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

Techniques d’enquête

Amélioration de l’erreur de mesure et de la représentativité dans les enquêtes non probabilistes

par Aditi Sen et Partha Lahiri

Date de diffusion : le 29 juin 2026

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

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

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Publication autorisée par la ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par la ministre de l'Industrie, 2026

L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2026
Vol. 52, nº 1, pp. 1-30
Statistique Canada, nº 12-001-X au catalogue

Amélioration de l'erreur de mesure et de la représentativité dans les enquêtes non probabilistes

Aditi Sen et Partha Lahiri[1]

Résumé

À l'ère des mégadonnées, les enquêtes non probabilistes sont de plus en plus nombreuses. Les techniques d'intégration de données faisant appel à la fois à des enquêtes probabilistes et non probabilistes sont largement utilisées afin d'obtenir de meilleures estimations pour des populations finies. Alors que la majorité des travaux de recherche existants ont porté sur la réduction du biais de sélection dans les enquêtes non probabilistes, la question de l'erreur de mesure dans ces enquêtes demeure relativement peu étudiée. Les méthodes statistiques visant à atténuer le biais de sélection ne permettent une estimation fiable que si l'on part du principe que les réponses aux enquêtes sont exactes. Motivée par une étude de cas récente de Kennedy, Mercer et Lau (2024), notre étude aborde le biais découlant à la fois des erreurs de mesure et des erreurs d'échantillonnage dans les enquêtes non probabilistes. Dans le présent article, nous proposons une nouvelle méthode d'intégration de données qui fait appel à plusieurs enquêtes probabilistes et non probabilistes et s'appuie sur des modèles d'apprentissage automatique pour construire un estimateur composite. L'estimateur composite proposé intègre les enquêtes probabilistes et non probabilistes lorsque les deux types d'enquêtes contiennent des variables réponses d'intérêt. Nous analysons le rendement de cet estimateur, par rapport à un estimateur composite existant dans la littérature, tant sur le plan analytique qu'empirique, à l'aide de données provenant de plusieurs enquêtes de Kennedy et coll. (2024). Enfin, nous déterminons les conditions dans lesquelles l'estimateur proposé donne de meilleurs résultats que les estimateurs fondés uniquement sur des enquêtes probabilistes.

Mots-clés : Correction du biais; enquête à participation volontaire; erreur de mesure; estimateur composite; intégration de données; modélisation prédictive.

1. Introduction

L'utilisation de l'inférence fondée sur le plan d'échantillonnage à partir d'échantillons probabilistes joue un rôle essentiel dans la production d'estimations fiables des caractéristiques d'une population finie (comme les moyennes et les totaux) depuis les travaux révolutionnaires de Neyman (1934). Cette approche présente l'attrait de pouvoir être appliquée à toute variable d'intérêt sans qu'il soit nécessaire de recourir à une modélisation explicite, lorsque les éléments suivants sont disponibles : une base de sondage (liste de toutes les unités de la population finie visée), un plan d'échantillonnage comportant des probabilités de sélection connues non nulles et les réponses d'enquête recueillies auprès de l'échantillon sélectionné. Malgré la présence de probabilités de sélection inégales, l'inférence fondée sur le plan demeure valide dans le cas d'échantillons probabilistes; par exemple, l'estimateur de Horvitz-Thompson (Horvitz et Thompson, 1952) fournit des estimations sans biais en intégrant les probabilités d'inclusion connues. En revanche, les échantillons non probabilistes posent un problème fondamentalement différent : le mécanisme de sélection est inconnu et la probabilité de sélection de certaines unités de la population peut être nulle. L'absence de renseignements relatifs au plan d'échantillonnage rend les méthodes traditionnelles fondées sur le plan inapplicables et motive l'élaboration d'autres stratégies inférentielles.

1. Aditi Sen est doctorante, Applied Mathematics & Statistics, and Scientific Computation. Courriel : asen123@umd.edu; Partha Lahiri est professeur, Department of Mathematics et The Joint Program in Survey Methodology, University of Maryland, College Park (MD) 20742, États-Unis. Courriel : plahiri@umd.edu.

Même s'il est souhaitable de produire des estimations fondées sur des échantillons probabilistes, cela n'est pas toujours possible en pratique du fait de contraintes budgétaires ou autres. La réalisation d'enquêtes probabilistes rigoureusement conçues peut s'avérer très coûteuse, notamment en raison du recrutement d'intervieweurs, du suivi des répondants sélectionnés sur une période prolongée malgré la baisse des taux de réponse et pour d'autres raisons. Ces facteurs, conjugués à l'afflux de mégadonnées à l'ère de l'informatique avancée, expliquent la popularité récente des enquêtes non probabilistes, même si celles-ci figurent depuis longtemps dans les ouvrages publiés sur les enquêtes par sondage (enquêtes-échantillons). Lorsque des réponses rapides peuvent être obtenues grâce à des enquêtes en ligne ou qu'il est possible d'échantillonner des personnes tout en respectant certaines répartitions globales (comme dans le cas d'un échantillonnage par quotas), on constate une tendance croissante à recourir à ces options plus pratiques. Pour refléter cet intérêt croissant, des numéros spéciaux de revues (*Techniques d'enquête* et le *Calcutta Statistical Association Bulletin*) ont récemment été consacrés à la recherche statistique sur l'échantillonnage non probabiliste.

La prudence est de mise lors de la production d'estimations relatives aux caractéristiques d'une population finie cible fondées sur de telles enquêtes non probabilistes, du fait du problème évident de biais de sélection qu'elles posent. Les chercheurs ont mis sur pied diverses méthodes statistiques visant à améliorer la représentativité d'enquêtes non probabilistes en les intégrant à des enquêtes probabilistes (appelées « enquêtes de référence »), qui ne contiennent généralement pas les variables d'intérêt, mais comportent des variables auxiliaires utiles. Pour obtenir un aperçu de ces approches, consulter, entre autres, Rao (2021) et Beaumont (2020). Plus précisément, lorsque des variables auxiliaires communes sont disponibles pour ces deux types d'enquêtes, certaines méthodes couramment utilisées comprennent les méthodes d'appariement et d'imputation massive (Rivers, 2007; Kim, Park, Chen et Wu, 2021), les méthodes de pondération selon la propension inverse (PPI) (Lee et Valliant, 2009; Valliant et Dever, 2011; Wang, Valliant et Li, 2021), les méthodes doublement robustes (Chen, Li et Wu, 2020) et d'autres encore. Wu (2022) présente un excellent survol de ces approches fondées sur des modèles et des hypothèses connexes. Plus récemment, Beaumont, Bosa, Brennan, Charlebois et Chu (2024) ont examiné des techniques de sélection de variables en parallèle avec la pondération selon la propension et à la poststratification de poids estimés. Contrairement à ces méthodes nécessitant la combinaison d'enquêtes probabilistes et non probabilistes, Pfeffermann (2015) ainsi que Kim et Morikawa (2023) s'appuient uniquement sur des enquêtes non probabilistes aux fins d'inférence.

Il est bien établi dans la littérature que l'application de méthodes d'intégration de données rend les enquêtes non probabilistes plus représentatives de la population finie, mais ces méthodes sont élaborées selon l'hypothèse de l'exactitude des réponses aux enquêtes. Kennedy et coll. (2024) soulignent que les enquêtes non probabilistes en ligne menées à des fins commerciales (appelées enquêtes à participation volontaire ou plus brièvement enquêtes volontaires) comprennent une quantité importante d'erreurs de mesure et que les biais dus à d'autres sources que l'erreur d'échantillonnage ne doivent pas être négligés. La présente étude s'inspire de l'étude d'étalonnage de 2021 (Mercer et Lau, 2023), conçue par le Pew

Research Center (ci-après le Pew), afin d'évaluer la précision des enquêtes en ligne produisant des estimations de la population générale adulte aux États-Unis ainsi que de sous-groupes démographiques clés, comme l'âge, le groupe ethnique, le genre et le niveau de scolarité. Les valeurs de référence proviennent de grandes enquêtes gouvernementales comme l'American Community Survey (ACS), la Current Population Survey (CPS), entre autres. Kennedy et coll. (2024) observent qu'après le calage, les réponses de certains sous-groupes dans le cadre d'enquêtes à participation volontaire demeurent fortement biaisées; contrairement aux enquêtes probabilistes, dont les estimations correspondantes concordent étroitement avec les valeurs de référence. Ce phénomène est attribué à la présence de fausses réponses; par exemple le fait de répondre « Oui » quelle que soit la question. Ces réponses, essentiellement non sincères ou frauduleuses, génèrent une importante erreur de mesure dans les enquêtes à participation volontaire. Pour obtenir de plus amples précisions sur les ouvrages publiés relatifs aux fausses réponses, le lecteur peut consulter Kennedy, Hatley, Lau, Mercer, Keeter, Ferno et Asare-Marfo (2021).

Dans Kennedy et coll. (2024), les auteurs utilisent la technique d'étalonnage pour repérer les sous-groupes d'enquêtes volontaires où de fausses réponses se concentrent, mais ils ne proposent aucune solution statistique pour corriger ce problème. Une solution rapide consistant à éliminer purement et simplement les répondants fournissant de fausses réponses pourrait aggraver la situation. Les auteurs soulignent la nécessité, pour les statisticiens travaillant avec des échantillons non probabilistes, d'aborder à la fois les problèmes de fausses réponses et de représentativité, et non pas seulement cette dernière. Notre objectif principal est d'intégrer deux domaines de recherche importants en matière d'échantillonnage non probabiliste : l'amélioration de la représentativité et la prise en compte des différences de mesure. Dans notre cadre, les enquêtes probabilistes et non probabilistes contiennent toutes deux les variables réponses d'intérêt, ainsi qu'un ensemble de variables auxiliaires communes et informatives. La disponibilité des variables réponses dans les deux types d'enquêtes est essentielle pour corriger le biais découlant des différences de mesure. La section 1.1 présente nos principales questions de recherche et les solutions proposées pour y répondre. Nous exposons en détail des hypothèses à la section 1.2. Par souci de concision, nous désignerons désormais les enquêtes probabilistes par ps et les enquêtes non probabilistes (notamment les enquêtes à participation volontaire) par nps.

1.1 Questions de recherche

Q.1 L'amélioration de la représentativité à elle seule permet-elle d'améliorer les estimations issues d'enquêtes nps?

Pour répondre à la question 1, nous appliquons la méthode de pondération selon la propension inverse (PPI) proposée par Chen et coll. (2020), en supposant que cette méthode corrige adéquatement le biais de sélection des enquêtes nps. Nous présentons un survol de la méthode PPI à la section 2. Pour obtenir de plus amples précisions sur les hypothèses (comme les données manquantes aléatoirement, MAR pour « missing at random » en anglais), le lecteur peut consulter Wu (2022). Il convient de souligner que Kennedy et coll. (2024) utilisent uniquement des techniques d'étalonnage dans leur étude de cas, et que des méthodes

fondées sur des modèles, comme la méthode PPI, ne sont pas employées pour améliorer la représentativité dans leurs enquêtes *nps*. Il importe également de mentionner que les méthodes de pondération concurrentes, mentionnées précédemment dans l'introduction, peuvent donner des résultats différents. Nous retenons ici la méthode PPI, l'une des plus couramment utilisées, sans approfondir la comparaison avec d'autres approches, puisque cela dépasse la portée principale de la présente étude.

Q.2 Comment corriger les différences de mesure dans les enquêtes nps?

Pour la question 2, nous élaborons une méthodologie visant à corriger les différences de mesure dans les enquêtes *nps* attribuables à de fausses réponses. L'étude de référence de Kennedy et coll. (2024) montre que, comparativement aux grandes enquêtes (Current Population Survey, American Community Survey, etc.), les enquêtes probabilistes (*ps*) ne sont pas sujettes à d'importantes erreurs de mesure, alors que les enquêtes non probabilistes (*nps*) le sont. Dans la présente étude, notre objectif est d'amener les niveaux d'erreur des enquêtes *nps* au même niveau que ceux observés pour les enquêtes *ps*. Ainsi, comme Kennedy et coll. (2024), nous considérons les enquêtes *ps* comme la référence, puisque l'on suppose que les erreurs de mesure sont les plus faibles dans leurs ensembles de données. De cette façon, la méthode proposée corrige les biais de mesure relatifs entre les deux types d'enquêtes *ps* et *nps* (et non les biais absolus). Même si nous reconnaissons que cette hypothèse puisse ne pas se vérifier parfaitement en pratique, notre démarche vise avant tout à améliorer la qualité des enquêtes *nps* en tirant parti des enquêtes *ps*, plutôt qu'à raffiner les enquêtes *ps* elles-mêmes. Par conséquent, la méthode proposée pour répondre à la question 2 consiste à corriger les « différences de mesure » entre les deux types d'enquêtes.

Q.3 Comment combiner les estimations provenant d'enquêtes ps et nps afin d'obtenir des estimations présentant un moindre biais?

