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

Techniques d'enquête

Plans de sondage et estimateurs pour les enquêtes multimodales fondées sur la collecte de données en personne

par J. Michael Brick et Jill M. DeMatteis

Date de diffusion : le 3 janvier 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.

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par le ministre de l'Industrie, 2024

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

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2023
Vol. 49, nº 2, pp. 513-538
Statistique Canada, nº 12-001-X au catalogue

Plans de sondage et estimateurs pour les enquêtes multimodales fondées sur la collecte de données en personne

J. Michael Brick et Jill M. DeMatteis[1]

Résumé

Les spécialistes de la recherche sur les enquêtes se tournent de plus en plus vers la collecte multimodale de données pour composer avec la baisse des taux de réponse aux enquêtes et l'augmentation des coûts. Une approche efficace propose des modes de collecte moins coûteux (par exemple sur le Web) suivis d'un mode plus coûteux pour un sous-échantillon des unités (par exemple les ménages) dans chaque unité primaire d'échantillonnage (UPE). Nous présentons deux solutions de rechange à cette conception classique. La première consiste à sous-échantillonner les UPE plutôt que les unités pour limiter les coûts. La seconde est un plan hybride qui comprend un échantillon (à deux degrés) par grappes et un échantillon indépendant sans mise en grappes. À l'aide d'une simulation, nous démontrons que le plan hybride comporte des avantages considérables.

Mots-clés : Biais; à plusieurs degrés; sous-échantillonnage; à deux phases; axé sur le Web.

1. Introduction

De nombreuses enquêtes reposent sur les modes de collecte de données par Internet et par la poste en raison de leurs coûts relativement faibles, mais les taux de réponse globaux peuvent être inférieurs à ceux qui seraient souhaités, et les estimations peuvent être entachées d'un biais de non-réponse considérable (Dillman, 2017; Brick, Kennedy, Flores-Cervantes et Mercer, 2021). L'introduction des interviews sur place comme mode de suivi peut largement augmenter les taux de réponse, mais à un coût plus élevé. L'augmentation des taux de réponse obtenue grâce au suivi en personne entraîne également une réduction du biais de non-réponse, en raison de l'introduction du mode en personne, ce qui crée une propension accrue à répondre aux enquêtes pour les sous-groupes qui sont moins susceptibles de répondre sur le Web ou par la poste. L'échantillonnage à deux phases est une méthode d'échantillonnage plus rentable de la collecte multimodale de données et d'interviews sur place. Hansen et Hurwitz (1946) ont présenté l'échantillonnage à deux phases comme méthode de traitement de la non-réponse fondée sur la théorie de l'échantillonnage à deux phases de Neyman (1938). Hansen et Hurwitz ont appliqué leur méthode au moyen d'un échantillon de première phase constitué d'établissements de vente au détail qui ont reçu un questionnaire par la poste, les non-répondants ayant été sous-échantillonnés pour la collecte en personne. Comme le volet en personne a abouti à un taux de réponse de presque 100 % après pondération pour le sous-échantillonnage, les estimations de l'enquête étaient sans biais.

Une application bien connue de cette stratégie d'échantillonnage est l'American Community Survey (ACS), une enquête où un échantillon de ménages est sélectionné dans les régions de totalisation et où les ménages doivent répondre à un questionnaire sur le Web ou par la poste; un suivi en personne se fait auprès

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

d'un sous-échantillon de non-répondants du secteur (Bureau du recensement des États-Unis, 2014; Bureau du recensement des États-Unis, 2019). Comme d'autres enquêtes aujourd'hui, l'ACS ne parvient pas à un taux de réponse complet, si bien que les poids d'enquête doivent être rajustés en fonction de la non-réponse et que les estimations peuvent rester entachées d'un biais de non-réponse.

Le présent article porte sur deux nouveaux plans de sondage pour les enquêtes multimodales auprès des ménages qui font appel à des interviews sur place pour accroître les taux de réponse. Les enquêtes axées sur ce type de plan multimodal sont peut-être plus courantes aux États-Unis et au Canada. L'approche type d'échantillonnage à deux phases consiste à sélectionner un échantillon d'unités primaires d'échantillonnage (UPE) et de ménages dans les UPE pour le ou les modes à faible coût, puis des sous-échantillons dans les diverses UPE pour le suivi en personne. Les deux nouvelles approches modifient les plans d'échantillonnage pour les modes à faible coût (sur le Web ou par la poste) afin d'améliorer l'efficacité des estimations, tout en conservant le plan de base en personne. Les plans d'échantillonnage qui reposent sur des modes plus coûteux avant les modes moins coûteux ne sont pas pris en considération (par exemple Bayart et Bonnel, 2015). Le premier nouveau plan de sondage que nous examinons est une variante où un grand échantillon d'UPE est prélevé et où les ménages de chaque UPE sont recrutés au moyen de modes à faible coût en première phase, mais où le suivi en personne se fait seulement dans un sous-échantillon d'UPE. Cette approche réduit l'effet de grappes en répartissant les répondants de l'échantillon de première phase sur un plus grand nombre d'UPE. Nous envisageons également une seconde approche qui consiste à sélectionner deux échantillons indépendants : un échantillon de ménages sans mise en grappes recrutés au moyen de modes à faible coût et un échantillon par grappes où tous les ménages échantillonnés sont recrutés de façon séquentielle au moyen de tous les modes. Les estimations se créent en regroupant des données des deux échantillons. Nous ne discutons pas des erreurs non dues à l'échantillonnage qui peuvent survenir lorsque plus d'un mode de collecte de données est utilisé (par exemple Goodman, Brown, Silverwood, Sakshaug, Calderwood, Williams et Ploubidis, 2022).

À la section 2, nous décrivons plus en détail les deux nouveaux plans de sondage. Nous présentons les estimateurs et décrivons leurs propriétés selon différents modèles de non-réponse à la section 3. La section 4 présente et décrit les résultats d'une étude par simulation que nous avons menée. Nous concluons à la section 5 par une discussion sur les répercussions des plans de sondage et des estimateurs, la présentation de recommandations et l'évocation de domaines de recherche future.

2. Plans de sondage

Toutes les applications que nous avons identifiées de l'échantillonnage à deux phases des ménages pour la non-réponse où des interviews sur place sont effectuées commencent par un échantillon de première phase

d'UPE (par grappes sur le plan géographique) et enchaînent avec un sous-échantillon de non-répondants en deuxième phase dans chacune des UPE de la première phase. Dans de nombreux cas, le sous-échantillonnage pour la non-réponse est une caractéristique adaptative ou réactive du plan plutôt qu'une approche de conception initiale (Groves et Heeringa, 2006; Heeringa, Wagner, Torres, Duan, Adams et Berglund, 2004; Wagner, Arrieta, Guyer et Ofstedal, 2014). La première phase consiste en un échantillon d'UPE avec un échantillon de ménages tirés d'une base de sondage fondée sur les adresses. Tous les ménages échantillonnés sont assujettis au protocole initial de collecte des données, qui peut comprendre la collecte sur le Web ou par la poste comme mode à faible coût.

Sous-échantillonnage d'unités à deux phases : Les non-répondants de l'échantillon de première phase sont sous-échantillonnés à l'intérieur de chaque UPE, et l'interview sur place est le mode retenu en seconde phase. L'ACS utilise ce plan. Nous appelons cette approche classique le *sous-échantillonnage d'unités à deux phases*. Särndal et Swensson (1987) ont étendu la théorie de Hansen et Hurwitz à des plans où l'échantillon de première phase n'était pas un échantillon aléatoire simple ou stratifié.

Le plan de sous-échantillonnage d'unités à deux phases pose deux problèmes statistiques. Le premier concerne l'augmentation de la variance des estimations attribuable au sous-échantillonnage (Kish, 1992), puisque les poids rendent compte du sous-échantillonnage. Le second est l'augmentation de la variance attribuable aux grappes, à supposer qu'il y ait une corrélation intraclasse positive pour les caractéristiques d'intérêt, parce que toutes les réponses sont regroupées dans les UPE échantillonnées. Ajoutons qu'un problème connexe lié aux coûts et au fonctionnement limite le nombre d'UPE échantillonnées en vue de rendre efficace la collecte de données en personne au sein des UPE.

Sous-échantillonnage d'UPE à deux phases : Un autre plan de sondage à deux phases consiste à sélectionner un grand nombre d'UPE et de ménages dans la première phase, puis à sélectionner un sous-échantillon d'UPE et tous les ménages dans ces UPE pour la seconde phase. Tous les ménages échantillonnés dans la première phase sont recrutés par Internet, mais un suivi en personne est effectué seulement pour les non-répondants du sous-échantillon des UPE lors de la deuxième phase. Nous appelons cette approche le *sous-échantillonnage d'UPE à deux phases*. Nous n'avons trouvé aucune enquête faisant appel à cette approche. Celle-ci présente des avantages et soulève des inquiétudes semblables à ceux du sous-échantillonnage d'unités à deux phases, mais offre l'avantage de réduire l'effet de grappes en répartissant les répondants de l'échantillon de première phase sur un plus grand nombre d'UPE.

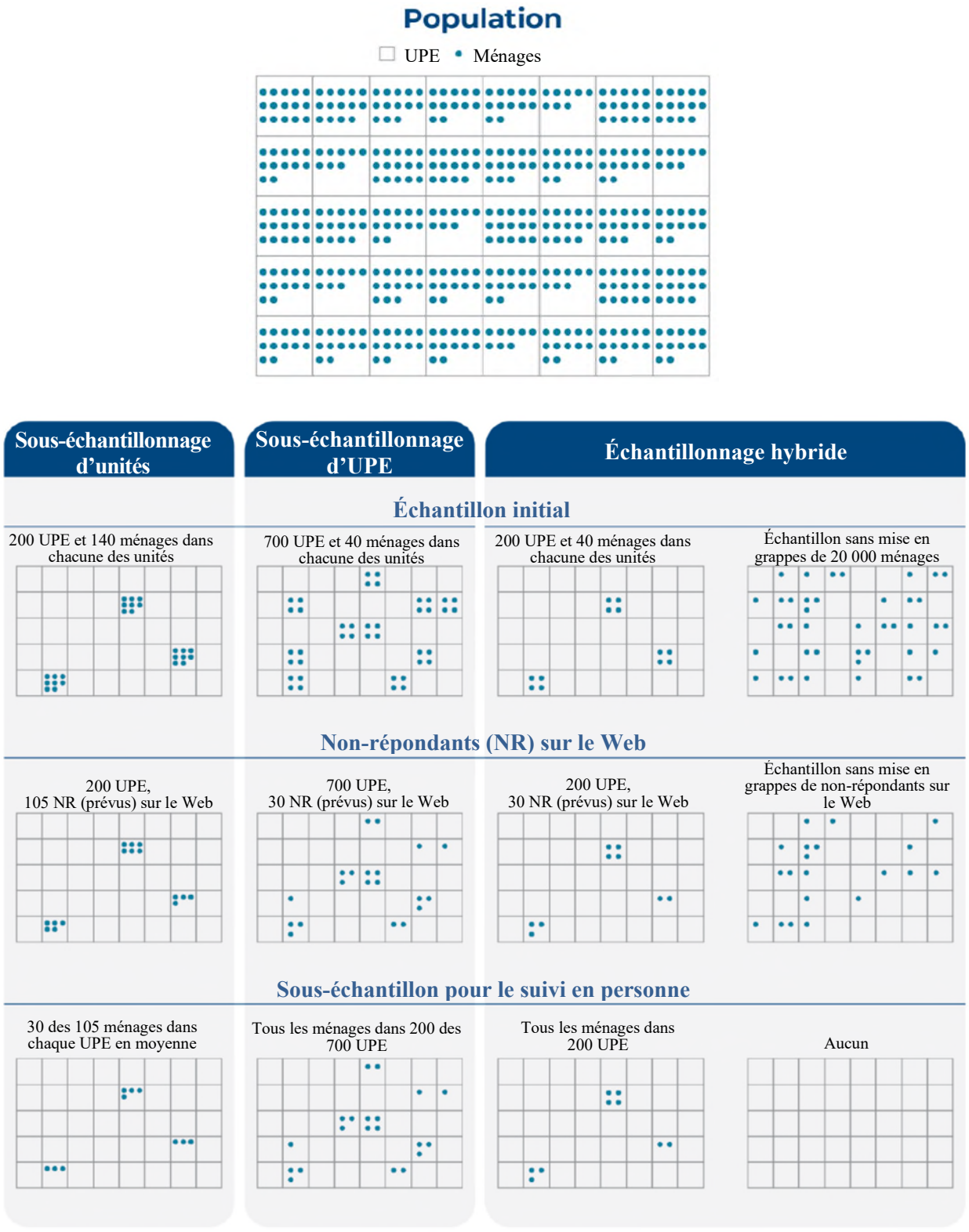
Échantillonnage hybride : Un plan de sondage différent comportant moins de restrictions que le sous-échantillonnage d'UPE à deux phases, qui consiste à prélever deux échantillons indépendants, puis à regrouper les estimations des deux échantillons pour produire les estimations finales. Le premier échantillon est un échantillon sans mise en grappes de ménages tirés de la base de sondage fondée sur les adresses, et

