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

Techniques d'enquête

Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples

par Sharon L. Lohr

Date de diffusion : le 6 janvier 2022

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

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

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, décembre 2021
Vol. 47, n° 2, pp. 247-285
Statistique Canada, n° 12-001-X au catalogue

Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples

Sharon L. Lohr[1]

Résumé

Les enquêtes à bases de sondage multiples, dans lesquelles des échantillons probabilistes indépendants sont sélectionnés dans chacune des Q bases de sondage, servent depuis longtemps à améliorer la couverture, réduire les coûts ou augmenter la taille des échantillons de sous-populations d'intérêt. Une grande partie de la théorie a été développée en supposant que (1) l'union des bases de sondage couvre la population d'intérêt, (2) un échantillon probabiliste avec réponse complète est sélectionné dans chaque base de sondage, (3) les variables d'intérêt sont mesurées dans chaque échantillon sans erreur de mesure, et (4) il existe suffisamment d'information pour tenir compte du chevauchement des bases de sondage lors du calcul des estimations. Après avoir passé en revue la conception, l'estimation et le calage des enquêtes effectuées à partir de bases de sondage multiples traditionnelles, je considère des modifications aux hypothèses qui permettent à une structure de bases de sondage multiples de servir de principe d'organisation pour d'autres méthodes de combinaison de données telles que l'imputation massive, l'appariement d'échantillons, l'estimation sur petits domaines et l'estimation par techniques de saisie-ressaisie. Enfin, je discute la façon dont les résultats de la recherche sur les enquêtes à l'aide de bases de sondage multiples peuvent être utilisés lors de la conception et de l'évaluation des systèmes de collecte de données qui intègrent plusieurs sources de données.

Mots-clés : Combinaison de données; intégration de données; enquête à double base de sondage; échantillonnage indirect; imputation massive; erreur de classification; plan de sondage; sous-couverture.

1. Introduction

Tout au long de sa carrière de 33 ans au Bureau du recensement américain et de sa carrière subséquente de 32 ans à Westat, Joe Waksberg a souvent eu recours à plusieurs sources de données en vue d'améliorer la qualité des estimations tout en réduisant les coûts. Il a utilisé des sources de données externes pour évaluer la couverture du recensement décennal des États-Unis (Marks et Waksberg, 1966; Waksberg et Pritzker, 1969), pour caler des poids d'enquête, et pour améliorer l'efficacité ou suréchantillonner les populations rares dans les plans de sondage (Hendricks, Igra et Waksberg, 1980; Cohen, DiGaetano et Waksberg, 1988; DiGaetano, Judkins et Waksberg, 1995; Waksberg, 1995; Waksberg, Judkins et Massey, 1997b).

À plusieurs reprises, Waksberg a intégré directement les données de deux enquêtes ou plus afin d'améliorer la couverture ou d'obtenir des tailles d'échantillon plus grandes pour les sous-populations (Waksberg, 1986; Burke, Mohadjer, Green, Waksberg, Kirsch et Kolstad, 1994; Waksberg, Brick, Shapiro, Flores-Cervantes et Bell, 1997a). Dans ces enquêtes à bases de sondage multiples, des échantillons indépendants ont été sélectionnés à partir de bases de sondage qui, ensemble, devaient couvrir la totalité ou la quasi-totalité de la population cible. Les données tirées des échantillons ont été combinées aux fins d'obtention d'estimations pour la population dans son ensemble et pour les sous-populations d'intérêt. Waksberg s'est intéressé au plan de ces enquêtes à bases de sondage multiples dans la

1. Sharon Lohr est professeure émérite à l'Arizona State University. Courriel : sharon.lohr@asu.edu.

perspective de maîtriser à la fois les erreurs dues à l'échantillonnage et celles non attribuables à l'échantillonnage. Il a ainsi découvert que l'utilisation de plusieurs bases de sondage répondait à la difficulté qu'il y avait à produire des estimations fiables dans un contexte de coûts de collecte de données accrus (avec une non-réponse plus élevée dans le cas des méthodes de collecte moins coûteuses) et de couverture de base de sondage incomplète.

De nos jours, les organismes statistiques et les organismes d'enquête sont confrontés aux mêmes types de défis que ceux auxquels Waksberga fait face, à savoir la diminution des taux de réponse et l'augmentation des coûts de la collecte des données d'enquête, mais à un degré plus élevé. Simultanément, l'apparition de nouvelles sources de données offre des occasions d'obtenir des renseignements sur des parties de populations d'intérêt, parfois avec une rapidité étonnante. De nombreux organismes utilisent maintenant ou étudient des méthodes d'intégration de données provenant de sources multiples afin d'améliorer l'exactitude ou l'actualité des estimations de population.

Je suis extrêmement honorée d'avoir été invitée à prononcer le discours Waksberg et, dans le présent article, je veux m'appuyer sur les réflexions de Waksberg au sujet des enquêtes à bases multiples pour discuter leur utilisation comme principe d'organisation de la combinaison d'informations provenant de sources multiples. Traditionnellement, les enquêtes à bases multiples intègrent des données de Q échantillons probabilistes $S_1, \ldots, S_Q$ qui sont sélectionnés indépendamment à partir de Q bases de sondage. Toutefois, on peut élargir la structure générale pour inclure des bases de sondage constituées d'enregistrements administratifs ou d'échantillons non probabilistes. La structure peut également être étendue à des situations dans lesquelles certaines sources de données ne mesurent pas les variables d'intérêt y, mais mesurent des covariables $\mathbf{x}$ qui peuvent servir à prédire y.

Plusieurs auteurs ont étudié des méthodes de combinaison de données provenant de sources multiples; voir par exemple Citro (2014), Lohr et Raghunathan (2017), National Academies of Sciences, Engineering, and Medicine (2017, 2018), Thompson (2019), Zhang et Chambers (2019), Beaumont (2020), Yang et Kim (2020) et Rao (2021). Les sources comprennent des échantillons probabilistes classiques, des ensembles de données administratives, des données des capteurs, des données des réseaux sociaux et des échantillons de commodité généraux.

Bien que les types de données (et la vitesse de collecte de certains types de données) aient changé au cours des dernières années, la structure élémentaire du problème de la combinaison des sources de données n'a pas changé depuis les premières enquêtes à double base de sondage. La section 2 traite de la structure et des hypothèses des enquêtes classiques à bases multiples au moyen de l'exemple de la *National Survey of America's Families* (NSAF, Enquête nationale sur les familles des États-Unis), une enquête à double base sur laquelle Waksberg a travaillée dans les années 1990. La section 3 examine les méthodes de calcul des estimations des caractéristiques de population à partir des enquêtes classiques à bases multiples où toutes les hypothèses sont respectées, y compris le cas particulier où un échantillon est

le recensement d'un sous-ensemble de la population. La section 4 traite ensuite de la façon dont la structure à bases multiples intègre plusieurs des méthodes actuelles de combinaison des données, parfois avec des hypothèses assouplies. La section 5 porte sur les problèmes posés par la conception de systèmes de collecte de données qui contrôlent pour les erreurs d'échantillonnage et de non-échantillonnage, et discute les orientations qui pourraient être données à de futurs travaux de recherche.

2. Structure et hypothèses classiques des enquêtes à bases multiples

Examinons d'abord un exemple de ce que j'appellerais une enquête « classique » à bases multiples – à savoir une enquête conçue pour prélever des échantillons probabilistes à partir de chacune des bases en nombre fixe – puis définissons la notation et les hypothèses qui serviront à décrire les estimateurs et leurs propriétés.

2.1 NSAF (National Survey of America's Families)

L'objectif de la *National Survey of America's Families* (NSAF) de 1997 était de fournir des informations sur les caractéristiques sociales et économiques de la population civile des États-Unis ne vivant pas en établissement institutionnel âgée de moins de 65 ans, en mettant l'accent sur l'obtention d'estimations fiables sur les personnes et les familles, en particulier les familles avec enfants, situées sous 200 % du seuil de pauvreté. L'enquête cherchait à obtenir des estimations pour l'ensemble du pays, ainsi que des estimations séparées pour 13 États, choisis à dessein pour leur variété en matière de région géographique, de parti politique dominant, de taille et de capacités budgétaires.

Pour répondre aux exigences de précision des estimations, il était souhaitable d'avoir une taille d'échantillon efficace d'environ 800 enfants pauvres dans chaque État. On aurait pu atteindre cet objectif en prélevant un échantillon de ménage dans une base aréolaire. Waksberg et coll. (1997b) ont déterminé que la présélection de ménages selon leur revenu et le sous-échantillonnage des ménages non pauvres constitueraient la méthode la plus rentable d'obtenir les tailles d'échantillon souhaitées dans un échantillon à base aréolaire, mais le coût serait élevé parce qu'on s'attend à ce que seulement une famille sur huit environ ait des enfants et se situe sous 200 % du seuil de pauvreté.

Les coûts de présélection seraient considérablement réduits si l'enquête pouvait être menée par téléphone au moyen de la composition aléatoire (CA). Cependant, étant donné que selon les données de la *Current Population Survey* (Enquête sur la population actuelle), environ 20 % des familles vivant dans la pauvreté n'ont pas de téléphone, la base de sondage obtenue par composition aléatoire devrait avoir un sous-dénombrement important de la population cible. De plus, les ménages situés sous 200 % du seuil de pauvreté sans téléphone peuvent avoir des niveaux de revenu ou des caractéristiques de santé différents de ceux des ménages également situés sous 200 % du seuil de pauvreté mais ayant un téléphone.

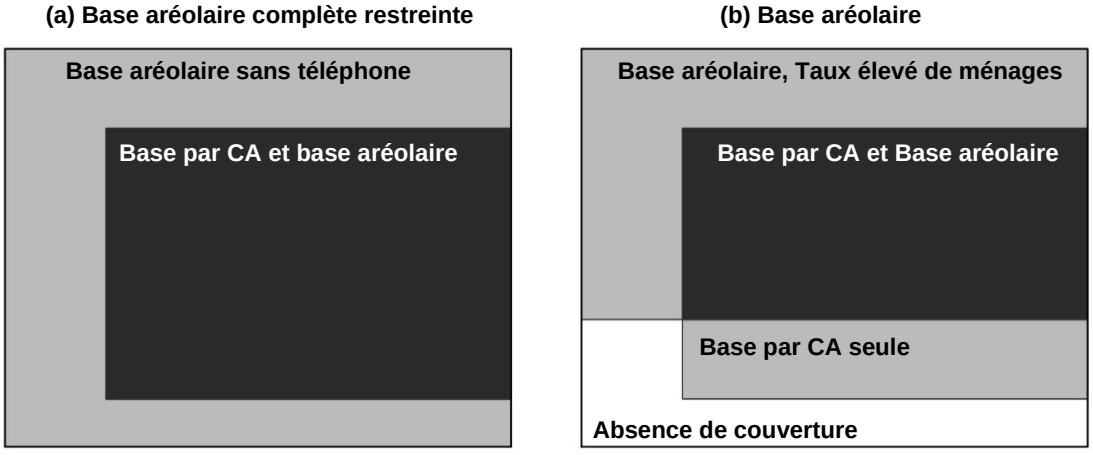
Ainsi, un échantillon de la base aréolaire donnerait une couverture élevée, mais entraînerait aussi des coûts élevés inacceptables. Une enquête par composition aléatoire coûterait moins, mais entraînerait un

sous-dénombrement important de la population d'intérêt. Waksberg et coll. (1997a) ont utilisé une enquête à double base de sondage, avec un échantillon tiré de la base aréolaire et un deuxième échantillon choisi indépendamment dans la base par CA, pour profiter des coûts plus faibles d'un échantillon par CA tout en couvrant également les ménages sans téléphone. La figure 2.1(a) montre la structure des deux bases de sondage.

Pour réduire encore les coûts, Waksberg et coll. (1997a) ont exclu des groupes d'îlots de recensement de la base aréolaire comprenant peu de ménages sans téléphone; selon le recensement de 1990, les régions exclues représentaient moins de 10 % des ménages sans téléphone dans chaque État. Avec cette exclusion, la base aréolaire et la base par CA contenaient chacune les ménages qui ne se trouvaient pas dans l'autre base, comme le montre la figure 2.1(b).

Les ménages avec téléphone appartenant aux groupes d'îlots non exclus étaient présents dans les deux bases de sondage. Si un échantillon probabiliste était prélevé dans chaque base de sondage, les ménages qui se chevauchent (la zone ombrée foncée de la figure 2.1(b)) pourraient être sélectionnés dans les deux échantillons. Les concepteurs de l'enquête pouvaient soit décider de mener des interviews auprès de tous les ménages de chaque échantillon, puis traiter la multiplicité de l'estimation (plan de chevauchement), soit éliminer dans une des bases les ménages se trouvant également dans l'autre base (plan de présélection).

Figure 2.1 Couverture de la base de la NSAF. La zone ombrée foncée se trouve dans les deux bases de sondage.



Waksberg et ses collaborateurs ont choisi une présélection. On a demandé aux ménages de l'échantillon de la base aréolaire s'ils avaient le téléphone, et seuls ceux qui n'avaient pas de téléphone ont été soumis à une interview détaillée. La réalisation des interviews détaillées était longue et coûteuse; la présélection des ménages ayant le téléphone au cours d'une courte interview a permis d'économiser des ressources qui ont pu servir à augmenter le nombre de ménages sans téléphone dans l'échantillon. Les

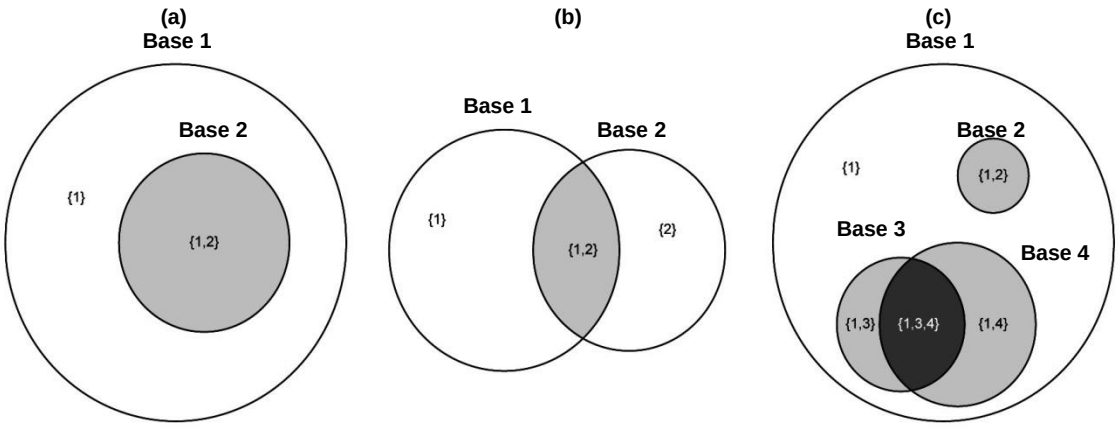
ménages ayant le téléphone ont été échantillonnés uniquement au moyen de la base de sondage par CA; les ménages de l'échantillon par CA sans enfants et situés au-dessus de 200 % du seuil de pauvreté ont été sous-échantillonnés. Comme une enquête de présélection a été utilisée, l'échantillon combiné des deux enquêtes était un échantillon stratifié, et des ressources ont été affectées aux deux échantillons à l'aide de formules d'échantillonnage stratifié, qui ont représenté le coût plus élevé de l'échantillonnage à partir de la base aréolaire.