Nous observons empiriquement que, même après correction des différences de mesure, les estimations de la moyenne pour une population finie obtenues à partir d'enquêtes *nps* peuvent être encore améliorées au moyen d'estimateurs composites conçus judicieusement. De tels estimateurs ont été étudiés dans le cadre d'estimations sur petits domaines (Schaible, 1978; Rao et Molina, 2015), lorsque les auteurs envisagent une combinaison pondérée d'un estimateur direct et d'un estimateur synthétique. Elliott et Haviland (2007) proposent également un estimateur pondéré fondé sur une enquête *nps* (échantillon de convenance en ligne) avec une enquête *ps* de manière à obtenir une erreur quadratique moyenne (EQM) inférieure à celle d'estimateurs reposant sur une enquête *ps* seule. Plus récemment, Kim et Tam (2021) ont proposé un estimateur d'intégration de données corrigé du biais, tirant parti des mégadonnées (que l'on peut aussi considérer comme une enquête *nps*) et d'une enquête *ps*, lorsque l'une d'elles fait l'objet d'une erreur de mesure.

Pour répondre à la question 3, nous proposons un nouvel estimateur composite, qui est une combinaison pondérée d'estimations sans biais provenant à la fois d'enquêtes *ps* et *nps*, visant à réduire à la fois le biais de sélection et les différences de mesure dans l'enquête *nps*. Dans la présente étude, nous démontrons analytiquement que l'EQM de l'estimateur proposé est inférieure à celle de l'estimateur d'Elliott et Haviland (2007). De plus, nous intégrons la modélisation prédictive à l'aide de modèles d'apprentissage automatique (AA) avancés (forêt aléatoire, optimisation de gradient) afin de tenir compte du biais des estimations fondées

sur la PPI. L'utilisation de modèles d'AA devient de plus en plus courante dans le contexte des enquêtes sur échantillon. La littérature récente portant sur l'échantillonnage, notamment les travaux de Toth et Eltinge (2011), de Breidt et Opsomer (2017), de Kern, Klausch et Kreuter (2019), ainsi que ceux de Dagdoug, Goga et Haziza (2021), ont recours à des techniques modernes de prédiction, comme les arbres de régression et les forêts aléatoires, aux fins de l'estimation d'enquête assistée par modèle.

L'estimateur proposé par Kim et Tam (2021) tient compte à la fois du biais de sélection et des erreurs de mesure dans les mégadonnées, sans supposer de mécanisme de MAR dans leur approche relative aux données manquantes. Toutefois, leur méthode exige la détermination d'unités se chevauchant et a recours à une pondération par calage, en traitant les mégadonnées comme un plan d'échantillonnage incomplet pour la population finie. Notre méthode est en revanche plus simple (elle ne nécessite pas de comparaison au niveau des unités) et repose sur l'hypothèse MAR pour la correction du biais fondée sur la PPI. Structurellement, notre estimateur se rapproche davantage de celui d'Elliott et Haviland (2007), mais s'en distingue par le fait qu'il minimise l'EQM dans une autre classe d'estimateurs. De plus, Elliott et Haviland (2007) attribuent le biais principalement à l'erreur d'échantillonnage (c'est-à-dire au biais de sélection), sans tenir compte de l'erreur de mesure, laquelle est étroitement liée à notre cadre analytique.

Q.4 Dans quels cas les estimateurs composites donnent-ils de meilleurs résultats que les estimateurs fondés sur une enquête ps seule?

D'après notre analyse des données, nous observons que l'estimateur composite proposé ne donne pas uniformément de meilleurs résultats que les estimateurs fondés uniquement sur l'enquête ps, lorsque cette dernière repose sur un échantillon suffisamment grand. Ainsi, pour répondre à la question Q4, nous avons tiré de plus petits échantillons d'une enquête ps disponible, que nous avons ensuite combinés à une enquête nps, afin d'illustrer un scénario où l'estimateur composite proposé donne de meilleurs résultats que l'estimateur fondé uniquement sur l'enquête ps.

1.2 Conditions et hypothèses

Après avoir présenté nos questions de recherche et les solutions proposées, nous passons maintenant à un examen détaillé des principales hypothèses sous-jacentes à notre méthodologie. L'approche proposée repose principalement sur deux hypothèses centrales (A1) et (A2).

(A1) La méthode de pondération selon la propension inverse (PPI) vise principalement à éliminer le biais de sélection dans les échantillons non probabilistes.

Cette hypothèse constitue la base de notre approche concernant la question 1, qui porte sur l'amélioration de la représentativité d'une enquête nps. La méthode PPI est largement utilisée à cette fin dans la littérature. Toutefois, cette méthode suppose un mécanisme de sélection MAR (Rubin, 1976). Si l'hypothèse MAR n'est pas respectée, l'estimation des scores de propension devient très difficile (Little et Rubin, 2019). De plus, la méthode PPI nécessite des variables auxiliaires solides pour corriger le biais de sélection. En l'absence de bonnes covariables, un biais de sélection résiduel peut subsister dans l'enquête nps même

après l'application de la méthode PPI. Plus précisément, les différences entre les enquêtes *ps* et *nps* après correction du biais de sélection seront une combinaison du biais de sélection résiduel dans les *nps*, des erreurs de mesure dans les *nps* et les *ps*, et de l'erreur d'échantillonnage. Toutefois, nous n'examinons pas davantage une telle situation dans la présente analyse, notre objectif principal étant de trouver de nouvelles techniques pour intégrer les deux sources de biais (sélection et mesure) dans les enquêtes *nps*.

(A2) Les échantillons probabilistes comportent une erreur de mesure négligeable.

Cette hypothèse est utilisée pour corriger les écarts de mesure dans les *nps*. La présence d'erreurs de mesure dans les *ps* est très probable, puisque toute enquête, aussi rigoureusement menée soit-elle, est sujette à ce type d'erreur. Une abondante littérature traite des erreurs de mesure dans les *ps*, en remontant aux travaux de Mahalanobis (1946), dans le contexte d'enquêtes agricoles et d'estimation de récoltes en Inde. Biemer (2009), dans un chapitre de son ouvrage, renvoie aux travaux classiques de Hansen, Hurwitz et Bershad (1961) sur l'estimation de la variance de l'erreur de mesure. Groves et Lyberg (2010) abordent les erreurs de mesure dans les *ps* dans le cadre du concept d'« erreur d'enquête totale ».

Plusieurs articles traitent du fait que les effets du mode d'enquête à l'origine d'erreurs de mesure dans les enquêtes probabilistes peuvent être atténués par une bonne conception du questionnaire ou corrigés au moyen de méthodes de pondération ou d'appariement après la collecte. Plus précisément, Jäckle, Roberts et Lynn (2010) présentent des méthodes d'évaluation des effets de mode d'enquête fondées sur des modèles à cotes proportionnelles, la modélisation par équations structurelles, etc. Suzer-Gurtekin, Heeringa et Vaillant (2012) ont recours à l'imputation multiple pour isoler analytiquement les effets du choix du mode et de la mesure. Schouten, Van den Brakel, Buelens, Van der Laan et Klausch (2013) proposent un nouveau plan expérimental permettant de décomposer les effets relatifs de mode d'enquête. Pour obtenir de plus amples renseignements, consulter Vannieuwenhuyze, Loosveldt et Molenberghs (2010), Buelens et Van den Brakel (2015), Klausch, Schouten, Buelens et Van den Brakel (2017) ainsi que les références qui y sont citées. En somme, l'évaluation et la correction d'erreurs de mesure dans les *ps* constituent un vaste domaine de recherche, qui n'est pas au centre de la présente étude. Dans celle-ci, nous nous intéressons principalement à l'amélioration d'une *nps* en tirant parti d'une *ps*. Par conséquent, nous ne poursuivons pas l'exploration des méthodes mentionnées ci-dessus et réservons ce sujet pour de futurs travaux de recherche.

Le reste du présent document est structuré de la manière suivante. La section 2 présente un survol de l'estimateur par PPI proposé par Chen et coll. (2020), utilisé pour améliorer la représentativité dans les *nps*. La section 3 brosse un tableau de la méthode que nous proposons pour traiter simultanément le biais de sélection et les différences de mesure au moyen d'estimateurs composites, élaborés en intégrant une *ps* et une *nps*, lorsque les deux enquêtes comportent la ou les variables d'intérêt. La section 4 expose l'étude de cas qui a motivé ces travaux, tirée du Pew, et fournit les renseignements nécessaires sur les enquêtes, notamment la méthode de collecte des données, la taille de l'échantillon, les données repères, la pondération, etc. La section 5 présente les résultats de notre analyse des données, en répondant successivement aux quatre questions de recherche soulevées précédemment. Enfin, nous formulons quelques remarques finales à la section 6.

2. Amélioration de la représentativité

L'objectif principal de cette section est de traiter la question de la représentativité des enquêtes *nps* et d'atténuer leur biais de sélection. À cette fin, nous utilisons la méthode de pondération selon la propension inverse (PPI) décrite dans Chen et coll. (2020). La méthode PPI présente un avantage sur le plan computationnel, car elle produit un seul ensemble de poids estimés pouvant être utilisés pour n'importe quelle variable réponse. Cette approche est similaire à la méthode fondée sur le plan d'échantillonnage de l'estimateur de Horvitz-Thompson, à la différence près que les probabilités de sélection sont ici estimées plutôt que connues à partir du plan d'échantillonnage. Pour appliquer cette méthode, il faut disposer de renseignements sur des variables auxiliaires communes (provenant à la fois des ps et des nps) ainsi que des poids d'enquête (des ps). Par souci d'exhaustivité, nous présentons ci-dessous un rappel de la méthode PPI telle qu'elle est décrite dans Chen et coll. (2020).

Soit S_A, un échantillon d'enquête *nps* de taille n_A et S_B, un échantillon d'enquête ps de taille n_B issus d'une population finie de taille N. d_i^B est le poids d'échantillonnage connu de la i^{e} unité provenant de l'enquête $ps, i = 1,\ldots,n_B$. L'indicateur de sélection pour S_A, noté R_i, est défini comme suit :

$$R_i = \begin{cases} 1 & \text{si } i \in S_A, \\ 0 & \text{si } i \notin S_A, \end{cases} \quad i = 1,\ldots,N.$$

Le score de propension de la i^{e} unité est défini par $\pi_i^A = P_q(R_i = 1 | x_i), i = 1,\ldots,N$, où q représente le modèle du mécanisme de sélection pour S_A et x est le vecteur des covariables. En appliquant les hypothèses de Chen et coll. (2020), nous avons $\pi_i^A > 0$; R_i et la réponse y_i sont indépendants compte tenu de x_i; R_i et R_j sont indépendants compte tenu de x_i et x_j, pour tout $i \neq j$. Le score de propension prend la forme suivante dans le cas du modèle logistique :

$$\pi_i^A = \pi(x_i, \theta_0) = \frac{\exp(x_i'\theta_0)}{1+\exp(x_i'\theta_0)}, \; i = 1,\ldots,N,$$

où θ_0 représente la vraie valeur des paramètres inconnus du modèle. L'estimateur du maximum de vraisemblance (EMV) du score de propension est $\hat{\pi}_i^A = \pi(x_i, \hat{\theta})$, où $\hat{\theta}$ maximise la fonction pseudo-logarithmique du rapport de vraisemblance suivante :

$$l^*(\theta) \equiv \sum_{i \in S_A} x_i'\theta - \sum_{i \in S_B} d_i^B \log\{1+\exp(x_i'\theta)\}.$$

La solution de l'équation ci-dessus est obtenue à l'aide de la procédure itérative de Newton-Raphson. Enfin, pour une variable de réponse y, l'estimateur par PPI de la moyenne de la population $\mu_Y = N^{-1}\sum_{i=1}^N y_i$ est obtenu comme suit :

$$\hat{\mu}_{CLW} \equiv (\hat{N}^A)^{-1} \sum_{i \in S_A} (y_i / \hat{\pi}_i^A), \tag{2.1}$$

où $\hat{N}^A = \sum_{i \in S_A} (\hat{\pi}_i^A)^{-1}$. Pour une expression détaillée de la variance de $\hat{\mu}_{CLW}$, nous invitons le lecteur à consulter Chen et coll. (2020).

Pour l'enquête ps, nous utilisons l'estimateur pondéré par le plan de sondage pour la variable réponse y, désigné par

$$\hat{\mu}_{cal} \equiv (\hat{N}^B)^{-1} \sum_{i \in S_B} d_i^B y_i, \tag{2.2}$$

où $\hat{N}^B = \sum_{i \in S_B} d_i^B$. La mesure de l'incertitude associée à cet estimateur peut être déduite à partir des renseignements relatifs au plan d'enquête. L'estimateur défini en (2.2) peut également être utilisé pour construire un estimateur pondéré à partir de l'enquête nps, en employant des poids calés plutôt que des poids de sondage. Ainsi, afin d'éviter toute confusion, nous utilisons l'abréviation « cal » pour l'estimateur défini en (2.2), car nous ferons appel au même estimateur pour l'enquête ps (avec les poids de sondage) et l'enquête nps (avec les poids calés du Pew) dans notre analyse des données.