ces ménages sont recrutés uniquement au moyen de modes à faible coût. Le deuxième échantillon repose sur un plan à deux degrés pour réduire les coûts de la collecte des données et consiste en un échantillon d'UPE et de ménages dans les UPE. Les données sont recueillies auprès des ménages faisant partie du deuxième échantillon en utilisant, de manière séquentielle, le mode de collecte sur le Web, puis celui en personne. Ce plan s'appuie sur des méthodes à double base (Lohr, 2011) et à deux phases. Comme les méthodes à double base, il consiste à sélectionner deux échantillons, mais à partir de la même base de sondage dans ce cas-ci, puis à combiner les données des deux échantillons. Tout comme pour l'échantillonnage à deux phases, seul un sous-échantillon de l'échantillon complet des ménages est assujetti au protocole de collecte intégrale des données. Nous appelons cela l'échantillonnage hybride à deux phases, ou de façon plus concise, l'*échantillonnage hybride*. Lorsque nous discutons des deux échantillons, nous les appelons, respectivement, échantillons *sans mise en grappes* et échantillons *par grappes* pour faciliter l'analyse.

Exemple : Nous présentons ci-dessous une illustration de chaque plan de sondage et considérons les effets sur la précision; plus loin, pour chacun de ces plans de sondage, nous examinerons le biais de divers estimateurs. À titre d'exemple simple de ces trois approches, supposons que l'objectif est de réaliser 10 000 interviews auprès des ménages, dont 70 % par Internet et 30 % en personne. Aux fins d'illustration, nous ne tenons pas compte des ajustements de pondération pour la non-réponse et des précisions sur les coûts qui sont abordés plus loin.

Les trois approches pour ce simple exemple sont illustrées à la figure 2.1. L'approche du sous-échantillonnage d'unités à deux phases consiste à prélever un échantillon de 200 UPE selon une probabilité proportionnelle au nombre de ménages, puis à sélectionner un nombre égal de ménages dans chaque UPE. Tous les ménages échantillonnés reçoivent un questionnaire à remplir sur le Web, et un sous-échantillon de non-répondants sur le Web dans chaque UPE fait l'objet d'un suivi en personne. Si nous supposons un taux de réponse de 25 % sur le Web et un taux de réponse de 50 % en personne, l'échantillonnage de 140 ménages par UPE donne 35 questionnaires remplis sur le Web. Le sous-échantillonnage d'environ 30 des quelque 105 non-répondants par UPE donne 15 interviews réalisées sur place pour un total de 50 réponses (35 par Internet et 15 en personne). L'effet de plan attribuable au sous-échantillonnage (pondération différentielle) est d'environ 1,44 (Kish, 1992) et atteint cette valeur maximale lorsque le résultat et les poids ne sont pas en corrélation. L'effet de plan attribuable à la mise en grappes est d'environ $1+\delta(\bar{m}-1)$, où $\bar{m}$ est le nombre moyen de ménages répondants par UPE et δ est la corrélation intraclasse. Selon 50 questionnaires remplis par UPE et $\delta = 0{,}02$, l'effet de plan attribuable à la mise en grappes est d'environ 2,0. L'effet de plan global est le produit des effets de pondération et de grappes, ou 2,9, et donne une taille d'échantillon efficace d'environ 3 500 (10 000/2,9).

Figure 2.1 Illustration de trois approches de suivi.



Notes : UPE = unités primaires d'échantillonnage; NR = non-répondants.

Dans l'approche de sous-échantillonnage d'UPE à deux phases, un échantillon de 700 UPE est sélectionné selon une probabilité proportionnelle au nombre de ménages, et 200 UPE sont sous-échantillonnées selon une probabilité égale. Le nombre d'UPE pour le sous-échantillon a été choisi de façon à égaler le nombre dans le plan de sous-échantillonnage d'unités, et un total de 700 UPE permet de tenir compte de tous les non-répondants dans les 200 UPE sous-échantillonnées, et ce, sans autre sous-échantillonnage à prévoir. Dans chaque UPE, un échantillon équiprobable de 40 ménages est sélectionné. Tous les ménages de l'échantillon reçoivent un questionnaire à remplir sur le Web, ce qui donne environ 7 000 questionnaires remplis sur le Web (700*40*0,25). Tous les non-répondants au questionnaire sur le Web dans le sous-échantillon des 200 UPE font l'objet d'un suivi en personne, ce qui donne 3 000 interviews réalisées sur place. L'effet de plan attribuable à la pondération est toujours de 1,44. En raison du nombre inégal de questionnaires remplis par grappe, nous utilisons l'effet de plan approximatif suggéré par Holt (1980), selon $m' = \sum m_i^2 \Big/ \sum m_i$ au lieu de $\bar{m}$, où m_i est le nombre de questionnaires remplis dans l'UPE i. Dans ce plan de sondage, $m_i = 10$ dans 500 UPE et $m_i = 25$ dans 200 UPE, donc $m' = 17{,}5$ et l'effet de plan attribuable aux grappes est de $1 + \delta\,(m' - 1) = 1{,}33$ lorsque $\delta = 0{,}02$. L'effet de plan global est de 1,9 et la taille d'échantillon efficace est d'un peu plus de 5 200, ce qui représente un important gain de précision par rapport à l'approche classique à deux phases.

Dans le cadre de l'approche hybride, un échantillon de 200 UPE (égal au nombre d'UPE pour lesquelles une interview sur place est réalisée dans les autres plans de sondage) est prélevé et 40 ménages par UPE dans l'échantillon par grappes sont sélectionnés, ce qui donne 2 000 questionnaires remplis sur le Web et 3 000 interviews sur place réalisées. L'échantillon sans mise en grappes est de 20 000, ce qui donne 5 000 questionnaires remplis sur le Web. L'effet de plan attribuable à la mise en grappes dans l'échantillon par grappes est de 1,48 lorsque $\delta = 0{,}02$ et il n'y a pas d'effet de pondération différentielle (en supposant qu'un ajustement uniforme de la non-réponse est appliqué à tous les répondants). En tout, il y a 7 000 questionnaires remplis sur le Web et 3 000 interviews sur place réalisées. On regroupe les deux échantillons en utilisant, par exemple, $\lambda = 0{,}7$ pour l'échantillon sans mise en grappes (puisqu'il représente 70 % du total) et $1 - \lambda = 0{,}3$ pour l'échantillon par grappes, de sorte que l'effet de plan global est de 1,14. (Un facteur de composition optimal, qui tient également compte de l'effet de grappes sur la précision de l'échantillon par grappes, pourrait être déterminé et appliqué; cependant, par souci de simplicité dans cette illustration, nous avons choisi des facteurs de composition proportionnels au nombre de questionnaires remplis dans le mode visé. Nous examinerons les effets des autres facteurs de composition plus loin dans l'article.) La taille de l'échantillon efficace est de 8 770, ce qui représente un important gain de précision par rapport aux deux approches à deux phases.

3. Modèles d'estimation et modèles de non-réponse

Modèles de non-réponse

Dans la documentation sur les modèles de non-réponse aux enquêtes, deux cadres ressortent : un cadre *déterministe* qui partitionne la population en deux groupes complets et s'excluant l'un l'autre, à savoir les

répondants et les non-répondants (comme il est décrit aux pages 359 à 363 dans Cochran, 1977); un cadre *stochastique* dans lequel chaque membre de la population reçoit une probabilité de réponse à une enquête particulière (par exemple voir Brick et Montaquila, 2009).

Dans la vision déterministe de la non-réponse pour un protocole donné de collecte de données, la population comprend un ensemble de ménages qui répondront à une demande sur le Web, un ensemble de ménages qui répondront à la demande en personne après une non-réponse sur le Web et un ensemble de ménages qui ne répondront pas du tout. Selon ce modèle, la population totale est

$$Y = N\left[\gamma_W \bar{Y}_W + \gamma_F \bar{Y}_F + (1-\gamma_W-\gamma_F)\,\bar{Y}_N\right], \tag{3.1}$$

où γ_W est la proportion de répondants par Internet, γ_F est la proportion de répondants en personne, mais non par Internet, et $\bar{Y}_W$, $\bar{Y}_F$ et $\bar{Y}_N$ sont les moyennes de population correspondantes de la caractéristique d'intérêt pour les ensembles de répondants sur le Web et en personne et les ensembles de non-répondants. La vision déterministe à partitionnement fixe de la population selon le mode de réponse dans le protocole de collecte des données est conçue à partir d'une perspective postérieure à la collecte des données, où des applications répétées du même protocole pourraient servir à déterminer ces constantes.

Le modèle stochastique qui s'harmonise à ce modèle déterministe présuppose que chaque unité de la population a un vecteur de propension à répondre, le premier élément étant la probabilité de répondre par Internet et le second, la probabilité de répondre en personne et non par Internet, disons $\phi_k = (\phi_{k,W}, \phi_{k,F})'$, où $0 \le \phi_{k,W} \le 1;\ 0 \le \phi_{k,F} \le 1;\ 0 < \phi_{k,W} + \phi_{k,F} \le 1$. Il faut mentionner que, si nous définissons $\phi_{k,F|W^C}$ comme étant la probabilité conditionnelle que l'unité k réponde en personne étant donné qu'elle n'a pas répondu par Internet, alors $\phi_{k,F} = (1-\phi_{k,W})\,\phi_{k,F|W^C}$. De plus, en prenant les espérances pour cette distribution de la réponse, on obtient $E_R \sum_{k\in U} \phi_{k,W} \equiv \gamma_W N;\ E_R \sum_{k\in U} \phi_{k,W} y_k \equiv Y_W$, où U désigne la population, $Y_W = \gamma_W N \bar{Y}_W$, et $E_R \sum_{k\in U} \phi_{k,F} \equiv \gamma_F N;\ E_R \sum_{K\in U} \phi_{k,F} y_k \equiv Y_F$ où $Y_F = \gamma_F N \bar{Y}_F$. En ayant un taux de réponse complet (c'est-à-dire que $\phi_{k,F|W^C} = 1$, de sorte que $\phi_{k,W} + \phi_{k,F} = 1 \forall k$), nous avons $Y = N[\gamma_W \bar{Y}_W + \gamma_F \bar{Y}_F]$.

