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

Techniques d’enquête

Techniques d’enquête
50-1

Date de diffusion : le 25 juin 2024

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Techniques d'enquête

Nº 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Juin 2024

•

Volume 50

•

Numéro 1

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans *The ISI Web of knowledge* (*Web of science*), *The Survey Statistician*, *Statistical Theory and Methods Abstracts* et *SRM Database of Social Research Methodology*, *Erasmus University*. On peut en trouver les références dans *Current Index to Statistics*, et *Journal Contents in Qualitative Methods*. La revue est également citée par *SCOPUS* sur les bases de données *Elsevier Bibliographic Databases*.

POLITIQUE DE RÉDACTION

Techniques d'enquête publie généralement des articles innovateurs de recherche théorique ou appliquée, et parfois des articles de synthèse, qui fournissent des idées nouvelles sur les méthodes statistiques pertinentes pour les bureaux nationaux de statistique et d'autres organismes statistiques. Les sujets d'intérêt sont mentionnés sur le site Web de la revue (www.statcan.gc.ca/techniquesdenquete). Les auteurs peuvent soumettre leurs articles à la section régulière de la revue ou à la section des notes courtes pour les contributions contenant moins de 3 000 mots, incluant les tableaux, les figures et la bibliographie. Bien que le processus d'examen puisse être simplifié pour les notes courtes, tous les articles sont soumis à une évaluation par des pairs. Cependant, les auteurs demeurent responsables du contenu de leur article et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée en version électronique deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le soumettre en français ou en anglais via le **portail de *Techniques d'enquête* sur le site Web de ScholarOne Manuscripts** (https://mc04.manuscriptcentral.com/surveymeth). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca/techniquesdenquete). Pour communiquer avec le rédacteur en chef, veuillez utiliser l'adresse suivante : (statcan.smj-rte.statcan@statcan.gc.ca).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 50, numéro 1, juin 2024

Table des matières

Numéro spécial pour les articles présentés à la 29e conférence Morris Hansen

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 1-3
Statistique Canada, nº 12-001-X au catalogue

Préface au numéro spécial pour les articles présentés lors de la 29e conférence Morris Hansen portant sur l'utilisation d'échantillons non probabilistes

Partha Lahiri[1]

L'article fondateur de Neyman a transformé le domaine des enquêtes par sondage, conduisant à l'adoption généralisée de l'échantillonnage probabiliste et des méthodes associées fondées sur le plan de sondage, particulièrement au sein des bureaux nationaux de statistique. Cependant, la mise en œuvre parfaite des méthodes fondées sur le plan repose sur une base de sondage parfaite de la population finie cible, sur des échantillons bien conçus avec des probabilités de sélection connues et non nulles, sur l'absence de non-réponse, sur l'absence d'erreurs de mesure et sur l'utilisation de poids de sondage pour corriger les probabilités inégales de sélection. Dans ces conditions, la convergence des estimateurs traditionnels fondés sur le plan et de leurs estimateurs de variance est assurée pour de grands échantillons, quelle que soit la validité de tout modèle qui aurait pu être utilisé pour construire ces estimateurs. Pour des échantillons de grande taille, l'approche d'échantillonnage probabiliste est en effet attrayante pour les praticiens d'enquête, car la même procédure d'estimation peut être utilisée pour traiter différents types de variables d'intérêt sans qu'il soit nécessaire de les modéliser séparément.

Les enquêtes par sondage probabiliste se heurtent à des difficultés, notamment la non-couverture, les erreurs de mesure, la baisse des taux de participation et les coûts élevés. À l'inverse, les enquêtes non probabilistes, comme les enquêtes volontaires, gagnent du terrain en raison de leur commodité et de leur rapport coût-efficacité. Dans l'échantillonnage non probabiliste, le mécanisme probabiliste de sélection est inconnu. De plus, les probabilités de sélection sont souvent nulles pour un sous-ensemble des unités de la population finie. Par conséquent, les méthodes traditionnelles fondées sur le plan de sondage ne peuvent pas être utilisées pour construire des estimations ou leurs mesures d'incertitude, et il faut s'appuyer sur des modèles dont les hypothèses ne sont pas toujours vérifiables. Il existe désormais un intérêt croissant pour l'intégration de données non probabilistes aux enquêtes probabilistes, dans le but d'atténuer ces défis et de tirer parti des atouts des deux approches.

En raison de l'importance croissante des enquêtes non probabilistes, le comité de la conférence Morris Hansen a décidé d'organiser la 29e conférence Morris Hansen sur le thème « Travailler avec des échantillons non probabilistes : évaluer et corriger le biais ». La Washington Statistical Society a inauguré la série de conférences Morris Hansen en 1990, soutenue par une subvention de Westat. Par la suite, le National Agricultural Statistics Service (NASS) s'est joint en tant que coparrain de l'événement et, depuis lors, a organisé la série de conférences presque chaque année à Washington, DC.

1. Partha Lahiri, Joint Program in Survey Methodology, University of Maryland, 1218 Lefrak Hall, College Park, Maryland 20742, États-Unis. Courriel : plahiri@umd.edu.

Compte tenu de la pandémie de Covid, la 29e conférence Morris Hansen s'est déroulée sous la forme d'un événement virtuel le 1er mars 2022. Le comité a invité Jean-François Beaumont, Courtney Kennedy et Yan Li, trois experts estimés dans le domaine des enquêtes non probabilistes, à donner des présentations basées sur leurs recherches récentes dans ce domaine. Ce numéro spécial contient des versions révisées des trois articles présentés lors de la 29e conférence Morris Hansen, ainsi que des discussions et des réponses des auteurs.

Le premier article rédigé par Kennedy, Mercer et Lau étudie les problèmes des erreurs de mesure associées aux enquêtes volontaires non probabilistes, fréquemment utilisées pour produire des estimations pour des domaines rares en raison de considérations de coût. Grâce à une étude comparative approfondie, les auteurs identifient des sous-groupes de la population caractérisés par un biais important dans les enquêtes volontaires, attribuant une partie de ce biais à de fausses réponses. Leurs résultats soulignent l'importance d'examiner les erreurs résultant de fausses réponses dans les enquêtes non probabilistes, soulignant la nécessité de s'attaquer non seulement au biais de sélection, mais également à la question des réponses erronées.

Le deuxième article rédigé par Li examine l'hypothèse d'échangeabilité conditionnelle, qui sert d'hypothèse centrale dans les méthodes d'ajustement fondées sur le score de propension. Plus précisément, Li explore la validité de l'hypothèse d'échangeabilité sous divers scores d'équilibrage et conçoit un score d'équilibrage adaptatif visant à obtenir des estimations sans biais de moyennes de la population finie.

Le troisième article rédigé par Beaumont, Bosa, Brennan, Charlebois et Chu représente une avancée significative dans le domaine des méthodes de pondération par l'inverse de la probabilité pour les échantillons non probabilistes visant à atténuer le biais de sélection. Leurs travaux de recherche englobent des techniques d'intégration de données incluant des méthodes paramétriques et des arbres de classification et de régression (CART), avec un accent particulier sur la prise en compte du plan de sondage probabiliste. Il convient de noter l'importance accordée à la sélection des variables dans le contexte des méthodologies proposées.

J'exprime ma sincère gratitude à Jean-François Beaumont, rédacteur en chef de *Techniques d'enquête*, pour avoir gracieusement accepté de consacrer un numéro spécial de *Techniques d'enquête* et pour m'avoir proposé d'en être le rédacteur invité. J'exprime également ma gratitude aux conférenciers Morris Hansen – Jean-François Beaumont, Courtney Kennedy et Yan Li – pour avoir accepté mon invitation à contribuer à ce numéro spécial sur la base de leurs présentations.

De plus, je remercie les auteurs des trois articles – Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois, Kenneth Chu, Yan Li, Courtney Kennedy, Andrew Mercer, Arnold Lau – ainsi que les commentateurs des articles – Vladislav Beresovsky, J. Michael Brick, Michael R. Elliott, Julie Gershunskaya, Jae Kwang Kim, Yonghyun Kwon, Takumi Saegusa, Aditi Sen et Changbao Wu – pour leurs commentaires perspicaces sur les articles. Leurs contributions ont stimulé des discussions intéressantes et enrichi la recherche présentée.

Je crois que les articles, discussions et réponses serviront de référence inestimable pour les recherches futures dans ce domaine dynamique et stimulant.

Partha Lahiri

Rédacteur invité

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 5-26
Statistique Canada, nº 12-001-X au catalogue

Étude de l'hypothèse selon laquelle les répondants aux enquêtes non probabilistes en ligne menées à des fins commerciales répondent en toute bonne foi

Courtney Kennedy, Andrew Mercer et Arnold Lau[1]

Résumé

Les méthodes statistiques élaborées pour les échantillons non probabilistes sont généralement axées sur la sélection non aléatoire comme principale raison pour laquelle les répondants à l'enquête peuvent différer systématiquement de la population cible. Selon une théorie bien établie, dans ces cas, si elle est conditionnée sur des variables auxiliaires nécessaires, la sélection peut devenir ignorable et les estimations d'enquête sont alors sans biais. Toutefois, cette logique repose sur l'hypothèse selon laquelle l'erreur de mesure est inexistante ou faible. Dans la présente étude, nous testons cette hypothèse de deux façons. Premièrement, nous utilisons une vaste étude d'étalonnage qui permet de déterminer les sous-groupes pour lesquels les erreurs dans les échantillons d'enquêtes non probabilistes menées en ligne à des fins commerciales sont particulièrement grandes d'une manière improbable en raison des effets de sélection. Nous présentons ensuite une étude de suivi qui porte sur une des causes des grandes erreurs : les fausses réponses (c'est-à-dire les réponses d'enquête qui sont frauduleuses, malveillantes ou non sincères d'une autre manière). Nous constatons que les fausses réponses, en particulier chez les répondants qui déclarent être jeunes ou d'origine hispanique, constituent un problème important et répandu dans les échantillons d'enquêtes non probabilistes menées en ligne à des fins commerciales, du moins aux États-Unis. La présente étude met en évidence la nécessité pour les statisticiens utilisant des échantillons non probabilistes établis à des fins commerciales de traiter les fausses réponses et les questions de représentativité, et pas uniquement ces dernières.

Mots-clés : Échantillons non probabilistes; enquêtes en ligne; erreur de mesure; étalonnage.

1. Introduction

Les statisticiens d'enquêtes savent depuis longtemps que les échantillons bruts provenant de panels non probabilistes en ligne établis à des fins commerciales ou de places de marché sont susceptibles d'être non représentatifs (par exemple Rivers, 2007; Dever, Rafferty et Valliant, 2008). La tâche du statisticien consiste alors à combiner les réponses des répondants avec des renseignements auxiliaires afin d'éliminer les différences entre l'échantillon et la population cible. Plusieurs méthodes ont été élaborées à cette fin (voir Elliott et Valliant, 2017, pour consulter une étude à ce sujet). Ces méthodes fonctionnent, du moins en théorie, quand les réponses des répondants dans les enquêtes non probabilistes sont authentiques et raisonnablement exactes. Cependant, de plus en plus de données probantes montrent que des proportions importantes de répondants aux enquêtes par échantillonnage non probabiliste fournissent de fausses données, y compris sur les variables que les statisticiens utilisent aux fins d'ajustement.

1.1 Se concentrer sur les enquêtes non probabilistes en ligne menées à des fins commerciales

Les échantillons non probabilistes peuvent prendre de nombreuses formes, comme un échantillon de personnes recrutées à partir d'une plateforme de média social ou un échantillon en boule de neige d'une

1. Courtney Kennedy, Andrew Mercer et Arnold Lau, Pew Research Center, 1615, L St., NW, Washington D.C., suite 800, 20036, États-Unis. Courriel : CKennedy@PewResearch.org.

population rare à risque. Toutefois, les sondages d'opinion publique américains sont dominés par une forme générale d'échantillonnage non probabiliste. Plus de 80 % des sondages d'opinion publique aux États-Unis sont actuellement réalisés au moyen de panels non probabilistes en ligne établis à des fins commerciales (Kennedy et coll., 2021). Les sondages d'opinion publique reposant sur des échantillons non probabilistes en ligne établis à des fins commerciales ne comportent presque jamais d'échantillon probabiliste complémentaire servant à effectuer un calage; il s'agit d'une caractéristique qui les distingue des études combinant statistiquement des échantillons probabilistes et non probabilistes (par exemple Elliott et Haviland, 2007).

Pour éviter la lourdeur du terme « panel non probabiliste en ligne établi à des fins commerciales », nous utiliserons le raccourci « panel volontaire ». Il faut rester prudent et ne pas confondre les données produites à des fins commerciales avec d'autres sources non probabilistes qualitativement différentes, comme celles de Beaumont (2022) et de Li (2022). Les échantillons d'enquêtes volontaires menées à des fins commerciales présentent un ensemble de problèmes uniques, qui sont au cœur de la présente étude.

À la section 2, nous donnons un aperçu de la littérature sur l'étalonnage des enquêtes non probabilistes et de celle portant sur les faux répondants. À la section 3, nous présentons une nouvelle étude d'étalonnage comparant les niveaux d'erreur moyens dans six panels d'enquête en ligne différents. Nous constatons que les erreurs sont particulièrement grandes pour certains sous-groupes et que ces mêmes sous-groupes sont enclins à déclarer des caractéristiques très inhabituelles. À la section 4, nous présentons une étude de suivi pour déterminer si les déclarations concernant les caractéristiques inhabituelles sont crédibles ou si elles prouvent la présence de fausses réponses. À la section 5, nous examinons certaines limites de ces collectes de données, et à la section 6, nous formulons des conclusions sur les répercussions de l'étude.

2. Contexte

Deux axes de recherche méthodologique ont été élaborés en matière de qualité des données dans les panels volontaires. Le premier vise à quantifier la taille moyenne des erreurs et à les comparer aux niveaux des échantillons probabilistes. Le deuxième axe de recherche porte sur l'erreur de mesure dans les réponses individuelles aux enquêtes causée par des répondants frauduleux, malveillants ou non sincères faisant partie de panels volontaires. Notre objectif dans le présent article est d'établir des liens entre les études sur ces deux axes au moyen de nouvelles données afin de mettre en évidence la façon dont ils se complètent, selon nous.

2.1 La littérature sur l'étalonnage

Bon nombre des études axées sur la représentativité visaient à examiner l'exactitude relative des estimations d'enquêtes probabilistes et non probabilistes par rapport aux repères démographiques disponibles. Par exemple, MacInnis et coll. (2018) ont constaté que les estimations pondérées d'un échantillon probabiliste en ligne comportaient des racines des erreurs quadratiques moyennes beaucoup plus petites que les estimations provenant de six panels volontaires différents. Dans l'ensemble, le nombre d'études dans lesquelles on constate une plus grande exactitude dans les estimations d'échantillons probabilistes (Malhotra

et Krosnick, 2007; Chang et Krosnick, 2009; Yeager et coll., 2011; Szolnoki et Hoffmann, 2013; Erens et coll., 2014; Sturgis et coll., 2016; Dutwin et Buskirk, 2017; Pennay et coll., 2018) est environ cinq fois supérieur au nombre d'études dans lesquelles on constate une exactitude similaire dans les estimations d'échantillons non probabilistes (Vavreck et Rivers, 2008; Ansolabehere et Schaffner, 2014).

D'autres études à cet égard ont porté sur l'efficacité de différentes méthodes statistiques pour réduire le biais de sélection dans les estimations tirées d'échantillons volontaires en ligne. Ces études indiquent de façon constante que même après l'emploi de méthodes statistiques plus sophistiquées, comme l'apprentissage automatique ou les méthodes doublement robustes, il reste souvent des erreurs de grande taille dans les estimations d'enquêtes non probabilistes (Dutwin et Buskirk, 2017; Mercer, Lau et Kennedy, 2018). Par exemple, Dutwin et Buskirk (2017) ont observé que « des techniques avancées, comme la pondération par la propension et l'appariement d'échantillons, n'ont pas amélioré ces mesures (non probabilistes) et, dans certains cas, ont empiré les choses » [*traduction*].

La littérature soulève la question suivante : pourquoi, même en cas de modélisation approfondie, les estimations des panels volontaires contiennent-elles souvent de grandes erreurs ? Cela s'explique-t-il par la spécification erronée des modèles ou l'absence de covariables clés, ou serait-ce parce que les données des panels volontaires sont imparfaites d'une manière qui échappe à la correction par modélisation statistique ? Outre qu'elle ne répond pas à ces questions, la littérature sur l'étalonnage comporte une autre limite. La plupart des études d'étalonnage ne tiennent compte que des estimations pour la population totale (par exemple tous les adultes aux États-Unis). Elles n'envisagent pas la possibilité que l'exactitude des estimations tirées d'enquêtes volontaires puisse varier entre les principaux sous-groupes (par exemple en fonction de l'âge, de la race ou de l'origine ethnique). Notre étude d'étalonnage (présentée à la section 3) vise à combler cette lacune en traitant de la variation des sous-groupes, puis en s'appuyant sur ces résultats pour mieux comprendre un deuxième corpus d'articles portant sur les enquêtes non probabilistes.

2.2 La littérature sur les faux répondants

Un autre axe de recherche a porté sur les répondants frauduleux, malveillants ou non sincères figurant dans des panels volontaires établis à des fins commerciales. L'ampleur du problème est alarmante : selon l'Insights Association (2022), les chercheurs doivent s'attendre à retirer de 15 % à 25 % des questionnaires remplis tirés des panels volontaires en raison de la mauvaise qualité des données. Geraci (2022) a affirmé que ce taux est encore plus élevé et qu'« il y a à peine 10 ans, les chercheurs devaient retirer de 5 % à 10 % de toutes les interviews des échantillons en ligne en raison de leur mauvaise qualité. Cette proportion se situe maintenant entre 35 % et 50 % » [*traduction*]. Cet axe de recherche ne vise pas à déterminer si les répondants aux enquêtes volontaires sont représentatifs d'une population plus grande, mais plutôt si leurs réponses sont crédibles.

Il est de plus en plus évident que les faux répondants ne sont pas simplement une nuisance (par exemple l'ajout de bruit aux estimations, la nécessité d'effectuer des interviews de remplacement). Ils peuvent aussi donner lieu à des estimations très biaisées et à de fausses conclusions. À titre d'exemple, Litman et coll.

(2021) ont montré que les rapports des Centers for Disease Control and Prevention concernant les taux élevés d'Américains qui ingéraient de l'eau de Javel pour se protéger contre la COVID-19 étaient attribuables à de faux répondants dans un échantillon d'enquête volontaire. Lopez et Hillygus (2018) ont conclu que les faux répondants avaient augmenté de façon erronée les estimations concernant la croyance du public dans les complots par un facteur de deux. Plus récemment, Westwood et coll. (2022) ont constaté que les faux répondants avaient gonflé de façon erronée les estimations concernant le soutien à la violence politique aux États-Unis de plus du double.

Si l'inquiétude sur la qualité des données dans les échantillons d'enquêtes volontaires n'est pas un phénomène nouveau (par exemple Downes-Le Guin, 2005; Baker et coll., 2010), la recherche sur l'étendue et l'effet des réponses non sincères s'est accélérée depuis quelques années pour plusieurs raisons, y compris les préoccupations relatives aux robots d'enquête (Baxter, 2016; Shanahan, 2018; McDowell, 2019; Puleston, 2019; Geraci, 2022), aux travailleurs étrangers qui dissimulent leur identité pour être admissibles aux enquêtes aux États-Unis (Kennedy et coll., 2018; Moss, 2018; Ahler, Roush et Sood, 2019), et aux personnes qui répondent plusieurs fois à une enquête à partir de différents appareils pour échapper aux vérifications de sécurité des panels (Ahler et coll., 2019; Kennedy et coll., 2021). Bien que les entreprises qui gèrent des panels volontaires établis à des fins commerciales soient conscientes de bon nombre de ces problèmes et tentent de les régler, les répondants frauduleux continuent de représenter une grande part des cas de fausses réponses dans les échantillons d'enquêtes volontaires en ligne.

2.3 Établir des liens entre les études sur les échantillons non probabilistes établis à des fins commerciales

Les méthodes statistiques élaborées pour les données d'enquête non probabiliste (par exemple Rivers, 2007; DiSogra et coll., 2011; Valliant et Dever, 2011; Valliant, 2020) supposent que leur tâche est de tirer parti des variables auxiliaires les plus efficaces. En d'autres termes, le problème à résoudre consisterait à examiner les façons pertinentes dont les répondants aux enquêtes volontaires diffèrent de la population cible. Selon une théorie bien établie, on peut éliminer le biais de sélection en conditionnant l'échantillon sur un ensemble adéquat de variables auxiliaires (Elliott et Valliant, 2017; Mercer et coll., 2017; Kohler, Kreuter et Stuart, 2019). L'hypothèse implicite de ces articles est que le fait de ne pas éliminer l'erreur dans les estimations d'enquête volontaire signifie simplement que nous n'avons pas trouvé les variables auxiliaires adéquates ou qu'elles ne sont pas disponibles. D'après cette logique, les progrès découlent de la découverte de nouvelles sources de données auxiliaires ou de l'élaboration de méthodes statistiques qui peuvent mieux tirer parti des données auxiliaires disponibles. Mais cette logique ne s'applique que si l'erreur de mesure est faible ou inexistante et, en général, on a toujours pensé que même si le comportement de satisfaction (peut-être la source la plus souvent étudiée d'erreur de mesure) peut être légèrement pire dans les enquêtes volontaires, il n'est pas un facteur important de l'erreur.

Le deuxième axe de recherche examiné soulève la possibilité que les faux répondants introduisent en fait une erreur de mesure beaucoup plus importante que ce que l'on pensait auparavant. Quand de 15 % à 50 % des données recueillies doivent être rejetées en raison de leur mauvaise qualité, mettre l'accent sur les

variables auxiliaires risque de nous empêcher d'avoir une vue d'ensemble. Il est peu probable que les quotas d'échantillonnage et la pondération soient efficaces si les variables utilisées contiennent de grandes erreurs introduites intentionnellement par des répondants de mauvaise foi.

Pour clarifier notre propos, nous n'affirmons pas que les faux répondants sont la seule source d'erreur dans les panels volontaires. Au lieu de cela, il est beaucoup plus probable qu'à la fois les faux répondants (c'est-à-dire les personnes répondant de façon erronée) et les répondants non représentatifs (c'est-à-dire les personnes répondant sincèrement, mais qui, dans l'ensemble, diffèrent de la population cible) contribuent à l'erreur. Notre étude vise deux objectifs : 1) montrer en quoi la littérature sur les faux répondants aide à expliquer la littérature sur l'étalonnage; 2) souligner la nécessité pour les statisticiens qui utilisent des échantillons tirés d'enquêtes volontaires de traiter la question des faux répondants *et* les problèmes de représentativité, et pas seulement ces derniers.

3. Étude d'étalonnage

Nous avons d'abord effectué un étalonnage pour déterminer la mesure dans laquelle l'exactitude des estimations tirées d'enquêtes en ligne diffère selon les sous-groupes. Les résultats nous ont amenés à formuler une hypothèse sur les fausses réponses. Comme la méthode d'étalonnage n'a pas été conçue pour détecter les fausses réponses, nous avons élaboré une collecte de données de suivi pour examiner directement cette hypothèse. Nous présenterons ces collectes de données l'une après l'autre.

3.1 Plan de sondage de l'étalonnage

L'étude d'étalonnage comportait des échantillons d'adultes américains tirés de chacune des six plateformes en ligne choisies (tableau 3.1). La taille des échantillons de répondants variait de 4 912 à 5 147. Trois des échantillons provenaient de panels volontaires établis à des fins commerciales. Les entreprises menant les enquêtes volontaires, qui utilisent régulièrement des quotas, ont reçu des quotas cibles pour l'âge × le sexe, la race × l'origine ethnique hispanique, le niveau de scolarité. Nous avons calculé les quotas cibles à partir de l'American Community Survey de 2019.

Tableau 3.1
Taille des échantillons et dates de la collecte selon la source.

Source	n	Dates de la collecte
Panel d'EFA 1	5 027	14 au 28 juin 2021
Panel d'EFA 2	5 147	14 au 27 juin 2021
Panel d'EFA 3	4 965	29 juin au 21 juillet 2021
Panel volontaire 1	4 912	15 au 25 juin 2021
Panel volontaire 2	4 931	11 au 27 juin 2021
Panel volontaire 3	4 955	11 au 26 juin 2021

Note : « Panel d'EFA » désigne un panel en ligne qui est recruté au moyen d'un échantillonnage probabiliste fondé sur des adresses. « Panel volontaire » désigne un panel non probabiliste en ligne établi à des fins commerciales ou de places de marché.

Les trois autres sources sont des panels d'enquête probabilistes dont les interviews sont effectuées en ligne, mais dont les participants sont recrutés hors ligne. La plupart, sinon la totalité, des participants de ces trois sources ont été recrutés au moyen d'un échantillonnage fondé sur des adresses (EFA) à partir du Computerized Delivery Sequence File des services postaux des États-Unis. Avant d'adopter le recrutement par EFA, des participants de deux de ces panels ont été recrutés hors ligne au moyen d'un échantillonnage par composition aléatoire de numéros de téléphone.

Un des échantillons des panels d'EFA a servi aux fins de la présente étude méthodologique ainsi qu'aux fins de recherches importantes qui ne font pas partie du projet. Les recherches importantes nécessitaient une taille d'échantillon plus grande. Tous les membres du panel ont été invités à participer à l'étude, et 10 606 d'entre eux ont rempli le questionnaire. Nous ne voulions pas que cette taille d'échantillon plus grande soit un facteur de confusion dans l'analyse (c'est-à-dire en permettant qu'un échantillon soit deux fois plus grand que les cinq autres). Pour régler ce problème, nous avons tiré un échantillon aléatoire stratifié à partir du panel complet en suivant la procédure standard du panel afin d'obtenir une taille d'échantillon cible de 5 000 questionnaires remplis. Seuls les répondants sélectionnés dans ce sous-échantillon ont été pris en compte dans la présente étude. Ce sont les cas qui auraient été obtenus si seules les personnes dans ce sous-échantillon avaient été invitées à participer à l'enquête. Le processus de sous-échantillonnage a été effectué indépendamment de toute analyse des données. Comme nous l'indiquons ci-dessous, les trois panels d'EFA se sont comportés de façon similaire, ce qui nous a donné l'assurance que le sous-échantillonnage du panel plus grand n'a pas nui à l'étude en produisant des résultats sensiblement différents.

Les panels d'EFA 1, 2 et 3 présentaient des taux de réponse propres à l'étude de 61 %, de 90 % et de 71 %, respectivement. Le taux de réponse cumulatif aux enquêtes (qui tient compte de la non-réponse au recrutement et à l'enquête actuelle, et de l'attrition des panels) était de 5 % pour le panel d'EFA 1, de 3 % pour le panel d'EFA 2 et de 7 % pour le panel d'EFA 3. Des taux de réponse comparables ne peuvent pas être calculés pour les échantillons des panels volontaires. Les données ont été recueillies par l'entreprise Ipsos. Un questionnaire commun a été fourni en anglais ou en espagnol pour les six échantillons.

Chaque panel d'enquête a sa propre méthode de pondération d'une enquête nationale réalisée auprès d'adultes américains. Étant donné que le projet porte sur la qualité des données, il était nécessaire d'éviter que des méthodes de pondération différentes soient un facteur de confusion dans les comparaisons entre échantillons. Nous avons donc utilisé une méthode de pondération standard pour les six échantillons. Pour les échantillons d'EFA, la première étape de la méthode de pondération consistait à ajuster les poids de base des panels en fonction des probabilités différentielles de sélection. Comme il n'existe pas de poids de sondage pour les échantillons des panels volontaires, on a attribué à ces cas un poids de départ de 1 et on les a traités comme s'il s'agissait d'échantillons aléatoires simples.

La deuxième étape consistait à caler les poids de départ pour chaque échantillon en fonction d'un ensemble commun de totaux de contrôle de la population. Les cibles de la population étaient l'âge × le sexe, le niveau de scolarité × le sexe, le niveau de scolarité × l'âge, la race ou l'origine ethnique × le niveau de scolarité, la naissance aux États-Unis par rapport à l'extérieur des États-Unis chez les Américains d'origine

hispanique et asiatique, le nombre d'années de vie aux États-Unis, la région de recensement × le fait qu'elle soit métropolitaine ou non, le bénévolat au cours de l'année précédente, le statut d'inscription sur les listes électorales, l'affiliation à un parti politique, la fréquence d'utilisation d'Internet et l'appartenance religieuse. Les six premiers repères provenaient de l'American Community Survey, les trois repères suivants, des suppléments à la Current Population Survey, et les trois derniers, de la National Public Opinion Reference Survey du Pew Research Center.

3.2 Analyse par étalonnage

Au total, 25 variables repères ont été utilisées dans la présente étude. Elles concernent de nombreux sujets différents, notamment le tabagisme, le service militaire, la possession d'un véhicule, la couverture en matière de soins de santé, le revenu, la participation à des programmes sociaux et la composition du ménage. Les repères ont été établis à partir de sources fédérales de grande qualité, fondées sur des enquêtes nationales ou des données administratives. Aucune des 25 variables repères ne faisait partie de la méthode de pondération. La liste complète des repères et des sources est fournie en annexe.

Pour estimer l'erreur d'enquête par rapport aux repères, nous avons calculé la valeur absolue moyenne de la différence entre l'estimation pondérée de l'enquête en ligne et le repère. Dans le cas d'une variable repère catégorique Y pour laquelle les catégories $c = 1, \ldots, k,$ nous supposons que $\overline{Y}_c$ désigne la « vraie » valeur de la population et que $\overline{y}_c$ désigne l'estimation de l'enquête de la proportion de la population appartenant à la catégorie c. Pour une enquête donnée, l'erreur absolue moyenne (EAM) pour Y est

$$\text{EAM} = \frac{\sum_{c=1}^{k} \left| \overline{y}_c - \overline{Y}_c \right|}{k}. \tag{3.1}$$

En d'autres termes, l'EAM de chaque panel traduit un processus en deux étapes. Premièrement, nous avons calculé la moyenne de la différence entre l'enquête et le repère pour toutes les catégories de réponses à la question. Deuxièmement, nous avons calculé la moyenne de ces 25 moyennes. Nous l'avons fait séparément pour chacun des six panels en ligne. Ainsi, tous les repères ont une influence égale dans l'analyse. Il n'y a aucune raison théorique qu'il en soit autrement. Les réponses indiquant un refus ou une incertitude ne font pas partie d'une catégorie d'étalonnage. Elles sont toutefois représentées dans le dénominateur des estimations de l'enquête en ligne parce qu'elles correspondent à la pratique réelle. Il est inhabituel pour les chercheurs travaillant sur les sondages d'opinion publique de rajuster les estimations uniquement en fonction des personnes qui fournissent une réponse autre que « Refus » ou « Pas certain ».

Les écarts moyens des repères (tableau 3.2) révèlent des constatations connues et d'autres nouvelles. Conformément aux études antérieures, l'erreur absolue moyenne estimée est systématiquement plus faible dans les échantillons probabilistes (3,0 points de pourcentage) que dans les échantillons non probabilistes (6,8 points de pourcentage). Dans ces deux groupes, les échantillons ont présenté des résultats à peu près similaires. Les erreurs absolues moyennes variaient de 2,6 à 3,5 dans les échantillons probabilistes fondés sur des adresses, et de 5,9 à 7,3 dans les échantillons non probabilistes des panels volontaires.

Tableau 3.2
Erreur absolue moyenne dans les estimations d'enquête en ligne pour 25 repères.

	Tous les adultes	18 à 29 ans	30 à 64 ans	65 ans et plus	Diplôme d'études secondaires ou niveau inférieur	Études collégiales partielles	Diplôme d'études collégiales	Blanc	Noir	Hispanique
Moyenne des panels d'EFA	3,0 (0,08)	4,0 (0,31)	3,2 (0,11)	3,1 (0,16)	4,0 (0,16)	3,1 (0,17)	2,5 (0,13)	2,7 (0,09)	4,4 (0,35)	4,2 (0,34)
Moyenne des panels volontaires	6,8 (0,09)	11,6 (0,27)	7,4 (0,12)	3,3 (0,13)	7,2 (0,14)	6,4 (0,17)	7,1 (0,18)	5,9 (0,11)	7,6 (0,27)	11,5 (0,29)
Panel d'EFA 1	2,6 (0,10)	2,9 (0,32)	2,9 (0,15)	2,7 (0,18)	3,2 (0,19)	2,8 (0,19)	2,3 (0,18)	2,4 (0,12)	4,1 (0,47)	3,4 (0,32)
Panel d'EFA 2	3,5 (0,11)	5,7 (0,45)	3,5 (0,16)	3,6 (0,22)	4,6 (0,23)	3,6 (0,22)	2,8 (0,15)	3,1 (0,12)	4,9 (0,42)	4,7 (0,42)
Panel d'EFA 3	2,9 (0,16)	3,3 (0,50)	3,3 (0,21)	2,8 (0,27)	4,2 (0,34)	2,9 (0,30)	2,5 (0,17)	2,7 (0,15)	4,2 (0,55)	4,4 (0,63)
Panel volontaire 1	7,1 (0,16)	11,9 (0,45)	8,1 (0,21)	3,4 (0,17)	7,0 (0,26)	7,3 (0,25)	7,9 (0,31)	6,4 (0,17)	8,2 (0,44)	11,7 (0,48)
Panel volontaire 2	7,3 (0,15)	12,8 (0,47)	8,1 (0,22)	3,5 (0,20)	7,3 (0,25)	6,6 (0,28)	8,5 (0,33)	6,6 (0,18)	8,6 (0,41)	11,7 (0,47)
Panel volontaire 3	5,9 (0,15)	10,2 (0,53)	6,1 (0,19)	3,0 (0,19)	7,3 (0,25)	5,3 (0,31)	5,0 (0,25)	4,8 (0,18)	6,1 (0,44)	11,2 (0,48)

Note : Les erreurs-types sont indiquées entre parenthèses. Les estimations concernant les adultes blancs et noirs sont fondées sur les personnes qui ne déclarent pas être d'origine hispanique.

Le fait qu'il y a une variation importante par sous-groupe est nouveau et peut-être sous-estimé dans le domaine. Pour les échantillons des panels volontaires, l'erreur absolue moyenne pour les estimations relatives aux jeunes adultes (18 à 29 ans) est plus de trois fois plus importante que l'erreur pour celles relatives aux adultes de 65 ans et plus (11,6 points de pourcentage par rapport à 3,3 points de pourcentage). Pour les échantillons fondés sur des adresses, en revanche, l'erreur absolue moyenne pour les estimations fondées sur les jeunes adultes et les aînés est nettement plus semblable (4,0 points de pourcentage par rapport à 3,1 points de pourcentage).

Le tableau 3.2 montre une tendance semblable en ce qui concerne les adultes d'origine hispanique. Pour les échantillons des panels volontaires, l'erreur absolue moyenne pour les estimations relatives aux adultes hispaniques est près du double de celle pour les estimations relatives aux adultes blancs (non hispaniques) [11,5 points de pourcentage par rapport à 5,9 points de pourcentage]. Cependant, pour ce qui est des échantillons fondés sur des adresses, l'erreur absolue moyenne pour les estimations fondées sur les adultes hispaniques et blancs est plus comparable (4,2 points de pourcentage et 2,7 points de pourcentage).

Autrement dit, après avoir fortement pondéré chacun des échantillons des panels volontaires, les estimations concernant les jeunes adultes et les adultes hispaniques étaient inexactes de plus de 10 points de pourcentage en moyenne. Ce sont des erreurs importantes et, dans la littérature, on n'explique pas clairement la raison pour laquelle elles sont concentrées dans ces deux groupes. Les estimations des panels volontaires concernant les adultes de 65 ans et plus, par exemple, sont relativement exactes, s'écartant des repères d'environ 3 points de pourcentage seulement. Bien entendu, d'autres variables démographiques pourraient être examinées. Nous avons également effectué l'analyse selon le sexe et le niveau de scolarité, mais la variance entre ces variables était beaucoup plus faible que les différences selon l'âge et l'origine ethnique. C'est pourquoi le niveau de scolarité et le sexe ne sont pas pris en compte dans la suite de la présente étude.

Le reste de l'analyse portera plutôt sur la raison pour laquelle ces erreurs sont si concentrées chez les personnes qui déclarent être jeunes ou d'origine hispanique.

3.3 Repères comportant les erreurs les plus grandes

Pour déterminer les raisons pour lesquelles les erreurs des enquêtes non probabilistes en ligne menées à des fins commerciales sont concentrées dans certains sous-groupes, nous examinons de plus près les variables pour lesquelles les erreurs sont les plus grandes. Le tableau 3.3 présente des estimations pondérées de la proportion d'adultes aux États-Unis recevant quatre prestations gouvernementales différentes : programme d'aide nutritionnelle supplémentaire, sécurité sociale, allocations de chômage et indemnisation des travailleurs. La principale constatation n'est pas simplement que les estimations d'enquête contiennent des erreurs, mais que les erreurs vont toutes dans le même sens. Plus précisément, les enquêtes non probabilistes en ligne menées à des fins commerciales contiennent proportionnellement un trop grand nombre de répondants déclarant recevoir ces prestations. La même tendance est observée pour les échantillons des panels d'EFA, mais l'ampleur des erreurs est radicalement différente.

Tableau 3.3
Estimations de la perception de quatre prestations gouvernementales différentes.

	Programme d'aide nutritionnelle supplémentaire	Sécurité sociale	Allocations de chômage	Indemnisation des travailleurs
Repère	*11,1 %*	*21,8 %*	*9,3 %*	*0,4 %*
Panel d'EFA 1	14,0 %	25,6 %	12,4 %	1,3 %
	(0,57)	(0,55)	(0,52)	(0,20)
Panel d'EFA 2	19,0 %	27,7 %	17,2 %	2,9 %
	(0,71)	(0,60)	(0,68)	(0,35)
Panel d'EFA 3	18,4 %	25,6 %	13,0 %	1,6 %
	(0,83)	(0,63)	(0,72)	(0,31)
Panel volontaire 1	29,9 %	38,7 %	18,8 %	10,0 %
	(0,77)	(0,74)	(0,68)	(0,50)
Panel volontaire 2	30,0 %	37,5 %	21,0 %	11,8 %
	(0,81)	(0,72)	(0,71)	(0,51)
Panel volontaire 3	21,7 %	34,7 %	16,9 %	7,6 %
	(0,67)	(0,73)	(0,65)	(0,47)

Note : Les erreurs-types sont indiquées entre parenthèses.

Par exemple, la perception de l'indemnisation des travailleurs est une caractéristique extrêmement rare chez les adultes aux États-Unis. La fréquence de perception de la population est inférieure à 1 %. Or, selon les échantillons non probabilistes en ligne établis à des fins commerciales, la fréquence de perception est de près de 10 %. De même, selon les échantillons non probabilistes en ligne établis à des fins commerciales, on estime que la fréquence de perception de l'aide nutritionnelle varie de 22 % à 30 %, tandis que le taux réel de la population est seulement de 11 %.

Ces résultats donnent à penser que les répondants aux enquêtes non probabilistes en ligne menées à des fins commerciales sont enclins à répondre « Oui » lorsqu'on leur demande s'ils ont certaines caractéristiques. Pour mieux illustrer ce phénomène, le tableau 3.4 présente la proportion pondérée de répondants

dans chaque échantillon qui ont déclaré avoir reçu au moins trois des quatre prestations gouvernementales mesurées dans l'enquête.

Encore une fois, il est important de rappeler qu'il est extrêmement rare de recevoir trois de ces prestations ou plus (fréquence de perception de 0,1 % dans la population). Les échantillons probabilistes des panels d'EFA montrent ces dynamiques, les estimations de la proportion d'adultes recevant trois de ces avantages ou plus variant de 0 % à 1 %. Toutefois, selon les échantillons non probabilistes en ligne établis à des fins commerciales, la fréquence de perception se situe de 6 % à 11 %.

Tableau 3.4
Pourcentage d'adultes déclarant avoir reçu au moins trois des quatre différentes prestations gouvernementales.

	Tous les adultes	18 à 29 ans	30 à 64 ans	65 ans et plus	Études secondaires ou niveau inférieur	Études collégiales	Diplôme d'études collégiales	Blanc	Noir	Hispanique
Repère	*0,1 %*	*0,1 %*	*0,1 %*	*0,2 %*	*0,2 %*	*0,1 %*	*0,0 %*	*0,1 %*	*0,2 %*	*0,1 %*
Panel d'EFA 1	0,8 %	1,1 %	0,8 %	0,4 %	1,2 %	0,9 %	0,3 %	0,3 %	2,7 %	1,1 %
	(0,16)	(0,51)	(0,20)	(0,17)	(0,33)	(0,33)	(0,11)	(0,09)	(0,96)	(0,50)
Panel d'EFA 2	1,2 %	2,0 %	1,2 %	0,7 %	2,2 %	1,2 %	0,3 %	0,4 %	3,1 %	2,7 %
	(0,24)	(0,78)	(0,28)	(0,39)	(0,55)	(0,38)	(0,10)	(0,15)	(1,08)	(0,86)
Panel d'EFA 3	1,4 %	2,4 %	1,4 %	0,4 %	2,0 %	1,7 %	0,3 %	1,0 %	2,3 %	2,9 %
	(0,30)	(1,08)	(0,39)	(0,18)	(0,68)	(0,63)	(0,21)	(0,32)	(1,17)	(1,33)
Panel volontaire 1	7,8 %	17,8 %	6,9 %	0,8 %	5,1 %	7,5 %	11,1 %	4,3 %	12,4 %	17,2 %
	(0,44)	(1,44)	(0,58)	(0,35)	(0,69)	(0,77)	(0,96)	(0,46)	(1,60)	(1,47)
Panel volontaire 2	9,0 %	18,0 %	8,9 %	0,2 %	6,0 %	7,7 %	13,7 %	6,9 %	10,8 %	16,9 %
	(0,42)	(1,42)	(0,59)	(0,12)	(0,60)	(0,85)	(0,87)	(0,48)	(1,47)	(1,47)
Panel volontaire 3	5,9 %	14,7 %	4,9 %	0,7 %	6,9 %	5,5 %	5,1 %	3,1 %	6,8 %	18,6 %
	(0,41)	(1,60)	(0,45)	(0,25)	(0,81)	(0,77)	(0,62)	(0,37)	(1,36)	(1,78)

Note : Les erreurs-types sont indiquées entre parenthèses. Les quatre prestations gouvernementales sont le programme d'aide nutritionnelle supplémentaire, la sécurité sociale, les allocations de chômage et l'indemnisation des travailleurs. Les estimations concernant les adultes blancs et noirs sont fondées sur les personnes qui ne déclarent pas être d'origine hispanique.

Le tableau 3.4 présente ces estimations pour les sous-groupes ayant les erreurs les plus importantes : les jeunes adultes et les Hispaniques. Pour ces groupes, l'erreur dans les estimations tirées d'enquêtes non probabilistes en ligne menées à des fins commerciales est stupéfiante. Selon les échantillons non probabilistes en ligne établis à des fins commerciales, environ 20 % des jeunes adultes et 20 % des personnes hispaniques reçoivent au moins trois de ces prestations. Ces résultats soulèvent la question centrale suivante : est-il plus probable que les jeunes répondants et les répondants hispaniques échantillonnés de façon non probabiliste présentent réellement cette caractéristique rare ou que ces répondants se présentent sous un faux jour (c'est-à-dire qu'ils fournissent de fausses réponses) ? Comme nous l'avons mentionné plus haut, il est crucial de répondre à cette question, car les techniques statistiques appliquées aux données d'enquêtes non probabilistes en ligne menées à des fins commerciales se fondent sur la première explication, même si cela dépasse l'entendement.

3.4 Supprimer les répondants apparemment faux des estimations

La question qui se pose alors naturellement est de savoir en quoi l'évaluation de l'exactitude change si les répondants apparemment faux sont exclus de l'analyse. À cette fin, nous avons répété l'exercice d'étalonnage de la section 3.2, mais cette fois-ci, nous avons supprimé tous les répondants qui ont déclaré

avoir reçu les quatre prestations gouvernementales mesurées (programme d'aide nutritionnelle supplémentaire, sécurité sociale, allocations de chômage et indemnisation des travailleurs). Nous avons ensuite repondéré chaque échantillon selon la procédure décrite ci-dessus. Cela a légèrement amélioré l'exactitude, mais c'était loin d'être une panacée. Les résultats sont donnés en annexe. L'écart moyen des repères pour les échantillons non probabilistes s'est amélioré de 21 % (de 6,8 points de pourcentage en moyenne à 5,4 points de pourcentage). Les estimations non probabilistes pour les jeunes adultes et les Hispaniques se sont améliorées, mais les erreurs moyennes sont demeurées plus élevées que pour les autres groupes démographiques (8,6 et 8,3 points de pourcentage en moyenne, respectivement, pour les jeunes adultes et les Hispaniques). La suppression des répondants apparemment faux n'a pas comblé l'écart d'exactitude entre les échantillons probabilistes et non probabilistes, puisque l'écart moyen des repères pour les échantillons probabilistes (2,8 points de pourcentage) est demeuré plus petit.

En somme, la suppression des répondants ayant un comportement de réponse extrêmement suspect contribue à l'exactitude au lieu de lui nuire, mais nous ne considérons pas qu'elle constitue une solution robuste. Ceux qui prétendent avoir reçu quatre prestations gouvernementales différentes ne sont qu'un sous-ensemble de tous les faux répondants possibles dans les échantillons non probabilistes. Des recherches antérieures (par exemple Kennedy et coll., 2021) indiquent que d'autres essais de repérage des faux répondants mèneraient à la découverte d'un ensemble de répondants problématiques différent, bien qu'il puisse se chevaucher partiellement avec le premier. Il est peu probable qu'il suffise d'un seul essai ou d'un seul type de réponse pour repérer toutes les fausses réponses.

4. Examiner directement les fausses réponses

Les résultats de l'étude d'étalonnage laissent croire que certains sous-groupes des panels volontaires sont enclins à fournir des données qui ne sont pas crédibles. Toutefois, l'étude d'étalonnage n'a pas été conçue pour distinguer les réponses crédibles des réponses non crédibles. Un essai de détection plus direct des fausses réponses serait nécessaire pour déterminer avec certitude si les tendances de l'analyse par étalonnage découlent de répondants « inhabituels, mais authentiques » aux enquêtes volontaires ou de faux répondants. Encore une fois, nous pensons que cette distinction est importante parce que les méthodes statistiques appliquées aux données des enquêtes volontaires supposent que leurs répondants peuvent être inhabituels, mais qu'ils sont authentiques.

4.1 Plan de sondage de l'enquête de suivi

En février 2022, nous avons réalisé une courte enquête (14 questions) auprès de $n = 569$ adultes aux États-Unis au moyen d'un panel volontaire différent des trois panels de l'étude d'étalonnage. Nous l'avons appelée « enquête de suivi », car elle visait à assurer un suivi et à approfondir les conclusions intrigantes de l'étude d'étalonnage. Si les tendances de l'étude d'étalonnage étaient reproduites dans ce quatrième panel volontaire distinct, cela constituerait une preuve solide d'un problème systémique dans le domaine des panels volontaires aux États-Unis. Aucun échantillon probabiliste n'a été pris en compte dans le suivi, car

ni l'étude d'étalonnage ni d'autres études (par exemple Kennedy et coll., 2021) n'ont permis de trouver des niveaux importants de faux répondants dans les échantillons probabilistes.

L'objectif de l'étude de suivi était de déterminer si les répondants prétendument hispaniques ou jeunes aux enquêtes volontaires sont enclins à répondre « Oui » quelle que soit la question. À cette fin, nous avons sélectionné des questions pour lesquelles une réponse « Oui » n'est pas crédible. La question « Êtes-vous titulaire d'un permis de pilotage de sous-marin SSGN ? » était posée et les réponses possibles étaient « Oui » ou « Non ». Un sous-marin de classe SSGN est un sous-marin nucléaire américain armé de missiles de croisière; la marine américaine en emploie seulement quatre. Si l'on compte les 425 000 membres en service actif et réservistes environ, l'ensemble de la marine américaine représente moins de 0,2 % de la population adulte des États-Unis, et ainsi la proportion d'adultes américains qualifiés pour piloter ce type de bâtiment est d'environ 0 %. Une autre question de l'enquête de suivi a été formulée de sorte à obtenir une batterie de réponses : « Parmi les activités suivantes, laquelle avez-vous faite au cours de la dernière semaine ? Cochez toutes les réponses qui s'appliquent ». La liste d'activités comprenait deux activités courantes (regarder la télévision et lire un livre) et quatre activités extraordinairement rares (acheter un jet privé, escalader un sommet de la chaîne de montagnes du Karakoram, apprendre à cuisiner des halušky et jouer au jaï alaï).

Le but de l'étude de suivi était de déterminer le nombre de répondants à l'enquête volontaire qui choisiraient les réponses non crédibles et si ce comportement était concentré chez les répondants qui déclarent être des personnes hispaniques ou de jeunes adultes, comme c'était le cas dans l'enquête d'étalonnage. Comme faire des inférences au sujet de la population des États-Unis n'était pas l'objectif de la recherche, l'analyse n'a pas été pondérée.

4.2 Résultats de l'enquête de suivi

L'enquête de suivi a permis de corroborer l'hypothèse *a posteriori* de l'enquête d'étalonnage, à savoir que les répondants aux enquêtes volontaires qui ont déclaré être jeunes ou hispaniques étaient enclins à donner de fausses réponses. Ils sont donc enclins à affirmer qu'ils ont certaines caractéristiques quand cela est tout simplement impossible en fonction des chiffres agrégés observés. La première colonne du tableau 4.1 montre que, dans l'ensemble, 5,3 % des répondants à l'enquête de suivi ont affirmé être titulaires d'un permis de pilotage de sous-marin nucléaire SSGN. Comme dans l'étude d'étalonnage, la fréquence de cette fausse allégation était particulièrement élevée chez les Hispaniques (23,7 %) et chez les moins de 30 ans (12,1 %).

La tendance était identique pour les répondants qui ont affirmé avoir effectué au moins une des activités extrêmement rares au cours de la semaine précédente. La proportion de répondants ayant déclaré avoir effectué au moins une des activités extrêmement rares au cours de la semaine précédente (acheter un jet privé, escalader un sommet de la chaîne de montagnes du Karakoram, apprendre à cuisiner des halušky ou jouer au jaï alaï) était considérablement plus élevée chez les 18 à 29 ans que chez les 30 ans et plus (t = 2,99, $p < 0{,}01$). De même, la proportion de répondants ayant déclaré avoir effectué au moins une activité extrêmement rare au cours de la semaine précédente était beaucoup plus élevée chez les Hispaniques que

chez les non-Hispaniques ($t = 5{,}11$; $p < 0{,}01$). Ces résultats montrent clairement ce que l'étalonnage indiquait : les répondants aux enquêtes volontaires menées à des fins commerciales dans ces sous-groupes sont enclins à donner de fausses réponses.

Tableau 4.1
Estimations des caractéristiques démographiques extrêmement rares tirées de l'enquête non probabiliste en ligne menée à des fins commerciales.