3. Estimateurs composites

Dans cette section, nous présentons une méthode permettant de combiner les estimations provenant d'enquêtes ps et nps, étant donné que la variable d'intérêt est présente dans les deux enquêtes. Pour formuler cet estimateur composite, nous adoptons le modèle décrit dans Elliott et Haviland (2007), selon lequel le biais associé à l'enquête nps est censé être connu. Sur le plan mathématique, soit $\overline{y}_1$, un estimateur pondéré selon le plan d'enquête de la moyenne de la population issu de l'enquête ps ($\hat{\mu}_{cal}$ dans notre cas); $\overline{y}_2$, un estimateur découlant de l'enquête nps ($\hat{\mu}_{CLW}$ dans notre cas). Soit μ, la vraie moyenne de l'estimateur de ps, tandis que l'estimateur de nps comprend un terme de biais additionnel ϵ. Dans ce modèle, ϵ est censé être connu, et μ est le paramètre d'intérêt inconnu. v_1 et v_2 sont respectivement les variances connues de $\hat{\mu}_{cal}$ et $\hat{\mu}_{CLW}$. La corrélation entre les estimations issues de ps et nps est censée être nulle, puisque les échantillons de ps et nps sont tirés indépendamment. Le modèle d'échantillonnage d'Elliott et Haviland (2007) peut alors s'écrire comme suit :

$$\begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \end{pmatrix} \sim \left(\begin{pmatrix} \mu \\ \mu + \epsilon \end{pmatrix}, \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix} \right). \tag{3.1}$$

En pratique, on peut utiliser un estimateur de ϵ par substitution, obtenu à partir de données historiques ou de connaissances provenant de spécialistes. Elliott et Haviland (2007) proposent d'estimer en pratique le biais comme la différence entre les estimations moyennes de ps et nps. v_1 et v_2 peuvent être respectivement obtenues à partir des estimations de la variance d'un estimateur fondé sur le plan (comme notre $\hat{\mu}_{cal}$) et de l'estimation de la variance de nps (comme $\hat{\mu}_{CLW}$ dérivé dans Chen et coll., 2020).

Dans Elliott et Haviland (2007), les auteurs proposent un estimateur composite de la forme :

$$\hat{\mu}_{EV} \equiv \frac{(\epsilon^2 + v_2)\bar{y}_1 + v_1\bar{y}_2}{\epsilon^2 + v_1 + v_2}, \tag{3.2}$$

qui est un estimateur avec biais; le biais résiduel étant donné par

$$b_{EV} \equiv E(\hat{\mu}_{EV}) - \mu = \frac{\epsilon v_1}{\epsilon^2 + v_1 + v_2}, \tag{3.3}$$

et la variance par

$$\mathrm{Var}(\hat{\mu}_{EV}) = \frac{v_1\{(\epsilon^2 + v_2)^2 + v_1 v_2\}}{(\epsilon^2 + v_1 + v_2)^2}. \tag{3.4}$$

Dans les équations ci-dessus, nous utilisons l'indice « EV », d'après le nom des auteurs, pour simplifier la notation. Ainsi, d'après (3.3) et (3.4), l'EQM de $\hat{\mu}_{EV}$ est donnée par

$$\mathrm{EQM}(\hat{\mu}_{EV}) \equiv b_{EV}^2 + \mathrm{Var}(\hat{\mu}_{EV}) = \frac{v_1(\epsilon^2 + v_2)}{\epsilon^2 + v_1 + v_2}. \tag{3.5}$$

$\hat{\mu}_{EV}$, tel que défini dans (3.2), est l'estimateur le plus efficace au sens où il réduit l'EQM dans une classe d'estimateurs avec biais.

Nous proposons un estimateur composite sans biais, qui est une combinaison pondérée des estimations moyennes issues de ps, données par $\hat{\mu}_{cal}$, et des estimations « corrigées du biais » issues de nps, données par $\hat{\mu}_{bc;CLW} \equiv \hat{\mu}_{CLW} - \epsilon$, où l'indice « bc » désigne des valeurs corrigées du biais. Cet estimateur composite, désigné par $\hat{\mu}_{comb}$ (où l'indice indique « combiné »), est défini comme suit :

$$\begin{aligned} \hat{\mu}_{comb}(w) &\equiv w\hat{\mu}_{cal} + (1-w)\hat{\mu}_{bc;CLW} \\ &= w\bar{y}_1 + (1-w)(\bar{y}_2 - \epsilon), \end{aligned} \tag{3.6}$$

où le poids $w \in (0,1)$ est déterminé en réduisant l'EQM de $\hat{\mu}_{comb}(w)$ donnée par

$$\begin{aligned} \mathrm{EQM}(\hat{\mu}_{comb}(w)) &\equiv E[w\bar{y}_1 + (1-w)(\bar{y}_2 - \epsilon) - \mu]^2 \\ &= w^2 v_1 + (1-w)^2 v_2. \end{aligned} \tag{3.7}$$

En dérivant (3.7) par rapport à w et en la posant égale à zéro, on obtient $w_1^* = v_2/(v_1 + v_2)$. Finalement, nous obtenons l'expression

$$\hat{\mu}_{comb} \equiv \left(\frac{v_2}{v_1 + v_2}\right)\bar{y}_1 + \left(\frac{v_1}{v_1 + v_2}\right)(\bar{y}_2 - \epsilon). \tag{3.8}$$

L'estimateur proposé en (3.8) est sans biais pour μ, puisqu'il combine deux estimateurs sans biais : $\hat{\mu}_{cal}$ et $\hat{\mu}_{bc;CLW}$. L'EQM de $\hat{\mu}_{comb}$ est obtenue à l'aide de la formule

$$\text{EQM}(\hat{\mu}_{comb}) = \frac{v_1 v_2}{v_1 + v_2}. \tag{3.9}$$

Ainsi, $\hat{\mu}_{EV}$ et $\hat{\mu}_{comb}$ réduisent tous deux l'EQM, mais dans des classes d'estimateurs disjointes. Pour comparer $\hat{\mu}_{EV}$ et $\hat{\mu}_{comb}$, nous comparons les EQM respectives données par (3.5) et (3.9). Puisque $\epsilon^2 > 0$, nous avons

$$v_2(\epsilon^2 + v_1 + v_2) = (\epsilon^2 + v_2)v_2 + v_1 v_2 < (\epsilon^2 + v_2)v_2 + v_1(\epsilon^2 + v_2) = (\epsilon^2 + v_2)(v_1 + v_2).$$

D'après ce qui précède, puisque

$$\frac{v_1 v_2}{v_1 + v_2} < \frac{v_1(\epsilon^2 + v_2)}{\epsilon^2 + v_1 + v_2}, \tag{3.10}$$

nous obtenons $\text{EQM}(\hat{\mu}_{comb}) < \text{EQM}(\hat{\mu}_{EV})$.

Nous constatons que dans l'exposé précédent le terme de biais ϵ doit être connu, éventuellement à partir d'autres sources. Supposons que $\hat{\epsilon}$, un estimateur de ϵ, soit construit à partir d'une enquête auxiliaire indépendante de $\bar{y}_2$. De plus, supposons $E(\hat{\epsilon}) = \epsilon$, $\text{Var}(\hat{\epsilon}) = v_b$. En vertu de ces hypothèses, l'estimateur de substitution combiné peut s'écrire comme suit :

$$\tilde{\mu}_{comb}(w) \equiv w\bar{y}_1 + (1-w)(\bar{y}_2 - \hat{\epsilon}), \tag{3.11}$$

où le poids $w \in (0,1)$. En utilisant la propriété d'absence de biais de $\hat{\epsilon}$, il s'ensuit que $E(\tilde{\mu}_{comb}(w)) = \mu$, c'est-à-dire, $\tilde{\mu}_{comb}(w)$ est un estimateur sans biais de μ pour toutes les valeurs de $w \in (0,1)$. L'EQM correspondante de $\tilde{\mu}_{comb}(w)$ est

$$\text{EQM}(\tilde{\mu}_{comb}(w)) = w^2 v_1 + (1-w)^2 (v_2 + v_b). \tag{3.12}$$

Minimiser l'EQM dans (3.12) par rapport à w donne le poids optimal $w_2^* = (v_2 + v_b)/(v_1 + v_2 + v_b)$. En substituant ce poids optimal dans (3.11), on obtient alors l'estimateur de substitution $\tilde{\mu}_{comb}$ défini par

$$\tilde{\mu}_{comb} \equiv \left(\frac{v_2 + v_b}{v_1 + v_2 + v_b}\right)\bar{y}_1 + \left(\frac{v_1}{v_1 + v_2 + v_b}\right)(\bar{y}_2 - \hat{\epsilon}) \tag{3.13}$$

et l'EQM de $\tilde{\mu}_{comb}$ est donnée par $\text{EQM}(\tilde{\mu}_{comb}) = v_1(v_2 + v_b)/(v_1 + v_2 + v_b)$. Il convient de mentionner que $\hat{\mu}_{comb}$ est un cas spécial de $\tilde{\mu}_{comb}$ avec $v_b = 0$, correspondant au cas dégénéré où l'estimateur du biais satisfait $\hat{\epsilon} = \epsilon$ avec une probabilité de 1. En comparaison avec l'estimateur proposé par Elliott et Haviland (2007), une manipulation algébrique similaire à celle de (3.10) montre que $\text{EQM}(\tilde{\mu}_{comb}) \leq \text{EQM}(\hat{\mu}_{EV})$ à condition que $v_b \leq \epsilon^2$. Cela indique que l'estimateur de substitution présente un meilleur rendement que celui d'Elliott et Haviland (2007) lorsque la variance de l'estimateur du biais est relativement faible par rapport au carré du biais. Une construction particulière de l'estimateur du biais $\hat{\epsilon}$ est présentée dans l'application sur des données réelles à la section 5.2.

4. Description des données

Pour la présente analyse, nous utilisons les six enquêtes issues de l'étude Benchmarking Study réalisée par le Pew Research Center en 2021. Les données et le rapport détaillé sont disponibles sur le site Web du Pew (lien). Les enquêtes de cette étude ont été réalisées entre le 14 juin et le 21 juillet 2021. Elles comprennent des interviews menées auprès de 29 937 adultes des États-Unis, soit environ 5 000 répondants par échantillon. Trois des enquêtes proviennent de différents panels en ligne à base probabiliste, dont l'American Trends Panel (ATP) du Pew. Les trois autres enquêtes proviennent de trois fournisseurs différents d'échantillons en ligne à participation volontaire. Ces enquêtes ont toutes été réalisées sur des plateformes en ligne.

Le tableau 4.1, tiré de Kennedy et coll. (2024), présente des renseignements détaillés sur la taille des échantillons et les périodes de collecte. Un questionnaire commun a été utilisé, en anglais ou en espagnol, pour les six enquêtes. Des renseignements démographiques, tels que l'âge, le groupe ethnique, le genre, le niveau de scolarité et l'état matrimonial, ont été recueillis, de même que des réponses à d'autres questions portant notamment sur la couverture d'assurance, l'usage du tabac, la tension artérielle, la citoyenneté américaine et le nombre d'enfants et d'adultes dans le ménage, etc. Ces six ensembles de données d'enquête peuvent être regroupés en deux types (nps et ps) comme suit :

i. Enquêtes non probabilistes (nps) : trois des six enquêtes sont des enquêtes à participation volontaire menées à des fins commerciales (panel volontaire 1, 2, 3). Les données ont été obtenues auprès de fournisseurs qui ont utilisé un échantillonnage par quotas. Le Pew a fourni aux fournisseurs des cibles de quotas fondées sur l'âge × le genre, le groupe ethnique × l'origine hispanique et le niveau de scolarité, selon l'enquête ACS de 2019. Dans ce cas, la collecte des données a été effectuée par l'entreprise Ipsos. Comme il s'agit de panels en ligne non probabilistes, les probabilités de sélection attribuées aux répondants sont inconnues.

ii. Enquêtes probabilistes (ps) : les trois autres enquêtes sont des enquêtes probabilistes. Même si les panels sont en ligne, les participants sont recrutés hors ligne; par exemple, au moyen d'un échantillonnage probabiliste. Plus précisément, le recrutement s'effectue à partir d'un échantillonnage fondé sur les adresses, établi à partir du fichier informatisé des adresses de livraison du service postal des États-Unis (U.S. Postal Service Computerized Delivery Sequence File). Ainsi, conformément à la nomenclature de Kennedy et coll. (2024), nous désignons ces enquêtes par les les noms ABS 1, 2, 3. Les taux de réponse propres à chaque étude pour les enquêtes ABS 1, 2, 3 étaient respectivement de 61 %, de 90 % et de 71 %.

Tableau 4.1
Tailles d'échantillon et périodes de collecte de six enquêtes du Pew

Type d'échantillon	Nom de l'échantillon	Identifiant	Taille de l'échantillon	Période de collecte
Probabiliste	Panel ABS 1	P_1	5 027	14 au 28 juin 2021
	Panel ABS 2	P_2	5 147	14 au 27 juin 2021
	Panel ABS 3	P_3	4 965	29 juin au 21 juillet 2021
Non probabiliste	Panel volontaire 1	O_1	4 912	15 au 25 juin 2021
	Panel volontaire 2	O_2	4 931	11 au 27 juin 2021
	Panel volontaire 3	O_3	4 955	11 au 26 juin 2021

Note : Address-based sampling (échantillonnage fondé sur les adresses) (ABS).
Source : Kennedy, Mercer et Lau (2024).