Estimateurs

Tout d'abord, supposons que nous effectuons un suivi en personne auprès de tous les non-répondants au questionnaire sur le Web dans un échantillon à deux degrés sans sous-échantillonnage. L'approche typique de l'estimation consiste à utiliser l'estimateur de Horvitz-Thompson (HT) et un ajustement pour la non-réponse. (Dans le cas présent, nous considérons un seul ajustement global, en ne tenant pas compte des ajustements de non-réponse et de calage plus complexes par souci de simplicité.) Cet estimateur du total est

$$\begin{aligned} \hat{t}_1 &= \sum_{k\in S} d_k \delta_k(W) \hat{R}^{-1} y_k \ + \ \sum_{k\in S} d_k \delta_k(F) \hat{R}^{-1} y_k \\ &= \hat{N}\left[\frac{\hat{\gamma}_W}{\hat{\gamma}_W + \hat{\gamma}_F} \bar{y}_W \ + \ \frac{\hat{\gamma}_F}{\hat{\gamma}_W + \hat{\gamma}_F} \bar{y}_F\right], \end{aligned} \tag{3.2}$$

où S désigne l'échantillon, d_k est la réciproque de la probabilité de sélection du ménage k (compte tenu des deux degrés de sélection), $\delta_k(W)=1$ si le ménage k répond à l'enquête sur le Web et 0 autrement, $\delta_k(F)=1$ si le ménage k répond à l'enquête en personne et 0 autrement, $\hat{N}=\sum_{k\in S} d_k$ et $\hat{R}=\hat{R}_W+(1-\hat{R}_W)\hat{R}_F$ selon $\hat{R}_W=\sum_{k\in S} d_k\delta_k(W)\big/\sum_{k\in S} d_k$ et $\hat{R}_F=\sum_{k\in S} d_k\delta_k(F)\big/\sum_{k\in S} d_k(1-\delta_k(W))$ pour les taux de réponse observés sur le Web et en personne. Veuillez noter que $\hat{R}_F$ est le taux de réponse conditionnel étant donné la non-réponse au questionnaire sur le Web. Tout au long du processus, nous estimons que $\hat{\gamma}_W=\hat{R}_W$ et que $\hat{\gamma}_F=(1-\hat{R}_W)\hat{R}_F$ d'après les taux de réponse observés lorsque ceux-ci dépendent du plan de sondage particulier.

L'estimateur $\hat{t}_1$ est sans biais si $\overline{Y}_N=\overline{Y}_F=\overline{Y}_W$ ou si le taux de réponse conditionnel en personne est de 100 %. Cette propriété est facilement démontrée si nous prenons les espérances tant pour la distribution d'échantillonnage que pour la distribution de réponse (Särndal et Swensson, 1987). Nous supposons tout au long du processus que les conditions types pour que l'estimateur de réponse complète approprié soit sans biais se vérifient également.

Contrairement à $\hat{t}_1$, qui applique un ajustement constant $\hat{R}^{-1}$ à tous les répondants, nous considérons également un estimateur couramment utilisé dans le contexte de l'échantillonnage à deux phases, qui permet d'ajuster seulement les répondants en personne :

$$\begin{aligned}\hat{t}_2 &= \sum_{k\in S} d_k\delta_k(W)\,y_k+\sum_{k\in S} d_k\delta_k(F)\hat{R}_F^{-1}y_k\\ &= \hat{N}\left[\hat{\gamma}_W\overline{y}_W+(1-\hat{\gamma}_W)\,\overline{y}_F\right].\end{aligned} \tag{3.3}$$

L'estimateur $\hat{t}_2$ est sans biais si $\overline{Y}_N=\overline{Y}_F$ ou si le taux de réponse conditionnel en personne est de 100 %.

Ensuite, nous examinons les plans de sous-échantillonnage et élargissons ces estimateurs pour tenir compte du sous-échantillonnage. Supposons que ω désigne la probabilité conditionnelle de sélection dans le sous-échantillon pour les interviews en personne. Au moyen du sous-échantillonnage d'unités à deux phases et du taux de sous-échantillonnage en seconde phase de ω $(0<\omega\le 1)$, les estimateurs élargis sont

$$\begin{aligned}\hat{t}_1 &= \sum_{k\in S} d_k\delta_k(W)\hat{R}^{-1}y_k+\sum_{k\in S} d_k\omega^{-1}\delta_k(F)\hat{R}^{-1}y_k\\ &= \hat{N}\left[\frac{\hat{\gamma}_W}{\hat{\gamma}_W+\hat{\gamma}_F}\overline{y}_W+\frac{\hat{\gamma}_F}{\hat{\gamma}_W+\hat{\gamma}_F}\overline{y}_F\right]\end{aligned} \tag{3.4}$$

et

$$\begin{aligned}\hat{t}_2 &= \sum_{k\in S} d_k\delta_k(W)\,y_k+\sum_{k\in S} d_k\omega^{-1}\delta_k(F)\,\hat{R}_F^{-1}y_k\\ &= \hat{N}\left[\hat{\gamma}_W\overline{y}_W+(1-\hat{\gamma}_W)\,\overline{y}_F\right].\end{aligned} \tag{3.5}$$

Dans cette expression, le taux de sous-échantillonnage est pris en considération dans l'estimation de γ_W (le tableau 3.1 montre la formulation de cet estimateur et d'autres estimateurs en ce qui concerne les poids).

Il convient de mentionner que les estimateurs obtenus aux équations (3.2) et (3.3) sont un cas particulier des estimateurs des équations (3.4) et (3.5), respectivement, où $\omega = 1$.

Tableau 3.1
Estimateurs, poids et modèles de non-réponse pour l'échantillonnage à deux phases et l'échantillonnage hybride.

Estimateur	Plan de sondage	Poids des répondants		Modèle de non-réponse
$\hat{t}_1$	À deux degrés	$d_k\hat{R}^{-1}$	Répondants sur le Web dans S_B	$\overline{Y}_N = \overline{Y}_F = \overline{Y}_W$
		$d_k\omega^{-1}\hat{R}^{-1}$	Répondants en personne dans S_B	
$\hat{t}_2$	À deux phases	d_k	Répondants en personne dans S_B	$\overline{Y}_N = \overline{Y}_F$
		$d_k\omega^{-1}\hat{R}_F^{-1}$	Répondants en personne dans S_B	
$\hat{t}_A$	À un degré, sans mise en grappes	$d_k\hat{R}_W^{-1}$	Tous les répondants (Web) dans S_A	$\overline{Y}_N = \overline{Y}_F = \overline{Y}_W$
$\hat{t}_{df,1}$	Hybride	$\lambda d_k\hat{R}_W^{-1}$	Tous les répondants (Web) dans S_A	$\overline{Y}_N = \overline{Y}_F = \overline{Y}_W$
		$(1-\lambda)d_k\hat{R}^{-1}$	Tous les répondants dans S_B	
$\hat{t}_{df,2}$	Hybride	κd_k	Tous les répondants (Web) dans S_A	$\overline{Y}_N = \overline{Y}_F$
		$(1-\kappa)d_k$	Répondants sur le Web dans S_B	
		$d_k\hat{R}_F^{-1}$	Répondants en personne dans S_B	

Pour les plans de sous-échantillonnage, les deux estimateurs sont sans biais si le taux de réponse conditionnel en personne est de 100 %; $\hat{t}_2$ est sans biais si $\overline{Y}_N = \overline{Y}_F$, alors que $\hat{t}_1$ exige la condition plus rigoureuse que $\overline{Y}_N = \overline{Y}_F = \overline{Y}_W$. Ce résultat s'applique au sous-échantillonnage d'unités et d'UPE à deux phases. Bien que Särndal et Swensson (1987) ne se soient pas penchés sur le sous-échantillonnage d'UPE à deux phases, leur preuve d'absence de biais s'applique, puisque le sous-échantillonnage qu'ils ont examiné n'est pas propre à une phase. Ainsi, l'extension à deux phases de l'estimateur $\hat{t}_2$ montré à l'équation (3.5) peut valoir pour le sous-échantillonnage d'unités à deux phases ou le sous-échantillonnage d'UPE à deux phases.

Dans le plan de sous-échantillonnage d'UPE, l'ajustement de la pondération dans $\hat{t}_2$, ω^{-1}, est le rapport entre le nombre d'UPE à la première phase et le nombre d'UPE à la seconde phase. Un autre ajustement de sous-échantillonnage est

$$\omega_s^{-1} = \frac{\sum_{k\in S} d_k\,(1-\delta_k(W))}{\sum_{k\in \text{sub-PSU}} d_k(1-\delta_k(W))}, \tag{3.6}$$

où le numérateur est la somme des poids de tous les répondants échantillonnés qui n'ont pas rempli le questionnaire sur le Web et où le dénominateur est la somme des poids des non-répondants au questionnaire sur le Web dans les UPE sous-échantillonnées. Cet ajustement tient compte de la taille des UPE sous-échantillonnées, cette taille étant le nombre de ménages sous-échantillonnés.

Pour l'échantillonnage hybride, nous utilisons la notation à double base pour simplifier la présentation. Supposons que S_A est un échantillon sans mise en grappes de ménages et S_B, un échantillon à deux degrés

de ménages échantillonnés dans les UPE. Le protocole pour S_A vise simplement le Web, tandis que pour S_B, il vise le Web et l'interview sur place pour tous les non-répondants sur le Web.

Comme S_A vise une collecte de données sur le Web à un seul degré, l'estimateur de Horvitz-Thompson ou l'équation (3.2) comptant les répondants sur le Web seulement est

$$\hat{t}_A = \frac{\sum_{k\in S_A} d_k}{\sum_{k\in S_A} d_k \delta_k(W)} \sum_{k\in S_A} d_k \delta_k(W) y_k = \hat{N}_{S_A} \overline{y}_{WA}, \tag{3.7}$$

où $\hat{N}_{S_A} = \sum_{k\in S_A} d_k$ et $\overline{y}_{WA}$ est la moyenne estimée pour les répondants sur le Web en fonction de l'échantillon sans mise en grappes S_A. Cet estimateur *sans mise en grappes* est sans biais si $\overline{Y}_N = \overline{Y}_F = \overline{Y}_W$, comme nous l'avons mentionné pour $\hat{t}_1$.

Si nous reformulons $\hat{t}_1$ pour un échantillon par grappes S_B, nous obtenons

$$\hat{t}_{B,1} = \hat{N}_{S_B} \left[\frac{\hat{\gamma}_W}{\hat{\gamma}_W + \hat{\gamma}_F} \overline{y}_{WB} + \frac{\hat{\gamma}_F}{\hat{\gamma}_W + \hat{\gamma}_F} \overline{y}_{FB} \right], \tag{3.8}$$

où $\hat{N}_{S_B} = \sum_{k\in S_B} d_k$, et $\overline{y}_{WB}$ et $\overline{y}_{FB}$ sont les moyennes estimées pour les répondants sur le Web et les répondants en personne, respectivement, en fonction de l'échantillon par grappes S_B.