2.2 Notation et hypothèses pour les enquêtes à bases multiples

Dans les enquêtes classiques à bases multiples comme la NSAF, il faut plusieurs hypothèses pour pouvoir obtenir des estimations sans biais des caractéristiques de la population, ainsi que des intervalles de confiance ayant des probabilités de couverture approximativement exactes.

Supposons qu'il y ait Q bases. Un domaine de population d est défini par les intersections des bases : le domaine $\{1, 3, 4\}$, par exemple, contient les unités de population qui sont dans les bases 1, 3 et 4, mais non dans les autres bases. Soit D l'ensemble des domaines possibles; selon le chevauchement des unités, D peut contenir entre 1 et $2^Q - 1$ domaines. La figure 2.2 présente trois exemples de relations entre bases de sondage. Quand la base de sondage 1 est complète, mais que la base de sondage 2 est incomplète, comme dans la figure 2.2(a), $D = \{\{1\}, \{1, 2\}\}$; toute unité de population se trouvant dans la base de sondage 2 est également dans la base de sondage 1. Pour une enquête à double base qui se chevauche comme celle de la figure 2.2(b), $D = \{\{1\}, \{2\}, \{1, 2\}\}$.

Figure 2.2 Structures à trois bases. (a) La base de sondage 1 a une couverture complète et la base de sondage 2 est incomplète. (b) Les bases de sondage 1 et 2 sont incomplètes mais se chevauchent. (c) La base 1 est complète; les bases 2, 3 et 4 sont toutes incomplètes, mais les bases 3 et 4 se chevauchent.



Définissons $\delta_i(d) = 1$ si l'unité i est dans le domaine d et 0 dans le cas contraire, et soit $\delta_i^{(q)} = 1$ si l'unité i est dans la base q et 0 dans le cas contraire. La base de sondage q a une taille de population $N^{(q)}$ et le domaine d a une taille de population N_d; ces tailles peuvent être connues ou inconnues. La population cible compte au total N unités.

On formule généralement les hypothèses suivantes afin de tirer des inférences des enquêtes classiques à bases multiples.

(A1) L'union des Q bases couvre la population cible.

(A2) L'échantillon S_q tiré de la base de sondage q est un échantillon probabiliste où l'unité i a une probabilité $\pi_i^{(q)}$ d'être dans S_q. Supposons que $w_i^{(q)}$ représente le poids final de l'unité i dans S_q; les options pour $w_i^{(q)}$ incluent le poids de sondage $1/\pi_i^{(q)}$, le poids de Hájek $N^{(q)}/[\hat{N}^{(q)}\pi_i^{(q)}]$ avec $\hat{N}^{(q)} = \sum_{j \in S_q} 1/\pi_j^{(q)}$, ou un poids ajusté pour compenser la non-réponse.

(A3) Les échantillons $S_1, \ldots, S_Q$ sont sélectionnés indépendamment.

(A4) L'appartenance au domaine de chaque unité i dans S_q, $\{\delta_i(d), d \in D\}$, est connue.

(A5) L'estimateur du total de la population dans le domaine d de S_q, $\hat{Y}_d^{(q)} = \sum_{i \in S_q} \delta_i(d) w_i^{(q)} y_i$, est approximativement sans biais pour $Y_d = \sum_{i=1}^{N} \delta_i(d) y_i$, pour toutes les bases q contenant un domaine d et pour toutes les variables y.

(A6) Il n'y a pas d'erreur de mesure. Si l'unité i est dans la base q et la base q', y_i aura la même valeur si elle est mesurée dans S_q que si elle est mesurée dans $S_{q'}$.

Il s'agit d'hypothèses fortes; un certain assouplissement de chaque hypothèse est possible pour des estimateurs donnés, comme nous le verrons à la section 3. Elles sont toutefois plus faibles que les hypothèses nécessaires pour certaines des autres méthodes possibles d'intégration de données. Le couplage d'enregistrements, par exemple, suppose implicitement que l'unité i dans la base de sondage q peut être appariée à une unité en particulier dans la base de sondage q'. Pour les enquêtes à bases multiples, il faut savoir si une unité échantillonnée dans la base de sondage q se trouve également dans d'autres bases, mais il n'est pas nécessaire d'identifier l'unité appariée.

2.3 Les hypothèses ont-elles été respectées dans la NSAF ?

Il est rare que les hypothèses d'une enquête soient respectées exactement dans la pratique, et la NSAF ne fait pas exception. L'hypothèse (A1) n'a pas été respectée en raison de l'exclusion des groupes d'îlots ayant un taux élevé de ménages avec téléphone. L'échantillon de la base aréolaire a produit moins de ménages sans téléphone que ce qui était attendu, peut-être en raison de l'erreur de mesure dans le recensement de 1990 ou des changements démographiques depuis 1990. De plus, les enquêtes menées après l'enquête à partir des données de la *Current Population Survey* (Enquête sur la population actuelle) de 1997 ont indiqué que les groupes d'îlots exclus de la base de sondage pouvaient avoir eu plus de

ménages sans téléphone que ce qui était attendu (Waksberg, Brick, Shapiro, Flores-Cervantes, Bell et Ferraro, 1998).

Bien que des échantillons probabilistes indépendants aient été prélevés de chaque base de sondage, on constate la présence de non-réponse pour chaque échantillon. Les taux de réponse estimés pour les enfants étaient de 65 % dans l'échantillon par CA et de 84 % dans l'échantillon de la base aréolaire. La procédure de pondération tentait de traiter le biais potentiel du sous-dénombrement et de la non-réponse. Les poids des ménages sans téléphone de l'échantillon aréolaire ont été ajustés par le quotient dans une tentative de compenser le sous-dénombrement des exclusions de groupe d'îlots. Les poids ajustés pour la non-réponse ont été calculés séparément pour les échantillons de la base aréolaire et de la base par CA, puis les échantillons combinés ont été poststratifiés sur les totaux de contrôle du *Census Bureau* (Brick, Shapiro, Flores-Cervantes, Ferraro et Strickler, 1999). Groves et Wissoker (1999) ont trouvé peu de preuves de biais résiduel dans leur analyse du biais dû à la non-réponse. L'une des différences, peu nombreuses, qu'ils rapportent est que les ménages de l'échantillon par CA nécessitant plus d'appels de contact et les ménages d'un sous-échantillon composé de non-répondants avaient une probabilité légèrement plus élevée de recevoir de l'aide alimentaire.

Dans la NSAF, on déterminait l'appartenance au domaine en demandant aux répondants du ménage de l'échantillon aréolaire s'ils avaient un téléphone qui fonctionnait. Si la réponse à cette question était exacte, l'hypothèse (A4) était respectée. Les enquêteurs ont tenté de réduire l'erreur de mesure pour l'hypothèse (A6) en demandant à des intervieweurs téléphoniques centralisés d'effectuer toutes les interviews détaillées; à cette fin, les ménages de la base aréolaire ont été interviewés au moyen d'un téléphone cellulaire apporté par le représentant sur le terrain. Comme les interviews dans le domaine $\{1, 2\}$ ont été obtenues uniquement à partir de l'échantillon par CA, on ne dispose pas de données permettant d'évaluer les erreurs de mesure possibles ou le biais de non-réponse relatif pour les deux échantillons.

Waksberg a utilisé des enquêtes à double base plusieurs fois avant la NSAF, principalement pour augmenter la taille des échantillons lors de l'échantillonnage de populations rares, mais il recommande leur utilisation seulement si un plan plus simple ne permet pas d'atteindre les objectifs de l'enquête. Il écrit ainsi : « Le prix est une complexité supplémentaire dans les opérations d'échantillonnage et la possibilité d'erreur si l'appariement des deux bases de sondage n'est pas réalisé avec soin... Mon instinct me dit qu'un plan plus complexe ne devrait être utilisé que s'il apporte un assez bon rendement » (Waksberg, 1986).

La complexité et les dépenses supplémentaires du plan à double base en valaient-elles la peine dans le cas de la NSAF ? Étant donné que les ménages avec téléphone ont été éliminés de l'échantillon aréolaire et que le rendement des ménages sans téléphone était inférieur aux attentes, seulement 1 488 des 44 461 ménages interviewés provenaient de l'échantillon aréolaire. Mais en raison du taux de pauvreté élevé des ménages sans téléphone, le pourcentage estimé d'enfants dans les ménages situés sous 200 % du

seuil de pauvreté était d'environ 3,6 points de pourcentage plus élevé avec l'échantillon complet qu'avec l'échantillon par CA seulement. Bien que, pour de nombreuses variables, la différence entre l'estimation de l'échantillon complet et l'estimation de l'échantillon par CA soit petite, cette différence n'aurait pas pu être évaluée sans l'échantillon aréolaire.

3. Estimation dans les enquêtes classiques à bases multiples

Le principal problème de l'inférence dans une enquête classique à bases multiples conçue de façon à répondre aux hypothèses (A1) à (A6) est la façon de tenir compte du chevauchement possible entre échantillons. Dans la NSAF, les ménages avec téléphone ont été exclus de l'échantillon aréolaire, mais dans de nombreuses applications, la présélection est infaisable ou il est plus rentable d'obtenir des données de l'échantillon complet sélectionné à partir de chaque base de sondage. Quand la combinaison de données n'a pas été prise en compte dans la conception d'enquêtes ou de sources de données séparées, le chevauchement dépend de la couverture de chaque source de données.

Avec un plan de sondage à chevauchement, les unités contenues dans plus d'une base de sondage ont plusieurs chances d'être sélectionnées dans l'échantillon. Un estimateur construit par addition des observations pondérées de chacun des Q échantillons,

$$\hat{Y}_{\text{concat}} = \sum_{q=1}^{Q} \sum_{i \in S_q} w_i^{(q)} y_i,$$

sera un estimateur biaisé de $Y = \sum_{i=1}^{N} y_i$ parce que les poids d'échantillon individuels ne reflètent pas les multiples chances de sélection des unités dans les domaines de chevauchement. Les méthodes d'estimation des totaux de population multiplient donc habituellement les poids d'enquête $w_i^{(q)}$ par un ajustement de la multiplicité $m_i^{(q)}$ satisfaisant $\sum_{q=1}^{Q} \delta_i^{(q)} m_i^{(q)} \approx 1$ pour chaque unité i, ce qui donne l'estimateur

$$\hat{Y} = \sum_{q=1}^{Q} \sum_{i \in S_q} w_i^{(q)} m_i^{(q)} y_i = \sum_{q=1}^{Q} \sum_{i \in S_q} \tilde{w}_i^{(q)} y_i, \tag{3.1}$$

où $\tilde{w}_i^{(q)} = w_i^{(q)} m_i^{(q)}$ est le poids ajusté pour la multiplicité.

3.1 Estimateur composite de Hartley

Hartley (1962) a été le premier auteur à présenter une théorie rigoureuse de l'estimation dans les enquêtes à double base de sondage, où les unités du domaine de chevauchement $\{1, 2\}$ pouvaient être échantillonnées à partir des deux bases de sondage. Son article de quatre pages est la source de plusieurs contributions importantes. Premièrement, Hartley a défini le problème en termes statistiques. Deuxièmement, il a proposé un estimateur optimal aux fins de combinaison des estimations provenant de deux enquêtes. Et troisièmement, il a étudié le problème de l'allocation des ressources aux différents

échantillons dans le plan, qu'il résout en tenant compte conjointement de l'allocation et de l'estimateur de façon à minimiser la variance du total estimé de la population assujetti à un coût fixe.

Hartley (1962) a estimé le total de la population $Y = \sum_{i=1}^{N} y_i$ au moyen de

$$\hat{Y}(\theta) = \hat{Y}_{\{1\}}^{(1)} + \hat{Y}_{\{2\}}^{(2)} + \theta \hat{Y}_{\{1,2\}}^{(1)} + (1-\theta)\,\hat{Y}_{\{1,2\}}^{(2)}. \tag{3.2}$$

Il a proposé de choisir θ pour minimiser $V\left[\hat{Y}(\theta)\right]$. Cela donne la valeur

$$\theta_H = \frac{V\left(\hat{Y}_{\{1,2\}}^{(2)}\right) + \mathrm{Cov}\left(\hat{Y}_{\{2\}}^{(2)}, \hat{Y}_{\{1,2\}}^{(2)}\right) - \mathrm{Cov}\left(\hat{Y}_{\{1\}}^{(1)}, \hat{Y}_{\{1,2\}}^{(1)}\right)}{V\left(\hat{Y}_{\{1,2\}}^{(1)}\right) + V\left(\hat{Y}_{\{1,2\}}^{(2)}\right)}. \tag{3.3}$$

L'estimateur dans (3.2) est de la forme en (3.1) avec des ajustements de poids pour tenir compte de la multiplicité

$$m_i^{(1)} = \delta_i\left(\{1\}\right) + \delta_i\left(\{1,2\}\right)\theta, \quad m_i^{(2)} = \delta_i\left(\{2\}\right) + \delta_i\left(\{1,2\}\right)(1-\theta).$$

Si l'on souhaite utiliser le facteur de composition optimal θ_H, les estimateurs peuvent être substitués aux covariances inconnues dans (3.3). Parce que θ_H dépend des covariances impliquant y, l'ajustement optimal de la multiplicité peut différer pour différentes variables, ce qui donne un ensemble de poids différent pour chacune. De plus, θ_H peut être inférieur à 0 ou supérieur à 1, ce qui peut entraîner des poids négatifs pour certaines observations. Ces caractéristiques se reportent à la généralisation de Q bases de sondage de l'estimateur optimal de Hartley étudié par Lohr et Rao (2006).

L'estimateur dans (3.2), dont la valeur fixe est de θ, est approximativement sans biais pour Y selon l'hypothèse (A5). Si les totaux de domaine estimés et les estimations des covariances dans (3.3) sont convergents, alors l'estimateur avec $\hat{\theta}_H$ est convergent pour Y. Saegusa (2019) a étudié l'estimateur de Hartley du point de vue de la théorie du processus empirique, pour établir une loi des grands nombres et un théorème de limite centrale quand S_1 et S_2 sont tous deux des échantillons aléatoires simples.

Hartley avait appliqué sa théorie dans le domaine de l'agriculture, et plusieurs des premières applications d'enquêtes à double base de sondage étaient destinées à des enquêtes sur l'agriculture ou les entreprises (Kott et Vogel, 1995), car il existait des bases listes comprenant les plus grandes activités agricoles ou opérations d'entreprises. Une enquête à double base comportant un échantillon disproportionnellement grand tiré de la base liste réduisait les coûts parce que (1) l'obtention de données d'une opération dans la base liste était souvent moins coûteuse que l'obtention de données d'une opération dans la base aréolaire et que (2) le suréchantillonnage de la base liste était analogue au suréchantillonnage de strates à variance élevée dans un échantillonnage stratifié, ce qui produisait une plus grande efficacité.

