

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

Estimation du score de propension et plan d'échantillonnage optimal lors de l'intégration d'échantillons probabilistes et de données non probabilistes

par Anders Holmberg, Lyndon Ang, Robert Clark et Bronwyn Loong

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Estimation du score de propension et plan d'échantillonnage optimal lors de l'intégration d'échantillons probabilistes et de données non probabilistes¹

Anders Holmberg, Lyndon Ang, Robert Clark et Bronwyn Loong²

Résumé

Alors que les sources de données non probabilistes ne sont pas une nouveauté dans les statistiques officielles, on constate un intérêt renouvelé à leur égard, que ce soit en raison de la diminution des taux de réponse aux enquêtes, de l'augmentation des coûts de collecte ou du désir de tirer parti des nouvelles sources de données découlant de la numérisation sociétale en cours. En raison de l'exclusion de certains segments de la population cible, l'inférence dérivée uniquement d'une source de données non probabilistes est susceptible d'entraîner des biais. Cette étude explore une solution à ce biais en intégrant des données non probabilistes à des échantillons probabilistes de référence. Des méthodes se concentreront sur la modélisation de la propension à l'inclusion dans l'ensemble de données non probabilistes avec l'aide de l'échantillon de référence connexe; la propension modélisée étant ensuite appliquée à une approche de pondération selon une probabilité inverse pour produire des estimations démographiques. On présume parfois que l'échantillon de référence est donné. Toutefois, dans cette présentation, une stratégie optimale est visée, c'est-à-dire la combinaison d'un estimateur fondé sur l'intégration des données et d'un plan d'échantillonnage pour l'échantillon probabiliste de référence. Des travaux récents sont présentés qui tirent parti des possibilités de bonne identification d'unités des enquêtes-entreprises pour étudier un estimateur fondé sur les propensions et dériver des probabilités de sélection optimales (inégaux) pour l'échantillon de référence.

Mots clés : Intégration des données; biais de sélection; probabilités de sélection optimales; mégadonnées; pondération par l'inverse de la probabilité.

1. Introduction

1.1 Contexte

L'un des objectifs des organismes de statistique nationaux (ONS) est de fournir des statistiques officielles fiables sur la situation économique, démographique, sociale et environnementale pour permettre une prise de décisions éclairée (Division de statistique des Nations Unies, 2014). Les données utilisées par les ONS peuvent provenir de tous les types de sources, selon les niveaux requis de qualité, d'actualité, de coût et de fardeau du répondant.

La numérisation continue de la société a engendré une quantité croissante de données recueillies sur les personnes, les entités et l'environnement. Ces nouvelles sources de données comprennent les mégadonnées, provenant, par exemple, de capteurs, de satellites et de systèmes administratifs, qui ont le potentiel de produire des statistiques plus efficaces. Voici quelques exemples tirés du Bureau de la statistique de l'Australie : les données sur la rémunération des employés soumises par les employeurs au bureau des impôts de l'Australie, qui servent à produire des statistiques sur les salaires et l'emploi (Australian Bureau of Statistics, 2024b, 2024d), des données sur les opérations bancaires pour aider à produire des statistiques sur les dépenses des ménages (Australian Bureau of Statistics, 2024c) et des données sur les contributions agricoles pour produire des statistiques sur la production agricole (Australian Bureau of Statistics, 2024a).

¹ Les points de vue exprimés dans cet article sont ceux des auteurs et ne reflètent pas nécessairement celles du Bureau de la statistique australien. Toute citation ou utilisation doit être clairement attribuée aux auteurs.

² Anders Holmberg, Australian Bureau of Statistics (ABS), Belconnen, ACT, 2617, Australie (anders.holmberg@abs.gov.au); Lyndon Ang, ABS et Australian National University, Canberra, ACT, 2600, Australie; Robert Clark, Australian National University, Canberra, ACT, 2600, Australie; Bronwyn Loong, Australian National University, Canberra, 2600, ACT, Australie

Les mégadonnées sont habituellement générées à des fins non statistiques au moyen d'instruments d'acquisition de données qui ne relèvent pas du contrôle des bureaux de la statistique. Par conséquent, ces données peuvent présenter un biais de sélection lorsque des groupes particuliers de la population cible sont moins susceptibles d'être inclus dans l'ensemble de données. Malgré leur volume, lorsque ces ensembles de données sont utilisés directement sans correction pour estimer les paramètres de population, comme les totaux et les moyennes, ils produisent probablement des estimations biaisées (voir, par exemple, Meng 2018).