Le regroupement des deux estimateurs dans le cas de l'échantillonnage hybride donne (en supposant que $\hat{N} = \hat{N}_{S_A} = \hat{N}_{S_B}$)

$$\begin{aligned} \hat{t}_{df,1} &= \lambda \hat{t}_A + (1-\lambda)\hat{t}_{B,1} \\ &= \hat{N}\left[\lambda \overline{y}_{WA} + (1-\lambda)\frac{\hat{\gamma}_W}{\hat{\gamma}_W + \hat{\gamma}_F} \overline{y}_{WB} + (1-\lambda)\frac{\hat{\gamma}_F}{\hat{\gamma}_W + \hat{\gamma}_F} \overline{y}_{FB} \right]. \end{aligned} \tag{3.9}$$

Cet estimateur, comme l'équation (3.2), n'accorde pas plus de poids aux répondants en personne et est sans biais lorsque $\overline{Y}_N = \overline{Y}_F = \overline{Y}_W$, comme nous l'avons mentionné précédemment. Le modèle de non-réponse plus rigoureux est nécessaire pour l'absence de biais, car bien que $\hat{t}_{B,1}$ estime le total pour la population de répondants sur le Web et en personne, $\hat{t}_A$ le fait seulement pour les répondants sur le Web.

Une autre possibilité est de regrouper deux estimateurs, mais d'utiliser la version à deux phases de l'équation (3.5) pour l'échantillon par grappes. Tout d'abord, nous regroupons les répondants sur le Web de S_A et de S_B, puis nous ajustons le poids des répondants en personne pour tenir compte de la non-réponse restante. Cet estimateur composite (où seuls les échantillons Web sont regroupés) est

$$\hat{t}_{df,2} = \hat{N}\hat{\gamma}_W \left[\kappa \overline{y}_{WA} + (1-\kappa)\, \overline{y}_{WB} \right] + \hat{N}(1-\hat{\gamma}_W)\, \overline{y}_{FB}. \tag{3.10}$$

Cet estimateur est sans biais lorsque $\overline{Y}_N = \overline{Y}_F$ ou lorsque le taux de réponse conditionnel en personne est de 100 % au moyen des mêmes arguments que ceux de l'équation (3.5). Cet estimateur est efficace, puisqu'il

permet d'inclure les observations de l'échantillonnage tant par grappes que sans mise en grappes. L'estimation de $\hat{\gamma}_W$ est la proportion observée de répondants sur le Web en fonction des deux échantillons.

Si le modèle de non-réponse posé est $\overline{Y}_N = \overline{Y}_F$ ou qu'il représente un taux de réponse complet en personne, $\hat{t}_2$ est alors un estimateur sans biais lorsque le plan de sondage est un sous-échantillonnage d'unités à deux phases ou un sous-échantillonnage d'UPE à deux phases et $\hat{t}_{df,2}$ est un estimateur sans biais lorsque le plan de sondage est un échantillonnage hybride. Des estimateurs plus efficaces ($\hat{t}_1$ et $\hat{t}_{df,1}$) sont disponibles, mais nécessitent l'hypothèse plus rigoureuse $\overline{Y}_N = \overline{Y}_F = \overline{Y}_W$.

Les seuls paramètres encore à préciser pour les estimateurs hybrides sont les facteurs de composition. L'approche habituelle est de retenir un facteur composite λ $(0 \le \lambda \le 1)$ égal à la taille d'échantillon efficace relative, et il s'agit souvent d'une approximation raisonnable. Pour $\hat{t}_{df,1}$, λ pourrait correspondre au ratio du nombre efficace de répondants dans S_A divisé par la somme de ce nombre et du nombre efficace de répondants dans S_B, où l'effet de plan pour S_B est estimé comme dans l'exemple qui précède. Une approche semblable pourrait être utilisée pour κ, mais comme seuls les répondants sur le Web sont regroupés, les tailles d'échantillon efficaces ne sont que celles des répondants sur le Web. Nous examinons différents facteurs de composition dans la simulation et constatons que les choix raisonnables ont peu d'effet sur les variances des estimations.

Estimation de la variance

Les méthodes d'estimation de la variance pour la plupart des estimateurs sont bien connues ou elles ne nécessitent que de légers ajustements pour traiter les approches de suivi de la non-réponse proposées dans le présent document (à savoir l'échantillonnage à deux degrés, les estimateurs sans mise en grappes et l'échantillonnage hybride). Par exemple, la théorie de l'estimation à double base est couverte par l'échantillonnage hybride (Lohr, 2011). L'estimation de la variance pour l'échantillonnage à deux phases de plans de sondage complexes a fait l'objet de plusieurs avancées théoriques récentes, notamment dans Hidiroglou (2001), Hidiroglou, Rao et Haziza (2009) et Beaumont, Béliveau et Haziza (2015).

L'estimation de la variance pour le sous-échantillonnage d'UPE à deux phases nécessite certaines modifications de la théorie de Beaumont, Béliveau et Haziza (2015) pour permettre une bonne estimation des variances. Le seul plan de sondage que nous avons examiné dans le cas présent consiste à échantillonner un grand nombre d'UPE en première phase, puis à prendre tous les non-répondants en seconde phase dans un sous-échantillon des UPE. Le sous-échantillonnage d'UPE en seconde phase ne dépend pas des résultats de l'échantillon en première phase (nous supposons également que chaque UPE de la première phase échantillonnée compte des non-répondants). Puisque ce plan de sondage permet de sélectionner tous les non-répondants dans les UPE sous-échantillonnées, il satisfait aux conditions d'invariance et d'indépendance (voir Särndal, Swensson et Wretman, 1992, page 134-135) pour l'échantillonnage à deux degrés. Dans la mesure où la fraction de sondage au premier degré est négligeable, la norme d'estimation de la variance avec remise peut s'appliquer.

Nous avons utilisé la linéarisation en série de Taylor pour notre simulation, mais Beaumont, Béliveau et Haziza (2015) ont décrit de quelle façon les méthodes de réplication s'appliquent tout aussi bien en l'occurrence. Le plan de sous-échantillonnage d'UPE à deux phases sélectionne normalement un échantillon stratifié d'UPE. (Notre simulation a porté sur un échantillon non stratifié, mais la généralisation aux échantillons stratifiés est simple.) Dans chaque strate, le nombre d'UPE sélectionnées est déterminé de manière à ce que les unités traitées comme strates de la variance (ou grappes) soient équilibrées par rapport au sous-échantillonnage de suivi. Cette combinaison des UPE donne une estimation sans biais de la variance (Lu, Brick et Sitter, 2006). Par exemple, en utilisant un sous-échantillon de 50 % d'UPE pour la seconde phase, l'approche consiste à échantillonner quatre UPE dans chaque strate et à les disposer par paires en deux unités de variance pour que chaque unité de variance ait une UPE sous-échantillonnée pour le suivi et l'autre non. Au moyen d'un sous-échantillon de 33 % d'UPE, six UPE sont échantillonnées par strate dans la première phase et deux unités de variance sont formées; chacune contient deux UPE non sous-échantillonnées pour le suivi et une sous-échantillonnée pour le suivi.

4. Étude par simulation

Pour évaluer les trois plans de sondage et les estimateurs, nous avons réalisé une étude par simulation. Dans la présente section, nous commençons par décrire l'approche que nous avons retenue pour générer les populations à simuler. Ensuite, nous exposons notre plan de simulation et fournissons une description des mesures que nous avons adoptées pour évaluer et comparer les plans et les estimateurs. Enfin, nous présentons les résultats de la simulation, en commençant par ceux des scénarios du plan hybride et en enchaînant avec une comparaison des plans de sous-échantillonnage avec le plan d'échantillonnage hybride.

Générer la population

Pour cette simulation, nous avons utilisé les données de l'échantillon de microdonnées à grande diffusion de 2015 à 2019 de l'ACS pour les 50 États et le District de Columbia, en traitant tous les répondants de l'ACS au sein des ménages comme étant notre population. La collecte de données de l'ACS commence par une invitation à répondre à l'enquête sur le Web, envoyée par la poste. Les non-répondants font l'objet d'un suivi d'abord par l'envoi d'un questionnaire papier (par la poste), puis par une interview téléphonique (ITAO) ou une interview sur place assistée par ordinateur (IPAO) pour un sous-échantillon de non-répondants (en personne). (Bien que l'ITAO ait été abandonnée comme mode de suivi pour l'ACS après septembre 2017, cela n'a pas d'importance dans notre simulation, car tous les répondants aux ITAO et aux IPAO sont traités comme des répondants en personne aux fins de la simulation.) Dans les fichiers de l'échantillon de microdonnées à grande diffusion de l'ACS, chaque ménage est associé à une grappe géographique appelée « zone de microdonnées à grande diffusion ». Nous avons utilisé les zones de microdonnées à grande diffusion comme unités primaires d'échantillonnage (UPE) pour les échantillons par grappes. Lors de la sélection des échantillons par grappes, nous avons échantillonné les UPE au moyen de probabilités proportionnelles au nombre de ménages dans les UPE et avons échantillonné les ménages au

moyen de probabilités égales dans les UPE, les probabilités conditionnelles de sélection étant proportionnelles à la réciproque des probabilités de sélection des UPE. Pour chaque échantillon sans mise en grappes, nous avons prélevé les ménages au moyen d'un échantillon aléatoire simple sans remise.

À la section précédente, nous avons discuté du modèle de non-réponse dans lequel l'estimateur est sans biais pour chaque estimateur. Pour examiner le rendement de chaque estimateur selon divers modèles de réponse, nous avons créé quatre pseudo-populations (A, B, C et D) en variant l'approche retenue pour déterminer les répondants et les non-répondants. Pour les pseudo-populations A, B et C, nous avons considéré que tous les répondants sur le Web à l'ACS étaient nos répondants axés sur le Web. Pour la pseudo-population A, nous avons désigné tous les répondants par la poste à l'ACS comme étant nos répondants en personne, et tous les répondants en personne à l'ACS comme étant nos non-répondants. Pour la pseudo-population B, parmi les répondants par la poste et en personne à l'ACS, nous avons désigné de façon aléatoire la moitié des répondants comme étant nos répondants en personne et l'autre moitié, nos non-répondants. La pseudo-population C a un taux de réponse de 100 % en raison du suivi en personne, tous les répondants par la poste et en personne à l'ACS étant nos répondants en personne. Notre motivation pour constituer la pseudo-population D était de créer une population ayant un taux de réponse plus faible au mode initial de prise de contact; par conséquent, nous avons défini tous les répondants par la poste à l'ACS comme étant nos répondants sur le Web et, comme nous l'avons fait en créant la pseudo-population B, nous avons désigné de façon aléatoire la moitié des répondants restants comme étant nos répondants en personne et l'autre moitié, nos non-répondants.