Plus tard, en raison du nombre croissant de téléphones cellulaires, le biais découlant de l'utilisation d'échantillons de téléphones fixes seuls est devenu préoccupant, ce qui a conduit à l'utilisation de sondages par téléphones à double base, avec un échantillon provenant d'une base de sondage par téléphone fixe et un deuxième provenant d'une base de sondage par téléphone cellulaire. Dans ce cas, les deux bases de sondage sont incomplètes, mais elles couvrent à elles deux la population des personnes

ayant un téléphone. Dans le cadre de ces enquêtes, il est important de s'interroger sur le traitement des personnes ayant les deux types de ligne téléphonique. La section suivante examine les choix de composition.

3.2 Ajustements de la pondération pour la multiplicité

L'estimateur optimal de Hartley, avec θ_H, utilise un ensemble différent de poids pour chaque variable de réponse, ce qui peut entraîner des incohérences internes entre estimateurs. Plusieurs auteurs ont proposé des estimateurs utilisant un seul ensemble de poids pour toutes les analyses. Ici, j'énumère brièvement certains des facteurs d'ajustement pour la multiplicité $m_i^{(q)}$ qui donnent un ensemble de poids pour l'estimateur général du total de la population dans (3.1). Les estimateurs obtenus sont approximativement sans biais pour le total de population Y selon les hypothèses (A1), (A4) et (A5). Ces estimateurs et d'autres sont étudiés en détail par Lohr (2011), Lu, Peng et Sahr (2013), Ferraz et Vogel (2015), Arcos, Rueda, Trujillo et Molina (2015) et Baffour, Haynes, Western, Pennay, Misson et Martinez (2016).

- L'estimateur de présélection $m_i^{(1)} = 1, m_i^{(2)} = 1 - \delta_i^{(1)}, \ldots, m_i^{(Q)} = \prod_{q=1}^{Q-1} \left(1 - \delta_i^{(q)}\right)$. Une unité échantillonnée à partir de la base q est écartée si elle se trouve dans l'une des bases $1, \ldots, q-1$. Cet estimateur est automatiquement utilisé en cas de plan de présélection comme celui de la NSAF. Avec un plan de chevauchement, son utilisation signifie que certaines observations de données sont écartées.
- Estimateur fondé sur la multiplicité, avec $m_i^{(q)} = 1/$ (nombre de bases contenant l'unité $i) = 1 \Big/ \sum_{q=1}^{Q} \delta_i^{(q)}$. Dans une enquête à double base, cela donne l'estimateur dans (3.2) avec $\theta = 1/2$. Mecatti (2007) a fait remarquer qu'avec l'estimateur fondé sur la multiplicité, l'hypothèse (A4) peut être remplacée par l'hypothèse un peu moins restrictive selon laquelle $\sum_{q=1}^{Q} \delta_i^{(q)}$ est connu pour chaque unité échantillonnée i.

  L'estimateur fondé sur la multiplicité peut également être considéré comme un cas particulier de la méthode généralisée du partage des poids (Deville et Lavallée, 2006) utilisant la matrice de liens normalisée, puisque le nombre de liens vers l'unité de population i est le nombre de bases de sondage contenant cette unité.
- L'estimateur à base unique (Bankier, 1986; Kalton et Anderson, 1986), qui considère les observations comme si elles avaient été échantillonnées à partir d'une base unique. Si des poids de probabilité inversés sont utilisés, avec $w_i^{(q)} = 1/\pi_i^{(q)}$, alors $m_i^{(q)} = \pi_i^{(q)} \Big/ \sum_{f=1}^{Q} \delta_i^{(f)} \pi_i^{(f)}$. Cet estimateur nécessite que la probabilité d'inclusion de l'unité i soit connue pour toutes les bases au nombre de Q, y compris les bases à partir desquelles l'unité n'a pas été échantillonnée. Les ajustements pour la multiplicité tiennent compte des probabilités d'inclusion pour les plans, mais non des variances relatives, qui sont touchées par la corrélation intra-grappe et la stratification dans chacun des échantillons.

- Estimateur de la taille d'échantillon efficace (TEE) (Chu, Brick et Kalton, 1999; O'Muircheartaigh et Pedlow, 2002), où l'estimateur de domaine de chaque base est pondéré par la taille d'échantillon efficace relative de cette base. Soit $n^{(q)}$ la taille de l'échantillon de la base q et soit $\text{deff}^{(q)}$ l'effet de plan pour une variable clé ou un effet de plan lissé pour plusieurs variables. La taille d'échantillon efficace pour S_q est $\tilde{n}^{(q)} = n^{(q)} / \text{deff}^{(q)}$ et l'ajustement fondé sur la multiplicité pour l'unité i est

$$m_i^{(q)} = \frac{\tilde{n}^{(q)}}{\sum_{f=1}^{Q} \delta_i^{(f)} \tilde{n}^{(f)}}.$$

  Cet estimateur tient compte des variances relatives des estimateurs de différents échantillons et est souvent plus efficace que les estimateurs de présélection, fondé sur la multiplicité et à base de sondage unique.

  L'estimateur par le pseudo-maximum de vraisemblance (PMV) de Skinner et Rao (1996) est de ce type quand les tailles de la base de sondage $N^{(q)}$ et les tailles de domaine N_d sont inconnues; Skinner et Rao (1996) ont recommandé d'utiliser l'effet de plan pour estimer $N_{\{1,2\}}$ afin d'établir la taille d'échantillon efficace pour un cas à double base de sondage. L'estimateur par le PMV équivaut asymptotiquement à un estimateur de la taille d'échantillon efficace poststratifiant sur les tailles de domaine N_d quand elles sont connues. Quand on connaît les tailles de base $N^{(q)}$ mais qu'on ne connaît pas $N_{\{1,2\}}$, l'estimateur par le PMV équivaut asymptotiquement au calage de l'estimateur de la taille d'échantillon efficace sur les tailles de domaine calculées à partir de la fonction de pseudo-ressemblance.

On peut calculer des estimations approximativement sans biais des variances pour tous les estimateurs pris en compte dans la présente section selon les hypothèses (A1) à (A6) et des conditions de régularité supplémentaires qui assurent la convergence des totaux estimés et des estimateurs de la variance à partir des Q échantillons. Skinner et Rao (1996) ont étudié les estimateurs de la variance par linéarisation; Chauvet (2016) a calculé les estimateurs de la variance par linéarisation pour l'Enquête sur le logement en France, qui tenait compte de la réduction de la variance due aux fractions de sondage élevées de certaines bases. Lohr et Rao (2000) ont élaboré une théorie permettant d'utiliser le jackknife avec des bases multiples, et Lohr (2007) et Aidara (2019) ont considéré les estimateurs de la variance bootstrap. Ces méthodes reposent sur l'hypothèse (A3) d'échantillons indépendants; Chauvet et de Marsac (2014) ont étudié le cas dans lequel les échantillons ont en commun des unités d'échantillonnage primaires, mais que des échantillons indépendants sont prélevés au deuxième degré du plan d'échantillonnage.

Le calcul des estimations de la variance par linéarisation nécessite un logiciel spécial qui met en œuvre les calculs de dérivée partielle pour les bases de sondage multiples. Cependant, on peut effectuer les calculs des méthodes d'estimation de la variance par répliques comme le jackknife et le bootstrap dans un logiciel d'enquête standard, en créant un seul ensemble de données contenant toutes les observations

concaténées et les poids $\tilde{w}_i^{(q)}$ des Q échantillons et en créant des poids de rééchantillonnage au moyen des méthodes standard pour des échantillons à plusieurs degrés stratifiés (Metcalf and Scott, 2009). L'ensemble de données concaténées comporte $\sum_{q=1}^{Q} H_q$ strates, où H_q est le nombre de strates pour S_q; les observations tirées de différents échantillons sont dans des strates différentes. Les méthodes de pondération de rééchantillonnage peuvent également inclure les effets du calage (voir la section 3.3) sur la variance.

Bien entendu, dans de nombreuses applications, il faut d'autres estimations de quantités que les totaux de population, et la théorie des bases multiples s'applique aux paramètres qui sont des fonctions lisses de totaux de domaine. Il se peut cependant qu'on recherche un facteur de composition différent quand l'intérêt primaire porte sur d'autres quantités que les totaux de population, il se peut aussi que des considérations particulières soient nécessaires aux fins d'autres types d'analyses. Les autres types d'analyses statistiques étudiés dans un contexte de bases multiples sont la régression linéaire (Lu, 2014b) et non paramétrique (Lu, Fu et Zhang, 2021), la régression logistique avec données ordinales (Rueda, Arcos, Molina et Ranalli, 2018), les fonctions de distribution empirique (Arcos, Martínez, Rueda et Martínez, 2017), l'estimation des flux bruts avec données manquantes (Lu et Lohr, 2010) et les tests du chi carré (Lu, 2014a).

Lu (2014b) constate que les paramètres de régression linéaire estimés au moyen des poids ajustés pour la multiplicité sont les coefficients de régression de la population finie $\mathbf{B}$ qui minimisent la somme des carrés $\sum_{i=1}^{N} (y_i - \mathbf{x}_i^T \mathbf{B})^2$. Toutefois, l'une des raisons pour lesquelles on utilise une enquête à bases de sondage multiples, plutôt qu'une base de sondage incomplète, est la crainte que les caractéristiques de la population puissent différer d'un domaine à l'autre. Lu (2014b) propose d'examiner les résidus séparément par domaine et d'ajuster des modèles de régression séparés par domaine pour évaluer la pertinence du modèle de régression.

3.3. Calage

L'estimateur par le PMV est calé en fonction des chiffres de population connus pour les bases de sondage et les domaines. Dans une enquête à double base où $N^{(1)}$ et $N^{(2)}$ sont connus, $\sum_{q=1}^{2} \sum_{i \in S_q} w_i^{(q)} m_{i,\text{PML}}^{(q)} \delta_i^{(f)} = N^{(f)}$ pour $f = 1, 2$. Si la taille du domaine de chevauchement $N_{\{1,2\}}$ est également connue, l'estimateur par le PMV est calé sur les trois tailles de domaine. Skinner (1991) a utilisé le calage avec l'estimateur à base de sondage unique, en appliquant l'estimation itérative par le quotient aux chiffres de la base de sondage de la population.

Ranalli, Arcos, Rueda et Teodoro (2016) ont étudié la théorie générale du calage pour les enquêtes à double base. Ils ont supposé qu'un vecteur d'information auxiliaire $\mathbf{x}$ est disponible avec des totaux de population connus $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i$ et ils ont calculé des poids de régression généralisée à bases multiples comme étant

$$c_i^{(q)} = \tilde{w}_i^{(q)} \left[1 + (\mathbf{X} - \hat{\mathbf{X}})^T \left(\sum_{f=1}^{Q} \sum_{k \in S_f} \alpha_k \tilde{w}_k^{(f)} \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \alpha_i \mathbf{x}_i \right], \tag{3.4}$$

où α_k est une constante arbitraire et $\hat{\mathbf{X}} = \sum_{f=1}^{Q} \sum_{k \in S_f} \tilde{w}_k^{(f)} \mathbf{x}_k$ estime $\mathbf{X}$ au moyen des poids ajustés pour la multiplicité. Dans des conditions de régularité, ils ont montré que pour l'estimateur à double base dans (3.2) avec une valeur θ fixe, la variance de l'estimateur par la régression généralisée $\hat{Y}_{\text{GR}} = \sum_{q=1}^{2} \sum_{i \in S_q} c_i^{(q)} y_i$ est calculée approximativement par

$$V(\hat{Y}_{\text{GR}}) \approx V\left[\sum_{q=1}^{2} \sum_{i \in S_q} \tilde{w}_i^{(q)} (y_i - \mathbf{x}_i^T \mathbf{B}) \right], \tag{3.5}$$

où $\mathbf{B} = \left(\sum_{i=1}^{N} \alpha_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{N} \alpha_i \mathbf{x}_i y_i$. La variance de l'estimateur dépend des résidus du modèle de régression, tout comme dans le cas à base de sondage unique.

Särndal et Lundström (2005) ont établi une distinction entre les types d'information auxiliaire utilisables dans le calage. InfoU désigne l'information disponible au niveau de la population. Un vecteur $\mathbf{x}^*$ peut être considéré comme InfoU si le total de population $\mathbf{X}^* = \sum_{i=1}^{N} \mathbf{x}_i^*$ est connu et $\mathbf{x}^*$ est observé pour chaque répondant de l'échantillon. InfoS désigne l'information disponible au niveau de l'échantillon, mais non au niveau de la population. Le vecteur $\mathbf{x}^o$ est considéré comme InfoS s'il est connu pour chaque membre de l'échantillon – à la fois les répondants et les non-répondants – mais que $\sum_{i=1}^{N} \mathbf{x}^o$ est inconnu.

Dans une enquête à bases multiples, les variables disponibles pour InfoU et InfoS peuvent différer d'une base à l'autre. Dans le cadre de la NSAF, on disposait de peu d'information auxiliaire pour les non-répondants dans l'échantillon par CA, mais l'information relative à l'adresse (par exemple les caractéristiques du groupe d'îlots) était connue pour tous les membres de l'échantillon à base aréolaire. L'inverse peut être vrai pour une enquête à double base dans laquelle la base 1 est une base aréolaire et la base 2 est une base liste. La base liste peut avoir une information riche utilisable aux fins d'ajustements de classe de pondération ou de calage, tandis que l'information auxiliaire de la base aréolaire peut être restreinte à l'information mesurée dans l'enquête pour laquelle les totaux de population sont connus à partir d'une source externe comme un recensement ou un registre de population.

Ranalli et coll. (2016) ont permis différentes informations InfoU dans les différentes bases de sondage; certaines des variables auxiliaires peuvent être connues pour des unités de tous les échantillons et pour l'ensemble de la population, tandis que d'autres variables peuvent être de la forme $x_i^* = x_i \delta_i^{(q)}$ avec un total $X^* = \sum_{i=1}^{N} x_i \delta_i^{(q)}$, le total de la variable x dans la base de sondage q. Le calage sur les chiffres de la base $N^{(q)}$ est par conséquent un cas particulier de la théorie générale de calage.

Cependant, les quantités différentes d'information pour les bases peuvent aussi avoir une incidence sur les ajustements fondés sur la multiplicité. Supposons que la base 2 possède de riches informations auxiliaires aux fins de calage tandis que la base 1 a peu d'informations. Le calage des poids $w_i^{(2)}$ avant la composition peut augmenter la taille d'échantillon efficace relative à partir de S_2 et ainsi augmenter la valeur de $\tilde{n}^{(2)} / \left(\tilde{n}^{(1)} + \tilde{n}^{(2)} \right)$ qui serait utilisée pour l'estimateur TEE.