Dans cet article, nous discutons des approches de quasi-randomisation pour corriger le biais de sélection dans des données non probabilistes, et en particulier les mégadonnées. Ces méthodes supposent que la légitimité des données non probabilistes est régie par un mécanisme de sélection inconnu, mais découvrable. L'idée est d'utiliser les renseignements auxiliaires disponibles provenant de la population (voir, p. exemple, Burakauskaitė et Čiginas 2023) ou d'un échantillon probabiliste de référence supplémentaire (voir, par exemple, Chen et coll. 2020 et Wang et coll. 2021) pour estimer la propension à la sélection dans l'ensemble de données non probabilistes. L'inverse de ces propensions est ensuite utilisé comme facteur de pondération pour pondérer les données non probabilistes et estimer les effectifs d'une population finie.

On suppose généralement que les échantillons de référence ont déjà été tirés; il peut s'agir d'enquêtes existantes entreprises à d'autres fins, mais qui recueillent les renseignements auxiliaires pertinents. Nous posons la question suivante : « Est-il logique de concevoir l'échantillon de référence plus efficacement pour les données non probabilistes? »

Dans certains cas, il pourrait être possible de déterminer les unités de l'échantillon de référence se trouvant également dans l'ensemble de données non probabilistes. Cela se produirait, par exemple, si un identificateur d'unité commun existait et était disponible à la fois dans l'ensemble de données non probabilistes et dans l'échantillon de référence. Nous explorons les possibilités de tirer parti de la capacité de relier les unités d'échantillonnage de référence et non probabilistes aux étapes du plan d'échantillonnage et de l'estimation.

Voici la structure de la suite de ce document : Nous décrivons la configuration de base de l'étude dans la deuxième section. Dans la troisième section, nous discutons de la répartition optimale de l'échantillon dans le cadre du plan de l'échantillon de référence. La quatrième section présente un estimateur de score de propension tirant parti de la capacité de déterminer les unités de l'échantillon de référence figurant également dans l'ensemble de données non probabilistes. La cinquième section examine le rendement de cet estimateur couplé aux idées de conception de la troisième section au moyen d'une étude de simulation. La sixième section présente quelques observations finales.

2. Configuration de base

Soit U , une population finie de taille N . Pour chaque unité i de la population, nous disposons des valeurs $(\mathbf{x}_i, y_i, \mathbf{z}_i)^T$ pour un élément de données d'intérêt y , quelques éléments de données auxiliaires $\mathbf{x}$ et quelques variables supplémentaires $\mathbf{z}$. Dans la présente étude, nous souhaitons estimer la moyenne de la population $\bar{Y} = \mu_y = N^{-1} \sum_{i=1}^N y_i$.

Supposons que nous disposons d'un échantillon probabiliste A (que nous appelons l'échantillon de « référence ») de taille n_A tirée de la population. A recueille des renseignements pour $\mathbf{x}$, mais pas pour y .³ Nous définissons $\pi_i^A = P(i \in A|U)$ comme étant la probabilité d'inclusion pour l'unité i de se trouver dans l'échantillon probabiliste, et $d_i^A = 1/\pi_i^A$ comme poids d'enquête pour $i \in A$. Les valeurs π_i^A et d_i^A sont connues à partir du plan d'échantillonnage probabiliste et peuvent dépendre de certaines des variables $\mathbf{z}$, par exemple par stratification de la population.

Nous avons également un échantillon non probabiliste B . B contient des renseignements sur $\mathbf{x}$ ainsi que sur y . Soit $\delta_i = I(i \in B)$, une variable indicatrice pour l'unité i incluse dans B . La taille de l'échantillon non probabiliste est $N_B = \sum_{i=1}^N \delta_i$. Contrairement à π_i^A , les probabilités d'inclusion $\pi_i^B = P(\delta_i = 1|U)$ sont inconnues et doivent être

³ L'ABS dispose d'applications de données agricoles (satellites) et comptables pour lesquelles il est impossible de recueillir les mêmes y que dans la source de données non probabilistes.

estimées. Nous définissons $C = U \setminus B$ comme les unités de la population qui ne sont pas incluses dans B . Il peut ou non y avoir chevauchement entre les deux échantillons A et B .