	Titulaire d'un permis de pilotage de sous-marin nucléaire	Comportement extrêmement rare
Tous les adultes	5,3 %	8,4 %
	(0,94)	(1,17)
18 à 29 ans	12,1 %*	17,2 %*
	(3,03)	(3,51)
30 ans et plus	3,5 %	6,2 %
	(0,87)	(1,13)
Hispanique	23,7 %*	28,0 %*
	(4,41)	(4,66)
Non-Hispanique	1,5 %	3,6 %
	(0,56)	(0,87)

Note : Les erreurs-types sont indiquées entre parenthèses. Le comportement extrêmement rare était défini comme le fait d'avoir effectué l'une des activités suivantes au cours de la semaine précédente : acheter un jet privé, escalader un sommet de la chaîne de montagnes du Karakoram, apprendre à cuisiner des halušky ou jouer au jaï alaï. L'astérisque (*) indique que la proportion est significativement plus élevée ($p < 0{,}01$) que celle du groupe complémentaire fondé sur un test t bilatéral, si l'on suppose des variances inégales.

4.3 Répercussions pour les statisticiens utilisant des données d'enquête volontaire

Ces résultats soulèvent une question importante. S'il est clair que les répondants hispaniques des enquêtes volontaires donnent une fausse réponse quand ils déclarent qu'ils pilotent un sous-marin nucléaire, pourraient-ils aussi fournir une fausse réponse quand ils déclarent être d'origine hispanique ? En général, il est impossible de valider l'origine ethnique des répondants lorsqu'il est question de données d'enquête volontaire. Cela dit, les données probantes sur les fausses réponses présentées dans l'étude laissent penser que toutes les réponses des répondants qui font des allégations invraisemblables devraient être considérées avec scepticisme. De fait, l'enquête de suivi ne comportait que 14 questions, ce qui élimine les excuses comme « peut-être que certaines personnes se fatiguent à la fin des longues enquêtes ». Ce n'est effectivement pas ce que montrent les données. Nous constatons plutôt que les personnes affirmant être d'origine hispanique déclarent aussi une série de caractéristiques invraisemblables. Plus important encore, cela s'est reproduit dans quatre panels non probabilistes établis à des fins commerciales. L'explication la plus simple, que nous considérons également comme la plus crédible, est que certains répondants aux enquêtes volontaires sont enclins à répondre « Oui », quelle que soit la question, ce qui se vérifie pour les variables d'ajustement comme pour les variables des résultats d'enquête.

Cela signifie pour les statisticiens qui se servent de ces données que certaines des variables qu'ils utilisent pour réduire le biais (en particulier les variables mesurées au moyen d'une question à laquelle on répond par oui ou par non) peuvent contenir de grandes erreurs. De plus, il se peut que ces erreurs soient concentrées dans certains sous-groupes, plutôt que réparties aléatoirement dans l'échantillon de répondants. Par conséquent, les techniques statistiques d'estimation au moyen de données d'enquêtes volontaires menées à

des fins commerciales pourraient ne pas fonctionner aussi bien que l'on croyait. Les études qui ne tiennent pas compte de l'existence de faux répondants dans ces types d'échantillons risquent fortement de surestimer les résultats de diverses méthodes de modélisation.

Il est moins facile de comprendre la raison pour laquelle les fausses réponses sont également courantes chez les jeunes répondants aux enquêtes volontaires. Comme l'âge n'est pas mesuré au moyen d'une question à laquelle on répond par oui ou par non, nous ne devrions pas nous attendre à trouver le même biais de positivité que celui qui semble en jeu dans l'origine ethnique hispanique. Dans l'enquête d'étalonnage, on a demandé aux répondants de choisir leur année de naissance dans un menu déroulant où les années étaient affichées de la plus récente à la plus lointaine. Dans l'étude de suivi, le fournisseur de l'échantillon a fourni une variable d'âge par intervalles. Il se peut que les faux répondants se contentent de choisir les réponses en haut de la liste. Il est également possible que les choix soient stratégiques, les faux répondants choisissant des réponses qui les rendent plus susceptibles d'être admissibles à une enquête ou de recevoir des incitatifs plus élevés. Bien qu'il semble possible que les faux répondants présentent une asymétrie vers des personnes jeunes, il y a peu de raisons de croire que les variables démographiques et d'autres variables auxiliaires sont mesurées avec plus d'exactitude que les variables importantes. En effet, la distinction entre ces deux types de variables n'est importante que pour les statisticiens.

5. Limites

La présente étude porte sur une catégorie d'enquêtes non probabilistes (c'est-à-dire les panels volontaires en ligne établis à des fins commerciales ou de places de marché) dans un pays (les États-Unis). Nous ne nous attendons pas à ce que les types d'erreurs observées dans l'étude soient présents dans des échantillons non probabilistes prélevés dans des circonstances qualitativement différentes. Dans les sources pour les enquêtes non probabilistes menées à des fins commerciales, les faux répondants sont récompensés pour leur mauvais comportement parce qu'ils reçoivent souvent des incitatifs ayant une valeur pécuniaire, mais il peut n'y avoir aucune structure de récompense de ce genre dans, par exemple, un échantillon sur mesure hors ligne d'une population à risque. Nos données ne nous permettent pas non plus de savoir si les résultats de l'étude s'appliquent aux panels non probabilistes établis à des fins commerciales dans d'autres pays.

Une autre limite de la présente étude concerne l'analyse par étalonnage présentée à la section 3. Bien que l'analyse par étalonnage soit utile pour évaluer l'exactitude des estimations d'enquête, elle comporte des limites. Les repères de l'étude sont tirés d'enquêtes financées par le gouvernement qui sont menées à grands frais et en portant une grande attention à leur qualité. Néanmoins, il s'agit d'enquêtes sujettes aux mêmes problèmes que les enquêtes en ligne. Les enquêtes ayant servi de repères ont des taux de réponse élevés, de l'ordre de 60 % ou plus. Ainsi, bien qu'il existe toujours, le risque de biais de non-réponse est généralement considéré comme étant plus faible pour ces enquêtes. Il est également pertinent d'observer que toutes les enquêtes, quel que soit le taux de réponse, sont sujettes à des erreurs de mesure. Les questions posées dans les enquêtes financées par le gouvernement sont soigneusement élaborées et mises à l'essai, mais elles ne sont pas à l'abri de certains des facteurs qui entraînent des problèmes de fiabilité et de validité dans toutes les enquêtes. Le contexte dans lequel une question est posée (par exemple les questions qui la précèdent) influe souvent sur les réponses. De plus, tous les éléments d'enquête peuvent faire l'objet d'un

certain biais de réponse, surtout le biais dû à la désirabilité sociale. En particulier si un intervieweur est présent, les répondants peuvent parfois modifier leurs réponses pour se présenter sous un jour plus favorable (par exemple en exagérant la fréquence de leur vote). Tous ces facteurs peuvent influer sur la comparabilité de mesures apparemment identiques enregistrées dans différentes enquêtes. L'évaluation de la qualité des données est, au mieux, un processus inexact. Il est donc important de garder à l'esprit que l'étalonnage fournit des mesures du biais estimé et dépend de l'ensemble particulier de mesures prises en compte.

6. Analyse

La présente étude est la première à reposer sur l'étalonnage pour cerner les sous-groupes dans lesquels les faux répondants sont concentrés. C'est aussi la première fois qu'il est démontré que les taux élevés de fausses données chez les personnes ayant déclaré être de jeunes adultes ou hispaniques semblent être un phénomène qui s'étend à l'échelle de l'industrie, du moins aux États-Unis. Les quatre panels volontaires établis à des fins commerciales ou de places de marché présentaient tous la même tendance. En revanche, les estimations provenant d'enquêtes volontaires pour les adultes de 65 ans et plus étaient relativement exactes (erreur absolue moyenne de 3,3 points de pourcentage). Cela indiquerait que les méthodes non probabilistes, comme les plans de sondage hybrides, peuvent être d'une efficacité différente selon le sous-groupe d'intérêt.

L'étude soulève également la question de savoir si les adultes hispaniques sont plus enclins que les autres adultes à fournir de fausses réponses. Nous ne pensons pas que cette explication des résultats observés soit crédible. Selon nous, la constatation la plus importante est de comprendre que selon toute vraisemblance, bon nombre de ces répondants « hispaniques » ne sont pas réellement hispaniques, ce qui entraîne de profondes répercussions pour les statisticiens d'enquêtes. Cela signifie qu'il ne faut pas se fier exclusivement à des techniques comme l'appariement d'échantillons, les modèles de propension ou les régressions hiérarchiques pour corriger les erreurs dans les échantillons d'enquêtes non probabilistes menées à des fins commerciales. En effet, toutes ces méthodes supposent que les répondants sont la personne qu'ils affirment être et donc que leurs renseignements démographiques sont mesurés avec une erreur nulle ou faible. La présente étude démontre que lorsque l'on utilise certains types de données non probabilistes (plus précisément ceux provenant des panels en ligne établis à des fins commerciales), cette hypothèse n'est pas confirmée.

Certains chercheurs qui utilisent des données obtenues en ligne à des fins commerciales connaissent le risque posé par les faux répondants et prennent des mesures pour l'atténuer. Toutefois, il faudrait d'autres recherches sur l'efficacité des pratiques actuelles parce que certains éléments prouvent qu'elles sont inadéquates. Kennedy et coll. (2021) ont constaté que 84 % des faux répondants avaient répondu correctement à une question de contrôle posée pour vérifier leur attention (ou une question « piège »), que 87 % avaient réussi la vérification concernant le temps de réponse trop rapide et que 76 % avaient réussi ces deux vérifications courantes de la qualité des données. Des techniques de détection plus sophistiquées ont été proposées (par exemple Jones, House et Gao, 2015), mais elles ne semblent pas avoir été largement adoptées.

Une préoccupation connexe est que les rapports publics sur les résultats des panels non probabilistes établis à des fins commerciales révèlent rarement si la menace que posent les faux répondants a été traitée et, dans ce cas, de quelle façon elle l'a été. À tout le moins, les chercheurs qui donnent des résultats fondés sur ce type de données devraient divulguer les mesures prises pour se prémunir contre les faux répondants et dans quelle mesure elles se sont avérées efficaces. Bien que certains organismes fournissent peut-être déjà ces renseignements, leur divulgation rigoureuse est loin d'être courante. Une plus grande sensibilisation et une plus grande transparence concernant l'existence des faux répondants dans les échantillons non probabilistes établis à des fins commerciales peuvent aider à réduire les cas de constatations erronées (par exemple Litman et coll., 2021; Westwood et coll., 2022) et à promouvoir une plus grande prudence lors de l'interprétation des constatations tirées d'échantillons non probabilistes.

Remerciements

Courtney Kennedy remercie le Morris Hansen Award Committee de l'avoir choisie pour présenter une contribution à la 29e conférence annuelle Morris Hansen. Tous les auteurs remercient Scott Keeter, le rédacteur en chef, et les examinateurs pour leurs propositions constructives.

Annexe

Tableau A.1
Variables d'étalonnage et source.

Variable	Source des repères	Formulation des questions
Maîtrise de l'anglais	American Community Survey de 2019	Parlez-vous une autre langue que l'anglais à la maison ? [Demander si la personne parle une autre langue que l'anglais à la maison.] Dans quelle mesure parlez-vous l'anglais ? Très bien; Bien; Mal; Pas du tout.
Citoyenneté	American Community Survey de 2019	Êtes-vous un(e) citoyen(ne) des États-Unis ?
Parent d'un enfant dans le ménage	National Health Interview Survey de 2020	Êtes-vous le parent ou le tuteur d'un enfant de moins de 18 ans ? [Demander si la personne est le parent ou le tuteur d'un enfant de moins de 18 ans.] Est-ce qu'un ou plusieurs de ces enfants de moins de 18 ans vivent actuellement dans votre ménage ?
État matrimonial	Current Population Survey de 2021, supplément de mars	Lequel de ces énoncés vous décrit le mieux ? Marié(e); Vivant avec un(e) partenaire; Divorcé(e); Séparé(e); Veuf(ve); Jamais marié(e).
Nombre d'adultes dans le ménage	American Community Survey de 2019	Combien de personnes, y compris vous-même, vivent dans votre ménage ? [Demander s'il y a plus d'une personne dans le ménage.] Combien de personnes, y compris vous-même, sont des adultes de 18 ans et plus ?
Nombre d'enfants dans le ménage	American Community Survey de 2019	Combien de personnes, y compris vous-même, vivent dans votre ménage ? [Demander s'il y a plus d'une personne dans le ménage.] Combien de personnes, y compris vous-même, sont des adultes de 18 ans et plus ?
Assurance maladie	National Health Interview Survey de 2020	Êtes-vous actuellement couvert(e) par une forme quelconque d'assurance maladie ou de régime de soins médicaux ?
Compte de retraite	Current Population Survey de 2021, supplément de mars	Au cours de l'année 2020, avez-vous eu un compte de retraite comme un régime 401(k) ou 403(b), un compte de retraite individuel ou un autre compte conçu expressément aux fins d'épargne-retraite ?

Tableau A.1 (suite)
Variables d'étalonnage et source.

Variable	Source des repères	Formulation des questions
Bons alimentaires	Current Population Survey de 2021, supplément de mars	À tout moment au cours de l'année 2020, avez-vous reçu, vous ou un membre de votre ménage, des prestations du programme d'aide nutritionnelle supplémentaire ou du programme de bons alimentaires, ou avez-vous utilisé une carte de prestations pour ces programmes ?
Sécurité sociale	Current Population Survey de 2021, supplément de mars	Au cours de l'année 2020, avez-vous reçu des paiements de sécurité sociale du gouvernement des États-Unis ?
Hypertension artérielle	National Health Interview Survey de 2020	Un médecin ou un autre professionnel de la santé vous a-t-il déjà dit que vous faisiez de l'hypertension artérielle, aussi appelée « haute pression » ?
Allergie alimentaire	National Health and Nutrition Examination Survey de 2009-2010	Avez-vous des allergies alimentaires ?
Antécédents de tabagisme	National Health Interview Survey de 2020	Avez-vous fumé au moins 100 cigarettes au cours de votre vie ? [Demander si la personne a déjà fumé 100 cigarettes.] À l'heure actuelle, fumez-vous des cigarettes ? Tous les jours; Certains jours; Pas du tout.
Antécédents de vapotage	National Health Interview Survey de 2020	Avez-vous déjà utilisé une cigarette électronique ou un autre produit de vapotage électronique, ne serait-ce qu'une fois, au cours de votre vie ? [Demander si la personne a déjà utilisé une cigarette électronique.] À l'heure actuelle, utilisez-vous des cigarettes électroniques ou d'autres produits de vapotage électroniques ? Tous les jours; Certains jours; Pas du tout.
Déménagement au cours de l'année précédente	Current Population Survey de 2021, supplément de mars	Viviez-vous dans cette maison ou cet appartement il y a un an ?
Type de résidence	American Community Survey de 2019	Quelle réponse décrit le mieux le bâtiment où vous vivez actuellement ? (Inclure tous les appartements, les plains-pieds, etc., même s'ils sont vacants.) Une maison mobile; Une maison unifamiliale non attenante; Une maison unifamiliale attenante à une ou plusieurs maisons; Un immeuble de deux appartements ou plus; Un bateau; Un véhicule récréatif, Une fourgonnette; etc.
Propriété immobilière	American Community Survey de 2019	Lequel des énoncés suivants décrit la maison, l'appartement ou la maison mobile où vous vivez ? Vous appartient ou appartient à un membre de votre ménage qui a une hypothèque ou un prêt (y compris les prêts sur la valeur nette de la propriété); Vous appartient ou appartient à un membre de votre ménage, franc et quitte (sans hypothèque ni prêt); Loué; Occupé sans paiement de loyer.
Nombre de véhicules	American Community Survey de 2019	Combien d'automobiles, de fourgonnettes et de camions d'une capacité d'une tonne ou moins sont stationnés à la propriété pour être utilisés par les membres de votre ménage ?
Situation d'emploi la semaine dernière	Current Population Survey de 2021, supplément de mars	La semaine dernière, avez-vous effectué du travail rémunéré ou à but lucratif ? [Demander si la personne n'a pas travaillé ou a refusé de travailler la semaine précédente.] La semaine dernière, aviez-vous un emploi à temps plein ou à temps partiel ? Indiquer tout emploi duquel vous étiez temporairement absent(e).
Travail touché par la COVID-19	Current Population Survey de 2021, supplément de mars	À tout moment au cours des quatre dernières semaines, avez-vous été dans l'impossibilité de travailler parce que votre employeur a fermé ses portes ou a perdu des occasions de faire du profit en raison du coronavirus ?
Avait un emploi l'an dernier	Current Population Survey de 2021, supplément de mars	Avez-vous occupé un emploi ou travaillé pour une entreprise à un moment quelconque en 2020 ?
Affiliation syndicale	Current Population Survey de 2021, supplément de mars	Êtes-vous membre d'un syndicat ou d'une association d'employés semblable à un syndicat ?
Allocations de chômage	Current Population Survey de 2021, supplément de mars	À tout moment au cours de 2020, avez-vous reçu des allocations de chômage d'un État ou du gouvernement fédéral ?
Indemnisation des travailleurs	Current Population Survey de 2021, supplément de mars	En 2020, avez-vous reçu des paiements d'indemnisation des travailleurs ou d'autres paiements en raison d'une blessure ou d'une maladie liée au travail ?
Statut de militaire ou d'ancien combattant	American Community Survey de 2019	Avez-vous déjà été en service actif dans les forces armées, les réserves ou la garde nationale des États-Unis ?

Tableau A.2
Erreur absolue moyenne dans les estimations d'enquête en ligne pour 25 repères, après suppression des cas apparemment faux.

	Tous les adultes	18 à 29 ans	30 à 64 ans	65 ans et plus	Diplôme d'études secondaires ou niveau inférieur	Études collégiales partielles	Diplôme d'études collégiales	Blanc	Noir	Hispanique
Moyenne des panels d'EFA	2,8 (0,08)	3,7 (0,30)	3,1 (0,10)	3,0 (0,17)	3,7 (0,16)	2,9 (0,18)	2,5 (0,13)	2,7 (0,09)	4,1 (0,37)	3,8 (0,32)
Moyenne des panels volontaires	5,4 (0,09)	8,6 (0,25)	6,1 (0,12)	3,3 (0,14)	6,2 (0,14)	5,3 (0,17)	5,0 (0,14)	4,9 (0,10)	6,4 (0,24)	8,3 (0,24)
Panel d'EFA 1	2,5 (0,09)	2,7 (0,33)	2,8 (0,13)	2,7 (0,18)	3,1 (0,19)	2,7 (0,18)	2,2 (0,18)	2,4 (0,12)	3,7 (0,45)	3,2 (0,30)
Panel d'EFA 2	3,3 (0,11)	5,3 (0,42)	3,4 (0,16)	3,5 (0,22)	4,3 (0,24)	3,5 (0,23)	2,8 (0,16)	3,0 (0,13)	4,6 (0,38)	4,3 (0,40)
Panel d'EFA 3	2,6 (0,15)	3,2 (0,43)	3,0 (0,19)	2,8 (0,28)	3,7 (0,31)	2,6 (0,29)	2,5 (0,18)	2,6 (0,15)	3,9 (0,60)	3,8 (0,57)
Panel volontaire 1	5,8 (0,15)	8,5 (0,39)	6,9 (0,21)	3,3 (0,18)	6,2 (0,26)	6,1 (0,23)	5,5 (0,25)	5,6 (0,17)	6,6 (0,38)	8,2 (0,40)
Panel volontaire 2	5,6 (0,15)	9,4 (0,41)	6,4 (0,21)	3,6 (0,19)	6,3 (0,27)	5,3 (0,29)	5,6 (0,23)	5,0 (0,16)	7,4 (0,40)	8,5 (0,44)
Panel volontaire 3	4,8 (0,15)	7,8 (0,44)	5,2 (0,19)	3,0 (0,18)	6,2 (0,23)	4,4 (0,29)	3,9 (0,23)	4,2 (0,16)	5,2 (0,41)	8,2 (0,41)

Note : Les cas apparemment faux sont définis dans la présente étude comme les personnes qui ont déclaré avoir reçu les quatre prestations gouvernementales mesurées.

Bibliographie

Ahler, D.J., Roush, C.E. et Sood, G. (2019). The micro-task market for lemons: Data quality on Amazon's Mechanical Turk. Présenté à l'Annual Meeting of the Midwest Political Science Association, 6 avril 2019.

Ansolabehere, S., et Schaffner, B. (2014). Does survey mode still matter? findings from a 2010 multi-mode comparison. *Political Analysis*, 22(3), 285-303.

Baker, R., Blumberg, S.J., Brick, J.M., Couper, M.P., Courtright, M., Dennis, J.M., Dillman, D., Frankel, M.R., Garland, P., Groves, R.M., Kennedy, C., Krosnick, J., Lavrakas, P.J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R.K. et Zahs, D. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, 74, 711-81.

Baxter, K. (2016). On the internet, nobody knows you're a bot participant: How bots are contaminating online research data and how we can stop them. Medium. Disponible sur https://medium.com/salesforce-ux/on-the-internet-nobody-knows-youre-a-bot-participant-327dd0da5ce7/.

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. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-fra.pdf.

Chang, L., et Krosnick, J.A. (2009). National surveys via RDD telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, 73, 4, 641-678.

Dever, J., Rafferty, A. et Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, 2, 47-62.

DiSogra, C., Cobb, C., Chan, E. et Dennis, J.M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings, Proceedings of the Survey Methods Section*, 4501-15. Alexandrie, Virginie: American Statistical Association.

Downes-Le Guin, T. (2005). Satisficing behavior in online panels. Présentation à la MRA Annual Conference & Symposium, Chicago, Illinois, États-Unis. http://www.sigmavalidation.com/tips/05_06_02_Online_ Panelists.pdf.

Dutwin, D., et Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81, 213-239.

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. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2007002/article/10498-fra.pdf.

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

Erens, B., Burkill, S., Couper, M.P., Conrad, F., Clifton, S., Tanton, C., Phelps, A., Datta, J., Mercer, C.H., Sonnenberg, P., Prah, P., Mitchell, K.R., Wellings, K., Johnson, A.M. et Copas, A.J. (2014). Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: A comparison with a probability sample interview survey. *Journal of Medical Internet Research*, 16(12), e:276.

Geraci, J. (2022). Poll-Arized: Why Americans Don't Trust the Polls and How to Fix Them Before It's Too Late. Houndstooth Press, 153.

Insights Association (2022). Online sample fraud: Causes, costs, and cures. En ligne, 11 février 2022.

Jones, M.S., House, L.A. et Gao, Z. (2015). Respondent screening and revealed preference axioms: Testing quarantining methods for enhanced data quality in web panel surveys. *Public Opinion Quarterly*, 79, 687-709.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. et Jewell, R. (2018). How Venezuela's economic crisis is undermining social science research − About everything. *Washington Post*, 7 novembre 2018. Disponible sur https://www.washingtonpost.com/news/monkey-cage/wp/2018/11/07/how-the-venezuelan-economic-crisis-is-undermining-social-science-research-about-everything-not-just-venezuela/.

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, 1050-1075.

Kohler, U., Kreuter, F. et Stuart, E.A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application*, 6, 149-172.

Li, Y. (2024). Hypothèse de l'échangeabilité dans des méthodes d'ajustement fondées sur le score de propension aux fins d'estimation de la moyenne de population au moyen d'échantillons non probabilistes. *Techniques d'enquête*, 50, 1, 45-65. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00008-fra.pdf.

Litman, L., Rosen, Z., Rosenzweig, C., Weinberger-Litman, S.L., Moss, A.J. et Robinson, J. (2021). Did people really drink bleach to prevent COVID-19? A tale of problematic respondents and a guide for measuring rare events in survey data. MedRxiv, DOI: https://doi.org/10.1101/2020.12.11.20246694.

Lopez, J., et Hillygus, D.S. (2018). Why so serious? survey trolls and misinformation. Présenté à l'Annual Meeting of the Midwest Political Science Association, Chicago.

MacInnis, B., Krosnick, J.A., Ho, A.S. et Cho, M. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, 82, 707-744.

Malhotra, N., et Krosnick, J. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to internet surveys with nonprobability samples. *Political Analysis*, 15(3), 286-323.

McDowell, B. (2019). Minimizing the impact of survey bots. Quirk's Media. Disponible sur https://www.quirks.com/articles/minimizing-the-impact-of-survey-bots.

Mercer, A.W., Kreuter, F., Keeter, S. et Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81, 250-271.

Mercer, A., Lau, A. et Kennedy, C. (2018). For weighting online opt-in samples, what matters most? Disponible sur https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

Moss, A. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it, CloudResearch. Disponible sur https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/.

Pennay, D.W., Neiger, D., Lavrakas, P.J. et Borg, K. (2018). The online panels benchmarking study: A total survey error comparison of findings from probability-based surveys and nonprobability online panel surveys in Australia (2e éd.). The Australian National University. http://csrm.cass.anu.edu.au/research/publications/methods-research-papers.

Puleston, J. (2019). Panel Hacking. Presented at the Annual Conference of the Association for Survey Computing. Disponible sur https://ascconference.org/wp-content/uploads/2019/04/11-Jon-Puleston-Panel-hacking-ASC-2019.pdf.

Rivers, D. (2007). Sampling for web surveys. Document présenté aux 2007 Joint Statistical Meetings, Salt Lake City, Utah, États-Unis.

Shanahan, T. (2018). Are you paying bots to take your online survey? Fors Marsh Group. Disponible sur https://www.forsmarshgroup.com/knowledge/news-blog/posts/2018/march/are-you-paying-bots-to-take-your-online-survey/.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B. et Smith, P. (2016). Report of the inquiry into the 2015 British general election opinion polls. Londres: Market Research Society and British Polling Council.

Szolnoki, G., et Hoffmann, D. (2013). Online, face-to-face and telephone surveys − Comparing different sampling methods in wine consumer research. *Wine Economics and Policy*, 2, 57-66.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263.

Valliant, R., et Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1), 105-37.

Vavreck, L., et Rivers, D. (2008). The 2006 Cooperative Congressional Election Study. *Journal of Elections, Public Opinion & Parties*, 18(4), 355-366.

Westwood, S.J., Grimmer, J., Tyler, M. et Nall, C. (2022). Current research overstates American support for political violence. *Proceedings of the National Academy of Sciences*, 119(12).

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A. et Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75, 709-747.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 27-31
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Étude de l'hypothèse selon laquelle les répondants aux enquêtes non probabilistes en ligne menées à des fins commerciales répondent en toute bonne foi »

J. Michael Brick[1]

Résumé

Devenue populaire pour certains types de projets de recherche par enquête, l'utilisation d'échantillons non probabilistes est rapide et peu coûteuse. Kennedy, Mercer et Lau examinent certains problèmes de qualité des données associés aux échantillons non probabilistes des panels volontaires, un type d'échantillon fréquemment utilisé aux États-Unis. Les auteurs montrent que les estimations obtenues à partir de ces échantillons posent de graves problèmes qui vont au-delà de la représentativité. Afin de bien évaluer tous les types d'enquêtes qui existent, il est important d'adopter le point de vue de l'erreur d'enquête totale.

Mots-clés : Erreur d'enquête totale; adéquation des données à leur utilisation; fabrication.

Kennedy, Mercer et Lau (KML) apportent une contribution utile à notre compréhension de la qualité des estimations obtenues à partir des panels volontaires. Le niveau de clarté de l'article est remarquable, surtout si l'on considère qu'une grande partie de la littérature liée aux panels volontaires porte sur les biais de sélection et est si complexe sur le plan technique que les praticiens peuvent avoir du mal à saisir pleinement les hypothèses clés et les implications connexes qui y sont énoncées.

Les recherches de KML visaient à examiner l'exactitude des estimations provenant des panels volontaires dans le cadre de la production d'estimations par domaine. La capacité de produire des estimations pour des domaines rares est un avantage considérable que les panels volontaires peuvent offrir, puisque ces types d'échantillons sont vraiment moins coûteux. La plupart des études qui ont été menées par le passé n'avaient pas permis d'examiner de près les estimations par domaine. Les constatations de KML mettent en lumière des aspects troublants des panels volontaires, des caractéristiques qui concordent avec ce que l'on retrouve dans la littérature existante, mais qui ne sont peut-être pas bien connues de bon nombre des utilisateurs de tels panels. Essentiellement, la qualité d'un grand nombre d'interviews volontaires est si médiocre que certains chercheurs suggèrent qu'il faudrait écarter de 15 % à 50 % d'entre elles. Je ne peux qu'imaginer ce que M. Deming aurait pensé d'une situation où un fournisseur cacherait l'existence de « défectuosités » de son produit à son client et remédierait à ce problème en s'en remettant à lui pour mener des inspections et éliminer les défauts observés. L'approche classique de M. Deming qui consiste à travailler sur le processus ou le système d'abord pour éviter toute forme de défectuosité n'est pas réalisable dans le cas des panels volontaires, car le processus est une boîte noire à laquelle le client n'a pas accès.

1. J. Michael Brick, J.M. Statistics and Data Science, Westat, 1600 Research Blvd, Rockville, MD 20850, États-Unis. Courriel : mikebrick@westat.com.

La première étude qu'ont menée KML les a poussés à formuler des hypothèses sur les faux répondants qui pourraient être mises à l'essai dans le cadre d'une plus petite étude de suivi. L'étude de suivi montre clairement qu'une grande part des répondants dans les panels volontaires ne sont pas crédibles. Les auteurs fournissent là encore les éléments probants que les utilisateurs de panels volontaires ne devraient pas ignorer.

Avant d'aborder certains renseignements détaillés, mentionnons que KML sous-entendent que l'« hypothèse » selon laquelle les répondants dans les panels volontaires se comportent comme des répondants dans des échantillons probabilistes est seulement une présomption, qui est erronée. Des examens d'enquêtes axés sur des échantillons probabilistes ont été effectués pendant des décennies afin de détecter les nombreuses sources d'erreur issues de l'utilisation du cadre de l'erreur d'enquête totale (EET). Est-il logique d'ignorer ce cadre simplement parce que la source de l'échantillon est différente ? Des critères classiques, tels que la couverture et les taux de réponse dans le cadre d'EET peuvent ne pas s'appliquer directement aux enquêtes volontaires, contrairement à beaucoup d'autres sources d'erreur.

Le plus récent rapport de l'American Association of Public Opinion Research (AAPOR, 2022) sur les mesures de la qualité révèle qu'il est encore bien difficile de cibler des échantillons en ligne de qualité, même si quelques progrès ont été réalisés en ce sens. Les utilisateurs n'ont que peu de motifs pour se sentir à l'aise en ce qui concerne la qualité d'un panel volontaire, quel qu'il soit. KML ne fournissent qu'un survol des mérites relatifs de leurs différents panels volontaires. Bien que la recherche d'un panel volontaire de haute qualité soit une mission que beaucoup se sont donnée de réaliser depuis plus d'une décennie, les données probantes accumulées jusqu'à présent laissent entendre que la cause est peine perdue.

KML ne tentent pas d'effectuer un examen approfondi de toutes les sources d'erreur des enquêtes volontaires, mais montrent plutôt avec clarté qu'il est possible que les hypothèses relatives aux enquêtes traditionnelles ne s'appliquent pas aux panels volontaires. La rapidité et la collecte de données à prix modique – les principaux avantages des panels volontaires – ne devraient pas être les seuls critères à prendre en considération au moment de décider de la façon de mener une enquête. L'exactitude des estimations est importante, ou du moins elle devrait l'être! KML véhiculent deux messages importants : 1) les rendements des panels volontaires, en ce qui concerne la précision des estimations pour les domaines, sont médiocres comparativement aux panels probabilistes de l'échantillonnage fondé sur les adresses; 2) l'une des raisons qui expliquent la qualité médiocre des panels volontaires est la présence de faux répondants. Chaque client potentiel d'un panel volontaire doit bien saisir ces éléments avant de décider si un panel volontaire est une source de données qui lui conviendrait.

Un panel volontaire pourrait être adapté à l'utilisation souhaitée si tout ce qu'il faut en tirer est une idée générale de la taille d'une estimation (« est-ce plus grand qu'une boîte à pain ? »). Les clients doivent toutefois être très conscients que l'augmentation de la taille de l'échantillon d'un panel volontaire n'améliorera pas l'exactitude des estimations, puisque les biais ne diminuent pas avec la taille de l'échantillon. KML montrent de plus que l'augmentation de la taille des échantillons pour produire des

estimations par domaine précises ne s'avère pas une méthode efficace. La présence de faux répondants ajoute des interférences qui faussent les estimations pour les domaines.

Hypothèses

KML démontrent que les biais d'échantillonnage et de sélection ne sont pas les seules différences qui subsistent entre les enquêtes par échantillon probabiliste et les enquêtes par panel volontaire. Commençons par examiner la question des répondants. Les deux axes de recherche actifs sur la qualité des répondants dans les panels volontaires portent sur les préoccupations concernant les répondants professionnels, ou, de façon plus générale, le conditionnement des panels (les répondants des panels volontaires sont-ils conditionnés après avoir répondu à d'autres enquêtes ?), ainsi que la validité des répondants (sont-ils ce qu'ils prétendent être ?). Les études de Hillygus, Jackson et Young (2014) et de Baker et coll. (2014) constituent des exemples de tels travaux. De nombreux utilisateurs agissent toutefois comme si des répondants de panels volontaires étaient semblables à des répondants d'échantillons probabilistes. KML montrent que cette hypothèse est injustifiée.

Pourquoi les répondants de panels volontaires ne se comporteraient-ils pas comme des répondants d'échantillons probabilistes ? La question consisterait peut-être à déterminer plutôt pourquoi nous croyons qu'ils sont similaires ? Les répondants des panels volontaires choisissent de se joindre à un panel ou de répondre aux questions d'une enquête bien précise sans y être invités. Les motivations de répondants issus d'échantillons probabilistes et de répondants issus de panels volontaires risquent d'être très différentes (Keusch, Batinic et Mayerhofer, 2014). Ceux qui prétendent que le choix de répondre à un panel est comme celui de répondre à une demande d'enquête directe au sein d'un échantillon probabiliste formulent une grande hypothèse qui n'est pas fondée. Mon hypothèse non vérifiée est que l'acte d'entrer en contact avec un ménage et de lui demander activement de se joindre à une enquête constitue sans doute le principal avantage qualitatif qu'offrent les panels probabilistes comparativement aux panels volontaires. De même, la littérature sur le conditionnement des panels à partir d'échantillons probabilistes n'est probablement pas très pertinente pour les panels volontaires (ou les panels probabilistes), car elle examine les effets sur les réponses au fil du temps à la même série générale de questions plutôt qu'à chaque demande d'enquête différente.

Dans le tableau 3.3, KML fournissent des preuves tangibles qui permettent d'établir qu'il existe des biais beaucoup plus importants dans des panels volontaires que dans des panels probabilistes. Supposons un modèle d'erreur de mesure simple où la réponse à l'enquête est soumise à l'erreur, mais que les données repères sont exemptes d'erreurs. Ce modèle est

$$y_{s,i} = \mu_i + \varepsilon_{s,i},$$

où $y_{s,i}$ est la réaction « 0-1 » à obtenir l'avantage du gouvernement pour l'enquête s et le répondant i, μ_i est la véritable valeur du répondant i et $\varepsilon_{s,i}$ est l'erreur dans l'enquête s du répondant i. Dans ce cas simple, le biais d'erreur de mesure est

$$\text{biais}_{\text{me}} = (1-\mu)\gamma_{\text{FP}} - \mu\gamma_{\text{FN}},$$

où μ est la moyenne de population finie de μ_i, γ_{FP} est la probabilité de faux positifs et γ_{FN} est la probabilité de faux négatifs. Le tableau 3.3 montre que les biais absolus des panels volontaires des quatre avantages gouvernementaux différents varient de 1,3 à 5,1 fois plus élevés que celui des échantillons de l'échantillonnage fondé sur les adresses, ce qui suppose que la qualité des panels volontaires est médiocre.

Si nous présumions que $\gamma_{\text{FP}} = \gamma_{\text{FN}}$, nous pourrions penser que ces conclusions sont raisonnables. Il existe toutefois de nombreux ouvrages sur ces taux tirés d'échantillons probabilistes qui montrent que γ_{FP} est *négligeable*, comparativement à γ_{FN}. Le tableau 3.3 dans l'étude de KML soulèvent ainsi de sérieuses questions quant à la surestimation des avantages des échantillons probabilistes. Celhay, Meyer et Mittag (2022) déclarent que les estimations du Supplemental Nutrition Assistance Program (SNAP) du Survey of Income and Program Participation (un échantillon probabiliste longitudinal) s'élèvent à $\gamma_{\text{FN}} = 0{,}180$ et à $\gamma_{\text{FP}} = 0{,}013$, ce qui entraîne une sous-estimation de la participation au SNAP. Cette constatation donne à penser que l'hypothèse selon laquelle un panel de nature probabiliste fonctionne comme un échantillon probabiliste ne se tient peut-être pas. Les conditions de l'enquête sont importantes, et nous devrions nous garder d'importer une hypothèse d'un milieu à un autre sans élément probant. Les données à elles seules ne nous permettent pas d'évaluer si la surestimation inattendue des avantages dans l'échantillonnage fondé sur les adresses est attribuable à une cause particulière ou à une certaine agrégation de causes. Il faudrait mener des études plus approfondies sur les panels probabilistes.

Faux répondants

Dans l'étude de suivi qu'ils ont menée, KML confirment leur hypothèse selon laquelle les Hispaniques et les jeunes répondants sont beaucoup plus susceptibles d'être des faux répondants. Le groupe se demande également, à juste titre, si ces personnes interrogées sont même véritablement hispaniques ou jeunes, puisque leurs réponses à ces éléments pourraient également être fabriquées de toute pièce.

Comme plusieurs enquêtes nécessitent le recours à des répondants qui sont jeunes et hispaniques, les auteurs suggèrent que le taux élevé de fabrication dans ces sous-groupes pourrait être lié à la motivation à faire partie de l'échantillon. Les faux répondants pourraient être motivés à déclarer qu'ils sont membres de ces sous-groupes pour obtenir les incitatifs promis ou d'autres récompenses. Cette idée semble plausible étant donné la forte probabilité de réponses bidon à d'autres éléments d'« essai » qui sont fournies par les membres de ces sous-groupes.

KML n'abordent pas en profondeur la question des faux répondants dans des panels volontaires en ce qui concerne les personnes qui ne sont pas membres de ces sous-groupes. Mais n'est-ce pas ce que l'on peut sous-entendre lorsque nous obtenons des résultats selon lesquels 3,5 % des personnes âgées de 35 ans et plus sont autorisées à exploiter un sous-marin nucléaire ? À tout le moins, l'utilisateur doit être conscient de la possibilité qu'une grande part des réponses sur *l'ensemble* soient fausses.

La solution consiste-t-elle à tenter de trouver un moyen d'éliminer ces faux cas ? KML n'approuvent pas cette solution et montrent que même si nous le voulions, il n'y aurait aucun moyen simple de le faire efficacement. Je suis entièrement d'accord. S'ils sont suffisamment motivés, les gens trouveront des moyens de déjouer les outils d'inspection. Si les panels volontaires souhaitent être acceptés comme produisant des résultats de haute qualité, ils ont un travail considérable à faire. KML nous ont rendu service en clarifiant les implications réelles des choix de sources d'échantillons qui sont à notre disposition.

Bibliographie

AAPOR (2022). Data Quality Metrics for Online Samples: Considerations for Study Design and Analysis. Téléchargé le 13 mars 2023 https://aapor.org/wp-content/uploads/2023/02/Task-Force-Report-FINAL.pdf.

Baker, R., Miller, C., Kachhi, D., Lange, K., Wilding-Brown, L. et Tucker, J. (2014). Validating respondents' identity in online samples. *Online Panel Research: Data Quality Perspective, A*, 441-456.

Celhay, P.A., Meyer, B.D. et Mittag, N. (2022). *What Leads to Measurement Errors? Evidence from Reports of Program Participation in Three Surveys*, (No. w29652). National Bureau of Economic Research.

Hillygus, D.S., Jackson, N. et Young, M. (2014). Professional respondents in nonprobability online panels. *Online Panel Research: Data Quality Perspective, A*, 219-237.

Keusch, F., Batinic, B. et Mayerhofer, W. (2014). Motives for joining nonprobability online panels and their association with survey participation behavior. *Online Panel Research: Data Quality Perspective, A*, 171-191.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 33-34
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Étude de l'hypothèse selon laquelle les répondants aux enquêtes non probabilistes en ligne menées à des fins commerciales répondent en toute bonne foi »

Michael R. Elliott[1]

Résumé

Kennedy, Mercer et Lau étudient la question des erreurs de déclaration faites par les répondants dans les échantillons non probabilistes et, ce faisant, mettent au jour un nouvel aspect, à savoir les erreurs volontaires de déclaration en ce qui concerne des caractéristiques démographiques. Ce constat laisse à penser que le bras de fer auquel on assiste entre les chercheurs et les personnes déterminées à nuire à la pratique des sciences sociales se poursuit. Les chercheurs doivent donc tenir compte de ces personnes lorsqu'ils utilisent des enquêtes probabilistes de grande qualité pour réduire les erreurs dans les échantillons non probabilistes.

Mots-clés : Erreurs de déclaration; données démographiques; étalonnage.

Saluons Kennedy, Mercer et Lau (KML) pour leur excellent article qui porte sur un problème sous-estimé concernant les échantillons non probabilistes, à savoir la présence de faux répondants. Certes, d'autres ont déjà abordé la question : Jamieson, Lupia, Amaya, Brady, Bautista, Clinton, Dever, Dutwin, Goroff, Hillygus, Kennedy, Langer, Lapinski, Link, Philpot, Prewitt, Rivers, Vavreck, Wilson et McNutt, 2023 font remarquer que d'autres études ont constaté la présence de tels répondants, citant notamment l'exemple de personnes qui [*Traduction*] ingèrent de l'eau de Javel pour se protéger contre la COVID-19 (Litman, Rosen, Hartman, Rosenzweig, Weinberger-Litman, Moss et Robinson, 2023), croient en des conspirations telles que le PizzaGate (Lopez et Hillygus, 2018), soutiennent la violence politique (Westwood, Grimmer, Tyler et Nall, 2022) ou appuient les opinions favorables à l'égard de Vladimir Poutine (Kennedy, Hatley, Lau, Mercer, Keeter, Ferno et Asare-Marfo, 2021). Toutefois, KML insistent davantage sur la gravité du problème et met au jour un nouvel aspect, à savoir les erreurs volontaires de déclaration en ce qui concerne des données démographiques de base. Il s'agit-là d'un point relativement important puisque ces données quantitatives, qui sont tirées du Recensement des États-Unis ou d'autres sources de données d'enquête gouvernementales et qui sont généralement considérées comme étant raisonnablement précises, sont souvent utilisées pour le calage ainsi que pour l'atténuation du biais de sélection ou du biais de non-réponse.

Nous sommes plusieurs, moi y compris, à préconiser le financement prudent et continu d'un ensemble d'enquêtes de grande qualité dont les données servent de repères sur lesquels caler les données des enquêtes non probabilistes (Wu, 2022). KML laissent entendre que ces enquêtes pourraient également devoir inclure des mesures visant à détecter les faux répondants. Ces mesures s'ajouteraient aux éventuelles étapes à suivre

1. Michael R. Elliott, Department of Biostatistics, University of Michigan, M4124 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109. Courriel : mrelliot@umich.edu.

pour pouvoir identifier directement de tels répondants dans l'enquête non probabiliste même, comme l'inclusion de questions d'identification (Petzel, Johnson et McKillip, 1973; Chandler, Rosenzweig, Moss, Robinson et Litman, 2019). Étant donné que les enquêtes non probabilistes peuvent facilement devenir la cible de personnes qui essaient d'influencer les résultats d'une étude pour toutes sortes de raisons, nous sommes loin d'en avoir fini avec ce bras de fer qui, au contraire, s'intensifie entre les chercheurs et les personnes déterminées à nuire à la pratique des sciences sociales.

Bibliographie

Chandler, J., Rosenzweig, C., Moss, A.J., Robinson, J. et Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51, 2022-2038.

Jamieson, K.H., Lupia, A., Amaya, A., Brady, H.E., Bautista, R., Clinton, J.D., Dever, J.A., Dutwin, D., Goroff, D.L., Hillygus, D.S., Kennedy, C., Langer, G., Lapinski, J.S., Link, M., Philpot, T., Prewitt, K., Rivers, D., Vavreck, L., Wilson, D.C. et McNutt, M.K. (2023). Protecting the integrity of survey research. *PNAS Nexus*, 2, 3, pgad049.

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, 1050-1075.

Litman, L., Rosen, Z., Hartman, R., Rosenzweig, C., Weinberger-Litman, S.L., Moss, A.J. et Robinson, J. (2023). Did people really drink bleach to prevent COVID-19? A guide for protecting survey data against problematic respondents. *Plos One*, 18, e0287837.

Lopez, J., et Hillygus, D.S. (2018). Why so serious?: Survey trolls and misinformation. *Why So Serious*. Disponible au SSRN: https://ssrn.com/abstract=3131087.

Petzel, T.P., Johnson, J.E. et McKillip, J. (1973). Response bias in drug surveys. *Journal of Consulting and Clinical Psychology*, 40, 437-439.

Westwood, S.J., Grimmer, J., Tyler, M. et Nall, C. (2022). Current research overstates American support for political violence. *Proceedings of the National Academy of Sciences*, 119, e2116870119.

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

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 35-39
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Étude de l'hypothèse selon laquelle les répondants aux enquêtes non probabilistes en ligne menées à des fins commerciales répondent en toute bonne foi »

Aditi Sen[1]

Résumé

La présente analyse résume les nouvelles constatations intéressantes de Kennedy, Mercer et Lau (KML) sur les erreurs de mesure dans les enquêtes à participation volontaire. Alors que KML éclairent les lecteurs au sujet des « fausses réponses » et des tendances qui peuvent s'y rattacher, cette analyse propose de combiner ces nouveaux résultats avec d'autres pistes de recherche sur l'échantillonnage non probabiliste, comme l'amélioration de la représentativité.

Mots-clés : Enquête à participation volontaire; erreur de mesure; qualité des données; intégration des données; pondération par l'inverse du score de propension.

Dans leur travail de recherche fondamental, KML se concentrent sur l'aspect important de l'erreur de mesure dans les enquêtes non probabilistes, notamment les enquêtes en ligne menées à des fins commerciales (appelées « enquêtes à participation volontaire »). Des méthodes avancées d'estimation des caractéristiques de la population à partir d'enquêtes non probabilistes sont fréquemment mises au point selon l'hypothèse d'exactitude des réponses aux enquêtes. En présence de réponses inexactes, lorsque cette hypothèse ne se vérifie pas, ces méthodes peuvent se révéler inadéquates. Ainsi, lorsque KML remettent en question l'exactitude des réponses individuelles aux enquêtes à participation volontaire, notre attention est attirée sur cette question sérieuse qui appelle des recherches plus approfondies sur le problème.

Le type d'enquête en ligne le plus populaire repose sur ce qu'on appelle un « panel volontaire ». Contrairement aux enquêtes probabilistes, où un échantillon représentatif de la population est tiré d'une base de sondage, les panels volontaires ne sont pas constitués à l'aide d'un plan d'échantillonnage probabiliste. Dans ce type de panel, les volontaires sont recrutés au moyen de diverses méthodes pratiques, mais non probabilistes, comme l'échantillonnage par quotas ou l'échantillonnage en boule de neige, et les personnes s'y joignent souvent en ayant droit à une mesure incitative quelconque. Les enjeux tels que l'augmentation du coût et la diminution des taux de réponse aux enquêtes probabilistes font l'objet de nombreuses discussions. À l'ère des mégadonnées et des capacités de programmation informatique rapides et efficaces, les enquêtes à participation volontaire suscitent beaucoup d'intérêt. Ces enquêtes coûtent moins cher et donnent la possibilité de recruter des volontaires pour les panels et de recevoir des réponses rapidement. Toutefois, rien ne garantit que ces échantillons représentent convenablement la population cible. De plus, comme KML l'ont souligné, il y a un risque que les réponses obtenues ne soient pas

1. Aditi Sen, University of Maryland College Park, États-Unis. Courriel : asen123@umd.edu.

authentiques. Il est possible que certains répondants soient motivés par la récompense ou qu'ils fournissent intentionnellement de mauvaises réponses. KML mettent en évidence les problèmes que posent les enquêtes à participation volontaire et trouvent des voies intéressantes pour de futures recherches dans le domaine des enquêtes non probabilistes.

Groves et Lyberg (2010) discutent du cadre de l'erreur d'enquête totale en s'inspirant des typologies de l'erreur décrites dans l'article de Deming publié en 1944 dans l'*American Sociological Review*, qui dresse une comparaison claire entre les composantes du biais de l'erreur et celles de la variance de l'erreur. Ils font également remarquer que les théories et les méthodes d'échantillonnage antérieures s'appliquent le mieux lorsque les erreurs non dues à l'échantillonnage sont faibles. Les erreurs non dues à l'échantillonnage comprennent, entre autres, les erreurs de non-réponse et les erreurs de mesure qui ne sont pas liées à la sélection de l'échantillon. Hansen, Hurwitz et Bershad (1961) mentionnent dans leur article que les opérations de collecte et de traitement des enquêtes-échantillons constituent le processus de mesure et sont une source d'erreurs de mesure. KML quantifient l'erreur de mesure dans les réponses individuelles en comparant la taille moyenne des erreurs dans les enquêtes à participation volontaire avec les enquêtes probabilistes et relie ainsi deux domaines de recherche méthodologique sur la qualité des données dans les panels volontaires.

L'article de KML porte essentiellement sur le domaine peu abordé des réponses aberrantes dans les enquêtes à participation volontaire, qui sont désignées sous l'appellation de « fausses réponses ». Le recours à l'étalonnage aide à formuler l'hypothèse selon laquelle ces réponses aux questions d'enquête sont aberrantes. Cette hypothèse est mise à l'épreuve par la suite à l'aide d'une « enquête de suivi », qui consiste à poser des questions fondées sur des événements rares, auxquelles une réponse affirmative est très improbable. Les auteurs travaillent avec six enquêtes en tout : trois d'entre elles sont des enquêtes commerciales à participation volontaire où les fournisseurs ont utilisé l'échantillonnage par quotas (l'American Community Survey [ACS] de 2019 étant utilisée comme cible) pour sélectionner les échantillons. Les trois autres sont des enquêtes probabilistes pour lesquelles des participants de panels sont recrutés au moyen d'un échantillonnage fondé sur les adresses. Dans l'étude par étalonnage, les réponses aux 25 questions communes à toutes ces enquêtes (traitées comme des estimations) sont comparées avec celles d'enquêtes gouvernementales (traitées comme de vraies valeurs) comme l'ACS, la Current Population Survey (CPS) et la National Health Interview Survey (NHIS) pour ce qui est de l'erreur absolue moyenne (EAM). Une réflexion allant dans le sens de l'enquête de suivi est la suivante : supposons que le questionnaire de l'enquête principale puisse être planifié de manière à inclure un ensemble de questions « de détection » spéciales (en plus des questions principales). Les réponses à ces questions serviraient à mesurer les « probabilités de fausse réponse » compte tenu des covariables. Ces probabilités de fausse réponse pourraient être utilisées pour sous-pondérer les réponses individuelles. Par exemple, on fournit les valeurs extrêmes de l'ajustement de poids, qui se situe entre 0 et 1, en utilisant les conditions suivantes : si la probabilité qu'une réponse soit fausse est 1, alors l'ajustement du poids est de 0, mais si la réponse est réputée fiable, alors l'ajustement du poids est de 1. On pourrait mettre au point une méthode pour intégrer

ces probabilités ainsi que la pondération habituelle de la probabilité de participation aux réponses, afin d'améliorer simultanément la représentativité et de tenir compte de la probabilité d'une fausse réponse.