Haziza et Lesage (2016) soutiennent qu'une procédure de pondération en deux étapes présente plusieurs avantages pour les enquêtes à base unique en présence de non-réponse. La première étape divise le poids de sondage pour l'unité i par sa propension à répondre estimée (souvent calculée à partir de l'information InfoS) et la deuxième étape cale les poids rajustés pour la non-réponse sur les totaux de contrôle de la population (disponibles à partir de l'information InfoU). En cas de non-réponse importante, les facteurs d'ajustement de la pondération de l'étape 1 sont souvent beaucoup plus élevés que ceux de l'étape 2; si le modèle de propension à répondre est exact, les ajustements de la pondération de l'étape 2 convergent vers 1 quand $n \to \infty$. La procédure en deux étapes est donc plus robuste face à la spécification erronée du modèle de calage.

Les mêmes considérations s'appliquent aux enquêtes à bases de sondage multiples. Une procédure en deux étapes, dans laquelle l'étape 1 ajuste les échantillons séparément pour la non-réponse et l'étape 2 cale les échantillons combinés, apporte de la robustesse au modèle de calage. Supposons que S_1 a une réponse complète; S_2 a une non-réponse, mais les propensions à répondre peuvent être prédites parfaitement à partir de la variable x. Ensuite, l'exécution d'un ajustement pour la non-réponse séparément pour chaque échantillon à l'étape 1 élimine le biais pour S_2 de sorte que l'hypothèse (A5) est satisfaite. Toutefois, si les données sont d'abord combinées, puis calées au moyen de (3.4), le calage peut modifier les poids des unités dans S_1 afin que soient respectées les contraintes de calage, ce qui introduit un biais pour les estimations réalisées à partir de S_1 tout en ne l'éliminant pas pour les estimations réalisées à partir de S_2. De nouveaux travaux de recherche sont nécessaires sur l'ordre des étapes d'ajustements des poids. Il pourrait être préférable d'effectuer deux étapes d'ajustements pour la non-réponse et de calage sur chaque échantillon séparément, puis d'ajuster les poids pour la multiplicité, et enfin d'effectuer un calage sur les totaux de population (y compris en recalant les variables de chaque base).

Une des conséquences de l'utilisation d'un estimateur avec chevauchement pour une enquête à bases multiples est que les ajustements fondés sur la multiplicité peuvent introduire une plus grande variation de poids et ainsi faire que les observations échantillonnées à partir d'une base aient des poids beaucoup plus grands que les observations échantillonnées à partir de plusieurs bases. Si, par exemple, une base liste (base 2 de la figure 2.2(a, b)) est suréchantillonnée de façon disproportionnée, les poids d'échantillonnage pour les observations dans le domaine $\{1\}$, qui sont échantillonnés uniquement à partir de la base 1, peuvent être grands par rapport aux poids pour les autres domaines. Wolter, Ganesh, Copeland, Singleton et Khare (2019) ont proposé d'utiliser un estimateur de rétrécissement, en estimant $Y_{\{1\}}$ par $\kappa\hat{Y}^{(1)}_{\{1\}} + (1-\kappa)\, N_{\{1\}} (\hat{Y}^{(2)}_{\{2\}} + \hat{Y}_{\{1,2\}}) \big/ N^{(2)}$, mais le rétrécissement peut introduire un biais, alors que la raison pour laquelle on utilise un plan à bases multiples plus compliqué au lieu de se contenter d'un échantillonnage à partir de la base 2 est justement d'éviter tout biais potentiel attribuable à l'omission du

domaine {1}. Une meilleure solution, si elle est faisable, consiste à tenir compte de la variation de poids lors de la conception du plan de l'enquête, comme nous l'avons vu à la section 5.

3.4 Échantillon probabiliste combiné au recensement d'un sous-ensemble de population

Lohr (2014) et Kim et Tam (2021) ont constaté que la situation à la figure 2.2(a) comprend le cas particulier où un échantillon probabiliste S_1 est tiré de la base de sondage 1 ayant une couverture complète, et l'échantillon S_2 de la base de sondage 2 est un recensement de domaine {1, 2}. Le domaine de chevauchement est défini par conséquent comme étant les unités dans S_2, qui peuvent provenir de dossiers administratifs ou d'un échantillon de commodité. Bien que S_2, considéré en soi, puisse avoir un biais de sous-couverture, dans une configuration à bases de sondage multiples, le biais est éliminé par la présence d'un échantillon de la base 1. Les unités dans S_2 ont $w_i^{(2)} = 1$ et se représentent seules; elles ne représentent aucune unité dans d'autres parties de la population. Quand $N^{(2)} / N$ est petit, disons à partir d'un petit échantillon de commodité, S_2 a peu d'effet sur les estimateurs à double base – presque toute la population est dans le domaine {1}. Toutefois, quand $N^{(2)} / N$ est grand, comme cela peut se produire quand la base de sondage 2 se compose de dossiers administratifs, la disponibilité de ces dossiers peut améliorer la précision de $\hat{Y}$ si les hypothèses (A1) à (A6) sont satisfaites.

Quand S_2 est un recensement sans erreur de mesure, $\hat{Y}_{\{1,2\}}^{(2)} = Y_{\{1,2\}}$. L'estimateur dans (3.2) est

$$\hat{Y}(\theta) = \hat{Y}_{\{1\}}^{(1)} + \theta \hat{Y}_{\{1,2\}}^{(1)} + (1-\theta)\, Y_{\{1,2\}}; \tag{3.6}$$

en supposant que $\theta = 0$ utilise le total de population connu de la base de sondage 2 et s'appuie sur la base 1 seulement pour estimer la partie de la population qui ne fait pas partie de la base 2.

Kim et Tam (2021) constatent que, puisque $Y_{\{1,2\}}$ est connu, il peut être utilisé comme total de calage d'InfoU. Ils proposent deux estimateurs par calage : un estimateur par le ratio $\hat{Y}_{\text{ratio}} = \hat{Y}^{(1)} Y_{\{1,2\}} \big/ \hat{Y}_{\{1,2\}}^{(1)}$ et un estimateur par calage par régression généralisée. Cependant, dans de nombreux plans, l'estimateur par le ratio sera moins efficace que $\hat{Y}(0)$ de (3.6) parce que

$$V\left(\hat{Y}_{\text{ratio}}\right) \approx V\left(\hat{Y}_{\{1\}}^{(1)}\right) + \left(\frac{Y_{\{1\}}}{Y_{\{1,2\}}}\right)^2 V\left[\hat{Y}_{\{1,2\}}^{(1)}\right] - 2\frac{Y_{\{1\}}}{Y_{\{1,2\}}} \operatorname{Cov}\left(\hat{Y}_{\{1\}}^{(1)}, \hat{Y}_{\{1,2\}}^{(1)}\right);$$

l'ajustement par le quotient peut introduire une variabilité supplémentaire à partir de $\hat{Y}_{\{1,2\}}^{(1)}$ qui est exclue à partir de $\hat{Y}(0)$.

Si l'on cale $\hat{Y}(\theta)$ sur $Y_{\{1,2\}} = \sum_{i=1}^{N} x_i$, pour $x_i = \delta_i^{(2)} y_i$, les poids de régression généralisée dans (3.4) deviennent

$$c_i^{(q)} = \tilde{w}_i^{(q)} \left[1 + \left(Y_{\{1,2\}} - \hat{Y}_{\{1,2\}}(\theta)\right)\left(\sum_{f=1}^{Q} \sum_{k \in S_f} \tilde{w}_k^{(f)} \delta_k^{(2)} y_k^2\right)^{-1} \delta_i^{(2)} y_i\right], \tag{3.7}$$

ce qui donne $\hat{Y}_{\text{GR}} = \hat{Y}(0)$ à partir de (3.6). De même, un calage sur le vecteur $\mathbf{x}_i = \left(1, \delta_i^{(2)}, \delta_i^{(2)} y_i\right)^T$ donne $\hat{Y}_{\text{GR}} = \hat{Y}_{\{1\}}^{(1)} N_{\{1\}} \Big/ \hat{N}_{\{1\}}^{(1)} + Y_{\{1,2\}}$.

Dans certains plans, la variance peut être réduite davantage encore. Montanari (1987, 1998) propose d'utiliser le coefficient de régression $\boldsymbol{\beta} = \left[V(\hat{\mathbf{X}})\right]^{-1} \text{Cov}\,(\hat{Y}, \hat{\mathbf{X}})$ à des fins de calage, ce qui donne l'estimateur

$$\hat{Y}_{\text{opt}} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})^T \boldsymbol{\beta}. \tag{3.8}$$

Rao (1994) a qualifié (3.8) d'estimateur par la régression optimal et a montré que $V(\hat{Y}_{\text{opt}}) \le V(\hat{Y}_{\text{GR}})$. Dans la situation à double base envisagée dans la présente section, avec $x_i = \delta_i^{(2)} y_i$,

$$\boldsymbol{\beta} = \frac{\text{Cov}\left(\hat{Y}^{(1)}, \hat{Y}_{\{1,2\}}^{(1)}\right)}{V\left(\hat{Y}_{\{1,2\}}^{(1)}\right)} = 1 + \frac{\text{Cov}\left(\hat{Y}_{\{1\}}^{(1)}, \hat{Y}_{\{1,2\}}^{(1)}\right)}{V\left(\hat{Y}_{\{1,2\}}^{(1)}\right)}$$

et

$$\begin{aligned} \hat{Y}_{\text{opt}} &= \hat{Y}^{(1)} + \left(Y_{\{1,2\}} - \hat{Y}_{\{1,2\}}^{(1)}\right) \left[1 + \frac{\text{Cov}\left(\hat{Y}_{\{1\}}^{(1)}, \hat{Y}_{\{1,2\}}^{(1)}\right)}{V\left(\hat{Y}_{\{1,2\}}^{(1)}\right)}\right] \\ &= \hat{Y}_{\{1\}}^{(1)} + \theta_H \hat{Y}_{\{1,2\}}^{(1)} + (1 - \theta_H)\, Y_{\{1,2\}}, \end{aligned} \tag{3.9}$$

où $\theta_H = -\text{Cov}\left(\hat{Y}_{\{1\}}^{(1)}, \hat{Y}_{\{1,2\}}^{(1)}\right) \Big/ V\left(\hat{Y}_{\{1,2\}}^{(1)}\right)$ est la valeur optimale de Hartley pour θ à partir de (3.3).

Bien que nous considérions habituellement que le facteur de composition θ se situe entre 0 et 1, θ_H peut être hors de cette fourchette. Pour donner un exemple conceptuel, supposons que la base de sondage 2 est une liste d'enfants recevant une aide alimentaire à l'école et que l'échantillon de la base 1 est un échantillon de grappes de ménages. Ensuite, les ménages dans lesquels un ou plusieurs enfants reçoivent une aide alimentaire ont certains membres dans le domaine $\{1, 2\}$ et d'autres dans le domaine $\{1\}$. Si y présente une corrélation intra-ménage élevée, nous nous attendons à ce que $\hat{Y}_{\{1\}}^{(1)}$ et $\hat{Y}_{\{1,2\}}^{(1)}$ soient corrélés positivement. Dans ce cas, l'estimateur optimal de Hartley donne des poids négatifs pour les unités du domaine $\{1, 2\}$ tirées de l'échantillon probabiliste.

Bien que $\hat{Y}_{\text{opt}}$ soit plus efficace dans des situations particulières comme l'échantillon de grappes décrit ci-dessus, cette variable dépend en pratique d'une estimation de la covariance, elle est optimale seulement pour cette variable y en particulier et elle peut avoir des poids négatifs. Des poids négatifs peuvent également se produire si l'on effectue un calage optimal avec la variable auxiliaire $\left(1, \delta_i^{(2)}, \delta_i^{(2)} y_i\right)$. En effet, ce calage donne les résultats de l'estimateur proposé par Fuller et Burmeister (1972). Ces estimateurs par la régression optimaux sont sensibles aux hypothèses du modèle et, de manière générale, je ne recommande pas leur utilisation.

Quand l'échantillon de la base de sondage 2 est un recensement et que les hypothèses (A1) à (A6) sont satisfaites, la précision des estimations de la population dépend entièrement du plan de S_1. Quand les

échantillons ne sont pas conçus pour faire partie d'une enquête à bases multiples (et parfois même s'ils le sont), il est probable qu'une ou plusieurs hypothèses ne soient pas respectées. Les hypothèses (A4) et (A6) sont particulièrement suspectes quand on souhaite combiner des données d'enquêtes qui n'ont pas été conçues en vue d'une combinaison. Même si deux enquêtes mesurent le chômage, elles peuvent utiliser des questions différentes, de sorte que les statistiques sur le chômage de S_2 mesurent un concept différent des statistiques de S_1. Une spécification erronée du domaine est également possible. On peut savoir qu'une unité du recensement S_2 est également dans la base de sondage 1 complète, mais il peut être difficile d'affirmer si une unité dans S_1 fait aussi partie des dossiers administratifs ou de l'échantillon de commodité qui sert de S_2. Ces problèmes sont abordés dans la section suivante.

4. Enquêtes à bases multiples et intégration de données

Rao (2021) étudie plusieurs méthodes d'intégration des données pour combiner l'information d'un échantillon probabiliste S_1, supposé provenir d'une base de sondage avec couverture complète, avec l'information d'un échantillon non probabiliste S_2, souvent le recensement d'une partie de la population comme à la section 3.4. Rao examine deux cas pour réaliser des inférences à propos de y : (1) y est observé dans les deux échantillons, et (2) l'information auxiliaire $\mathbf{x}$ est observée dans les deux échantillons, mais y est observé seulement dans S_2. Dans la présente section, j'examine diverses méthodes d'intégration de données dans la perspective du paradigme à bases multiples et des hypothèses de la section 2.2.

4.1 Estimation sur petits domaines

L'estimation sur petits domaines peut être considérée comme un cas particulier de problème d'estimation à double base dans lequel l'hypothèse (A6) n'est pas satisfaite. Ici, S_1 est un échantillon probabiliste tiré de la base de sondage 1 et la base 2 est souvent une source de données administratives. On suppose que les deux bases ont une couverture complète de la population, mais la variable d'intérêt y est mesurée seulement dans S_1. L'information auxiliaire $\mathbf{x}$ utilisée pour prédire y est mesurée dans les deux échantillons. Beaumont et Rao (2021) ont discuté l'intégration des échantillons probabilistes et non probabilistes au moyen de l'estimateur de Fay-Herriot (1979) avec des techniques d'estimation sur petits domaines.