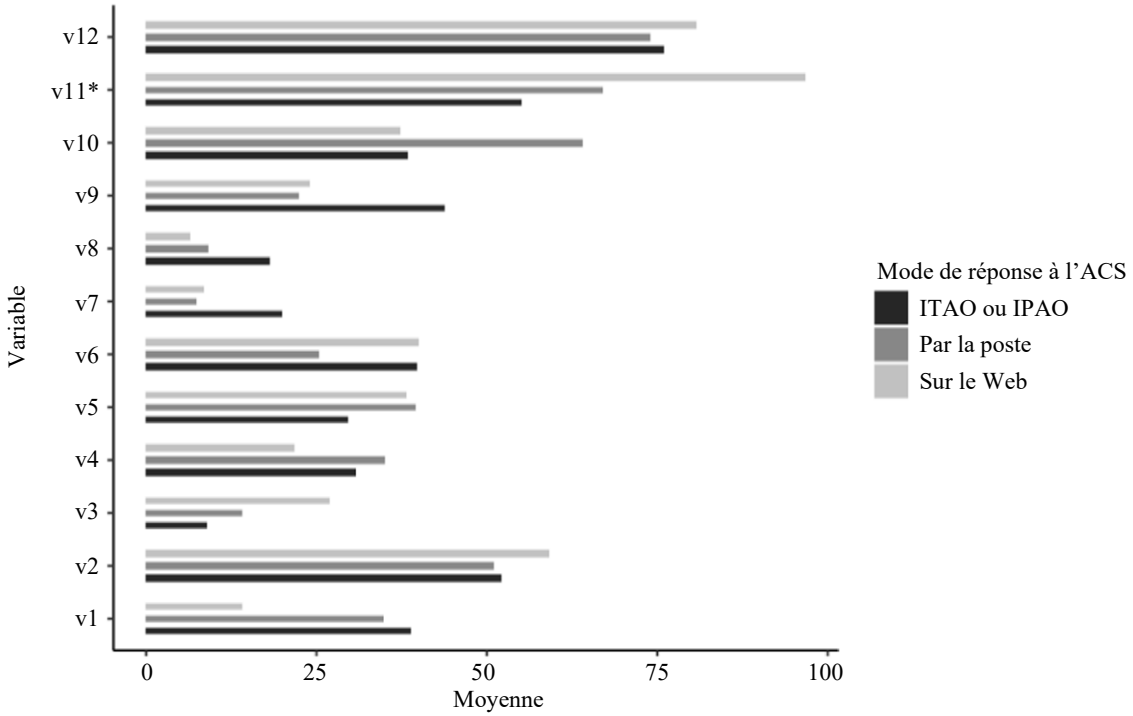
Nous avons inclus plusieurs variables démographiques et socioéconomiques au niveau du ménage dans la présente simulation. La figure 4.1 donne une brève description de chacune des variables et présente les moyennes de population de chacune des variables selon le mode de réponse à l'ACS, où l'ensemble de données complet de l'échantillon de microdonnées à grande diffusion de l'ACS est la population. Pour la pseudo-population A, nous nous attendons à ce que tous les estimateurs soient biaisés dans une certaine mesure, puisque les non-répondants diffèrent des deux groupes de répondants. Le suivi en personne, lorsqu'il est correctement pris en compte dans l'estimateur, devrait réduire le biais dans les estimations pour les variables v1, v3, v8 et v11 d'après la figure 4.1. Pour les pseudo-populations B et D, le modèle de réponse $\overline{Y}_N = \overline{Y}_F$ est valable en soi, de sorte que les estimateurs fondés sur ce modèle de réponse devraient être sans biais. Pour la pseudo-population C, il n'y a pas de non-réponse après le suivi en personne, donc les estimateurs qui pondèrent bien l'échantillon de suivi devraient être sans biais.

Plan de sondage en simulation

Les simulations consistaient à prélever des échantillons des diverses pseudo-populations (comme l'illustre le tableau 4.1) et, pour chaque variable, à calculer les totaux estimés à l'aide de chacun des estimateurs présentés à la section 3. Le tableau 4.2 énumère les scénarios de simulation. Les scénarios sont étiquetés en fonction de la combinaison de paramètres; par exemple, le scénario B2U est le scénario qui repose sur la pseudo-population B et qui permet de sélectionner un échantillon par grappes seulement et qui procède par sous-échantillonnage d'unités à deux phases pour le suivi de la non-réponse en personne.

Pour les scénarios A1A, B1A, C1A et D1A, dont chacun permet d'examiner le plan hybride au moyen de différentes pseudo-populations, nous avons sélectionné des échantillons par grappes et sans mise en grappes d'une taille de 2 500 dans chaque cas et de 50 UPE pour l'échantillon par grappes. Pour permettre une comparaison directe entre l'échantillonnage hybride (plus précisément, le scénario B1A) et les sous-échantillonnages d'unités et d'UPE, nous avons inclus les scénarios B2P et B2U.

Figure 4.1 Moyennes de population pour les variables tirées de l'échantillon de microdonnées à grande diffusion de 2015 à 2019 de l'American Community Survey, selon le mode de réponse. Au total, 49,3 % des réponses proviennent du Web, 29,1 %, de la poste et 21,7 %, du suivi en personne.



Notes : ACS = American Community Survey; ITAO = Interviews téléphoniques assistées par ordinateur; IPAO = Interviews sur place assistées par ordinateur.
Définitions des variables : v1 – diplôme d'études secondaires; v2 – études collégiales partielles; v3 – baccalauréat ou grade supérieur; v4 – ménage d'une seule personne; v5 – ménage de deux personnes; v6 – ménage de trois personnes ou plus; v7 – personne de référence hispanique; v8 – personne de référence noire et non hispanique; v9 – locataire; v10 – personne de 60 ans ou plus; v11* – revenu du ménage/110 000; v12 – pourcentage en région urbaine.

Tableau 4.1
Définition des pseudo-populations.

Pseudo-population	Catégorie de réponse des pseudo-populations selon le mode de réponse à l'ACS		
	Répondants sur le Web de l'ACS	Répondants par la poste de l'ACS	Répondants par ITAO ou IPAO de l'ACS
A	Répondants sur le Web	Répondants en personne	Non-répondants
B	Répondants sur le Web	50 % de répondants en personne, 50 % de non-répondants	50 % de répondants en personne, 50 % de non-répondants
C	Répondants sur le Web	Répondants en personne	Répondants en personne
D	50 % de répondants en personne, 50 % de non-répondants	Répondants sur le Web	50 % de répondants en personne, 50 % de non-répondants

Notes : ACS = American Community Survey; ITAO = Interviews téléphoniques assistées par ordinateur; IPAO = Interviews sur place assistées par ordinateur.

De plus amples renseignements sur la taille des échantillons et les fractions de sous-échantillonnage utilisées dans chacun de ces plans de sondage sont présentés en annexe (tableau A.1). Ces plans de sondage ont été élaborés de manière à obtenir le même nombre total prévu de réponses obtenues (3 127,7) et le même nombre selon le mode.

Nous avons réalisé 5 000 itérations indépendantes de chaque scénario. Pour les scénarios du plan hybride, nous avons calculé chacun des estimateurs du tableau 3.1 ($\hat{t}_1$, $\hat{t}_2$, $\hat{t}_A$, $\hat{t}_{df,1}$ et $\hat{t}_{df,2}$). Pour les deux scénarios de sous-échantillonnage, nous avons calculé $\hat{t}_1$ et $\hat{t}_2$. Pour le scénario de sous-échantillonnage d'UPE à deux phases (B2P), nous avons également calculé la variante de l'estimateur $\hat{t}_2$ qui utilise l'autre ajustement de sous-échantillonnage ω_s^{-1} obtenu dans l'équation (3.6).

Tableau 4.2
Scénarios de simulation.

Scénario	Pseudo-population	Plan de sondage	Suivi de la non-réponse
A1A	A	1 (échantillons indépendants par grappes et sans mise en grappes)	A (tous les répondants)
B1A	B	1 (échantillons indépendants par grappes et sans mise en grappes)	A (tous les répondants)
C1A	C	1 (échantillons indépendants par grappes et sans mise en grappes)	A (tous les répondants)
D1A	D	1 (échantillons indépendants par grappes et sans mise en grappes)	A (tous les répondants)
B2P	B	2 (échantillon par grappes seulement)	P (sous-échantillonnage d'unités primaires d'échantillonnage à deux phases)
B2U	B	2 (échantillon par grappes seulement)	U (sous-échantillonnage d'unités à deux phases)

Pour chaque itération, nous avons calculé le biais relatif (BR), le coefficient de variation (c.v.), la racine de l'erreur quadratique moyenne relative (REQMR) et un indicateur de la couverture du total de population par l'intervalle de confiance (IC) [pour un IC normal à 95 %] pour chaque estimateur, et nous avons fait la moyenne de chacune de ces mesures pour toutes les itérations. Le biais relatif (BR) et la REQMR ont été calculés par rapport au paramètre de la population (c'est-à-dire en divisant le biais et la REQMR, respectivement, par le total de population). Nous avons utilisé la linéarisation en série de Taylor pour calculer les estimations de la variance, comme nous l'avons vu précédemment. Le BR, le c.v., la REQMR et la couverture de l'IC sont les mesures que nous avons employées pour évaluer les estimateurs.

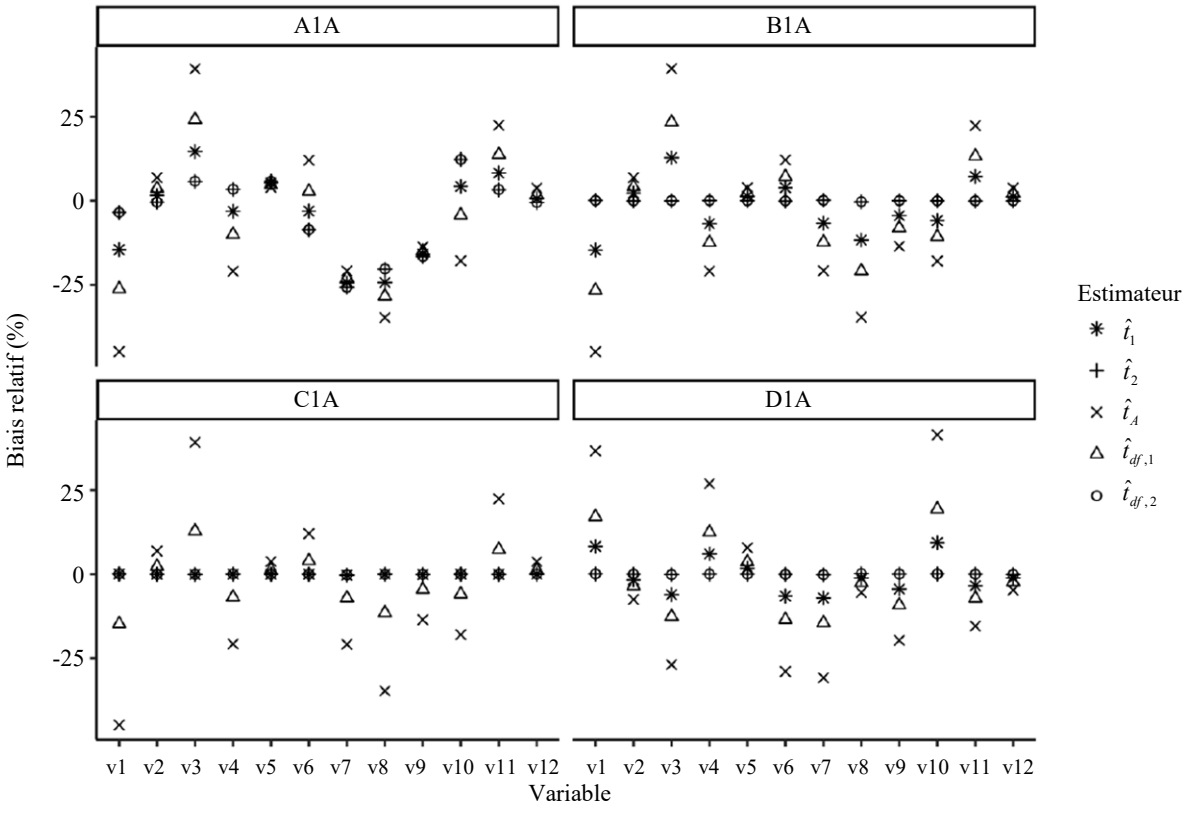
Résultats : scénarios du plan hybride

Les figures 4.2 et 4.3 présentent les résultats du BR et de la REQMR; les résultats du c.v. et de l'IC figurent dans arXiv:2303.13303v1 [stat.ME] pour chacun des scénarios de simulation du plan hybride. Dans le scénario A1A, à savoir le scénario ayant le biais inhérent le plus important en raison de la façon dont la pseudo-population est définie, les résultats de la simulation démontrent que tous les estimateurs sont biaisés pour au moins certaines caractéristiques. Pour les caractéristiques comme celles de v1, de v3, de v8 et de v11, où le suivi en personne devrait réduire le biais, nous constatons que les estimateurs qui intègrent bien le suivi en personne ($\hat{t}_2$ et $\hat{t}_{df,2}$) sont généralement moins biaisés; pour les variables comme celles de v5,