Nous formulons certaines hypothèses concernant le mécanisme de sélection des unités incluses dans B , afin de faciliter la production d'inférences à l'aide de ces ensembles de données; voir par exemple Chen et coll. (2020) et Yang et coll. (2021). Ces hypothèses sont les suivantes :

H1 Ignorabilité : dépend du fait que l'ensemble des covariables $\mathbf{x}_i$, δ_i et y_i sont indépendantes.

H2 Positivité : dépend de $\mathbf{x}_i$, $P(\delta_i = 1|\mathbf{x}_i) > 0, \forall i \in U$.

H3 Indépendance : dépend de $\mathbf{x}_i$ et $\mathbf{x}_j$, $\delta_i \perp \delta_j, i \neq j, \forall i, j \in U$.

L'ignorabilité signifie que $P(\delta_i = 1|\mathbf{x}, y) = P(\delta_i = 1|\mathbf{x})$. En d'autres mots, la sélection dans B est ignorable selon les covariables $\mathbf{x}$. Cette hypothèse est semblable au scénario de données manquantes au hasard (MAR) de Rubin (1976). Andridge et coll. (2019) et Boonstra et coll. (2021) désignent ce type de processus de sélection comme étant une sélection aléatoire (SAR).

L'hypothèse de positivité énonce que selon $\mathbf{x}$, chaque unité de la population a une chance non nulle d'être incluse dans B . Cela ne s'applique pas toujours; par exemple, dans un grand ensemble de données généré par une plateforme de médias sociaux, seules les personnes membres de la plateforme ont une chance d'être incluses dans le grand ensemble de données.

En plus de ce qui précède, nous présumons que nous pouvons coupler les unités dans B soit avec A soit avec la base de population. Cette hypothèse peut être satisfaite s'il existe un identificateur d'unité commun disponible dans B et dans la base de population, ou s'il est possible de coupler de façon probabiliste les enregistrements à l'aide d'un ensemble commun de variables de couplage. À l'occasion, nous assouplissons cette hypothèse afin qu'elle soit suffisante pour déterminer les unités se trouvant dans A qui sont également dans B . Cette hypothèse assouplie peut être satisfaite en incluant une question dans l'instrument d'enquête pour A , par laquelle on demande si le répondant figure également dans l'ensemble de données B .

Nous souhaitons estimer l'inconnu π_i^B , qui dans la présente étude est supposé respecter un modèle paramétrique de sorte que $\pi_i^B = \pi(\mathbf{x}_i, \boldsymbol{\theta}_0)$, où $\boldsymbol{\theta}_0$ sont les valeurs réelles des paramètres du modèle inconnu. Une fois les valeurs finales $\hat{\boldsymbol{\theta}}$ obtenues, on peut calculer $\hat{\pi}_i^B$ et l'intégrer dans un estimateur pondéré selon la probabilité inversée (IPW), tel que l'estimateur de type Hájek (Chen et coll. 2020)

$$\hat{\mu}_{IPW} = \frac{1}{N_B} \sum_{i \in B} \frac{y_i}{\hat{\pi}_i^B} \quad (1)$$

Les estimations de la variance peuvent être calculées pour $\hat{\mu}_{IPW}$, par exemple, en utilisant le cadre asymptotique élaboré par Chen et coll. (2020).

3. Plan de sondage optimal pour l'échantillon de référence

3.1 Réimaginer l'échantillon de référence comme supplément pour corriger le biais de sélection

L'objectif habituel d'une enquête probabiliste consiste à recueillir des renseignements sur un ou des sujets précis d'intérêt pour le statisticien de l'enquête (voir, p. exemple, Cochran, 1977, chapitre 1). Elle sera généralement conçue pour fournir des estimations efficaces pour (un sous-ensemble) de variables clés présentant un intérêt et, dans notre contexte, elles seront distinctes des variables recueillies dans l'ensemble de données non probabilistes.

Nous suggérons qu'il pourrait être possible de concevoir efficacement l'échantillon de référence A , afin d'obtenir spécifiquement un résultat optimal (le coût le plus bas pour une variance donnée, ou vice versa) pour l'estimation $\hat{\mu}_{IPW}$ produite à l'aide de B . Voici quelques scénarios qui pourraient être attrayants :

- Une enquête existante A recueille certains renseignements au sujet d'une population qu'un nouvel ensemble de données administratives vaste, mais incomplet, comprend également. Pour réduire le fardeau de réponse, nous pouvons cesser de recueillir y dans A et nous fier plutôt à B . L'échantillon A peut ensuite être réutilisé pour fournir les renseignements sur les caractéristiques $\mathbf{x}$ à utiliser pour corriger le biais de sélection.
- Nous aimerions tirer parti d'une nouvelle source de données disponible B , mais nous ne disposons pas de suffisamment de renseignements auxiliaires pour $\mathbf{x}$, afin d'aider à corriger le biais de sélection. Dans ce cas, nous pourrions choisir d'élaborer une enquête probabiliste A pour faciliter la collecte de $\mathbf{x}$.