Un estimateur composite sur petits domaines (Rao et Molina, 2015) de la moyenne de population η_a dans le domaine a a la forme

$$\hat{\eta}_a = \theta_a \hat{\eta}_a^{(1)} + (1-\theta_a)\,\hat{\eta}_a^{(2)},$$

où $\hat{\eta}_a^{(1)}$ est l'estimateur direct pour la moyenne de l'échantillon dans le domaine a à partir de S_1 (qui peut avoir une grande variance ou ne pas exister), $\hat{\eta}_a^{(2)} = \mathbf{x}_a^T \hat{\boldsymbol{\beta}}$ est une valeur prédite à partir d'un modèle de régression, et θ_a est un facteur de composition. Pour l'estimateur de Fay-Herriot, θ_a dépend de la précision relative des deux estimateurs sous un modèle de régression supposé dont les paramètres sont

estimés à partir de S_1. Pour l'estimateur $\hat{\eta}_a$, la variable y est mesurée différemment dans les deux bases – on utilise les valeurs prédites pour la base de sondage 2 – et différents facteurs de composition sont utilisés dans différents domaines.

4.2 Imputation massive et appariement d'échantillons

Supposons que S_1 est un échantillon probabiliste à réponse complète de la base 1, mais que la variable d'intérêt y n'est pas mesurée dans S_1. Toutefois, y est mesuré dans S_2 à partir de la base de sondage 2, et les variables auxiliaires $\mathbf{x}$ sont mesurées dans les deux échantillons. Supposons que $\tilde{y}_i$ est la valeur prédite de y_i à partir d'un modèle d'imputation, en liant y_i à $\mathbf{x}_i$, qui est développée sur S_2 et supposons que $\tilde{Y}^{(1)} = \sum_{i \in S_1} w_i^{(1)} \tilde{y}_i$ et $\tilde{Y}_d^{(1)} = \sum_{i \in S_1} w_i^{(1)} \delta_i(d)\, \tilde{y}_i$ sont la population estimée et les totaux de domaine d provenant de S_1 au moyen des valeurs imputées.

Comme dans l'estimation sur petits domaines, l'imputation massive convient au contexte à double base en assouplissant l'hypothèse (A6) selon laquelle il n'y a pas d'erreur de mesure. Kim et Rao (2012) ainsi que Chipperfield, Chessman et Lim (2012) ont examiné une situation où les deux bases de sondage sont complètes et S_1 et S_2 sont tous deux des échantillons probabilistes. Les bases de sondage peuvent différer – la base 1, par exemple, peut être une base aréolaire et la base 2 un registre de population – mais on suppose que les deux ont une couverture complète. Chipperfield et coll. (2012) ont utilisé un estimateur composite

$$\hat{Y}_{\text{imp}} = \theta \tilde{Y}^{(1)} + (1-\theta)\, \hat{Y}^{(2)}, \tag{4.1}$$

où la valeur optimale du facteur de composition θ minimise la variance (compte tenu de la variabilité d'échantillonnage et d'imputation). Kim et Rao (2012) ont proposé d'ajouter une correction du biais avec l'estimateur

$$\tilde{Y}^{(1)} + \sum_{i \in S_2} w_i^{(2)} (y_i - \tilde{y}_i);$$

cet estimateur est de la même forme que (4.1) avec $\theta = 1$ s'il est requis que les paramètres estimés dans le modèle d'imputation satisfassent $\sum_{i \in S_2} w_i^{(2)} (y_i - \tilde{y}_i) = 0$.

Si le modèle d'imputation produit des prédictions exactes et sans biais pour y_i, la combinaison des échantillons augmente la taille d'échantillon efficace aux fins de calcul des estimations. Quand les deux échantillons sont des échantillons probabilistes avec couverture complète, il est possible d'effectuer des diagnostics de modèle sur S_2. Chipperfield et coll. (2012) ont proposé plusieurs diagnostics, y compris la mise à l'essai du modèle d'imputation sur de petits domaines, l'examen de la possibilité de prédire l'appartenance à l'enquête à partir de la valeur de y_i (pour S_2) ou $\tilde{y}_i$ (pour S_1), et l'étude de la sensibilité de l'erreur quadratique moyenne à différents niveaux de biais dans $\tilde{Y}^{(1)}$. Toutefois, la sensibilité des diagnostics dépend de la qualité et de la taille de S_2. Si S_2 est petit par rapport à S_1, S_1 peut contenir des sous-populations qui ne sont pas bien représentées dans S_2 et qui sont mal ajustées par le modèle d'imputation.

La situation se complique quand la base de sondage 2 est incomplète ou quand S_2 comporte un biais de sélection. Quand le domaine $\{1\}$ n'est pas vide, comme dans la figure 2.2(a), alors l'estimateur composite avec des valeurs imputées devient

$$\hat{Y}_{\text{imp}} = \tilde{Y}^{(1)}_{\{1\}} + \theta \tilde{Y}^{(1)}_{\{1,2\}} + (1-\theta)\,\hat{Y}^{(2)}_{\{1,2\}}. \tag{4.2}$$

Les propriétés de l'estimateur dans (4.2) dépendent de la mesure dans laquelle le modèle d'imputation prédit correctement les valeurs de y_i dans S_1. Plusieurs méthodes d'imputation ont été proposées. Avec l'appariement d'échantillons (Rivers, 2007), $\tilde{y}_i$ pour l'observation i dans S_i est défini comme étant égal à la valeur de y_i du plus proche voisin de l'observation (pour ce qui est des valeurs de $\mathbf{x}$) dans S_2. Envisageant la situation dans laquelle S_2 est un échantillon de commodité, Rivers (2007) a pris $\theta = 1$ dans (4.2) et utilisé l'information dans S_2 dans le seul but de trouver les valeurs imputées $\tilde{y}_i$ pour S_1. Yang, Kim et Hwang (2021) ont étudié les propriétés théoriques d'estimateurs imputés massivement qui utilisent les méthodes du plus proche voisin.

S'appuyant sur les travaux de Lee (2006), Lee et Valliant (2009) ainsi que Valliant et Dever (2011) concernant l'utilisation de la pondération par le score de propension pour estimer des caractéristiques de population à partir d'un échantillon non probabiliste, Chen, Li et Wu (2020) ont proposé un estimateur « doublement robuste » pour une situation où $\mathbf{x}_i$ est mesuré dans les deux enquêtes, mais y_i est mesuré seulement dans l'échantillon non probabiliste S_2. Soit $R_i^{(2)} = 1$ si l'unité de population i est dans S_2 et 0 sinon. Sous des hypothèses fortes selon lesquelles (1) $R_i^{(2)}$ et y_i sont indépendants étant données les covariables $\mathbf{x}_i$, (2) $\pi_i^{(2)} = P\left(R_i^{(2)} = 1\right) > 0$ pour toutes les unités de population i, et (3) $R_i^{(2)}$ et $R_j^{(2)}$ sont indépendantes conditionnellement étant donné $\mathbf{x}$, ils ont estimé $\pi_i^{(2)}$ comme une fonction de $\mathbf{x}_i$, en utilisant l'information dans S_1, et ils ont proposé l'estimateur

$$\hat{Y}_{\text{DR}} = \sum_{i \in S_1} w_i^{(1)} \tilde{y}_i + \sum_{i \in S_2} \frac{1}{\hat{\pi}_i^{(2)}} (y_i - \tilde{y}_i),$$

où $\tilde{y}_i$ est une prédiction par imputation pour les valeurs inconnues de y dans S_1 (développé au moyen de l'information dans S_2). L'estimateur $\hat{Y}_{\text{DR}}$ est approximativement sans biais pour Y si le modèle d'imputation ou le modèle de prédiction $\pi_i^{(2)}$ est correct. Si le modèle d'imputation est correct, alors le premier terme de $\hat{Y}_{\text{DR}}$ est approximativement sans biais pour Y et le second terme a la valeur attendue 0. Si le modèle prédisant $\pi_i^{(2)}$ est correct, alors $\sum_{i \in S_2} y_i / \hat{\pi}_i^{(2)}$ est approximativement sans biais pour Y et $E\left[\sum_{i \in S_1} w_i^{(1)} \tilde{y}_i - \sum_{i \in S_2} \tilde{y}_i / \hat{\pi}_i^{(2)}\right] \approx 0$. Toutefois, si aucun des modèles n'est correct, $\hat{Y}_{\text{DR}}$ peut avoir un biais important.

Kim et Tam (2021) ont considéré une extension de la situation de la section 3.4 dans laquelle y_i n'est pas mesuré dans S_1, ou est mesuré autrement que dans S_2, et ils ont proposé de substituer une valeur imputée $\tilde{y}_i$ à la valeur y_i dans les estimateurs à partir de S_1 dans (3.6), de façon à obtenir l'estimateur dans (4.2) avec $\theta = 0$. Ils ont calé cet estimateur sur la taille de domaine connue $N_{\{1,2\}}$.

4.3 Imputation et NSAF

Les estimateurs de la section 4.2 imputent une valeur prédite $\tilde{y}_i$ pour la valeur inconnue de y_i dans S_1. Ils ont tous une hypothèse forte selon laquelle le modèle d'imputation développé sur S_2 s'applique aux unités du domaine {1}. Comme Lu (2014b) l'a constaté lors de l'étude de la régression pour des enquêtes à double base, les relations entre $\mathbf{x}$ et y peuvent différer d'un domaine à l'autre. Ainsi, un modèle d'imputation élaboré sur un échantillon à partir d'une base de sondage incomplète, ou sur un échantillon comportant un biais de sélection, peut donner de mauvaises prédictions pour y dans d'autres parties de la population. De plus, sans données sur y dans la partie de population imputée, il se peut qu'il soit impossible d'évaluer la qualité des prédictions.

Une enquête à double base a été menée pour la NSAF parce que les enquêteurs craignaient que les caractéristiques d'intérêt puissent différer chez les ménages avec et sans téléphone. Soit $y_i = 1$ si l'enfant i se trouve dans un ménage qui est sous 200 % du seuil de pauvreté, et 0 sinon. En utilisant l'échantillon complet des deux bases de sondage (*Urban Institute et Child Trends*, 2007), on estime que 42,2 % des enfants vivaient dans des ménages situés sous 200 % du seuil de pauvreté, avec une erreur type de 0,5 %. Le pourcentage estimé de l'échantillon par CA était de 38,6 % et le pourcentage estimé de l'échantillon aréolaire était de 93,4 %. Les enfants des ménages sans téléphone, échantillonnés à partir de la base aréolaire, étaient beaucoup plus susceptibles de vivre dans la pauvreté.

Supposons maintenant que la NSAF n'ait pas mesuré les variables de la pauvreté et du revenu dans l'échantillon aréolaire et que y_i ait été imputé au moyen des relations de régression développées dans l'échantillon par CA. Dans de nombreuses enquêtes, les seules informations disponibles aux fins de développement d'un modèle d'imputation sont les variables démographiques. L'ajustement d'un modèle de régression logistique à l'échantillon par CA qui prédit y à partir de la race (selon les catégories : blanche, noire et autre) et l'attribution de chaque enfant d'un échantillon aréolaire à la catégorie ayant la probabilité prédite la plus élevée donnent une estimation de 30,5 % d'enfants de l'échantillon aréolaire vivant dans la pauvreté, soit une valeur plus basse que celle de l'échantillon par CA. Si l'on ajoute au modèle une variable indiquant que l'enfant vit dans un ménage monoparental, le pourcentage estimé de l'échantillon de la région atteint 51,9 %. Ces deux estimations, ainsi que les estimations calculées au moyen d'imputations de la moyenne cellulaire, sont nettement en deçà du pourcentage de 93,4 % obtenu à partir des données réelles.

Bien entendu, le problème s'explique par l'information auxiliaire insuffisante qui ne permet pas de fournir une bonne prédiction de la pauvreté dans l'échantillon aréolaire. La principale caractéristique des données, et la raison pour laquelle Waksberg et ses collaborateurs ont utilisé une enquête à double base, est que l'absence de téléphone est fortement associée à la pauvreté. Cette association ne peut pas être estimée à partir de l'échantillon par CA, dans lequel tous les ménages ont un téléphone. Il serait possible de développer un modèle d'imputation au moyen de l'information provenant d'autres enquêtes, comme la *Current Population Survey*, dans laquelle les ménages avec et sans téléphone sont échantillonnés, mais je n'ai pas pu trouver de modèle d'imputation prédisant y à partir de variables fondées sur d'autres facteurs que le revenu dans l'échantillon par CA qui fournisse de bonnes prédictions.

Les ménages sans téléphone représentaient une petite partie de la population de la NSAF, mais les différences entre les relations multivariées dans les ménages avec et sans téléphone étaient si grandes que l'imputation ne réduisait que légèrement le biais. Cependant, si la pauvreté n'avait pas été mesurée pour l'échantillon sans téléphone, et si les statistiques publiées s'étaient fondées seulement sur les imputations, il n'y aurait pas eu de moyen de détecter le biais.

4.4 Classification erronée de domaine

L'un des principaux défis liés à la combinaison de données au moyen d'une méthode à bases multiples consiste à déterminer l'appartenance au domaine (ou la multiplicité) des unités dans les sources de données. Cette difficulté se pose y compris pour les enquêtes conçues pour utiliser des bases de sondage multiples.

La NSAF a été conçue comme une enquête de présélection où les ménages avec téléphone étaient exclus de l'échantillon aréolaire. Tous les ménages échantillonnés à partir de la base 2, la base par CA, ont été classés correctement puisqu'il a été communiqué avec eux par téléphone. Le plus difficile était d'obtenir la bonne classification de domaine pour les ménages de l'échantillon de la base aréolaire. Les questions initiales de présélection demandaient si le ménage avait des téléphones en état de marche; les ménages qui répondaient par la négative étaient transférés à l'intervieweur par téléphone qui a effectué l'interview détaillée. L'intervieweur par téléphone a mené une autre brève interview de présélection et a reposé des questions sur les services téléphoniques. 7 % supplémentaires des ménages ont été exclus après avoir répondu aux questions plus détaillées concernant la possession d'une ligne téléphonique. Certains avaient en effet dit à l'intervieweur en personne qu'ils n'avaient pas de téléphone parce qu'ils pensaient que l'intervieweur voulait l'emprunter. D'autres avaient mal compris la question au sujet de la possession d'un téléphone : ainsi, un répondant ayant répondu aux questions de présélection dans le salon croyait que la question concernait seulement les téléphones dans le salon et n'a pas mentionné le téléphone dans la chambre à coucher (Cunningham, Shapiro et Brick, 1999). Même s'il est possible que la deuxième interview de présélection ait corrigé une erreur de classification s'expliquant par le fait que des répondants aient dit par erreur ne pas posséder de téléphone pendant la présélection, il n'y avait pas de remède contre les erreurs de classification possibles en raison de répondants ayant dit à la présélection qu'ils avaient le téléphone alors que ce n'était pas le cas. Une classification erronée dans ce sens peut expliquer en partie pourquoi les enquêteurs avaient un échantillon de ménages sans téléphone de taille plus petite que ce qu'ils attendaient.