Fait intéressant, les travaux de recherche de KML nous révèlent la présence de tendances dans les fausses réponses. Comme l'indiquent les auteurs, leur étude est la première à utiliser l'étalonnage pour déterminer les sous-groupes associés à une forte probabilité de fournir de fausses réponses. Lorsqu'ils sont sous-groupés par variables démographiques des répondants, un à la fois, principalement l'âge et l'origine ethnique, on observe, à partir des trois enquêtes à participation volontaire, que les répondants jeunes (de 18 à 29 ans) et ceux d'origine hispanique sont plus susceptibles de suivre de telles tendances. Bien sûr, ces regroupements sont subjectifs; il existe d'autres variables, comme le sexe et le niveau de scolarité, qui n'ont pas d'incidence manifeste sur la mise en évidence de ces différences frappantes. Il est entendu que le fait de prendre en compte l'interaction entre plusieurs domaines et de les subdiviser réduirait considérablement la taille de l'échantillon. Dans de tels scénarios, on pourrait envisager d'appliquer des techniques d'estimation sur petits domaines. Ghosh (2020) présente un excellent examen des différents modèles et méthodes d'estimations sur petits domaines. On peut se demander s'il est possible de mettre au point un outil statistique, en se servant de méthodes d'apprentissage automatique comme les arbres de régression, afin de faciliter le repérage des variables importantes pour détecter les fausses réponses. Cela pourrait aider à découvrir l'effet des interactions entre les variables, plutôt que d'examiner une variable à la fois, comme le font KML. Dans le contexte de regroupements à l'aide de l'apprentissage automatique, Loh (2011) passe en revue certains algorithmes largement disponibles sur les arbres de classification et de régression.

La littérature sur les enquêtes non probabilistes a mis l'accent sur l'amélioration de la représentativité, c'est-à-dire la réduction du biais de sélection afin de rendre un échantillon plus représentatif de la population. Beaucoup de travail a été fait sur les méthodes de pondération par l'inverse du score de propension, où la propension est définie comme la probabilité de participation des unités de population dans l'échantillon non probabiliste. Valliant et Dever (2011), Chen, Li et Wu (2020), Wang, Valliant et Li (2021), Savitsky, Williams, Gershunskaya, Beresovsky et Johnson (2023) établissent des méthodes consistant à combiner ou à cumuler des enquêtes non probabilistes avec des enquêtes probabilistes ou de référence pour estimer les probabilités de participation, qui sont autrement inconnues pour les enquêtes non probabilistes du fait qu'on ignore quel est leur mécanisme de sélection. Ici, des hypothèses sur l'ignorabilité et la stricte positivité des scores de propension doivent être formulées. Ces méthodes permettent généralement de définir une fonction de log-vraisemblance en créant une variable indicatrice de succès si une unité est présente dans l'échantillon non probabiliste. L'information sur l'ensemble de la population finie étant inconnue, la modification subséquente de la vraisemblance en une pseudo-vraisemblance et l'utilisation d'une enquête de référence dépendent de la méthodologie de combinaison. Ainsi, les chercheurs estiment le score de propension à l'aide de différentes techniques d'intégration des données et appuient le rendement des estimateurs en ce qui concerne le biais et la variance à l'aide d'études de simulation et d'ensembles de données réelles. Plus précisément, Chen, Li et Wu (2020) calculent un estimateur doublement robuste pour les moyennes de population finie, le nom doublement robuste provenant de deux modèles : l'un étant le modèle de score de propension et l'autre, le modèle de régression des résultats.

Pour placer tous ces éléments dans le contexte de l'analyse sur l'article de KML, il serait intéressant de voir l'effet de l'utilisation des idées élaborées dans les articles mentionnés ci-dessus pour estimer les poids des enquêtes à participation volontaire. À l'heure actuelle, les poids en question sont établis au moyen du calage pour assurer la correspondance avec les caractéristiques de la population, mais de telles procédures d'estimation comportant l'intégration des données auront-elles une incidence sur les erreurs de mesure attribuables aux fausses réponses ? L'article de KML met en lumière le fait qu'il n'est pas conseillé d'utiliser directement les réponses des enquêtes à participation volontaire. Il incombe aux chercheurs de vérifier la crédibilité de ces réponses à l'aide d'enquêtes probabilistes existantes, comme l'analyse comparative avec des enquêtes gouvernementales. Pour les statisticiens d'enquête, il serait utile de savoir comment combiner adéquatement les enquêtes à participation volontaire avec d'autres sources qui aident à valider leur crédibilité et à améliorer les réponses ou à éliminer les fausses réponses. Essentiellement, les idées exceptionnelles formulées par KML orientent les lecteurs vers une piste de réflexion unique, qui met l'accent sur les erreurs non dues à l'échantillonnage. Cela, en combinaison avec la mise au point récente de méthodologies sur l'amélioration de la représentativité des échantillons non probabilistes, nous donne des résultats de recherche novateurs auxquels nous pouvons nous attendre.

Bibliographie

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.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition*, nouvelle série, numéro spécial sur la "Statistical Data Integration", 1-67.

Groves, R.M., et Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74, 5, 849-879. Doi: https://doi.org/10.1093/poq/nfq065.

Hansen, M.H., Hurwitz, W.N. et Bershad, M.A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 32^{e} session, 38, Partie 2, 359-74.

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining Knowl Discov*, 1, 14-23. Doi: https://doi.org/10.1002/widm.8.

Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovsky, V. et Johnson, N.G. (2023). Methods for combining probability and nonprobability samples under unknown overlaps. Doi: https://doi.org/10.48550/arXiv.2208.14541.

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

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40(4), 5237-5250.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 41-44
Statistique Canada, nº 12-001-X au catalogue

Réponse des auteurs aux commentaires sur l'article « Étude de l'hypothèse selon laquelle les répondants aux enquêtes non probabilistes en ligne menées à des fins commerciales répondent en toute bonne foi »

Courtney Kennedy, Andrew Mercer et Arnold Lau[1]

Résumé

Nos commentaires répondent aux points de discussion soulevés par Sen, Brick et Elliott. Nous évaluons les avantages et les inconvénients potentiels de la suggestion de Sen de recourir à l'apprentissage automatique pour repérer les faux répondants au moyen d'interactions et de combinaisons improbables de variables. Nous rejoignons la réflexion de Brick sur l'incidence des faux répondants sur les enquêtes non probabilistes menées à des fins commerciales. Enfin, nous examinons les solutions proposées par Elliott pour relever le défi exposé dans notre étude.

Mots-clés : Enquêtes non probabilistes menées à des fins commerciales; études d'étalonnage; panels d'enquête.

Nous remercions la direction de la revue d'avoir accueilli ce dialogue et les intervenants, de nous avoir fait part de leurs commentaires judicieux. Chacun d'entre eux apporte un point de vue unique. Sen relie notre étude à d'autres tendances dans les statistiques d'enquête. Brick propose des réflexions sérieuses sur l'état des enquêtes non probabilistes menées à des fins commerciales et aide à situer notre travail dans ce contexte. Elliott fait avancer la discussion sur les solutions au défi exposé dans notre étude.

Sen fait remarquer que les groupes démographiques mis en évidence dans notre étude sont subjectifs et non exhaustifs, car les chercheurs pourraient également s'intéresser au niveau de scolarité, aux variables géographiques, etc. Nous sommes d'accord avec l'idée générale et reconnaissons qu'il serait possible d'obtenir de nouveaux renseignements en élargissant le champ des variables en corrélation avec les fausses réponses. Nous apprécions également le fait qu'elle évoque l'apprentissage automatique comme un moyen possible de repérer les faux répondants grâce aux interactions et aux combinaisons improbables de variables. Le fait que l'apprentissage automatique soit évolutif et qu'il puisse s'adapter à l'évolution du comportement des répondants en fait une piste de recherche potentiellement fructueuse pour de futures études. Par ailleurs, nous doutons que la modélisation de petits domaines et les estimateurs doublement robustes soient capables d'améliorer la précision. Des études antérieures ont montré que pour les échantillons volontaires, ces méthodes n'offrent que des améliorations marginales par rapport aux méthodes d'étalonnage plus courantes (Mercer, Lau et Kennedy, 2018; Valliant, 2020). Il se peut que leur utilité soit limitée en raison du fait que, si ces méthodes sont excellentes pour corriger les problèmes liés à la sélection, elles sont mal adaptées pour traiter le problème des faux répondants, qui est fondamentalement une question d'erreur de mesure.

1. Courtney Kennedy, Andrew Mercer et Arnold Lau, Pew Research Center, 1615, L St., NW, Washington D.C., suite 800, 20036, États-Unis. Courriel : CKennedy@PewResearch.org.

Brick propose plusieurs réflexions de haut niveau axées sur l'industrie qui trouvent un écho chez nous. Il souligne ce qui suit : « La recherche d'un panel volontaire de haute qualité a été largement poursuivie pendant plus d'une décennie, mais les preuves recueillies jusqu'à présent indiquent qu'il s'agit d'une chimère » [traduction]. En effet, notre étude ainsi que celle de Geraci (2022), Enns et Rothschild (2022) et d'autres donnent à penser que l'émergence d'un tel panel volontaire devient moins probable et non plus probable. Auparavant, les statisticiens, dans ce contexte, se concentraient sur la modélisation afin de rendre les échantillons d'enquêtes non probabilistes menées à des fins commerciales plus représentatifs. Aujourd'hui, ils doivent relever un défi supplémentaire : déterminer quelles interviews sont réelles et lesquelles sont fausses.

Nous sommes également sensibles au fait que Brick souligne le rôle du fournisseur de données et nous saisissons dans quelle mesure le client (par exemple le chercheur) assume la charge du repérage et de la correction des types d'erreurs que nous consignons. Notre étude indique que les allégations de nettoyage des données figurant sur les sites Web des fournisseurs donnent une fausse impression de protection contre cette menace. En effet, si les vérifications de la qualité des fournisseurs fonctionnaient, les faux répondants ne figureraient pas dans les échantillons des clients, et des études comme la nôtre n'existeraient pas. Il est impératif que les chercheurs soient conscients de la menace que représentent les faux répondants, en particulier pour les estimations par domaine et les estimations de population complète qui donnent des résultats étonnants. À notre avis, cette menace est devenue si importante que les chercheurs qui publient des estimations ponctuelles à l'aide d'échantillons non probabilistes établis à des fins commerciales devraient inclure une analyse approfondie sur leur façon de gérer les faux répondants. Les rédacteurs en chef des revues ont probablement un rôle à jouer dans la promotion de cette pratique.

L'un des commentaires de Brick à propos de notre étude était particulièrement intéressant. En examinant le tableau 3.2, Brick constate dans quelle mesure la littérature portant sur la participation aux programmes montrent que la probabilité d'une fausse déclaration négative tend à être nettement plus élevée que la probabilité d'une fausse déclaration positive. Mais le tableau 3.2 montre la tendance inverse, de façon spectaculaire pour les échantillons des panels volontaires, et de façon plus discrète mais toujours perceptible pour les panels Internet recrutés par échantillonnage basé sur l'adresse. Nous sommes d'accord avec Brick pour dire que ce résultat contraire indique que les panels en ligne (tant volontaires que recrutés par échantillonnage basé sur l'adresse) ont un rendement différent de celui des échantillons probabilistes plus rigoureux sur ces résultats. Pour les panels volontaires, nous disposons d'une hypothèse raisonnablement solide : les faux répondants ont tendance à répondre par l'affirmative (par exemple « Oui », « D'accord ») quelle que soit leur situation réelle, parce qu'ils veulent se qualifier pour les enquêtes futures et gagner plus d'argent. Pour les panels Internet recrutés par échantillonnage basé sur l'adresse, cependant, nous n'avons connaissance d'aucune hypothèse permettant de prédire de fausses déclarations positives. Nous pensons que les différences entre les échantillons probabilistes rigoureux et les panels Internet probabilistes ne sont pas fondamentales par nature, et qu'elles sont probablement fonction des différences de mode, du conditionnement des panels et d'autres phénomènes bien connus dans la littérature concernant les méthodes

d'enquête. Cela dit, nous sommes d'accord avec Brick pour dire que la détermination des mécanismes précis à l'origine de ces différences semble être un terreau fertile pour le développement théorique et les travaux de recherche à venir.

Tous les intervenants ont proposé des solutions possibles au problème de la qualité des données exposé dans notre étude. Comme le laissent entendre les remarques de Brick, une solution consiste à décider simplement de ne pas utiliser d'échantillons non probabilistes établis à des fins commerciales. S'ils sont incontestablement moins chers et plus rapides, de nombreuses études (par exemple Dutwin et Buskirk, 2017; KML; MacInnis, Krosnick, Ho et Cho, 2018; Pennay, Neiger, Lavrakas et Borg, 2018; Yeager, Krosnick, Chang, Javitz, Levendusky, Simpser et Wang, 2011) montrent qu'ils sont moins exacts. Comme Brick, nous n'approuvons pas l'utilisation d'échantillons volontaires en supposant que l'on peut éliminer les cas de fausse déclaration. Nous sommes d'accord avec son observation selon laquelle, s'ils sont suffisamment motivés, les acteurs malveillants continueront à trouver des moyens de contourner les outils d'inspection dans les sources en ligne qui permettent aux personnes d'adhérer au processus.

Sen évoque la possibilité de réduire la pondération des répondants dont on a constaté qu'ils étaient susceptibles d'être faux à l'aide de questions de détection. Les perspectives de cette approche semblent dépendre du degré de validité des données fournies par les faux répondants. Pour les échantillons non probabilistes dans lesquels l'erreur de mesure est susceptible de provenir davantage de la satisfaction que de la fraude, cette approche semble prometteuse. Pour les échantillons d'enquêtes volontaires menées à des fins commerciales présentant des signes de fausses réponses (par exemple les cas des personnes répondant « Oui » quelle que soit la question), il est moins certain que le fait de conserver les cas de fausse réponse, même dans une capacité pondérée à la baisse, améliorerait les erreurs quadratiques moyennes. Heureusement, ces questions peuvent être mises à l'essai et, comme Sen, nous serions heureux de les approfondir.

Elliott rejoint Wu (2022) dans son plaidoyer en faveur d'enquêtes permanentes suffisamment rigoureuses pour produire des données repères de haute qualité à utiliser pour calibrer des enquêtes moins rigoureuses. Nous soutenons cette proposition avec enthousiasme. Au Pew Research Center, nous avons pris des mesures modestes dans ce sens, en créant une enquête annuelle multimodale basée sur les adresses, conçue pour produire des estimations de données repères en temps opportun sur l'affiliation des Américains à un parti politique, leur appartenance religieuse et leur utilisation de la technologie (Pew Research Center, 2022). Cette étude multimodale témoigne de la plus grande rigueur que nous puissions appliquer à l'aide des ressources de notre établissement institutionnel, mais le type d'investissement proposé par Elliott permettrait d'obtenir des plans d'échantillonnage beaucoup plus poussés (par exemple en ayant une phase de collecte de données en personne). Il est clair pour nous que ces nouvelles études d'étalonnage sont nécessaires pour améliorer les échantillons probabilistes à très faible taux de réponse comme les trois qui sont présentés dans notre étude. La question de savoir si les études d'étalonnage peuvent sauver les échantillons d'enquêtes non probabilistes menées à des fins commerciales reste, à notre avis, ouverte, étant donné le défi que pose, pour les variables de pondération et de résultat, les répondants qui font intentionnellement une fausse déclaration.

Bibliographie

Dutwin, D., et Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81, 213-239.

Enns, P., et Rothschild, J. (2022). Do you know where your survey data come from? Outsourcing data collection poses huge risks for public opinion. Medium, accessible à l'adresse https://medium.com/3streams/surveys-3ec95995dde2.

Geraci, J. (2022). *Poll-arized: Why Americans Don't Trust the Polls and How to Fix Them Before It's Too Late*. Houndstooth Press, p. 153.

MacInnis, B., Krosnick, J.A., Ho, A.S. et Cho, M. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, 82, 707-744.

Mercer, A., Lau, A. et Kennedy, C. (2018). For weighting online opt-in samples, what matters most? *Pew Research Center*. http://www.pewresearch.org/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

Pennay, D.W., Neiger, D., Lavrakas, P.J. et Borg, K. (2018). The online panels benchmarking study: A total survey error comparison of findings from probability-based surveys and nonprobability online panel surveys in Australia (2e Éds.) The Australian National University. https://csrm.cass.anu.edu.au/research/publications/methods-research-papers?search_term=The+online+panels+benchmarking.

Pew Research Center (2022). National Public Opinion Reference Survey (NPORS). Accessible à l'adresse https://www.pewresearch.org/methods/fact-sheet/national-public-opinion-reference-survey-npors/.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), 231-263.

Wu, C. (2022). Inférence statistique avec des échantillons d'enquête non probabiliste. *Techniques d'enquête*, 48, 2, 307-338. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00002-fra.pdf. Avec discussion disponible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2022002-fra.htm.

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A. et Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75, 709-747.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 45-65
Statistique Canada, nº 12-001-X au catalogue

Hypothèse de l'échangeabilité dans des méthodes d'ajustement fondées sur le score de propension aux fins d'estimation de la moyenne de population au moyen d'échantillons non probabilistes

Yan Li[1]

Résumé

Des échantillons non probabilistes émergent rapidement pour aborder des sujets prioritaires urgents dans différents domaines. Ces données sont actuelles, mais sujettes à un biais de sélection. Afin de réduire le biais de sélection, une littérature abondante portant sur la recherche sur les enquêtes a étudié l'utilisation de méthodes d'ajustement par le score de propension (SP) pour améliorer la représentativité de la population des échantillons non probabilistes, au moyen d'échantillons d'enquête probabilistes utilisés comme références externes. L'hypothèse d'échangeabilité conditionnelle (EC) est l'une des principales hypothèses requises par les méthodes d'ajustement fondées sur le SP. Dans le présent article, j'examine d'abord la validité de l'hypothèse de l'EC conditionnellement à plusieurs estimations de scores d'équilibrage qui sont utilisées dans les méthodes d'ajustement fondées sur le SP existantes. Un score d'équilibrage adaptatif est proposé aux fins d'estimation sans biais des moyennes de population. Les estimateurs de la moyenne de population selon les trois hypothèses de l'EC sont évalués au moyen d'études de simulation de Monte Carlo et illustrés au moyen de l'étude sur la séroprévalence du SRAS-CoV-2 des National Institutes of Health pour estimer la proportion d'adultes aux États-Unis qui présentaient des anticorps de la COVID-19 du 1er avril au 4 août 2020.

Mots-clés : Appariement par scores de propension; échantillon par quota; estimateur de la variance par linéarisation en séries de Taylor; étude de séroprévalence du SRAS-CoV-2; pondération par scores de propension; score d'équilibrage.

1. Introduction

Des échantillons non probabilistes émergent rapidement pour aborder des sujets prioritaires urgents des sujets prioritaires urgents dans différents domaines (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau, 2013; Kennedy, Mercer, Keeter, Hatley, McGeeney et Gimenez, 2016). Ces données sont actuelles, mais sujettes à un biais de sélection. Les participants sont souvent autosélectionnés et se portent volontaires pour participer à une étude sans probabilités de sélection préétablies. Les exemples comprennent des échantillons épidémiologiques composés de volontaires qui ne sont pas choisis aléatoirement et qui ne sont donc généralement pas représentatifs d'une population. De plus, les volontaires sont souvent sujets à des « effets de volontaires sains » (Pinsky, Miller, Kramer, Church, Reding, Prorok, Gelmann, Schoen, Buys, Hayes et Berg, 2007), ce qui se traduit habituellement par des estimations plus faibles de l'incidence de la maladie et de la mortalité chez les volontaires que dans la population générale. Un autre exemple concerne les données recueillies à partir de panels Web à échantillonnage probabiliste, qui peuvent donner un taux d'attrition élevé et dans lesquels les taux de non-réponse sont souvent de 90 % ou plus (Baker et coll., 2013). Bien qu'une non-réponse élevée ne soit pas nécessairement un indice du biais dans les

1. Yan Li, Joint Program in Survey Methodology and Department of Epidemiology and Biostatistics, University of Maryland, College Park. Courriel : yli6@umd.edu.

réponses (Groves et Peytcheva, 2008; Brick et Tourangeau, 2017), le biais de sélection est très préoccupant parce que la composition des panels Web diffère souvent de celle de la population sous-jacente.

Contrairement aux échantillons non probabilistes, les enquêtes probabilistes basées sur la population sont conçues pour produire des estimations presque sans biais des caractéristiques de la population. Elles s'appuient sur des plans d'échantillonnage probabiliste, comme l'échantillonnage stratifié par grappes à plusieurs degrés, pour sélectionner des échantillons. Les échantillons qui en résultent, s'ils sont correctement pondérés par les poids d'enquête, peuvent représenter fidèlement la population cible; ils sont par conséquent moins sujets au biais de sélection.

Afin de réduire le biais de sélection des échantillons non probabilistes, une littérature abondante portant sur la recherche sur les enquêtes a étudié l'utilisation de méthodes d'ajustement par le score de propension (SP) pour améliorer la représentativité de la population des échantillons non probabilistes, au moyen des échantillons d'enquête probabiliste comme références externes (Elliott et Valliant, 2017). Différentes méthodes d'ajustement basées sur le SP ont été élaborées; elles peuvent être regroupées en deux catégories : 1) les méthodes de pondération par l'inverse du SP (par exemple Chen, Li et Wu, 2020; Elliott, 2013; Valliant et Dever, 2011) ou les méthodes de pondération par cotes inverses (par exemple Wang, Valliant et Li, 2021) (*pondération* par SP); 2) les méthodes d'appariement par SP ou par logarithme du risque du SP (*appariement* par SP) (par exemple Lee et Valliant, 2009; Wang, Graubard, Katki et Li, 2022; Rivers, 2007).

Les méthodes de pondération par SP établissent une pseudo-pondération pour chaque unité d'échantillon non probabiliste comme étant l'inverse de sa propension à la participation. Elles corrigent le biais de sélection selon les vrais modèles de propension, bien qu'elles puissent être sensibles à la spécification erronée du modèle de propension (Valliant, 2020) et produire des estimations présentant de grandes variances en raison des poids extrêmes (Stuart, 2010). En revanche, les méthodes d'appariement par SP s'appuient sur le score de propension pour mesurer la similarité dans les distributions des covariables comprises dans le modèle de propension entre l'enquête probabiliste et l'échantillon non probabiliste; elles tendent par conséquent à être moins sensibles à la spécification erronée du modèle de propension. De plus, comme les méthodes d'appariement par SP évitent les poids extrêmes, elles produisent des estimations présentant des variances plus petites. Pour obtenir une synthèse exhaustive sur les autres méthodes d'analyse des échantillons non probabilistes et d'intégration des données, consulter Beaumont (2020), Rao (2021) et Valliant (2020).

Les méthodes d'ajustement fondées sur le SP (par exemple Chen et coll., 2020) nécessitent les hypothèses clés suivantes pour faire des inférences d'échantillons non probabilistes. Premièrement, l'échantillon de l'enquête de référence, par la pondération, représente correctement la population finie (PF) d'intérêt. Deuxièmement, toutes les unités de la PF ont une propension à la participation positive (c'est-à-dire que tous les membres de la population ont une propension positive à participer à des échantillons non probabilistes). Troisièmement, l'échangeabilité conditionnelle (EC) se vérifie sans covariables non mesurées, c'est-à-dire que la probabilité que chaque membre de la PF participe à l'échantillon non probabiliste n'est pas liée à son résultat, conditionnellement à toutes les covariables

mesurées. Quatrièmement, le fait d'être échantillonné dans le cadre de l'enquête de référence et de participer à l'échantillon non probabiliste sont indépendants. Toutes ces hypothèses sont essentielles. Dans le présent article, nous nous intéressons à l'hypothèse de l'EC et nous examinons plusieurs scores d'équilibrage (c'est-à-dire des fonctions de covariables) qui satisfont à l'hypothèse de l'EC.

Dans les études observationnelles pour les inférences causales, les chercheurs tentent généralement d'ajuster pour toutes les covariables mesurées, afin d'imiter une expérience complètement randomisée et supposent que de tels ajustements sont suffisants pour des estimations sans biais des effets du traitement. Cette hypothèse est connue sous le nom d'« échangeabilité des assignations de traitement » (Rubin, 1978). Toutefois, la recherche sur les enquêtes vise à faire une inférence au sujet des paramètres de la PF, et il existe peu de recherches sur la suffisance de l'hypothèse mentionnée ci-dessus. Certaines études (par exemple celle de Wang et coll., 2021) ont mentionné la nécessité de faire des hypothèses sur le fait que la propension à la participation est ignorable étant donné un ensemble de variables d'ajustement. Cependant, dans ces études, on s'est contenté d'observer la présence ou l'absence d'estimations biaisées, et on a rarement examiné si et dans quelle mesure on va à l'encontre de l'hypothèse de l'EC quand on fait une inférence à propos des paramètres de la population finie.

La contribution du présent article consiste à : 1) étudier la validité de l'hypothèse de l'EC qui est conditionnelle à diverses estimations de scores d'équilibrage qui sont utilisées dans les méthodes actuelles d'ajustement par le SP, y compris les méthodes de pondération par le SP et d'appariement par SP, pour les inférences d'échantillons non probabilistes; 2) élaborer un score d'équilibrage adaptatif pour l'hypothèse de l'EC afin d'en améliorer l'efficacité. Dans l'article, nous n'élaborons pas de nouvelles méthodes d'ajustement fondées sur le SP, mais nous étudions divers scores d'équilibrage qui satisfont à l'hypothèse de l'EC. La pondération par SP en utilisant la méthode de la propension logistique ajustée (PLA) est utilisée à des fins d'illustration. Le score d'équilibrage établi peut également être utilisé dans des méthodes d'appariement par SP comme le lissage par la méthode du noyau (Wang et coll., 2022). Les estimateurs PLA, supposant l'échangeabilité des résultats conditionnelle à divers scores d'équilibrage, sont évalués via des études de simulation de Monte Carlo et illustrés au moyen de l'étude sur la séroprévalence du SARS-CoV-2 des National Institutes of Health pour estimer la proportion d'adultes des États-Unis présentant des anticorps contre la COVID-19 du 1er avril au 4 août 2020.

2. Hypothèse de l'échangeabilité conditionnelle

2.1 Notation

Considérons une population finie (PF) cible comme un échantillon aléatoire de N personnes tirées d'un modèle de superpopulation, indexé par $U = \{1, 2, \ldots, N\}$, avec des observations sur une variable d'étude y et un vecteur de covariables $\mathbf{x}$. Soit $\{y_i, \mathbf{x}_i : i \in C\}$ les observations dans l'échantillon non probabiliste de personnes, où $C \subset U$ de taille n_c. Nous cherchons à estimer la moyenne de la PF $\overline{Y}_N = \frac{1}{N}\sum_{i \in U} y_i$ au moyen

de l'échantillon non probabiliste C. Le problème est que nous observons C, qui, cependant, peut ne pas être un échantillon représentatif tiré de U. Par conséquent, $E_C(y \mid U) \neq \overline{Y}_N$, où l'indice C désigne le caractère aléatoire en raison du processus de participation à l'échantillon non probabiliste inconnu tiré de U. Soit $E(y \mid C) = E_U(E_C(y \mid U))$ et $E(y \mid U) = E_U(\overline{Y}_N)$, où l'indice U désigne l'espérance sous le modèle de superpopulation. L'espérance de y dans C peut différer de celle dans U, à savoir $E(y \mid C) \neq E(y \mid U)$ en raison du biais de sélection de l'échantillon non probabiliste C.

2.2 Hypothèse de l'échangeabilité conditionnelle et score d'équilibrage

Pour obtenir un estimateur convergent par rapport au plan de sondage de $\overline{Y}_N$ au moyen de C, l'EC suppose

$$E\{y \mid b(\mathbf{x}), C\} = E\{y \mid b(\mathbf{x}), U\}, \tag{2.1}$$

où $b(\mathbf{x})$ est une fonction des covariables $\mathbf{x}$, qu'on appelle le score d'équilibrage.

L'hypothèse de l'EC (2.1) indique que conditionnellement au score d'équilibrage $b(\mathbf{x})$, c'est-à-dire une fonction de covariables mesurées, le résultat a la même espérance dans C que dans U. Autrement dit, les unités de l'échantillon non probabiliste ayant la même valeur de score d'équilibrage $b(\mathbf{x})$ représenteraient le même nombre d'unités de la PF. Intuitivement, si deux personnes avaient la même propension à la participation, elles représenteraient le même nombre d'unités de la PF. Par conséquent, un choix naturel de $b(\mathbf{x})$ est la propension à la participation $P(i \in C \mid \mathbf{x}, U)$, c'est-à-dire la probabilité que l'unité de la PF i participe à C conditionnellement à la valeur de $\mathbf{x}$.

De façon plus générale, le *critère de base* pour choisir un score d'équilibrage est que $b(\mathbf{x})$ est plus fin que, sinon égal à, $P(i \in C \mid \mathbf{x}, U)$ pour que l'hypothèse de l'EC soit valide (2.1). Par conséquent, le choix le plus fin de score d'équilibrage est $b(\mathbf{x}) = \mathbf{x}$ et le moins fin est $b(\mathbf{x}) = P(i \in C \mid \mathbf{x}, U)$ ou sa fonction monotone. Par conséquent, les $b(\mathbf{x})$ choisis doivent pouvoir permettre de distinguer les unités C ayant des propensions à la participation différentes.

Dans l'inférence causale (Rosenbaum et Rubin, 1983), l'hypothèse de l'échangeabilité conditionnelle indique que le résultat est échangeable entre le groupe *traité* et le groupe *témoin*, conditionnellement à toutes les covariables mesurées. La distribution des covariables dans le groupe traité est appariée à celle dans le groupe témoin au moyen de méthodes de pondération par SP ou d'appariement par SP, selon un modèle de *propension de l'attribution du traitement (treatment assignment propensity model)*. On estime ensuite l'effet du traitement en comparant les moyennes des deux groupes après pondération ou appariement. De façon analogue, dans les inférences d'échantillons non probabilistes, la distribution des covariables dans l'échantillon non probabiliste est appariée à celle dans la PF selon un modèle de *propension à la participation* (*échantillon non probabiliste*). Au lieu d'estimer l'effet du traitement, on estime la moyenne de la PF en supposant l'échangeabilité du résultat entre l'échantillon non probabiliste et la PF après la pondération par SP ou l'appariement par SP. Pour en savoir plus, les lecteurs trouveront dans Mercer, Kreuter, Keeter et Stuart (2017) des précisions sur les parallèles entre l'inférence causale et l'inférence d'échantillon non probabiliste.

3. Scores d'équilibrage existants

3.1 Estimation de $P(i \in C \mid \mathbf{x}, U)$

On peut estimer directement la propension à la participation $P(i \in C \mid \mathbf{x}, U)$ si les covariables $\mathbf{x}$ sont connues pour toutes les personnes dans U. Malheureusement, nous n'avons pas la mesure de $\mathbf{x}$ pour l'ensemble de U, mais on peut estimer sa distribution à partir d'un échantillon probabiliste S de taille n_s, $\{\mathbf{x}_i : i \in S\}$. Différentes méthodes de modélisation de la propension où S est utilisé comme enquête de référence ont été proposées (Chen et coll., 2020; Kern, Li et Wang, 2021). À titre d'illustration, nous supposons un modèle de régression logistique

$$\log\left\{\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right\} = \mathbf{B}^T g(\mathbf{x}_i), \quad \text{pour } i \in U, \tag{3.1}$$

où le score de propension $p(\mathbf{x}_i)$ est la propension de l'unité i à faire partie de l'échantillon non probabiliste par rapport à la population finie, selon une approximation de l'échantillon d'enquête pondéré, désigné par S_w. De même, $\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)} = P(i \in C \mid \mathbf{x}_i, U)$. $g(\mathbf{x}_i)$ est une fonction connue des covariables observées, et $\mathbf{B}$ sont les coefficients de régression inconnus qu'il faut estimer; voir dans Wang et coll. (2021, section 2.3) la justification du modèle de propension (3.1). Nous définissions w_i comme le poids d'échantillon de l'unité $i \in S$. Quand on résout $S(\mathbf{B}) = \left\{\sum_{i' \in C}(1-p(\mathbf{x}_{i'}))\, g(\mathbf{x}_{i'}) - \sum_{i \in S} w_i p(\mathbf{x}_i)\, g(\mathbf{x}_i)\right\} = 0$ pour $\mathbf{B}$, l'estimation est désignée par $\hat{\mathbf{B}}_w$. L'indice w indique que les poids de l'enquête de référence servent à estimer $\mathbf{B}$ dans le modèle de propension (3.1). La propension à la participation $P(i \in C \mid \mathbf{x}, U)$ pour $i \in C \cup S$ peut être estimée par $\exp(\mathbf{x}_i \hat{\mathbf{B}}_w) = \frac{\hat{p}(\mathbf{x}_i, \hat{\mathbf{B}}_w)}{1-\hat{p}(\mathbf{x}_i, \hat{\mathbf{B}}_w)}$, $\hat{p}(\mathbf{x}_i, \hat{\mathbf{B}}_w)$ étant l'estimation du score de propension $p(\mathbf{x}_i)$.

3.2 Hypothèse de l'échangeabilité conditionnelle à $b(\mathbf{x}; \hat{\mathbf{B}}_w)$

Pour satisfaire à l'hypothèse de l'EC (2.1), le score d'équilibrage doit être aussi fin ou plus fin que le taux de participation estimé. D'après les conclusions de Wang et coll. (2022), le prédicteur linéaire, c'est-à-dire une transformation logarithmique naturelle de la propension à la participation estimée, est utilisé comme score d'équilibrage, à savoir $b(\mathbf{x}_i; \hat{\mathbf{B}}_w) = \hat{\mathbf{B}}_w^T g(\mathbf{x}_i) = \log \hat{p}(i \in C \mid \mathbf{x}_i, U)$. Alors, selon le modèle de propension (3.1), on suppose que $b(\mathbf{x}) = b(\mathbf{x}; \hat{\mathbf{B}}_w)$ dans (2.1), c'est-à-dire que

$$E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_w), C\} = E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_w), U\}$$

se vérifie approximativement. Comme dans ce qui suit, nous estimons la moyenne de la population au moyen de plusieurs méthodes existantes d'ajustement fondées sur le SP. Par exemple, la méthode de pondération par SP par la propension logistique ajustée ou PLA (Wang et coll., 2021) pondère l'unité i dans C par l'inverse de $\hat{p}(i \in C \mid \mathbf{x}_i, U) = \exp(b(\mathbf{x}_i; \hat{\mathbf{B}}_w))$. Un autre exemple est la méthode d'appariement par SP par pondération du noyau (Wang et coll. 2022), qui apparie les unités dans C et S selon la similarité dans $b(\mathbf{x}; \hat{\mathbf{B}}_w)$. Il a été prouvé que les estimations par la PLA et par pondération du noyau sont approximativement sans biais selon l'hypothèse de l'EC conditionnellement à $b(\mathbf{x}; \hat{\mathbf{B}}_w)$.

Toutefois, un inconvénient important de cette méthode est l'inflation de la variance potentiellement importante dans $b(\mathbf{x};\hat{\mathbf{B}}_w)$ en raison de la variabilité (Scott et Wild, 2001; Li, Graubard et DiGaetano, 2011) des poids de l'enquête de référence différentiels par rapport aux poids de l'échantillon non probabiliste (= 1) dans l'estimation du paramètre du modèle $\mathbf{B}$. Aux fins de réduction de la variance, les poids de l'enquête n'ont pas été pris en compte dans l'estimation de $\mathbf{B}$ (Wang, Graubard, Katki et Li, 2020; Lee et Valliant, 2009).

3.3 Hypothèse de l'échangeabilité conditionnelle à $b(\mathbf{x};\hat{\mathbf{B}}_0)$

Supposons que le modèle de propension (3.1) est ajusté aux données combinées de l'échantillon non probabiliste et de l'enquête non pondérées $(C \cup S)$ et que le score d'équilibrage qui en résulte est $b(\mathbf{x}_i,\hat{\mathbf{B}}_0)=\hat{\mathbf{B}}_0^T g(\mathbf{x}_i)$, $\hat{\mathbf{B}}_0$ ayant été obtenu en résolvant l'équation d'estimation $S(\mathbf{B})=\left\{\sum_{i'\in C}(1-p(\mathbf{x}_{i'}))\,g(\mathbf{x}_{i'})-\sum_{i\in S}p(\mathbf{x}_i)\,g(\mathbf{x}_i)\right\}=0$ pour $\mathbf{B}$, sans tenir compte des poids de l'échantillon probabiliste. Par conséquent, l'EC fondée sur $\hat{\mathbf{B}}_0$, supposée par les méthodes existantes de pondération par SP ou d'appariement par SP (Wang et coll., 2020; Lee et Valliant, 2009; Kern et coll., 2021), est

$$E\{y \mid b(\mathbf{x};\hat{\mathbf{B}}_0), C\} = E\{y \mid b(\mathbf{x};\hat{\mathbf{B}}_0), U\}.$$

Lorsque les poids d'enquête ne sont pas utilisés, le score d'équilibrage estimé $b(\mathbf{x},\hat{\mathbf{B}}_0)$ peut être plus stable que $b(\mathbf{x},\hat{\mathbf{B}}_w)$. La question consiste à déterminer dans quelle mesure l'hypothèse de l'EC conditionnellement à $b(\mathbf{x},\hat{\mathbf{B}}_0)$ est plausible dans des problèmes réels.

Mentionnons que $b(\mathbf{x},\hat{\mathbf{B}}_0)$ produit une distribution de $\mathbf{x}$ équilibrée entre C et S, et que par conséquent, l'échangeabilité de la distribution de y (tous les $\mathbf{x}$ étant équilibrés) se vérifie entre C et S, ce qui toutefois ne suffit pas pour obtenir une estimation sans biais de la moyenne de la PF. En effet, il est nécessaire d'avoir l'échangeabilité de la distribution de y entre C et U conditionnellement à $b(\mathbf{x},\hat{\mathbf{B}}_0)$. Nous savons d'après la section 2.2 que $P(i\in C \mid \mathbf{x},U)$ est le score d'équilibrage le plus grossier satisfaisant à (2.1) et que $b(\mathbf{x},\hat{\mathbf{B}}_w)$ produit approximativement une distribution de y équilibrée entre C et U. Selon les critères de base du choix de score d'équilibrage, le score d'équilibrage $b(\mathbf{x},\hat{\mathbf{B}}_0)$ doit être aussi fin ou plus fin que $b(\mathbf{x},\hat{\mathbf{B}}_w)$. Un exemple en est que $b(\mathbf{x}_i;\hat{\mathbf{B}}_0)=\hat{\mathbf{B}}_0^T g(\mathbf{x}_i)$ est une fonction linéaire de $b(\mathbf{x}_i;\hat{\mathbf{B}}_w)=\hat{\mathbf{B}}_w^T g(\mathbf{x}_i)$, c'est-à-dire $\hat{\mathbf{B}}_w^T=\text{const.}\times\hat{\mathbf{B}}_0^T$. Supposons que l'enquête de référence S suréchantillonne, de par son plan de sondage, un groupe minoritaire, par exemple les femmes afro-américaines. Cette relation linéaire exige que la distribution du même groupe minoritaire, défini par la race ou l'origine ethnique et le genre dans l'échantillon non probabiliste, soit proportionnelle à celle de l'enquête de référence. Or, en réalité, nous n'avons aucun contrôle sur l'échantillonnage non probabiliste et, par conséquent, la relation linéaire ne se vérifie que par hasard. L'estimateur fondé sur $b(\mathbf{x};\hat{\mathbf{B}}_0)$ est efficace, mais il peut être biaisé.

Exemple hypothétique

À des fins d'illustration, supposons qu'un échantillon non probabiliste et un échantillon d'enquête sont sélectionnés par échantillonnage avec probabilité proportionnelle à la taille (PPT), avec la mesure de taille

pour l'unité i de la PF définie, respectivement, par $s_{ic} = \exp(x_{i1}B_1 + x_{i2}B_2)$ pour la participation à l'échantillon non probabiliste et $s_{is} = \exp(x_{i1}B_1' + x_{i3}B_3)$ pour la sélection de l'échantillon d'enquête. Supprimons l'indice i et supposons que $B_1 \approx B_1'$. La probabilité qu'une unité de la PF participe à l'échantillon non probabiliste (p_c) par rapport à sélection dans l'enquête (p_s) est

$$\begin{aligned}\log\left(\frac{p_c}{p_s}\right) &= \log\left(\frac{n_c s_c}{\sum_U s_c} \Big/ \frac{n_s s_s}{\sum_U s_s}\right) = \log\left(\frac{n_c \sum_U s_s}{n_s \sum_U s_c} \times \frac{s_c}{s_s}\right), \\ &= \text{const.} + x_1(B_1 - B_1') + x_2 B_2 - x_3 B_3 = \text{const.} + \mathbf{x}^T \mathbf{B}_0,\end{aligned}$$

où $\mathbf{x} = (x_1, x_2, x_3)^T$ et $\mathbf{B}_0 = (B_1 - B_1', B_2, -B_3)^T$. En ajustant un modèle logistique, y compris toutes les variables $\mathbf{x}$, à l'échantillon combiné (enquête non probabiliste et enquête non pondérée), un score d'équilibrage estimé serait

$$b(\mathbf{x}; \hat{\mathbf{B}}_0) = \mathbf{x}^T \hat{\mathbf{B}}_0.$$

Mentionnons que $\hat{\mathbf{B}}_0$ inclut l'effet x_1 atténué dans la construction $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ en raison de la distribution similaire de x_1 dans S et dans C. Par conséquent, les scores d'équilibrage estimés ne peuvent pas distinguer les unités C ayant différentes propensions à la participation par x_1, et donc $E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0), C\} \neq E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0), U\}$, ce qui conduit à une estimation biaisée de $\bar{Y}_N$.

Dans la section suivante, nous proposons un score d'équilibrage adaptatif qui ajuste $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ pour que ce soit une fonction monotone de l'estimation de $P(i \in C \mid \mathbf{x}, U)$ pour l'estimation sans biais de la moyenne de la PF.

4. Score d'équilibrage adaptatif

Nous proposons un score d'équilibrage ajusté en trois étapes. La première étape consiste à adapter un modèle de régression logistique à l'échantillon $C \cup S$ combiné sans poids, donné par (Wang et coll., 2020)

$$\log\left\{\frac{p(i \in C \mid \mathbf{x}_i, U)}{p(i \in S \mid \mathbf{x}_i, U)}\right\} = \log\left\{\frac{p^*(\mathbf{x}_i)}{1 - p^*(\mathbf{x}_i)}\right\} = \mathbf{B}_0^T g(\mathbf{x}_i) \quad \text{pour } i \in U \tag{4.1}$$

et les estimations du paramètre du modèle $\mathbf{B}_0$ sont désignées par $\hat{\mathbf{B}}_0$, où $p^*(\mathbf{x}_i)$ est la propension à être dans C par rapport à être dans S pour l'unité i. Comme nous l'avons vu à la section 3.3, $b(\mathbf{x}; \hat{\mathbf{B}}_0) = \hat{\mathbf{B}}_0^T g(\mathbf{x}_i)$ équilibre la distribution de $\mathbf{x}$ entre C et S. Si l'on ne tient pas compte des poids de l'échantillon dans l'analyse, $\hat{\mathbf{B}}_0$ tend à être plus efficace que $\hat{\mathbf{B}}_w$. Il peut toutefois y avoir violation de l'hypothèse de l'EC (2.1) quand $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ n'arrive pas à équilibrer la distribution dans $\mathbf{x}$ entre C et U.

La deuxième étape vise à élaborer un facteur de correction du biais pour ajuster $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ de façon à ce que la distribution équilibrée dans $\mathbf{x}$ entre C et U (obtenue par approximation par l'enquête de référence pondérée S_w) puisse être atteinte. En tant que dispositif *de calcul*, on construit une pseudo-population de $S^* \cup U$ où S^* est un double de S qui a les mêmes distributions conjointes de covariables $\mathbf{x}$ et le même

résultat y que l'original S. Dans la pseudo-population $S^* \cup U$, S^* et S sont traités comme deux ensembles différents. Nous modélisons $q(\mathbf{x}_i)$ comme étant la probabilité que l'unité i soit incluse dans S à partir de la pseudo-population, c'est-à-dire

$$q(\mathbf{x}_i) = p(i \in S \mid \mathbf{x}_i, S^* \cup U) = \frac{p(i \in S \mid \mathbf{x}_i, U)}{1 + p(i \in S \mid \mathbf{x}_i, U)}.$$

Supposons un modèle logistique

$$\log\{p(i \in S \mid \mathbf{x}_i, U)\} = \log\left\{\frac{q(\mathbf{x}_i)}{1 - q(\mathbf{x}_i)}\right\} = \boldsymbol{\gamma}^T g(\mathbf{x}_i), \quad \text{pour } i \in U \tag{4.2}$$

où $\boldsymbol{\gamma}$ désigne les paramètres du modèle, estimés par la résolution de l'équation estimant $S(\boldsymbol{\gamma}) = \sum_{i \in S}(1 - q(\mathbf{x}_i) - w_i q(\mathbf{x}_i))\, g(\mathbf{x}_i) = 0$ pour $\boldsymbol{\gamma}$. L'estimation est désignée par $\hat{\boldsymbol{\gamma}}_w$ et mesure les effets de $g(\mathbf{x})$ sur la sélection de l'échantillon S. Nous l'utilisons pour corriger les effets déformés ou manquants de $g(\mathbf{x})$ sur la propension à la participation de l'échantillon non probabiliste C dans $b(\mathbf{x}; \hat{\mathbf{B}}_0)$, en particulier pour les variables intervenant à la fois dans les processus d'échantillonnage S et de participation C.

À l'étape 3, le nouveau score d'équilibrage estimé est construit comme étant

$$b(\mathbf{x}_i; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w) = (\hat{\boldsymbol{\gamma}}_w^T + \hat{\mathbf{B}}_0^T)\, g(\mathbf{x}_i) \quad \text{pour } i \in U.$$

Comme nous l'avons mentionné, l'addition des modèles (4.1) et (4.2) donne le modèle (3.1), le premier membre étant égal à

$$\log\left\{\frac{p(i \in C \mid \mathbf{x}_i, U)}{p(i \in S \mid \mathbf{x}_i, U)}\right\} + \log\{p(i \in S \mid \mathbf{x}_i, U)\} = \log\{P(i \in C \mid \mathbf{x}_i, U)\},$$

une fonction monotone de la propension à la participation, et le deuxième membre la même forme fonctionnelle $g(\mathbf{x}_i)$ que dans le modèle (3.1). Nous savons que $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ dans le modèle (3.1), bien que satisfaisant à l'hypothèse de l'EC (2.1), peut être inefficace en raison des poids différentiels dans l'analyse. Au lieu d'ajuster le modèle (3.1) directement aux données combinées de l'échantillon non probabiliste et de l'enquête pondérée $(C \cup S_w)$ pour obtenir $\hat{\mathbf{B}}_w$, nous construisons le score d'équilibrage ajusté $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w)$ basé sur $\hat{\mathbf{B}}_0$ et $\hat{\boldsymbol{\gamma}}_w$ en trois étapes. Ce score d'équilibrage ajusté est une fonction monotone (logarithme naturel) de la propension à la participation de l'échantillon C, et par conséquent la distribution de y est échangeable entre C et U, c'est-à-dire que

$$E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T), C\} = E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T), U\}$$

se vérifie approximativement.

Comme dans ce qui suit, des méthodes d'ajustement fondées sur le SP peuvent servir à créer des pseudo-pondérations pour les unités dans C en se basant sur le nouveau score d'équilibrage adaptatif $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$. Les méthodes de pondération par SP pondèrent chaque unité dans C par l'inverse du taux de participation

estimé. En revanche, les méthodes d'appariement par SP apparient les unités C et S en s'appuyant sur le score d'équilibrage adaptatif, puis distribuent les poids de sondage dans S aux unités C selon leurs similarités. Par exemple, la méthode de pondération par la PLA (Wang et coll., 2021) crée des pseudo-pondérations.

$$\hat{w}_j^{\text{PLA}} = \exp^{-1}(b(\mathbf{x}_j; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)) \quad \text{pour } j \in C.$$

Le lissage par la méthode du noyau (Wang et coll., 2022) crée des pseudo-pondérations en additionnant les poids fractionnels distribués à partir de chaque unité d'enquête $i \in S$,

$$\hat{w}_j^{\text{KW}} = \sum_{i \in S} w_i K_{ij} \quad \text{avec} \quad K_{ij} = \frac{K\left(\frac{d_{ij}}{h}\right)}{\sum_{l \in C} K\left(\frac{d_{il}}{h}\right)} \quad \text{pour} \quad j \in C,$$

où $K(\cdot)$ est une fonction de noyau arbitraire telle que la fonction de densité normale standard, h est la largeur de bande associée à $K(\cdot)$, et la distance $d_{ij} = b(\mathbf{x}_i; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T) - b(\mathbf{x}_j; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ mesure la similarité dans la distribution de $\mathbf{x}$ entre l'unité d'échantillonnage non probabiliste $j \in C$ et l'unité d'enquête $i \in S$.

La moyenne de population peut ensuite être estimée par

$$\bar{y} = \frac{1}{\sum_{j \in C} \hat{w}_j} \sum_{j \in C} \hat{w}_j y_j, \tag{4.3}$$

où $\hat{w}_j$ peut être $\hat{w}_j^{\text{ALP}}$ ou $\hat{w}_j^{\text{KW}}$.