Une question raisonnable est : « Si nous concevons l'enquête de référence pour permettre de produire des statistiques à partir de y , pourquoi ne pas simplement recueillir y dans l'enquête? ». En effet, si seul un petit nombre de données supplémentaires doivent être recueillies, le coût supplémentaire pourrait être minime. Toutefois, si tout élément de donnée supplémentaire est si substantiel que cela introduit des augmentations importantes des coûts et du fardeau du répondant, un compromis est nécessaire. Parfois, il est également pratiquement impossible de recueillir certaines variables dans A , en raison de l'impossibilité de concevoir la collecte, du coût de la collecte des éléments de données et/ou du fardeau de réponse.

Les économies augmentent encore si A peut être utilisé pour résoudre le biais de sélection pour plus d'un ensemble de données non probabilistes. Cela permet à l'ONS de réduire le nombre d'enquêtes à mener. Auparavant, l'ONS peut avoir mené une série d'enquêtes; chacune d'entre elles recueillant des données sur un ensemble particulier de sujets. L'ONS se fierait maintenant davantage à une série de données non probabilistes, complétées par des renseignements auxiliaires provenant d'une « enquête sur les profils de population » A , pour aider à résoudre les problèmes de biais de sélection.

3.2 Plan de sondage optimal utilisant $\hat{\mathbf{x}}$

La variance de l'estimateur IPW $\hat{\mu}_{IPW}$ dépend généralement de plusieurs facteurs. Ceux-ci comprennent le modèle utilisé pour π_i^B (p. ex. un modèle logistique) et le schéma de sondage utilisé pour A (p. ex. l'échantillonnage de Poisson). Pour une expression de la variance donnée pour l'estimateur IPW, il est possible de dériver des probabilités de sélection optimales. Ces probabilités auront tendance à nécessiter de connaître $\mathbf{x}_i$ pour $i \in U$, et les probabilités optimales seront proportionnelles à une fonction des variables $\mathbf{x}$. Il se peut que nous n'ayons pas toujours de renseignements pour tous les $\mathbf{x}$. Il se peut que nous n'ayons des renseignements disponibles au niveau de la population que pour certaines variables $\mathbf{z}$, par exemple les variables de la base de sondage pour le plan d'échantillonnage.

Comment pourrions-nous mettre en œuvre un plan d'échantillonnage optimal en pratique, lorsque $\mathbf{x}$ n'est peut-être pas disponible pour l'ensemble de la population, ou qu'il n'est que partiellement disponible? Nous adoptons la proposition de Clark et Steel (2022) pour réduire au minimum la variance prévue de $\hat{\mu}_{IPW}$, dépendant des renseignements disponibles $\mathbf{z}$. En pratique, les étapes suivantes peuvent être suivies :

1. Utiliser les enregistrements disponibles i qui comprennent à la fois $\mathbf{x}_i$ et $\mathbf{z}_i$ pour imputer $\hat{\mathbf{x}}_i$ pour le reste de la population.
2. Produire $\hat{\pi}_i^{A,opt}$ en utilisant $\mathbf{x}_i$ et, au besoin, les $\hat{\mathbf{x}}_i$ imputés.
3. Répéter les étapes 1 et 2 m fois.
4. Calculer la moyenne des valeurs m de $\hat{\pi}_i^{A,opt}$ et l'utiliser dans le plan d'échantillonnage.

L'une des approches pour créer les valeurs estimées consisterait à appliquer la méthode des k plus proches voisins (kNN) pour imputer en masse les valeurs $\hat{\mathbf{x}}_i$ lorsqu'elles sont manquantes, en utilisant $\mathbf{z}_i$ comme mesure de distance. L'idée d'imputation de masse à l'aide de la méthode kNN a déjà été explorée, voir par exemple Rivers (2007) et Yang et coll. (2021). Une autre solution qui pourrait être moins intensive sur le plan des calculs serait d'utiliser l'imputation par enregistrement donneur, où les catégories de donneur sont conçues en fonction de $\mathbf{z}$.