Dans les enquêtes téléphoniques à double base de sondage, on détermine généralement le domaine de la figure 2.2(b) (cellulaire seulement, ligne filaire seulement, ou les deux) en demandant au répondant de quels autres téléphones il dispose et, parfois, le temps d'utilisation relative de chaque type de téléphone. Brick, Flores-Cervantes, Lee et Norman (2011) ont constaté que leurs échantillons avec téléphone filaire et leurs échantillons avec téléphone cellulaire avaient tous deux des proportions estimées plus faibles d'utilisateurs doubles que ce qui était attendu d'après les statistiques recueillies sur la possession de téléphone dans le cadre de la *National Health Interview Survey* (Enquête nationale sur la santé réalisée par

interviews des États-Unis). Ils ont supposé que cela était attribuable aux personnes ayant accès aux deux types de téléphone, mais qui utilisaient rarement l'un des deux.

L'appartenance à un domaine peut être inconnue ou difficile à estimer en cas de combinaison de sources de données existantes. Dans certains cas, par exemple en cas de combinaison de listes administratives, on pourrait coupler des enregistrements, ou les fichiers de données peuvent contenir de l'information qui indique si l'unité se trouve dans d'autres bases. Dans d'autres, on dispose de peu ou pas d'information sur l'appartenance au domaine. Comment savoir si un participant à une enquête par panel à participation volontaire fait aussi partie d'une base de bénéficiaires de l'Assurance maladie si aucune question sur l'Assurance maladie n'est posée dans l'enquête ?

Lohr (2011) constate que même une petite quantité de classification erronée de domaine peut entraîner des biais importants dans les estimateurs à double base. De plus, le calage sur des nombres de domaine fondés sur des classifications erronées peut aggraver le biais. Elle propose une méthode d'ajustement du biais attribuable à la classification erronée de domaine, en supposant que les probabilités de classification erronée P (observation classée dans le domaine d' | observation réellement dans le domaine d) sont connues ou peuvent être estimées avec exactitude pour différents sous-groupes de population. Lin, Liu et Stokes (2019) ont étudié une méthode semblable utilisant les probabilités de classification erronée P (observation réellement dans le domaine d | observation classée dans le domaine d').

Il serait possible d'utiliser des méthodes à bases de sondage multiples quand l'appartenance au domaine est inconnue si la probabilité que l'unité i soit dans le domaine d peut être estimée à partir de l'information auxiliaire $\mathbf{x}_i$ connue pour toutes les unités échantillonnées. Kim et Tam (2021) proposent de remplacer l'appartenance à un domaine inconnu par un estimateur pour la situation présentée à la section 3.4, où S_2 est le recensement d'un sous-ensemble de la population. Ils établissent $\tilde{\delta}_i(\{1,2\}) = 1$ si la probabilité prédite que l'unité $i \in S_1$ soit dans le domaine $\{1,2\}$, $\hat{P}\left[\delta_i(\{1,2\}) = 1 | \mathbf{x}_i\right]$, dépasse 1/2, et ils estiment le total de population pour le domaine $\{1\}$ comme étant $\sum_{i \in S_1} w_i^{(1)} \left[1 - \tilde{\delta}_i(\{1,2\})\right] y_i$.

Quand l'appartenance à un domaine est imputée, l'erreur quadratique moyenne dépend de l'exactitude des imputations de domaine ainsi que des caractéristiques du plan et du biais de non-réponse dans S_1. D'autres recherches sont nécessaires pour établir les propriétés statistiques des estimateurs quand l'appartenance à un domaine est estimée. Il peut aussi être souhaitable d'étudier d'autres estimateurs qui utilisent directement les probabilités prédites pour estimer le total dans le domaine $\{1\}$ comme étant $\sum_{i \in S_1} w_i^{(1)} \hat{P}\left[\delta_i(\{1,2\}) = 0 | \mathbf{x}_i\right] y_i$.

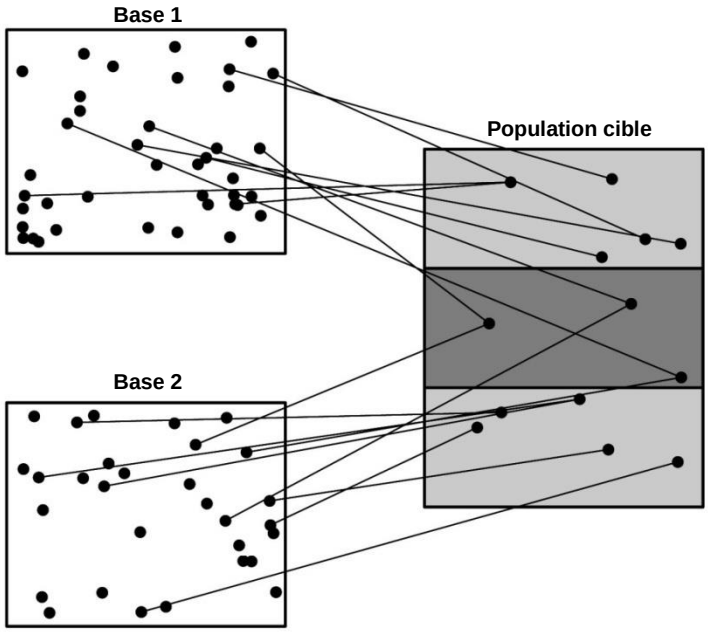
Dever (2018) a utilisé l'appariement d'échantillons pour évaluer le chevauchement de la base pour un échantillon probabiliste S_1, tiré d'une base de sondage fondée sur l'adresse, et un échantillon non probabiliste S_2 recruté à partir de sites de médias sociaux. Elle s'est intéressée au pourcentage de répondants dans S_1 qui n'avaient pas d'appariement proche dans S_2. Bien que cette procédure ne fournisse pas d'estimation sans biais de la taille du domaine $\{1\}$, un pourcentage élevé de cas sans appariement pour de grands échantillons peut indiquer que S_2 représente une population différente de S_1.

4.5 Sondage indirect et estimation de la saisie-ressaisie

Les sections 4.2 à 4.4 ont examiné des extensions d'estimateurs à bases multiples qui ont assoupli les hypothèses (A2), (A4) et (A6). Cependant, toutes ces extensions supposaient qu'au moins une des bases, ou l'union des bases, avait une couverture complète. Examinons maintenant un exemple où l'hypothèse (A1) de couverture complète est assouplie et des bases multiples sont utilisées pour estimer la taille de la population.

Dans un échantillonnage indirect, la population cible se compose d'unités qui sont liées à des unités de la base de sondage, mais qui ne sont pas nécessairement dans cette base (Lavallée, 2007) – les unités de la population cible sont échantillonnées indirectement par les liens aux unités d'échantillonnage de la base de sondage. Lavallée et Rivest (2012) ont étendu cette idée à l'échantillonnage à bases multiples. Supposons par exemple que la population cible se compose de travailleurs en soins à domicile, qui fournissent des soins rémunérés à domicile à des personnes âgées, malades ou handicapées. La base 1 pourrait être une liste de personnes recevant des prestations d'assurance maladie et la base 2 pourrait être une liste d'aides-soignants à domicile provenant d'organismes d'emploi ou de délivrance de permis d'exercer. On demande aux personnes de l'échantillon de la base 1 de donner l'identité des travailleurs qui leur fournissent des soins à domicile, qui sont ensuite interviewés. Un échantillon de travailleurs de la base 2 est également interviewé. Les travailleurs en soins à domicile identifiés dans l'échantillon de la base de sondage 1 peuvent avoir des liens avec plusieurs personnes de la base de sondage 1 et peuvent aussi être dans la base 2. De même, les personnes de l'échantillon de la base de sondage 2 peuvent également avoir des liens vers des unités de la base 1. La figure 4.1 présente un exemple de structure de couplage.

Figure 4.1 Échantillonnage indirect à deux bases de sondage liées à la population cible. Les unités dans la zone ombrée foncée ont des liens avec les deux bases.



Dans un échantillonnage indirect, les Q bases de sondage peuvent contenir différents types d'unités (la situation avec différents types d'unités est également étudiée par Hartley, 1974). Nous ne nous intéressons pas au chevauchement des bases de sondage (représentées comme ne se chevauchant pas à la figure 4.1 parce qu'elles contiennent différents types d'unités), mais au chevauchement pour les unités de la population cible. Les unités échantillonnées dans la population cible ont plusieurs chances de sélection si elles sont liées à plusieurs unités dans une ou deux bases de sondage.

Soit $l_{j,k}^{(q)} = 1$ si l'unité j de la base q est liée à l'unité k de la population cible, et soit $L_k^{(q)}$ le nombre total de liens entre l'unité k dans la population cible et la base q (on suppose qu'il est possible de la connaître par interrogation de l'unité k). On peut alors trouver un estimateur $\hat{Y}^{(q)}$ pour chaque base en utilisant les liens comme suit :

$$\hat{Y}^{(q)} = \sum_{j \in S^{(q)}} w_j^{(q)} \sum_k \frac{l_{j,k}^{(q)}}{L_k^{(q)}} y_k = \sum_k u_k^{(q)} y_k,$$

où

$$u_k^{(q)} = \sum_{j \in S^{(q)}} w_j^{(q)} \frac{l_{j,k}^{(q)}}{L_k^{(q)}}.$$

Dans le contexte de notre exemple, la personne j dans S_1 pourrait dire qu'elle reçoit des soins à domicile rémunérés de la part du fournisseur k, ce qui donne $l_{j,k}^{(1)} = 1$. On demanderait alors au fournisseur de soins à domicile lié combien d'autres personnes pour lesquelles il travaille sont bénéficiaires de l'assurance maladie (en supposant qu'il possède cette information ou qu'elle peut être déterminée à partir d'autres sources), ce qui donne la valeur $L_k^{(1)}$. La quantité $u_k^{(q)}$ additionne les poids des unités dans S_q avec des liens à l'unité k, ce qui ajuste pour la multiplicité des liens à cette base. Si $w_j^{(q)} = 1/\pi_j^{(q)}$, alors

$$E[u_k^{(q)}] = E\left[\sum_{j \in S^{(q)}} w_j^{(q)} \frac{l_{j,k}^{(q)}}{L_k^{(q)}}\right] = a_k^{(q)},$$

où $a_k^{(q)} = 1$ si le membre de la population cible k est lié à au moins une unité dans la base q et 0 sinon.

On peut ensuite estimer les caractéristiques de la population de fournisseurs de soins à domicile au moyen de méthodes à bases multiples, en supposant que l'unité k liée à partir de S_q fournit une information exacte sur (1) le nombre de liens à des membres de la base $q\,(L_k^{(q)})$, nécessaire aux ajustements fondés sur la multiplicité avec la base q, et (2) le fait qu'elles soient aussi liées, ou pas, à une ou plusieurs autres bases ($a_k^{(f)}$ pour $f \neq q$), nécessaire aux ajustements pour tenir compte de la multiplicité du couplage à partir de différentes bases.

Lavallée et Rivest (2012) ont fait remarquer que si l'union des deux bases a une couverture incomplète – l'hypothèse (A1) n'est pas respectée – les échantillons des deux bases peuvent servir à estimer la taille de la population cible. Soit $\hat{T}^{(q)} = \sum_k u_k^{(q)}$ pour $k = 1, 2$. Alors, $E[\hat{T}^{(q)}]$ est le nombre de membres de la population cible qui peuvent être couplés à partir de la base de sondage q. Chaque échantillon fournit également une estimation du nombre d'unités de la population cible qui peuvent être couplées à partir des deux bases de sondage : $\hat{T}_{\{1,2\}}^{(1)} = \sum_k u_k^{(1)} a_k^{(2)}$ et $\hat{T}_{\{1,2\}}^{(2)} = \sum_k u_k^{(2)} a_k^{(1)}$. On peut les

composer pour obtenir un estimateur $\hat{T}_{\{1,2\}}$ du nombre de personnes dans la population cible qui peuvent être saisies à partir des deux bases.

On peut employer l'estimateur de la taille de la population par saisie-ressaisie de Lincoln-Petersen. Sous l'hypothèse forte que le fait d'être saisi par la base 1 est indépendant du fait d'être saisi par la base 2, le nombre total de fournisseurs de soins à domicile peut être estimé par $\hat{T}^{(1)}\hat{T}^{(2)}/\hat{T}_{\{1,2\}}$. Dans certains cas, quand l'hypothèse d'indépendance n'est pas satisfaite pour l'ensemble de la population, elle peut l'être approximativement pour des sous-populations dont les nombres estimés peuvent être additionnés. S'il y a plus de deux bases, des modèles loglinéaires peuvent servir à étudier les associations entre bases (Lohr, 2022, chapitre 14); Zhang (2019) a présenté un modèle pour une situation dans laquelle les bases peuvent contenir des unités mal classées.

Alleva, Arbia, Falorsi, Nardelli et Zuliani (2020) ont proposé d'utiliser un échantillonnage indirect à bases multiples pour estimer le nombre de personnes infectées par le SRAS-CoV-2 au cours des premiers stades de la pandémie de COVID-19 en 2020. Cette information était nécessaire pour estimer les paramètres de transmissibilité et d'infection dans des modèles épidémiologiques. Dans cette application, la base de sondage 1 comprend les personnes dont les infections ont été vérifiées (renseignements provenant peut-être d'hôpitaux, de centres de quarantaine ou de cliniques), et la base 2 comprend les autres personnes; les personnes dans S_2 sont soumises à un test de dépistage du SRAS-CoV-2. L'échantillon couplé comprend les personnes qui ont eu des contacts au cours des 14 derniers jours avec une personne dans S_1 ou avec un membre de S_2 ayant obtenu un résultat positif au test.

5. Conception de systèmes de collecte de données

La section 4 traite de la façon dont on peut penser les estimateurs des données intégrées dans une structure d'enquête à bases multiples. Cette structure peut également être utilisée lors de la conception de systèmes de collecte de données utilisant des sources multiples. Hartley (1962) a calculé les valeurs de $n^{(1)}$, $n^{(2)}$ et θ qui minimisent la variance de $\hat{Y}(\theta)$ dans (3.2) quand S_1 et S_2 sont tous deux des échantillons aléatoires simples. Il est possible d'étendre sa méthode de base pour examiner les effets des choix de plans d'échantillonnage dans d'autres situations en tenant compte des erreurs quadratiques moyennes dans diverses hypothèses de biais possibles.

De nombreux travaux ont porté sur le plan optimal et les effets de la non-réponse pour les sondages à double base téléphonie cellulaire/téléphonie filaire. Brick, Dipko, Presser, Tucker et Yuan (2006) et Brick et coll. (2011) ont étudié les erreurs non dues à l'échantillonnage; Lu, Sahr, Iachan, Denker, Duffy et Weston (2013) ont réalisé une étude par simulations pour calculer l'erreur quadratique moyenne anticipée selon divers modèles de coûts et biais potentiels. En étudiant l'allocation des ressources dans les enquêtes téléphoniques à double base avec non-réponse, Lohr et Brick (2014) ont constaté que pour certaines structures de coûts, une enquête de présélection, dans laquelle les répondants ayant des lignes filaires sont éliminés de l'échantillon des répondants avec téléphone cellulaire, était plus rentable qu'une enquête avec chevauchement. Levine et Harter (2015) ont présenté des résultats graphiques pour fournir une orientation en matière d'allocation, en tenant compte de l'inflation de la variance par rapport à la variation de poids. Chen, Stubblefield et Stoner (2021) ont examiné le problème de plan représenté par le suréchantillonnage

des populations minoritaires dans les enquêtes téléphoniques à double base de sondage, au moyen de méthodes d'allocation optimales à partir d'un échantillonnage stratifié. La plupart de ces articles sont centrés sur la réduction de la variance des estimations pour un coût fixe et ne tiennent pas compte des effets du biais potentiel.