Pondération : Kennedy et coll. (2024) ont utilisé une approche standard de pondération par calage pour les six enquêtes. Pour l'enquête nps, du fait de l'absence de poids de sondage, un poids initial de 1 a été attribué à chaque répondant. Pour l'enquête ps, les poids de base du panel ont été corrigés afin de tenir compte des probabilités différentielles de sélection, puis utilisés dans le processus de calage. Plusieurs variables de contrôle démographique (tirées des enquêtes ACS, CPS et National Public Opinion Reference Survey du Pew), comme l'âge × sexe, la race/l'appartenance ethnique × le niveau de scolarité, etc., ont été utilisées pour caler les poids initiaux de chaque échantillon sur un ensemble commun de totaux de contrôle de la population.

Données repères : certaines questions d'enquête ont été retenues aux fins d'analyse de comparaison avec les données repères, lesquelles provenaient de sources comme les enquêtes National Health and Nutrition Examination Survey (NHANES), National Health Interview Survey (NHIS), ACS et CPS ainsi que l'outil de suivi des données de l'enquête sur la COVID du Pew et des données de projection relatives aux élections. Ces valeurs de référence sont fournies à l'échelle nationale, ainsi que selon certaines catégories de variables démographiques, comme l'âge (trois catégories : 18 à 29 ans, 30 à 64 ans, 65 ans et plus), le groupe ethnique (trois catégories : Blancs non hispaniques, Noirs non hispaniques et Hispaniques) et le niveau de scolarité (trois catégories : diplôme d'études secondaires ou niveau de scolarité inférieur, études collégiales ou universitaires partielles, diplôme d'un collège ou d'une université). Aux fins de l'analyse de l'erreur de mesure, ces valeurs de référence sont comparées aux estimations issues des six enquêtes pour les variables réponses d'intérêt, soit globalement ou selon les sous-groupes démographiques susmentionnés. Nous précisons cette analyse dans la section Évaluation qui suit.

4.1 Évaluation

Afin d'évaluer les estimateurs tant du point de vue du biais de représentativité que des différences de mesure, nous utilisons l'écart quadratique moyen (MSD pour « Mean Squared Deviation/Difference » en anglais) comme critère d'évaluation, un MSD plus faible indiquant un meilleur rendement de l'estimateur. Nous définissons cette erreur en fonction de l'écart par rapport aux valeurs de référence comme suit. Soit m, le nombre de variables de référence pour une enquête (ps ou nps) et k le nombre de sous-groupes fondés sur des facteurs d'intérêt (c'est-à-dire des variables démographiques, telles que l'âge, le groupe ethnique, etc.) Si $\overline{Y}_{jc}$ est la valeur pour la question j et le sous-groupe c provenant de la source de référence (considérée comme la valeur réelle) et $\hat{\overline{Y}}_{jc}$ est l'estimation correspondante provenant de l'enquête (où $j = 1,\ldots,m;\ c = 1,\ldots,k$), alors le MSD pour l'estimateur $\hat{\overline{Y}}$ est défini comme suit :

$$\mathrm{MSD}(\hat{\overline{Y}}) \equiv m^{-1} \sum_{j=1}^{m} \left[k^{-1} \sum_{c=1}^{k} \left(\hat{\overline{Y}}_{jc} - \overline{Y}_{jc} \right)^2 \right]. \tag{4.1}$$

À titre d'exemple, pour évaluer l'incidence de l'amélioration du biais de sélection, nous calculons $\hat{\mu}_{CLW}$ de nps et calculons $\mathrm{MSD}(\hat{\mu}_{CLW})$. Nous comparons $\mathrm{MSD}(\hat{\mu}_{CLW})$ à $\mathrm{MSD}(\hat{\mu}_{cal})$ de ps. Les évaluations peuvent être effectuées à l'échelle globale de l'enquête, comme dans (4.1), ou au niveau d'un sous-groupe selon un facteur d'intérêt, ou encore au niveau d'une question, ou des deux, en retirant le facteur de calcul

de la moyenne en conséquence. Dans la présente analyse des données, seules des questions binaires sont prises en compte, et le niveau de la question n'est pas considéré comme composante additionnelle dans le calcul du MSD.

5. Analyse des données

Avant de procéder à notre analyse empirique, nous tenons compte des principaux résultats de Kennedy et coll. (2024). Les auteurs utilisent la mesure de l'erreur absolue moyenne (EAM) et observent que les valeurs de l'EAM sont plus élevées dans les enquêtes à participation volontaire que dans les enquêtes fondées sur les adresses (ABS). À l'échelle des sous-groupes, les valeurs de l'EAM présentent un profil distinct dans les enquêtes à participation volontaire, profil qui est absent des enquêtes ABS. Par exemple, chez les jeunes de 18 à 29 ans et les adultes hispaniques, les valeurs de l'EAM sont plus élevées dans les enquêtes à participation volontaire, probablement en raison de « fausses réponses ». Ce phénomène est plus marqué pour les questions portant sur les prestations gouvernementales telles que la sécurité sociale, les bons alimentaires, les prestations de chômage ou les indemnités d'accident du travail. Une forte proportion de jeunes de 18 à 29 ans et d'adultes hispaniques déclarent avoir reçu au moins trois de ces quatre types de prestations, ce qui est très rare dans la population réelle. Selon Kennedy et coll. (2024), dans les enquêtes à participation volontaire, cette proportion se situe entre 6 % et 11 %, alors que l'incidence réelle dans la population n'est que de 0,1 %. Le fait de connaître ces résultats nous incite à recourir à des méthodes avancées de pondération pour les échantillons à participation volontaire, conjointement avec des corrections de biais liées aux différences de mesure. Dans les sections qui suivent, nous décrivons nos méthodes d'analyse des données. L'ensemble de notre code est disponible sur ce site Web (lien).

5.1 Estimateurs par pondération selon la propension inverse à partir d'échantillons d'enquêtes à participation volontaire

Nous utilisons la méthode présentée à la section 2, en nous appuyant sur le paquet `R nonprobsvy` pour calculer $\hat{\mu}_{CLW}$ pour trois échantillons d'enquêtes à participation volontaire O_1, O_2, O_3, avec respectivement P_1, P_2, P_3 comme référence. Nous soulignons qu'il n'existe aucun appariement intrinsèque entre les enquêtes O_j et P_j ; c'est-à-dire que toute combinaison de O_j et P_k, pour $j, k = 1, 2, 3$, peut être utilisée afin d'améliorer la représentativité. Dans notre analyse, nous utilisons les indices appariés O_j et P_j pour $j = 1, 2, 3$, mais nous nous attendons à obtenir des résultats similaires avec d'autres combinaisons. Cette attribution garantit également l'indépendance entre les paires, ce qui est essentiel pour estimer $\hat{\epsilon}$ dans l'estimateur composite (3.13) au moyen de sources de données externes de manière indépendante. Dans le modèle logistique du score de propension, nous considérons l'âge, le genre, le groupe ethnique, le niveau de scolarité et la région comme variables auxiliaires. Pour les échantillons d'enquêtes à participation volontaire, nous comparons deux estimateurs en termes de MSD : l'un est la moyenne pondérée de l'enquête reposant sur les poids calés du Pew ($\hat{\mu}_{cal}$) et l'autre est l'estimateur par PPI ($\hat{\mu}_{CLW}$). Pour les échantillons ABS, seul le premier estimateur s'applique ($\hat{\mu}_{cal}$). La liste des estimateurs que nous comparons dans cette analyse de données, accompagnée des formules détaillées, est présentée au tableau 5.1. Pour ces estimateurs,

les valeurs de MSD sont calculées à l'aide de (4.1) avec $m = 12$ questions de référence comportant uniquement des réponses « Oui » / « Non ». Des renseignements détaillés relatifs à ces questions figurent au tableau 5.2.

Tableau 5.1
Liste des estimateurs de la moyenne d'une population finie, avec le numéro de l'équation correspondante pour la formule détaillée et les types d'enquêtes

N°	Estimateur	Renseignements sur la méthode d'estimation	Formule	Type d'enquête
1	$\hat{\mu}_{cal}$	Moyenne pondérée à l'aide de poids de sondage ou de poids calés	(2.2)	*ps* ou *nps*
2	$\hat{\mu}_{CLW}$	Estimateur par PPI de Chen, Li et Wu (2020)	(2.1)	*nps*
3	$\hat{\mu}_{bc;CLW}$	Version corrigée du biais de différence de mesure de $\hat{\mu}_{CLW}$, où le biais est calculé à partir de (5.1)	(5.2)	*nps*
4	$\hat{\mu}_{EV}$	Estimateur composite d'Elliott et Haviland (2007), où le biais est calculé à partir de (5.1)	(3.2)	*ps* et *nps*
5	$\tilde{\mu}_{comb}$	Estimateur composite proposé combinant $\hat{\mu}_{cal}$ et $\hat{\mu}_{bc;CLW}$, où le biais est calculé à partir de (5.1)	(3.13)	*ps* et *nps*
6	$\hat{\mu}_{AA;CLW}$	Estimateur par PPI de Chen et coll. (2020) au moyen des réponses prédites	(5.4)	*ps* et *nps*
7	$\hat{\mu}_{AA;comb}$	Estimateur composite proposé combinant $\hat{\mu}_{cal}$ et $\hat{\mu}_{AA;CLW}$	(5.5)	*ps* et *nps*
8	$\hat{\mu}_{AA;EV}$	Version de $\hat{\mu}_{EV}$, où le biais correspond à la différence entre $\hat{\mu}_{cal}$ et $\hat{\mu}_{CLW}$	(5.6)	*ps* et *nps*

Notes : - Dans la colonne 5, « ou » indique qu'un estimateur est défini soit pour *ps* soit pour *nps* et « et » indique qu'un estimateur est défini en utilisant à la fois *ps* et *nps*.
- Apprentissage automatique (AA); Chen, Li et Wu (CLW); Elliott et Haviland (EV); pondération selon la propension inverse (PPI).

Tableau 5.2
Renseignements relatifs aux 12 questions (à réponse « Oui » ou « Non » seulement) utilisées comme référence

N° de question		Libellé de la question (traductions libres)
1.	Assurances	Êtes-vous actuellement couvert(e) par une forme quelconque d'assurance maladie ou de régime de soins de santé?
2.	Tension artérielle	Un médecin ou un autre professionnel de la santé vous a-t-il déjà dit que vous faisiez de l'hypertension, aussi appelée tension ou pression artérielle élevée?
3.	Parent	Êtes-vous le parent ou le tuteur d'un ou de plusieurs enfants de moins de 18 ans?
4.	Allergies alimentaires	Avez-vous des allergies alimentaires?
5.	Emploi l'année passée	Avez-vous occupé un emploi ou travaillé à votre compte à un moment quelconque en 2020?
6.	Régime de retraite	À un moment quelconque en 2020, aviez-vous un régime d'épargne-retraite tel qu'un 401(k), un 403(b), un IRA (régime individuel d'épargne-retraite) ou un autre compte conçu expressément pour l'épargne en vue de la retraite?
7.	Prestations d'assurance chômage	À un moment quelconque en 2020, avez-vous reçu des prestations d'assurance chômage provinciales ou fédérales?
8.	Indemnité d'accident du travail	En 2020, avez-vous reçu des paiements au titre d'une indemnité pour accident de travail ou d'autres paiements à la suite d'une blessure ou d'une maladie liée au travail?
9.	Bons alimentaires	À un moment quelconque en 2020, avez-vous reçu, vous ou une personne de votre ménage, des prestations du SNAP (programme d'aide alimentaire complémentaire) ou du programme de bons alimentaires, ou utilisé une carte de prestations SNAP ou des bons alimentaires?
10.	Sécurité sociale	En 2020, avez-vous reçu des prestations de la sécurité sociale du gouvernement des États-Unis?
11.	Syndicat	Êtes-vous membre d'un syndicat ou d'une association de travailleurs semblable à un syndicat?
12.	Citoyenneté américaine	Êtes-vous citoyen(ne) des États-Unis?

Source : Tiré du questionnaire publié sur le site Web du Pew Research Center https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/(lien).

Au tableau 5.3, nous comparons $\hat{\mu}_{cal}$ et $\hat{\mu}_{CLW}$, en termes de MSD, exprimé selon une échelle de 10^2, afin d'en faciliter la représentation. Nous constatons que, pour les enquêtes à participation volontaire, $\text{MSD}(\hat{\mu}_{CLW})$ est supérieur à $\text{MSD}(\hat{\mu}_{cal})$. Par exemple, pour O_1, $\text{MSD}(\hat{\mu}_{CLW})$ est 1,168, alors que

$\text{MSD}(\hat{\mu}_{cal})$ est 1,039. Ainsi, aucune amélioration n'est observée, au niveau global de l'enquête, dans les enquêtes à participation volontaire lorsque la méthode PPI est utilisée plutôt que des poids calés. En conséquence, comparativement aux enquêtes ABS, les valeurs de MSD des enquêtes à participation volontaire demeurent très élevées. Par exemple, pour O_1, $\text{MSD}(\hat{\mu}_{CLW})$ est 1,168, alors que pour P_1, $\text{MSD}(\hat{\mu}_{cal})$ est 0,080. Ensuite, au tableau 5.4, nous comparons les valeurs de MSD des mêmes estimateurs selon les sous-groupes d'âge et de groupe ethnique. Nous observons ici que, sauf dans quelques cas, $\text{MSD}(\hat{\mu}_{CLW})$ est supérieur à $\text{MSD}(\hat{\mu}_{cal})$. Les exceptions comprennent 6 cas sur 18 : le groupe d'âge de 18 à 29 ans pour O_1, le groupe d'âge de 65 ans et plus pour O_2 et O_3, la population blanche pour O_1, et la population noire pour O_2 et O_3. Enfin, une conclusion semblable à celle du tableau 5.3 peut être tirée : l'utilisation de la méthode PPI ne présente pas d'amélioration par rapport à l'estimateur fondé sur des poids calés. De plus, la comparaison entre les enquêtes à participation volontaire et les enquêtes ABS montre que des biais propres à certains sous-groupes sont présents dans les échantillons d'enquêtes à participation volontaire, comme l'indique le Pew. Par exemple, pour O_1 dans le groupe d'âge de 18 à 29 ans, $\text{MSD}(\hat{\mu}_{CLW})$ est 2,312, comparativement à une valeur de 0,158 de $\text{MSD}(\hat{\mu}_{cal})$ pour la même catégorie dans P_1.