Pour estimer la variance de $\bar{y}$, nous supposons la taille de PF $N \to \infty$ et considérons le caractère aléatoire attribuable à l'échantillonnage de S et au processus de participation de C tiré de U. On élabore l'estimateur de la variance par linéarisation en séries de Taylor (LT) pour tenir compte de la variabilité attribuable à l'estimation des scores de propension $p^*(\mathbf{x}_i)$ et $q(\mathbf{x}_i)$ aux étapes 1 et 2. La technique de linéarisation en séries de Taylor est couramment utilisée dans les ouvrages publiés portant sur les enquêtes pour calculer des estimateurs de variance convergents par rapport au plan de sondage (Li, Graubard, Huang et Gastwirth, 2015; Li et Graubard, 2012). En supposant l'indépendance entre le fait d'être échantillonné dans l'enquête de référence et la participation à l'échantillon non probabiliste, la variance de $\bar{y}$ peut être estimée par (Korn et Graubard, 1999)

$$\text{var}_{\text{TL}}(\bar{y}) \cong \text{var}\left(\sum_{j \in C} z_j\right) + \text{var}\left(\sum_{i \in S} z_i\right), \tag{4.4}$$

où z_j (ou z_i) est l'écart de Taylor pour la j^{e} (ou i^{e}) unité dans C (ou dans S) que nous avons calculé en prenant la dérivée de $\bar{y}$ par rapport au poids de sondage (Shah, 2004). Par exemple, quand $\hat{w}_j = \hat{w}_j^{\text{PLA}}$, l'écart de Taylor pour l'unité $j \in C$ est

$$z_j = \frac{\partial}{\partial w_j} \bar{y} = \frac{\hat{w}_j (y_j - \bar{y})}{\sum_{l \in C} \hat{w}_l} + \frac{\sum_{l \in C} (y_l - \bar{y})}{\sum_{l \in C} \hat{w}_l} \left(\frac{\partial}{\partial w_j} \hat{w}_l\right)$$

et

$$\frac{\partial}{\partial w_j}\hat{w}_l = \left(\frac{\partial}{\partial \hat{\theta}}\hat{w}_l\right)\left(\frac{\partial}{\partial w_j}\hat{\theta}\right) = -\hat{w}_l x_l \left(\frac{\partial}{\partial w_j}\hat{\theta}\right),$$

où $\hat{\theta}$ désigne les paramètres estimés du modèle, qui peuvent être $\hat{B}_0$, $\hat{B}_w$, ou $\hat{B}_0 + \hat{\gamma}_w$, par exemple

$$\frac{\partial}{\partial w_j}(\hat{B}_0 + \hat{\gamma}_w) = (1-\hat{p}_j^*)\, x_j \left(\sum_{j' \in C \cup S} \hat{p}_{j'}^*(1-\hat{p}_{j'}^*)\, x_{j'}\, x_{j'}^T\right)^{-1},$$

où $\hat{p}_j^*$ pour $j \in C$ est le score de propension estimé pour l'unité j dans le modèle (4.1).

Pour l'unité $i \in S$, l'écart de Taylor est

$$z_i = \frac{\sum_{j \in C}(y_j - \bar{y})}{\sum_{j \in C}\hat{w}_j}\left(\frac{\partial}{\partial w_i}\hat{w}_j\right),$$

et

$$\frac{\partial}{\partial w_i}\hat{w}_j = \frac{\partial}{\partial \hat{\theta}}\hat{w}_j \frac{\partial}{\partial w_i}\hat{\theta} = -\hat{w}_j x_j \left(\frac{\partial}{\partial w_i}\hat{\theta}\right),$$

où $\hat{\theta}$ peut être $\hat{B}_0$, $\hat{B}_w$, ou $\hat{B}_0 + \hat{\gamma}_w$ par exemple

$$\begin{aligned}\frac{\partial}{\partial w_i}(\hat{B}_0 + \hat{\gamma}_w) \quad &= -\hat{p}_i^* x_i \left(\sum_{j' \in C \cup S} \hat{p}_{j'}^*(1-\hat{p}_{j'}^*)\, x_{j'} x_{j'}^T\right)^{-1} \\ &\quad + \left(1-\hat{q}_i - \hat{q}_i w_i\right) x_i \left(\sum_{j' \in S}(1+w_{j'})\,\hat{q}_{j'}\,(1-\hat{q}_{j'})\, x_{j'} x_{j'}^T\right)^{-1},\end{aligned}$$

où $\hat{q}_i$ pour $i \in S$ est le score de propension estimé pour l'unité i dans le modèle (4.2). L'écart de Taylor pour chaque unité mesure la variation de l'estimateur non linéaire, $\bar{y}$ dans notre cas, comme si l'unité avait été supprimée de l'échantillon. L'estimateur de la variance par linéarisation en séries de Taylor de $\bar{y}$ est ensuite calculé approximativement par (4.4), où $\text{var}\left(\sum_{i \in S} z_i\right)$ prend en compte la variabilité attribuable à l'échantillonnage complexe de S. D'après les conclusions de Wang et coll. (2021), on peut prouver que $\bar{y}$ est convergent par rapport au plan de sondage et $\text{var}_{\text{TL}}(\bar{y}) = O\left(\frac{1}{n_c}\right) + O\left(\frac{1}{n_s}\right)$. Les sections 5 et 6 présentent les estimations par la propension logistique ajustée pour illustrer les hypothèses d'échangeabilité conditionnelles à différents scores d'équilibrage. De même, il est possible de dériver les estimateurs de la variance des estimations par pondération de noyau avec les scores d'équilibrage adaptatif, que nous donnerons dans un futur article.

5. Étude par simulations

5.1 Génération de la population

Des études par simulations sont menées pour évaluer les estimations par la PLA basées sur le score d'équilibrage ajusté $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$, ainsi que $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ et $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ à des fins de comparaison. Nous générons

une population finie (PF) de taille $N = 1\ 000\ 000$ avec trois covariables indépendantes $x_1,\ x_2,\ x_3,$ chacune suivant une loi normale standard $N(0,1).$ Le résultat binaire $\mathbf{Y}$ est généré avec la moyenne définie par

$$P(Y=1)=\frac{\exp(\beta_0+x_1\beta_{x_1}+x_2\beta_{x_2}+x_1x_2\beta_{x_1x_2})}{1+\exp(\beta_0+x_1\beta_{x_1}+x_2\beta_{x_2}+x_1x_2\beta_{x_1x_2})}, \tag{5.1}$$

où $\beta_y=(\beta_0,\beta_{x_1},\beta_{x_2},\beta_{x_1x_2})^T$ sont les paramètres du modèle de résultat spécifiés comme étant $\beta_0=-1,\ \beta_{x_1}=0{,}8,\ \beta_{x_2}=0{,}2,\ \beta_{x_1x_2}=0{,}5.$ La moyenne du résultat binaire est d'environ 30 %. Les résultats ont montré une tendance similaire quand $\beta_0=-2$ ou $-3;$ ils ne sont donc pas indiqués.

5.2 Sélection de l'échantillon probabiliste S

Nous sélectionnons un échantillon aléatoire probabiliste S de taille n_s avec remise à partir de la population finie en utilisant un échantillonnage avec probabilité proportionnelle à la taille (PTT), la mesure de la taille de la k^{e} personne de la PF (mos_k) étant définie par

$$\text{mos}_k=\exp\left[a\times(\alpha_0+x_{k1}\alpha_{x_1}+x_{k2}\alpha_{x_2}+x_{k3}\alpha_{x_3}++x_{k1}x_{k2}\alpha_{x_1x_2}+x_{k1}x_{k3}\alpha_{x_1x_3})\right] \tag{5.2}$$

de sorte que la probabilité d'inclusion soit

$$p(k\in S\mid x;U)=\frac{n_s\times\text{mos}_k}{\sum_{k\in U}\text{mos}_k},$$

et que le poids d'échantillon correspondant soit l'inverse de la probabilité d'inclusion, c'est-à-dire $w_k=\frac{\sum_{k\in U}\text{mos}_k}{n_s\times\text{mos}_k}.$ Nous spécifions $\left(\alpha_0,\alpha_{x_1},\alpha_{X_2},\alpha_{x_3},\alpha_{x_1x_2},\alpha_{x_1x_3}\right)=(-1,\ 0{,}5,\ 0,\ 0{,}5,\ 0,\ -0{,}2)$ et supposons $a=0{,}5,$ 1 ou 1,5 pour faire varier le coefficient de variation (CV) des poids de sondage dans S (désignés par w_s), correspondant respectivement à $\text{CV}(w_s)=0{,}38,$ 0,86 ou 1,5. Notons que les variables de sélection dans l'échantillonnage S sont x_1 et $x_3,$ et que l'échantillon probabiliste pondéré par $w_k,\ S,$ se rapproche de la PF.

5.3 Sélection de l'échantillon non probabiliste C

Le processus de sélection sous-jacent pour l'échantillonnage C est inconnu. Nous sélectionnons C de taille $n_c=2\ 500$ dans la PF au moyen de l'échantillonnage PPT avec $\text{mos}_k,$ donné par (5.2) et spécifié pour inclure trois scénarios : 1) un échantillon par quota qui a la même distribution conjointe de x_1 et x_2 que dans la PF, désigné par Quota.$x_1x_2;$ 2) un échantillon par quota qui a la même distribution de x_2 que dans la PF, désigné par Quota.$x_2;$ et 3) un échantillon de volontaires dont les distributions dans x_1 ou x_2 sont différentes de celles dans la PF, désigné par Volontaire. La variable x_3 n'est pas prédictive du résultat et, par conséquent, n'induit aucun biais dans l'estimation de la moyenne de la PF (Li, Irimata, He et Parker, 2022). Le tableau 5.1 résume les paramètres du modèle pour la génération de résultat dans (5.1), la sélection de l'échantillon probabiliste S et la sélection de trois échantillons non probabilistes dans (5.2). Nous changeons la taille de l'échantillon probabiliste $n_s=1\ 250,\ 2\ 500,\ 3\ 750$ et la taille de l'échantillon non probabiliste est fixée à $n_c=2\ 500.$ Les poids de sondage associés aux unités C sont masqués dans l'analyse.

Tableau 5.1
Spécifications des paramètres du modèle pour la génération de résultat, la sélection de l'échantillon probabiliste (*S*) et la sélection de l'échantillon non probabiliste (*C*).

Modèle	Ordonnée à l'origine	x_1	x_2	x_3	x_1x_2	x_1x_3
Résultat	-1	0,8	0,2	0	0,5	0
Sélection de l'échantillon S	-1	0,5	0	0,5	0	-0,2
Participation de l'échantillon C						
Quota. x_1x_2	-1	0	0	0,5	0	0
Quota. x_2	-1	0,5	0	0	0	-0,2
Volontaire	-1	0,5	0,5	0,5	0	-0,2

Les trois estimations par la propension logistique ajustée (4.3) fondées sur le score d'équilibrage adaptatif $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$, celui non pondéré $b(\mathbf{x};\hat{\mathbf{B}}_0)$ et celui pondéré $b(\mathbf{x};\hat{\mathbf{B}}_w)$ sont calculées pour chacune des $R=$ 1 000 exécutions de simulation et évaluées comme suit :

- Biais relatif (RelBias%) = Biais (= moyenne des R moyennes simulées − moyenne de la population) divisé par la moyenne de la population $\times 100\,\%$.
- Variance empirique (VE) = Variance des R moyennes simulées $\times 10^4$.
- Ratio de variance (RV) = variance LT/variance empirique.

Pour construire les scores d'équilibrage estimés, la fonction $g(\mathbf{x}_i)$ dans (3.1) à (4.2) comprend non seulement les effets principaux de x_1, x_2, x_3, mais aussi leurs effets d'interaction par paires. On s'attend à ce que les estimations par la PLA fondées sur $b(\mathbf{x};\hat{\mathbf{B}}_w)$ soient approximativement sans biais, mais avec une variance gonflée en raison des poids différentiels; les estimations par la PLA fondées sur $b(\mathbf{x};\hat{\mathbf{B}}_0)$ ont la plus petite variance, mais elles peuvent être biaisées. En revanche, on s'attend à ce que les estimations par la PLA fondées sur le score d'équilibrage adaptatif $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ soient approximativement sans biais et comportent une plus petite variance selon les vrais modèles de propension.

5.4 Résultats

Le tableau 5.2 présente le biais relatif (%) des estimations par la PLA basées sur les trois scores d'équilibrage au moyen d'échantillons non probabilistes de Quota. x_1x_2, Quota. x_2 et Volontaire. À des fins de comparaison, nous incluons également les estimations non pondérées. Nous pouvons faire trois observations : 1) Comme on pouvait s'y attendre, les estimations non pondérées sont sans biais pour Quota. x_1x_2, mais fortement biaisées pour les échantillons Quota. x_2 et Volontaire. Ce résultat est conforme aux conclusions de Li et coll. (2022). 2) Pour corriger cela, les scores d'équilibrage de $b(\mathbf{x};\hat{\mathbf{B}}_w)$ ou $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ apparient l'échantillon Quota. x_2 ou Volontaire avec la distribution conjointe de x_1 et x_2 dans la PF et, par conséquent, produisent des estimations approximativement sans biais pour les trois échantillons non probabilistes. 3) En revanche, score non pondéré $b(\mathbf{x};\hat{\mathbf{B}}_0)$ donne des estimations biaisées puisqu'il ne s'agit pas d'une fonction monotone ou plus fine de la propension à la participation estimée des trois échantillons non probabilistes.

Tableau 5.2
Biais relatif (%) des estimations par la propension logistique ajustée de la moyenne de la population ($\overline{Y} = 0{,}3$) avec les tailles d'échantillon probabiliste et non probabiliste $n_s = n_c = 2\,500$ et $\mathrm{CV}(w_s) = 0{,}86$.

	Quota. $x_1 x_2$ $\mathrm{CV}(w_c) = 0{,}53$	Quota. x_2 $\mathrm{CV}(w_c) = 0{,}6$	Volontaire $\mathrm{CV}(w_c) = 1{,}10$
Sans pondération	-1,33	24,33	33,67
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	-1,33	-1,33	-1,33
$b(\mathbf{x}; \hat{\mathbf{B}}_0)$	19,33	19,33	19,33
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	-1,33	-1,33	-1,33

Ensuite, nous comparons au tableau 5.3 les deux estimations PLA sans biais avec $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ et $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ pour ce qui est de leur efficacité quand on fait varier les coefficients de variation (CV) des poids de l'échantillon probabiliste $\mathrm{CV}(w_s) = 0{,}38$, 0,86 ou 1,50. Nous pouvons faire trois observations. Premièrement, quand $\mathrm{CV}(w_s)$ augmente, la variance augmente comme nous nous y attendions. Par exemple, quand nous utilisons $b(\mathbf{x}; \hat{\mathbf{B}}_w)$, la variance empirique augmente pour passer de 1,00 à 1,12 puis à 1,30 pour Quota. $x_1 x_2$. Deuxièmement, quand $\mathrm{CV}(w_s)$ augmente, le gain d'efficacité de $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ sur $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ augmente. Par exemple, la différence relative des deux variances empiriques passe de 1 % (= (1 − 0,99) / 1,00) à 4 % (= (1,12 − 1,07) / 1,12) à 12 % (= (1,3 − 1,14) / 1,30) quand $\mathrm{CV}(w_s)$ augmente pour passer de de 0,38 à 0,86 puis à 1,5 pour Quota. $x_1 x_2$. Troisièmement, si l'on compare les trois échantillons non probabilistes, le gain d'efficacité de $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ sur $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ est le plus grand pour Quota. $x_1 x_2$. Intuitivement, les pseudo-pondérations créées pour Quota. $x_1 x_2$ sont non informatives et ajoutent donc une variance supplémentaire en raison de l'estimation de $b(\mathbf{x}; \hat{\mathbf{B}}_w)$.

Tableau 5.3
Variance empirique ($\times 10^4$) de deux estimations par la propension logistique ajustée sans biais selon des coefficients variables de variation de poids d'échantillon probabiliste $\mathrm{CV}(w_s)$, $n_s = n_c = 2\,500$.

	Quota. $x_1 x_2$	Quota. x_2	Volontaire
	$\mathrm{CV}(w_s) = 0{,}38$		
Sans pondération	0,81	0,94	0,81
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,00	0,97	1,44
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	0,99	0,98	1,45
	$\mathrm{CV}(w_s) = 0{,}86$		
Sans pondération	0,85	0,90	0,99
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,12	1,00	1,62
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1,07	1,02	1,64
	$\mathrm{CV}(w_s) = 1{,}50$		
Sans pondération	0,85	0,90	0,99
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,30	1,11	1,72
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1,14	1,07	1,68

Le tableau 5.4 présente la variance empirique (VE) dans le panneau de gauche et le ratio des variances (RV) dans le panneau de droite pour les estimations par la PLA quand nous varions la taille des échantillons probabilistes (n_s = 1 250; 2 500; 3 750) ayant une taille d'échantillon non probabiliste fixe n_c = 2 500. Nous pouvons faire trois observations. Premièrement, la variance empirique diminue quand n_s augmente : par exemple, la VE des estimations par la PLA avec $b(\mathbf{x}; \hat{B}_w)$ pour Quota. $x_1 x_2$ diminue pour passer de 1,33 à 1,08 puis à 0,99. Toutefois, la différence diminue, ce qui signifie une réduction de la VE plus grande de 0,25 (= 1,33 − 1,08) – quand n_s augmente pour passer de 1 250 à 2 500, comparativement à une baisse modérée de 0,09 (= 1,08 − 0,99) quand n_s augmente pour passer de 2 500 à 3 750. Ce résultat est attribuable au fait que $\mathrm{Var}(\bar{y}) = O\left(\frac{1}{n_c}\right) + O\left(\frac{1}{n_s}\right)$ est dominé par $O\left(\frac{1}{n_c}\right)$ quand $n_s > n_c$ et, par conséquent, le gain d'efficacité est modéré si n_s est augmenté une fois que $n_s > n_c$. Deuxièmement, si nous comparons les deux scores d'équilibrage, $b(\mathbf{x}; \hat{B}_0, \hat{\gamma}_w^T)$ est plus efficace que $b(\mathbf{x}; \hat{B}_w)$ quand n_s est petit. Intuitivement, quand $n_s < n_c$, la variance des estimations par la PLA est dominée par l'échantillon probabiliste S, qui a des poids de sondage différentiels w_s et induit donc une grande variabilité lors de l'estimation de $\hat{B}_w$. Cela se produit particulièrement pour les échantillons par quota où les poids de sondage utilisés pour l'estimation de $b(\mathbf{x}; \hat{B}_w)$ sont approximativement non informatifs et ajoutent par conséquent une variance supplémentaire. Troisièmement, l'estimateur de la variance par linéarisation en séries de Taylor (LT) proposé obtient généralement de bons résultats, le ratio des variances se rapprochant de 1 (voir la partie de droite dans le tableau 5.4). Toutefois, la variance LT fondée sur $b(\mathbf{x}; \hat{B}_0, \hat{\gamma}_w^T)$, surestime la variance pour Quota. $x_1 x_2$ quand n_s est petit. On constate que le RV pour $b(\mathbf{x}; \hat{B}_0, \hat{\gamma}_w^T)$ est plus proche de 1 quand n_s augmente ou quand $\mathrm{CV}(w_s)$ est petit (résultats non indiqués).

Tableau 5.4
Variance empirique ($\times 10^4$) et ratio des variances de deux estimations par la propension logistique ajustée sans biais avec des tailles d'échantillon probabiliste variables de n_s, $\mathrm{CV}(w_s)$ = 0,86 et n_c = 2 500.

	Variance empirique (VE)			Ratio des variances (RV)		
	Quota. $x_1 x_2$	Quota. x_2	Volontaire	Quota. $x_1 x_2$	Quota. x_2	Volontaire
Sans pondération	0,81	0,87	1,03	1,02	0,95	0,86
			n_s = 1 250			
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,33	1,18	1,77	1,02	0,98	0,96
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1,17	1,08	1,80	1,41	1,35	1,14
			n_s = 2 500			
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,08	0,98	1,65	1,06	1,01	0,93
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1,00	0,96	1,69	1,31	1,23	1,03
			n_s = 3 750			
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	0,99	0,94	1,60	1,08	1,00	0,93
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	0,95	0,94	1,63	1,26	1,15	1,00

En résumé, au moyen d'études par simulations, on a observé que des estimations par la PLA fondées sur $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ et $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ sont approximativement sans biais et d'une efficacité comparable quand l'échantillon probabiliste de référence a une grande taille d'échantillon n_s ou des poids de sondage stables

avec un petit $\mathrm{CV}(w_s)$. En revanche, quand l'échantillon probabiliste de référence a un petit n_s ou des poids de sondage variables, $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ a tendance à produire des estimations plus efficaces, en particulier pour les échantillons par quota. Les spécialistes des enquêtes doivent choisir une enquête de référence ayant une taille suffisamment grande et des poids de sondage stables, car le gain d'efficacité obtenu en augmentant n_s est modéré une fois que $n_s > n_c$.

6. Analyse des données de séropositivité au SARS-CoV-2 des National Institutes of Health

Le principal objectif de l'étude sur la séropositivité au SARS-CoV-2 est d'estimer la prévalence de la séropositivité aux anticorps du virus SARS-CoV-2 dans la population cible composée d'adultes de 18 ans et plus vivant aux États-Unis qui n'ont pas reçu de diagnostic de COVID-19 pendant la première phase de la pandémie, d'avril à août 2020. Dans les semaines après l'annonce de l'étude, plus de 460 000 personnes se sont portées volontaires. L'étude ne pouvait toutefois se permettre qu'un sous-ensemble de ces volontaires. Un échantillon par quota a été sélectionné en fonction de 6 variables de quota, à savoir le groupe d'âge, la race, le sexe, l'origine ethnique, la densité de la population et la région géographique, pour qu'il corresponde approximativement à la répartition de ces variables chez les adultes des États-Unis. Quelque 8 058 participants ont répondu au questionnaire sur les facteurs cliniques et fourni des échantillons de sang servant à évaluer la séropositivité. L'échantillon prélevé dans le cadre de l'étude sur la séropositivité au SRAS-CoV-2 a été appelé « échantillon de la COVID ». Bien que l'échantillon de la COVID ait été un échantillon aléatoire ayant des probabilités de sélection connues tiré du bassin de volontaires de Kalish, Klumpp-Thomas, Hunsberger, Baus, Fay, Siripong, Wang, Hicks, Mehalko, Travers, Drew, Pauly, Spathies, Ngo, Adusei, Karkanitsa, Croker, Li, Graubard, Czajkowski, Belliveau, Chairez, Snead, Frank, Shunmugavel, Han, Giurgea, Rosas, Bean, Athota, Cervantes-Medina, Gouzoulis, Heffelfinger, Valenti, Caldararo, Kolberg, Kelly, Simon, Shafiq, Wall, Reed, Ford, Lokwani, Denson, Messing, Michael, Gillette, Kimberly, Reis, Hall, Esposito, Memoli et Sadtler (2021), ce bassin de volontaires est un échantillon non aléatoire de la population cible aux États-Unis et peut présenter un biais de sélection élevé.

Pour aider à corriger le biais de sélection, nous utilisons le Behavioral Risk Factor Surveillance System (BRFSS ou Système de surveillance des facteurs de risque comportementaux, Centers for Disease Control and Prevention, 2022) comme enquête de référence. Le BRFSS se compose d'enquêtes annuelles à l'échelle des États américains, qui sont combinées en une enquête représentative nationale comportant des observations à grande échelle au niveau de l'État. En plus des 6 variables de quota, 10 variables démographiques et liées à la santé sont recueillies dans le BRFSS, lesquelles sont également prédictives de la séropositivité, mais ne sont pas utilisées dans l'échantillonnage par quota. Après avoir supprimé les observations ayant des valeurs manquantes pour une ou plusieurs des 16 variables, n_s = 367 165 participants au total ont été inclus dans l'analyse. Le CV des poids de sondage du BRFSS est $\mathrm{CV}(w_s)$ = 1,92.

Le tableau 6.1 montre la distribution pondérée de l'échantillon pour les 16 variables du BRFSS et l'échantillon de la COVID. Comme on s'y attendait, les distributions des 6 variables de quota dans les deux échantillons sont très proches. Pour les 10 variables démographiques et liées à la santé, la plupart des distributions diffèrent considérablement entre les deux échantillons. En général, les participants de l'échantillon de la COVID ont tendance à être plus scolarisés, propriétaires de leur logement, employés et en meilleure santé. À titre d'exemple, 84 % des participants de l'échantillon de la COVID, comparativement à 29 % pour le BRFSS pondéré, possèdent un diplôme d'études collégiales ou de niveau supérieur. Par conséquent, un biais de sélection existe dans l'échantillon de la COVID, et notre objectif est de réduire le biais de sélection dans l'estimation de la séropositivité non diagnostiquée au SRAS-CoV-2.

Le tableau 6.2 montre les estimations par la PLA de la prévalence de la séropositivité non diagnostiquée fondées sur les trois scores d'équilibrage. Comme nous l'avons indiqué, l'estimation par la PLA fondée sur $b(\mathbf{x};\hat{\mathbf{B}}_w)$ a permis de détecter un taux de séropositivité de 4,65 %, proche du taux de 4,67 % détecté au moyen de $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$. Les deux erreurs-types correspondantes sont également proches (0,78 contre 0,77). En revanche, le $b(\mathbf{x};\hat{\mathbf{B}}_0)$ non pondéré donne un taux de séropositivité de 3,95 %, proche de la moyenne non pondérée de 3,77 %, les deux présentant un biais de sélection. Il est intéressant de mentionner que le score d'équilibrage adaptatif $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ a produit des pseudo-pondérations stables pour l'échantillon de la COVID avec $\mathrm{CV}(\hat{w}_c) = 2{,}24$, proche du 2,25 du $b(\mathbf{x};\hat{\mathbf{B}}_0)$ non pondéré, et que les deux sont inférieurs à $\mathrm{CV}(\hat{w}_c) = 2{,}33$ produit par le $b(\mathbf{x};\hat{\mathbf{B}}_w)$ pondéré.

Tableau 6.1
Distribution des covariables (%) dans l'échantillon de la COVID par rapport à celles du Behavioral Risk Factor Surveillance System (BRFSS).

	Échantillon de la COVID-19	BRFSS pondéré		Échantillon de la COVID-19	BRFSS pondéré		Échantillon de la COVID-19	BRFSS pondéré
Groupe d'âge			Milieu urbain/rural			Vacciné contre la grippe		
18 à 44 ans	41,6	42,9	Urbain	94,7	93,2	Oui	73,8	51,3
45 à 69 ans	42,6	41,8	Rural	5,3	6,8	Non	26,2	48,7
70 à 95 ans	15,8	15,2	Présence d'enfants			Maladie cardiovasculaire		
Sexe			Oui	32,5	34,7	Oui	4,1	9,5
Homme	47,4	47,8	Non	67,5	65,3	Non	95,9	90,5
Femme	52,6	52,2	Scolarité			Maladie pulmonaire		
Race			<= Études secondaires	2,6	39,4	Oui	18,8	18,7
Blanc seulement	77,5	74,8	Études collégiales	13,8	31,5	Non	81,2	81,3
Noir seulement	9,4	12,6	>= Études collégiales	83,6	29,1	Immunisé		
Autres	13,1	12,5	Propriétaire			Oui	23,4	31,1
Origine ethnique			Propriétaire	75,2	68,8	Non	76,6	68,9
Hispanique	15,9	14,1	Locataire	20,2	25,6	Diabète		
Non hispanique	84,1	85,9	Autre	4,7	5,6	Oui	5,5	11,9
Région			Emploi			Non	94,5	88,1
Nord-est	16,7	17,1	A un emploi	71,2	57,4	Assurance maladie		
Midwest	15,8	17,6	Inactif	23,8	32,2	Oui	97,4	89,0
Centre de l'Atlantique	20,8	17,3	Sans emploi	5,0	10,4	Non	2,6	11,0
Sud/Centre	14,2	15,7						
Montagnes/Sud-ouest	15,5	15,3						
Ouest/Pacifique	17,0	16,9						

Tableau 6.2
Taux de séropositivité non diagnostiquée chez les adultes des États-Unis du 1er avril au 4 août 2020.

	$CV(\hat{w}_c)$	Estimations (%)	Erreur-type* ($\times 10^{-2}$)
Sans pondération	0,00	3,77	0,22
$b(\mathbf{x};\hat{\mathbf{B}}_0)$	2,25	3,94	0,52
$b(\mathbf{x};\hat{\mathbf{B}}_w)$	2,33	4,65	0,78
$b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$	2,24	4,67	0,77

* : pour tenir compte de la variabilité attribuable à l'estimation de $\mathbf{B}$, $\mathbf{B}_0$ ou $\boldsymbol{\gamma}$.

7. Conclusion et discussion

Dans le présent article, nous avons examiné l'échangeabilité du résultat conditionnelle au $b(\mathbf{x};\hat{\mathbf{B}}_w)$ pondéré et au $b(\mathbf{x};\hat{\mathbf{B}}_0)$ non pondéré sont utilisés dans les méthodes existantes de pondération et d'appariement fondées sur le score de propension pour les inférences d'échantillons non probabilistes. Nous proposons un score d'équilibrage adaptatif $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ pour corriger le biais potentiel dans $b(\mathbf{x};\hat{\mathbf{B}}_0)$ en trois étapes : 1) estimer le score d'équilibrage non pondéré $b(\mathbf{x};\hat{\mathbf{B}}_0)$; 2) estimer le facteur de correction du biais $b(\mathbf{x};\hat{\gamma}_w)$; 3) construire $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)=b(\mathbf{x};\hat{\mathbf{B}}_0)+b(\mathbf{x};\hat{\boldsymbol{\gamma}}_w)$, qui est une fonction monotone de la propension à la participation estimée.

Le critère de base pour choisir le score d'équilibrage est qu'il doit être plus fin que, sinon égal à la propension à la participation afin d'équilibrer la distribution de x entre l'échantillon non probabiliste et la population finie. $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ et $b(\mathbf{x};\hat{\mathbf{B}}_w)$ produisent tous deux des estimations sans biais d'une efficacité comparable, $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ étant plus efficace pour les échantillons par quota quand l'enquête de référence est petite ou a des poids d'échantillon variables. Les spécialistes des enquêtes doivent choisir comme enquête de référence un échantillon suffisamment grand ayant des poids d'échantillon stables. Notons que le gain d'efficacité obtenu en augmentant la taille de l'échantillon probabiliste n_s est modéré une fois que $n_s > n_c$.

Nous avons cerné deux limites : 1) le score d'équilibrage adaptatif est construit en supposant l'exactitude du modèle de propension par régression logistique aux étapes 1 et 2 pour obtenir les scores d'équilibrage non pondérés et le facteur de correction du biais; 2) dans les deux étapes, la régression logistique est supposée avoir la même forme fonctionnelle. En conséquence, nous proposons les deux prolongements suivants de l'article dans de futurs travaux de recherche : 1) Permettre une forme fonctionnelle différente à l'étape 2, où nous modélisons la probabilité pour la sélection de l'échantillon de référence, à partir de la forme fonctionnelle supposée à l'étape 1. À l'aide de variables de sélection connues et de la probabilité de sélection pour chaque unité de l'enquête de référence, il est possible de mettre en œuvre des diagnostics de modèle comme une courbe ROC (courbe caractéristique de la performance d'un test) pour faciliter la sélection du modèle. 2) Construire plusieurs modèles de propension. Un modèle de régression logistique a été adapté pour estimer les scores de propension aux étapes 1 et 2 de la section 4. Toutefois, la spécification erronée du modèle de régression logistique pourrait donner des scores de propension mal estimés qui vont à l'encontre de l'hypothèse (2.1) et, par conséquent, donnent des estimations biaisées. Les méthodes non

paramétriques, comme les méthodes d'apprentissage automatique, peuvent fournir des solutions de rechange, qui assouplissent les spécifications du modèle paramétrique supposé concernant la sélection des variables, la forme fonctionnelle et la sélection des termes des polynômes et des interactions multidirectionnelles spécifiées dans la modélisation paramétrique.

Dans l'article, nous avons discuté de la façon de construire des scores d'équilibrage qui satisfont à l'hypothèse de l'EC de sorte que la distribution du résultat soit échangeable entre l'échantillon non probabiliste et la population finie. Notons que le score d'équilibrage est une fonction des covariables observées $\mathbf{x}$ qui sont recueillies à la fois dans l'échantillon non probabiliste C et dans l'enquête de référence S. S'il manque des covariables importantes dans S ou C, alors quel que soit le score d'équilibrage choisi, les estimations moyennes de la PF sont inévitablement biaisées. Il reste des éléments importants à prendre en considération : Quelles variables doivent être recueillies dans C et S? Comment les questions de l'enquête seront-elles harmonisées dans la collecte des données de C et S? Et comment les erreurs de mesure ou de déclaration peuvent-elles être réduites dans la conception du questionnaire? Le mode de traitement de ces questions peut être essentiel pour satisfaire à l'hypothèse de l'EC dans les méthodes d'ajustement fondées sur le score de propension pour l'analyse des échantillons non probabilistes. En résumé, comme l'exige l'hypothèse de l'échangeabilité conditionnelle, il est important d'avoir des enquêtes de référence de grande qualité qui permettent de recueillir des ensembles complets de variables comportant un minimum d'erreurs de mesure et de déclaration, qui ont une taille d'échantillon suffisamment grande et qui sont bien conçues avec des poids d'échantillonnage informatifs et stables.

Remerciements

Nous remercions Barry Graubard pour ses discussions stimulantes et Lingxiao Wang pour son aide en matière de simulation. La présente étude est financée par la subvention NIH R03 CA252782.

Bibliographie

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. et Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1(2), 90-143.

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

Brick, J., et Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33, 735-752. DOI: https://doi.org/10.1515/jos-20170034.

Centers for Disease Control and Prevention (2022). Behavioral Risk Factor Surveillance System: Annual survey data. Atlanta, Géorgie: Centers for Disease Control and Prevention, US Department of Health and Human Services. Récupéré de http://www.cdc. gov/brfss/annual_data/annual_data.htm.

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.

Elliott, M. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2.

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

Groves, R., et Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, 72(2), 167-189. DOI: https://doi.org/10.1093/poq/nfn011.

Kalish, H., Klumpp-Thomas, C., Hunsberger, S., Baus, H.A., Fay, M.P., Siripong, N., Wang, J., Hicks, J., Mehalko, J., Travers, J., Drew, M., Pauly, K., Spathies, J., Ngo, T., Adusei, K.M., Karkanitsa, M., Croker, J.A., Li, Y., Graubard, B.I., Czajkowski, L., Belliveau, O., Chairez, C., Snead, K.R., Frank, P., Shunmugavel, A., Han, A., Giurgea, L.T., Rosas, L.A., Bean, R., Athota, R., Cervantes-Medina, A., Gouzoulis, M., Heffelfinger, B., Valenti, S., Caldararo, R., Kolberg, M.M., Kelly, A., Simon, R., Shafiq, S., Wall, V., Reed, S., Ford, E.W., Lokwani, R., Denson, J.-P., Messing, S., Michael, S.G., Gillette, W., Kimberly, R.P., Reis, S.E., Hall, M.D., Esposito, D., Memoli, M.J. et Sadtler, K. (2021). Undiagnosed SARS-CoV-2 seropositivity during the first six months of the COVID-19 pandemic in the United States. *Sci Transl Med*, 13(601), eabh3826.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K. et Gimenez, A. (2016). *Evaluating Online Nonprobability Surveys.* Washington, DC: Pew Research Center.

Kern, C., Li, Y. et Wang, L. (2021). Boosted kernel weighting – Using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, 9(5), 1088-1113.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys.* New York: John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/9781118032619.

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

Li, Y., et Graubard, B. (2012). Pseudo semiparametric maximum likelihood estimation exploiting gene environment independence for population-based case-control studies with complex samples. *Biostatistics*, 13, 711-723.

Li, Y., Graubard, B. et DiGaetano, R. (2011). Weighting methods for population-based case-control studies with complex sampling. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 60, 165-185.

Li, Y., Graubard, B., Huang, P. et Gastwirth, J. (2015). Extension of the Peters–Belson method to estimate health disparities among multiple groups using logistic regression with survey data. *Statistics in Medicine*, 34, 595-612.

Li, Y., Irimata, K.E., He, Y. et Parker, J. (2022). Variable inclusion strategies through directed acyclic graphs to adjust health surveys subject to selection bias for producing national estimates. *Journal of Official Statistics*, 38(3), 1-27.

Mercer, A.W., Kreuter, F., Keeter, S. et Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81, 250-271. DOI: https://doi.org/10.1093/poq/nfw060.

Pinsky, P.F., Miller, A., Kramer, B.S., Church, T., Reding, D., Prorok, P., Gelmann, E., Schoen, R.E., Buys, S., Hayes, R.B. et Berg, C.D. (2007). Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *American Journal of Epidemiology*, 165(8), 874-881.

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

Rivers, D. (2007). Sampling for web surveys. Document présenté aux *Joint Statistical Meetings - Section on Survey Research Methods*.

Rosenbaum, P., et Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Rubin, D. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6, 34-58.

Scott, A., et Wild, C. (2001). The analysis of clustered case-control studies. *Journal of the Royal Statistical Society Series C*, 50, 389-401.

Shah, B.V. (2004). Commentaires à propos de l'article "Estimateurs de variance par linéarisation pour des données d'enquête" par A. Demnati et J.N.K. Rao. *Techniques d'enquête*, 30, 1, 18. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2004001/article/6991-fra.pdf.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1).

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263.

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

Wang, L., Graubard, B.I., Katki, H.A. et Li, Y. (2020). Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *Journal of the Royal Statistical Society Series A*, 183, 1293-1311.

Wang, L., Graubard, B.I., Katki, H.A. et Li, Y. (2022). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. *Revue Internationale de Statistique*, 90, 146-164.

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250. DOI: https://doi.org/10.1002/sim.9122.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 67-74
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Hypothèse de l'échangeabilité dans des méthodes d'ajustement fondées sur le score de propension aux fins d'estimation de la moyenne de population au moyen d'échantillons non probabilistes »

Jae Kwang Kim et Yonghyun Kwon[1]

Résumé

La construction de pseudo-poids pour l'intégration des données peut être comprise dans le cadre de l'échantillonnage à deux phases. Au moyen du cadre d'échantillonnage à deux phases, nous abordons deux approches de l'estimation des scores de propension et mettons au point une nouvelle façon de construire la fonction de score de propension pour l'intégration des données en utilisant la méthode de maximum de vraisemblance conditionnelle. Les résultats d'une étude de simulation limitée sont aussi présentés.

Mots-clés : Intégration des données; fonction de score de propension; pseudo-poids; échantillonnage à deux phases.

1. Introduction

Nous souhaitons féliciter Yan Li d'avoir été sélectionné à titre de conférencier à la Conférence Morris Hansen et d'avoir donné une présentation intéressante sur l'intégration des données. L'intégration des données est un domaine de recherche émergent pour combiner de multiples sources de données de façon défendable. Dans l'intégration des données, en utilisant un échantillon probabiliste indépendant comme échantillon de calage, il est possible de réduire le biais de sélection dans l'échantillon de commodité. Cependant, les outils statistiques pour l'intégration des données sont limités. C'est pourquoi je salue la tentative de Li de mettre au point un outil statistique supplémentaire pour l'intégration des données.

Le recours à la fonction de score d'équilibrage pour contrôler le biais de sélection dans l'échantillon non probabiliste est une assez bonne idée. La façon de construire la fonction de score d'équilibrage dans le contexte de l'intégration des données peut être plus compliquée. Li a reconnu que la méthode d'estimation du score de propension (SP) de Chen, Li et Wu (2020) peut être inefficace, car la procédure d'estimation exige d'utiliser les poids d'enquête dans l'échantillon probabiliste. Au lieu d'utiliser l'estimation pondérée, Li a proposé une méthode d'estimation non pondérée et a ensuite mis au point une méthodologie pour la correction des biais. L'estimation non pondérée du SP est aussi envisagée par Elliott et Valliant (2017) et a été adoptée par certains spécialistes. Dans le présent article, nous aimerions clarifier deux approches existantes de l'estimation des scores de propension et trouver une façon défendable de construire la fonction de score de propension pour l'intégration des données.

L'article est structuré de la façon suivante. Dans la section 2, nous présentons un cadre d'échantillonnage à deux phases pour l'intégration des données et l'approche selon un modèle de SP conditionnel est présentée.

1. Jae Kwang Kim et Yonghyun Kwon, Department of Statistics, Iowa State University, Ames (Iowa) 50011, États-Unis. Courriel : jkim@iastate.edu.

La section 3 présente une autre approche, appelée l'approche selon un modèle non conditionnel. L'étude par simulation est présentée à la section 4. Des conclusions sont présentées à la section 5.

2. Approche selon un modèle de SP conditionnel

Nous utilisons la configuration prise en considération dans Yang, Kim et Hwang (2021) où l'échantillon A est un échantillon probabiliste observant $\mathbf{x}$ et l'échantillon B est l'échantillon non probabiliste observant $(\mathbf{x}, y)$. Le tableau 2.1 présente la configuration générale des deux structures d'échantillon pour l'intégration des données. Comme indiqué au tableau 2.1, l'échantillon B n'est pas représentatif de la population cible.

Tableau 2.1
Structure de données pour l'intégration des données et la fusion des données.

Intégration des données				
Échantillon	**Type**	***X***	***Y***	**Représentatif ?**
A	Échantillon probabiliste	✓		Oui
B	Échantillon non probabiliste	✓	✓	Non

La formulation est assez semblable à l'échantillonnage à deux phases :

1. L'échantillon de première phase $S_1 \equiv A \cup B$ est sélectionné à partir de U et $\mathbf{x}_i$ est observé pour toutes les unités dans l'échantillon S_1.
2. L'échantillonnage de deuxième phase $S_2 = B$ est sélectionné à partir de S_1 et y_i est observé pour toutes les unités dans l'échantillon S_2.

Contrairement à l'échantillonnage à deux phases classique, nous ne connaissons pas la probabilité d'inclusion de premier ordre de S_1. Nous connaissons seulement la probabilité d'inclusion de premier ordre de l'échantillon A. En d'autres termes, $\pi_i^{(A)} = P(i \in A \mid i \in U)$ est la probabilité d'inclusion de premier ordre (connue) de l'échantillon A.

Supposons que $\pi_i^{(B)} = P(i \in B \mid i \in U)$ est la probabilité d'inclusion de premier ordre (inconnue) de l'échantillon B. Notez que la probabilité d'inclusion de premier ordre de S_1 peut être représentée par

$$\begin{aligned} P(i \in S_1 \mid i \in U) &= P(i \in A \cup B \mid i \in U) \\ &= P(i \in A \mid i \in U) + P(i \in B \mid i \in U) - P(i \in A \mid i \in U)P(i \in B \mid i \in U) \\ &= \pi_i^{(A)} + \pi_i^{(B)} - \pi_i^{(A)}\pi_i^{(B)} \end{aligned} \tag{2.1}$$

où la dernière égalité découle de l'indépendance des deux échantillons. Nous pouvons donc exprimer la probabilité d'inclusion conditionnelle pour l'échantillon de deuxième phase par

$$P(i \in S_2 \mid i \in S_1) = \frac{P(i \in B \mid i \in U)}{P(i \in A \cup B \mid i \in U)} = \frac{\pi_i^{(B)}}{\pi_i^{(A)} + \pi_i^{(B)} - \pi_i^{(A)}\pi_i^{(B)}}. \tag{2.2}$$

Maintenant, puisque nous observons $\mathbf{x}_i$ pour $i \in S_1 = A \cup B$, nous pouvons faire un modèle statistique pour la probabilité d'inclusion conditionnelle dans l'équation (2.2) en tant que fonction de $\mathbf{x}$. Supposons que

$$P(i \in S_2 \mid i \in S_1) = p(\mathbf{x}_i;\phi) \tag{2.3}$$

est le modèle statistique pour la probabilité d'inclusion conditionnelle avec le paramètre inconnu ϕ. Nous pouvons estimer ϕ par une analyse non pondérée. Autrement dit :

$$\hat{\phi} = \arg\max_{\phi} \sum_{i \in S_1} \left[\delta_i \log p(\mathbf{x}_i;\phi) + (1-\delta_i) \log\{1 - p(\mathbf{x}_i;\phi)\} \right],$$

où $\delta_i = \mathbb{I}(i \in B)$ est la fonction indicatrice de l'événement $i \in B$. Si un modèle de régression logistique avec $\text{logit}\{p(\mathbf{x}_i;\phi)\} = \mathbf{x}_i'\phi$ est utilisé dans l'équation (2.3), alors $\hat{\phi}$ peut être obtenu en trouvant la solution à

$$\sum_{i \in B} \{1 - p(\mathbf{x}_i;\phi)\}\mathbf{x}_i - \sum_{i \in A} p(\mathbf{x}_i;\phi)\mathbf{x}_i = \mathbf{0}.$$

Cette estimation non pondérée est pleinement justifiée, car le modèle de probabilité d'inclusion conditionnelle (2.3) est conditionnel à l'échantillon de première phase $S_1 = A \cup B$. Puisque le modèle de propension dans l'équation (2.3) est conditionnel à l'échantillon de première phase, on peut l'appeler le modèle de score de propension (SP) conditionnel.

Maintenant, puisque l'équation (2.3) est le modèle pour la probabilité d'inclusion conditionnelle dans l'équation (2.2), nous pouvons obtenir

$$\frac{\pi_i^{(B)}}{\pi_i^{(A)} + \pi_i^{(B)} - \pi_i^{(A)}\pi_i^{(B)}} = p(\mathbf{x}_i;\hat{\phi}),$$

ce qui suppose que

$$\frac{1}{\hat{\pi}_i^{(B)}} = 1 + \frac{1}{\pi_i^{(A)}} \left\{ \frac{1}{p(\mathbf{x}_i;\hat{\phi})} - 1 \right\}. \tag{2.4}$$

Donc, $\hat{w}_i^{(B)} = 1/\hat{\pi}_i^{(B)}$ dans l'équation (2.4) peut être utilisé comme pseudo-poids final pour les éléments faisant partie de l'échantillon B.

En pratique, nous ne pouvons pas utiliser l'équation (2.4) directement, car les probabilités d'inclusion de premier ordre sont inconnues en dehors de l'échantillon. Une façon de résoudre ce problème est d'estimer $w_i^{(A)} = 1/\pi_i^{(A)}$ à l'aide de

$$\tilde{w}_i^{(A)} = E\{w_i^{(A)} \mid \mathbf{x}_i, I_i^{(A)} = 1\} \tag{2.5}$$

suivant le résultat de Pfeffermann et Sverchkov (1999). Ainsi, l'équation (2.4) peut être remplacée par

$$\frac{1}{\hat{\pi}_i^{(B)}} = 1 + \tilde{w}_i^{(A)} \left\{ \frac{1}{p(\mathbf{x}_i;\hat{\phi})} - 1 \right\}. \tag{2.6}$$

Li a utilisé un modèle paramétrique pour $E(\pi^{(A)} \mid \mathbf{x}) = \overline{\pi}^{(A)}(\mathbf{x};\gamma)$ et mis au point une méthode de pseudo-maximum de vraisemblance pour estimer γ à partir de l'échantillon. Une fois $\hat{\gamma}$ obtenu, nous pouvons utiliser l'équation (2.6) avec $\tilde{w}_i^{(A)} = 1/\tilde{\pi}(\mathbf{x}_i;\hat{\gamma})$.

Au lieu d'utiliser l'équation (2.6), Elliott et Valliant (2017) ont proposé d'utiliser

$$\frac{1}{\hat{\pi}_i^{(B)}} = \frac{1}{\hat{\pi}_i^{(A)}}\left\{\frac{1}{p(\mathbf{x}_i;\hat{\phi})} - 1\right\} \tag{2.7}$$

où

$$\hat{\pi}_i^{(A)} = E\{\pi_i^{(A)} \mid \mathbf{x}_i, I_i^{(A)} = 1\}. \tag{2.8}$$

Cependant, $\tilde{w}_i^{(A)} \neq 1/\hat{\pi}_i^{(A)}$ en général et le pseudo-poids dans l'équation (2.7) n'est pas justifié sur le plan théorique.

3. Approche selon un modèle de SP non conditionnel

Une autre approche du modèle de SP est de présumer un modèle statistique pour $\pi_i^{(B)} = P(i \in B \mid i \in U)$ comme

$$\pi_i^{(B)} = \pi_B(\mathbf{x}_i;\phi) \tag{3.1}$$

pour un certain paramètre ϕ. Ce modèle de SP non conditionnel a été utilisé par Chen et coll. (2020) et Wang, Valliant et Li (2021), où la méthode de pseudo-maximum de vraisemblance a été utilisée pour estimer ϕ.

Si nous voulons améliorer l'efficacité des estimateurs de ϕ, nous pouvons considérer la méthode du maximum de vraisemblance comme suit. Premièrement, si $\pi_i^{(A)}$ sont disponibles dans S_1, en utilisant l'équation (3.1), nous pouvons dériver le modèle de probabilité d'inclusion conditionnelle suivant :

$$\pi_{2i|1}(\phi) = \frac{\pi_B(\mathbf{x}_i;\phi)}{\pi_i^{(A)} + \pi_B(\mathbf{x}_i;\phi) - \pi_i^{(A)} \cdot \pi_B(\mathbf{x}_i;\phi)}. \tag{3.2}$$

Dans la deuxième étape, nous pouvons calculer l'estimateur du maximum de vraisemblance conditionnelle de ϕ à partir de l'échantillon combiné par

$$\hat{\phi} = \arg\max_{\phi} \sum_{i \in S_1}\left[\delta_i \log \pi_{2i|1}(\phi) + (1-\delta_i)\log\{1-\pi_{2i|1}(\phi)\}\right], \tag{3.3}$$

où $\pi_{2i|1}(\phi)$ est défini dans l'équation (3.2). L'estimateur du maximum de vraisemblance conditionnelle dans l'équation (3.3) repose sur l'hypothèse selon laquelle nous pouvons déterminer les unités qui appartiennent à l'intersection de A et B. Une fois $\hat{\phi}$ obtenu à partir de la méthode du maximum de vraisemblance conditionnelle, nous pouvons utiliser $\hat{w}_i^{(B)} = 1/\pi^{(B)}(\mathbf{x}_i;\hat{\phi})$ comme pseudo-poids pour l'échantillon B. Cette méthode de maximum de vraisemblance conditionnelle a aussi été utilisée par Savitsky et coll. (2022) en posant l'hypothèse selon laquelle $\pi_i^{(A)}$ sont disponibles dans l'échantillon B.

Si $\pi_i^{(A)}$ ne sont pas disponibles en dehors de l'échantillon A, nous ne pouvons pas construire la probabilité d'inclusion conditionnelle dans l'équation (3.2). Dans ce cas, nous pouvons remplacer $\pi_i^{(A)}$ par $\tilde{\pi}_i^{(A)} = 1/\tilde{w}_i^{(A)}$, où $\tilde{w}_i^{(A)}$ est défini dans l'équation (2.5), et calculer

$$\pi_{2i|1}(\phi) = \frac{\pi_B(\mathbf{x}_i;\phi)}{\tilde{\pi}_i^{(A)} + \pi_B(\mathbf{x}_i;\phi) - \tilde{\pi}_i^{(A)} \cdot \pi_B(\mathbf{x}_i;\phi)} \tag{3.4}$$

pour appliquer la méthode de maximum de vraisemblance conditionnelle ci-dessus dans l'équation (3.3). Les pseudo-poids finaux sont donnés par $\hat{w}_i^{(B)} = 1/\pi_B(\mathbf{x}_i;\hat{\phi})$ et $\hat{\phi}$ est calculé par l'équation (3.3).

Au lieu de la méthode de maximum de vraisemblance, les pseudo-poids pour l'échantillon B peuvent être construits pour satisfaire

$$\sum_{i\in B} \frac{1}{\pi_B(\mathbf{x}_i;\phi)} \mathbf{x}_i = \sum_{i\in A} \frac{1}{\pi_i^{(A)}} \mathbf{x}_i. \tag{3.5}$$

La condition (3.5) est souvent appelée la propriété de calage. La propriété de calage est une propriété souhaitable pour tout pseudo-poids. Une fois que $\hat{\phi}$ est calculé à partir de l'équation de calage dans l'équation (3.5), le pseudo-poids final pour l'échantillon B est donné par $\hat{w}_i^{(B)} = 1/\pi_B(\mathbf{x}_i;\hat{\phi})$.