Les données d'imputation peuvent être obtenues d'une enquête existante A recueillant $\mathbf{x}_i$, ou d'une étude pilote. De plus, si nous pouvons coupler les enregistrements dans B avec la population de la base, l'ensemble de données

couplées contiendra alors à la fois $\mathbf{x}_i$ et $\mathbf{z}_i$, $i \in B$. L'ensemble de données couplées peut alors fournir $\mathbf{x}_i$ pour $i \in B$, et les valeurs imputées $\hat{\mathbf{x}}_i$ pour $i \in U \setminus B$ peuvent être obtenues à partir de l'enquête existante ou de l'étude pilote.

4. Estimation du score de propension, chevauchement exclu

N'oublions pas que nous souhaitons estimer la propension inconnue de la sélection dans un ensemble de données non probabilistes π_i^B , et ces propensions sont présumées i) satisfaire la quantité de données manquantes SAR et ii) respecter un modèle paramétrique dont les paramètres $\boldsymbol{\theta}_0$ doivent être estimés. La fonction de vraisemblance de la population de $\pi_i^B = \pi(\mathbf{x}_i; \boldsymbol{\theta}_0)$ est :

$$L(\boldsymbol{\theta}) = \prod_{i \in U} (\pi_i^B)^{\delta_i} (1 - \pi_i^B)^{(1-\delta_i)} \quad (2)$$

et la fonction du logarithme du rapport de vraisemblance correspondante est

$$l(\boldsymbol{\theta}) = \sum_{i \in U} \delta_i \log \pi_i^B + \sum_{i \in U} (1 - \delta_i) \log(1 - \pi_i^B) \quad (3)$$

Les deux composantes de (3) peuvent être estimées de diverses façons. Nous présumons qu'il est possible de connaître l'appartenance à B de chaque unité de l'échantillon de référence A . Il convient de souligner que la première composante de (3) peut être calculée directement à l'aide de B . La deuxième composante peut être estimée à l'aide de $A \setminus B$, c'est-à-dire par l'échantillon dans A qui n'est pas également dans B . La pseudo-vraisemblance obtenue est donnée par

$$l_{CE}(\boldsymbol{\theta}) = \sum_{i \in B} \log \pi_i^B + \sum_{i \in A \setminus B} d_i^A \log(1 - \pi_i^B) \quad (4)$$

Dans la présente étude, nous appelons cela la fonction de pseudo-vraisemblance chevauchement exclu (CE). Selon un modèle logistique pour les scores de propension, $\pi_i^B = \exp(\mathbf{x}_i^T \boldsymbol{\theta}_0) / (1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta}_0))$. Ensuite, la fonction de pseudo-vraisemblance (4) devient

$$l_{CE}(\boldsymbol{\theta}) = \sum_{i \in B} \{\mathbf{x}_i^T \boldsymbol{\theta} - \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta}))\} - \sum_{i \in A \setminus B} d_i^A \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta})) \quad (5)$$

La fonction de score $S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l_{CE}(\boldsymbol{\theta})$ peut être dérivée, et l'estimateur de maximum de pseudo-vraisemblance $\hat{\boldsymbol{\theta}}$ peut être trouvé numériquement en appliquant la procédure itérative Newton-Raphson pour résoudre $S(\boldsymbol{\theta}) = 0$.

De plus amples renseignements sur l'approche du CE, y compris des détails sur un estimateur de variance remplaçant l'estimateur IPW obtenu (1) se trouvent dans Ang et coll. (2024).

4.1 Plan de fractionnement de population pour l'échantillon de référence

Dans la section précédente, nous avons supposé qu'il était possible de déterminer les unités dans A étant également dans B . Dans la présente section, nous énonçons une hypothèse plus poussée selon laquelle il est possible de déterminer quelles unités de la *base de sondage de la population* ne sont pas dans B . Cela pourrait être possible, par exemple, lorsqu'un identificateur d'unité commun existe dans la base de sondage et dans B . Dans ce cas, nous pouvons décider de concevoir l'échantillon de référence A uniquement en fonction de la sous-population $U \setminus B$. Cette approche présente des similitudes avec un scénario de « population fragmentée » (voir Zhang, 2019), où U est divisée en une « base de sondage » incomplète B et $C = U \setminus B$; B est entièrement échantillonnée et A est un échantillon de C .