Tableau 5.3
Valeurs du MSD (selon une échelle de 10^2) de deux estimateurs de moyenne de population, $\hat{\mu}_{cal}$ et $\hat{\mu}_{CLW}$, calculées à partir de 12 questions de référence provenant de six enquêtes

Estimateur	P_1	P_2	P_3	O_1	O_2	O_3
$\text{MSD}(\hat{\mu}_{cal})$	0,080	0,210	0,097	1,039	1,024	0,480
$\text{MSD}(\hat{\mu}_{CLW})$	-	-	-	1,168	1,237	0,564

Notes : - Consulter le tableau 5.1 pour obtenir la définition ou la formule des estimateurs et la définition du MSD dans l'équation (4.1).
- Chen, Li et Wu (CLW); Mean Squared Deviation/Difference (écart quadratique moyen) (MSD).

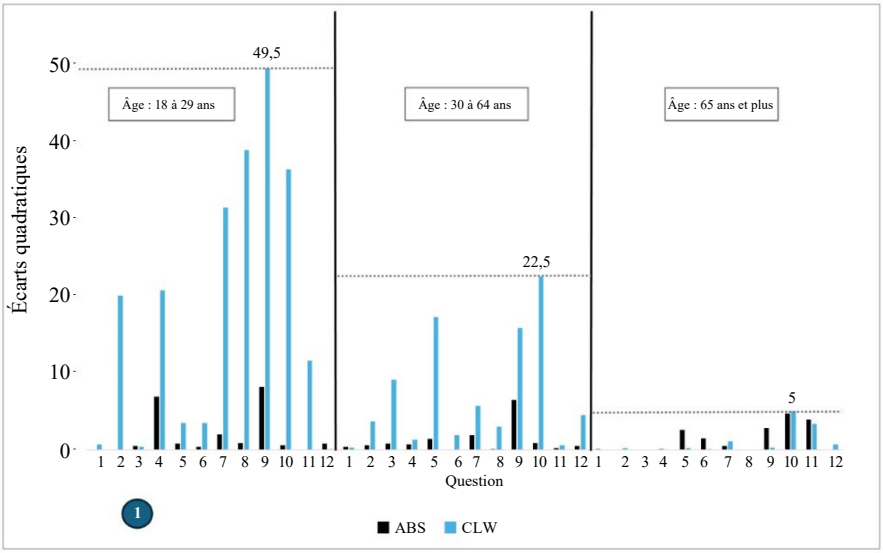
Tableau 5.4
Valeurs du MSD au niveau des sous-groupes (selon une échelle de 10^2) de $\hat{\mu}_{cal}$ et $\hat{\mu}_{CLW}$ pour trois sous-groupes d'âge (18 à 29 ans, 30 à 64 ans, 65 ans et plus) et trois sous-groupes ethniques (Blancs, Noirs et Hispaniques)

	Facteur	Groupe	P_1	P_2	P_3	O_1	O_2	O_3
$\text{MSD}(\hat{\mu}_{cal})$	âge	18 à 29 (ans)	0,158	0,507	0,177	2,869	3,162	1,790
		30 à 64 (ans)	0,104	0,239	0,117	1,345	1,407	0,567
		65 et plus (ans)	0,130	0,185	0,135	0,189	0,162	0,102
$\text{MSD}(\hat{\mu}_{CLW})$	âge	18 à 29 (ans)	-	-	-	2,312	3,901	1,930
		30 à 64 (ans)	-	-	-	1,497	1,705	0,728
		65 et plus (ans)	-	-	-	0,318	0,146	0,097
$\text{MSD}(\hat{\mu}_{cal})$	groupe ethnique	Blancs	0,064	0,152	0,090	1,034	0,907	0,390
		Noirs	0,281	0,502	0,376	1,099	1,121	0,499
		Hispaniques	0,102	0,375	0,236	2,625	2,341	2,147
$\text{MSD}(\hat{\mu}_{CLW})$	groupe ethnique	Blancs	-	-	-	0,913	1,136	0,471
		Noirs	-	-	-	1,351	0,877	0,464
		Hispaniques	-	-	-	3,441	2,781	2,444

Notes : - Les estimations pour les adultes blancs et noirs sont fondées sur les personnes qui ne déclarent pas être hispaniques; les valeurs pour les catégories « asiatiques/autres groupes ethniques » (représentant moins de 5 % des répondants dans chaque enquête) ne sont pas présentées.
- Chen, Li et Wu (CLW); Mean Squared Deviation/Difference (écart quadratique moyen) (MSD).

Pour comprendre plus en détail l'effet de la méthode PPI sur la réduction du biais de sélection, nous calculons l'écart quadratique entre les estimateurs ($\hat{\mu}_{CLW}$ et $\hat{\mu}_{cal}$) à partir des valeurs de référence pour trois sous-groupes d'âge et pour 12 questions. Dans la présente section, nous décrivons uniquement les résultats d'écart quadratique pour les enquêtes O_3 et P_3, car dans la section suivante, nous utiliserons les autres enquêtes (O_1, O_2 et P_1, P_2) afin de construire un estimateur du biais. Nous utiliserons désormais des exposants sur les estimateurs pour indiquer la source de l'échantillon, par exemple P_1, P_2, P_3 ou O_1, O_2, O_3. À la figure 5.1, les barres bleu pâle représentent les écarts quadratiques pour $\hat{\mu}_{CLW}^{O_3}$ et les barres noires ceux pour $\hat{\mu}_{cal}^{P_3}$. Nous indiquons la plus grande valeur des barres bleues pour chaque groupe d'âge au moyen d'une ligne pointillée. Les valeurs d'écart quadratique figurant dans cette figure, ainsi que dans le reste des figures de la section 5, sont ajustées selon une échelle de 10^3, pour faciliter la visualisation. Des conclusions semblables à celles obtenues précédemment peuvent être tirées ici : parmi les trois groupes d'âge, le groupe des 18 à 29 ans présente le plus grand écart entre la hauteur des deux types de barres. On observe que la plus grande valeur d'écart quadratique pour $\hat{\mu}_{CLW}^{O_3}$ est 49,473 dans le groupe des 18 à 29 ans. Cette valeur provient de la question 9 portant sur les bons alimentaires. Le biais est souvent important dans les questions délicates, subjectives ou auxquelles il est difficile de répondre lorsqu'elles dépendent de la mémoire du répondant ou lorsque les définitions des concepts sont ambiguës. Selon ces critères, les estimations pour la question 9 sont plus susceptibles de présenter des écarts quadratiques élevés par rapport aux valeurs de référence. Dans l'ensemble, on constate que les barres bleues sont nettement plus hautes que les barres noires. Nous en concluons donc que l'amélioration de la représentativité à elle seule ne suffit pas à améliorer les estimations issues des enquêtes à participation volontaire.

Figure 5.1 Écarts quadratiques entre deux estimations de moyennes de population par rapport à la valeur de référence $\bar{Y}$, soit $(\bar{Y} - \hat{\mu}_{cal}^{P_3})^2$ en noir et $(\bar{Y} - \hat{\mu}_{CLW}^{O_3})^2$ en bleu, pour 12 questions et trois groupes d'âge (18 à 29 ans, 30 à 64 ans et 65 ans et plus)



Notes : - Les plus grandes valeurs des barres bleues dans chaque groupe d'âge sont indiquées par des lignes pointillées. L'axe des Y est ajusté selon une échelle de 10^3.
- Address-based sampling (échantillonnage fondé sur les adresses) (ABS); Chen, Li et Wu (CLW).

5.2 Estimateurs par pondération selon la propension inverse corrigés pour contrer le biais découlant des différences de mesure

Dans cette section, nous abordons le biais attribuable aux différences de mesure. Les résultats de Kennedy et coll. (2024) indiquent que l'ensemble des échantillons des enquêtes à participation volontaire présentent un biais similaire. Les auteurs confirment que ce biais de mesure est lié aux effets de l'âge et du groupe ethnique, ce que nous avons également vérifié par des modèles d'ANOVA (analyse de variance). Ainsi, suivant l'analyse empirique de Kennedy et coll. (2024), nous supposons que la distribution du biais est semblable entre les échantillons d'enquêtes à participation volontaire. Forts de cette connaissance préalable des facteurs à l'origine du biais et de l'hypothèse de similarité du biais entre les échantillons d'enquêtes à participation volontaire, nous calculons une estimation du biais à l'aide de O_1, O_2 et P_1, P_2, en utilisant O_3 pour la correction du biais et P_3 pour la validation. Rappelons que, selon le modèle défini à l'équation (3.1), $\hat{\mu}_{CLW} - \hat{\mu}_{cal}$ fournit un estimateur sans biais de ϵ. Nous définissons donc une estimation du biais au niveau de la question × groupe d'âge comme la moyenne des différences pour les paires, entre nps et ps. Pour la question j et le groupe d'âge c, on obtient ainsi

$$\begin{aligned}\hat{\epsilon}^{jc} &\equiv 4^{-1}\{(\hat{\mu}_{CLW}^{jc;O_1} - \hat{\mu}_{cal}^{jc;P_1}) + (\hat{\mu}_{CLW}^{jc;O_1} - \hat{\mu}_{cal}^{jc;P_2}) + (\hat{\mu}_{CLW}^{jc;O_2} - \hat{\mu}_{cal}^{jc;P_1}) + (\hat{\mu}_{CLW}^{jc;O_2} - \hat{\mu}_{cal}^{jc;P_2})\} \\ &= 2^{-1}\{(\hat{\mu}_{CLW}^{jc;O_1} - \hat{\mu}_{cal}^{jc;P_1}) + (\hat{\mu}_{CLW}^{jc;O_2} - \hat{\mu}_{cal}^{jc;P_2})\},\end{aligned} \tag{5.1}$$

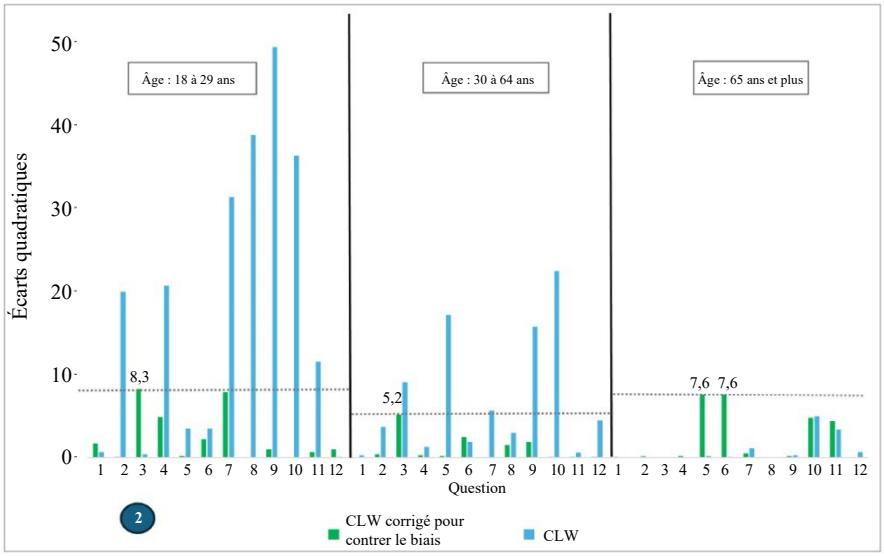
où $j = 1,\ldots,12;\ c = 1, 2, 3.$ En raison de la structure additive du biais dans le modèle défini à l'équation (3.1), nous calculons les estimations corrigées du biais pour la question j et le groupe d'âge c dans O_3 comme suit :

$$\hat{\mu}_{bc;CLW}^{jc;O_3} \equiv \hat{\mu}_{CLW}^{jc;O_3} - \hat{\epsilon}^{jc}, \quad j = 1,\ldots,12;\ c = 1, 2, 3. \tag{5.2}$$

Dans cette section, nous décrivons la méthode de correction du biais en mettant l'accent sur le facteur âge, mais une correction similaire peut être effectuée selon le facteur groupe ethnique. Nous ne considérons pas les deux facteurs simultanément (âge × groupe ethnique) pour la correction du biais, faute de points de référence à un niveau de désagrégation aussi fin.