de v7 et de v9, où nous nous attendrions à ce que le suivi en personne accroisse le biais (voir la figure 4.1), aucune augmentation du biais n'est apparente pour les estimateurs qui intègrent le suivi en personne. Dans ce scénario, la couverture de l'IC est généralement très faible (bien inférieure au niveau nominal à 95 %) en raison du biais dont sont entachées les estimations. L'estimateur composite du plan hybride $\hat{t}_{df,2}$ est généralement comparable à l'estimateur $\hat{t}_2$ pour la REQMR.

Dans le cas des pseudo-populations B et D, les répondants sur le Web diffèrent des répondants en personne et des non-répondants, mais l'hypothèse $\overline{Y}_N = \overline{Y}_F$ se vérifie dans les espérances. Une différence clé entre ces deux pseudo-populations est que le taux de réponse prévu sur le Web est considérablement plus élevé dans la pseudo-population B que dans la pseudo-population D (49 % par rapport à 29 %). Comme il était à prévoir pour ce modèle de non-réponse, les estimateurs $\hat{t}_2$ et $\hat{t}_{df,2}$ ont le plus petit biais relatif et les meilleurs taux de couverture de l'intervalle de confiance. Pour la plupart des caractéristiques examinées, l'estimateur $\hat{t}_{df,2}$ a la plus petite REQMR. Les estimateurs qui ajustent les poids des répondants sur le Web en fonction de la non-réponse, $\hat{t}_4$, $\hat{t}_{df,1}$ et $\hat{t}_1$, présentent un biais plus important et une couverture médiocre de l'intervalle de confiance.

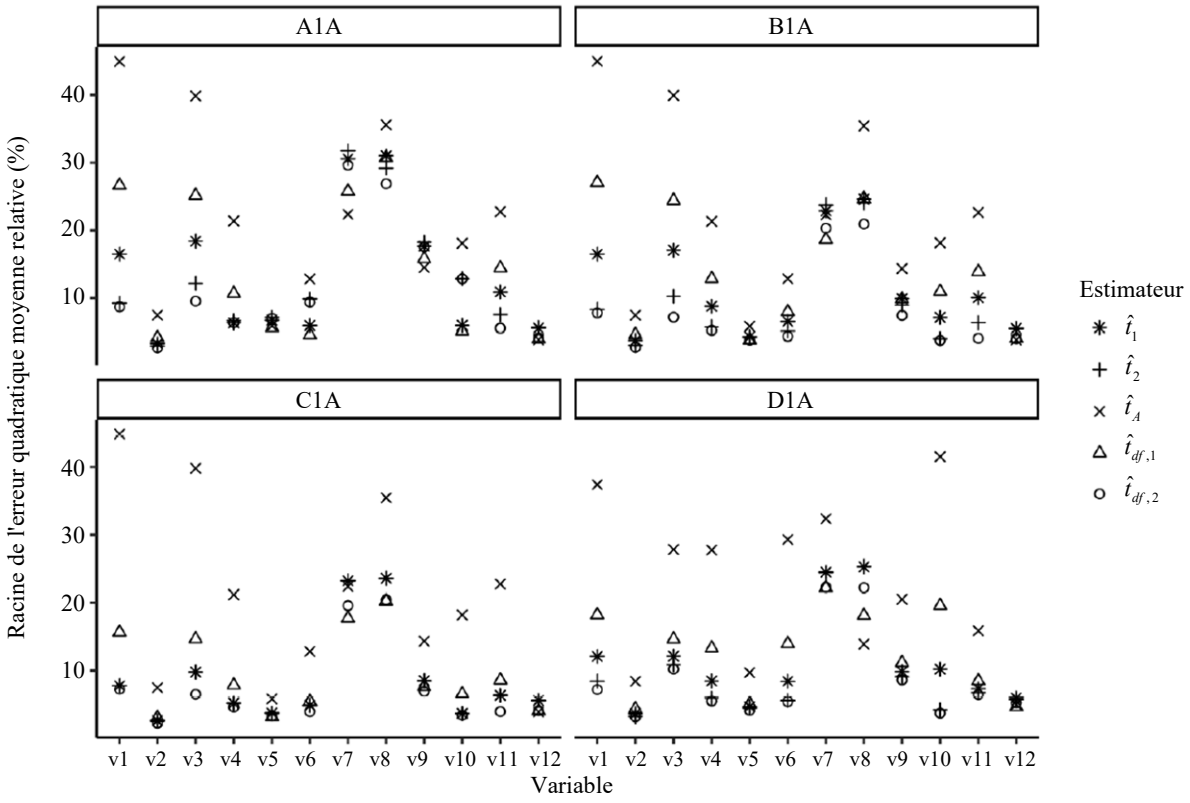
Figure 4.2 Biais relatifs des divers estimateurs pour chacun des scénarios du plan hybride.



En ce qui concerne la pseudo-population C, on obtient un taux de réponse complet après le suivi en personne. Cependant, les caractéristiques des répondants sur le Web diffèrent généralement de celles des répondants en personne. Dans le scénario C1A, le biais relatif est négligeable pour $\hat{t}_2$ et $\hat{t}_{df,2}$, c'est-à-dire les estimateurs qui supposent que $\overline{Y}_N = \overline{Y}_F$. Dans ce scénario, étant donné que $R = 1$, l'estimateur $\hat{t}_1$ se réduit à l'estimateur de Horvitz-Thompson (pondération de base) et présente des propriétés semblables à celles de $\hat{t}_2$ et de $\hat{t}_{df,2}$. De plus, la couverture de l'intervalle de confiance est la plus proche des niveaux nominaux pour ces trois estimateurs, tandis que pour les autres estimateurs, elle demeure généralement très faible.

Comme il est indiqué à la section 1, le choix des facteurs de composition pour les estimateurs du plan hybride est un autre aspect à prendre en considération. Dans cette optique, nous avons calculé $\hat{t}_{df,2}$ pour chaque itération des scénarios du plan hybride, et ce, de deux manières, soit par le facteur de composition quasi optimal décrit à la section 2, soit par le facteur de composition fixé à $\kappa = 0{,}2$, ce qui est loin du facteur optimal. Les résultats présentés dans arXiv:2303.13303v1 [stat.ME] montrent que le choix du facteur de composition n'a pratiquement aucun effet sur le biais; pour tous les scénarios sauf le scénario A1A, il en va de même pour l'effet du facteur de composition sur la couverture de l'intervalle de confiance. Le choix du facteur de composition a effectivement un léger effet sur le c.v. et la REQMR; le facteur optimal donne lieu à des estimations moins variables.

Figure 4.3 Racine de l'erreur quadratique moyenne relative de chaque estimateur pour chacun des scénarios du plan hybride.



Résultats : comparaison entre les plans de sous-échantillonnage et le scénario d'échantillonnage hybride

Ensuite, nous examinons les scénarios de sous-échantillonnage pour le suivi de la non-réponse, le scénario B2U (sous-échantillonnage d'unités) et le scénario B2P (sous-échantillonnage d'UPE). Comme ces scénarios prévoient uniquement un échantillon par grappes, les seuls estimateurs qui leur sont pertinents sont $\hat{t}_1$ et $\hat{t}_2$. Le plan hybride est une solution de rechange aux plans de sous-échantillonnage; à la section précédente, nous avons démontré que, comme on peut s'y attendre selon la théorie de l'estimation à double base de sondage, l'estimateur composite $\hat{t}_{df,2}$ donne les meilleurs résultats pour ce plan. Pour les trois scénarios, l'estimateur $\hat{t}_2$ a un biais négligeable et l'estimateur $\hat{t}_1$, un biais évident (voir le tableau A.2 en annexe). Par conséquent, notre évaluation comporte une comparaison des résultats qui concernent $\hat{t}_2$ pour les scénarios B2U (sous-échantillonnage d'unités) et B2P (sous-échantillonnage d'UPE) et $\hat{t}_{df,2}$ pour le scénario B1A (plan hybride sans sous-échantillonnage). Pour ces comparaisons, les figures 4.4 et 4.5 présentent le BR et la REQMR. Les résultats pour le c.v. et la couverture de l'intervalle de confiance figurent dans arXiv:2303.13303v1 [stat.ME].

Figure 4.4 Biais relatifs des divers estimateurs pour le scénario du plan hybride par rapport aux scénarios de sous-échantillonnage.