4. Étude par simulation

Une étude par simulation limitée est menée pour comparer la performance des estimateurs, y compris les méthodes proposées dans l'article de Li. Dans la simulation, nous générons une population finie avec $y_i \sim \text{Bernoulli}(p_i)$, $p_i = \text{expit}(-1 + 0{,}8x_{1i} + 0{,}2x_{2i} + 0{,}5x_{1i}x_{2i})$ avec (x_1, x_2, x_3) obtenu par la distribution normale. La taille de la population finie est $N =$ 5 000.

À partir de la population finie, l'échantillon A est généré à répétition par l'échantillon de probabilité proportionnelle à la taille avec mesure de la taille

$$mos_i = \exp(-1 + 0{,}5x_{1i} + 0{,}5x_{3i} - 0{,}2x_{1i}x_{3i})$$

selon une taille d'échantillon $n_A = 250$. De plus, l'échantillon B est sélectionné à répétition par échantillonnage aléatoire stratifié avec deux strates, où la strate 1 est $U_1 = i \in U : x_{1i} > 0$ et la strate 2 est $U_2 = i \in U : x_{1i} \le 0$. Dans la strate 1, les échantillons $n_{B1} = 0{,}7n_B$ sont sélectionnés par échantillonnage aléatoire simple. Dans la strate 2, les échantillons $n_{B2} = 0{,}3n_B$ sont sélectionnés par échantillonnage aléatoire simple. La taille de l'échantillon B est choisie pour être $n_B =$ 250 ou $n_B =$ 2 500 de sorte que le ratio d'échantillonnage est soit 5 % ou 50 %. Les poids de sondage pour l'échantillon A sont disponibles dans l'échantillon A, mais pas dans l'échantillon B. La variable à l'étude y est disponible seulement dans l'échantillon B. La covariable des principaux effets et leurs effets d'interaction par paires $(x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3)$ sont disponibles dans $A \cup B$.

Nous comparons les estimateurs suivants :

Moyenne C Moyenne d'échantillon de l'échantillon non probabiliste C. *Non pondéré* dans l'article.

WBS Estimateur de PLA (propension logistique ajustée) reposant sur la méthode des scores d'équilibrage pondéré, proposé par Wang et coll. (2021).

ABS Estimateur de PLA reposant sur la méthode des scores d'équilibrage adaptatif, proposé par Li.

CLW Estimateur de PPI (pondération probabiliste inverse) de Chen et coll. (2020) reposant sur le modèle de régression logistique pour $\pi_i^{(B)}$.

Cal Estimateur de calage qui satisfait l'équation (3.5) à l'aide d'un modèle de régression logistique pour $\pi_i^{(B)}$.

CPS Estimateur de pseudo-poids proposé dans l'équation (2.6) reposant sur le modèle de probabilité d'inclusion conditionnelle et les poids pondérés dans l'équation (2.5). Le modèle de régression logistique est utilisé pour le modèle de probabilité d'inclusion conditionnelle, et la régression de Poisson a été utilisée pour les poids pondérés de l'échantillon A dans l'équation (2.5).

UCPS Estimateur du pseudo-poids proposé à la section 3 reposant sur le modèle de régression logistique pour $\pi_i^{(B)}$ avec $\hat{\phi}$ estimé par la méthode de maximum de vraisemblance conditionnelle dans l'équation (3.3).

Bien que l'échantillon B soit sélectionné en utilisant l'échantillonnage stratifié, les scores de propension de **WBS**, **ABS**, **CLW**, **CPS** et **UCPS** ont été établis à partir du modèle logistique, et nous avons permis l'erreur de spécification du modèle dans le modèle de réponse de $\pi^{(B)}$.

Les résultats de la simulation après 1 000 simulations sont présentés dans le tableau 4.1. Quand n_B = 250, les estimateurs ABS, CPS et UCPS ont tendance à surclasser tous les autres estimateurs pris en compte. Quand n_B = 2 500, les estimateurs CPS et UCPS sont meilleurs que les autres estimateurs pris en compte. Les méthodes ABS et WBS sont développées en fonction de l'hypothèse selon laquelle le chevauchement entre les deux échantillons est négligeable, mais cette hypothèse ne tient pas pour n_B = 2 500, car le taux d'échantillonnage pour l'échantillon B, n_B / N = 0,5, est non négligeable.

Tableau 4.1
Biais, écart type et racine de l'erreur quadratique moyenne après 1 000 répétitions.

	n_B = 250			n_B = 2 500		
	BIAIS	**ET**	**REQM**	**BIAIS**	**ET**	**REQM**
Moyenne C	0,0533	0,0252	0,0589	0,0514	0,0052	0,0517
WBS	0,0087	0,0275	0,0289	0,0053	0,0139	0,0149
ABS	0,0097	0,0264	0,0281	0,0097	0,0130	0,0162
CLW	0,0084	0,0278	0,0291	-0,0081	0,0234	0,0248
Cal	0,0061	0,0284	0,0291	0,0080	0,0140	0,0161
CPS	0,0095	0,0263	0,0279	0,0035	0,0116	0,0121
UCPS	0,0094	0,0263	0,0280	0,0035	0,0116	0,0121

5. Conclusion

Des hypothèses de modèle pour l'échantillon non probabiliste sont utilisées pour construire des pseudo-poids. Les hypothèses de modèle peuvent être classées en deux groupes, l'un étant l'approche selon un modèle de SP conditionnel et l'autre, l'approche selon un modèle de SP non conditionnel. L'approche selon un modèle de SP conditionnel est attrayante sur le plan du calcul, mais les facteurs de pondération pour l'échantillon A devraient être construits correctement. Dans l'approche selon un modèle de SP non conditionnel, la méthode de pseudo-maximum de vraisemblance de Chen et coll. (2020) a été utilisée. La méthode de Li est plus efficace que la méthode de pseudo-maximum de vraisemblance, pourvu que le taux d'échantillonnage pour l'échantillon B soit négligeable. Dans le présent article, nous proposons une autre approche qui repose sur la méthode de maximum de vraisemblance conditionnelle comme méthode d'estimation efficace, qui peut se justifier même lorsque le taux d'échantillonnage pour l'échantillon B est non négligeable. Le calcul pour la méthode de maximum de vraisemblance conditionnelle est quelque peu utilisé. Beaumont et coll. (2024) ont indépendamment proposé une méthode très similaire, appelée la méthode de maximum de vraisemblance de l'échantillon. Une analyse plus approfondie de la méthode proposée sera présentée ailleurs.

Remerciements

Nous remercions madame Partha Lahiri pour l'invitation à participer à cette discussion et les deux examinateurs anonymes ainsi que l'éditeur pour leurs commentaires constructifs. La recherche a été financée en partie par une subvention de la US National Science Foundation (Numéro de subvention : 2242820) et une subvention de la Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

Bibliographie

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. et Chu, K. (2024). Réponse des auteurs aux commentaires sur l'article « Traitement d'échantillons non probabilistes en podérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada » : De nouvelles avancées concernant les méthodes de vraisemblance pour l'estimation des probabilités de participation pour des échantillons non probabilistes. *Techniques d'enquête*, 50, 1, 139-160. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2024001/article/00001-fra.pdf.

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

Elliott, M.R., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Pfeffermann, D., et Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā B*, 61, 166-186.

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

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

Yang, S., Kim, J.K. et Hwang, Y. (2021). Intégration de données d'enquêtes probabilistes et de mégadonnées aux fins d'inférence de population finie au moyen d'une imputation massive. *Techniques d'enquête*, 47, 1, 29-58. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021001/article/00004-fra.pdf.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 75-80
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Hypothèse de l'échangeabilité dans des méthodes d'ajustement fondées sur le score de propension aux fins d'estimation de la moyenne de population au moyen d'échantillons non probabilistes » :

Inférence causale, échantillon non probabiliste et population finie

Takumi Saegusa[1]

Résumé

Dans certains articles sur les échantillons non probabilistes, l'hypothèse de l'échangeabilité conditionnelle est jugée nécessaire pour une inférence statistique valide. Cette hypothèse repose sur une inférence causale, bien que son cadre de résultat potentiel diffère grandement de celui des échantillons non probabilistes. Nous décrivons les similitudes et les différences entre deux cadres et abordons les enjeux à prendre en considération lors de l'adoption de l'hypothèse d'échangeabilité conditionnelle dans les configurations d'échantillons non probabilistes. Nous examinons aussi le rôle de l'inférence de la population finie dans différentes approches de scores de propension et de modélisation de régression des résultats à l'égard des échantillons non probabilistes.

Mots-clés : Inférence causale; population finie; Échantillon non probabiliste.

1. Introduction

Je félicite la professeure Yan Li pour un autre important ajout à sa recherche active sur les échantillons non probabilistes. Dans son article, la professeure Li a classé la recherche existante sur les échantillons non probabilistes dans les catégories suivantes : 1) les méthodes de pondération du score de propension et 2) les méthodes d'appariement par scores de propension, et indiqué que l'hypothèse d'échangeabilité conditionnelle (EC) était requise pour la première catégorie. Après avoir examiné les méthodes existantes en vue de l'hypothèse d'EC, la professeure Li a proposé le nouveau score d'équilibrage adaptatif pour s'assurer que l'hypothèse d'EC tenait. Étant donné la cristallisation de l'abondance d'articles sur les échantillons non probabilistes et l'inférence causale, son article exige une quantité considérable de connaissances pour comprendre les concepts complexes. Le but premier de notre analyse sera d'examiner les concepts de base et les enjeux fondamentaux que la présentation de la professeure Li n'a abordés que légèrement.

La présente analyse est structurée de la manière suivante. À la section 2, nous analysons l'hypothèse d'échangeabilité conditionnelle dans l'inférence causale. Nous décrivons les différences de cadres probabilistes dans l'inférence causale et les échantillons non probabilistes, et abordons les enjeux à prendre en considération lors de l'adoption de l'hypothèse d'échangeabilité conditionnelle dans les échantillons non

1. Takumi Saegusa, Department of Mathematics, University of Maryland, College Park, Maryland 20742, États-Unis d'Amérique. Courriel : tsaegusa@umd.edu.

probabilistes. Dans la section 3, nous décrivons deux approches principales dans les problèmes de données manquantes, y compris l'inférence causale. Nous abordons ensuite les enjeux liés au rôle de l'inférence de la population finie découlant de l'hypothèse d'échangeabilité conditionnelle dans différentes approches.

2. Inférence causale

Premièrement, nous analysons le lien entre l'hypothèse d'EC et l'inférence causale. Dans le présent article, l'hypothèse d'EC est formulée comme l'équation

$$E[y \mid b(x), C] = E[y \mid b(x), U] \tag{2.1}$$

où $b(x)$ est une fonction des covariables x que l'on appellera un score d'équilibrage, U est une population finie et $C \subset U$ est un échantillon non probabiliste. Bien que défini simplement, le critère de son choix dans l'article indique que le score d'équilibrage semble être implicitement déterminé pour satisfaire l'hypothèse d'EC. Qui plus est, il est énoncé comme fait sans autre discussion que toute quantité (y compris le score de propension) inférieure au score de propension satisfait l'hypothèse d'EC comme score d'équilibrage. Une importante analyse qui aide à comprendre ces concepts est celle qui a été rédigée conjointement par la professeure Li (Wang, Graubard, Katki et Li, 2022), qui est, autant que nous sachions, le premier article qui a présenté explicitement les scores d'équilibrage et l'échangeabilité conditionnelle dans l'inférence causale dans la littérature sur les échantillons non probabilistes. Dans Wang, Graubard, Katki et Li (2022), toutefois, ces concepts étaient directement empruntés des travaux de Rosenbaum et Rubin (1983) sur l'inférence causale, et l'on y affirmait que les résultats sur les scores de propension tenaient dans l'environnement non probabiliste sans analyse formelle. Comme les définitions de l'hypothèse d'EC et du score d'équilibrage dans l'article sont différentes de celles de Rosenbaum et Rubin (1983), et comme le cadre contrefactuel de Rosenbaum et Rubin (1983) est relativement différent de la configuration des échantillons non probabilistes, il est intéressant d'accorder une attention particulière aux similitudes et aux différences entre l'inférence causale et les échantillons non probabilistes.

Pour ce faire, nous résumons d'abord brièvement Rosenbaum et Rubin (1983) où les variables d'intérêt sont les résultats potentiels $(Y(0), Y(1))$, covariables X et l'attribution de traitement $Z \in \{0,1\}$. Le score d'équilibrage $b(x)$ dans Rosenbaum et Rubin (1983) a été défini comme la fonction des covariables $X = x$ qui satisfait l'indépendance conditionnelle entre X et l'attribution de traitement Z étant donné $b(X)$ (c'est-à-dire $X \perp Z \mid b(X)$). Il a été démontré que le score de propension en traitement est un score d'équilibrage, et que toute fonction de x qui peut être cadrée dans le score de propension est aussi un score d'équilibrage. Comme le laisse supposer la définition, il n'y a aucune exigence relativement au lien entre les résultats potentiels et les covariables. L'hypothèse qui relie ces variables est l'échangeabilité conditionnelle en ce qui a trait aux covariables (ou forte ignorabilité de Rosenbaum et Rubin (1983)), définie différemment comme l'indépendance conditionnelle entre les résultats potentiels et l'attribution de traitement étant donné les covariables (c'est-à-dire $(Y(0), Y(1)) \perp Z \mid X$). Le résultat principal est que

l'échangeabilité conditionnelle en ce qui a trait à la covariable X laisse supposer une échangeabilité conditionnelle en ce qui a trait à un score d'équilibrage $b(X)$. Autrement dit, à partir de l'hypothèse d'échangeabilité conditionnelle clé étant donné les covariables x on peut réduire l'information de x à un score d'équilibrage. Les scores d'équilibrage $b(x)$ ne sont significatifs qu'en présence d'échangeabilité conditionnelle en ce qui a trait aux covariables x. Une conséquence de ce résultat est que la différence entre deux résultats potentiels est expliquée uniquement par l'attribution de traitement.

Une façon naturelle d'appliquer ces résultats à la configuration d'échantillon non probabiliste est de considérer la sélection à l'échantillon non probabiliste comme attribution de traitement, et les résultats de l'échantillon non probabiliste C et du reste dans la population finie (c'est-à-dire $U \setminus C$) comme deux résultats potentiels. Dans cette configuration, l'échangeabilité conditionnelle de Rosenbaum et Rubin (1983) suppose l'échangeabilité conditionnelle en ce qui a trait au score de propension de sorte que C et $U \setminus C$ soient comparables étant donné le score de propension. En revanche, la professeure Li suppose immédiatement la comparabilité de C et U étant donné le score de propension. Du point de vue de l'inférence causale, la comparabilité de Rosenbaum et Rubin (1983) est une conséquence d'une hypothèse vérifiable sur le plan conceptuel alors que la professeure Li commence avec la comparabilité désirée en la supposant. Si, au lieu de cela, on commence à partir de l'échangeabilité conditionnelle comme dans Rosenbaum et Rubin (1983), il se pourrait qu'un résultat ne soit tout de même pas satisfaisant, car deux échantillons (soit C et $U \setminus C$) demeurent différents par « traitement » de participation dans un échantillon non probabiliste. Par exemple, si les échantillons non probabilistes sont des dossiers d'hôpitaux ou des participants à un certain programme éducatif, les deux échantillons sont différents en raison de la réception de soins par l'hôpital ou de l'effet éducatif. Même si nous ne trouvons pas un tel « traitement » qui différencie l'échantillon non probabiliste et le reste, la comparabilité conditionnelle entre C et $U \setminus C$ ne correspond pas nécessairement à la population finie U. Pour obtenir la bonne population cible, il faut obtenir une répartition du score de propension dans la population finie U. Cette tâche n'est pas simple à effectuer comme décrit ci-dessous en ce qui a trait à la représentation des probabilités du score de propension.

Une autre approche est de dévier de l'inférence causale en commençant à partir de l'indépendance conditionnelle entre Y et sélection de Z dans C étant donné X au lieu de l'échangeabilité conditionnelle avec des résultats potentiels. Dans ce cas, toutes les dérivations de fait demeurent valides pour conclure le résultat que $Y \perp Z \mid X$ suppose que $Y \perp Z \mid b(X)$ comme souhaité. Cependant, une nouvelle hypothèse d'indépendance conditionnelle est simplement l'hypothèse standard des données manquantes au hasard dans le problème de données manquantes, qui est aussi adoptée par Chen, Li et Wu (2020) dans leur recherche sur les échantillons non probabilistes. L'hypothèse des données manquantes au hasard est connue de nombreux statisticiens et est plus facile à examiner que l'hypothèse d'échangeabilité conditionnelle de la professeure Li. Si cette approche est celle qui est implicitement adoptée dans Wang, Graubard, Katki et Li (2022), ainsi que dans le présent article, il convient d'aborder les avantages supplémentaires de cette approche par rapport à l'hypothèse des données manquantes au hasard en plus de l'écart entre $U \setminus C$ et U

pour la comparabilité. Si une approche différente est adoptée, un lien non vérifié entre les scores d'équilibrage et l'hypothèse d'EC (2.1) devrait être dérivé de manière explicite. Par ailleurs, nous aimerions souligner que Chen, Li et Wu (2020) n'est pas le seul article qui n'utilise pas l'hypothèse d'EC de la professeure Li pour les méthodes de pondération du score de propension (voir par exemple Kim et Morikawa [2023] pour le cas de données manquantes non ignorables).

Comme mentionné ci-dessus, la comparabilité de C et $U \setminus C$ permet une estimation fiable du modèle de régression selon C pour les éléments dans $U \setminus C$ mais l'estimation de $\overline{Y}_N$ exige une estimation cohérente des scores de propension pour U pour relier la régression étant donné X à l'ensemble de la population U. Cependant, une simple estimation du score de propension n'est pas possible, car X n'est pas disponible pour tous les éléments dans $U \setminus C$. La variable X est disponible dans un échantillon de référence S de U avec un plan d'échantillonnage connu, mais S n'est pas une solution de rechange simple pour $U \setminus C$ car les éléments dans S peuvent aussi être dans un échantillon non probabiliste C. Pour solutionner cet enjeu complexe, Wang, Valliant et Li (2021) ont découvert le lien entre le score de propension dans C par rapport à U et le score de propension dans C par rapport à l'échantillon empilé de C et U où les mêmes éléments dans C et S sont traités différemment (pour une dérivation rigoureuse, voir Savitsky, Williams, Gershunskaya, Beresovsky et Johnson [2023]). Au moyen de cette relation, la professeure Li a modélisé le dernier score de propension par régression binaire pour estimer le premier. L'événement pour le dernier score de propension pour un échantillon empilé est construit artificiellement et conceptuellement difficile à modéliser. Cet enjeu augmente la possibilité accrue d'erreur de spécification du modèle, qui invaliderait l'estimation cohérente avec le plan d'échantillonnage de $\overline{Y}_N$. L'événement pour le premier score de propension est l'événement initial et est naturel à modéliser. Cette approche a été adoptée par Savitsky, Williams, Gershunskaya, Beresovsky et Johnson (2023).

3. Inférence de la population finie

Un autre concept que nous voulons aborder est le rôle de la population finie dans les échantillons non probabilistes. Le but de l'article est de mettre au point un estimateur cohérent avec le plan d'échantillonnage de la moyenne de la population finie $\overline{Y}_N$. Aux fins de convergence par rapport au plan de sondage, l'on suppose une série de conditions dans la séquence des populations finies avec toutes les variables sauf la sélection dans les échantillons traitée de façon non aléatoire. En revanche, l'approche basée sur le modèle traite la population finie comme réalisation aléatoire à partir de la super population, et modèle la relation stochastique parmi les variables. Dans la recherche sur les données manquantes, par ailleurs, deux approches principales (et leurs combinaisons) pour l'estimation sont la modélisation du score de propension et la modélisation de régression du résultat. Une approche plus convenable à l'égard de l'approche fondée sur le plan de sondage est la modélisation du score de propension qui modélise la sélection dans les échantillons étant donné les covariables, car il est possible de considérer des sélections aléatoires alors que toutes les

autres variables peuvent être traitées fixes. En revanche, la modélisation de régression du résultat présume une répartition pour Y étant donné X, et elle convient à l'approche basée sur un modèle.

La professeure Li a fait une tentative difficile de lier l'approche de régression du résultat à l'approche fondée sur le plan de sondage. Il convient de noter que l'attente conditionnelle peut être considérée comme une régression avec des variables conditionnelles comme covariables. De cette perspective, l'approche semble être purement l'approche fondée sur un modèle en fonction de la régression du résultat. Cependant, la professeure Li a tenté de soigneusement mettre au point l'attente conditionnelle étape par étape en commençant par une population finie et un échantillon non probabiliste. Si la condition était purement fondée sur un modèle, la variable y dans la condition (2.1) est simplement une variable aléatoire à partir de la super population. Dans l'approche conditionnelle de l'article, cette variable y devrait être clairement définie par rapport à la population finie U et l'échantillon non probabiliste C au moyen d'indices. Si y est un choix aléatoire d'une variable d'un échantillon S de U, $E_S[y|U]=\sum_{i\in U}\pi_i Y_i$ où π_i désigne la probabilité d'inclusion pour l'unité i. Dans ce cas, l'échantillon autopondéré S satisfait $E_S[y|U]=\bar{Y}_N$ mais un échantillon stratifié S, par exemple, ne satisfait pas cette équation en général. Autrement dit, il se pourrait que l'enjeu de biais allégué ne soit pas unique à un échantillon non probabiliste. Pour mesurer pleinement la condition d'échangeabilité conditionnelle, une définition claire de y dans C ou U est plus que souhaitable. En outre, il est souhaitable d'élucider la manière dont la condition fondée sur un modèle de l'hypothèse d'EC mène au résultat fondé sur le plan de sondage malgré l'écart conceptuel.

Bibliographie

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. Extrait de https://doi-org.proxy-um.researchport.umd.edu/10.1080/01621459.2019.1677241. Doi: 10.1080/01621459.2019.1677241.

Kim, J., et Morikawa, K. (2023). An empirical likelihood approach to reduce selectionbias in voluntary samples. À paraître dans *Calcutta Statistical Association Bulletin*, 35.

Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55. Extrait de https://doi-org.proxy-um.researchport.umd.edu/10.1093/biomet/70.1.41. Doi: 10.1093/biomet/70.1.41.

Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovsky, V. et Johnson, N.G. (2023). *Methods for Combining Probability and Nonprobability Samples Under Unknown Overlaps*.

Wang, L., Graubard, B.I., Katki, H.A. et Li, Y. (2022). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. *Revue Internationale de Statistique*, 90, 146-164.

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40(24), 5237-5250. Extrait de https://doi-org.proxy-um.researchport.umd.edu/10.1002/sim.9122. Doi: 10.1002/sim.9122.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 81-86
Statistique Canada, nº 12-001-X au catalogue

Réponse de l'auteur aux commentaires sur l'article « Hypothèse de l'échangeabilité dans des méthodes d'ajustement fondées sur le score de propension aux fins d'estimation de la moyenne de population au moyen d'échantillons non probabilistes »

Yan Li[1]

Résumé

Dans cette réplique, je réponds aux commentaires des participants à l'analyse, M. Takumi Saegusa, M. Jae-Kwang Kim et Mme Yonghyun Kwon. Les commentaires de M. Saegusa, qui portent sur les différences entre l'hypothèse d'échangeabilité conditionnelle (EC) pour les inférences causales et l'hypothèse d'EC pour les inférences de population finie au moyen d'échantillons non probabilistes ainsi que sur la distinction entre les méthodes fondées sur le plan et celles fondées sur un modèle pour l'inférence de population finie au moyen d'échantillons non probabilistes, sont examinés et clarifiés dans le contexte de mon article. Je réponds ensuite au cadre exhaustif de M. Kim et de Mme Kwon pour classer les méthodes actuelles d'estimation des scores de propension (SP) en méthodes conditionnelles et inconditionnelles. J'étends leurs études par simulations pour varier les poids de sondage, permettre des modèles de SP incorrectement précisés, et inclure un estimateur supplémentaire, à savoir l'estimateur par la propension logistique ajustée mis à l'échelle (Wang, Valliant et Li (2021), noté sWBS). Dans mes simulations, on observe que l'estimateur sWBS dépasse de façon constante les autres estimateurs ou leur est comparable dans le modèle de SP incorrectement précisé. L'estimateur sWBS, ainsi que les estimateurs WBS ou ABS décrits dans mon article, ne supposent pas que les unités superposées dans les échantillons de référence probabiliste et non probabiliste sont négligeables, et ils n'exigent pas non plus l'identification des unités superposées, comme le nécessitent les estimateurs proposés par M. Kim et Mme Kwon.

Mots-clés : Échangeabilité conditionnelle; inférences causales; score de propension; essais randomisés; études d'observation; étude de séroprévalence du SRAS-CoV-2.

Je tiens à remercier les participants à la discussion pour leurs observations perspicaces sur mon article ainsi que pour les excellentes références supplémentaires qu'ils ont citées. Je commencerai par répondre aux commentaires de M. Saegusa sur deux éléments importants. Le premier compare les différences entre l'hypothèse d'échangeabilité conditionnelle (EC) pour les inférences causales et l'hypothèse d'EC pour les inférences de population finie au moyen d'échantillons non probabilistes. Le deuxième élément porte sur la distinction entre les méthodes fondées sur le plan et celles fondées sur un modèle aux fins d'inférence de population finie au moyen d'échantillons non probabilistes.

1. Réponse aux commentaires de M. Saegusa

L'hypothèse d'EC dans l'inférence causale et l'inférence de population finie

M. Saegusa a fourni une explication détaillée de l'hypothèse d'EC pour l'estimation des effets causaux des traitements dans des essais randomisés et des études d'observation. La condition clé est la satisfaction

1. Yan Li, Joint Program in Survey Methodology and Department of Epidemiology and Biostatistics, University of Maryland, College Park. Courriel : yli6@umd.edu.

de l'échangeabilité conditionnelle pour que soient effectuées des inférences causales. Dans les essais randomisés, l'EC des résultats potentiels est obtenue par l'attribution aléatoire de traitements. En revanche, dans l'analyse des études d'observation, l'EC est supposée plutôt que garantie afin que soient tirées les conclusions causales. L'hypothèse d'EC dans les études d'observation affirme que la distribution des résultats potentiels (pour différents traitements), compte tenu de toutes les covariables *observées*, est échangeable entre les groupes de traitement. Dans ce cas, il n'y a pas de covariables *non observées* qui influencent à la fois l'attribution du traitement et le résultat d'intérêt; voir Rubin (2007) pour en savoir plus sur l'estimation de l'effet causal au moyen d'essais randomisés et d'études d'observation. Notons que contrairement aux *essais randomisés*, dans le contexte de l'inférence de population finie (FP) au moyen d'échantillons non probabilistes, l'EC est supposée plutôt que garantie, ce qui est semblable à l'hypothèse d'EC nécessaire aux fins de l'inférence causale dans les études d'observation.

M. Saegusa a correctement établi un lien entre l'autosélection dans les échantillons non probabilistes et l'attribution du traitement dans l'inférence causale, et il a défini que le fait d'être autosélectionné dans l'échantillon non probabiliste C par rapport à dans le reste de la population finie (c.-à-d. U\C) comme étant les deux traitements. Toutefois, dans l'inférence de PF au moyen d'échantillons non probabilistes, nous souhaitons estimer la moyenne de PF pour un seul résultat, plutôt que les effets du traitement, c.-à-d. la différence entre deux résultats potentiels selon les traitements C et U\C. La situation n'est pas celle de plusieurs traitements appliqués à des groupes « semblables » pour obtenir plusieurs résultats. On a plutôt un seul résultat potentiel réalisé. Par conséquent, l'hypothèse d'EC dans l'inférence de PF diffère de l'EC dans les études d'observation, où elle affirme que la distribution du résultat (*unique*) est échangeable entre l'échantillon non probabiliste C et la population finie U, compte tenu de toutes les covariables *observées*. Selon cette hypothèse d'EC, on peut alors inférer la moyenne de PF au moyen de C sans observer le résultat dans U\C.

En résumé, l'EC dans l'inférence de PF est semblable à l'EC dans l'inférence causale quand on utilise des études d'observation. Cependant, contrairement à l'inférence causale, l'échangeabilité se rapporte à un seul résultat entre C et U dans l'inférence de PF.

Méthodes fondées sur un modèle ou sur le plan de sondage aux fins d'inférences de population finie

M. Saegusa a bien exposé les différences fondamentales entre les méthodes fondées sur un modèle et celles fondées sur le plan aux fins de l'inférence de PF. Les méthodes fondées sur un modèle traitent le résultat comme une variable aléatoire, tandis que les méthodes fondées sur le plan considèrent l'indicateur de sélection (dans l'échantillon) comme un indicateur aléatoire (avec un résultat constant). Dans l'article, un ensemble de pseudo-poids fondés sur le plan a été construit pour l'échantillon non probabiliste afin d'estimer la moyenne de la PF selon l'hypothèse d'EC de $E\{y \mid b(\mathbf{x}), C\} = E\{y \mid b(\mathbf{x}), U\}$, où l'espérance $E(.)$ est par rapport à deux niveaux de caractère aléatoire de (1) la réalisation aléatoire de la PF à partir d'une superpopulation et de (2) l'autosélection aléatoire dans C à partir de la population finie U.

M. Saegusa indique ensuite qu'une « définition claire de y dans C et/ou U est souhaitable ». Cependant, les résultats pour obtenir une estimation sans biais de la moyenne de PF s'appliquent tant que la PF est la réalisation aléatoire d'une superpopulation. Seule l'existence d'une fonction de distribution avec des moments finis appropriés de variables est nécessaire pour la superpopulation. Il n'est pas nécessaire de préciser une forme spécifique du modèle paramétrique; voir par exemple Graubard et Korn (2002) pour en savoir plus.

2. Réponse aux commentaires de M. Jae-Kwang Kim et de Mme Yonghyun Kwon

Dans ce qui suit, j'aborde la discussion très judicieuse de M. Kim et de Mme Kwon, dans laquelle ils ont d'abord présenté un cadre exhaustif établi pour classer les méthodes actuelles d'estimation des scores de propension (SP) en méthodes conditionnelles et inconditionnelles. Ils ont ensuite mené des études par simulations comparant différents estimateurs, et je suis heureux de constater que l'estimateur ABS (échantillonnage fondé sur des adresses) que nous avons proposé a bien fonctionné dans leurs simulations.

Les méthodes conditionnelles comprennent deux phases : la première phase consistant à échantillonner $S_1 = A \cup B$ à partir de U et la deuxième phase à échantillonner B à partir de S_1. Les paramètres du modèle ϕ de la probabilité d'inclusion conditionnelle pour la deuxième phase $P(i \in S_2 \mid i \in S_1) = p(x_i;\phi)$ sont estimés par

$$\hat{\phi} = \arg\max \sum_{i \in S_1} \left[\delta_i \log\left(p(x_i;\phi)\right) + (1-\delta_i)\log\left(1-p(x_i;\phi)\right) \right],$$

où $\delta_i = I(i \in B)$ est la fonction indicatrice de l'événement $i \in B$. Selon un modèle statistique, disons un modèle de régression logistique $\text{logit}\{p(x_i;\phi)\} = x_i'\phi$, on peut obtenir $\hat{\phi}$ en résolvant pour ϕ

$$\sum_{i \in B} \left(1 - p(x_i;\phi)\right) x_i - \sum_{i \in A \text{ et } i \notin B} p(x_i;\phi)\, x_i = 0.$$

Notons que les unités qui se superposent et qui sont sélectionnées dans les deux échantillons de A et de B doivent être identifiées et retirées de l'échantillon A pour la deuxième sommation ci-dessus. Involontairement, « et $i \notin B$ » a été omis dans la deuxième sommation de la discussion. À partir de l'estimation $\hat{\phi}$, M. Kim et Mme Kwon ont proposé le pseudo-poids du score de propension conditionnel ou SPC pour la i^{e} unité dans B, donné par

$$\hat{w}_i^{(B)} = 1 + w_i^{(A)} \left(\frac{1}{p(x_i;\hat{\phi})} - 1 \right),$$

où $w_i^{(A)}$ est souvent inconnu et estimé selon un modèle paramétrique en pratique.

Les méthodes inconditionnelles comprenaient seulement une étape. L'estimateur par le maximum de vraisemblance conditionnelle de ϕ a été estimé à partir de l'échantillon combiné $S_1 = A \cup B$. Tout comme

l'estimateur SPC, la méthode proposée du score de propension inconditionnel (SPIC) repose également sur l'hypothèse que les unités qui appartiennent à l'intersection de A et B peuvent être identifiées.

Le SP conditionnel et le SP inconditionnel proposés ont été évalués au moyen d'études par simulations, où il a été tenu compte de tailles d'échantillon variées dans l'échantillon B sélectionné au moyen d'un échantillonnage aléatoire simple stratifié (EASS) avec une variable de stratification par catégorie. Bien que simple, ce plan est intelligent. Il s'harmonise avec le véritable modèle de SP sous-jacent pour toutes les méthodes examinées, ce qui assure une comparaison équitable. Pour mieux évaluer le rendement des estimations proposées, nous avons étendu les études par simulations en incluant un estimateur supplémentaire selon le même plan d'échantillonnage EASS, mais en faisant varier les poids de sondage. Rappelons que la taille de la population est $N = 5\ 000$, que la taille de l'échantillon A est $n_A = 250$, et celle de l'échantillon B varie entre $n_B = 250$ et 2 500. Nous considérons les trois estimateurs qui ont les plus petites erreurs quadratiques moyennes (EQM) dans le tableau 4.1 de leur discussion : WBS, ABS et SPC. Puisque le SPIC fonctionne de la même façon que le SPC, il n'est pas pris en compte ici. Rappelons que WBS désigne l'estimateur par la propension logistique ajustée proposé par Wang et coll. (2021). Dans le même article, les auteurs ont également proposé l'estimateur WBS mis à l'échelle, noté sWBS, dans lequel les poids mis à l'échelle sont la valeur de 1 pour les unités de l'échantillon B et $n_s w_i^{(A)} \Big/ \sum_{i \in S_A} w_i^{(A)}$ pour l'unité i de l'échantillon A. Les fractions de sondage varient à l'intérieur des strates d'échantillonnage. Dans la strate 1, $n_{B1} = f_1 n_B$ échantillons sont sélectionnés par échantillonnage aléatoire simple. Dans la strate 2, $n_{B2} = (1 - f_1)\, n_B$ échantillons sont sélectionnés par échantillonnage aléatoire simple. La valeur de f_1 varie et est 0,7, 0,8 et 0,9 pour produire des valeurs différentes du coefficient de variation des poids de sondage de l'EASS (CVWT pour l'anglais *coefficient of variation of the weights*). Dans l'analyse du score de propension, nous considérons deux modèles : (M1) les effets principaux (x_1, x_2, x_3) et leurs effets d'interaction par paires; (M2) les effets principaux (x_{1c}, x_2, x_3) et leurs effets d'interaction par paires, où $x_{1c} = I\{x_1 < 0\}$, la fonction indicatrice de l'événement $x_1 < 0$ où x_1 sont générés à partir d'un $N(0,1)$. Notons que M2 est conforme au plan d'EASS tandis que M1 est incorrectement précisé par l'inclusion de la variable continue x_1 dans l'analyse du SP.

Quatre observations sont formulées : (1) les quatre estimateurs sont approximativement sans biais selon le vrai modèle de SP; (2) les estimateurs ABS et SPC ont des performances similaires pour une petite taille d'échantillon de $n_B = 250$ selon les deux modèles; (3) quand la taille de l'échantillon est grande $n_B = 2\ 500$, le SPC a toujours le plus petit ET et la plus petite EQM dans le vrai modèle. Ces résultats correspondent aux attentes puisqu'il y a un fort pourcentage d'unités superposées dans les deux échantillons. Par conséquent, la méthode du SPC gagne en efficacité, car elle suppose que les unités superposées peuvent être identifiées ; (4) dans le modèle de SP incorrectement précisé, sWBS présente toujours le biais le plus petit, particulièrement quand le coefficient de variation des poids de sondage est grand. En revanche, le SPC a le biais le plus grand et l'ET le plus grand quand le coefficient de variation des poids de sondage et n_B sont grands. Le biais et la perte d'efficacité du SPC peuvent être attribués à la spécification erronée du modèle de $P(i \in B \mid i \in A \cup B)$, la taille limitée de l'échantillon obtenue par la suppression des unités superposées

(~50 %) de l'échantillon A, et les poids de sondage variables. Le SPC est sensible à la spécification erronée du modèle, en particulier quand n_B et le coefficient de variation des poids de sondage sont grands.

En résumé, sous le vrai modèle de score de propension, les estimateurs ABS et SPC ont des performances similaires quand n_B est petit. Quand n_B est grand, l'estimateur SPC est plus efficace en raison du nombre croissant d'unités sélectionnées et identifiées dans les deux échantillons, A et B. En cas de spécification erronée du modèle de SP, sWBS (Wang et coll., 2021) a de meilleures performances ou des performances comparables à celles des autres estimateurs. Il faudrait étudier de façon plus approfondie les effets de divers modèles de SP incorrectement précisés ou les valeurs de mise à l'échelle sur les performances de l'estimateur sWBS. Ensuite, on a élaboré les estimateurs ABS, WBS, sWBS, ainsi que SPC, sans supposer que les unités superposées dans les deux échantillons sont négligeables. Pour un échantillon de grande taille $n_B = 2\ 500$, le taux d'échantillonnage pour l'échantillon B, $n_B/N = 50$ %, est non négligeable. Tous les estimateurs, comme le montre le tableau 1, sont approximativement sans biais selon le vrai modèle de SP, qui prouve empiriquement que les quatre méthodes ne nécessitent pas l'hypothèse selon laquelle les unités superposées dans les deux échantillons sont négligeables. Enfin, du point de vue pratique, il importe que le lecteur sache que l'estimateur SPC exige l'identification des unités superposées. Or, cela peut être impossible dans de nombreuses situations. Par exemple, dans l'étude de la NIH sur la séropositivité au SARS-CoV-2 abordée dans mon article, ces renseignements d'identification n'avaient pas été recueillis.

Tableau 1
Biais, erreur type et erreur quadratique moyenne (× 100) selon un EASS avec variation du CV des poids de sondage (CVWT) après 5 000 répétitions.

	Modèle de SP correctement précisé (x_{1c})						Modèle de SP incorrectement précisé (x_1)					
	n_B = 250			n_B = 2 500			n_B = 250			n_B = 2 500		
	BIAIS	ET	EQM	BIAIS	ET	EQM	BIAIS	ET	EQM	BIAIS	ET	EQM
CVWT = 0,44												
Moyenne C	-3,25	2,98	4,41	-3,33	0,94	3,46	4,84	2,87	5,63	4,74	0,93	4,83
WBS	0,04	3,60	3,60	-0,04	1,42	1,42	0,51	3,06	3,10	0,23	1,54	1,56
sWBS	0,02	3,56	3,56	-0,05	1,41	1,41	0,33	3,05	3,07	0,18	1,62	1,63
ABS	0,02	3,54	3,54	-0,04	1,38	1,38	0,56	2,98	3,04	0,56	1,45	1,55
SPC	0,01	3,53	3,53	0,01	1,20	1,20	0,55	2,99	3,04	0,12	1,36	1,36
CVWT = 0,75												
Moyenne C	-4,97	2,89	5,76	-4,99	0,92	5,07	7,10	2,97	7,70	7,10	0,95	7,16
WBS	-0,03	4,16	4,16	0	1,56	1,56	1,03	3,18	3,34	0,47	1,55	1,62
sWBS	-0,08	4,09	4,09	-0,03	1,55	1,55	0,39	3,23	3,25	0,18	1,64	1,65
ABS	-0,08	4,08	4,08	-0,02	1,52	1,52	1,13	3,15	3,34	1,11	1,52	1,88
SPC	-0,08	4,07	4,07	0,05	1,34	1,34	1,10	3,16	3,34	-0,41	1,64	1,69
CVWT = 1,33												
Moyenne C	-6,58	2,88	7,19	-6,66	0,90	6,72	9,49	3,07	9,98	9,45	0,97	9,50
WBS	0,11	5,65	5,65	0,00	1,91	1,91	2,74	3,49	4,44	1,39	1,67	2,17
sWBS	0,06	5,54	5,54	-0,03	1,89	1,89	1,07	3,78	3,93	-0,05	1,85	1,85
ABS	0,03	5,49	5,49	-0,04	1,86	1,86	2,60	3,50	4,36	1,94	1,69	2,58
SPC	0,03	5,49	5,49	0,03	1,71	1,71	2,54	3,54	4,36	-3,00	3,10	4,31

Bibliographie

Graubard, B.I., et Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17(1), 73-96.

Rubin, D.B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat Med.*, 26(1), 20-36. Doi: 10.1002/sim.2739. PMID: 17072897.

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. Doi: 10.1002/sim.9122. PMID: 34219260; PMCID: PMC8526388.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 87-121
Statistique Canada, nº 12-001-X au catalogue

Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada

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

Résumé

Les organismes nationaux de statistique étudient de plus en plus la possibilité d'utiliser des échantillons non probabilistes comme solution de rechange aux échantillons probabilistes. Toutefois, il est bien connu que l'utilisation d'un échantillon non probabiliste seul peut produire des estimations présentant un biais important en raison de la nature inconnue du mécanisme de sélection sous-jacent. Il est possible de réduire le biais en intégrant les données de l'échantillon non probabiliste aux données d'un échantillon probabiliste, à condition que les deux échantillons contiennent des variables auxiliaires communes. Nous nous concentrons sur les méthodes de pondération par l'inverse de la probabilité, lesquelles consistent à modéliser la probabilité de participation à l'échantillon non probabiliste. Premièrement, nous examinons le modèle logistique ainsi que l'estimation par la méthode du pseudo maximum de vraisemblance. Nous proposons une procédure de sélection de variables en fonction d'un critère d'information d'Akaike (AIC) modifié qui tient compte de la structure des données et du plan d'échantillonnage probabiliste. Nous proposons également une méthode simple fondée sur le rang pour former des strates a posteriori homogènes. Ensuite, nous adaptons l'algorithme des arbres de classification et de régression (CART) à ce scénario d'intégration de données, tout en tenant compte, encore une fois, du plan d'échantillonnage probabiliste. Nous proposons un estimateur de la variance bootstrap qui tient compte de deux sources de variabilité : le plan d'échantillonnage probabiliste et le modèle de participation. Nos méthodes sont illustrées au moyen de données recueillies par approche participative et de données d'enquête de Statistique Canada.

Mots-clés : Critère d'information d'Akaike; arbres de classification et de régression; modèle logistique; probabilité de participation; intégration de données statistiques; sélection de variables.

1. Introduction

Statistique Canada et d'autres organismes statistiques partout dans le monde étudient de plus en plus les échantillons non probabilistes. De fait, Statistique Canada a récemment mené plusieurs enquêtes non probabilistes pour évaluer les répercussions de la pandémie de COVID-19 sur différents aspects de la vie de la population canadienne. Les données de ces enquêtes non probabilistes ont été recueillies auprès de visiteurs du site Web de Statistique Canada qui ont répondu volontairement à un questionnaire en ligne. La principale raison pour laquelle on a envisagé cette approche non probabiliste, appelée « approche participative » à Statistique Canada, au lieu d'une enquête probabiliste, est la réduction importante de temps et de coût dans la production de statistiques. Un autre avantage important est la nature non intrusive de l'approche participative, puisque la participation se fait sur une base volontaire. Toutefois, il est bien connu que l'utilisation d'un échantillon non probabiliste seul, comme un échantillon recueilli par approche participative, peut produire des estimations présentant un biais important en raison de la nature inconnue du mécanisme de sélection (ou de participation) sous-jacent. Pour réduire ce biais de participation, on peut

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

combiner les données d'un échantillon non probabiliste aux données d'un échantillon probabiliste, idéalement de grande taille. Les méthodes d'estimation qui combinent des données provenant d'un échantillon probabiliste et d'un échantillon non probabiliste relèvent du domaine de l'intégration de données statistiques.

Nous examinons un scénario d'intégration de données dans lequel les variables d'intérêt ne sont disponibles que dans l'échantillon non probabiliste. Cependant, un vecteur de variables auxiliaires est observé dans les deux échantillons et utilisé aux fins de réduction du biais. Une des méthodes d'inférence possible dans ce scénario repose sur un modèle pour les variables d'intérêt, ainsi que sur l'hypothèse que l'échantillon non probabiliste n'est pas informatif par rapport au modèle. L'approche de prédiction pour les populations finies (par exemple Royall, 1970; Valliant, Dorfman and Royall, 2000) est une voie possible pour l'intégration de données. Si un modèle linéaire entre les variables d'intérêt et les variables auxiliaires se vérifie, on peut le mettre en œuvre en pondérant l'échantillon non probabiliste au moyen d'un calage sur les totaux de population connus ou les totaux estimés à partir de l'enquête probabiliste (par exemple Elliott et Valliant, 2017; Valliant, 2020). L'appariement statistique est une autre méthode fondée sur un modèle (voir Yang, Kim et Hwang, 2021, pour une publication récente). Il consiste à imputer les valeurs manquantes des variables d'intérêt dans l'échantillon probabiliste au moyen de données de l'échantillon non probabiliste. La méthode est appelée appariement d'échantillons (par exemple Rivers, 2007) quand l'imputation par donneur est utilisée pour remplacer les valeurs manquantes. L'approche de prédiction avec totaux estimés et l'appariement statistique donnent des estimateurs identiques pour un modèle linéaire dont la variance de l'erreur est liée linéairement aux variables auxiliaires (Beaumont, 2020). Étant donné que les deux méthodes reposent sur un modèle pour les variables d'intérêt, elles ne sont pas nécessairement pratiques en présence de multiples variables d'intérêt, car il faut déterminer et valider un modèle pour chacune d'elles.

Une autre approche d'inférence repose sur un modèle pour l'indicateur de participation plutôt que sur un modèle pour les variables d'intérêt. Cette approche est plus intéressante en présence de multiples variables d'intérêt, car il n'y a qu'un seul indicateur de participation, et donc un seul modèle à choisir et à valider. On obtient les estimations en pondérant chaque participant de l'échantillon non probabiliste par l'inverse de sa probabilité de participation estimée. Cette approche est souvent désignée dans la littérature par les expressions « pondération par l'inverse de la probabilité » ou « pondération par le score de propension ». Nous nous concentrons sur cette approche. Si les valeurs des variables auxiliaires sont observées pour l'ensemble de la population, le problème est essentiellement identique à celui de la pondération pour la non-réponse aux enquêtes, et il est possible d'appliquer directement des méthodes de pondération de la non-réponse pour pondérer l'échantillon non probabiliste.

En général, les variables auxiliaires ne sont observées que pour les participants de l'échantillon non probabiliste. Chen, Li et Wu (2020) ont proposé une méthode simple et prometteuse pour régler ce problème. Elle exige que les variables auxiliaires soient également observées dans un échantillon probabiliste et suppose que la fonction logistique est utilisée pour modéliser la probabilité de participation. Une solution de rechange à celle de Chen, Li et Wu (2020) consiste à créer un échantillon groupé, en combinant

l'échantillon probabiliste et non probabiliste, et à modéliser l'indicateur de participation en supposant qu'il n'y a pas de chevauchement entre les deux échantillons (par exemple Lee, 2006; Valliant et Dever, 2011; et Ferri-Garcia et Rueda, 2018). Chen, Li et Wu (2020) ont observé que cette méthode de regroupement donne lieu à un estimateur biaisé de la probabilité de participation. Toutefois, Beaumont (2020) a souligné qu'elle donne des probabilités de participation estimées approximativement équivalentes à celles de Chen, Li et Wu (2020) quand toutes les probabilités de participation sont petites et que l'échantillon probabiliste est adéquatement pondéré. Wang, Valliant et Li (2021) ont proposé une extension de la méthode de regroupement pour tenir compte d'un chevauchement non négligeable entre l'échantillon probabiliste et non probabiliste. Elliott et Valliant (2017) ont proposé une autre méthode de pondération par l'inverse de la probabilité qui repose sur l'échantillon groupé. Elle suppose également l'absence de chevauchement entre les deux échantillons et exige que les poids de sondage probabiliste soient disponibles dans l'échantillon non probabiliste. Des synthèses récentes sur les méthodes d'intégration de données statistiques sont présentées dans Beaumont (2020), Lohr (2021), Rao (2021), Valliant (2020), Wu (2022) et Yang et Kim (2020).

Le choix des variables auxiliaires est essentiel aux fins de réduction du biais. Idéalement, elles doivent être associées à la fois à l'indicateur de participation et aux variables d'intérêt. Chen, Li et Wu (2020) ont supposé que les variables auxiliaires étaient données. En pratique, plusieurs variables auxiliaires peuvent être disponibles dans les deux échantillons, souvent catégoriques, et il n'est pas toujours évident de déterminer celles qui sont pertinentes ainsi que les interactions appropriées. Les outils de sélection de variables pourraient être utiles, mais ils doivent être adaptés au scénario d'intégration de données envisagé dans le présent article. En particulier, ils doivent tenir compte du plan de sondage utilisé pour sélectionner l'échantillon probabiliste et de tout ajustement des poids de sondage, comme les ajustements pour la non-réponse et le calage. Nous proposons une procédure de sélection pas à pas qui atteint cet objectif. Elle repose sur une modification du critère d'information d'Akaike (AIC) semblable à celle élaborée par Lumley et Scott (2015) pour l'estimation de paramètres d'un modèle à partir de données d'enquête probabiliste. Bahamyirou et Schnitzer (2021) ont envisagé de recourir à la méthode LASSO (opérateur de sélection et de contraction par moindres valeurs absolues) comme solution de rechange. Cette technique utilise habituellement la validation croisée pour déterminer le paramètre de pénalité. L'élaboration de méthodes de validation croisée qui tiennent compte adéquatement de la combinaison d'un échantillon probabiliste et non probabiliste et du plan d'échantillonnage probabiliste nécessite des travaux supplémentaires.

Le modèle logistique peut parfois produire quelques probabilités estimées qui sont très petites, ce qui peut mener à des poids très élevés et des estimations potentiellement instables. Une solution courante à ce problème dans le contexte de la non-réponse aux enquêtes consiste à créer des groupes homogènes et à pondérer chaque répondant (participant) dans un groupe donné par l'inverse du taux de réponse (participation) estimé dans le groupe. Les poids qui en résultent possèdent une propriété de calage (voir la section 3.3), qui tend à limiter l'ampleur des poids les plus grands. La création de groupes homogènes fournit également une certaine robustesse face à une spécification erronée du modèle, comme l'illustrent Haziza et Lesage (2016) dans le contexte de la non-réponse aux enquêtes.