Nous comparons $\hat{\mu}_{bc;CLW}^{jc;O_3}$ et $\hat{\mu}_{CLW}^{jc;O_3}$ en termes d'écarts quadratiques par rapport aux valeurs de référence afin d'évaluer l'effet de la correction du biais de mesure lié à l'âge. À la figure 5.2, tout en conservant la cohérence des couleurs avec la figure précédente, nous traçons les écarts quadratiques de $\hat{\mu}_{CLW}^{jc;O_3}$ sous forme de barres bleues et ceux des estimations corrigées du biais $\hat{\mu}_{bc;CLW}^{jc;O_3}$ sous forme de barres vertes. La plus haute barre verte de chaque groupe d'âge est indiquée par une ligne pointillée. On constate que, dans la plupart des cas (c'est-à-dire pour la cellule question × groupe d'âge), les barres vertes sont plus courtes que les barres bleues. La plus haute barre verte dans le groupe d'âge 18 à 29 ans est de 8,315, ce qui est nettement inférieur à la valeur maximale des barres bleues dans ce même groupe (49,472, comme indiqué à la figure 5.1). Ainsi, l'analyse empirique montre que la correction des différences de mesure, combinée à l'amélioration de la représentativité, permet d'obtenir des estimations plus précises.

Figure 5.2 Écarts quadratiques entre deux estimations de $\hat{\mu}_{CLW}^{O_3}$ et leurs versions corrigées pour contrer le biais $\hat{\mu}_{bc;CLW}^{O_3}$, soit $(\bar{Y} - \hat{\mu}_{CLW}^{O_3})^2$ en bleu et $(\bar{Y} - \hat{\mu}_{bc;CLW}^{O_3})^2$ en vert, pour 12 questions et trois groupes d'âge (18 à 29 ans, 30 à 64 ans et 65 ans et plus)



Notes : - Les plus grandes valeurs des barres vertes dans chaque groupe d'âge sont indiquées par des lignes pointillées. L'axe des Y est ajusté selon une échelle de 10^3.
- Chen, Li et Wu (CLW).

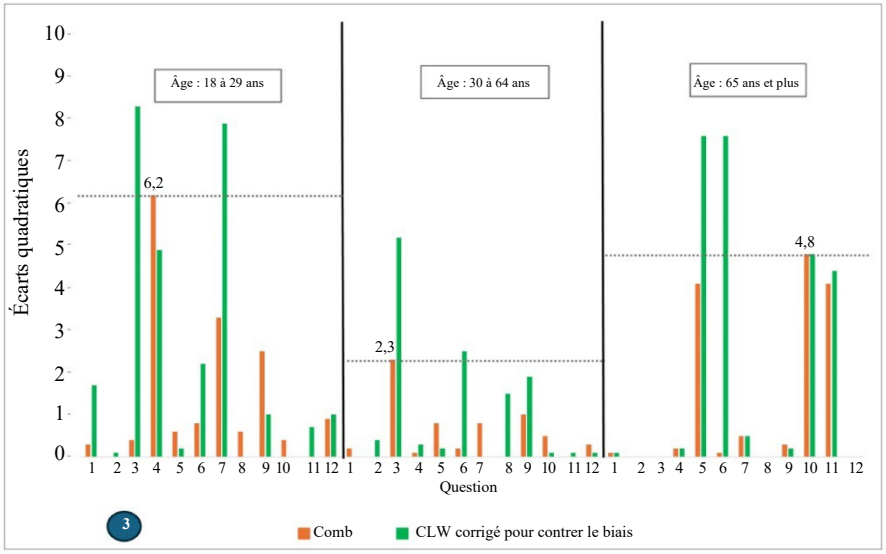
5.3 Estimateur composite à partir d'échantillons ABS et à participation volontaire

Nous calculons les estimateurs composites $\hat{\mu}_{EV}$ et $\tilde{\mu}_{comb}$ décrits à la section 3. Pour assurer une comparaison équitable, nous utilisons pour les deux estimateurs la même estimation du biais $\hat{\epsilon}$, telle que définie à la section 5.2 (équation 5.1), ainsi que les valeurs de v_1, v_2 correspondant aux variances estimées de $\hat{\mu}_{CLW}$ et $\hat{\mu}_{cal}$ (obtenues respectivement à l'aide des paquets `R nonprobsvy` et `survey`). Nous présentons deux figures dans cette section : figures 5.3 et 5.4. Dans la première, nous comparons l'estimateur composite proposé ($\tilde{\mu}_{comb}$) à l'estimateur corrigé du biais de la section précédente ($\hat{\mu}_{bc:CLW}^{O_3}$). Dans la seconde, nous comparons notre estimateur composite proposé ($\tilde{\mu}_{comb}$) à l'estimateur composite déjà existant $\hat{\mu}_{EV}$. À la figure 5.3, nous représentons en barres orange les écarts quadratiques de $\tilde{\mu}_{comb}$ par rapport à la valeur de référence, et en barres vertes ceux de $\hat{\mu}_{bc:CLW}^{O_3}$ (comme à la figure 5.2). En moyenne, les barres orange sont plus basses que les barres vertes, ce qui indique que l'estimateur composite proposé donne de meilleurs résultats que l'estimateur par PPI corrigé pour contrer le biais. La valeur maximale de la barre orange pour le groupe d'âge de 18 à 29 ans est de 6,182, comparativement à 8,315 pour la barre verte (voir la figure 5.2).

À la figure 5.4, nous comparons de façon analogue $\tilde{\mu}_{comb}$ et $\hat{\mu}_{EV}$ selon leurs écarts quadratiques avec les valeurs de référence. Là encore, en moyenne, les valeurs correspondant à $\tilde{\mu}_{comb}$ (barres orange) sont inférieures à celles de $\hat{\mu}_{EV}$ (barres bleu foncé). Les lignes pointillées montrent que le plus important écart quadratique par rapport à la valeur de référence est encore plus petit pour $\tilde{\mu}_{comb}$ que pour $\hat{\mu}_{EV}$ dans les groupes d'âge de 18 à 29 ans et de 30 à 64 ans. En revanche, dans le groupe de 65 ans et plus, le rendement

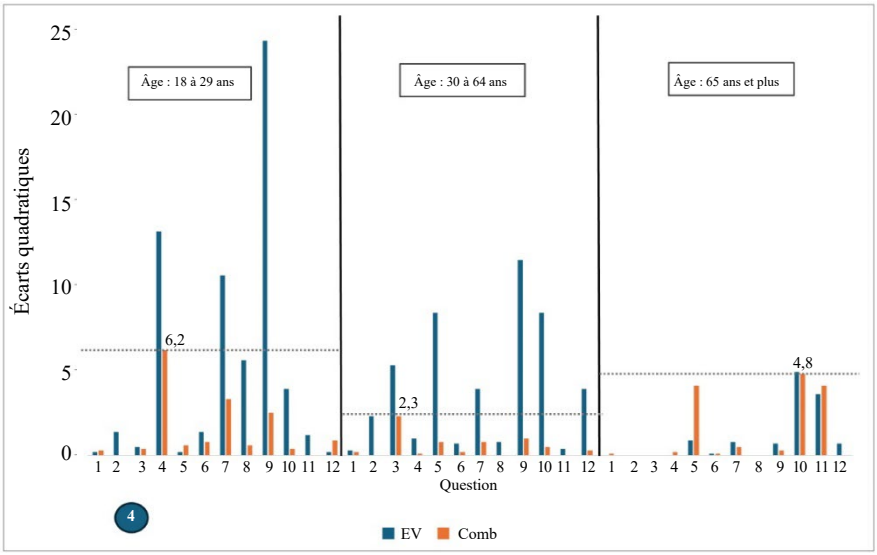
de $\tilde{\mu}_{comb}$ et $\hat{\mu}_{EV}$ est similaire. Nous concluons donc que l'estimateur composite proposé offre un meilleur rendement que celui décrit antérieurement dans la littérature.

Figure 5.3 Écarts quadratiques par rapport à la valeur de référence de $\hat{\mu}_{bc:CLW}^{O_3}$ (à l'aide de O_3) et de $\tilde{\mu}_{comb}$ (à l'aide de O_3 et P_3), soit $(\bar{Y} - \hat{\mu}_{bc:CLW}^{O_3})^2$ en vert et $(\bar{Y} - \tilde{\mu}_{comb})^2$ en orange, pour 12 questions et trois groupes d'âge (18 à 29 ans, 30 à 64 ans et 65 ans et plus)



Notes : - Les plus grandes valeurs des barres orange dans chaque groupe d'âge sont indiquées par des lignes pointillées. L'axe des Y est ajusté selon une échelle de 10^3.
- Chen, Li et Wu (CLW).

Figure 5.4 Écarts quadratiques par rapport à la valeur de référence des estimateurs composites $\hat{\mu}_{EV}$ et $\tilde{\mu}_{comb}$, soit $(\bar{Y} - \hat{\mu}_{EV})^2$ en bleu foncé et $(\bar{Y} - \tilde{\mu}_{comb})^2$ en orange, pour 12 questions et trois groupes d'âge (18 à 29 ans, 30 à 64 ans et 65 ans et plus)



Notes : - Les plus grandes valeurs des barres orange dans chaque groupe d'âge sont indiquées par des lignes pointillées. L'axe des Y est ajusté selon une échelle de 10^3.
- Elliott et Haviland (EV).

5.4 Comparaison de l'estimateur composite avec les enquêtes ABS

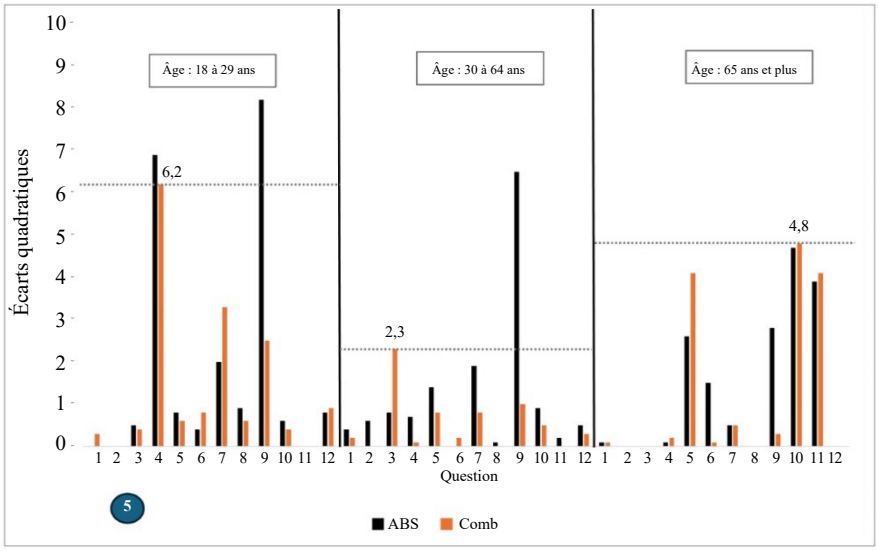
À la suite de l'exposé sur les estimateurs composites présentés à la section 5.3, on pourrait être tenté de se demander si l'estimateur composite proposé (fondé à la fois sur ps et nps) donne de meilleurs résultats que l'estimateur fondé uniquement sur ps. Dans la présente section, nous examinons si le rendement de l'estimateur composite peut être encore amélioré. La figure 5.5 illustre une comparaison des écarts quadratiques de $\tilde{\mu}_{comb}$ (barres orange) et de $\hat{\mu}_{cal}^{P_3}$ (barres noires), en maintenant la même codification de couleurs que dans les figures précédentes. Pour le groupe d'âge de 18 à 29 ans, on constate que l'estimateur fondé uniquement sur ps donne de meilleurs résultats que l'estimateur composite proposé pour certaines questions (à savoir les questions 1, 6, 7, 12). On observe également que pour la question 10 (portant sur la sécurité sociale) dans le groupe d'âge de 65 ans et plus, tous les estimateurs présentés à la section 5 affichent des valeurs similaires pour les écarts quadratiques par rapport à la valeur de référence (figures 5.1 à 5.5). Cela peut s'expliquer par le fait que les personnes âgées ou retraitées sont plus susceptibles de recevoir de telles prestations et moins enclines à fournir de fausses réponses. Par conséquent, les estimations fondées sur nps sont déjà très proches de ps.