Dans les années 1980, plusieurs articles ont étudié les structures d'erreur et les plans d'échantillonnage pour les enquêtes à double base de sondage, habituellement en complétant un échantillon tiré d'une base par CA par un échantillon de base aréolaire dont la couverture était supposée complète. Biemer (1984) et Choudhry (1989) ont examiné des plans optimaux en théorie et au moyen d'études par simulations. Groves et Lepkowski (1985, 1986) et Traugott, Groves et Lepkowski (1987) ont étudié les plans à double base en vue de réduire au minimum l'erreur quadratique moyenne quand les estimations de la base par CA peuvent être biaisées. Lepkowski et Groves (1986) ont constaté qu'à mesure que le biais augmentait dans l'échantillon par CA, son allocation optimale diminuait, atteignant une allocation nulle quand le biais était de 9 % du pourcentage estimé anticipé.

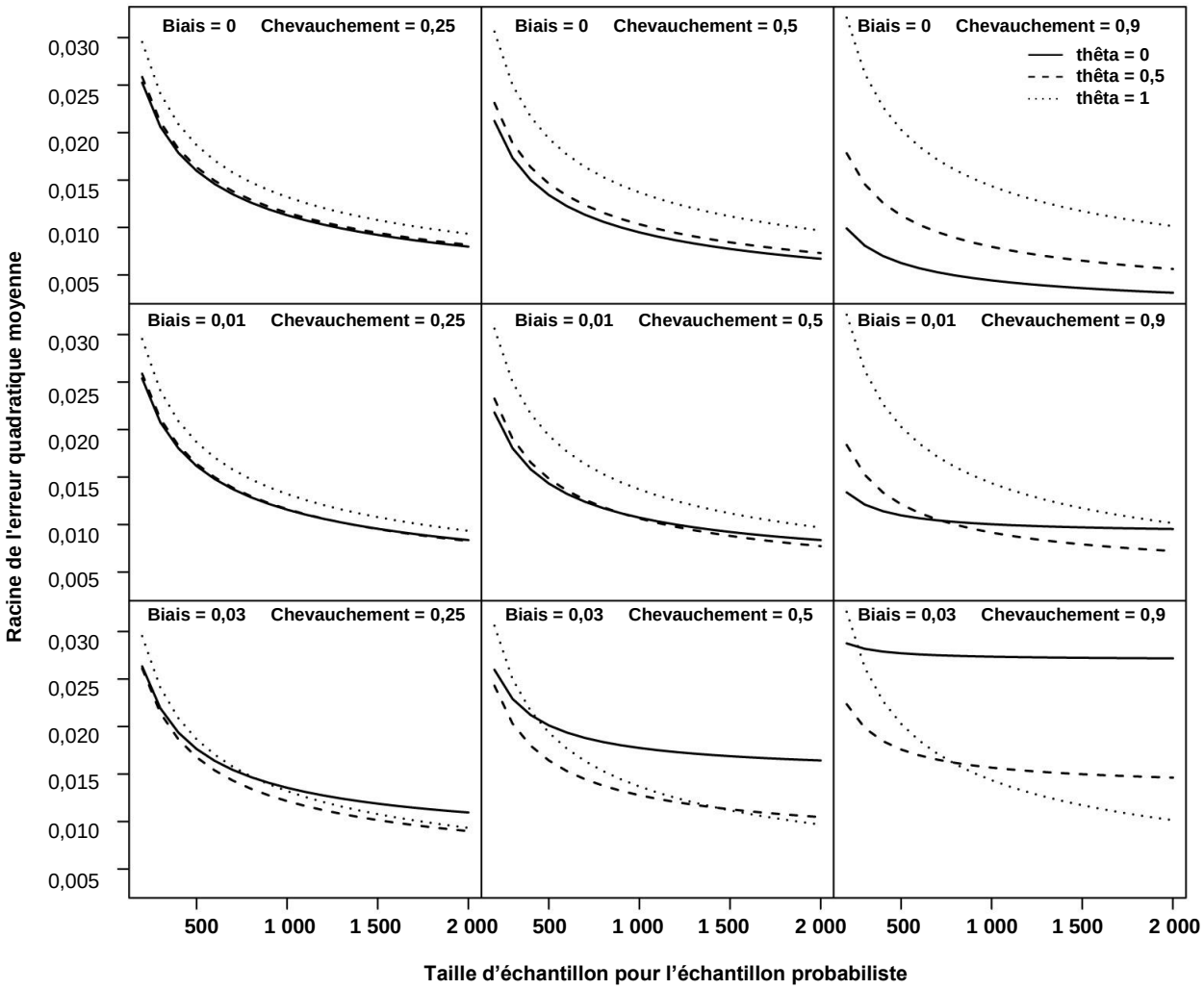
Un léger biais peut avoir un effet semblable pour la situation examinée à la section 3.4, où un recensement est tiré de la base de sondage 2 incomplète et un échantillon probabiliste de grande qualité est tiré de la base 1 complète. Les graphiques de la figure 5.1 montrent la racine de l'erreur quadratique moyenne (REQM) pour une proportion estimée quand S_1 est un échantillon aléatoire simple de taille n et S_2 est un recensement du domaine $\{1, 2\}$, pour les combinaisons de taille de chevauchement $N_{\{1, 2\}} / N$ dans {0,25; 0,5; 0,9} et le biais dans {0; 0,01; 0,03}. La proportion de la population est de 0,2 dans le domaine $\{1\}$ et de 0,3 dans le domaine $\{1, 2\}$, et la proportion de la population globale est estimée au moyen de $\hat{Y}(\theta)/N$ pour $\hat{Y}(\theta)$ dans (3.2). Les lignes montrent la REQM de chaque n pour $\theta = 1$ (S_2 n'est pas utilisé du tout), $\theta = 0$ (la proportion estimée dans le domaine $\{1, 2\}$ provient de S_2 et S_1 contribue uniquement à l'estimation de la proportion dans le domaine $\{1\}$), et $\theta = 1/2$. Dans la ligne inférieure des graphiques, le biais de S_2 commence à dominer la REQM y compris pour des tailles d'échantillon relativement petites à partir de S_1. Une faible quantité de biais de mesure peut annuler l'avantage supposé de l'intégration des données. Cet exemple suppose que l'erreur dans S_2 est due au biais de mesure, mais il n'est pas sans faire penser à l'exemple de Meng (2018), qui montre que même quand le biais de sélection d'un échantillon de commodité est petit, une taille d'échantillon aléatoire simple de 400 peut contenir plus d'informations utiles qu'un échantillon de commodité dont la taille est de 500 millions.

Comme Thompson (2019) le fait remarquer, bon nombre des méthodes élaborées pour combiner des données provenant de sources multiples étaient conçues pour des situations en particulier, avec des solutions adaptées aux circonstances particulières du problème posé. On ne doit pas s'attendre à ce que ces méthodes fonctionnent aussi bien, en moyenne, pour d'autres situations, en raison des effets de la régression vers la moyenne. Avant d'adopter une méthode de combinaison de données, il peut être souhaitable d'effectuer des études par simulations supplémentaires, qui tiennent compte des résultats quand les hypothèses du modèle ne sont pas satisfaites.

Lohr et Raghunathan (2017) ont discuté les problèmes de conception de systèmes de collecte de données exploitant plusieurs sources de données, en s'intéressant particulièrement à situation dans laquelle

une enquête probabiliste est utilisée conjointement avec des sources de données administratives qui couvrent une partie de la population. Ils ont envisagé l'utilisation de sources administratives pour (1) améliorer la base pour l'échantillon probabiliste, (2) fournir des informations contextuelles aux fins d'interprétation des données de l'enquête, (3) fournir de l'information aux fins de suivi de la non-réponse et d'évaluation du biais, et (4) concevoir tout le système de collecte de données de façon à tirer profit de la collecte de données peu coûteuse permise par certaines des bases tout en obtenant une couverture complète à partir de l'enquête probabiliste. Une réflexion sur le problème de plan dans le paradigme à bases multiples peut être utile pour le dernier point. Lohr et Raghunathan (2017) ont avancé que quand la base de sondage 1 est complète mais que son échantillonnage est coûteux, tandis que la base 2 est incomplète mais son échantillonnage moins coûteux, comme la situation examinée dans la section 3.4 du présent article, il peut être souhaitable d'employer une enquête de présélection à deux phases pour l'échantillon tiré de la base 1 et de compter sur l'échantillon tiré de la base 2 afin d'obtenir l'information pour le domaine $\{1, 2\}$. C'est la stratégie que Waksberg et ses collaborateurs ont adoptée dans la conception du plan de la NSAF.

Figure 5.1 Racine de l'erreur quadratique moyenne de la proportion de population estimée selon différents niveaux de chevauchement et de biais.



Quand une erreur de mesure ou de classification erronée de domaine est possible, il est toutefois préférable d'utiliser un plan plus robuste. Les plans optimaux pour les enquêtes à double base allouent les ressources de façon à réduire au minimum la variance des totaux de population estimés d'intérêt pour les coûts fixes. Les plans optimaux selon les hypothèses (A1) à (A6) ne sont pas nécessairement optimaux quand certaines de ces hypothèses ne sont pas respectées. La structure à bases multiples permet de tenir compte des performances potentielles du plan en cas d'assouplissement des hypothèses.

Hartley (1962) a démontré qu'une enquête à double base de sondage augmentait considérablement l'efficacité dans les situations de la figure 2.2(a, b) quand les données peuvent être obtenues à peu de frais à partir de la base 2 et que $N_{\{1,2\}}/N$ est grand. Toutefois, quand la base de sondage 1 est complète et que les coûts sont comparables ou que $N_{\{1,2\}}/N$ est petit, la complexité supplémentaire représentée par l'utilisation d'une enquête à double base peut l'emporter sur ses avantages. Si, en outre, il est probable qu'il y ait une spécification erronée de domaine ou si y est mesurée différemment d'une enquête à l'autre, une enquête à double base sera plus compliquée qu'un seul échantillon de la base 1 et peut produire des estimations biaisées.

En revanche, l'utilisation de sources de données multiples peut aider à évaluer les erreurs non liées à l'échantillonnage. Hartley (1974) a écrit que lorsqu'il a présenté son travail sur les enquêtes à bases multiples à une conférence, un participant à la discussion a laissé entendre qu'une comparaison « plus juste » consisterait à comparer la variance d'un échantillon à double base avec celle d'un seul échantillon ayant le même coût et provenant de la base de sondage incomplète mais bon marché. Hartley a répondu (page 107) : « La difficulté à cet égard, bien entendu, est que le biais attribuable au caractère incomplet peut être d'une ampleur qui rendrait l'enquête à base unique inutile. Si aucune information a priori sur ce biais n'est disponible, l'enquête à double base peut en fait être considérée comme une méthode économique pour *mesurer* ce biais *et* l'éliminer. »

C'est pourquoi il peut être souhaitable de concevoir le système de collecte de données avec plusieurs objectifs, à savoir : (1) obtenir des estimations des principales quantités de population avec une petite erreur quadratique moyenne, (2) évaluer les erreurs non dues à l'échantillonnage à partir des sources de données, et (3) fournir de l'information pour améliorer les futurs plans d'enquête. Voici certaines des questions à prendre en considération :

- Qualité et stabilité des sources de données. La théorie classique du plan d'enquête à bases multiples suppose que les bases de sondage sont fixes. Toutefois, il pourrait être souhaitable d'utiliser des sources de données de rechange dans lesquelles la base change au fil du temps (par exemple des prix moissonnés sur le Web) de façon à fournir des informations plus actuelles en coordination avec une enquête probabiliste. Des éléments théoriques sont nécessaires sur les modalités. Si on s'appuie sur des données fournies par une source externe, ces données continueront-elles d'être disponibles, et sous la même forme ?
- Mesure de l'appartenance au domaine. Dans la mesure du possible, il faut recueillir de l'information auprès de chaque source afin de déterminer avec exactitude l'appartenance au

domaine. Si les éléments d'information recueillis dans des sources administratives ne sont pas modifiables, il est parfois possible d'ajouter des éléments à des échantillons probabilistes qui permettent de déterminer le domaine.

- Redondance. Dans la situation décrite à la section 3.4, où le recensement d'une partie de la population est complété par un échantillon probabiliste, un plan de présélection pourrait être optimal pour S_1. Toutefois, un plan de présélection ne permet pas d'évaluer les différences potentielles dans les mesures des deux échantillons. On peut souhaiter un certain degré de chevauchement entre les sources de données afin d'évaluer les différences entre les estimations de domaines de différentes sources.

  Lorsqu'un modèle d'imputation est élaboré pour y en fonction des relations entre y et $\mathbf{x}$ à partir d'une source de données dont la couverture est incomplète, il y a un risque que ce modèle ne s'applique pas aux autres parties de la population. Il peut être souhaitable de prélever un petit échantillon de la partie non couverte de la population pour évaluer le modèle.

- Quantité relative d'information pour différents domaines. Quand les sources de données comprennent des dossiers administratifs ou de grands échantillons de commodité, il peut y avoir beaucoup plus d'information sur certaines parties de la population que sur d'autres. Il faut alors savoir comment obtenir des renseignements fiables sur les parties manquantes de la population. Quand cette information provient d'un échantillon, il peut y avoir une variation de poids élevée. Levine et Harter (2015) ont étudié la question de la variation de poids dans les enquêtes téléphoniques à double base. On peut réduire une partie de la variation de poids en obtenant d'autres sources de données administratives sur les sous-populations sous-représentées, mais il reste le risque que, comme les organisations recourent de moins en moins aux échantillons probabilistes coûteux, certaines sous-populations soient absentes de toutes les sources.

- Robustesse des hypothèses de plan. Les plans optimaux en théorie se révèlent souvent moins efficaces en pratique. L'examen des performances anticipées du plan en cas de non-respect des hypothèses peut être utile pour modifier un plan théoriquement optimal. Dans certains cas, la combinaison d'informations de plusieurs sources peut donner des estimations moins bonnes que si une seule source est exploitée, ou il peut être décidé que les gains découlant de la combinaison des données ne valent pas la peine de fournir cet effort supplémentaire.

  Waksberg (1998) fait judicieusement remarquer : « Ne traitez pas les procédures statistiques comme des opérations mécaniques; préparez-vous à l'imprévu. » Le fait de posséder un plan suffisamment robuste face aux hypothèses offre une certaine souplesse en cas de problèmes imprévus.