Une voie possible pour la création de groupes homogènes consiste à adapter l'algorithme des arbres de classification et de régression (CART), élaboré par Breiman, Friedman, Olshen et Stone (1984), au scénario d'intégration de données étudié dans le présent article. Le fait que les variables auxiliaires et leurs interactions sont choisies automatiquement est un avantage important des méthodes basées sur les arbres de régression. Chu et Beaumont (2019) ont élaboré un algorithme permettant la croissance d'un arbre qui tient compte des poids de sondage. Ils ont appelé l'algorithme « nppCART » parce qu'il intègre les données de l'échantillon non probabiliste et probabiliste. L'élagage est un aspect important de l'algorithme CART, qui sert à éviter le surajustement et à améliorer l'efficacité des estimations qui en résultent. L'élagage repose souvent sur des techniques de validation croisée, mais, comme nous l'avons indiqué ci-dessus, ces techniques n'ont pas encore été étendues au scénario d'intégration de données étudié dans le présent article. Nous envisageons ici plutôt une modification du critère AIC, semblable à celle de Lumley et Scott (2015), qui tient adéquatement compte du plan d'échantillonnage probabiliste et de tout ajustement des poids de sondage, et nous l'utilisons pour élaborer une procédure d'élagage.

Dans la section 2, nous présentons le problème d'intégration de données dc même que certaines notations. L'estimation des probabilités de participation est abordée dans les sections 3 et 4. Dans la section 3, nous examinons plus particulièrement le modèle logistique et nous décrivons la procédure de sélection de variables que nous proposons ainsi qu'une méthode simple fondée sur le rang, appelée méthode Frank, pour la création de groupes homogènes. Dans la section 4, nous décrivons l'algorithme nppCART et la procédure d'élagage proposée. L'estimation bootstrap de la variance de nos estimateurs est abordée à la section 5. Une évaluation empirique de nos méthodes utilisant des données réelles est présentée à la section 6. La dernière section présente quelques observations finales.

2. Scénario d'intégration de données

Examinons l'estimation du total de la population $\theta = \sum_{k \in U} y_k$, où U est l'ensemble des unités de la population et y_k est la valeur d'une variable d'intérêt y pour l'unité k de la population. Nous supposons que y_k est observé sans erreur dans un échantillon non probabiliste $s_{\text{NP}} \subset U$. En plus de y_k, un vecteur de variables auxiliaires $\mathbf{x}_k$ est également observé pour chaque unité $k \in s_{\text{NP}}$. L'indicateur de participation à l'échantillon non probabiliste est noté par δ_k, c'est-à-dire $\delta_k = 1$ si $k \in s_{\text{NP}}$, et $\delta_k = 0$ autrement. Un échantillon probabiliste s_P, réalisé au moyen d'un certain plan d'échantillonnage probabiliste, est également disponible. Les variables auxiliaires $\mathbf{x}_k$ sont observées pour chaque unité $k \in s_P$, mais la variable d'intérêt y_k et l'indicateur de participation δ_k sont absents de l'échantillon probabiliste.

L'objectif est d'estimer θ selon le scénario d'intégration de données ci-dessus, c'est-à-dire en utilisant les valeurs y observées dans l'échantillon non probabiliste ainsi que les valeurs de $\mathbf{x}$ observées dans les deux échantillons. La pondération par l'inverse de la probabilité nécessite de modéliser la probabilité de participation $p_k = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$, qu'on suppose strictement supérieure à 0. L'estimateur de θ selon cette méthode est $\hat{\theta}_{\text{NP}} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k$, où $\hat{w}_k^{\text{NP}} = \hat{p}_k^{-1}$ est le poids de sondage non probabiliste, aussi appelé le pseudo poids de sondage, du participant k, et $\hat{p}_k$ est un estimateur convergent de p_k. Une hypothèse

essentielle pour la validité de cette approche est que le mécanisme de participation n'est pas informatif, c'est-à-dire $\Pr(\delta_k = 1 \mid \mathbf{x}_k, y_k) = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$. La disponibilité de variables auxiliaires associées à la fois à δ_k et y_k est cruciale pour rendre cette hypothèse plausible et réduire le biais de participation.

On peut ensuite caler le poids de sondage non probabiliste $\hat{w}_k^{\mathrm{NP}}$ (par exemple Deville et Särndal, 1992) pour obtenir des gains d'efficacité plus grands ainsi qu'une propriété de double robustesse (par exemple Chen, Li et Wu, 2020; Valliant, 2020). Le calage du poids de sondage non probabiliste $\hat{w}_k^{\mathrm{NP}}$ peut être particulièrement efficace quand des variables auxiliaires fortement prédictives de y_k sont disponibles, lesquelles ont été exclues de la modélisation de p_k. Nous nous concentrons ensuite sur la modélisation et l'estimation de la probabilité de participation p_k.

3. Estimation de la probabilité de participation au moyen d'un modèle logistique

Le modèle le plus courant pour la probabilité de participation $p_k = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$ est le modèle logistique $p_k(\boldsymbol{\alpha}) = \left[1 + \exp(-\mathbf{x}_k'\boldsymbol{\alpha})\right]^{-1}$, où $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus du modèle. En supposant qu'on observe $\mathbf{x}_k$ pour tous les $k \in U$, et que δ_k sont mutuellement indépendants, on peut trouver un estimateur de $\boldsymbol{\alpha}$ en résolvant l'équation d'estimation sans biais obtenue par la méthode du maximum de vraisemblance :

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum_{k \in U} \left[\delta_k - p_k(\boldsymbol{\alpha})\right] \mathbf{x}_k = \sum_{k \in s_{\mathrm{NP}}} \mathbf{x}_k - \sum_{k \in U} p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}. \tag{3.1}$$

L'estimateur du maximum de vraisemblance qui en résulte est désigné par $\tilde{\boldsymbol{\alpha}}$ et satisfait $\mathbf{U}(\tilde{\boldsymbol{\alpha}}) = \mathbf{0}$. La probabilité de participation estimée est désignée par $\tilde{p}_k = p_k(\tilde{\boldsymbol{\alpha}})$.

L'équation d'estimation (3.1) ne peut pas être utilisée quand le vecteur des variables auxiliaires $\mathbf{x}_k$ est observé seulement pour $k \in s_{\mathrm{NP}}$ et manquant pour $k \in U - s_{\mathrm{NP}}$. Chen, Li et Wu (2020) ont proposé d'estimer $\sum_{k \in U} p_k(\boldsymbol{\alpha})\mathbf{x}_k$ dans (3.1) au moyen d'une enquête probabiliste. L'équation d'estimation du pseudo maximum de vraisemblance qui en résulte est

$$\hat{\mathbf{U}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}, \tag{3.2}$$

où w_k est le poids de sondage probabiliste pour l'unité $k \in s_P$. Par souci de simplicité, nous supposons dans nos développements théoriques que $w_k = \pi_k^{-1}$, où π_k est la probabilité que l'unité k de la population soit sélectionnée dans s_P. Ce poids garantit que $E_d\left[\hat{\mathbf{U}}(\boldsymbol{\alpha})\right] = \mathbf{U}(\boldsymbol{\alpha})$, où l'indice d indique que l'espérance est prise par rapport au plan d'échantillonnage probabiliste. Par conséquent, l'équation d'estimation (3.2) est sans biais par rapport au modèle de participation et au plan d'échantillonnage probabiliste. En pratique, le poids de sondage w_k est souvent obtenu après ajustement du poids de sondage initial, π_k^{-1}, pour tenir compte de la non-réponse et du calage. L'équation d'estimation (3.2) exige de connaître le vecteur $\mathbf{x}_k$ pour toutes les unités $k \in s_{\mathrm{NP}}$ et $k \in s_P$, mais pas pour toutes les unités $k \in U$. Sa solution donne l'estimateur du pseudo maximum de vraisemblance $\hat{\boldsymbol{\alpha}}$, qui satisfait $\hat{\mathbf{U}}(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$. La probabilité de participation estimée qui

en résulte est désignée par $\hat{p}_k = p_k(\hat{\boldsymbol{\alpha}})$. Il convient de mentionner que l'équation d'estimation (3.2) peut ne pas avoir de solution. Cela est plus susceptible de se produire quand n^{NP}/N est grand et que l'échantillon probabiliste est petit (voir Beaumont, 2020). Cela n'a pas posé de problème dans nos expériences, puisque n^{NP}/N était inférieur à 1 %. Beaumont (2020) soutient qu'il serait possible de réduire l'occurrence de solutions inexistantes en remplaçant le modèle logistique par le modèle exponentiel.

Chen, Li et Wu (2020) ont examiné le cas dans lequel les variables auxiliaires sont données. En pratique, il peut être nécessaire de choisir des variables auxiliaires pertinentes et leurs interactions parmi un grand ensemble de variables auxiliaires candidates. Dans les applications que nous avons expérimentées jusqu'à présent, les variables auxiliaires candidates sont souvent catégoriques (par exemple niveau de scolarité, état matrimonial, etc.). Le croisement à l'aveugle de toutes ces variables peut donner un grand nombre de groupes avec de nombreux petits groupes, même vides. C'est ce qui nous a incités à trouver des méthodes permettant de sélectionner les variables auxiliaires pertinentes et leurs interactions.

Nous considérons une procédure de sélection pas à pas qui cherche à minimiser une version modifiée du critère AIC, qui tient adéquatement compte du plan d'échantillonnage probabiliste servant à tirer s_P. La justification de ce critère AIC modifié est fournie dans la section 3.1, et notre procédure de sélection est décrite dans la section 3.2. La section 3.3 permet d'examiner un cas particulier important du modèle logistique : le modèle des groupes homogènes. À la section 3.4, nous proposons une méthode simple fondée sur le rang pour créer les groupes homogènes. Enfin, à la section 3.5, la méthode récente de Wang, Valliant et Li (2021) est abordée et comparée à celle de Chen, Li et Wu (2020).

3.1 Critère d'information d'Akaike modifié pour le modèle logistique qui tient compte du plan d'échantillonnage probabiliste

Prenons d'abord le cas où $\mathbf{x}_k$ est connu pour toutes les unités de population $k \in U$. En supposant que δ_k sont mutuellement indépendants, nous pouvons écrire le logarithme de la fonction de vraisemblance comme suit :

$$\begin{aligned} l(\boldsymbol{\alpha}) &= \sum_{k\in U} \delta_k \log[p_k(\boldsymbol{\alpha})] + (1-\delta_k)\log[1-p_k(\boldsymbol{\alpha})] \\ &= \sum_{k\in s_{\mathrm{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right] + \sum_{k\in U} \log[1-p_k(\boldsymbol{\alpha})]. \end{aligned}$$

Définissons $l_0(\boldsymbol{\alpha}) = E_m[l(\boldsymbol{\alpha})]$, où l'indice m indique que l'espérance est prise par rapport au vrai modèle de participation qui est inconnu. L'estimateur du maximum de vraisemblance $\tilde{\boldsymbol{\alpha}}$ maximise $l(\boldsymbol{\alpha})$ et nous désignons par $\boldsymbol{\alpha}_0$ la valeur de $\boldsymbol{\alpha}$ qui maximise $l_0(\boldsymbol{\alpha})$. Dans des conditions de régularité, l'estimateur du maximum de vraisemblance $\tilde{\boldsymbol{\alpha}}$ est convergent pour $\boldsymbol{\alpha}_0$ selon le modèle, c'est-à-dire $\sqrt{N}\left(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) = O_p(1)$, où N est la taille de la population.

Le critère d'information d'Akaike (AIC) est un estimateur de $-2E_m[l_0(\tilde{\boldsymbol{\alpha}})]$. Il est bien connu qu'un estimateur convergent de $-2E_m[l_0(\tilde{\boldsymbol{\alpha}})]$ est

$$\text{AIC} = -2l(\tilde{\boldsymbol{\alpha}}) + 2q, \tag{3.3}$$

où q est le nombre de paramètres du modèle (ou le nombre de variables auxiliaires). L'équation (3.3) est l'expression originale du critère d'information d'Akaike et la plus répandue en pratique.

Examinons maintenant le cas où $\mathbf{x}_k$ est connu seulement pour $k \in s_{\text{NP}}$ et $k \in s_P$. Chen, Li et Wu (2020) ont proposé le logarithme de la fonction de pseudo vraisemblance

$$\hat{l}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right] + \sum_{k \in s_P} w_k \log[1 - p_k(\boldsymbol{\alpha})]. \tag{3.4}$$

L'utilisation de $w_k = \pi_k^{-1}$ garantit que $E_d[\hat{l}(\boldsymbol{\alpha})] = l(\boldsymbol{\alpha})$ et $E_{md}[\hat{l}(\boldsymbol{\alpha})] = l_0(\boldsymbol{\alpha})$. Dans des conditions de régularité, l'estimateur du pseudo maximum de vraisemblance $\hat{\boldsymbol{\alpha}}$, qui maximise $\hat{l}(\boldsymbol{\alpha})$ dans (3.4), est convergent pour $\boldsymbol{\alpha}_0$ selon le modèle et le plan de sondage, c'est-à-dire $\sqrt{n^P}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = O_p(1)$, où n^P est la taille de l'échantillon probabiliste.

Pour l'estimation par la methode du pseudo maximum de vraisemblance, le AIC peut être défini comme un estimateur de

$$-2E_{md}[l_0(\hat{\boldsymbol{\alpha}})] = -2E_{md}[\hat{l}(\hat{\boldsymbol{\alpha}})] + 2E_{md}[\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}})].$$

Dans l'annexe 1, nous fournissons la preuve que

$$E_{md}[\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}})] \approx q + \text{tr}\left[E_m\{\text{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\}[-\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1}\right], \tag{3.5}$$

où la fonction $\hat{\mathbf{U}}(\boldsymbol{\alpha}) = \partial\hat{l}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ est donnée dans (3.2) pour le modèle logistique, et $\mathbf{H}_0(\boldsymbol{\alpha}) = \partial^2 l_0(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'$. Nos développements suivent étroitement ceux de Lumley et Scott (2015). À partir de (3.5) et (A.3) dans l'annexe 1, un estimateur convergent de $-2E_{md}[l_0(\hat{\boldsymbol{\alpha}})]$ est

$$\text{AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2q + 2\text{tr}\left\{\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)][-\hat{\mathbf{H}}(\hat{\boldsymbol{\alpha}})]^{-1}\right\}, \tag{3.6}$$

où $\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]$ est un estimateur convergent par rapport au plan de $\text{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]$ et $\hat{\mathbf{H}}(\boldsymbol{\alpha}) = \partial^2\hat{l}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'$. Pour le modèle logistique,

$$\hat{\mathbf{H}}(\boldsymbol{\alpha}) = -\sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha})[1 - p_k(\boldsymbol{\alpha})]\mathbf{x}_k\mathbf{x}_k'. \tag{3.7}$$

L'expression du AIC (3.6) est semblable à celle donnée dans Lumley et Scott (2015), mais ils ont omis le terme $2q$. Ce terme est négligeable comparativement au troisième terme du côté droit de (3.6) quand la fraction de sondage n^P/N est négligeable. Cependant, le terme $2q$ peut ne pas être négligeable comparativement au troisième terme de (3.6), même quand n^P/N est de petite taille. Cela a tendance à se produire quand les probabilités de participation $p_k(\boldsymbol{\alpha})$ sont petites, ce qui est habituellement le cas des enquêtes en ligne à participation volontaire, comme les enquêtes par approche participative de Statistique Canada. Par conséquent, le terme $2q$ ne devrait pas être négligé en règle générale, à moins que la taille de

l'échantillon non probabiliste soit considérablement plus grande que la taille de l'échantillon probabiliste. Une autre raison de garder $2q$ dans l'expression (3.6) est qu'elle se réduit à l'expression habituelle du AIC (3.3) quand l'échantillon probabiliste est un recensement. Le dernier terme du côté droit de (3.6) peut donc être interprété comme une pénalité pour avoir utilisé un échantillon probabiliste au lieu d'un recensement complet dans l'équation d'estimation (3.2). Plus l'échantillon probabiliste est petit, plus l'effet de la pénalité est grand sur le AIC (3.6).

3.2 Sélection pas à pas des variables auxiliaires et des interactions simples

Dans la section empirique 6, nous utilisons une procédure pas à pas qui repose sur le AIC (3.6) pour sélectionner les variables auxiliaires (effets principaux) et les interactions simples. Notre procédure commence par le modèle naïf, qui comprend seulement l'ordonnée à l'origine. À chaque étape de la procédure, une variable (effet principal ou interaction simple) est incluse dans le modèle ou, si elle est déjà incluse, est supprimée du modèle. L'inclusion ou la suppression de la variable qui donne la plus grande réduction du AIC (3.6) est sélectionnée. Une interaction est admissible à l'inclusion seulement si les deux effets principaux ont déjà été sélectionnés, et un effet principal est admissible à l'élimination seulement s'il ne soutient aucune interaction. La procédure s'arrête quand aucune variable ne peut être ajoutée ni supprimée du modèle, ce qui signifie qu'aucune autre réduction du AIC (3.6) n'est possible.

L'un des problèmes posés par le choix des variables auxiliaires dans un modèle de participation est qu'il ne tient pas compte des relations entre les variables auxiliaires et les variables d'intérêt. Par conséquent, il se peut qu'une variable auxiliaire qui serait faiblement associée à la participation, mais fortement associée à certaines des variables d'intérêt soit éliminée du modèle de participation final. Cela pourrait avoir un effet négatif sur la réduction du biais de l'estimateur $\hat{\theta}_{\mathrm{NP}}$ du paramètre de population finie θ. Il est donc recommandé de considérer des méthodes de sélection des variables qui tendent vers un surajustement, comme le AIC, afin de réduire le risque d'omettre une variable auxiliaire pertinente. Un surajustement modéré peut permettre de mieux contrôler le biais au détriment d'une augmentation possible de la variance. Notre intention est d'éviter un surajustement exagéré afin de stabiliser $\hat{\theta}_{\mathrm{NP}}$. Comme nous l'avons souligné à la section 2, il est également possible de régler le problème de sélection des variables ci-dessus en calant les poids par l'inverse de la probabilité $\hat{w}_k^{\mathrm{NP}}$ au moyen de variables de calage prédictives des variables d'intérêt.

3.3 Le modèle des groupes homogènes

Considérons une partition de la population U en G groupes, $U_g,\ g=1,...,G,$ et supposons que $s_{\mathrm{NP},g}$ et $s_{P,g}$ sont les ensembles des unités $k \in U_g$ qui font partie respectivement de l'échantillon probabiliste et de l'échantillon non probabiliste. Dans le modèle des groupes homogènes, on suppose que la probabilité de participation est constante pour toutes les unités $k \in U_g,$ c'est-à-dire $p_k \equiv p_g,\ k \in U_g,\ g=1,...,G.$ Le modèle des groupes homogènes peut être considéré comme un cas particulier du modèle logistique pour lequel $q=G,\ \boldsymbol{\alpha}'=(\alpha_1,\ldots,\alpha_g,\ldots,\alpha_G)$ et $\mathbf{x}_k'=(x_{1k},\ldots,x_{gk},\ldots,x_{Gk}),$ où x_{gk} est une variable binaire qui est

égale à 1 si $k \in U_g$ et à 0, autrement. Par conséquent, pour une unité $k \in U_g$, $p_k(\boldsymbol{\alpha}) = p(\alpha_g) \equiv p_g = [1+\exp(-\alpha_g)]^{-1}$, et donc $\alpha_g = \log[p_g/(1-p_g)]$. Pour ce modèle, le logarithme de la fonction de pseudo vraisemblance (3.4) se réduit à

$$\hat{l}(\boldsymbol{\alpha}) = \sum_{g=1}^{G} n_g^{\mathrm{NP}} \log\left[\frac{p(\alpha_g)}{1-p(\alpha_g)}\right] + \hat{N}_g \log[1-p(\alpha_g)], \tag{3.8}$$

où n_g^{NP} est la taille de $s_{\mathrm{NP},g}$ et $\hat{N}_g = \sum_{k \in s_{P,g}} w_k$ est la taille estimée de la population dans le groupe g obtenue à partir de l'échantillon probabiliste. L'estimateur du pseudo maximum de vraisemblance $\hat{\boldsymbol{\alpha}}' = (\hat{\alpha}_1, \ldots, \hat{\alpha}_g, \ldots, \hat{\alpha}_G)$, qui maximise $\hat{l}(\boldsymbol{\alpha})$ dans (3.8), est tel que $\hat{\alpha}_g = \log[\hat{p}_g/(1-\hat{p}_g)]$, $g = 1, \ldots, G$, où

$$\hat{p}_g = \frac{n_g^{\mathrm{NP}}}{\hat{N}_g}. \tag{3.9}$$

À partir de (3.8), nous pouvons écrire $\hat{l}(\hat{\boldsymbol{\alpha}})$ comme étant

$$\hat{l}(\hat{\boldsymbol{\alpha}}) = \sum_{g=1}^{G} \hat{N}_g [\hat{p}_g \log(\hat{p}_g) + (1-\hat{p}_g)\log(1-\hat{p}_g)]. \tag{3.10}$$

Pour le modèle des groupes homogènes, la fonction d'estimation $\hat{\mathbf{U}}(\boldsymbol{\alpha})$ dans (3.2) se réduit à $[\hat{\mathbf{U}}(\boldsymbol{\alpha})]' = [\hat{U}_1(\alpha_1), \ldots, \hat{U}_g(\alpha_g), \ldots, \hat{U}_G(\alpha_G)]$, où

$$\hat{U}_g(\alpha_g) = n_g^{\mathrm{NP}} - \hat{N}_g p(\alpha_g). \tag{3.11}$$

De plus, à partir de (3.7), la matrice $\hat{\mathbf{H}}(\hat{\boldsymbol{\alpha}})$ se réduit à une matrice diagonale, dont le g^{e} élément sur la diagonale est donné par

$$\hat{H}_g(\hat{\alpha}_g) = -\hat{N}_g \hat{p}_g (1-\hat{p}_g). \tag{3.12}$$

Supposons que $\boldsymbol{\alpha}_0' = (\alpha_{0,1}, \ldots, \alpha_{0,g}, \ldots, \alpha_{0,G})$. Au moyen de (3.11) et (3.12), le AIC (3.6) devient

$$\mathrm{AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2G + 2\sum_{g=1}^{G} \frac{\hat{v}_d[\hat{U}_g(\alpha_{0,g})]}{\hat{N}_g \hat{p}_g (1-\hat{p}_g)}, \tag{3.13}$$

où $\hat{v}_d[\hat{U}_g(\alpha_{0,g})]$ est un estimateur convergent par rapport au plan de $\mathrm{var}_d[\hat{U}_g(\alpha_{0,g})]$. Au moyen de (3,11), un estimateur de variance convergent est

$$\hat{v}_d[\hat{U}_g(\alpha_{0,g})] = \hat{p}_g^2 \hat{v}_d(\hat{N}_g), \tag{3.14}$$

où $\hat{v}_d(\hat{N}_g)$ est un estimateur convergent par rapport au plan de $\mathrm{var}_d(\hat{N}_g)$. Au moyen de (3.14), le AIC (3.13) peut être réécrit comme étant

$$\mathrm{AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2G + 2\sum_{g=1}^{G} n_g^{\mathrm{NP}} \frac{[\mathrm{cv}_d(\hat{N}_g)]^2}{1-\hat{p}_g}, \tag{3.15}$$

où $\text{cv}_d(\hat{N}_g) = \sqrt{\hat{v}_d(\hat{N}_g)} \Big/ \hat{N}_g$ est le coefficient de variation estimé de $\hat{N}_g$. Encore une fois, le dernier terme du côté droit de (3.13) ou (3.15) peut être interprété comme une pénalité attribuable à l'estimation des tailles de population inconnues N_g, $g = 1, \ldots, G$, à l'aide d'un échantillon probabiliste.

Au moyen de (3.9), nous obtenons le poids de sondage non probabiliste d'une unité $k \in s_{\text{NP},g}$ comme étant

$$\hat{w}_k^{\text{NP}} = \hat{p}_k^{-1} = \frac{\hat{N}_g}{n_g^{\text{NP}}}. \tag{3.16}$$

Le poids de sondage non probabiliste (3.16) montre l'importance d'éviter les groupes pour lesquels n_g^{NP} est très petit, voire nul, afin de réduire l'occurrence de poids extrêmes. Au moyen de (3.16), l'estimateur pondéré par l'inverse de la probabilité du total de population θ peut s'écrire comme étant

$$\hat{\theta}_{\text{NP}} = \sum\nolimits_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k = \sum_{g=1}^{G} \hat{N}_g \bar{y}_g^{\text{NP}}, \tag{3.17}$$

où $\bar{y}_g^{\text{NP}} = \sum\nolimits_{k \in s_{\text{NP},g}} y_k \Big/ n_g^{\text{NP}}$ est la moyenne de la variable d'intérêt y sur les unités qui sont dans $s_{\text{NP},g}$. L'estimateur (3.17) est simplement un estimateur post-stratifié et satisfait les équations de calage $\sum\nolimits_{k \in s_{\text{NP},g}} \hat{w}_k^{\text{NP}} = \hat{N}_g$, $g = 1, \ldots, G$. Les groupes (post-strates) sont construits pour être homogènes par rapport à l'indicateur de participation. S'ils sont également homogènes par rapport à la variable d'intérêt, alors l'estimateur post-stratifié (3.17) a une propriété de double robustesse (par exemple voir Chen, Li et Wu, 2020; et Valliant, 2020).

Jusqu'à présent, nous avons supposé que l'appartenance à un groupe est prédéterminée pour chaque unité de la population. En pratique, les groupes homogènes sont souvent définis après avoir observé les données de l'échantillon. Il existe plusieurs méthodes de construction de groupes homogènes dépendants de l'échantillon. À la section 3.4, nous proposons une méthode simple fondée sur le rang qui partitionne l'échantillon non probabiliste par rapport aux probabilités de participation estimées à partir d'un modèle logistique. Une extension de l'algorithme CART, nppCART, est décrite dans la section 4. Une fois que l'échantillon probabiliste et l'échantillon non probabiliste ont été partitionnés en groupes dépendants de l'échantillon, les poids peuvent être calculés au moyen de (3.16) comme si l'appartenance au groupe était fixe.

3.4 Méthode fondée sur le rang pour la création des groupes homogènes

La première étape de cette méthode consiste à estimer les probabilités de participation au moyen d'un modèle logistique (avec ou sans sélection pas-à-pas). Nous désignons par $\hat{p}_k^{\text{logistic}} = p_k(\hat{\boldsymbol{\alpha}})$ ces probabilités de participation estimées, qui sont calculées pour chaque $k \in s_{\text{NP}}$ et $k \in s_P$. L'idée est alors de former G groupes qui sont homogènes par rapport à $\hat{p}_k^{\text{logistic}}$ afin de rendre le modèle des groupes homogènes plausible. Une fois les groupes formés, les probabilités estimées $\hat{p}_k^{\text{logistic}}$ sont éliminées et les poids de sondage non probabiliste sont calculés au moyen de (3.16).

De nombreuses méthodes permettent de partitionner s_{NP} en groupes homogènes. Une méthode simple et populaire consiste à former des groupes avec un nombre égal de participants (par exemple Eltinge et Yansaneh, 1997, forment des groupes avec un nombre égal d'unités d'échantillonnage dans le contexte de la non-réponse aux enquêtes). Cette méthode équivaut à déterminer les limites des groupes à partir d'intervalles de largeur égale dans la plage de r_k, $k \in s_{\text{NP}}$, où r_k est le rang de $\hat{p}_k^{\text{logistic}}$. Nous proposons ci-dessous une généralisation de cette méthode qui conserve la simplicité de l'attribution des unités en fonction de leur rang, tout en permettant une certaine souplesse afin que les classes n'aient pas besoin d'être de taille égale.

Plutôt que de faire des classes de largeur égale dans la plage de r_k, nous proposons de former G classes de largeur égale dans la plage de $f(r_k)$, une fonction monotone du rang r_k. Nous appelons cette procédure la méthode Frank. Toutes les unités de l'échantillon non probabiliste qui se trouvent dans une classe donnée sont attribuées au même groupe. Toute fonction non linéaire f donnera donc de plus petits groupes (moins d'unités) lorsque la pente est plus forte, et de plus grands groupes lorsque la pente est plus aplatie. Nous proposons la fonction

$$f(r_k) = \log\left(1 + a\frac{r_k}{n^{\text{NP}}}\right),$$

$k \in s_{\text{NP}}$, où n^{NP} est la taille de l'échantillon non probabiliste et a est une constante prédéfinie non négative qui détermine le degré de non-linéarité. Cette fonction est concave vers le bas, avec une pente plus grande et des groupes plus petits pour les unités dont les rangs sont plus petits. La constante a détermine la taille de cet effet, avec une grande valeur (par exemple $a = 100$) fournissant des groupes de taille plus inégale. La limite quand a tend vers 0 du côté droit rend cette fonction linéaire et donne ainsi des groupes de taille égale. Le rang peut être défini en ordre croissant de $\hat{p}_k^{\text{logistic}}$ ($r_k = 1$ pour le plus petit $\hat{p}_k^{\text{logistic}}$, $r_k = 2$ pour le second plus petit $\hat{p}_k^{\text{logistic}}$, etc.), auquel cas les unités ayant les probabilités estimées les plus petites se trouvent dans les plus petits groupes, ou en ordre décroissant de $\hat{p}_k^{\text{logistic}}$ ($r_k = 1$ pour le plus grand $\hat{p}_k^{\text{logistic}}$, $r_k = 2$ pour le second plus grand $\hat{p}_k^{\text{logistic}}$, etc.), auquel cas les unités ayant les probabilités estimées les plus grandes seront dans les plus petits groupes. La méthode Frank ressemble à la formation de groupes de largeur égale, mais les groupes sont rassemblés vers une extrémité ou l'autre, selon que les $\hat{p}_k^{\text{logistic}}$ sont triés en ordre croissant ou décroissant. La figure A.1(A) de l'annexe 2 illustre la méthode Frank pour $a = 10$, $G = 15$ et $n^{\text{NP}} = 31\,415$, soit la taille de l'échantillon non probabiliste utilisé dans notre étude empirique à la section 6.

Après le fractionnement de l'échantillon non probabiliste en groupes, chaque unité de l'échantillon probabiliste doit être attribuée à l'un des groupes. Étant donné que la fonction f est monotone, chaque groupe contient des unités de l'échantillon non probabiliste avec des valeurs de $\hat{p}_k^{\text{logistic}}$ comprises dans un certain intervalle, et les intervalles de deux groupes différents ne se chevauchant pas de telle sorte que les groupes peuvent être triés en fonction de leur valeur moyenne de $\hat{p}_k^{\text{logistic}}$. La limite entre deux groupes consécutifs est le point milieu entre le plus grand $\hat{p}_k^{\text{logistic}}$ du groupe ayant la plus petite moyenne et le plus petit $\hat{p}_k^{\text{logistic}}$ de l'autre groupe. Une fois toutes les limites déterminées, chaque unité de l'échantillon probabiliste $k \in s_P$ est attribuée au groupe avec les limites qui couvrent $\hat{p}_k^{\text{logistic}}$.

L'application de la méthode Frank exige de déterminer des valeurs appropriées de a et de G ainsi que de trier $\hat{p}_k^{\text{logistic}}$, $k \in s_{\text{NP}}$, en ordre croissant ou décroissant, avant de calculer les rangs r_k. Chaque choix possible donne un ensemble différent de groupes. Nous proposons de déterminer les valeurs de a et G, et l'ordre de tri, en examinant différentes options et en choisissant celle qui donne la plus petite valeur du AIC (3.15). Cette proposition est examinée de façon empirique dans la section 6.3.

3.5 Pondération par la propension logistique ajustée

Comme nous l'avons indiqué dans l'introduction, Wang, Valliant et Li (2021) ont proposé une extension de la méthode de regroupement pour tenir compte d'un chevauchement non négligeable entre l'échantillon probabiliste et l'échantillon non probabiliste. La justification de leur méthode, appelée pondération par la propension logistique ajustée, ne repose pas sur une approche de vraisemblance pure, mais elle produit tout de même une équation d'estimation sans biais par rapport à md donnée par

$$\hat{\mathbf{U}}^{\text{ALP}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \frac{1}{1+p_k^{\text{ALP}}(\boldsymbol{\alpha})}\,\mathbf{x}_k - \sum_{k \in s_P} w_k \frac{p_k^{\text{ALP}}(\boldsymbol{\alpha})}{1+p_k^{\text{ALP}}(\boldsymbol{\alpha})}\,\mathbf{x}_k = \mathbf{0}, \tag{3.18}$$

où $p_k^{\text{ALP}}(\boldsymbol{\alpha}) = \exp(\mathbf{x}_k'\boldsymbol{\alpha})$. L'équation d'estimation (3.18) n'est pas équivalente à (3.2). Cependant, si toutes les probabilités de participation sont petites, les deux équations d'estimation devraient produire des estimations semblables des probabilités de participation.

Une différence importante entre Wang, Valliant et Li (2021) et Chen, Li et Wu (2020) est le choix du modèle de participation. Chen, Li et Wu (2020) modélisent la probabilité de participation au moyen d'une fonction logistique, tandis que Wang, Valliant et Li (2021) considèrent une fonction exponentielle. Le modèle logistique est plus naturel, car il garantit que les probabilités de participation estimées sont toujours comprises dans l'intervalle $(0,1)$. Par contre, le modèle exponentiel peut produire des probabilités estimées supérieures à 1. Wang, Valliant et Li (2021) ont mené une étude par simulation pour évaluer leur méthode. Leurs résultats montrent que (3.18) donne des estimations des moyennes de population qui sont plus robustes face à une mauvaise spécification du modèle que celles de (3.2). Cette robustesse pourrait s'expliquer par l'utilisation du modèle exponentiel.

Pour le modèle des groupes homogènes, nous avons vu à la section 3.3 que la solution de (3.2) donne $p_k(\hat{\boldsymbol{\alpha}}) = \hat{p}_g = n_g^{\text{NP}} \big/ \hat{N}_g$, pour chaque unité $k \in U_g$. On peut facilement montrer que la solution de (3.18) pour le modèle des groupes homogènes donne également $p_k^{\text{ALP}}(\hat{\boldsymbol{\alpha}}) = \hat{p}_g = n_g^{\text{NP}} \big/ \hat{N}_g$, pour chaque unité $k \in U_g$. L'équivalence entre (3.2) et (3.18) pour le modèle des groupes homogènes laisse entendre qu'en général, les deux méthodes peuvent produire des estimations similaires de θ, particulièrement quand les probabilités estimées sont utilisées uniquement dans le but de créer des groupes homogènes (par exemple au moyen de la méthode Frank décrite à la section 3.4).

Wang, Valliant et Li (2021) ont également proposé une version normalisée de leur méthode de pondération par la propension logistique ajustée. Bien que l'équation d'estimation normalisée ne soit plus sans biais par rapport à md, les auteurs ont montré son efficacité dans une étude par simulation dans le cas

de l'estimation de moyennes de population. Nous avons évaulé la méthode de pondération par la propension logistique ajustée, y compris sa version normalisée, dans nos expériences empiriques. Les estimations obtenues (non incluses) étaient proches des estimations du pseudo maximum de vraisemblance de Chen, Li et Wu (2020), particulièrement après la création de groupes homogènes. Cette observation n'est pas surprenante si l'on considère que la taille de l'échantillon non probabiliste est inférieure à 1 % de la taille de la population dans nos expériences et que les probabilités de participation estimées ont tendance à être assez petites. D'autres travaux de recherche sont nécessaires afin d'effectuer une comparaison plus en profondeur de la méthode de pondération par la propension logistique ajustée et de la méthode du pseudo maximum de vraisemblance.

Un des objectifs de la présente étude était d'élaborer une procédure de sélection des variables applicable au scénario d'intégration des données décrit à la section 2. Wang, Valliant et Li (2021) n'ont pas abordé le problème de la sélection des variables. Un critère d'information d'Akaike fondé sur Lumley et Scott (2015) ne convient pas avec la méthode de pondération par la propension logistique ajustée (ni sa version normalisée) parce que l'équation d'estimation sous-jacente n'est pas justifiée par une approche de vraisemblance pure. Toutefois, si la propension logistique ajustée était préférable dans un contexte donné, la sélection des variables pourrait d'abord reposer sur la méthode de la pseudo vraisemblance de Chen, Li et Wu (2020), puis la pondération par la propension logistique ajustée pourrait ensuite être appliquée au moyen des variables auxiliaires sélectionnées.

4. Estimation de la probabilité de participation au moyen de nppCART

La procédure de croissance des arbres CART, élaborée par Breiman, Friedman, Olshen et Stone (1984), est un algorithme de partitionnement récursif binaire qui minimise une certaine fonction objectif. Pour une variable dépendante binaire comme δ_k, l'impureté de l'entropie est une fonction objectif appropriée. Pour une partition donnée, U_g, $g = 1, \ldots, G$, l'impureté de l'entropie est donnée par

$$I = -\sum_{g=1}^{G} \frac{N_g}{N} [\tilde{p}_g \log(\tilde{p}_g) + (1-\tilde{p}_g) \log(1-\tilde{p}_g)],$$

où N_g est la taille de U_g, $N = \sum_{g=1}^{G} N_g$ et $\tilde{p}_g = n_g^{\mathrm{NP}} \big/ N_g$. L'impureté de l'entropie ne peut pas être calculée quand N_g est inconnu. Nous proposons de remplacer N_g par l'estimateur pondéré par les poids de sondage $\hat{N}_g$. Cela donne la fonction objectif calculable

$$\hat{I} = -\sum_{g=1}^{G} \frac{\hat{N}_g}{\hat{N}} [\hat{p}_g \log(\hat{p}_g) + (1-\hat{p}_g) \log(1-\hat{p}_g)], \tag{4.1}$$

où $\hat{p}_g$ est donné dans (3.9) et $\hat{N} = \sum_{g=1}^{G} \hat{N}_g$. L'impureté de l'entropie estimée (4.1) est proportionnelle au logarithme de la fonction de pseudo vraisemblance (3.10) pour le modèle des groupes homogènes puisque $\hat{I} = -\hat{l}(\hat{\boldsymbol{\alpha}}) \big/ \hat{N}$.

L'algorithme de partitionnement récursif binaire commence par examiner tous les fractionnements possibles de l'échantillon non probabiliste s_{NP} en deux groupes. Un fractionnement désigne toute partition binaire de s_{NP} qui repose sur les catégories ou les valeurs numériques d'une des variables auxiliaires candidates. Par exemple, un fractionnement pourrait être « SEXE = homme » et « SEXE = femme » ou « ÂGE < 25 » et « ÂGE ≥ 25 ». Pour chaque fractionnement de s_{NP}, l'échantillon probabiliste s_P est également fractionné au moyen de la même partition binaire. Un fractionnement est considéré comme inadmissible et est rejeté s'il satisfait à l'un des trois critères d'arrêt suivants :

i) $n_g^{\text{NP}} < C_{\text{NP}}$, pour $g = 1$ ou $g = 2$, où $C_{\text{NP}} \geq 1$ est une constante prédéterminée précisant le nombre minimum de participants dans un groupe;

ii) $n_g^{\text{NP}} \geq \hat{N}_g$, pour $g = 1$ ou $g = 2$;

iii) $n_g^P < C_P$, pour $g = 1$ ou $g = 2$, où n_g^P est la taille de $s_{P,g}$ et $C_P \geq 1$ est une constante prédéterminée précisant le nombre minimum d'unités de l'échantillon probabiliste dans un groupe.

Ensuite, l'impureté de l'entropie estimée (4.1) avec $G = 2$ est calculée pour chaque fractionnement admissible, et le meilleur de ces fractionnements admissibles, c'est-à-dire celui qui a la plus petite valeur de (4.1), est sélectionné afin de former les deux premiers groupes. Si tous les fractionnements sont inadmissibles ou si le meilleur fractionnement ne diminue pas la fonction objectif (4.1), alors le partitionnement en deux groupes n'est pas effectué.

Après la détermination des deux premiers groupes initiaux, la même opération de fractionnement est répétée pour chacun des deux groupes, et ainsi de suite, couche par couche, jusqu'à ce que tous les groupes ne puissent plus être fractionnés selon les critères d'arrêt. Nous disons que ce processus donne un arbre à pleine maturité, bien que ce soit un léger abus de langage, puisque des critères d'arrêt limitent sa croissance. La procédure ci-dessus, dont la version précédente a été appelée nppCART par Chu et Beaumont (2019), est essentiellement identique à l'algorithme CART d'origine, à l'exception de l'utilisation de l'entropie estimée (4.1) et des trois critères d'arrêt ci-dessus. Le critère d'arrêt (i) permet de s'assurer que le poids de sondage non probabiliste $\hat{w}_k^{\text{NP}}$ dans (3.16) ne devienne pas extrême. Le critère d'arrêt (ii) permet de s'assurer que la probabilité estimée $\hat{p}_g$ soit toujours inférieure à 1. Le dernier critère est ajouté pour s'assurer que l'estimateur $\hat{N}_g$ ne soit pas trop instable.

Chu et Beaumont (2019) ont élaboré un programme R qui met en œuvre l'algorithme nppCART. Ils ont montré dans une étude par simulation que cet algorithme était efficace pour réduire le biais de participation, bien que l'estimateur post-stratifié qui en résultait (3.17) ait une variance un peu plus grande que ses concurrents. Cette instabilité peut s'expliquer par un surajustement, c'est-à-dire la création d'un trop grand nombre de groupes. La recommandation habituelle pour éviter le surajustement est d'élaguer l'arbre après sa croissance. Habituellement, l'élagage se fait en deux étapes. Dans la première étape, on détermine une séquence finie de sous-arbres emboîtés de taille décroissante et d'impureté croissante, en commençant par l'arbre mature qui comprend le nombre maximal de groupes et en terminant par le sous-arbre dégénéré qui

contient seulement un groupe. À la deuxième étape, le meilleur de ces sous-arbres emboîtés est sélectionné, souvent par validation croisée à K blocs. Cette approche d'élagage équivaut à pénaliser la fonction objectif en ajoutant un terme de pénalité défini comme le produit d'un paramètre de pénalité positif et du nombre de groupes. La validation croisée est ensuite généralement utilisée pour déterminer une valeur optimale pour le paramètre de pénalité. On trouve plus de détails sur l'élagage dans Breiman, Friedman, Olshen et Stone (1984) et également dans Izenman (2008, chapitre 9). Dans le contexte de la non-réponse aux enquêtes, les arbres de classification et de régression ont été étudiés par Phipps et Toth (2012) et Lohr, Hsu et Montaquila (2015).

Cependant, comme souligné dans l'introduction, les méthodes classiques de validation croisée ne peuvent pas être appliquées directement au scénario d'intégration des données étudié dans le présent article, et ce sujet nécessite d'autres recherches. Comme solution de rechange à la validation croisée afin de sélectionner le meilleur sous-arbre, dans un ensemble de sous-arbres emboîtés de taille décroissante et d'impureté croissante, nous proposons de choisir le sous-arbre qui minimise le AIC (3.15). Ce AIC tient compte du plan d'échantillonnage probabiliste par l'entremise de l'estimation de la variance de $\hat{N}_g$ par rapport au plan (voir la section 5). Cette variance pouvait être facilement estimée dans nos expériences à la section 6 au moyen des poids bootstrap disponibles. Tout comme la sélection de variables, dont il est question à la section 3.2, l'élagage vise à éviter les surajustements excessifs afin de stabiliser $\hat{\theta}_{\text{NP}}$.

5. Estimation de la variance bootstrap

Il ne suffit pas de produire des estimations pondérées par l'inverse de la probabilité des paramètres de population finie, il est également important de fournir aux utilisateurs des indicateurs de la qualité de ces estimations. Nous proposons une procédure bootstrap pour estimer la variance des estimateurs pondérés par l'inverse de la probabilité en mettant l'accent sur l'estimateur post-stratifié (3.17). La variance peut être utile, mais elle présente certaines limites puisqu'elle est calculée sous l'hypothèse selon laquelle le modèle de participation est correctement spécifié et les estimateurs pondérés par l'inverse de la probabilité sont sans biais. L'absence de biais dépend de façon critique de la disponibilité et du choix approprié des variables auxiliaires afin de rendre l'hypothèse de participation non informative raisonnable. Bien qu'une certaine quantité de biais semble inévitable en pratique, le calcul des estimations de la variance peut néanmoins fournir des renseignements utiles à des fins de comparaison et d'évaluation, comme l'illustre la section 6.

L'estimateur de la variance bootstrap que nous proposons tient compte de deux sources de variabilité : le plan d'échantillonnage probabiliste et le modèle de participation. Nous supposons que B poids bootstrap $w_k^{(b)}, \ b = 1, \ldots, B,$ sont disponibles pour chaque unité $k \in s_P,$ et que ces poids bootstrap capturent correctement de la variabilité due au plan d'échantillonnage probabiliste. Par exemple, nous supposons que ces poids bootstrap peuvent servir à obtenir un estimateur convergent par rapport au plan de $\text{var}_d(\hat{N}_g)$ comme suit :

$$\hat{v}_d^{\text{boot}}(\hat{N}_g) = \frac{1}{B}\sum_{b=1}^{B}(\hat{N}_g^{(b)} - \hat{N}_g)^2, \tag{5.1}$$

où $\hat{N}_g^{(b)} = \sum_{k\in s_{P,g}} w_k^{(b)}$ est la b^{e} réplique bootstrap de $\hat{N}_g$. Les poids bootstrap de Rao, Wu et Yue (1992) sont souvent utilisés dans les enquêtes sociales menées par Statistique Canada. Ils sont valides pour des plans d'échantillonnage stratifiés à plusieurs degrés quand les fractions de sondage au premier degré sont petites et peuvent intégrer des ajustements de poids, comme des ajustements pour la non-réponse et le calage. Beaumont et Émond (2022) ont proposé une extension de la méthode qui n'exige pas de petites fractions de sondage au premier degré.

Le mécanisme de participation inconnu est modélisé selon un plan d'échantillonnage de Poisson, où l'on suppose que les unités de la population participent indépendamment les unes des autres avec probabilités p_k, $k \in U$. Pour l'échantillonnage de Poisson, Beaumont et Patak (2012) ont souligné que des poids bootstrap valides pour les unités de l'échantillon $k \in s_{\text{NP}}$ peuvent être écrits comme étant $p_k^{-1}a_k^{(b)}$, $b = 1, \ldots, B$, à condition que les facteurs bootstrap $a_k^{(b)}$ soient générés indépendamment les uns des autres au moyen d'une distribution qui n'est pas lourdement asymétrique avec une moyenne de un et une variance de $1 - p_k$. Pour un échantillon non probabiliste, la probabilité de participation réelle p_k est inconnue, mais elle peut être remplacée par un estimateur convergent $\hat{p}_k$. Suivant Beaumont et Émond (2022), qui ont étudié le bootstrap en présence de non-réponse aux enquêtes, nous proposons donc de générer les facteurs bootstrap $a_k^{(b)}$, $k \in s_{\text{NP}}$ et $b = 1, \ldots, B$, indépendamment les uns des autres en utilisant la distribution gamma avec une moyenne de un et une variance de $1 - \hat{p}_k$. Le choix de la distribution gamma vise à garantir des facteurs bootstrap $a_k^{(b)}$ qui ne sont pas négatifs.

L'estimateur bootstrap de la variance de l'estimateur pondéré par l'inverse de la probabilité $\hat{\theta}_{\text{NP}}$, $\text{var}_{md}(\hat{\theta}_{\text{NP}})$, est donné par

$$\hat{v}_{md}^{\text{boot}}(\hat{\theta}_{\text{NP}}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_{\text{NP}}^{(b)} - \hat{\theta}_{\text{NP}})^2, \tag{5.2}$$

où $\hat{\theta}_{\text{NP}}^{(b)}$ est la b^{e} réplique bootstrap de $\hat{\theta}_{\text{NP}}$. En supposant que le modèle logistique est utilisé avec des variables auxiliaires fixes, la b^{e} réplique bootstrap de $\hat{\theta}_{\text{NP}} = \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k$, avec $\hat{w}_k^{\text{NP}} = [p_k(\hat{\boldsymbol{\alpha}})]^{-1}$, est $\hat{\theta}_{\text{NP}}^{(b)} = \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)} y_k$, où $\hat{w}_k^{\text{NP},(b)} = a_k^{(b)} / p_k(\hat{\boldsymbol{\alpha}}^{(b)})$, et $\hat{\boldsymbol{\alpha}}^{(b)}$ est la solution de la b^{e} réplique bootstrap de l'équation d'estimation (3.2) :

$$\hat{\mathbf{U}}^{(b)}(\boldsymbol{\alpha}) = \sum_{k\in s_{\text{NP}}} a_k^{(b)} \mathbf{x}_k - \sum_{k\in s_P} w_k^{(b)} p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}.$$

En supposant maintenant que le modèle des groupes homogènes est utilisé, la b^{e} réplique bootstrap de l'estimateur post-stratifié (3.17) peut s'écrire comme suit :

$$\hat{\theta}_{\text{NP}}^{(b)} = \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)} y_k = \sum_{g=1}^{G} \hat{N}_g^{(b)} \overline{y}_g^{\text{NP},(b)}, \tag{5.3}$$

où $\hat{w}_k^{\mathrm{NP},(b)} = a_k^{(b)} \hat{N}_g^{(b)} \big/ n_g^{\mathrm{NP},(b)}$, pour $k \in s_{\mathrm{NP},g}$, $n_g^{\mathrm{NP},(b)} = \sum_{k \in s_{\mathrm{NP},g}} a_k^{(b)}$ et $\bar{y}_g^{\mathrm{NP},(b)} = \sum_{k \in s_{\mathrm{NP},g}} a_k^{(b)} y_k \big/ n_g^{\mathrm{NP},(b)}$. La réplique bootstrap (5.3) est valide à condition que les groupes homogènes soient fixes. Cette simplification est souvent effectuée lors de l'estimation de la variance des estimateurs ajustés pour la non-réponse aux enquêtes, même quand les groupes homogènes sont déterminés de façon adaptative à partir des données observées de l'échantillon. Dans notre contexte, il ne serait pas simple d'élaborer une procédure bootstrap qui tient compte correctement de l'élagage ou de la sélection de variables. En particulier, un bootstrap double pourrait être nécessaire si les estimateurs de la variance par rapport au plan employés dans le AIC (3.6) ou (3.15) étaient obtenus au moyen de poids bootstrap. Traiter les variables auxiliaires ou les groupes homogènes comme étant fixes alors qu'ils ne le sont pas devrait tendre à sous-estimer la variance $\mathrm{var}_{md}(\hat{\theta}_{\mathrm{NP}})$. Bien que l'on s'attende à ce que l'ampleur de la sous-estimation soit petite ou modérée, d'autres recherches sont nécessaires sur ce sujet.

6. Évaluation empirique de méthodes au moyen de données réelles

Nous avons évalué et comparé les méthodes de pondération par l'inverse de la probabilité, abordées aux sections 3 et 4, au moyen de données réelles. Dans la section 6.1, nous présentons les trois sources de données utilisées dans nos recherches. Les méthodes sont décrites dans la section 6.2 et les résultats sont présentés dans les sections 6.3 et 6.4.

6.1 Sources de données et variables

Après le début du confinement lié à la COVID-19 en mars 2020, Statistique Canada a mené une série d'enquêtes par approche participative pour répondre aux besoins urgents en information sur la vie de la population canadienne. Chaque enquête par approche participative a permis de recueillir des données auprès des personnes qui ont consulté le site Web de Statistique Canada et qui ont répondu volontairement à un court questionnaire en ligne. Renaud et Beaumont (2020) présentent plus en détail les expériences d'approche participative menées par Statistique Canada.