Dans notre étude de cas, ps comporte déjà une taille d'échantillon importante (près de 5 000). Ainsi, une enquête ps à elle seule suffit généralement à estimer les caractéristiques d'une population finie. Afin d'expliquer les situations où le rendement de l'estimateur composite proposé s'avère nettement meilleur en termes de MSD, nous générons des échantillons d'enquête ps de plus petite taille. À cette fin, nous tirons des échantillons de P_3 selon une procédure d'échantillonnage stratifié avec répartition proportionnelle, les strates étant formées à partir d'une segmentation basée sur les poids d'enquête de P_3, soit (0; 0,5), (0,5; 1) et ainsi de suite. Nous calculons ensuite les valeurs de MSD de $\hat{\mu}_{EV}$, $\tilde{\mu}_{comb}$ et $\hat{\mu}_{cal}$ pour des échantillons plus petits tirés de P_3, tout en gardant l'échantillon original O_3 de nps inchangé. Il convient de mentionner que les estimateurs $\hat{\mu}_{CLW}$ et $\hat{\mu}_{bc;CLW}$ (calculés uniquement à partir de O_3) demeurent inchangés, puisque la taille d'échantillon de O_3 reste la même. Leurs valeurs respectives de MSD sont de 0,564 et de 0,183. Le tableau 5.5 présente les valeurs de MSD (ajustées selon une échelle de 10^2, comparables à celles du tableau 5.3), ainsi que la variation relative du MSD (en pourcentage, indiquée entre parenthèses) des estimateurs par rapport à $\hat{\mu}_{cal}$, définie comme suit :

$$\frac{\mathrm{MSD}(\hat{\mu}_{cal}) - \mathrm{MSD}(\hat{\mu})}{\mathrm{MSD}(\hat{\mu})} \times 100, \tag{5.3}$$

où $\hat{\mu} \in \{\hat{\mu}_{EV}, \tilde{\mu}_{comb}\}$.

Figure 5.5 Écarts quadratiques de $\hat{\mu}_{cal}^{P_3}$ (à l'aide de P_3) et de $\tilde{\mu}_{comb}$ (à l'aide de P_3 et O_3) avec la valeur de référence, soit $(\bar{Y} - \hat{\mu}_{cal}^{P_3})^2$ en noir et $(\bar{Y} - \tilde{\mu}_{comb})^2$ en orange, pour 12 questions et trois groupes d'âge (18 à 29 ans, 30 à 64 ans et 65 ans et plus)



Notes : - Les plus grandes valeurs des barres orange dans chaque groupe d'âge sont indiquées par des lignes pointillées. L'axe des Y est ajusté selon une échelle de 10^3.
- Address-based sampling (échantillonnage fondé sur les adresses) (ABS).

Tableau 5.5
Valeurs de MSD (selon une échelle de 10^2) de l'estimateur $\hat{\mu}_{cal}$ (à partir de P_3 et ses échantillons), $\tilde{\mu}_{comb}$ et $\hat{\mu}_{EV}$ (à partir de P_3 et ses échantillons, combinés avec O_3), partant de la taille d'échantillon originale (4 912) et diminuant jusqu'à 100

Taille de l'échantillon de ps	$\hat{\mu}_{cal}$	$\hat{\mu}_{EV}$	$\tilde{\mu}_{comb}$
4 912	0,097	0,337 (-71,284)	0,101 (-3,767)
1 000	0,231	0,295 (-21,718)	0,158 (45,896)
500	0,399	0,385 (3,642)	0,289 (38,293)
100	1,073	0,748 (43,476)	0,626 (71,399)

Notes : - Les nombres présentés entre parenthèses correspondent à la variation relative de l'écart quadratique moyen (MSD) (en %) des estimateurs par rapport à $\hat{\mu}_{cal}$, définie à l'équation (5.3).
- Elliott et Haviland (EV).

Nous observons que, lorsque la taille de l'échantillon de ps diminue, l'estimateur composite proposé $\tilde{\mu}_{comb}$ présente des valeurs de MSD inférieures à celles de $\hat{\mu}_{cal}$. Par exemple, pour une taille d'échantillon de ps de 100, $\text{MSD}(\tilde{\mu}_{comb})$ est 0,626, soit bien inférieure à la valeur de $\text{MSD}(\hat{\mu}_{cal})$ de 1,073. Cela se reflète également dans la variation relative de MSD entre ces deux estimateurs (71,399 %). Cependant, pour l'échantillon initial P_3, on observe que $\text{MSD}(\tilde{\mu}_{comb}) > \text{MSD}(\hat{\mu}_{cal})$ et la variation relative de MSD est de -3,767 %. Un phénomène similaire est observé pour $\hat{\mu}_{EV}$, où $\text{MSD}(\hat{\mu}_{EV}) < \text{MSD}(\hat{\mu}_{cal})$ pour des tailles d'échantillon de ps de 500 et de 100. Il est intéressant de souligner que l'écart MSD de $\hat{\mu}_{EV}$ est toujours supérieur à celui de $\tilde{\mu}_{comb}$. Enfin, nous concluons que l'estimateur composite proposé n'est pas supérieur à l'estimateur fondé uniquement sur ps, lorsque ps repose sur un échantillon de taille suffisante. Cependant,

lorsque la taille de l'échantillon de ps diminue, la variance de $\hat{\mu}_{cal}$ augmente, et l'estimateur composite $\tilde{\mu}_{comb}$ donne alors de meilleurs résultats que $\hat{\mu}_{cal}$.

5.5 Modèle prédictif pour la correction du biais

À la section 3, nous mentionnons que, pour calculer des estimateurs composites, il faut connaître le biais ϵ à partir de sources auxiliaires ou de données historiques. C'est pourquoi, à la section 5.2, nous estimons le biais à partir de O_1, O_2, P_1, P_2 et appliquons la correction du biais dans O_3. Il est important de mentionner que l'estimateur du biais proposé par Elliott et Haviland (2007) ne peut pas être appliqué à notre cas, puisque notre estimateur composite deviendrait alors identique à l'estimateur de ps lui-même; c'est-à-dire que, dans l'équation (3.13), en utilisant $\hat{\epsilon} = \bar{y}_2 - \bar{y}_1$, nous obtenons $\tilde{\mu}_{comb} = \bar{y}_1$. Lorsque d'autres sources d'information sur le biais ne sont pas disponibles pour obtenir $\hat{\epsilon}$, nous proposons une autre solution dans la présente section. À cette fin, nous ajustons un modèle sur P_3 (en associant les réponses aux variables auxiliaires) et, à l'aide des estimations paramétriques ajustées issues de ce modèle, nous prédisons les réponses au niveau des unités dans O_3 (ou, de façon équivalente, les probabilités de répondre « Oui »). Cette approche d'intégration des données peut être reliée à ce qui est proposé dans Sen et Lahiri (2025) dans le contexte de l'estimation d'une proportion dans une population finie. La différence dans ce cas est que la prédiction est faite sur l'enquête non probabiliste, quand dans Sen et Lahiri (2025) les réponses sont prédites pour une plus vaste enquête probabiliste.

Pour la construction du modèle, nous utilisons le paquet `R caret` (Kuhn, Wing, Weston, Williams, Keefer, Engelhardt, Cooper, Mayer, Kenkel, R Core Team, Benesty, Lescarbeau, Ziem, Scrucca, Tang, Candan et Hunt, 2020). À l'aide de `caret`, nous ajustons des modèles d'apprentissage automatique (AA), par forêt aléatoire (FA) et optimisation de gradient (OG), en utilisant la validation croisée (VC) comme méthode de rééchantillonnage et en optimisant des hyperparamètres, tels que le nombre d'itérations d'optimisation, la profondeur maximale de l'arborescence, le taux de rétrécissement, etc. Le tableau 5.6 présente les renseignements détaillés pertinents concernant la construction des modèles. Nous observons, d'après ce tableau, que les modèles FA et OG affichent un rendement similaire sur le plan de la précision (0,85) pour les données d'essai, mais que le modèle OG présente une valeur supérieure de l'aire sous la courbe (ASC) (0,92). Pour le choix du modèle final, nous tenons compte, en plus de l'ASC et de la précision, du critère d'écart MSD considéré dans les sections précédentes. Enfin, comme le modèle OG présente un MSD inférieur et une ASC supérieure à ceux du modèle FA, nous le retenons comme modèle prédictif final. Les valeurs finales des paramètres d'ajustement utilisées pour le modèle sont : `n.trees` = 1 000, `interaction.depth` = 3, `shrinkage` = 0,05 et `n.minobsinnode` = 10, et `logLoss` a été utilisé pour sélectionner le modèle optimal en utilisant la valeur la plus petite. Les niveaux de question figurent parmi les variables les plus importantes pour la classification. La figure 5.6 illustre la valeur `logLoss` selon différents paramètres d'ajustement ainsi que la fonction d'efficacité du récepteur (courbe ROC) des deux modèles de classification binaire (FA et OG).

Nous appliquons ensuite la même méthode PPI (décrite à la section 2) aux probabilités prédites issues du modèle retenu (OG) afin d'obtenir un nouvel estimateur de la moyenne de population de nps. Cet estimateur s'exprime comme suit :

$$\hat{\mu}_{AA;CLW} \equiv (\hat{N}^A)^{-1} \sum_{i \in S_A} (\hat{p}_i^y / \hat{\pi}_i^A), \tag{5.4}$$

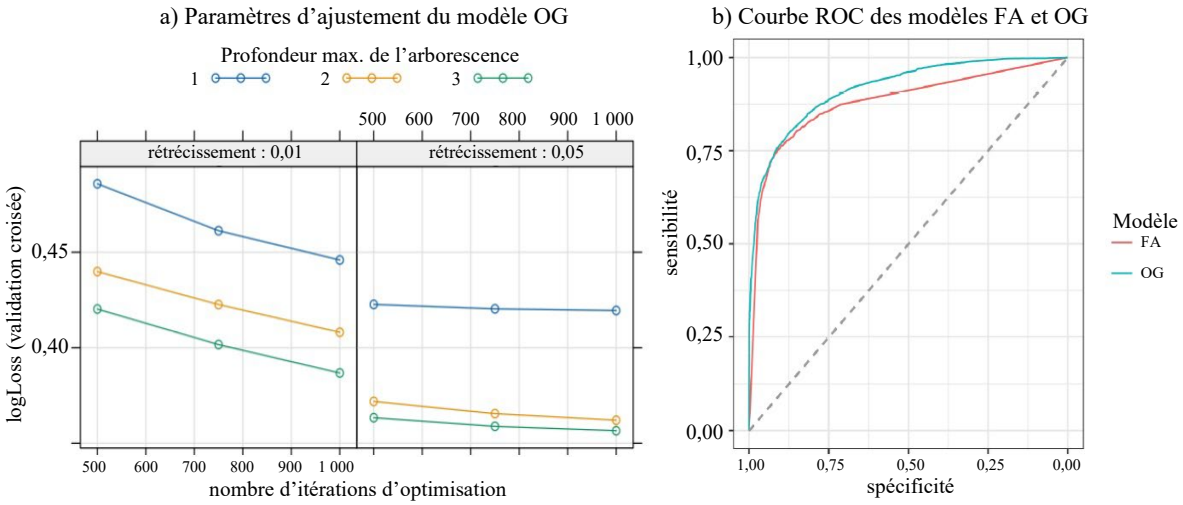
où $\hat{p}_i^y$ représente la probabilité estimée que la personne i réponde « Oui » à la variable y étant donné les caractéristiques auxiliaires de i, $\hat{\pi}_i^A$ est obtenu selon la méthode PPI, et $\hat{N}^A = \sum_{i \in S_A} (\hat{\pi}_i^A)^{-1}$. La différence entre $\hat{\mu}_{AA;CLW}$ défini dans (5.4) et $\hat{\mu}_{CLW}$ défini dans (2.1) est que, dans (5.4), on utilise des valeurs de probabilité estimées $(\hat{p}^y)$ plutôt que les réponses réelles (y).

Tableau 5.6
Renseignements détaillés relatifs à la modélisation prédictive sur l'échantillon P_3 de ps de 4 912 répondants, dans le but d'obtenir des estimations corrigées du biais à partir de l'échantillon O_3 de nps

Structure des données	Question × données au niveau du répondant
Taille (apprentissage + essai)	59 000 observations réparties en ensemble d'apprentissages (80 %) et d'essais (20 %)
Type de variable réponse	Binaire (les refus ne sont pas pris en compte)
Variables auxiliaires (nombre de niveaux)	Question (12), âge (3), groupe ethnique (5), niveau de scolarité (3), genre (3), région (4)
Modèles prédictifs (ASC, exactitude)	Forêt aléatoire (0,88, 0,85), optimisation de gradient (0,92, 0,85)

Note : Aire sous la courbe (ASC).

Figure 5.6 Résultats de la modélisation prédictive sur l'échantillon probabiliste P_3



Note : Forêt aléatoire (FA); optimisation de gradient (OG); Receiver Operating Characteristic (fonction d'efficacité du récepteur) (ROC).

Par conséquent, dans ce cas, le nouvel estimateur composite est construit comme dans (3.8), mais au moyen de (5.4) au lieu de (2.1). L'estimateur proposé $\hat{\mu}_{AA;comb}$ est défini comme suit :

$$\hat{\mu}_{AA;comb} \equiv \left(\frac{v_2}{v_1 + v_2}\right)\hat{\mu}_{cal} + \left(\frac{v_1}{v_1 + v_2}\right)\hat{\mu}_{AA;CLW}. \tag{5.5}$$

Nous comparons l'estimateur ci-dessus à une autre version de $\hat{\mu}_{EV}$, que nous notons $\hat{\mu}_{AA;EV}$. Bien qu'aucune modélisation ne soit utilisée dans la construction de cet estimateur, nous conservons le même indice « AA » pour faciliter la comparaison. Dans cet estimateur, le biais inconnu est estimé par la différence entre les moyennes pondérées provenant de ps et nps, comme le proposent Elliott et Haviland (2007). Ainsi, $\hat{\mu}_{AA;EV}$ est défini comme suit :

$$\hat{\mu}_{AA;EV} \equiv \frac{\{(\bar{y}_1 - \bar{y}_2)^2 + v_2\}\bar{y}_1 + v_1\bar{y}_2}{(\bar{y}_1 - \bar{y}_2)^2 + v_1 + v_2}. \tag{5.6}$$

Le tableau 5.7 présente les valeurs des deux estimateurs composites $\hat{\mu}_{AA;comb}$ (proposé) et $\hat{\mu}_{AA;EV}$ (existant) ainsi que leurs valeurs respectives de MSD pour les 12 questions retenues pour l'analyse. Des valeurs de référence sont également fournies avec les estimations. On observe que, pour toutes les questions, $\text{MSD}(\hat{\mu}_{AA;comb}) < \text{MSD}(\hat{\mu}_{AA;EV})$.

