisent plus efficacement $\hat{p}_k$, par exemple, en tirant parti d'un modèle pour les variables de l'enquête afin d'obtenir une propriété de double robustesse (par exemple Chen, Li et Wu, 2020; Chambers, Ranjbar, Salvati et Pacini, 2022). L'exemple le plus simple, qui est courant, se présente quand les poids $1/\hat{p}_k$ sont calés sur des totaux de population connus ou estimés de variables auxiliaires. L'estimateur de totaux de population qui en résulte est doublement robuste en ce sens qu'il est valide sous le modèle de participation ou sous un modèle linéaire entre les variables d'enquête et les variables auxiliaires.

Notre point de vue est que les statisticiens d'enquête devraient commencer par obtenir les estimations les plus efficaces possibles de p_k, $k \in s_{\text{NP}}$, avant de les utiliser pour l'estimation de paramètres de population finie. Il s'agit exactement du même point de vue que celui adopté par de nombreux statisticiens d'enquête pour estimer des paramètres de population finie au moyen de données tirées d'un échantillon probabiliste. Ils commencent par leur meilleure estimation possible de la probabilité de sélection dans l'échantillon, π_k, puis ils l'utilisent pour dériver des estimateurs efficaces des paramètres de population finie (par exemple au moyen de techniques de calage). Il se trouve que la probabilité π_k est habituellement connue pour les échantillons probabilistes et qu'il n'est pas nécessaire de l'estimer.

Dans cette dernière remarque, nous aimerions saisir l'occasion pour remercier chaleureusement Prof. Partha Lahiri, rédacteur en chef invité de ce numéro spécial, de tous les efforts qu'il a déployés afin d'organiser cette excellente collection d'articles, avec leur discussion, qui ont été présentés lors de la conférence Morris Hansen Memorial Lecture de 2022.

Bibliographie

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 539-553.

Beaumont, J.-F. (2020). Les Enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? *Techniques d'enquête*, 46, 1, 1-30. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2020001/article/00001-fra.pdf.

Beresovsky, V. (2019). On application of a response propensity model to estimation from web samples. Dans ResearchGate.

Chambers, R.L. (2023). Le principe de l'information manquante – Un paradigme d'analyse de données désordonnées d'enquête par sondage. *Techniques d'enquête*, 49, 2, 237-278. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023002/article/00018-fra.pdf.

Chambers, R., Ranjbar, S., Salvati, N. et Pacini, B. (2022). Weighting, informativeness and causal inference, with an application to rainfall enhancement. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 185, 1584-1612.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Elliott, M.R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2, 813-845.

Elliott, M.R. (2022). Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste ». *Techniques d'enquête*, 48, 2, 347-358. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022002/article/00004-fra.pdf.

Elliott, M., et Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.

Gershunskaya, J., et Lahiri, P. (2023). Discussion of "Probability vs. nonprobability sampling: From the birth of survey sampling to the present day", by Graham Kalton. *Statistics in Transition New Series*, 24, 3, 31-37.

Kim, J.K., et Kwon, Y. (2024). Commentaires à propos de l'article « Hypothèse de l'échangeabilité dans des méthodes d'ajustement fondées sur le score de propension aux fins d'estimation de la moyenne de population au moyen d'échantillons non probabilistes ». *Techniques d'enquête*, 50, 1, 67-74. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00007-fra.pdf.

Lumley, T., et Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3, 1-18.

Orchard, T., et Woodbury, M.A. (1972). A missing information principle: Theory and application. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697-715.

Pfeffermann, D., Krieger, A.M. et Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Pfeffermann, D., et Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, 61, 166-186.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovsky, V. et Johnson, N.G. (2022). Methods for combining probability and nonprobability samples under unknown overlaps. https://doi.org/10.48550/arXiv.2208.14541.

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.