- Information auxiliaire. Plusieurs des méthodes d'intégration des données reposent sur des renseignements auxiliaires pour effectuer des imputations ou prédire l'appartenance à un domaine. Mercer, Lau et Kennedy (2018) ont soutenu que, pour le calage, la richesse de l'information auxiliaire est beaucoup plus importante que la méthode même utilisée pour le

calage, et il en va de même pour d'autres méthodes de combinaison de données. Le fait d'avoir une information auxiliaire riche (au-delà des variables démographiques) permet de meilleurs modèles d'intégration des données et une meilleure évaluation de leurs performances.

Waksberg affirme qu'un statisticien d'enquête doit examiner l'ensemble du problème et non pas seulement le plan optimal pour mesurer une seule variable. Il explique qu'un statisticien réalisant un échantillonnage devrait « penser non seulement aux questions précises qui sont posées, mais aussi aux aspects plus généraux de ces questions, et ainsi se demander si les questions sont logiques et peuvent être résolues, ou si elles devraient être modifiées ou précisées. Voici la façon dont j'ai essayé de faire réfléchir les personnes avec lesquelles je travaille : Voici une question, comment répondez-vous à cette question précise ? Est-ce la bonne question ? Quelles statistiques obtiendrez-vous par une interprétation étroite de la question, et existe-t-il une meilleure façon de procéder ? » (Morganstein et Marker, 2000, page 304).

Dans le présent article, j'ai avancé que les enquêtes à bases multiples peuvent servir de structure organisationnelle pour la conception et l'évaluation des systèmes d'intégration de données. Cela pourrait contribuer à clarifier les forces et les faiblesses de chaque source et, peut-être, à trouver une meilleure façon de procéder.

Remerciements

Je suis reconnaissante au Comité du Prix Waksberg de m'avoir choisie pour cet honneur, à Mike Brick pour nos discussions utiles, et au rédacteur associé et à deux examinateurs dont les suggestions constructives ont amélioré l'article.

Bibliographie

Aidara, C.A.T. (2019). Quasi random resampling designs for multiple frame surveys. *Statistica,* 79, 321-338.

Alleva, G., Arbia, G., Falorsi, P.D., Nardelli, V. et Zuliani, A. (2020). A sampling approach for the estimation of the critical parameters of the SARS-CoV-2 epidemic: An operational design. https://arxiv.org/ftp/arxiv/papers/2004/2004.06068.pdf, dernière consultation le 28 mars 2021.

Arcos, A., Martínez, S., Rueda, M. et Martínez, H. (2017). Distribution function estimates from dual frame context. *Journal of Computational and Applied Mathematics,* 318, 242-252.

Arcos, A., Rueda, M., Trujillo, M. et Molina, D. (2015). Review of estimation methods for landline and cell phone surveys. *Sociological Methods & Research*, 44, 458-485.

Baffour, B., Haynes, M., Western, M., Pennay, D., Misson, S. et Martinez, A. (2016). Weighting strategies for combining data from dual-frame telephone surveys: Emerging evidence from Australia. *Journal of Official Statistics*, 32, 549-578.

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

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 Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, 11-22.

Biemer, P.P. (1984). Methodology for optimal dual frame sample design. Bureau of the Census SRD Research Report CENSUS/SRD/RR-84/07.

Brick, J.M., Flores-Cervantes, I.F., Lee, S. et Norman, G. (2011). Erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double. *Techniques d'enquête,* 37, 1, 1-16. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11443-fra.pdf.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. et Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Brick, J.M., Shapiro, G., Flores-Cervantes, I., Ferraro, D. et Strickler, T. (1999). *1997 NSAF Snapshot Survey Weights*. Washington, DC: Urban Institute.

Burke, J., Mohadjer, L., Green, J., Waksberg, J., Kirsch, I.S. et Kolstad, A. (1994). Composite estimation in national and state surveys. Dans *Proceedings of the Survey Research Methods Section*, 873-878. Alexandrie, Virginie: American Statistical Association.

Chauvet, G. (2016). Variance estimation for the 2006 French housing survey. *Mathematical Population Studies*, 23, 147-163.

Chauvet, G., et de Marsac, G.T. (2014). Méthodes d'estimation sur bases de sondage multiples dans le cadre de plans de sondage à deux degrés. *Techniques d'enquête,* 40, 2, 367-378. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014002/article/14090-fra.pdf.

Chen, S., Stubblefield, A. et Stoner, J.A. (2021). Oversampling of minority populations through dual-frame surveys. *Journal of Survey Statistics and Methodology*, 9, 626-649.

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

Chipperfield, J., Chessman, J. et Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. *Australian & New Zealand Journal of Statistics*, 54, 223-238.

Choudhry, G.H. (1989). Cost-variable optimization of dual frame design for estimating proportions. Dans *Proceedings of the Survey Research Methods Section*, 566-571. Alexandrie, Virginie: American Statistical Association.

Chu, A., Brick, J.M. et Kalton, G. (1999). Weights for combining surveys across time or space. *Bulletin of the International Statistical Institute*, 2, 103-104.

Citro, C.F. (2014). Des modes multiples pour les enquêtes à des sources de données multiples pour les estimations. *Techniques d'enquête,* 40, 2, 151-181. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014002/article/14128-fra.pdf.

Cohen, S.B., DiGaetano, R. et Waksberg, J. (1988). Sample design of the NMES survey of American Indians and Alaska Natives. Dans *Proceedings of the Survey Research Methods Section*, 740-745. Alexandrie, Virginie: American Statistical Association.

Cunningham, P., Shapiro, G. et Brick, J.M. (1999). *1997 NSAF In-Person Survey Methods*. Washington, DC: Urban Institute.

Dever, J.A. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. Dans *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*. https://nces.ed.gov/FCSM/pdf/A4_Dever_2018FCSM.pdf, dernière consultation le 7 juillet 2021.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : Les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête,* 32, 2, 185-196. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9551-fra.pdf.

DiGaetano, R., Judkins, D. et Waksberg, J. (1995). Oversampling minority school children. Dans *Proceedings of the Survey Research Methods Section*, 503-508. Alexandrie, Virginie: American Statistical Association.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Ferraz, C., et Vogel, F. (2015). Multiple frame sampling. Dans *Handbook on Master Sampling Frames for Agricultural Statistics: Frame Development, Sample Design and Estimation*, 89-106. Rome: Food and Agriculture Organization of the United Nations.

Fuller, W.A., et Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. Dans *Proceedings of the Social Statistics Section*, 245-249. Alexandrie, Virginie: American Statistical Association.

Groves, R.M., et Lepkowski, J.M. (1985). Dual frame, mixed mode survey designs. *Journal of Official Statistics*, 1, 263-286.

Groves, R.M., et Lepkowski, J.M. (1986). An experimental implementation of a dual frame telephone sample design. Dans *Proceedings of the Survey Research Methods Section*, 340-345. Alexandrie, Virginie: American Statistical Association.

Groves, R.M., et Wissoker, D. (1999). *1997 NSAF Early Nonresponse Studies*. Washington, DC: Urban Institute.

Hartley, H.O. (1962). Multiple frame surveys. Dans *Proceedings of the Social Statistics Section*, 203-206. Alexandrie, Virginie: American Statistical Association.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Series C,* 36, 99-118.

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

Hendricks, S., Igra, A. et Waksberg, J. (1980). Ethnic stratification in the California Hypertension Survey. Dans *Proceedings of the Survey Research Methods Section*, 680-685. Alexandrie, Virginie: American Statistical Association.

Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society. Series A (General),* 149, 65-82.

Kim, J.K., et Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika,* 99, 85-100.

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

Kott, P.S., et Vogel, F.A. (1995). Multiple-frame business surveys. Dans *Business Survey Methods*, (Éds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott), 185-203. New York: John Wiley & Sons, Inc.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.

Lavallée, P., et Rivest, L.-P. (2012). Capture-recapture sampling and indirect sampling. *Journal of Official Statistics,* 28, 1-27.

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

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.

Lepkowski, J.M., et Groves, R.M. (1986). A mean squared error model for dual frame, mixed mode survey design. *Journal of the American Statistical Association,* 81, 930-937.

Levine, B., et Harter, R. (2015). Optimal allocation of cell-phone and landline respondents in dual-frame surveys. *Public Opinion Quarterly*, 79, 91-104.

Lin, D., Liu, Z. et Stokes, L. (2019). Méthode de correction de l'erreur d'appartenance à une base dans les estimateurs à double base de sondage. *Techniques d'enquête,* 45, 3, 581-605. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019003/article/00008-fra.pdf.

Lohr, S.L. (2007). Recent developments in multiple frame surveys. Dans *Proceedings of the Survey Research Methods Section*, 3257-3264. Alexandrie, Virginie: American Statistical Association.

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.

Lohr, S.L. (2014). When should a multiple frame survey be used? *The Survey Statistician,* 69 (janvier), 17-21.

Lohr, S.L. (2022). *Sampling: Design and Analysis, Third Edition*. Boca Raton, FL: CRC Press.

Lohr, S.L., et Brick, J.M. (2014). Allocation for dual frame telephone surveys with nonresponse. *Journal of Survey Statistics and Methodology*, 2, 388-409.

Lohr, S.L., et Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

Lohr, S.L., et Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association,* 101, 1019-1030.

Lu, B., Peng, J. et Sahr, T. (2013). Estimation bias of different design and analytical strategies in dual-frame telephone surveys: An empirical evaluation. *Journal of Statistical Computation and Simulation,* 83, 2352-2368.

Lu, B., Sahr, T., Iachan, R., Denker, M., Duffy, T. et Weston, D. (2013). Design and analysis of dual-frame telephone surveys for health policy research. *World Medical & Health Policy*, 5, 217-232.

Lu, Y. (2014a). Tests du khi-carré dans les enquêtes à base de sondage double. *Techniques d'enquête,* 40, 2, 353-366. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014002/article/14096-fra.pdf.

Lu, Y. (2014b). Regression coefficient estimation in dual frame surveys. *Communications in Statistics – Simulation and Computation*, 43, 1675-1684.

Lu, Y., et Lohr, S. (2010). L'estimation des flux bruts dans les enquêtes à base de sondage double. *Techniques d'enquête,* 36, 1, 13-24. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2010001/article/11248-fra.pdf.

Lu, Y., Fu, Y. et Zhang, G. (2021). Nonparametric regression estimators in dual frame surveys. *Communications in Statistics – Simulation and Computation,* 50, 854-864.

Marks, E., et Waksberg, J. (1966). Evaluation of coverage in the 1960 Census of Population through case-by-case checking. Dans *Proceedings of the Social Statistics Section*, 62-70. Alexandrie, Virginie: American Statistical Association.

Mecatti, F. (2007). Un estimateur à base de sondage unique fondé sur la multiplicité pour les sondages à bases multiples. *Techniques d'enquête,* 33, 2, 171-178. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2007002/article/10492-fra.pdf.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12, 685-726.

Mercer, A., Lau, A. et Kennedy, C. (2018). *For Weighting Online Opt-In Samples, What Matters Most?* Washington, DC: Pew Research.

Metcalf, P., et Scott, A. (2009). Using multiple frames in health surveys. *Statistics in Medicine*, 28, 1512-1523.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *Revue Internationale de Statistique*, 55, 191-202.

Montanari, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête,* 24, 1, 71-79. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1998001/article/3911-fra.pdf.

Morganstein, D., et Marker, D. (2000). A conversation with Joseph Waksberg. *Statistical Science*, 15, 299-312.

National Academies of Sciences, Engineering, and Medicine (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: National Academies Press.

National Academies of Sciences, Engineering, and Medicine (2018). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: National Academies Press.

O'Muircheartaigh, C., et Pedlow, S. (2002). Combining samples vs. cumulating cases: A comparison of two weighting strategies in NLSY97. Dans *Proceedings of the Survey Research Methods Section*, 2557-2562. Alexandrie, Virginie: American Statistical Association.

Ranalli, M.G., Arcos, A., Rueda, M.d.M. et Teodoro, A. (2016). Calibration estimation in dual-frame surveys. *Statistical Methods & Applications*, 25, 321-349.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

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

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation, 2nd Ed*. Hoboken, NJ: Wiley.

Rivers, D. (2007). Sampling for web surveys. Document présenté à la Joint Statistical Meetings.

Rueda, M.d.M., Arcos, A., Molina, D. et Ranalli, M.G. (2018). Estimation techniques for ordinal data in multiple frame surveys with complex sampling designs. *Revue Internationale de Statistique*, 86, 51-67.

Saegusa, T. (2019). Large sample theory for merged data from multiple sources. *The Annals of Statistics*, 47, 1585-1615.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Hoboken, NJ: Wiley.

Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Thompson, M.E. (2019). Combining data from new and traditional sources in population surveys. *Revue Internationale de Statistique*, 87, S79-S89.

Traugott, M.W., Groves, R.M. et Lepkowski, J.M. (1987). Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion Quarterly*, 51, 522-539.

Urban Institute and Child Trends (2007). *National Survey of America's Families (NSAF), 1997*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributeur].

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

Waksberg, J. (1986). Discussion of papers on new approaches in telephone sample design. Dans *Proceedings of the Survey Research Methods Section*, 367-369. Alexandrie, Virginie: American Statistical Association.

Waksberg, J. (1995). Distribution of poverty in census block groups (BGs) and implications for sample design. Dans *Proceedings of the Survey Research Methods Section*, 497-502. Alexandrie, Virginie: American Statistical Association.

Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the US Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 119-135.

Waksberg, J., et Pritzker, L. (1969). Changes in census methods. *Journal of the American Statistical Association,* 64, 1141-1149.

Waksberg, J., Judkins, D. et Massey, J.T. (1997b). Suréchantillonnage géographique dans les enquêtes démographiques aux États-Unis. *Techniques d'enquête,* 23, 1, 69-80. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997001/article/3107-fra.pdf.

Waksberg, J., Brick, J.M., Shapiro, G., Flores-Cervantes, I. et Bell, B. (1997a). Dual-frame RDD and area sample with particular focus on low-income population. Dans *Proceedings of the Survey Research Methods Section*, 713-718. Alexandrie, Virginie: American Statistical Association.

Waksberg, J., Brick, J.M., Shapiro, G., Flores-Cervantes, I., Bell, B. et Ferraro, D. (1998). Corrections visant à tenir compte de la non-réponse et de la sous-couverture dans une enquête à base de sondage duale. *Recueil : Symposium 97, Nouvelles orientations pour les enquêtes et les recensements*, 219-226. Ottawa: Statistique Canada.

Wolter, K.M., Ganesh, N., Copeland, K.R., Singleton, J.A. et Khare, M. (2019). Estimation tools for reducing the impact of sampling and nonresponse errors in dual-frame RDD telephone surveys. *Statistics in Medicine*, 38, 4718-4732.

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.

Zhang, L.-C. (2019). Log-linear models of erroneous list data. Dans *Analysis of Integrated Data* (Éds., L.-C. Zhang et R.L. Chambers), 197-218. Boca Raton, FL: CRC Press.

Zhang, L.-C., et Chambers, R.L. (Eds.) (2019). *Analysis of Integrated Data*. Boca Raton, FL: CRC Press.