Selon le plan de fractionnement de la population, la pseudo-vraisemblance (4) change seulement en ce sens que la somme du deuxième terme est maintenant sur l'ensemble de A , et d_i^A sont des poids qui sont égaux à $(N - N_B)$:

$$l_{SPOE}(\boldsymbol{\theta}) = \sum_{i \in B} \log \pi_i^B + \sum_{i \in A} d_i^A \log(1 - \pi_i^B) \quad (6)$$

Le logarithme du rapport de vraisemblance l_{SPOE} peut être maximisé de même similaire en l_{CE} , par exemple en supposant d'abord qu'un modèle logistique se maintient pour les scores de propension et en résolvant l'équation de score obtenue à l'aide de l'approche Newton-Raphson.

Le scénario ci-dessus serait approprié si les données dans B sont de grande qualité; il est alors possible de les utiliser directement pour faire des inférences. Lorsque les données de B comportent des erreurs de mesure, ou si l'on soupçonne que la quantité de données manquantes dans B dépend également de y , nous pourrions vouloir concevoir l'échantillon A de façon à ce qu'au moins une partie de l'échantillon chevauche B . La collecte de y dans A (ou au moins dans la partie chevauchante) serait bénéfique pour la formation d'un modèle visant à corriger l'erreur de mesure dans B , ou de permettre l'inclusion de y dans le modèle de propension. La meilleure façon de concevoir le chevauchement entre A et B en cas d'enjeux de qualité dans B est laissé pour de futures recherches.

5. Étude de simulation

Une étude de simulation a été menée pour examiner le rendement des estimations IPW établies à l'aide des scores de propension EC. La configuration de la simulation est semblable à celle utilisée dans Chen et coll. (2020). Dans notre simulation, nous tenons compte d'une population finie de taille $N = 200\ 000$ qui comprend un élément de donnée d'intérêt y et des variables auxiliaires x , liées par le modèle

$$y_i = f(x, \beta) = 2 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \sigma \epsilon_i \quad (7)$$

avec $x_{1i} = z_{1i}$, $x_{2i} = z_{2i} + 0,3x_{1i}$, $x_{3i} = 0,2(x_{1i} + x_{2i})$, $x_{4i} = z_{4i} + 0,1(x_{1i} + x_{2i} + x_{3i})$ et $z_{1i} \sim \text{Bernoulli}(0,5)$, $z_{2i} \sim \text{Uniforme}(0,2)$, $z_{3i} \sim \text{Exponentiel}(1)$ et $z_{4i} \sim \chi^2(4)$. Les termes d'erreur ϵ_i sont indépendants et identiquement distribués $N(0,1)$. Les scores de propension réelle π_i^B sont obtenus par

$$\log\left(\frac{\pi_i^B}{1-\pi_i^B}\right) = \theta_0 + 0,1x_{1i} + 0,2x_{2i} + 0,3x_{3i} + 0,4x_{4i} \quad (8)$$

où l'ordonnée à l'origine θ_0 est choisie pour s'assurer que la taille de l'échantillon cible pour N_B est atteinte. L'échantillonnage de Poisson est utilisé pour sélectionner l'échantillon B , tandis que la méthode d'échantillonnage randomisé systématique de probabilité proportionnelle à la taille (PPS) est utilisée pour l'échantillon probabiliste A . En général, les probabilités d'inclusion pour π_i^A sont proportionnelles à $z_i = c + x_{3i}$, où c est choisi pour veiller à ce que le ratio $\max(z_i)/\min(z_i) = 10$. Différents ensembles de simulations ont été exécutés pour trois valeurs de N_B : 2 000, 50 000 et 140 000. La taille de l'échantillon pour A a été maintenue fixe à $n_A = 5\ 000$ pour toutes les exécutions de simulation.

$R = 2\ 000$ simulations ont été exécutées pour chaque scénario de taille d'échantillon. Le rendement des différents estimateurs a été évalué au moyen du calcul du biais relatif en pourcentage de Monte-Carlo (%BR) et du pourcentage de racine carrée de l'erreur quadratique moyenne relative (%REQMR) :

$$\%BR = \frac{1}{R} \sum_{r=1}^R \frac{\hat{Y}_r - \bar{Y}}{\bar{Y}} \times 100$$

$$\%REQMR = \frac{\sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{Y}_r - \bar{Y})^2}}{\bar{Y}} \times 100$$

où $\hat{Y}_r$ est l'estimation de la moyenne de la population calculée pour la r^e simulation et $\bar{Y} = \mu_y$ est la moyenne réelle de la population.