Nous avons examiné l'utilisation de l'Enquête sur la population active (EPA) comme moyen de réduire le biais de participation des estimations par approche participative. À l'exception du recensement, l'EPA est l'enquête sociale probabiliste la plus importante menée par Statistique Canada, comprenant un échantillon d'environ 56 000 ménages sélectionnés chaque mois. Des données sont recueillies pour toutes les personnes admissibles au sein des ménages répondants. Le taux de réponse des ménages était d'environ 90 % avant la pandémie, mais il est tombé à environ 70 % en juin 2020. Dans notre étude empirique, nous avons utilisé les données de l'échantillon de juin 2020 de l'EPA, qui contient les réponses de 87 779 personnes. L'EPA repose sur un plan de sondage stratifié à plusieurs degrés et un estimateur composite de régression (voir Gambino, Kennedy et Singh, 2001). Des poids bootstrap de Rao, Wu et Yue (1992) sont produits et mis à la disposition des utilisateurs aux fins d'estimation de la variance.

Parallèlement aux expériences d'approche participative, Statistique Canada a également lancé une série d'enquêtes probabilistes par panel en ligne : la Série d'enquêtes sur les perspectives canadiennes (SEPC). L'échantillon de la SEPC provient de personnes ayant déjà répondu à l'EPA. L'échantillon probabiliste initial de juin 2020 de la SEPC était relativement grand, plus de 30 000 personnes ayant été sélectionnées, mais le taux de recrutement/réponse global était assez faible, avoisinant les 15 %, ce qui s'est traduit par 4 209 répondants en juin 2020. La SEPC est présentée de façon détaillée dans Baribeau (2020).

En juin 2020, des participants d'expériences d'approche participative réalisées précédemment ont également été choisis au hasard et ont reçu le même questionnaire que les répondants de la SEPC; 31 415 participants ont répondu au questionnaire. Cela a permis de comparer les estimations de cet échantillon non probabiliste par approche participative à celles de l'échantillon probabiliste de la SEPC.

Le tableau 6.1 présente les estimations naïves de l'approche participative et les estimations de la SEPC pour neuf proportions sélectionnées. Pour les deux premières proportions, les estimations de l'EPA sont également disponibles et sont très proches des estimations correspondantes de la SEPC. Cela n'est pas inattendu, car la non-réponse dans la SEPC est ajustée en fonction du niveau de scolarité et de la situation d'emploi. Les deux enquêtes probabilistes montrent de grandes différences en comparaison aux estimations naïves de l'approche participative pour ces deux proportions. Les cinq proportions suivantes montrent également des différences significatives entre les estimations naïves de l'approche participative et les estimations de la SEPC, tandis que les estimations des deux sources sont semblables pour les deux dernières proportions.

Tableau 6.1
Proportions d'intérêt.

Proportion	Description	Estimation naïve de l'approche participative	Estimation de la SEPC	Estimation de l'EPA
θ_1	Proportion de personnes titulaires d'un diplôme universitaire.	64,5 %	30,6 %	30,2 %
θ_2	Proportion de personnes qui occupaient un emploi rémunéré ou travaillaient dans une entreprise pendant la semaine de référence.	65,4 %	50,1 %	50,3 %
θ_3	Proportion de personnes dont le lieu de travail habituel est un lieu fixe hors de leur domicile.	50,2 %	40,2 %	-
θ_4	Proportion de personnes qui ont travaillé la plupart de leurs heures à leur domicile pendant la semaine de référence.	45,6 %	19,3 %	-
θ_5	Proportion de personnes qui déclarent avoir un revenu « plus que suffisant » pour répondre aux besoins de leur ménage.	32,1 %	15,9 %	-
θ_6	Proportion de personnes qui sont « très susceptibles » de se faire vacciner contre la COVID-19 lorsque le vaccin sera disponible.	74,2 %	57,3 %	-
θ_7	Proportion de personnes qui sont « très préoccupées » par le risque pour la santé posé par les rassemblements en grands groupes.	70,0 %	54,4 %	-
θ_8	Proportion de personnes qui « craignent d'être une cible parce qu'elles exposent les autres à un risque » parce qu'elles ne portent pas toujours un masque en public.	9,9 %	9,8 %	-
θ_9	Proportion de personnes qui déclarent avoir commandé la même quantité de mets à emporter qu'auparavant.	45,6 %	46,2 %	-

Dans un premier temps, nous avons utilisé les données de l'EPA de juin 2020 pour réduire le biais de participation des estimations naïves de l'approche participative en utilisant les méthodes de pondération par l'inverse de la probabilité traitées dans les sections 3 et 4. Les variables auxiliaires candidates disponibles dans les échantillons de l'approche participative et de l'EPA étaient le groupe d'âge (13 niveaux), le sexe (2 niveaux), la région économique (56 niveaux), le niveau de scolarité (8 niveaux), le statut d'immigrant (3 niveaux), la taille du ménage (6 niveaux), l'état matrimonial (6 niveaux) et la situation d'emploi (3 niveaux). L'annexe 3 comprend davantage de renseignements sur ces huit variables auxiliaires. Ensuite, nous avons utilisé des poids de sondage non probabiliste pour calculer des estimations ajustées de l'approche participative pour les neuf proportions définies au tableau 6.1 et nous les avons comparées à celles obtenues uniquement au moyen de l'échantillon probabiliste de la SEPC. Ces résultats sont fournis dans la section 6.3. Il convient de mentionner qu'une proportion est définie comme étant $\theta = N^{-1}\sum_{k\in U} y_k$, où y_k est une variable binaire d'intérêt, et est estimée par $\hat{\theta}_{\text{NP}} = \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k \Big/ \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP}}$. Pour les deux premières proportions du tableau 6.1, la variable d'intérêt y_k peut être dérivée à partir de variables auxiliaires. Nous nous attendons donc à ce que les méthodes de pondération réussissent à éliminer le biais de participation pour ces proportions.

Dans une deuxième étape, nous avons obtenu des estimations ajustées de l'approche participative en utilisant les données de la SEPC de juin 2020 plutôt que les données de l'EPA en utilisant les mêmes variables auxiliaires candidates que ci-dessus. Notre objectif était d'évaluer l'effet de l'utilisation d'un plus petit échantillon probabiliste sur la réduction du biais. Ces résultats sont fournis dans la section 6.4.

6.2 Méthodes

Nous avons évalué les huit méthodes décrites dans le tableau 6.2 ci-dessous. Pour les méthodes 3, 5 et 6, qui font intervenir un modèle logistique utilisant la procédure de sélection pas à pas décrite à la section 3.2, tous les effets principaux et les interactions simples ont été considérés comme variables candidates à inclure ou à supprimer du modèle. Pour ces méthodes, l'estimateur $\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]$ nécessaire pour calculer le AIC (3.6), a été obtenu au moyen des poids bootstrap comme suit :

$$\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)] = \frac{1}{B}\sum_{b=1}^{B} [\hat{\mathbf{U}}^{*(b)}(\hat{\boldsymbol{\alpha}})]\,[\hat{\mathbf{U}}^{*(b)}(\hat{\boldsymbol{\alpha}})]',$$

où

$$\hat{\mathbf{U}}^{*(b)}(\hat{\boldsymbol{\alpha}}) = \sum_{k\in s_{\text{NP}}} \mathbf{x}_k - \sum_{k\in s_P} w_k^{(b)} p_k(\hat{\boldsymbol{\alpha}})\,\mathbf{x}_k.$$

Pour les méthodes 4, 5, 6 et 8, l'estimateur $\hat{v}_d(\hat{N}_g)$, nécessaire pour calculer le AIC (3.15), est obtenu à partir de (5.1). Pour les méthodes 6, 7 et 8, qui utilisent nppCART, nous avons spécifié les valeurs $C_{\text{NP}} = 5$ et $C_P = 5$ dans les critères d'arrêt (i) et (iii) présentés à la section 4.

Tableau 6.2
Description des méthodes.

Méthode	Modèle	Sélection pas à pas	Groupes homogènes	Description
1	Ordonnée à l'origine	-	-	Modèle logistique naïf avec seulement l'ordonnée à l'origine (ou modèle des groupes homogènes avec un seul groupe).
2	Logistique	-	-	Modèle logistique comprenant tous les effets principaux, mais aucune interaction.
3	Logistique	Oui	-	Modèle logistique avec sélection pas à pas des effets principaux et des interactions simples au moyen de la minimisation du AIC (3.6).
4	Logistique	-	Frank	Méthode 2 suivie de la création de groupes homogènes au moyen de la méthode Frank, décrite à la section 3.4, avec tri en ordre croissant, $a = 10$ et le nombre de groupes minimisant approximativement le AIC (3.15).
5	Logistique	Oui	Frank	Méthode 3 suivie de la création de groupes homogènes au moyen de la méthode Frank, décrite à la section 3.4, avec tri en ordre croissant, $a = 10$ et le nombre de groupes minimisant approximativement le AIC (3.15).
6	Logistique	Oui	nppCART avec élagage	Méthode 3 suivie de la création de groupes homogènes au moyen de nppCART avec élagage minimisant le AIC (3.15); une seule variable auxiliaire est fournie à nppCART : la probabilité de participation estimée à partir du modèle logistique.
7	-	-	nppCART sans élagage	nppCART fondé sur toutes les variables auxiliaires candidates sans élagage.
8	-	-	nppCART avec élagage	nppCART fondé sur toutes les variables auxiliaires candidates avec élagage minimisant le AIC (3.15).

6.3 Résultats de l'intégration des données de l'approche participative à l'échantillon probabiliste de l'EPA

Résultats de la sélection pas à pas pour le modèle logistique

En utilisant l'EPA comme échantillon probabiliste, notre procédure de sélection pas à pas, décrite à la section 3.2, a abouti à la sélection de tous les effets principaux de même que 15 interactions simples pour un total de 395 paramètres de modèle. Six effets principaux sont entrés dans le modèle avant la première interaction, selon l'ordre suivant : niveau de scolarité, région économique, statut d'immigrant, sexe, groupe d'âge et taille du ménage. Ensemble, ces six effets principaux représentent plus de 95 % de la réduction totale du AIC (différence entre le AIC des méthodes 1 et 3). À elle seule, la variable du niveau de scolarité représente plus de 40 % de la réduction totale du AIC. Pour ces données, il semble donc que les interactions ne soient pas aussi importantes que les effets principaux dans la réduction du AIC. Cela suggère qu'un modèle incluant tous les effets principaux, mais aucune interaction, pourrait être raisonnable.

Comparaison des valeurs du critère d'information d'Akaike

Le tableau 6.3 montre les valeurs du critère d'information d'Akaike relatif (RAIC) pour les huit méthodes décrites dans le tableau 6.2. Le AIC relatif est défini comme étant

$$\text{RAIC} = \frac{\text{AIC}_0 - \text{AIC}}{\text{AIC}_0} \times 100\ \%,$$

où AIC_0 est la valeur du AIC (3.6) pour le modèle naïf comprenant seulement l'ordonnée à l'origine. Pour les méthodes 1, 2 et 3, le AIC relatif est calculé au moyen du AIC (3.6), tandis qu'on le calcule au moyen du AIC (3.15) pour les méthodes 4 à 8 en supposant que les groupes sont fixes. Le AIC relatif peut être interprété de la même façon que le coefficient de détermination pour une régression linéaire : il est de 0 pour le modèle naïf, il augmente quand le AIC diminue et il est toujours inférieur à 1. Il peut toutefois prendre des valeurs négatives contrairement au coefficient de détermination. Si le AIC relatif d'un modèle est plus grand que celui d'un modèle concurrent, cela indique que ses variables auxiliaires sont de meilleurs prédicteurs de la participation. Le tableau 6.3 montre également le nombre de paramètres q du modèle ou le nombre de groupes $G; q$ est indiqué pour les méthodes 1, 2 et 3 et G est indiqué pour les méthodes 4 à 8.

Tableau 6.3
Valeurs du AIC relatif en pourcentage.

Méthode	Modèle	Sélection pas à pas	Groupes homogènes	AIC relatif (%)	q ou G	Proportion (%) du AIC issue du 1er terme	Proportion (%) du AIC issue du 2e terme	Proportion (%) du AIC issue du 3e terme
1	Ordonnée à l'origine	-	-	0	1	100,00	0,00	0,00
2	Logistique	-	-	10,7	90	99,90	0,04	0,06
3	Logistique	Oui	-	11,1	395	99,59	0,18	0,23
4	Logistique	-	Frank	10,7	100	99,89	0,05	0,07
5	Logistique	Oui	Frank	11,3	100	99,88	0,05	0,07
6	Logistique	Oui	nppCART avec élagage	12,2	1 276	97,99	0,59	1,42
7	-	-	nppCART sans élagage	11,9	3 165	96,23	1,45	2,33
8	-	-	nppCART avec élagage	12,5	1 772	97,58	0,82	1,60

Le AIC relatif varie de 10,7 % à 12,5 % pour les méthodes 2 à 8. Toutes ces méthodes offrent donc une amélioration significative par rapport à la méthode naïve. En comparant les méthodes 2 et 3, nous constatons que la prise en compte des interactions simples n'a produit qu'une légère amélioration du AIC relatif, comme mentionné ci-dessus. L'utilisation de la méthode Frank pour créer des groupes homogènes n'a pas amélioré considérablement le AIC relatif. Cela indique que le modèle logistique était raisonnable pour ces données. L'utilisation de l'algorithme nppCART a entraîné une augmentation du AIC relatif, bien que non considérable. Cela pourrait indiquer que nppCART offre une certaine robustesse. Cependant, nppCART a également donné lieu à un nombre de groupes beaucoup plus grands que d'autres méthodes, même après l'élagage. Étant donné que le AIC (3.15) suppose que les groupes sont fixes (bien qu'ils ne le soient pas), cette amélioration du AIC relatif ne doit pas être surinterprétée.

Le tableau 6.3 montre aussi la proportion du AIC qui provient de chacun des trois termes du côté droit de (3.6) ou (3.15). Sans surprise, le premier terme, $-2\hat{l}(\hat{\boldsymbol{\alpha}})$, est la composante dominante du AIC. L'importance relative des deux autres termes augmente avec q ou G. Ces deux termes ont une importance semblable, bien que le troisième terme soit toujours légèrement plus grand que le deuxième. Dans cette application, aucun des termes ne devrait être omis dans le calcul du AIC.

La méthode Frank

Les figures A.1(B) et A.1(C) de l'annexe 2 illustrent la méthode Frank pour la création des groupes homogènes pour la méthode 5 du tableau 6.2. La figure A.1(B) présente un graphique de $\hat{p}_k^{\text{logistic}}$ en fonction du rang r_k pour l'échantillon probabiliste et l'échantillon non probabiliste. Il montre également les limites correspondantes, par rapport aux rangs, pour $G = 15$ et différentes valeurs de a, et pour les deux ordres de tri. La figure A.1(B) montre que les groupes contenant de plus petites valeurs de $\hat{p}_k^{\text{logistic}}$ sont sous-représentés dans l'échantillon non probabiliste, comparativement à l'échantillon probabiliste, parce que ces unités sont moins susceptibles de participer. La figure A.1(B) illustre également que le tri en ordre croissant produit des groupes qui sont plus près d'être de taille égale dans l'échantillon probabiliste, particulièrement quand a est grand. Cela a l'avantage de réduire l'occurrence de groupes qui contiennent trop peu d'unités dans l'échantillon probabiliste, ce qui pourrait entraîner des poids instables. Une valeur de $a = 5$ ou $a = 10$, ainsi qu'un tri en ordre croissant, semble offrir un compromis approprié pour les deux échantillons.

La figure A.1(C) montre les valeurs du AIC (3.15) en fonction du nombre de groupes G pour quelques valeurs de a et les deux ordres de tri. Il semble que l'ordre de tri entraîne une différence significative sur le AIC et que des valeurs inférieures soeient obtenues quand $\hat{p}_k^{\text{logistic}}$, $k \in s_{\text{NP}}$, sont triés en ordre croissant. La figure A.1(C) ne montre pas une grande sensibilité au choix de a mais les meilleures valeurs semblent se situer près de $a = 10$. On note que le nombre optimal de classes est proche de 100 dans cette application, ce qui est beaucoup plus élevé que la valeur de 5 souvent recommandée (par exemple Eltinge et Yansaneh, 1997). En se fondant sur ces résultats, nous avons choisi un tri par ordre croissant et nous avons utilisé $a = 10$ et $G = 100$ quand la méthode Frank a été appliquée avec les données de l'EPA. Nous avons choisi un plus petit nombre de groupes avec les données de la SEPC (voir la section 6.4).

Avec ces données, la formation de groupes comptant un nombre égal de participants $(a = 0)$ était légèrement inférieure à $a = 10$ pour ce qui est du AIC (voir la figure A.1(C)). Cependant, les deux valeurs de a ont donné des estimations similaires (résultats non présentés).

Comparaisons des estimations

Le tableau 6.4 présente les estimations et leurs erreurs-types bootstrap (en italique) pour chacune des neuf proportions du tableau 6.1 et pour chaque méthode décrite au tableau 6.2. L'erreur-type bootstrap est la racine carrée de l'estimation de la variance bootstrap donnée dans (5.2). La b^{e} réplique bootstrap de la proportion estimée $\hat{\theta}_{\text{NP}} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k \Big/ \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}}$ est $\hat{\theta}_{\text{NP}}^{(b)} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)} y_k \Big/ \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)}$. Pour les méthodes 4 à 8, les poids bootstrap $\hat{w}_k^{\text{NP},(b)}$ sont obtenus en vertu de la simplification selon laquelle les groupes homogènes sont fixes. Les erreurs-types bootstrap ne sont pas calculées pour les méthodes 2 et 3. Les estimations de la SEPC et leurs erreurs-types fondées sur le plan sont également fournies à des fins de comparaison à la dernière ligne du tableau 6.4. On croit que les estimations de la SEPC sont moins biaisées que les estimations ajustées de l'approche participative, puisqu'elles sont obtenues à partir d'une enquête probabiliste, malgré un faible taux de réponse (environ 15 %), avec des ajustements de poids dus à la non-réponse et au calage.

À partir des estimations et des erreurs-types du tableau 6.4, nous faisons les observations suivantes :

- Les méthodes 2 à 8 sont toutes à peu près équivalentes.
- Pour les sept premières proportions, dans lesquelles les estimations naïves (méthode 1) sont significativement différentes des estimations de la SEPC, les méthodes 2 à 8 donnent des estimations ajustées de l'approche participative plus proches des estimations de la SEPC, ce qui indique une réduction non négligeable du biais. En effet, pour les trois premières proportions, les estimations ajustées de l'approche participative ne sont pas très différentes des estimations de la SEPC. Cela n'est pas surprenant pour les deux premières proportions, puisque les variables d'intérêt peuvent être dérivées à partir des variables auxiliaires. En revanche, cette observation est particulièrement intéressante pour la troisième proportion. Pour les proportions 4 à 7, la réduction du biais n'est pas si spectaculaire, bien que non négligeable; les estimations ajustées de l'approche participative se situent entre les estimations naïves et les estimations de la SEPC.
- Pour les deux dernières proportions, les estimations naïves, les estimations ajustées de l'approche participative et les estimations de la SEPC sont toutes semblables. On observe un léger écart, mais non alarmant, entre les estimations ajustées de l'approche participative et les estimations de la SEPC pour la dernière proportion dans les méthodes 2 et 3, qui n'utilisent pas de groupes homogènes. Dans l'ensemble, il est rassurant de constater que la pondération par l'inverse de la probabilité n'a pas introduit de biais importants pour les deux dernières proportions.
- Enfin, les erreurs-types de la méthode naïve sont beaucoup plus petites que celles des autres méthodes. Cela indique que les estimations naïves sont probablement plus stables. Cependant, l'erreur-type ne tient pas compte du biais et elle ne doit pas être le principal critère pour choisir la méthode qui convient le mieux.

Tableau 6.4
Estimations et erreurs-types (en italique) en pourcentage.

Méthode	Modèle	Sélection pas à pas	Groupes homogènes	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
1	Ordonnée à l'origine	-	-	64,5	65,4	50,2	45,6	32,1	74,2	70.0	9.9	45.6
				0,27	*0,26*	*0,27*	*0,28*	*0,27*	*0,24*	*0.26*	*0.17*	*0.28*
2	Logistique	-	-	29,7	50,2	40,4	28,0	23,5	67,9	62.4	11.4	43.5
				-	-	-	-	-	-	-	-	-
3	Logistique	Oui	-	28,9	48,2	39,8	26,6	23,3	68,1	64.1	10.2	42.3
				-	-	-	-	-	-	-	-	-
4	Logistique	-	Frank	32,4	52,1	40,6	29,5	23,5	68,0	63.5	10.7	44.9
				0,41	*0,76*	*0,70*	*0,58*	*0,60*	*0,74*	*0.78*	*0.49*	*0.77*
5	Logistique	Oui	Frank	30,8	51,4	39,8	28,5	22,4	67,9	64.0	10.3	44.4
				0,35	*0,86*	*0,78*	*0,63*	*0,59*	*0,82*	*0.89*	*0.54*	*0.87*
6	Logistique	Oui	nppCART avec élagage	30,9	50,7	39,5	28,4	22,9	67,8	63.7	10.4	44.5
				0,36	*0,84*	*0,78*	*0,70*	*0,79*	*1,02*	*1.00*	*0.62*	*1.02*
7	-	-	nppCART sans élagage	30,2	52,7	40,6	28,0	24,3	69,3	65.4	9.4	46.8
				0,29	*0,88*	*0,91*	*0,46*	*0,82*	*0,91*	*0.96*	*0.42*	*0.74*
8	-	-	nppCART avec élagage	30,2	52,5	40,5	28,0	23,8	69,4	65.2	9.3	47.0
				0,29	*0,87*	*0,91*	*0,47*	*0,81*	*0,90*	*1.03*	*0.39*	*0.78*
Estimation de la SEPC				**30.6**	**50.1**	**40.2**	**19,3**	**15,9**	**57,3**	**54,4**	**9,8**	**46,2**
				0.87	***1.25***	***1.14***	***0,97***	***0,87***	***1,41***	***1,33***	***0,86***	***1,42***

Au moyen de ces données, les méthodes 2 à 8 ont donné des résultats comparables. Cela peut s'expliquer par la grande taille de l'échantillon probabiliste de l'EPA. Afin d'étudier le comportement des méthodes de pondération par l'inverse de la probabilité quand l'échantillon probabiliste est plus petit, nous avons remplacé l'EPA par l'échantillon probabiliste de la SEPC. Les résultats de ce cas sont discutés ci-dessous.

6.4 Résultats de l'intégration des données de l'approche participative à l'échantillon probabiliste de la SEPC

Résultats de la sélection pas à pas pour le modèle logistique

Quand nous avons utilisé la SEPC comme échantillon probabiliste, notre procédure de sélection pas-à-pas a sélectionné de nouveau tous les effets principaux, mais seulement 10 interactions simples pour un total de 254 paramètres de modèle. Tous les effets principaux sauf un sont entrés dans le modèle avant la première interaction, selon l'ordre suivant : niveau de scolarité, taille du ménage, région économique, sexe, statut d'immigrant, groupe d'âge et état matrimonial. Pour ces données, les interactions simples n'étaient, encore une fois, pas aussi importantes que les effets principaux pour réduire le AIC.

Comparaison des valeurs du critère d'information d'Akaike

Le tableau 6.5 présente les valeurs du critère d'information d'Akaike (AIC) relatif pour les huit méthodes décrites dans le tableau 6.2. En comparant les méthodes 2 et 3, nous constatons que la prise en compte des interactions simples n'a produit qu'une légère amélioration du AIC relatif. Pour ces données, la création de groupes homogènes a entraîné une augmentation non négligeable du AIC relatif. En particulier, quand un modèle logistique est utilisé avec une sélection pas à pas, le AIC relatif est de 12,1 et augmente à 18,5 après la formation de groupes homogènes au moyen de nppCART. L'utilisation de nppCART sans modèle logistique (méthodes 7 et 8) a également produit un AIC relatif plus grand que les méthodes 2 et 3. L'effet de l'élagage demeure négligeable avec ces données puisque les AIC relatifs des méthodes 7 et 8 sont semblables. Néanmoins, l'élagage a réduit le nombre de groupes de 600 à 451. Le remplacement de l'échantillon de l'EPA par l'échantillon de la SEPC a entraîné une réduction du nombre de groupes pour les méthodes 4 à 8; cela n'est pas surprenant puisque la taille de l'échantillon de la SEPC est beaucoup plus petite que la taille de l'échantillon de l'EPA.

Le tableau 6.5 montre également la proportion du AIC qui provient de chacun des trois termes du côté droit de (3.6) ou (3.15). Encore une fois, le premier terme, $-2\hat{l}(\hat{\boldsymbol{\alpha}})$, est la composante dominante du AIC, et l'importance relative des deux autres termes augmente avec q ou G. Compte tenu de la petite taille de l'échantillon de la SEPC, le troisième terme, qui peut être considéré comme une pénalité due à l'utilisation d'un échantillon probabiliste au lieu d'un recensement, est maintenant relativement plus grand que le deuxième terme $2q$ (ou $2G$). Le deuxième terme pourrait donc être omis, comme dans Lumley et Scott (2015), bien que le fait de le négliger n'apporte aucun avantage en terme de calcul informatique.

Comparaisons des estimations

Le tableau 6.6 présente les estimations et leurs erreurs-types bootstrap (en italique) pour chacune des neuf proportions du tableau 6.1 et pour chaque méthode décrite au tableau 6.2. Nous faisons les observations suivantes :

- Pour les deux premières proportions, les variables d'intérêt peuvent être dérivées de variables auxiliaires, et nous nous attendons à ce que les méthodes de pondération par l'inverse de la probabilité éliminent entièrement le biais. Les méthodes 7 et 8 (nppCART sans modèle logistique) ont essentiellement éliminé l'écart entre les estimations naïves et les estimations de la SEPC. Les autres méthodes n'ont pas connu autant de succès, bien que la méthode 4 (modèle logistique avec effets principaux suivi de la méthode Frank) ait obtenu d'assez bons résultats.
- La méthode 2 a semblé surajuster les estimations naïves pour les trois premières proportions. La formation de groupes homogènes (méthode 4) a corrigé ce surajustement.
- Les méthodes 2 et 3 (modèle logistique sans groupes homogènes) sont quelque peu erratiques. Cela peut s'expliquer par des poids de sondage non probabiliste variables et extrêmes, en particulier pour la méthode 3. Le coefficient de variation des poids de sondage non probabiliste est indiqué dans le tableau 6.7 pour chaque méthode. Il est de 7,5 et de 39,7 respectivement pour les méthodes 2 et 3, alors qu'il ne dépasse pas 5,5 pour toutes les autres méthodes. Cela montre l'importance de la formation de groupes homogènes pour réduire les poids extrêmes. Par comparaison, quand l'EPA est utilisée comme échantillon probabiliste, le coefficient de variation des poids de sondage non probabiliste est de 4,7 et de 6,3 respectivement pour les méthodes 2 et 3, et il n'est pas supérieur à 4,0 pour toutes les autres méthodes.
- Les méthodes qui utilisent la sélection pas à pas ont eu tendance à sous-ajuster quand des groupes homogènes ont été formés (méthodes 5 et 6), particulièrement pour la première proportion. Cela n'était pas attendu étant donné les grandes valeurs du AIC relatif dans le tableau 6.5. Toutefois, le AIC relatif indique seulement la force de l'association entre les variables auxiliaires et la participation. Il ne tient pas compte de la force de l'association entre les variables auxiliaires et les variables d'intérêt, qui peut avoir une incidence sur l'ampleur du biais de participation et de la variance.
- En comparant les méthodes 5 et 6, nous constatons que la création de groupes homogènes au moyen de la méthode Frank et de nppCART a produit des estimations semblables, les estimations de nppCART ayant tendance à être légèrement plus proches des estimations de la SEPC, peut-être en raison du plus grand nombre de groupes utilisé avec nppCART.
- L'élagage n'a pas montré d'améliorations considérables dans nos expériences puisque les méthodes 7 et 8 ont produit des estimations similaires.
- Dans l'ensemble, nppCART avec ou sans élagage (méthodes 7 et 8) semble être la méthode la plus stable et la plus fiable pour réduire le biais de participation, suivie de près par la méthode 4 (modèle logistique avec les effets principaux seulement ainsi que la méthode Frank).

Il est intéressant de constater que les estimations par nppCART du tableau 6.6 (méthodes 7 et 8) ne sont pas très différentes des estimations correspondantes du tableau 6.4 fondées sur l'échantillon probabiliste de

l'EPA. Cela suggère qu'on peut réussir à réduire le biais au moyen d'un petit échantillon probabiliste, même s'il demeure préférable d'utiliser un échantillon probabiliste plus grand. Pour nppCART, l'utilisation de l'EPA comme échantillon probabiliste a donné des résultats légèrement meilleurs que l'utilisation de la SEPC. Pour d'autres méthodes, les différences se sont parfois avérées beaucoup plus grandes et l'utilisation de l'EPA a fourni de meilleures estimations. Cela peut constituer un argument en faveur de nppCART lorsque la taille de l'échantillon probabiliste est petite.

Tableau 6.5
Valeurs du AIC relatif en pourcentage.

Méthode	Modèle	Sélection pas à pas	Groupes homogènes	AIC relatif (%)	q ou G	Proportion (%) du AIC issue du 1er terme	Proportion (%) du AIC issue du 2e terme	Proportion (%) du AIC issue du 3e terme
1	Ordonnée à l'origine	-	-	0	1	100,00	0,00	0,00
2	Logistique	-	-	11,2	90	98,45	0,04	1,50
3	Logistique	Oui	-	12,1	254	96,27	0,12	3,62
4	Logistique	-	Frank	13,4	20	98,18	0,01	1,80
5	Logistique	Oui	Frank	15,9	16	99,35	0,01	0,64
6	Logistique	Oui	nppCART avec élagage	18,5	384	96,43	0,19	3,38
7	-	-	nppCART sans élagage	14,3	600	95,93	0,28	3,78
8	-	-	nppCART avec élagage	14,4	451	96,27	0,21	3,51

Tableau 6.6
Estimations et erreurs-types (en italique) en pourcentage.

Méthode	Modèle	Sélection pas à pas	Groupes homogènes	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
1	Ordonnée à l'origine	-	-	64,5	65,4	50,2	45,6	32,1	74,2	70.0	9.9	45.6
				0,28	*0,28*	*0,29*	*0,29*	*0,28*	*0,25*	*0.25*	*0.17*	*0.28*
2	Logistique	-	-	21,3	44,4	34,4	24,4	22,8	69,1	61.3	10.2	44.9
				-	-	-	-	-	-	-	-	-
3	Logistique	Oui	-	29,4	43,4	28,3	29,8	27,4	78,4	71.8	10.1	27.6
				-	-	-	-	-	-	-	-	-
4	Logistique	-	Frank	34,1	50,9	39,4	30,2	25,8	70,8	66.6	9.8	45.1
				0,59	*0,61*	*0,56*	*0,51*	*0,50*	*0,55*	*0.58*	*0.36*	*0.59*
5	Logistique	Oui	Frank	43,6	54,6	41,8	34,3	27,4	71,7	67.9	9.7	44.6
				0,67	*0,54*	*0,50*	*0,55*	*0,43*	*0,44*	*0.47*	*0.30*	*0.47*
6	Logistique	Oui	nppCART avec élagage	42,0	54,0	41,2	34,2	27,3	70,8	67.4	10.1	44.6
				0,81	*0,77*	*0,73*	*0,71*	*0,63*	*0,69*	*0.66*	*0.44*	*0.70*
7	-	-	nppCART sans élagage	30,8	48,9	39,1	28,5	27,7	71,5	64.9	8.9	47.1
				0,98	*1,38*	*1,41*	*0,80*	*1,35*	*1,23*	*1.46*	*0.56*	*1.49*
8	-	-	nppCART avec élagage	30,8	49,8	38,7	29,3	27,1	71,5	65.2	9.3	46.8
				0,98	*1,27*	*1,28*	*0,78*	*1,24*	*1,20*	*1.41*	*0.80*	*1.35*
Estimation de la SEPC				**30.6**	**50.1**	**40.2**	**19,3**	**15,9**	**57,3**	**54,4**	**9,8**	**46,2**
				0.87	***1.25***	***1.14***	***0,97***	***0,87***	***1,41***	***1,33***	***0,86***	***1,42***

Tableau 6.7
Coefficients de variation des poids de sondage non probabiliste.

Échantillonnage probabiliste	Méthode 1	Méthode 2	Méthode 3	Méthode 4	Méthode 5	Méthode 6	Méthode 7	Méthode 8
SEPC	0	7,5	39,7	1,8	1,4	2,2	5,5	5,0
EPA	0	4,7	6,3	2,6	3,0	3,6	4,0	3,9

7. Conclusion

Nous avons étendu la méthode du pseudo maximum de vraisemblance de Chen, Li et Wu (2020) qui intègre des données d'un échantillon non probabiliste et d'un échantillon probabiliste. Nous avons élaboré une procédure de sélection de variables pour le modèle logistique et une extension de l'algorithme CART, soit nppCART. Inspirées par Lumley et Scott (2015), nos extensions reposent sur un AIC modifié qui tient adéquatement compte du plan d'échantillonnage probabiliste. Dans le cadre de nos recherches, nous avons constaté que le terme de pénalité supplémentaire dû à l'utilisation d'un échantillon probabiliste au lieu d'un recensement n'était pas négligeable.

Sans surprise, nos expériences ont montré que les méthodes de pondération par l'inverse de la probabilité peuvent réduire le biais de participation, mais qu'il reste parfois un biais important. Pour le grand échantillon probabiliste de l'EPA, toutes les méthodes ont donné des résultats similaires. Des différences importantes entre les méthodes ont été observées quand le plus petit échantillon probabiliste de la SEPC a été utilisé. En particulier, nos expériences ont montré l'importance de la création de groupes homogènes pour réduire l'occurrence de poids extrêmes et améliorer la stabilité et la robustesse des estimations. Pour ce qui est du petit échantillon probabiliste, la prise en compte des interactions simples a quelque peu réduit le AIC, mais n'a généralement pas présenté d'avantages pour les estimations. Les effets principaux ont semblé plus importants que les interactions simples pour réduire le AIC avec nos données. Dans l'ensemble, la meilleure méthode de réduction du biais a été l'algorithme nppCART, suivi de près par l'utilisation d'un modèle logistique avec effets principaux seulement ainsi que la création de groupes homogènes. Cependant, des conclusions différentes pourraient être tirées en présence de domaines plus petits ou d'autres ensembles de données.

Il est bien connu que les estimateurs pondérés par l'inverse de la probabilité peuvent être inefficaces, particulièrement quand les variables d'intérêt sont faiblement liées aux poids. On peut atténuer ce problème par un calage sur des totaux de population connus ou des totaux estimés à partir de l'échantillon probabiliste. Le calage sera particulièrement efficace quand des variables auxiliaires fortement liées aux variables d'intérêt sont disponibles et exclues du modèle de participation. Ce n'était pas le cas dans nos expériences. Le lissage de poids est une autre solution visant à améliorer l'efficacité des estimateurs pondérés par l'inverse de la probabilité qui peut être utile quand on ne dispose pas de variables de calage aussi puissantes. Il consiste à remplacer les poids par des prédictions obtenues en modélisant les poids conditionnellement aux variables d'intérêt. Dans le contexte de l'intégration d'un échantillon probabiliste et non probabiliste, le lissage de poids a été étudié par Ferri-Garcia, Beaumont, Bosa, Charlebois et Chu (2021).

Des méthodes basées sur les arbres de régression plus sophistiquées que l'algorithme CART, comme les forêts aléatoires, sont proposées dans la littérature. Étant donné les bonnes performances de nppCART dans nos expériences, il pourrait être judicieux d'étendre ces méthodes au scénario d'intégration des données étudié dans le présent article et de les évaluer. D'autres recherches sont nécessaires sur le sujet.

Il n'existe probablement pas de méthode de pondération par l'inverse de la probabilité qui soit uniformément meilleure que toutes les autres méthodes. Toutes les techniques sont utiles et peuvent faire partie de la boîte à outils du statisticien. Cependant, il serait utile d'élaborer des indicateurs de réduction du biais qui aideraient les statisticiens à choisir la meilleure méthode pour un échantillon probabiliste et un échantillon non probabiliste donnés. Bien que le AIC relatif et le coefficient de variation des poids de sondage non probabiliste soient deux indicateurs utiles, ils ne dressent pas un tableau complet, car ils ne disent rien à propos de la force de l'association entre les variables auxiliaires et les variables d'intérêt. Il serait intéressant d'explorer l'utilisation de méthodes d'appariement statistique avec des modèles non paramétriques (par exemple des forêts aléatoires) pour chaque variable d'intérêt conditionnellement aux variables auxiliaires. Les estimations obtenues devraient être plus efficaces que les méthodes de pondération par l'inverse de la probabilité parce qu'elles seraient adaptées à chaque variable d'intérêt. En pratique, l'application de cette stratégie d'appariement statistique serait fastidieuse, car il faudrait élaborer et valider un modèle différent pour chaque estimation produite. On pourrait toutefois calculer et utiliser quelques estimations par appariement statistique pour évaluer les méthodes de pondération par l'inverse de la probabilité. On pourrait s'attendre à ce qu'une meilleure méthode de pondération par l'inverse de la probabilité tende généralement à produire des estimations plus proches des estimations par appariement statistique. Une procédure possible pour rapprocher les deux méthodes consisterait à caler les poids par l'inverse de la probabilité de façon à ce que les estimations qui en résultent correspondent exactement aux estimations par appariement statistique sélectionnées.

Annexe 1

Démonstration de la preuve de l'équation (3.5)

Au moyen de développements du premier degré en série de Taylor, nous obtenons

$$\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}}) = \left[\hat{l}(\boldsymbol{\alpha}_0) - l_0(\boldsymbol{\alpha}_0)\right] + \left[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0) - \mathbf{U}_0(\boldsymbol{\alpha}_0)\right]' \left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) + o_p\left(\frac{N}{n^P}\right) \tag{A.1}$$

et

$$\hat{\mathbf{U}}(\hat{\boldsymbol{\alpha}}) = \hat{\mathbf{U}}(\boldsymbol{\alpha}_0) + \hat{\mathbf{H}}(\boldsymbol{\alpha}_0)\left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) + o_p\left(\frac{N}{\sqrt{n^P}}\right), \tag{A.2}$$

où $\mathbf{U}_0(\boldsymbol{\alpha}) = \partial l_0(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$. En plus de (A.1) et (A.2), nous supposons également que

$$\hat{\mathbf{H}}(\boldsymbol{\alpha}) = \mathbf{H}_0(\boldsymbol{\alpha}) + o_p\left(N\right) \tag{A.3}$$

selon le modèle et le plan de sondage. En notant que $\mathbf{U}_0(\boldsymbol{\alpha}_0) = \mathbf{0}$ et $\hat{\mathbf{U}}(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$, nous obtenons à partir de (A.1), (A.2) et (A.3),

$$\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}}) = \left[\hat{l}(\boldsymbol{\alpha}_0) - l_0(\boldsymbol{\alpha}_0)\right] + \left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right)' \left[-\mathbf{H}_0(\boldsymbol{\alpha}_0)\right]\left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) + o_p\left(\frac{N}{n^P}\right). \tag{A.4}$$

En ignorant le terme d'ordre plus faible et en prenant l'espérance des deux côtés de (A.4) :

$$E_{md}\left[\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}})\right] \approx \text{tr}\left[-\mathbf{H}_0(\boldsymbol{\alpha}_0)\,\text{var}_{md}(\hat{\boldsymbol{\alpha}})\right], \tag{A.5}$$

où $\text{var}_{md}(\hat{\boldsymbol{\alpha}}) = E_{md}[(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)']$. En utilisant (A.2) et (A.3) et en ignorant les termes d'ordre plus faible, nous pouvons approcher cette variance comme suit :

$$\begin{aligned}
\text{var}_{md}(\hat{\boldsymbol{\alpha}}) &\approx [\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1}\,\text{var}_{md}[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\,[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} \\
&= [\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1}\left\{\text{var}_m[\mathbf{U}(\boldsymbol{\alpha}_0)] + E_m\{\text{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\}\right\}[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} \\
&= -[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} + [\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} E_m\{\text{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\}\,[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1},
\end{aligned} \tag{A.6}$$

où $\mathbf{U}(\boldsymbol{\alpha}) = \partial l(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ est donné dans l'équation (3.1) pour le modèle logistique. La dernière équation dans (A.6) résulte d'une propriété bien connue de la matrice d'information de Fisher $-\mathbf{H}_0(\boldsymbol{\alpha}_0)$ (si l'on suppose que le vrai modèle se trouve dans la même famille paramétrique que le modèle postulé). L'utilisation de (A.6) dans (A.5) donne le résultat (3.5).

Annexe 2

Illustration de la méthode Frank

La figure A.1 ci-dessous contient trois sous-figures, les figures A.1(A), A.1(B) et A.1(C), qui illustrent le comportement de la méthode Frank pour les données décrites à la section 6.1 et pour la méthode 5 décrite à la section 6.2 quand l'EPA est utilisée comme échantillon probabiliste. La description de chaque sous-figure est donnée ci-dessous :

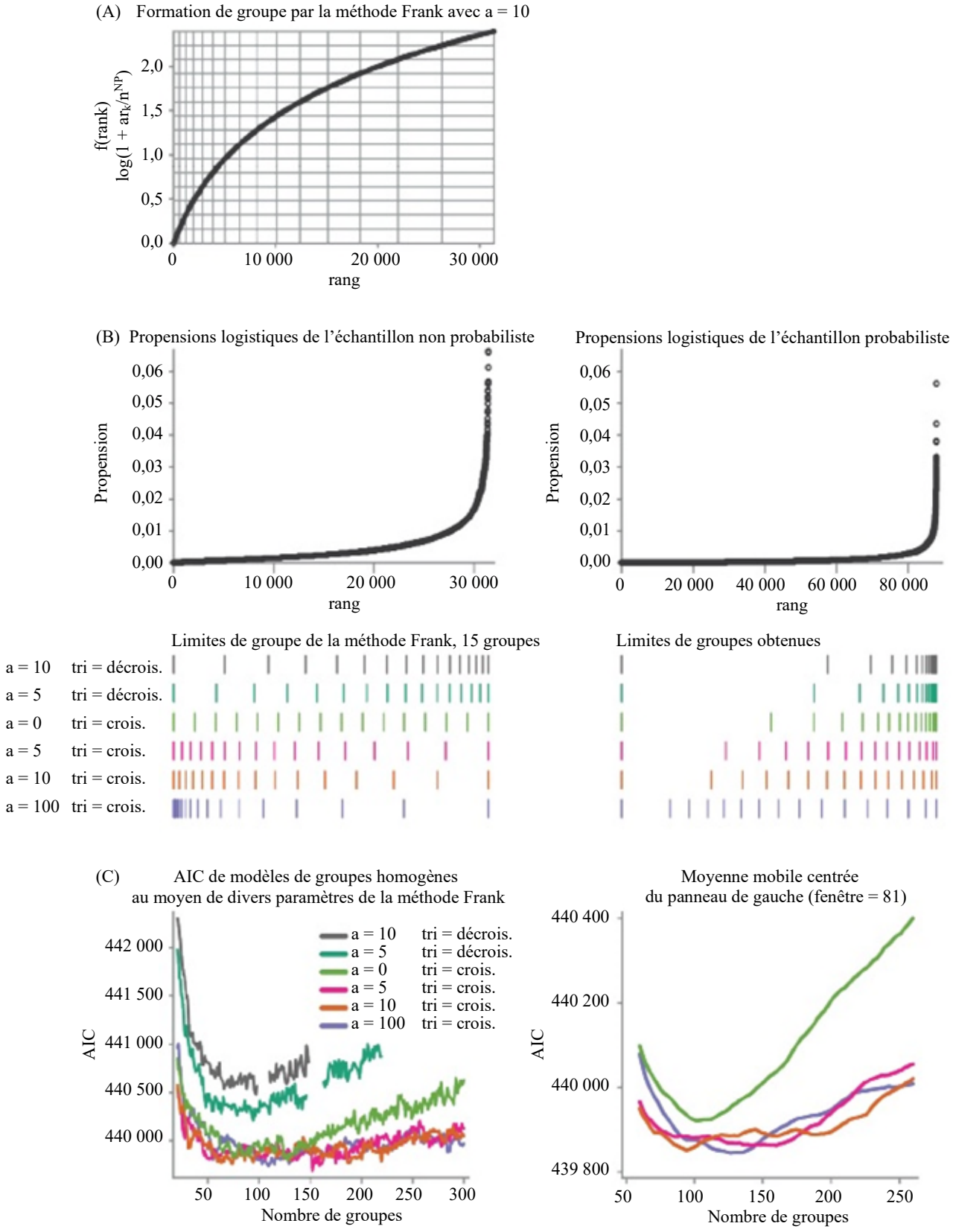
(A) Méthode Frank avec $a = 10$, $G = 15$ et $n^{\text{NP}} = 31\,415$. Le rang, r_k, est sur l'axe horizontal et la fonction du rang, $f(r_k) = \log(1 + a\,r_k/n^{\text{NP}})$, est sur l'axe vertical. Les classes sont de largeur égale dans la plage de $f(r_k)$. La constante a détermine la forme de la fonction. Quand a augmente, elle devient de plus en plus non linéaire et les groupes sont plus regroupés d'un côté.

(B) Les panneaux supérieurs montrent les valeurs triées de $\hat{p}_k^{\text{logistic}}$ pour l'échantillon non probabiliste (à gauche) et l'échantillon probabiliste (à droite). Quinze groupes sont formés à partir de l'échantillon

non probabiliste au moyen de la méthode Frank avec différentes valeurs de a et les deux ordres de tri, ce qui donne différentes limites de groupe représentées par les barres de couleur dans les panneaux inférieurs. Pour l'échantillon non probabiliste (panneau inférieur gauche), quand le rang est défini par ordre croissant de $\hat{p}_k^{\text{logistic}}$, les groupes sont plus petits pour les petites valeurs de $\hat{p}_k^{\text{logistic}}$. Quand le rang est défini en ordre décroissant de $\hat{p}_k^{\text{logistic}}$, les groupes sont plus petits pour les grandes valeurs de $\hat{p}_k^{\text{logistic}}$. L'augmentation de a accroît le regroupement, tandis que $a = 0$ donne des groupes de taille égale.

(C) Le AIC (3.15) par rapport au nombre de groupes pour différentes valeurs de a et les deux ordres de tri. Le tri de $\hat{p}_k^{\text{logistic}}$ en ordre croissant donne des valeurs plus petites du AIC, sans grande sensibilité aux variations de la valeur de a. Le AIC est minimisé avec environ 100 groupes pour tous les paramétrages. Le panneau de droite lisse le panneau de gauche au moyen d'un filtre de moyenne mobile centrée dont la taille de la fenêtre est de 81. Les courbes lissées montrent que la méthode Frank donne des résultats légèrement meilleurs que les groupes de taille égale $(a = 0)$, surtout lorsque le nombre de groupes est plus élevé que le nombre optimal, ce qui ajoute une certaine robustesse au choix du nombre de groupes. Quand le nombre de groupes est grand et que $\hat{p}_k^{\text{logistic}}$ est trié par ordre décroissant, certains groupes ne contiennent aucune unité dans l'échantillon probabiliste. Par conséquent, $\hat{p}_g$ n'est pas défini pour ces groupes, et le AIC ne peut pas être calculé.

Figure A.1 Illustration de la méthode Frank.



Annexe 3

Variables auxiliaires

Groupe d'âge : groupes d'âge de cinq ans, débutant par le groupe des 15 à 19 ans puis terminant par le groupe des 75 ans et plus.

Sexe : Hommes - Femmes.

Niveau de scolarité : 8 catégories (sans diplôme d'études secondaires; études secondaires; études postsecondaires partielles; certificat ou diplôme d'une école de métiers; collège communautaire, CÉGEP, etc.; certificat universitaire inférieur au baccalauréat; baccalauréat; grade supérieur au baccalauréat).

Région économique : région géographique infraprovinciale divisant le pays. Cette variable comprend 73 niveaux, mais certains ont été regroupés en raison d'un nombre insuffisant de répondants; 56 niveaux ont été utilisés dans les modèles.

Statut d'immigrant : 3 niveaux (né au Canada; immigrant reçu; immigrant non reçu).

Taille du ménage : nombre de personnes dans le ménage, quel que soit leur âge, limité à 6.

État matrimonial : 6 niveaux (marié(e); conjoint(e) de fait; veuf ou veuve; séparé(e); divorcé(e); célibataire; jamais marié(e)).

Situation d'emploi : 3 niveaux (personne occupée et au travail au moins une partie de la semaine de référence; personne occupée, mais absente du travail; personne sans emploi).

Bibliographie

Bahamyirou, A., et Schnitzer, M.E. (2021). Data integration through outcome adaptive LASSO and a collaborative propensity score approach. *arXiv preprint arXiv:2103.15218*.

Baribeau, B. (2020). Trial by COVID for Statistics Canada's web panel pilot. Document interne, Statistique Canada.

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

Beaumont, J.-F., et Émond, N. (2022). A bootstrap variance estimation method for multistage sampling and two-phase sampling when Poisson sampling is used at the second phase. *Stats*, 5, 339-357.

Beaumont, J.-F., et Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *Revue Internationale de Statistique*, 80, 127-148.

Breiman, L., Friedman, J.H., Olshen, R.A. et Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Boca Raton, Floride.

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

Chu, K., et Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, mai 2019.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

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

Eltinge, J.L., et Yansaneh, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 1, 37-45. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997001/article/3103-fra.pdf.

Ferri-Garcia, R., Beaumont, J.-F., Bosa, K., Charlebois, J. et Chu, K. (2021). Weight smoothing for nonprobability surveys. *TEST* (publié en ligne).

Ferri-Garcia, R., et Rueda, M.d.M. (2018). Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT*, 42, 159-182.

Gambino, J., Kennedy, B., et Singh, M.P. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application. *Techniques d'enquête*, 27, 1, 69-79. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001001/article/5855-fra.pdf.

Haziza, D., et Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, *32*, 129-145.

Izenman, A.J. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. New York: Springer Science & Business Media.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web survey. *Journal of Official Statistics*, 22, 329-349.

Lohr, S.L. (2021). Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples. *Techniques d'enquête*, 47, 2, 247-285. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00008-fra.pdf.

Lohr, S., Hsu, V. et Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandrie, Virginie.

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

Phipps, P., et Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, 6, 772-794.

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

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 2, 225-234. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1992002/article/14486-fra.pdf.

Renaud, M., et Beaumont, J.-F. (2020). Crowdsourcing during a pandemic: The Statistics Canada experience. Document présenté au Comité consultative sur les méthodes statistiques, Statistique Canada, 27 octobre 2020.

Rivers, D. (2007). Sampling from web surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandrie, Virginie.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 2, 231-263.

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

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New-York: John Wiley & Sons, Inc.