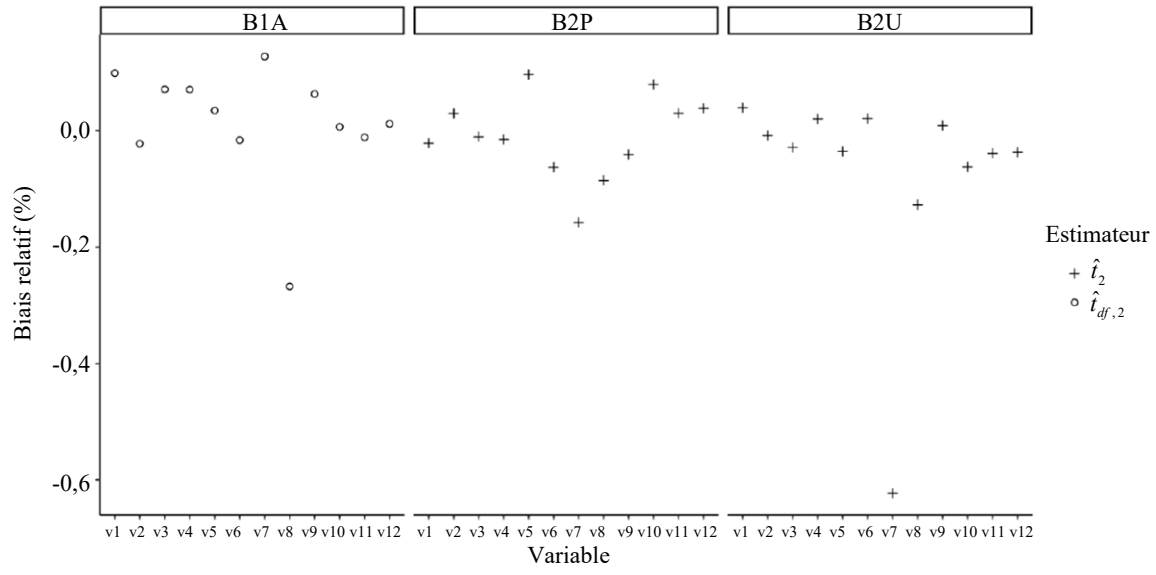
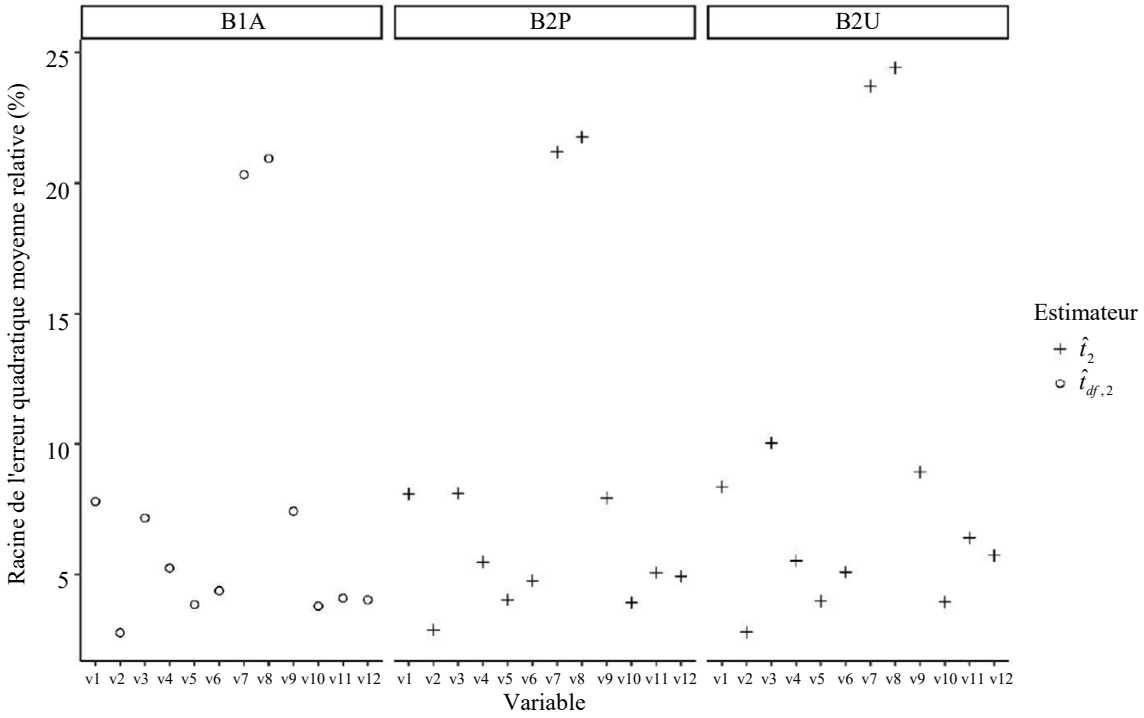


Figure 4.5 Racine de l'erreur quadratique moyenne relative des divers estimateurs pour le scénario du plan hybride par rapport aux scénarios de sous-échantillonnage.



Il n'y a pas de différences appréciables de biais pour $\hat{t}_{df,2}$ et $\hat{t}_2$ d'un plan de sondage à l'autre. L'estimateur $\hat{t}_2$ dans les plans de sous-échantillonnage est généralement plus variable que l'estimateur composite $\hat{t}_{df,2}$ dans le plan hybride pour ce qui est du c.v. et de la REQMR. Les couvertures de l'intervalle de confiance sont comparables pour les trois scénarios et leurs estimateurs respectifs.

Pour le scénario de sous-échantillonnage d'UPE, nous avons évalué la variation de l'estimateur $\hat{t}_2$ qui repose sur l'autre ajustement de sous-échantillonnage ω_s^{-1} donné à l'équation (3.4) en comparant celui-ci à l'estimateur qui repose sur la réciproque du taux de sous-échantillonnage d'UPE comme ajustement ω^{-1}. Le résultat de cette comparaison présenté dans arXiv:2303.13303v1 [stat.ME] est que les estimateurs sont comparables pour le biais, mais que l'autre estimateur se comporte légèrement mieux pour le c.v., la REQMR et la couverture de l'intervalle de confiance.

Le tableau A.2 de l'annexe présente les mesures sommaires réparties en moyenne pour les variables de l'ensemble des estimateurs se rattachant à chaque scénario.

5. Discussion

À mesure que les taux de réponse ont diminué et que les coûts des enquêtes ont augmenté, les spécialistes de la recherche sur les enquêtes ont cherché des approches pour lutter contre ces tendances. Dans le présent

document, nous avons présenté des approches concernant les plans de sondage et des approches d'estimation pour la collecte multimodale de données et nous avons utilisé une étude par simulation pour examiner leur efficacité. Même si nous nous sommes concentrés sur le Web pour la première phase de la collecte des données, nos constatations et nos recommandations touchent aussi des plans de sondage où l'envoi par la poste (employée seule ou en combinaison avec le Web, comme dans une approche axée sur le Web) est utilisé en première phase.

Les protocoles de collecte des données qui intègrent des interviews sur place ont longtemps été considérés comme la règle d'or, car ils obtiennent généralement des taux de réponse plus élevés que les autres modes et réduisent le biais par l'obtention de réponses de sous-groupes qui ne sont généralement pas couverts avec d'autres modes (par exemple ménages sans téléphone non couverts par des enquêtes téléphoniques ou ménages sans accès Internet non couverts par des enquêtes en ligne). Cependant, les interviews sur place sont coûteuses par rapport aux autres modes, ce qui fait que les coûts sont prohibitifs pour de nombreuses études où il a fallu plutôt privilégier les enquêtes en ligne ou téléphonique ou les questionnaires sur papier ou des combinaisons de ces modes moins coûteux.

Nous avons présenté deux plans de sondage et des estimateurs connexes qui sont des solutions de rechange au sous-échantillonnage d'unités à deux phases et qui pourraient ouvrir la porte à la collecte de données en personne pour certaines études; pour d'autres études, les approches que nous avons décrites pourraient faciliter un choix plus discriminant de l'interview sur place comme façon de contenir les coûts tout en maintenant des normes de qualité élevées. L'approche du sous-échantillonnage d'UPE fait seulement appel à un échantillon par grappes; on sélectionne initialement un plus grand nombre d'UPE et on opte pour le suivi en personne avec un sous-échantillon d'UPE. L'estimateur $\hat{t}_2$ a bien fonctionné pour les deux plans de sous-échantillonnage, tandis que $\hat{t}_1$ accusait un biais important et une mauvaise couverture de l'intervalle de confiance. De toute évidence, essayer de réduire la variance en utilisant $\hat{t}_1$ est susceptible d'entraîner des biais dans de nombreuses situations.

L'approche du plan hybride combine deux échantillons indépendants, l'un sans mise en grappes et l'autre par grappes, le mode initial de collecte de données le moins coûteux étant appliqué aux deux échantillons, mais le suivi plus coûteux en personne auprès des non-répondants étant utilisé uniquement dans l'échantillon par grappes. Par conséquent, l'approche de l'échantillonnage hybride offre comme avantage l'obtention du même nombre de réponses à un coût moindre pour un même taux de réponse global pondéré et une réduction du biais comme plan de sondage avec un suivi complet de la non-réponse. Avec le plan hybride, l'estimateur hybride $\hat{t}_{df,2}$ est sans biais dans le modèle de non-réponse $\overline{Y}_N = \overline{Y}_F$ ou là où le suivi en personne mène à un taux de réponse de 100 %.

Grâce à la combinaison des paramètres employés dans l'étude par simulation, l'estimateur composite $\hat{t}_{df,2}$ dans le scénario du plan hybride avait une REQMR qui était, en moyenne, de 8 % inférieure à celle de l'estimateur $\hat{t}_2$ dans le scénario du sous-échantillonnage d'UPE et de 14 % inférieure dans le scénario du sous-échantillonnage d'unités. Les trois scénarios ont été conçus pour occasionner à peu près le même coût

pour le même nombre attendu de réponses obtenues par mode et avec un suivi en personne dans le même nombre d'UPE.

Les résultats de simulation font voir que le $\hat{t}_{df,2}$ du plan hybride est supérieur à ceux des deux plans à deux phases dans les conditions que nous avons examinées; le sous-échantillonnage d'UPE serait supérieur à cet égard au sous-échantillonnage d'unités. De plus, le plan hybride présente des avantages pratiques, surtout en cas d'incertitude concernant les taux de réponse par mode. Grâce au plan hybride, il est relativement simple d'accroître la taille de l'échantillon sans mise en grappes pour traiter un taux de réponse en ligne moindre que prévu. Avec les plans de sondage à deux phases, il y a plus de complications, comme les répercussions sur le coût et le calendrier du suivi en personne en raison de l'augmentation de l'échantillon de première phase.

La simulation a toujours démontré les lacunes bien connues de l'estimateur $\hat{t}_1$ qui est couramment utilisé dans la pratique, où il n'y a pas de distinction chez les répondants selon le mode pendant le calcul des rajustements de pondération. Le modèle de non-réponse qui suppose l'égalité des moyennes entre les répondants sur le Web, les répondants en personne et les non-répondants est implicite dans cet estimateur. Cela est plus restrictif que le modèle d'égalité des moyennes entre les répondants en personne et les non-répondants à la base de l'estimateur $\hat{t}_2$ et de l'estimateur du plan hybride $\hat{t}_{df,2}$. Brick et coll. (2021) traitent de cette question et tiennent compte en partie des déséquilibres de composition des répondants en recourant à un ajustement adaptatif pour les modes. Dans une étude s'appuyant sur le Web avec un suivi en personne, l'approche de Brick et coll. (2021) peut être mise en œuvre par une repondération des répondants sur le Web et en personne en vue de conserver certaines des qualités de réduction du biais de l'ajustement des répondants en personne tout en réduisant les variances en ajustant les poids des répondants sur le Web également. D'autres recherches sont nécessaires pour explorer cette approche avec les enquêtes en personne.

Nous avons aussi analysé les propriétés théoriques des estimateurs dans le contexte des plans de sondage présentés dans la présente étude. Cependant, comme pour toute simulation, nous n'avons pas examiné tous les scénarios possibles. Dans la conception de notre simulation, nous nous sommes concentrés sur les aspects que nous croyions les plus susceptibles d'influer sur le rendement relatif des estimateurs.

Les estimateurs que nous avons présentés comprennent un ajustement unique ayant pour effet de traiter l'échantillon comme une seule catégorie de pondération. Dans la pratique, nous ne nous attendons pas à ce que l'hypothèse $\overline{Y}_N = \overline{Y}_F$ se vérifie en général, mais cette hypothèse pourrait être mieux estimée à l'intérieur de catégories, c'est-à-dire $\overline{Y}_{N_C} = \overline{Y}_{F_C}$, où l'indice C indique la catégorie. Si des variables auxiliaires peuvent être trouvées et sont disponibles à la fois pour les répondants et les non-répondants de sorte que ce modèle de non-réponse se vérifie (au moins approximativement), alors ces catégories serviraient normalement à calculer les ajustements de pondération afin de réduire le biais. D'autres recherches seront nécessaires à un élargissement des estimateurs présentés en fonction de cette situation. Il pourrait ainsi y avoir une repondération par catégorie pour la non-réponse, et ce, tant pour l'échantillonnage hybride que pour le sous-échantillonnage d'UPE. De même, les estimateurs calés doivent être soumis à une évaluation.