Les résultats ont été produits pour les estimateurs suivants :

- les estimations CE - IPW à l'aide de l'estimateur proposé du score de propension chevauchement exclu;
- des probabilités CE optimal - sélection optimale ont été produites, en supposant tous les renseignements auxiliaires disponibles;
- les probabilités CE optimal « anticipé » - sélection optimale ont été estimées, en supposant que tous les renseignements auxiliaires manquaient, sauf pour x_3 ;
- le fractionnement de la population CE – échantillon A est tiré de $U \setminus B$, et l'approche de CE est utilisée;
- CLW - estimateur IPW de type Hájek dans Chen et coll. (2020);
- WVL - estimateur ALP proposé par Wang et coll. (2021)

5.1 Résultats de simulation

Le tableau 5.1-1 présente les valeurs de %BR et de %REQMR pour chacun des estimateurs selon les différentes tailles d'échantillons pour B . Les approches de CE proposées, ainsi que l'estimateur CLW, ne comprennent pas de biais. L'estimateur WVL est à peu près sans biais, malgré un soupçon de biais pour la plus petite taille d'échantillon pour B .

Tableau 5.1-1
Rendement comparatif des estimateurs

	$N_B = 2\ 000$		$N_B = 50\ 000$		$N_B = 140\ 000$	
	%BR	%REQMR	%BR	%REQMR	%BR	%REQMR
CE	-0,06	5,80	-0,01	0,77	-0,02	0,32
CE optimal	0,10	5,70	0,00	0,72	-0,01	0,27
CE optimal « anticipé »	-0,01	5,83	-0,02	0,78	0,00	0,29
Fractionnement de la population CE	-0,14	5,84	0,00	0,75	-0,01	0,22
CLW	-0,17	6,16	-0,07	1,27	-0,06	0,74
WVL	1,10	5,59	0,47	0,88	0,04	0,53

Note : Les valeurs en caractères gras indiquent la REQMR la plus basse pour une valeur particulière de N_B

En ce qui concerne le rendement des estimateurs relativement à la REQMR, l'approche WVL donne la REQMR la plus faible lorsque N_B est très petit. Toutefois, pour les échantillons de plus grande taille, les approches de CE dépassent à la fois l'approche WVL et l'estimateur CLW. L'optimisation du plan d'échantillonnage pour A réduit la REQMR comparativement à une situation où nous n'utilisons pas les probabilités de sélection optimales, comme nous nous y attendons. La conception de A de façon à le tirer uniquement de $U \setminus B$ entraîne d'autres gains en matière d'exactitude; cette approche se traduit par la REQMR la plus faible pour le plus grand N_B . Il reviendra aux recherches futures d'essayer également de produire des probabilités de sélection optimale selon l'approche de fractionnement de la population.

Il est encourageant de constater que notre utilisation de probabilités de sélection optimale « anticipées » peut également être efficace. Dans la simulation, pour la plus grande taille d'échantillon de N_B , nous avons été en mesure d'atteindre une petite réduction de la REQMR en utilisant ces probabilités optimales estimées plutôt qu'un plan non optimisé pour A .

6. Conclusion

Des approches de quasi-randomisation ont été élaborées pour permettre des inférences de population finie à partir de données non probabilistes. Ces approches utilisent les renseignements auxiliaires disponibles au sein de la population ou à partir d'un échantillon probabiliste de référence pour estimer la propension à la sélection dans l'ensemble de données non probabilistes.

Dans cet article, nous avons proposé plusieurs idées pour concevoir plus efficacement l'échantillon de référence en conjonction avec un estimateur pondéré selon la probabilité inversée IPW connexe qui utilise les propensions estimées produites par la pseudo-vraisemblance à chevauchement exclu (CE). Le plan optimal de l'échantillon de référence A pour la collecte de x offre l'occasion à l'ONS d'adapter son programme d'enquêtes, facilitant l'utilisation de multiples ensembles de données non probabilistes (en particulier de grands ensembles) comme source principale de statistiques, complétées par un petit échantillon de référence auxiliaire.