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

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

Yang, S., et Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

Yang, S., Kim, J.K. et Hwang, Y. (2021). Intégration de données d'enquêtes probabilistes et de mégadonnées aux fins d'inférence de population finie au moyen d'une imputation massive. *Techniques d'enquête*, 47, 1, 33-64. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021001/article/00004-fra.pdf.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 123-132
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada »

Julie Gershunskaya et Vladislav Beresovsky[1]

Résumé

Beaumont, Bosa, Brennan, Charlebois et Chu (2024) proposent des méthodes novatrices de sélection de modèles aux fins d'estimation des probabilités de participation pour des unités d'échantillonnage non probabiliste. Notre examen portera principalement sur le choix de la vraisemblance et du paramétrage du modèle, qui sont essentiels à l'efficacité des techniques proposées dans l'article. Nous examinons d'autres méthodes fondées sur la vraisemblance et la pseudo-vraisemblance pour estimer les probabilités de participation et nous présentons des simulations mettant en œuvre et comparant la sélection de variables fondée sur le critère d'information d'Akaike (AIC). Nous démontrons que, dans des scénarios pratiques importants, la méthode fondée sur une vraisemblance formulée sur les échantillons non probabiliste et probabiliste groupés qui sont observés offre un meilleur rendement que les autres solutions fondées sur la pseudo-vraisemblance. La différence de sensibilité du AIC est particulièrement grande en cas de petites tailles de l'échantillon probabiliste et de petit chevauchement dans les domaines de covariables.

Mots-clés : Échantillon non probabiliste; probabilités de participation; vraisemblance de l'échantillon ; combinaison de données.

En s'appuyant sur les travaux récents portant sur les méthodes d'intégration de données, Beaumont et coll. (2024) proposent et appliquent à des données réelles des méthodes novatrices de sélection de modèles aux fins d'estimation des probabilités de participation pour des unités d'échantillonnage non probabiliste. Nous félicitons les auteurs pour leur contribution inspirante et opportune et nous sommes ravis d'avoir l'occasion de commenter et d'examiner les méthodes proposées dans l'article.

Les statisticiens d'enquête doivent de plus en plus composer avec la nécessité d'extraire des renseignements utiles des données recueillies sans plan d'enquête probabiliste correctement planifié. Parallèlement, nous assistons à l'élaboration rapide de méthodes d'apprentissage automatique permettant de traiter efficacement des ensembles multidimensionnels de covariables. Progressivement, on saisit que l'apprentissage automatique peut servir à traiter l'estimation à partir d'échantillons non probabilistes. L'article ouvre la voie à l'adaptation de ces méthodes dans un scénario où l'on combine un échantillon probabiliste et un échantillon non probabiliste. Les auteurs proposent une formule générale modifiée du critère d'information d'Akaike (AIC) qui tient compte du plan d'échantillonnage probabiliste dans le scénario des échantillons combinés. Ils calculent également l'expression du AIC pour le cas particulier des groupes homogènes et l'appliquent pour faire un choix parmi les partitions dans les méthodes fondées sur le rang pour former les groupes. Enfin, les auteurs adaptent l'algorithme de croissance des arbres de

1. Julie Gershunskaya et Vladislav Beresovsky, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE Washington, DC 20212, États-Unis. Courriel : Gershunskaya.Julie@bls.gov.

classification et de régression (CART) en utilisant une pseudo-vraisemblance comme fonction « objectif » et en appliquant le AIC modifié pour élaguer l'arbre.

Notre examen portera principalement sur le choix de la vraisemblance et du paramétrage du modèle, qui sont essentiels à l'efficacité des techniques proposées dans l'article. À la section 1, nous examinons plusieurs méthodes d'estimation des probabilités de participation proposées au cours des dernières années et nous fournissons des expressions du AIC pour le cas des groupes homogènes. À la section 2, nous présentons des simulations mettant en œuvre et comparant la sélection de variables basée sur le AIC pour ces méthodes. Enfin, nous présentons quelques conclusions à la section 3.

1. Méthodes d'estimation des probabilités de participation

Nous commençons par les deux méthodes fondées sur la pseudo-vraisemblance examinées dans l'article, puis nous discutons d'une méthode fondée sur la vraisemblance exacte avant de proposer une modification à la méthode de pondération par la propension logistique ajustée de Wang, Valliant et Li (2021); voir également l'étude connexe dans Gershunskaya et Lahiri (2023). Nous nous penchons ensuite sur le cas des groupes homogènes et calculons puis comparons les expressions du AIC pour chacune des approches dans ce cas particulier important et relativement simple. Dans les pages qui suivent, à moins d'indication explicite, nous adoptons la notation de l'article commenté.

1.1 Méthode de Chen, Li et Wu

En supposant que l'échantillon probabiliste et l'échantillon non probabiliste sont sélectionnés à partir de la même population finie U, Chen, Li et Wu (2020) [ci-après CLW] commencent par écrire un logarithme du rapport de vraisemblance sur les unités dans U, par rapport à la variable de Bernoulli δ_k :

$$\ell^{\mathrm{CLW}}(\boldsymbol{\alpha}) = \sum_{k\in U}\left\{\delta_k \log\left[p_k(\boldsymbol{\alpha})\right] + (1-\delta_k)\log\left[1-p_k\left(\boldsymbol{\alpha}\right)\right]\right\}, \tag{1.1}$$

où δ_k est l'indicateur d'inclusion de l'échantillon non probabiliste de l'unité k, $p_k(\mathbf{x_k}) = P\{\delta_k = 1 \mid \mathbf{x_k}\}$, et $\boldsymbol{\alpha}$ est le vecteur de paramètre dans un modèle de régression logistique pour p_k, où $\mathrm{logit}(p_k(\boldsymbol{\alpha})) = \boldsymbol{\alpha}^T\mathbf{x_k}$.

Étant donné que les unités de la population finie ne sont pas observées, CLW regroupent la somme dans (1.1) en la présentant comme une somme en deux parties : la partie 1 comprend la somme sur les unités d'échantillonnage non probabiliste, s_{NP}, et la partie 2 est la somme sur la population finie U :

$$\ell^{\mathrm{CLW}}(\boldsymbol{\alpha}) = \sum_{k\in s_{\mathrm{NP}}}\log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k\left(\boldsymbol{\alpha}\right)}\right] + \sum_{k\in U}\log\left[1-p_k\left(\boldsymbol{\alpha}\right)\right]. \tag{1.2}$$

CLW utilisent une méthode de pseudo-vraisemblance en remplaçant la somme sur la population finie par son estimation fondée sur l'échantillon probabiliste :

$$\hat{\ell}^{\mathrm{CLW}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right] + \sum_{k \in s_P} w_k \log\left[1-p_k(\boldsymbol{\alpha})\right], \tag{1.3}$$

où les poids $w_k = \pi_k^{-1}$ sont des valeurs inverses des probabilités d'inclusion de l'échantillon de référence π_k. On obtient les estimations en résolvant les équations d'estimation respectives fondées sur la pseudo-vraisemblance.

Il convient de mentionner que dans cette méthode, la vraisemblance est formulée par rapport à l'indicateur δ_k, bien que cette variable ne soit pas observée.

1.2 Méthode de pondération par la propension logistique ajustée

Pour leur méthode de pondération par la propension logistique ajustée (PLA), Wang et coll. (2021) présentent un concept imaginaire composé de deux parties : ils *empilent* ensemble l'échantillon non probabiliste s_{NP} (partie 1) et la population finie U (partie 2). Étant donné que les unités de l'échantillonnage non probabiliste appartiennent à la population finie, elles apparaissent deux fois dans l'ensemble empilé. Ils formulent une vraisemblance de Bernoulli pour la variable R_k, où $R_k = 1$ si l'unité k appartient à la partie 1 de l'ensemble empilé; et $R_k = 0$ autrement :

$$\ell^{\mathrm{PLA}}(\boldsymbol{\gamma}) = \sum_{k \in s_{\mathrm{NP}}} \log\left[p_{Rk}(\boldsymbol{\gamma})\right] + \sum_{k \in U} \log\left[1-p_{Rk}(\boldsymbol{\gamma})\right], \tag{1.4}$$

où $\boldsymbol{\gamma}$ est le vecteur de paramètre dans un modèle de régression logistique pour $p_{Rk}(\mathbf{x_k}) = P\{R_k = 1 \mid \mathbf{x_k}\}$. Puisque la population finie n'est pas disponible, ils appliquent une méthode de pseudo-vraisemblance :

$$\hat{\ell}^{\mathrm{PLA}}(\boldsymbol{\gamma}) = \sum_{k \in s_{\mathrm{NP}}} \log\left[p_{Rk}(\boldsymbol{\gamma})\right] + \sum_{k \in s_P} w_k \log\left[1-p_{Rk}(\boldsymbol{\gamma})\right], \tag{1.5}$$

qui donne une estimation de p_{Rk}. Cependant, l'objectif réel est de trouver des probabilités p_k plutôt que p_{Rk}. À la deuxième étape de la méthode de pondération par la propension logistique ajustée, les estimations de p_k sont calculées à partir de l'identité

$$p_{Rk} = \frac{p_k}{1+p_k}. \tag{1.6}$$

Wang et coll. (2021) ont constaté que, dans leurs simulations, l'estimateur de la méthode de pondération par la propension logistique ajustée était plus efficace que celui de la méthode de CLW, surtout quand la taille de l'échantillon non probabiliste était beaucoup plus grande que la taille de l'échantillon probabiliste.

1.3 Méthode de régression logistique implicite

Examinons maintenant une méthode de vraisemblance exacte formulée pour les échantillons non probabiliste et probabiliste groupés. Savitsky, Williams, Gershunskaya et Beresovsky (2023) proposent

d'empiler ensemble les deux échantillons et de considérer que la variable indicatrice $z_k = 1$ si l'unité k appartient à l'échantillon non probabiliste (partie 1) et que $z_k = 0$ si l'unité k appartient à l'échantillon probabiliste (partie 2). Dans cette construction d'échantillons empilés, en cas de chevauchement entre les deux échantillons, s_{NP} et s_P, alors les unités qui se chevauchent sont ajoutées à l'ensemble empilé, s, deux fois : une fois dans l'échantillon non probabiliste (où $z_k = 1$) et une fois dans l'échantillon probabiliste de référence (où $z_k = 0$). Nous n'avons pas besoin de savoir quelles unités se chevauchent ou si des unités se chevauchent. Ils utilisent les premiers principes pour prouver l'existence d'une relation entre les probabilités $p_{zk}(\mathbf{x_k}) = P\{z_k = 1 \mid \mathbf{x_k}\}$ d'être dans la partie 1 de l'ensemble empilé, d'une part, et les probabilités d'inclusion, p_k et π_k, d'autre part :

$$p_{zk} = \frac{p_k}{\pi_k + p_k}. \tag{1.7}$$

Elliott (2009) et Elliott et Valliant (2017) ont calculé l'expression (1.7) en supposant des échantillons sans chevauchement s_{NP} et s_P. Le calcul donné dans Savitsky et coll. (2023) n'exige pas cette hypothèse.

Pour obtenir des estimations de p_k, Beresovsky (2019) propose une méthode intitulée « régression logistique implicite (RLI) » pour permettre l'estimation de p_k directement à partir de la vraisemblance formulée sur l'échantillon combiné. Dans la méthode de RLI, les probabilités $p_k = p_k(\boldsymbol{\alpha})$ sont paramétrées comme étant $\text{logit}(p_k(\boldsymbol{\alpha})) = \boldsymbol{\alpha}^T \mathbf{x_k}$ (comme dans la méthode de CLW), et l'identité (1.7) sert à présenter p_{zk} comme une fonction composée de $\boldsymbol{\alpha}$; c'est-à-dire $p_{zk} = p_{zk}(p_k(\boldsymbol{\alpha})) = p_k(\boldsymbol{\alpha})/(\pi_k + p_k(\boldsymbol{\alpha}))$. Le logarithme du rapport de vraisemblance de la variable de Bernoulli observée z_k est

$$\ell^{\text{RLI}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\left[p_{zk}(p_k(\boldsymbol{\alpha}))\right] + \sum_{k \in s_P} \log\left[1 - p_{zk}\left(p_k(\boldsymbol{\alpha})\right)\right]. \tag{1.8}$$

On obtient les équations de score à partir de (1.8), en prenant les dérivées, par rapport à $\boldsymbol{\alpha}$, de la fonction composée $p_{zk} = p_{zk}(p_k(\boldsymbol{\alpha}))$. Ainsi, les estimations de p_k sont obtenues directement à partir de (1.8) en une seule étape.

Il convient de mentionner que, pour la méthode de RLI, les probabilités d'inclusion de l'échantillon probabiliste π_k sont supposées connues pour toutes les unités de l'ensemble combiné, ce qui est possible pour de nombreuses enquêtes probabilistes. Si elles ne sont pas immédiatement disponibles, les probabilités π_k pour les unités dans s_{NP} peuvent être déterminées si des variables du plan d'échantillonnage probabiliste sont disponibles pour les unités de l'échantillonnage non probabiliste. Comme l'indiquent Elliott et Valliant (2017), π_k peut être estimé au moyen d'un modèle de régression. Savitsky et coll. (2023) ont utilisé la technique de modélisation bayésienne pour obtenir à la fois π_k et p_k. Par ailleurs, si les probabilités π_k ne sont pas disponibles pour la partie non probabiliste de l'échantillon combiné, on peut appliquer une méthode de pseudo-vraisemblance dite « pseudo-RLI », comme cela est expliqué ci-dessous à la section 1.4.

1.4 Méthode de pseudo-régression logistique implicite

On peut modifier la méthode d'estimation de Wang et coll. (2021) en une procédure d'estimation en une étape semblable à la méthode de RLI : p_k peut être paramétré au moyen de la fonction de lien logistique comme étant $\text{logit}(p_k(\boldsymbol{\alpha})) = \boldsymbol{\alpha}^T \mathbf{x}_\mathbf{k}$, tandis que les probabilités p_{Rk} dans (1.6) peuvent être considérées comme une fonction composée, $p_{Rk} = p_{Rk}(p_k(\boldsymbol{\alpha})) = p_k(\boldsymbol{\alpha})/(1+p_k(\boldsymbol{\alpha}))$. La pseudo-vraisemblance prend la forme :

$$\hat{\ell}^{\text{PRLI}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\left[p_{Rk}(p_k(\boldsymbol{\alpha}))\right] + \sum_{k \in s_P} w_k \log\left[1 - p_{Rk}(p_k(\boldsymbol{\alpha}))\right]. \tag{1.9}$$

Ce changement dans l'estimation des paramètres du modèle rend la méthode plus efficace et moins biaisée que la méthode de pondération par la propension logistique ajustée. Il permet aussi d'éviter les cas où les estimations de p_k deviennent supérieures à 1, comme cela peut se produire dans la méthode de pondération par la propension logistique ajustée, où l'estimation est effectuée en deux étapes.

Il faut prendre note que, bien que les formules (1.3) et (1.9) fondées sur la pseudo-vraisemblance reposent exactement sur le même ensemble de données observées, ces expressions sont très différentes. Nous nous attendons à ce que la méthode de pseudo-RLI donne des estimations plus efficaces parce qu'elle est fondée sur une vraisemblance correctement formulée pour ce qui est de la variable de Bernoulli observée R_k, tandis que la vraisemblance de la méthode de CLW est donnée pour ce qui est de la variable non observée δ_k. Nos simulations (qui ne sont pas comprises dans l'analyse) confirment que la méthode de pseudo-RLI offre un meilleur rendement que la méthode de CLW. L'effet sur le rendement du AIC est illustré dans les simulations de la section 2.

1.5 Groupes homogènes

Les auteurs ont présenté les expressions de logarithme du rapport de vraisemblance et du AIC selon la méthode de CLW pour le cas particulier des groupes homogènes. Nous élargissons maintenant leur approche aux méthodes de pseudo-RLI et de RLI.

Pour la méthode de pseudo-RLI, il est facile de constater que, pour une partition donnée, les estimations des probabilités de participation p_g du groupe g sont $\hat{p}_g = n_g^{\text{NP}} / \hat{N}_g$, où $\hat{N}_g = \sum_{k \in s_P} w_k$ (comme dans la méthode de CLW). Le AIC est donné par

$$\text{AIC}^{\text{PRLI}} = -2\hat{\ell}^{\text{PRLI}}(\hat{\boldsymbol{\alpha}}) + 2G + 2\sum_{g=1}^{G} \left[\frac{1-\hat{p}_g}{1+\hat{p}_g}\right] n_g^{\text{NP}} \frac{\left[\hat{cv}_d(\hat{N}_g)\right]^2}{1-\hat{p}_g}, \tag{1.10}$$

où le logarithme du rapport de vraisemblance est

$$\hat{\ell}^{\text{PRLI}}(\hat{\boldsymbol{\alpha}}) = \sum_{g=1}^{G} \hat{N}_g \left[\hat{p}_g \log\frac{\hat{p}_g}{1+\hat{p}_g} + \log\frac{1}{1+\hat{p}_g}\right]. \tag{1.11}$$

Il convient de mentionner que les derniers termes utilisés dans les formules du AIC selon les méthodes de CLW et de pseudo-RLI diffèrent d'un facteur $(1-\hat{p}_g)/(1+\hat{p}_g)<1$. Cela signifie que, pour une partition donnée, le terme de pénalisation dans la méthode de pseudo-RLI est toujours plus petit que celui dans la méthode de CLW.

Dans la méthode de RLI, les estimations de p_g ne sont pas disponibles sous une forme fermée. On peut les trouver en résolvant les équations :

$$\sum_{k\in s_g}\frac{\pi_{gk}}{\pi_{gk}+p_g}=n_g^P, \tag{1.12}$$

où les π_{gk} sont supposés connus, s_g est la partie de l'échantillon combiné qui appartient au groupe g, $g=1,\ldots,G$. Étant donné que la RLI est fondée sur une vraisemblance exacte, la formule du AIC pour la méthode de RLI ne contient pas le troisième terme et est une expression du AIC type :

$$\mathrm{AIC}^{\mathrm{RLI}}=-2\hat{\ell}^{\mathrm{RLI}}(\hat{\boldsymbol{\alpha}})+2G, \tag{1.13}$$

où

$$\hat{\ell}^{\mathrm{RLI}}(\hat{\boldsymbol{\alpha}})=\sum_{g=1}^{G}\left[\sum_{k\in s_g^{\mathrm{NP}}}\log\frac{\hat{p}_g}{\pi_{gk}+\hat{p}_g}+\sum_{k\in s_g^{P}}\log\frac{\pi_{gk}}{\pi_{gk}+\hat{p}_g}\right]. \tag{1.14}$$

Comparons les termes de pénalisation des trois méthodes pour un ensemble donné de groupes homogènes. (Le partitionnement lui-même dépend de la vraisemblance utilisée, mais nous ne considérons pas cet effet pour le moment.) Nous supposons que, tous les autres facteurs étant égaux, plus la pénalisation est petite, mieux le AIC fonctionne. Par conséquent, nous nous attendons à ce que le critère fondé sur la RLI offre un meilleur rendement que la méthode de pseudo-RLI ou de CLW. Nous nous attendons aussi à ce que la méthode de pseudo-RLI fonctionne mieux que la méthode de CLW, surtout quand l'échantillon non probabiliste est relativement grand et que $\hat{p}_g$ dans (1.10) et dans la formule (3.15) de l'article examiné se rapproche de 1. Les résultats des simulations de la section 2 corroborent ce raisonnement.

2. Simulations

Nous avons réalisé une expérience de simulation pour étudier le rendement du AIC dans les méthodes de CLW, de pseudo-RLI et de RLI.

Pour chaque unité $k=1\ldots,N$ d'une population finie U de taille $N=$ 10 000, nous avons généré des covariables x_{1k} et x_{2k} comme variables normales standards indépendantes.

Nous utilisons un échantillonnage de Poisson avec des probabilités de participation p_k pour sélectionner l'échantillon non probabiliste tiré de la population U. Les probabilités p_k sont générées comme étant

$$\text{logit}(p_k) = \alpha_0 + \alpha_1 x_{1k} + \alpha_2 x_{2k}, \tag{2.1}$$

où nous établissons des valeurs de coefficient précises pour différents scénarios de simulation, comme suit :

- le fait de paramétrer $\alpha_0 = -5$ produit un échantillon s_{NP} de taille approximative $n^{\text{NP}} = 100$; le fait de paramétrer $\alpha_0 = -2{,}5$ produit un échantillon s_{NP} de taille approximative $n^{\text{NP}} = 1\,000$;
- α_1 a été paramétré à 1 pour tous les scénarios;
- α_2 a été paramétré à des valeurs sur une grille allant de $-0{,}3$ à 0,3, qui correspondent à une série de scénarios. Il convient de mentionner que le paramétrage $\alpha_2 = 0$ correspond au cas où p_k est indépendant de x_{2k}; des valeurs plus grandes de α_2 produisent une dépendance plus forte à x_{2k}.

Pour l'échantillon probabiliste s_P, nous avons envisagé des scénarios où la taille de l'échantillon est $n^P = 100$ ou $n^P = 1\,000$. L'échantillon probabiliste est généré au moyen de la probabilité proportionnelle à la taille sans remise, où la mesure de la taille est définie comme suit :

$$\text{logit}(m_k) = 1 + \beta x_{1k}. \tag{2.2}$$

Dans les modèles multivariés comportant un grand nombre de variables auxiliaires et d'interactions, il est probable que l'échantillon probabiliste et l'échantillon non probabiliste se chevaucheraient très peu dans certains des domaines définis par les variables. Firth (1993) et Heinze et Schemper (2002) ont démontré qu'un petit chevauchement ou une séparation peut entraîner des estimations instables des paramètres du modèle. Dans ce cas, il est essentiel d'utiliser une méthode efficace de sélection des variables. C'est la raison pour laquelle nous avons inclus dans nos simulations des scénarios de petit chevauchement et de chevauchement élevé dans les domaines de variables. Nous paramétrons les valeurs de coefficient β pour réguler le degré de chevauchement dans l'étendue de la covariable x_1. Pour simuler le chevauchement « élevé », nous avons établi que $\beta = 1$ (de sorte que $\beta = \alpha_1$); pour le scénario de « petit » chevauchement, nous avons établi que $\beta = -1$.

Le tableau 2.1 présente un résumé des scénarios de simulation examinés, S1 à S4, caractérisés par des combinaisons de chevauchement élevé ou petit et des tailles d'échantillon différentes. Nous avons appliqué les trois méthodes (CLW, pseudo-RLI et RLI) pour chaque scénario afin de choisir entre deux modèles :

$$\begin{aligned} &\text{le modèle complet : } \text{logit}(p_k) = \alpha_0 + \alpha_1 x_{1k} + \alpha_2 x_{2k}, \\ &\text{et le modèle abrégé : } \text{logit}(p_k) = \alpha_0 + \alpha_1 x_{1k}. \end{aligned} \tag{2.3}$$

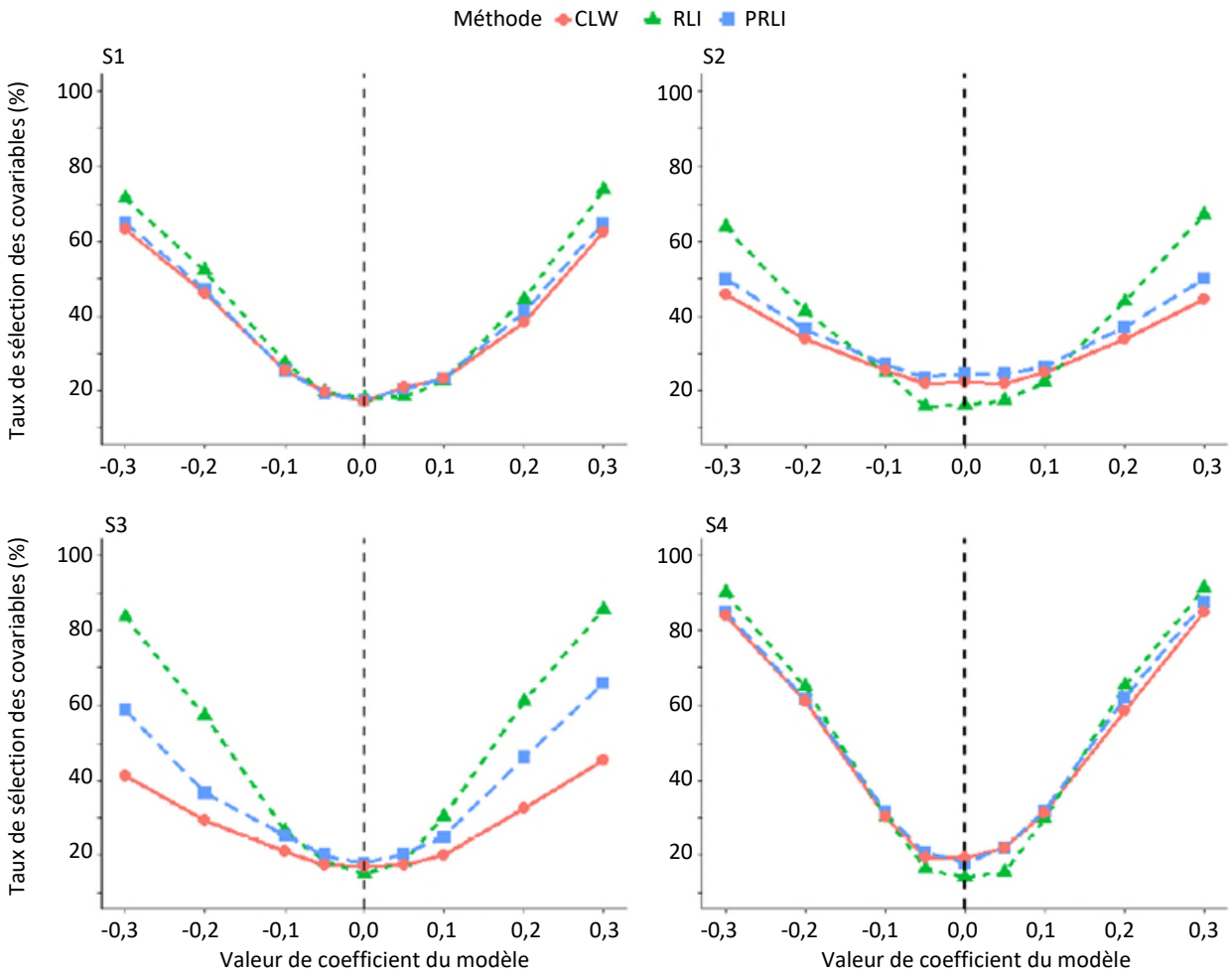
Tableau 2.1
Résumé des scénarios de simulation examinés.

	Chevauchement	n^P	n^{NP}
S1	Élevé	100	100
S2	Petit	100	100
S3	Petit	100	1 000
S4	Petit	1 000	100

Dans chaque cas, nous calculons le AIC pour le modèle complet et le modèle abrégé, et nous choisissons le modèle ayant le AIC le plus petit. Nous répétons cet essai $B = 500$ fois pour chaque scénario et la valeur de α_2, et nous trouvons le pourcentage de fois r que le modèle complet est choisi, $p = 100r/B$.

Les graphiques de la figure 2.1 correspondent aux quatre scénarios du tableau 2.1. Nous représentons graphiquement le pourcentage p par rapport aux valeurs de coefficient α_2. Pour les valeurs absolues plus élevées de α_2, le pourcentage plus élevé p serait préférable; quand α_2 est proche de 0, les valeurs plus petites de p indiqueraient un meilleur rendement. La courbe aux points rouges montre les résultats de la méthode de CLW, la courbe aux carrés bleus, ceux de la méthode de pseudo-RLI, et la courbe aux triangles verts, ceux de la méthode de RLI.

Figure 2.1 Rendement relatif du AIC selon les scénarios S1 à S4 pour les méthodes de CLW (points rouges), de RLI (triangles verts) et de pseudo-RLI (carrés bleus).



Pour tous les scénarios examinés, la méthode de RLI présente le meilleur rendement : pour les valeurs absolues plus grandes de α_2, le AIC fondé sur la RLI choisit plus fréquemment le modèle complet; quand α_2 est plus proche de 0, le critère fondé sur la RLI sélectionne le modèle complet le moins de fois. Dans le scénario de chevauchement élevé (S1) ou celui de taille relativement grande de l'échantillon probabiliste (S4), les trois méthodes produisent des résultats semblables. Cependant, quand l'échantillon probabiliste est relativement petit et que le chevauchement est petit (S2 et S3), le rendement de l'essai fondé sur la méthode de RLI est nettement meilleur que celui des deux autres méthodes. Dans la plupart des cas, l'essai fondé sur la méthode de pseudo-RLI est légèrement meilleur que la méthode de CLW. Quand la taille de l'échantillon non probabiliste est grande par rapport à celle de l'échantillon probabiliste (S3), la différence de rendement augmente : quand la valeur absolue de α_2 est proche de 0,3, le critère fondé sur la méthode de CLW choisit le modèle complet dans seulement environ 40 à 50 % des cas, tandis que le critère fondé sur la méthode de pseudo-RLI le choisit dans environ 60 % des cas et l'essai fondé sur la méthode de RLI, dans approximativement 85 % des cas.

3. Conclusions

Nous félicitons les auteurs pour leur contribution à l'adaptation des algorithmes de sélection de modèles aux problèmes d'intégration de données. À cette fin, le choix d'une fonction « objectif » est important. Dans notre analyse, nous avons examiné plusieurs autres fonctions de vraisemblance proposées récemment. La méthode de RLI fondée sur la vraisemblance exacte a offert un meilleur rendement que les solutions de rechange fondées sur la pseudo-vraisemblance dans un cas pratique important qui se rapporte à de petites tailles d'échantillon probabiliste et à un petit chevauchement dans les domaines de covariables.

Nous constatons un désavantage de la méthode de CLW quand l'échantillon probabiliste est petit et que l'échantillon non probabiliste est relativement grand. Dans ce cas, nous avons également remarqué des problèmes de convergence avec l'algorithme de Newton-Raphson dans la méthode de CLW.

La méthode de RLI exige que les probabilités d'inclusion de l'échantillon probabiliste soient disponibles pour les unités de l'échantillonnage non probabiliste. Si ces probabilités peuvent être calculées à partir des variables disponibles du plan de sondage, la méthode de RLI serait alors la méthode privilégiée. Sinon, la méthode de pseudo-RLI semble être une option viable.

Bibliographie

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. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-fra.pdf.

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

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.

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

Elliott, M.R., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27-38.

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

Heinze, G., et Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21, 2409-2419.

Savitsky, T.D., Williams, M.R., Gershunskaya, J. et Beresovsky, V. (2023). Methods for combining probability and nonprobability samples under unknown overlaps. *Statistics in Transition New Series*, 24(5), 1-34.

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(4), 5237-5250.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 133-138
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Traitement d'échantillons non probabilistes en pondérant par l'inverse de la probabilité, avec application aux données recueillies par approche participative de Statistique Canada »

Changbao Wu[1]

Résumé

Nous proposons des comparaisons entre trois méthodes paramétriques d'estimation des probabilités de participation ainsi que de brefs commentaires à propos des groupes homogènes et de la poststratification.

Mots-clés : Pondération de probabilité inverse; probabilité de participation; échantillon groupé; poststratification; pseudo-maximum de vraisemblance; échantillon probabiliste de référence.

Beaumont, Bosa, Brennan, Charlebois et Chu (2024) ont présenté un examen approfondi des méthodes de pondération de probabilité inverse pour les échantillons non probabilistes, qui comprend des méthodes paramétriques et des méthodes de classification par arborescence et qui met principalement l'accent sur la sélection de variables. Il s'agit d'un sujet de recherche important, étant donné l'augmentation constante des demandes d'utilisation d'échantillons non probabilistes dans les domaines appliqués et les statistiques officielles depuis quelques années. Le présent article constitue une contribution opportune à l'étude et à la comparaison de différentes méthodologies au moyen d'un ensemble de données réelles. Je tiens à remercier le rédacteur en chef invité, M. Partha Lahiri, de son invitation. Je suis heureux de l'occasion qui m'est offerte de participer à une courte analyse à ce propos. À la section 1, je présente des comparaisons entre trois méthodes paramétriques d'estimation des probabilités de participation et de pondération de probabilité inverse. Je formule également de brefs commentaires sur l'utilisation de groupes homogènes aux fins de poststratification à la section 2.

1. Les méthodes de Chen, Li et Wu (2020), de Valliant et Dever (2011) et de Wang, Valliant et Li (2021)

Il s'agit de trois méthodes paramétriques fréquemment mentionnées dans les ouvrages récents sur la pondération de probabilité inverse au moyen des probabilités de participation estimées pour des échantillons non probabilistes. Les trois méthodes comportent des différences conceptuelles ainsi que des similitudes concernant les résultats numériques quand la taille de l'échantillon est petite par rapport à la taille de la population.

1. Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (Ontario) N2L 3G1. Courriel : cbwu@uwaterloo.ca.

Le fondement des estimateurs pondérés de probabilité inverse (PPI) a été construit selon un échantillonnage probabiliste au moyen de l'estimateur de Horvitz-Thompson. En suivant la notation de Beaumont et coll. (2024), supposons que $U = \{1, 2, \ldots, N\}$ est la population finie de taille N. Soit d le plan de sondage probabiliste pour la sélection d'un échantillon probabiliste s. Selon le plan de sondage probabiliste, la variable de l'indicateur d'inclusion de l'échantillon $\delta_k = I(k \in s)$ est définie pour chaque unité k dans la population cible U, c'est-à-dire pour $k = 1, 2, \ldots, N$, où $I(\cdot)$ est la fonction indicatrice, et les probabilités d'inclusion de l'échantillonnage $\pi_k = P(\delta_k = 1 \mid U) = P(k \in s \mid U)$ peuvent être calculées en fonction du plan de sondage donné, d. La distribution conjointe de $(\delta_1, \delta_2, \ldots, \delta_N)$, selon un échantillonnage répété, définit les caractéristiques du plan de sondage, et l'estimateur de Horvitz-Thompson $\hat{\theta}_{\text{HT}} = \sum_{k \in s} y_k / \pi_k$ pour le total de la population $\theta = \sum_{k \in U} y_k$ est uniquement sans biais par rapport au plan parmi une classe d'estimateurs linéaires. L'établissement de ce résultat fondamental dans l'échantillonnage probabiliste suppose deux choses : i) l'hypothèse de positivité, c'est-à-dire $\pi_k > 0$ pour tous les k dans U, de sorte que $\hat{\theta}_{\text{HT}}$ puisse être réécrit comme étant $\hat{\theta}_{\text{HT}} = \sum_{k \in U} \delta_k y_k / \pi_k$; ii) l'indicateur d'inclusion de l'échantillon δ_k est indépendant de y_k et $E_d(\delta_k \mid U) = \pi_k$, où l'espérance E_d est prise par rapport au plan de sondage d.

Soit s_{NP} un échantillon non probabiliste de taille n_{NP} tiré de la population U. Supposons que $\{(y_k, \mathbf{x}_k), k \in s_{\text{NP}}\}$ est l'ensemble de données de l'échantillon non probabiliste. Encore une fois, l'indicateur de participation de l'échantillon $\delta_k = I(k \in s_{\text{NP}})$ est défini pour chaque unité k dans la population cible U, c'est-à-dire pour $k = 1, 2, \ldots, N$. Contrairement à l'échantillonnage probabiliste, les probabilités de participation $p_k = P(\delta_k = 1 \mid U)$ pour l'échantillon non probabiliste s_{NP} sont inconnues et doivent donc être estimées, ce qui nécessite des hypothèses sous la forme p_k et un modèle supposé, désigné par q, pour le mécanisme de participation. Le modèle q mène aux spécifications de la distribution conjointe de $(\delta_1, \delta_2, \ldots, \delta_N)$. Deux composantes sont couramment supposées pour le modèle q : i) les probabilités de participation satisfont $p_k = P(\delta_k = 1 \mid \mathbf{x}_k, y_k) = P(\delta_k \mid \mathbf{x}_k) > 0, \ i = 1, 2, \ldots, N$; ii) les indicateurs d'inclusion de l'échantillon $\delta_1, \delta_2, \ldots, \delta_N$ sont conditionnellement indépendants étant donné $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$.

Il convient de souligner le fait que les estimateurs PPI pour les échantillons non probabilistes doivent être construits et évalués selon le modèle supposé q sur le mécanisme de participation et selon la distribution induite sur $(\delta_1, \delta_2, \ldots, \delta_N)$. C'est là que les différences conceptuelles entre les trois méthodes peuvent être clairement établies. La méthode de Chen et coll. (2020) pour estimer p_k est fondé sur la fonction de vraisemblance complète $\prod_{k=1}^{N} p_k^{\delta_k} (1 - p_k)^{1-\delta_k}$. Dans le cas d'une forme paramétrique prédéfinie $p_k = p(\mathbf{x}_k, \boldsymbol{\alpha})$, l'estimateur par pseudo-maximum de vraisemblance $\hat{\boldsymbol{\alpha}}$ est calculé et évalué selon un modèle paramétrique supposé q sur $(\delta_1, \delta_2, \ldots, \delta_N)$ ainsi que le plan de sondage, d, pour l'échantillon probabiliste de référence s_P. L'estimateur PPI $\hat{\theta}_{\text{PPI}} = \sum_{k \in s_{\text{NP}}} y_k / \hat{p}_k$, où $\hat{p}_k = p(\mathbf{x}_k, \hat{\boldsymbol{\alpha}})$, est convergent pour θ selon la randomisation conjointe du modèle q et le plan de sondage d. Les résultats sont rigoureusement établis, sans restriction sur la « fraction de sondage » n_{NP} / N ou la forme paramétrique pour $p_k = p(\mathbf{x}_k, \boldsymbol{\alpha})$, qu'elle

découle d'un modèle de régression logistique ou de tout autre modèle convenant à une variable réponse binaire.

L'article de Valliant et Dever (2011) a représenté la première tentative sérieuse d'aborder l'estimation des probabilités de participation dans le scénario à deux échantillons décrit dans Beaumont et coll. (2024). Par la suite, plusieurs articles s'en sont inspirés, notamment de la part de Chen et coll. (2020) et de Wang et coll. (2021). La méthode proposée était fondée sur l'ajustement d'un modèle de régression logistique pondéré par les poids d'enquête à l'échantillon groupé $s_{\text{NP}} \cup s_P$, la « variable réponse » étant définie comme $D_k = 1$ si $k \in s_{\text{NP}}$ et $D_k = 0$ si $k \in s_P$, pour $k \in s_{\text{NP}} \cup s_P$, selon la supposition d'une absence de chevauchement entre s_{NP} et s_P. Il est évident que les D_i sont définis au moyen des s_{NP} et s_P donnés et qu'ils sont conceptuellement différents des indicateurs de participation de l'échantillon $(\delta_1, \delta_2, \ldots, \delta_N)$. Dans un modèle paramétrique supposé ξ sur les D_k où $\pi(\mathbf{x}_k, \boldsymbol{\gamma}) = P(D_k = 1 \mid s_{\text{NP}} \cup s_P)$, l'« estimateur PPI théorique » $\hat{\theta} = \sum_{k \in s_{\text{NP}}} y_k / \pi(\mathbf{x}_k, \boldsymbol{\gamma})$ doit être évalué d'abord par rapport au modèle ξ, ce qui donne $E_\xi(\hat{\theta} \mid s_{\text{NP}}, s_P) = E_\xi\left\{\sum_{k \in s_{\text{NP}} \cup s_P} D_k y_k / \pi(\mathbf{x}_k, \boldsymbol{\gamma}) \mid s_{\text{NP}}, s_P\right\} = \sum_{k \in s_{\text{NP}} \cup s_P} y_k$. En créant un ensemble de poids pour l'échantillon groupé $s_{\text{NP}} \cup s_P$ pour l'analyse de régression logistique « pondérée par les poids d'enquête » sur les D_k et sans connaissance préalable de la manière dont s_{NP} a été sélectionné, Valliant et Dever (2011) ont simplement attribué la valeur « 1 » à chaque $k \in s_{\text{NP}}$, ce qui suppose essentiellement que les unités dans s_{NP} sont échangeables pour ce qui est du modèle q. L'estimateur PPI de Valliant et Dever (2011), quand il est évalué selon le modèle q pour $(\delta_1, \delta_2, \ldots, \delta_N)$, n'est pas convergent sauf si s_{NP} est un échantillon aléatoire simple tiré de la population, comme le montrent Chen et coll. (2020).

Dans l'article le plus récent de Wang et coll. (2021), une stratégie qui va dans le sens opposé à celle de Valliant et Dever (2011) a été adaptée. Au lieu de regrouper les deux échantillons, les auteurs ont d'abord créé une population élargie $s^*_{\text{NP}} \cup U$, où s^*_{NP} se compose du même ensemble d'unités que dans $s_{\text{NP}} \subset U$, mais ces unités sont considérées différemment dans l'union de s^*_{NP} et U. Les auteurs ont défini la variable indicatrice $R_k = 1$ si $k \in s^*_{\text{NP}}$ et $R_k = 0$ si $k \in U$, et ont défini davantage la probabilité $\pi_k = P(R_k = 1 \mid s^*_{\text{NP}} \cup U) = P(k \in s^*_{\text{NP}} \mid k \in s^*_{\text{NP}} \cup U)$, pour tous les $k \in s^*_{\text{NP}} \cup U$. J'ai éprouvé beaucoup de difficulté à mettre cette formulation dans un cadre conceptuel approprié, puisque la variable indicatrice R_k dépend de $s_{\text{NP}} = \{i \mid i \in U \text{ et } \delta_i = 1\}$, qui dépend du vecteur complet des indicateurs d'inclusion de l'échantillon $(\delta_1, \delta_2, \ldots, \delta_N)$. Le type de modèle de probabilité qui sous-tend $P(\cdot)$ dans la définition de $\pi_k = P(R_k = 1 \mid s^*_{\text{NP}} \cup U)$ n'apparaît pas clairement. Cela m'a conduit à une longue réflexion pour essayer de comprendre les arguments qui sous-tendent l'identité $p_k = P(\delta_k = 1 \mid U) = \pi_k / (1 - \pi_k)$ (Wang et coll., 2021, page 5 241, équation (9)). L'identité suppose qu'un modèle de régression logistique sur les R_k entraînerait un modèle sur les δ_k au moyen de la fonction de lien logarithmique, ce qui constituerait une source potentielle de préoccupations quand la fraction de sondage n_{NP}/N est grande.

De fait, les trois méthodes produisent des résultats numériques semblables pour les probabilités de participation estimées quand la fraction de sondage n_{NP}/N est petite. Cela peut s'expliquer par la

vérification des calculs dans chacune des méthodes. Selon un modèle paramétrique $p_k = p(\mathbf{x}_k, \boldsymbol{\alpha})$ et l'hypothèse de l'indépendance conditionnelle de $(\delta_1, \delta_2, \ldots, \delta_N)$ étant donné $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, la fonction de logarithme du rapport de vraisemblance complète pour α est donnée par

$$\ell(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\{p(\mathbf{x}_k, \boldsymbol{\alpha})\} + \sum_{k \in U - s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}. \tag{1.1}$$

Le deuxième terme du côté droit de (1.1), désigné par

$$L_2 = \sum_{k \in U - s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\},$$

n'est pas calculable à partir de l'échantillon non probabiliste s_{NP}, puisqu'il suppose $\mathbf{x}_k$ pour tous les k qui ne sont pas dans l'échantillon s_{NP}. Bien qu'elles soient distinctes sur le plan conceptuel, en matière de calcul, les trois méthodes diffèrent seulement par rapport à la façon dont le terme L_2 est traité.

Soit $\{(\mathbf{x}_k, d_k), k \in s_P\}$ l'ensemble de données de l'échantillon probabiliste de référence, où les d_k sont les poids d'enquête pour s_P. En traitant L_2 comme un total d'une population de taille $N - n_{\text{NP}}$, la méthode de Valliant et Dever (2011) équivaut à estimer L_2 par

$$\hat{L}_2^{(1)} = \sum_{k \in s_P} w_k \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\},$$

où $w_k = d_k(\hat{N} - n_{\text{NP}})/\hat{N}$ sont les poids remis à l'échelle qui satisfont $\sum_{k \in s_P} w_k = \hat{N} - n_{\text{NP}}$ et $\hat{N} = \sum_{k \in s_P} d_k$ est la taille estimée de la population pour U. En observant que L_2 peut être réécrit comme étant $L_2 = \sum_{k \in U} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\} - \sum_{k \in s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}$, la méthode de Chen et coll. (2020) remplace L_2 par

$$\hat{L}_2^{(2)} = \sum_{k \in s_P} d_k \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\} - \sum_{k \in s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\},$$

qui est sans biais par rapport au plan (ou convergent par rapport au plan de sondage, selon que les d_k sont les poids de sondage de base ou les poids calés ou ajustés), indépendamment de la fraction de sondage n_{NP}/N. La méthode de Wang et coll. (2021) pour estimer $\pi_k = P(R_k = 1 | s_{\text{NP}} \cup U)$ revient à remplacer L_2 par

$$\hat{L}_2^{(3)} = \sum_{k \in s_P} d_k \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}.$$

Cela dépasse nettement la cible, puisque $\hat{L}_2^{(3)}$ est une estimation pour $\sum_{k \in U} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}$ (ou cela est en dessous de la cible si nous tenons compte du fait que $\log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\} < 0$ pour tous les k). Cependant, l'utilisation de $\hat{L}_2^{(3)}$ pour remplacer L_2 dans $\ell(\boldsymbol{\alpha})$ donne une fonction de logarithme du rapport de vraisemblance qui ressemble à un scénario hypothétique où l'échantillon s_{NP} est tiré d'une population plus grande $s_{\text{NP}} \cup U$. Il faut ajuster le π_k qui en résulte pour le rapprocher de la probabilité réelle de participation p_k.

Il est évident que les trois quantités $\hat{L}_2^{(1)}$, $\hat{L}_2^{(2)}$ et $\hat{L}_2^{(3)}$ ne diffèrent pas trop quand la fraction de sondage n_{NP}/N est petite, ce qui donne des probabilités de participation estimées semblables dans ces scénarios. La dernière étape d'ajustement de la méthode de Wang et coll. (2021), c'est-à-dire $p_k = \pi_k/(1-\pi_k)$, donne également des résultats semblables, puisque nous avons habituellement $\pi_k = O(n_{\text{NP}}/N)$, ce qui signifie $p_k/\pi_k = 1+o(1)$ quand $n_{\text{NP}}/N = o(1)$.

2. Groupes homogènes et poststratification

Dans la pratique, les variables auxiliaires qui sont prises en compte dans les enquêtes probabilistes ou non probabilistes sont souvent catégoriques ou ordinales, particulièrement dans les enquêtes portant sur des populations humaines pour lesquelles on recueille régulièrement des renseignements de base sur les variables démographiques et les indicateurs socioéconomiques. Quand les variables pertinentes pour caractériser le mécanisme de participation sont discrètes, l'estimateur PPI équivaut à un estimateur poststratifié; voir, par exemple, l'analyse détaillée proposée dans la section 5 de Wu (2022). L'estimateur poststratifié (PPI) repose sur des probabilités de participation uniformes dans la même strate de second niveau, ce qui élimine efficacement l'effet des valeurs extrêmes des probabilités de participation estimées qui apparaissent souvent avec un modèle paramétrique en présence de variables auxiliaires continues, et quand l'estimateur a une forme simple et facile à utiliser.

Toutefois, la formation de groupes homogènes en poststrates pose deux difficultés majeures. La première est le grand nombre de groupes initiaux quand de nombreuses variables auxiliaires discrètes sont disponibles dans les ensembles de données. Les méthodes de sélection de variables dont il est question dans Beaumont et coll. (2024) pourraient avoir une utilité pratique pour réduire le nombre de groupes pour l'estimateur PPI définitif. La deuxième est posée par les scénarios dans lesquels il y a un grand nombre de variables auxiliaires continues et discrètes mixtes. Certaines méthodologies élaborées dans la littérature sur les données manquantes et l'inférence causale pourraient être adaptées aux échantillons non probabilistes. La section 5 de Wu (2022) contient une brève analyse sur les méthodes fondées sur le rang. La méthode fondée sur le rang décrite dans Beaumont et coll. (2024) est dans le même esprit. En raison de son importance, ce sujet nécessiterait des recherches plus poussées.

La sélection de variables au moyen du critère d'information d'Akaike ou d'autres critères similaires nécessite une fonction de vraisemblance réelle. Beaumont et coll. (2024) ont démontré l'utilité de la fonction de pseudo-vraisemblance de Chen et coll. (2020) dans le contexte de la sélection de variables. Je me réjouis à l'idée de voir des progrès dans cette direction et j'ai hâte de voir plus d'efforts de recherche menés à cette fin.

Remerciements

La présente recherche a été soutenue par des subventions du Conseil de recherches en sciences naturelles et en génie du Canada et de l'Institut canadien des sciences statistiques.

Bibliographie

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. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-fra.pdf.

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

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

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

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

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 139-160
Statistique Canada, nº 12-001-X au catalogue

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

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

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

Résumé

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

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

1. Observations générales

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

La variance (2.6) est minimisée quand

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Lissage des poids

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

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

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

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

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

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

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

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

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

2.3 Sélection des variables

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5. Méthode de vraisemblance de l'échantillon

5.1 Chevauchement connu entre les deux échantillons

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

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

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

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

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

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

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

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

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

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

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

où

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

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

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

5.2 Chevauchement inconnu entre les deux échantillons

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.3 Sélection des variables

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

6. Fonction d'estimation unifiée

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

7. Conclusion

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

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

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

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

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

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

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

Bibliographie

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 39, No. 4, December 2023

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 40, No. 1, March 2024

Letter to Editor

Articles

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 51, No. 3, September/septembre 2023

Special Issue in Honour of Nancy Reid/Numéro spécial en l'honneur de Nancy Reid

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 51, No. 4, December/décembre 2023

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Les auteurs désirant faire paraître un article sont invités à le soumettre en français ou en anglais via le **portail de *Techniques d'enquête* sur le site Web de ScholarOne Manuscripts** (https://mc04.manuscriptcentral.com/surveymeth). Avant de soumettre l'article, prière d'examiner un numéro récent de *Techniques d'enquête* et de noter les points ci-dessous. Les articles doivent être soumis en Word ou Latex, préférablement en Word avec MathType pour les expressions mathématiques. Une version pdf est également requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé et introduction

2.1 Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
2.2 Le dernier paragraphe de l'introduction devrait contenir une brève description de chacune des sections.

3. Rédaction

3.1 Éviter les notes au bas des pages et les abréviations.
3.2 Limiter l'utilisation d'acronymes. Si un acronyme est utilisé, il doit être défini lors de sa première utilisation.
3.3 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$, etc.
3.4 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées par un chiffre arabe à la droite si l'auteur y fait référence plus loin. Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, l'équation (4.2) est la deuxième équation importante de la section 4.
3.5 Des caractères gras devraient normalement être utilisés pour distinguer les vecteurs et les matrices des valeurs scalaires.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés avec des chiffres arabes et porter un titre aussi explicatif que possible en haut des tableaux ou des figures. Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, le tableau 3.1 est le premier tableau de la section 3.
4.2 Une description textuelle détaillée des figures pourrait être requise à des fins d'accessibilité si le message transmis par l'image n'est pas suffisamment expliqué dans le texte.

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple : Cochran (1977, page 164).
5.2 La première fois qu'une référence est citée dans le texte, le nom de chacun des auteurs doit être écrit. Pour les fois suivantes, le nom de chacun des auteurs peut être écrit à nouveau. Cependant, si la référence contient trois auteurs ou plus, les noms du deuxième auteur et des auteurs suivants peuvent être remplacés par « et coll. ».
5.3 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les articles d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

6. Communications brèves

6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots, incluant les tableaux, les figures et la bibliographie.