Tableau 5.7
Estimateurs composites $\hat{\mu}_{AA;comb}$ (proposé) et $\hat{\mu}_{AA;EV}$ (existant) et leurs valeurs d'écart quadratique moyen (MSD) (selon une échelle de 10^3) pour 12 questions binaires de référence présentées au tableau 5.2

Question	Référence	$\hat{\mu}_{AA;EV}$	MSD ($\hat{\mu}_{AA;EV}$)	$\hat{\mu}_{AA;comb}$	MSD ($\hat{\mu}_{AA;comb}$)
1. Assurances	90,800	88,979	4,450	90,755	4,447
2. Tension artérielle	31,100	37,725	5,271	35,709	5,109
3. Parent	26,000	22,336	10,189	25,977	10,099
4. Allergies alimentaires	9,400	14,038	0,984	12,823	0,966
5. Emploi l'année passée	64,200	58,262	20,105	63,476	19,817
6. Régime de retraite	49,900	49,866	25,016	51,119	25,015
7. Prestations d'assurance chômage	9,300	16,846	0,970	12,189	0,825
8. Indemnité d'accident du travail	0,400	6,637	0,284	1,144	0,010
9. Bons alimentaires	11,100	21,549	5,328	16,416	5,247
10. Sécurité sociale	21,800	29,898	26,574	26,189	26,413
11. Syndicat	5,600	10,410	0,657	8,055	0,595
12. Citoyenneté américaine	92,500	96,299	4,413	96,064	4,332

Note : Apprentissage automatique (AA); Elliott et Haviland (EV); Mean Squared Deviation/Difference (écart quadratique moyen) (MSD).

Enfin, nous résumons les deux cas de correction du biais :

Scénario 1 – le biais est calculé à partir d'autres sources;

Scénario 2 – les estimations corrigées du biais sont prédites par modélisation.

Une comparaison entre les deux scénarios ne serait pas équitable, étant donné que de meilleurs résultats sont attendus lorsque le biais est connu à partir d'autres sources. Le tableau 5.8 présente les valeurs de MSD d'estimateurs comparables dans ces deux scénarios. Dans le scénario 1, nous comparons $\tilde{\mu}_{comb}$, $\hat{\mu}_{EV}$, $\hat{\mu}_{CLW}$ et $\hat{\mu}_{bc;CLW}$. Alors que dans le scénario 2, nous comparons $\hat{\mu}_{AA;comb}$, $\hat{\mu}_{AA;EV}$ et $\hat{\mu}_{AA;CLW}$. On observe que, dans

les deux cas, les estimateurs proposés ($\tilde{\mu}_{comb}$ et $\hat{\mu}_{AA;comb}$) affichent des valeurs de MSD inférieures à celles des estimateurs composites existants ($\hat{\mu}_{EV}$ et $\hat{\mu}_{AA;EV}$) ainsi qu'à celles d'autres estimateurs ($\hat{\mu}_{CLW}$, $\hat{\mu}_{bc;CLW}$ et $\hat{\mu}_{AA;CLW}$). Il convient également de mentionner que la correction du biais est efficace dans les deux situations, puisqu'elle réduit de manière importante les valeurs de MSD. En consultant le tableau 5.3 de la section 5.1, nous constatons que la valeur de MSD de $\hat{\mu}_{cal}^{P_3}$ est de 0,097. Ainsi, les estimateurs composites proposés ($\tilde{\mu}_{comb}$, $\hat{\mu}_{AA;comb}$) parviennent à obtenir une précision comparable pour une enquête *nps* (combinée à une enquête *ps*) à celle que l'on obtiendrait normalement pour une enquête *ps* de taille similaire.

Tableau 5.8
Comparaison des valeurs de l'écart quadratique moyen (MSD) (selon une échelle de 10^2, fournies entre parenthèses) des estimateurs proposés et concurrents, calculées à partir de 12 questions de référence provenant de O_3 de *nps* et de P_3 de *ps* (dans le cas des estimateurs composites) pour deux cas de correction du biais

	Estimateur composite proposé	Estimateur composite existant	Autre
Scénario 1 :	$\tilde{\mu}_{comb}$ (0,101)	$\hat{\mu}_{EV}$ (0,337)	$\hat{\mu}_{CLW}$ (0,564) $\hat{\mu}_{bc;CLW}$ (0,183)
Scénario 2 :	$\hat{\mu}_{AA;comb}$ (0,092)	$\hat{\mu}_{AA;EV}$ (0,355)	$\hat{\mu}_{AA;CLW}$ (0,099)

Notes : - Consulter le tableau 5.1 pour obtenir de plus amples précisions sur les estimateurs.
- Apprentissage automatique (AA); Chen, Li et Wu (CLW); Elliott et Haviland (EV); Mean Squared Deviation/Difference (écart quadratique moyen) (MSD).

6. Discussion

Dans le présent article, nous abordons le problème de l'amélioration de l'inférence pour une population finie au moyen de la construction d'estimateurs composites intégrant des données provenant d'enquêtes probabilistes et non probabilistes. Notre objectif est d'atténuer le biais découlant de problèmes de représentativité et d'erreurs de mesure dans les enquêtes non probabilistes. À cette fin, nous proposons deux méthodes de correction du biais : i) la correction des différences de mesure, applicable lorsque des sources de données additionnelles autres qu'une enquête probabiliste ou non probabiliste sont disponibles; ii) la correction du biais fondée sur un modèle, au moyen de modèles d'apprentissage automatique avancés (méthode appropriée dans les situations où seules une enquête probabiliste et une enquête non probabiliste sont accessibles). Nous illustrons l'efficacité de ces approches au moyen d'une étude de cas fondée sur les données de Kennedy et coll. (2024), qui montre que les estimateurs proposés offrent une précision accrue par rapport aux grandes enquêtes gouvernementales considérées comme des références. La méthode i) constitue une solution simple sur le plan informatique, tandis que la méthode ii) est plus largement applicable en pratique, notamment dans les contextes où les sources de données sont limitées.

Nous démontrons également que l'estimateur composite proposé présente un rendement amélioré (plus précisément, un moindre écart MSD), lorsque la taille de l'échantillon de l'enquête probabiliste est relativement plus petite que celle de l'enquête non probabiliste. Toutefois, la réduction de la taille de l'échantillon probabiliste peut poser des défis en matière d'estimation sur petits domaines. Dans notre étude de cas, nous

observons que certains sous-groupes (définis selon des combinaisons de question × groupe d'âge) ne comportent aucune observation lorsque la taille de l'échantillon probabiliste est réduite considérablement. Dans de telles situations, les techniques de modélisation sur petits domaines, telles que celles présentées par Nandram et Rao (2024) et par Franco et Bell (2013), s'avèrent particulièrement utiles. Aborder cette question constitue une piste de recherche intéressante pour l'avenir.

La méthodologie proposée dans le présent document a été mise à l'essai dans le cadre d'une étude de cas portant sur des enquêtes à participation volontaire menées aux États-Unis. Le rendement de la méthode pourrait varier selon les pays, les modes de collecte ou les contextes culturels, où la nature des biais de sélection et des erreurs de mesure peut différer de façon marquée. En outre, la méthodologie proposée repose sur certaines hypothèses, décrites dans la présente étude, accompagnées de références traitant d'éventuels non-respects de ces hypothèses. Comme orientation pour de futurs travaux de recherche, il serait souhaitable de concevoir des estimateurs composites qui tiennent compte des erreurs de mesure tant dans les enquêtes probabilistes que non probabilistes. Dans l'ensemble, notre approche à double volet, qui vise à la fois les biais de sélection et les différences de mesure, offre un cadre prometteur pour améliorer la méthodologie d'enquête dans un contexte de dépendance croissante à l'égard d'enquêtes non probabilistes.

Remerciements

Les auteurs tiennent à remercier Andrew Mercer, spécialiste principal en méthodologie de recherche au Pew Research Center, pour son soutien à cette étude. Ils souhaitent également exprimer leur reconnaissance aux deux examinateurs anonymes et au rédacteur associé pour leurs commentaires constructifs concernant une version antérieure du présent article.

Bibliographie

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. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2020001/article/00001-fra.pdf.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. et Chu, K. (2024). Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada. *Techniques d'enquête*, 50(1), 87-121. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2024001/article/00004-fra.pdf.

Biemer, P. (2009). Chapter 12 – Measurement errors in sample surveys. *Handbook of Statistics*, 29, 281-315.

Breidt, F.J. et Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190-205.

Buelens, B. et van den Brakel, J.A. (2015). Measurement error calibration in mixed-mode sample surveys. *Sociological Methods & Research*, 44(3), 391-426.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Dagdoug, M., Goga, C. et Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542), 1234-1251.

Elliott, M.N. et Haviland, A. (2007). Utilisation d'un échantillon de convenance électronique comme complément à un échantillon probabiliste. *Techniques d'enquête*, 33(2), 233-238. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2007002/article/10498-fra.pdf.

Franco, C. et Bell, W.R. (2013). Applying bivariate binomial/logit normal models to small area estimation. JSM 2013 - Survey Research Methods Section.

Groves, R.M. et Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5), 849-879.

Hansen, M., Hurwitz, W. et Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute, 32nd Session*, 38, 359-374.

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

Jäckle, A., Roberts, C. et Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *Revue Internationale de Statistique*, 78(1), 3-20.

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J. et Asare-Marfo, D. (2021). Strategies for detecting insincere respondents in online polling. *Public Opinion Quarterly*, 85(4), 1050-1075.

Kennedy, C., Mercer, A. et Lau, A. (2024). Étude de l'hypothèse selon laquelle les répondants aux enquêtes non probabilistes en ligne menées à des fins commerciales répondent en toute bonne foi. *Techniques d'enquête*, 50(1), 5-26. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2024001/article/00013-fra.pdf.

Kern, C., Klausch, T. et Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, 13(1), 73-93.

Kim, J.K. et Morikawa, K. (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. *Calcutta Statistical Association Bulletin*, 75(1), 8-27.

Kim, J.K., Park, S., Chen, Y. et Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3), 941-963.

Kim, J.K. et Tam, S.M. (2021). Data integration by combining big data and survey sample data for finite population inference. *Revue Internationale de Statistique*, 89(2), 382-401.

Klausch, T., Schouten, B., Buelens, B. et van den Brakel, J. (2017). Adjusting measurement bias in sequential mixed-mode surveys using re-interview data. *Journal of Survey Statistics and Methodology*, 5(4), 409-432.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. et Hunt, T. (2020). Package 'caret'. *The R Journal*, 223(7), 48.

Lee, S. et Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3), 319-343.

Little, R.J. et Rubin, D.B. (2019). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Mercer, A. et Lau, A. (2023). Comparing two types of online survey samples. *Pew Research Center*. Accessible à l'adresse : https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/.

Nandram, B. et Rao, J.N.K. (2024). Bayesian integration for small areas by supplementing a probability sample with a non-probability sample. *Statistics and Applications*, ISSN 2454-7395 (en ligne), 22(1), 343-374.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 123-150.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, 3(4), 425-483.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83(1), 242-272.

Rao, J.N.K. et Molina, I. (2015). *Small Area Estimation, 2nd Edition*. New York: John Wiley & Sons, Inc.

Rivers, D. (2007). Sampling for web surveys. *Joint Statistical Meetings*, American Statistical Association, Alexandria, Virginie.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Schaible, W.L. (1978). Choosing weights for composite estimators for small area statistics. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 741, 746.

Schouten, B., van den Brakel, J.A., Buelens, B., van der Laan, J. et Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, 42(6), 1555-1570.

Sen, A. et Lahiri, P. (2025). Estimation of finite population proportions for small areas – A statistical data integration approach. *Journal of Survey Statistics and Methodology*, 13(3), 309-332.

Suzer-Gurtekin, Z.T., Heeringa, S.G. et Vaillant, R. (2012). Investigating the bias of alternative statistical inference methods in sequential mixed-mode surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 4711-4725.

Toth, D. et Eltinge, J.L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496), 1626-1636.

Valliant, R. et Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1), 105-137.

Vannieuwenhuyze, J., Loosveldt, G. et Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74(5), 1027-1045.

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.*, 40(24), 5237-5250.

Wu, C. (2022). Inférence statistique avec des échantillons d'enquête non probabiliste. *Techniques d'enquête*, 48(2), 307-338. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022002/article/00002-fra.pdf.