Ces idées de conception ont été examinées de façon empirique dans le cadre d'une étude de simulation. Les résultats indiquent que lorsque des données pour x sont disponibles dans la base de sondage de la population, l'utilisation de probabilités de sélection optimale pour le plan de sondage offre des avantages sur le plan de l'efficacité. Lorsque x n'est pas entièrement disponible dans la base de sondage, mais que B est grand, l'utilisation de probabilités de sélection optimale « anticipées » en fonction des valeurs prédites $\hat{x}$ peut également entraîner des améliorations. Lorsque nous pouvons coupler B avec la base de sondage de la population, il peut être souhaitable de sélectionner l'échantillon A uniquement dans $U \setminus B$.

Dans la présente étude, l'estimateur de score de propension EC que nous avons utilisé en conjonction avec les idées de plan d'échantillonnage s'applique lorsque nous sommes en mesure de déterminer les unités dans A se trouvant également dans B . L'estimateur produit des estimations impartiales selon le bon modèle de travail et efficaces, surtout lorsque B est grand.

Bibliographie

- Ang, L., R. Clark, B. Loong et A. Holmberg. (2024), « Estimating Propensities of Selection for Big Datasets via Data Integration », présenté à la 7^e International Conference on Establishment Statistics, Glasgow, Royaume-Uni. Document accessible au lien suivant : <https://arxiv.org/abs/2501.04185>.
- Australian Bureau of Statistics (2024a), « Australian Agriculture: Broadacre Crops, 2022-23 financial year | Australian Bureau of Statistics ». Accessible à l'adresse <https://www.abs.gov.au/statistics/industry/agriculture/australian-agriculture-broadacrecrops/latest-release>.
- Australian Bureau of Statistics (2024b), « Monthly Employee Earnings Indicator, October 2023 to March 2024 | Australian Bureau of Statistics ». Accessible à l'adresse <https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/monthlyemployee-earnings-indicator/latest-release>.
- Australian Bureau of Statistics (2024c), « Monthly Household Spending Indicator, June 2024 | Australian Bureau of Statistics ». Accessible à l'adresse <https://www.abs.gov.au/statistics/economy/finance/monthly-household-spendingindicator/latest-release>.
- Australian Bureau of Statistics (2024d), « Weekly Payroll Jobs, Week ending 15 June 2024 | Australian Bureau of Statistics ». Accessible à l'adresse <https://www.abs.gov.au/statistics/labour/jobs/payrolljobs/latest-release>.
- Burakauskaitė, I., et A. Čiginas (2023), « An Approach to Integrating a Non-Probability Sample in the Population Census », *Mathematics* 11(8), p. 1782. Numéro DOI : <https://doi.org/10.3390/math11081782>.
- Chen, Y., P. Li et C. Wu. (2020), « Doubly Robust Inference With Nonprobability Survey Samples », *Journal of the American Statistical Association*, 115(532), 2011-2021. Numéro DOI : <https://doi.org/10.1080/01621459.2019.1677241>.
- Clark, R.G. et D.G. Steel. (2022), « Sample design for analysis using high-influence probability sampling », *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(4), p. 1733-1756. Numéro DOI : <https://doi.org/10.1111/rssa.12916>.
- Cochran, W.G. (1977), *Sampling Techniques*, 3^e édition, New York, NY: John Wiley & Sons, Inc.

- Meng, X.-L. (2018), « Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election », *The Annals of Applied Statistics*, 12(2), p. 685-726. Numéro DOI : <https://doi.org/10.1214/18-AOAS1161SF>.
- Rivers, D. (2007), Sampling for web surveys, dans « Joint Statistical Meetings 2007, Section on Survey Research Methods ».
- Division de statistique des Nations Unies (2014), « Principes fondamentaux de la statistique officielle », Document accessible à l'adresse <https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx>. Date de consultation : 2024-11.
- Wang, L., R. Valliant et Y. Li. (2021), « Adjusted logistic propensity weighting methods for population inference using nonprobability-based epidemiologic cohorts », *Statistics in Medicine*, 40(24), p. 5237-5250. Numéro DOI : <https://doi.org/10.1002/sim.9122>
- Yang, S., J.-K. Kim et Y. Hwang. (2021), « Intégration de données d'enquêtes probabilistes et de mégadonnées aux fins d'inférence de population finie au moyen d'une imputation massive », *Techniques d'enquête*, 47(1), p. 29-58. Article accessible à l'adresse <https://www150.statcan.gc.ca/n1/pub/12-001-x/2021001/article/00004-fra.htm>.
- Zhang, L.-C. (2019), « On valid descriptive inference from non-probability sample », *Statistical Theory and Related Fields*, 3(2), p. 103-113. Numéro DOI : <https://doi.org/10.1080/24754269.2019.1666241>.