Nous concluons par quelques réflexions sur un autre plan de sondage qui, à notre avis, justifie un examen plus approfondi, celui d'un échantillon sans grappes combiné à un échantillon d'UPE existantes, où on commence par le Web et on enchaîne avec un suivi en personne auprès des non-répondants en ligne dans les UPE échantillonnées. Une application possible de ce plan de sondage est le renouvellement de l'échantillon dans une étude longitudinale où l'échantillon initial était un échantillon par grappes d'UPE. Une autre application est lorsque du personnel formé n'est disponible que pour un échantillon existant d'UPE. Dans ces contextes, on pourrait envisager de sélectionner un échantillon sans grappes en essayant d'enquêter sur le Web d'abord et en faisant un suivi en personne auprès des non-répondants en ligne dans l'échantillon sans grappes qui se trouvent à l'intérieur des UPE visées. Nous croyons que convient en pareil cas l'estimateur $\hat{t}_2$ présenté en (3.5) avec une modification du second terme de manière à remplacer le ω^{-1} par la réciproque de la probabilité de sélection des UPE. D'autres travaux devraient permettre d'examiner plus à fond ce plan et les propriétés de l'estimateur dans ce contexte.

Remerciements

Nous sommes reconnaissants envers Sharon Lohr pour ses réflexions et commentaires très utiles sur une ébauche de ce manuscrit. Nous apprécions aussi grandement les commentaires et les suggestions du rédacteur en chef, du rédacteur associé et des trois examinateurs.

Documentation supplémentaire

Dans le document supplémentaire (disponible à l'adresse https://doi.org/10.48550/arXiv.2303.13303), nous présentons d'autres résultats de simulation. Plus précisément, ce document présente les c.v. et les taux de couverture de l'intervalle de confiance pour toutes les comparaisons, les résultats de l'examen des autres facteurs de composition et ceux de l'ajustement de l'autre sous-échantillonnage à l'équation (3.4).

Annexe

Tableau A.1
Paramètres des plans de suivi de la non-réponse en simulation.

	Total pour l'échantillon sans grappes	Total pour l'échantillon par grappes	
		Par UPE	Dans l'ensemble
Aucun sous-échantillonnage pour les unités de suivi de la non-réponse (scénario B1A)			
Taille de l'échantillon sans grappes	2 500		
Nombre d'UPE			50
Nombre d'unités échantillonnées par UPE			50
Nombre prévu de réponses obtenues			
Web		24,0	1 201,0
En personne (dans les 50 UPE)		14,5	725,7
Total	1 201,0	38,5	1 926,7

Note : UPE = Unités primaires d'échantillonnage.

Tableau A.1(suite)
Paramètres des plans de suivi de la non-réponse en simulation.

	Total pour l'échantillon sans grappes	Total pour l'échantillon par grappes	
		Par UPE	Dans l'ensemble
Sous-échantillonnage d'unités à deux phases (scénario B2U)			
Taille de l'échantillon sans grappes	0		
Nombre d'UPE			50
Nombre d'unités échantillonnées par UPE			100
Fraction de sous-échantillonnage de suivi des cas de non-réponse			0,5
Nombre prévu de réponses obtenues			
Web		48,0	2 402,1
En personne (dans les 50 UPE)		14,5	725,7
Total	0	62,6	3 127,7
Sous-échantillonnage d'UPE à deux phases (scénario B2P)			
Taille de l'échantillon sans grappes	0		
Nombre d'UPE			
Nombre total d'UPE dans l'échantillon			100
UPE sous-échantillonnées pour les unités de suivi des cas de non-réponse			50
Nombre d'unités échantillonnées par UPE			50
Nombre prévu de réponses obtenues par UPE			
Web		24,0	2 402,1
En personne (seulement dans les 50 UPE sous-échantillonnées pour les unités de suivi des cas de non-réponse)		14,5	725,7
Total dans les UPE sous-échantillonnées		38,5	
Nombre total prévu de réponses obtenues	0		3 127,7

Note : UPE = Unités primaires d'échantillonnage.

Tableau A.2
Moyenne des mesures sommaires des estimateurs par scénario.

Statistique	Scénario	2 phases	Sans grappes	2 degrés	df-1	df-2 (en opt.)	df-3 (non en opt.)	2 phases (UPE)
BR	A1A	-3,76 %	-5,36 %	-4,20 %	-4,66 %	-3,73 %	-3,76 %	
	B1A	-0,02 %	-5,34 %	-1,78 %	-3,19 %	0,01 %	-0,01 %	
	C1A	-0,03 %	-5,35 %	-0,03 %	-1,79 %	-0,01 %	-0,03 %	
	D1A	0,00 %	-2,19 %	-0,49 %	-1,02 %	0,01 %	0,00 %	
	B2P	-0,01 %		-1,78 %				-0,02 %
	B2U	-0,07 %		-1,82 %				
c.v.	A1A	7,31 %	4,28 %	7,19 %	4,74 %	5,95 %	6,69 %	
	B1A	8,16 %	4,27 %	7,70 %	4,96 %	6,82 %	7,55 %	
	C1A	7,79 %	4,27 %	7,79 %	5,44 %	6,39 %	7,15 %	
	D1A	8,51 %	5,61 %	8,23 %	5,98 %	7,76 %	8,17 %	
	B2P	7,37 %		6,39 %				7,18 %
	B2U	8,15 %		7,54 %				

Notes : Dans le cas de l'intervalle de confiance, les valeurs pour chaque variable sont divisées par la moyenne sur tous les scénarios.
BR = biais relatifs; c.v. = coefficient de variation; UPE = unités primaires d'échantillonnage.

Tableau A.2(suite)
Moyenne des mesures sommaires des estimateurs par scénario.

Statistique	Scénario	2 phases	Sans grappes	2 degrés	df-1	df-2 (en opt.)	df-3 (non en opt.)	2 phases (UPE)
REQMR	A1A	12,78 %	20,81 %	13,25 %	14,40 %	11,69 %	12,27 %	
	B1A	9,12 %	20,77 %	11,43 %	13,58 %	7,66 %	8,45 %	
	C1A	8,71 %	20,75 %	8,71 %	9,53 %	7,16 %	8,00 %	
	D1A	9,48 %	22,47 %	11,03 %	12,80 %	8,66 %	9,11 %	
	B2P	8,18 %		10,27 %				7,97 %
	B2U	9,08 %		11,33 %				
IC	A1A	72,58 %	12,86 %	75,06 %	39,85 %	65,24 %	68,93 %	
	B1A	94,96 %	12,89 %	81,40 %	39,46 %	94,66 %	94,60 %	
	C1A	94,05 %	12,95 %	94,05 %	63,57 %	93,42 %	93,43 %	
	D1A	94,70 %	18,48 %	83,41 %	49,66 %	94,36 %	94,30 %	
	B2P	94,72 %		75,19 %				95,31 %
	B2U	94,91 %		79,97 %				
BSA (BR)	A1A	8,85 %	20,12 %	9,99 %	13,16 %	8,86 %	8,85 %	
	B1A	0,09 %	20,08 %	6,57 %	11,95 %	0,07 %	0,08 %	
	C1A	0,05 %	20,06 %	0,05 %	6,63 %	0,05 %	0,05 %	
	D1A	0,06 %	21,03 %	4,75 %	9,81 %	0,08 %	0,06 %	
	B2P	0,06 %		6,64 %				0,07 %
	B2U	0,09 %		6,64 %				
IC normal	A1A	1,16	0,71	1,16	0,76	0,93	1,04	
	B1A	1,20	0,71	1,18	0,76	0,99	1,08	
	C1A	1,08	0,71	1,08	0,75	0,85	0,96	
	D1A	1,23	0,93	1,18	0,85	1,12	1,16	
	B2P	1,07		0,95				1,07
	B2U	1,19		1,13				

Notes : Dans le cas de l'intervalle de confiance, les valeurs pour chaque variable sont divisées par la moyenne sur tous les scénarios.
UPE = unités primaires d'échantillonnage; REQMR = racines de l'erreur quadratique moyenne relative; IC = intervalles de confiance; BSA (BR) = biais standardisés absolus (biais relatifs).

Bibliographie

Bayart, C., et Bonnel, P. (2015). How to combine survey media (web, telephone, face-to-face): Lyon and Rhône-Alps case study. *Transportation Research Procedia*, 11, 118-135.

Beaumont, J.-F., Béliveau, A. et Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology*, 3(4), 524-542.

Brick, J.M., Kennedy, C., Flores-Cervantes, I. et Mercer, A.W. (2021). An adaptive mode adjustment for multimode household surveys. *Journal of Survey Statistics and Methodology*, 10(4), 1024-1047.

Brick, J.M., et Montaquila, J.M. (2009). Nonresponse and weighting, Elsevier. *Handbook of Statistics*, 29, 163-185.

Bureau du recensement des États-Unis (2014). American Community Survey design and methodology (Janvier 2014). U.S. Census Bureau. (Accessible à l'adresse https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html).

Bureau du recensement des États-Unis (2019). American Community Survey: Accuracy of the data. U.S. Census Bureau. (Accessible à l'adresse https://www2.census.gov/programs-surveys/acs/tech_docs/accuracy/ACS_ Accuracy_of_Data_2019.pdf).

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Dillman, D.A. (2017). Inciter les participants aux enquêtes à mode mixte à répondre sur le Web : les promesses et les défis. *Techniques d'enquête*, 43, 1, 3-34. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2017001/article/14836-fra.pdf.

Goodman, A., Brown, M., Silverwood, R.J., Sakshaug, J.W., Calderwood, L., Williams, J. et Ploubidis, G.B. (2022). The impact of using the Web in a mixed-mode follow-up of a longitudinal birth cohort study: Evidence from the National Child Development Study. *Journal of the Royal Statistical Society, Series A,* 185(3), 822-850.

Groves, R.M., et Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169(3), 439-457.

Hansen, M.H., et Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41(236), 517-529.

Heeringa, S.G., Wagner, J., Torres, M., Duan, N., Adams, T. et Berglund, P. (2004). Sample designs and sampling methods for the Collaborative Psychiatric Epidemiology Studies (CPES). *International Journal of Methods in Psychiatric Research*, 13(4), 221-240.

Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 2, 157-169. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001002/article/6091-fra.pdf.

Hidiroglou, M.A., Rao, J.N.K. et Haziza, D. (2009). Variance estimation in two-phase sampling. *Australian & New Zealand Journal of Statistics*, 51(2), 127-141.

Holt, D.H. (1980). Discussion de l'article de V. Verma, C. Scott et C. O'Muircheartaigh, Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society, Series A*, 143(4), 468-469.

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, 8(2), 183.

Lohr, S.L. (2011). Autres plans de sondage : échantillonnage avec bases de sondage multiples chevauchantes. *Techniques d'enquête*, 37, 2, 213-232. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011002/article/11608-fra.pdf.

Lu, W.W., Brick, J.M. et Sitter, R.R. (2006). Algorithms for constructing combined strata variance estimators. *Journal of the American Statistical Association*, 101(476), 1680-1692.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201), 101-116.

Särndal, C.-E., et Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 3, 279-294.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics.

Wagner, J., Arrieta, J., Guyer, H. et Ofstedal, M.B. (2014). Does sequence matter in multimode surveys: Results from an experiment. *Field Methods*, 26(2), 141-155.