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

Techniques d’enquête

Tendances et orientations de la théorie et de la méthodologie des enquêtes par sondage

par J.N.K. Rao et Sharon L. Lohr

Date de diffusion : le 30 juin 2025

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Techniques d'enquête, juin 2025
Vol. 51, nº 1, pp. 87-122
Statistique Canada, nº 12-001-X au catalogue

Tendances et orientations de la théorie et de la méthodologie des enquêtes par sondage

J.N.K. Rao et Sharon L. Lohr[1]

Résumé

Rao (1999) a résumé les tendances de la théorie et de la méthodologie des enquêtes par sondage au tournant du siècle. Nous présentons un examen à jour de certaines tendances actuelles relatives aux plans d'enquête et aux méthodes d'estimation pour le 50e anniversaire de *Techniques d'enquête*. On observe, parmi les récentes innovations dans les plans d'enquête, la recherche sur l'anticipation des erreurs non dues à l'échantillonnage à l'étape de la conception et l'élaboration de plans d'échantillonnage adaptatif et équilibré pour profiter des renseignements détaillés de la base de sondage ou des données recueillies pendant le processus de l'enquête. Les méthodes non paramétriques et les méthodes d'apprentissage automatique sont de plus en plus utilisées pour la vérification des données de même que pour l'estimation assistée par un modèle et les ajustements pour la non-réponse. Des modèles d'estimation sur petits domaines ont été élargis pour y intégrer des renseignements spatiaux et des renseignements tirés de séries chronologiques, augmenter la souplesse et la robustesse des modèles de couplage et de variance, procéder à un étalonnage selon des estimateurs directs sur grands domaines et (pour les modèles au niveau de l'unité) tenir compte des plans d'échantillonnage informatif. La disponibilité croissante de vastes ensembles de données administratives, de données de capteurs, de données satellitaires et d'échantillons de commodité a stimulé la recherche sur la façon d'utiliser ces sources – à elles seules et lorsqu'elles sont intégrées dans des échantillons probabilistes. Nous concluons en abordant certaines frontières de la recherche sur les enquêtes.

Mots-clés : Apprentissage automatique; conception de l'enquête; échantillons non probabilistes; estimation sur petits domaines; intégration des données; plan d'enquête; qualité de l'enquête.

1. Introduction

Nous remercions le rédacteur en chef de nous avoir invités à participer au présent numéro, qui célèbre l'apport de *Techniques d'enquête*. Dès son premier numéro, paru en juin 1975, la revue a fourni une tribune pour la recherche « sur toutes les phases de l'élaboration des méthodes d'enquête » (Platek, 1999, page 123). Sa grande accessibilité et son accent porté sur les problèmes pratiques ont fait de la revue une ressource essentielle pour les spécialistes de la recherche sur les enquêtes du monde entier.

Dans le présent article, nous résumons quelques-unes des tendances de la recherche sur les enquêtes au cours des 50 ans d'existence de *Techniques d'enquête*. Nous nous penchons ensuite sur certaines orientations futures possibles. Le présent article est la mise à jour de l'article « Some current trends in sample survey theory and methods » (Rao, 1999) [en anglais seulement], qui a permis d'examiner l'état des enquêtes par sondage au tournant du siècle, et est axé sur les aspects de la discipline qui ont émergé ou changé depuis. Rao (2005), Brick (2011) et Rao et Fuller (2017) ont présenté d'autres examens de récentes recherches sur l'échantillonnage.

Ce genre d'examen est nécessairement limité à quelques sujets et ne peut mentionner tous les progrès importants survenus au cours du dernier quart de siècle. Nous invitons le lecteur à consulter d'autres articles

1. J.N.K. Rao, professeur de recherche distingué, Carleton University, Ottawa (Ontario), Canada, K1S 5B6. Courriel : jrao@math.carleton.ca. Sharon Lohr, professeure émérite, Arizona State University. Courriel : sharon.lohr@asu.edu.

pour prendre connaissance des examens menés sur d'importantes avancées en matière de protection des renseignements personnels, d'imputation et d'erreur de mesure. Bien que l'accent porte toujours sur l'amélioration des estimations du total de population $Y = \sum_{i=1}^{N} y_i$ d'une variable à l'étude y pour une population de taille N, l'accent est mis de plus en plus sur les nouvelles utilisations de l'information auxiliaire $\mathbf{x}$ pour le plan de sondage, l'estimation et l'ajustement pour la non-réponse. L'interaction entre les avancées technologiques et la théorie statistique ouvre de nouvelles avenues pour la recherche sur les enquêtes.

Les étapes principales, pour une enquête par sondage, déterminées par Rao et Fuller (2017, page 1) – la collecte, le traitement, l'estimation et l'analyse des données – ne changent pas, bien que les procédures et l'accent mis dans ces étapes aient changé au fil des ans. La section 2 aborde de récentes innovations ayant trait au plan d'enquête, à la collecte des données et au traitement des données d'enquête, en insistant plus particulièrement sur les nouveautés qui donnent lieu à des améliorations de la qualité globale d'une enquête. La section 3 décrit les progrès réalisés en matière d'estimation au moyen d'une pondération, d'une modélisation et d'une estimation de la variance améliorées. Elle porte également sur les nouveautés qui visent à traiter la non-réponse. La section 4 traite de certaines innovations récentes dans l'estimation sur petits domaines, la section 5 décrit les travaux sur l'intégration des données provenant de sources multiples et la section 6 résume les tendances et dégage certaines possibilités de recherche future.

2. Plan d'enquête et collecte des données

Rao (1999) a insisté sur l'importance du plan d'enquête complet, qui vise à minimiser l'erreur quadratique moyenne (EQM) des estimations, en tenant compte non seulement de l'erreur d'échantillonnage, mais aussi des erreurs découlant de la collecte des données, du traitement des données, de l'erreur de mesure, du sous-dénombrement et de la non-réponse.

2.1 Collecte et traitement des données

À la fin du 20e siècle, la plupart des enquêtes étaient menées en personne (si une couverture complète était souhaitée) ou par téléphone (comme option à moindre coût). Rao (1999, page 4) a remarqué que les enquêtes par téléphone étaient devenues populaires au cours des récentes années. Bien que des enquêtes par téléphone soient encore menées dans certaines régions, bon nombre d'organismes d'enquête se sont tournés vers d'autres méthodes alors que les taux de réponse par téléphone ont diminué rapidement (Kennedy et Hartig, 2019, ont fait état de taux de réponse par téléphone de 6 % en 2018).

Deux méthodes d'enquête en particulier sont devenues plus courantes en Amérique du Nord depuis l'an 2000. Les listes d'adresses mises à jour au moyen des renseignements des services postaux des États-Unis fournissent maintenant une couverture presque complète des ménages américains (Harter, Battaglia, Buskirk, Dillman, English, Fahimi, Frankel, Kennel, McMichael, McPhee, Montaquila, Yancey et Zukerberg, 2016). La mise au point de cette base de sondage fondée sur l'adresse a donné lieu à un nombre accru d'enquêtes-ménages dans lesquelles le premier contact se fait par la poste. Le suivi des cas de non-réponse

peut alors se faire à l'aide de méthodes affichant des taux de réponse plus élevés, telles que les visites en personne. Dans le Recensement américain de 2020 et le Recensement canadien de 2021, on a fait le premier contact auprès des ménages à l'aide d'une lettre, offert des méthodes multiples pour répondre au questionnaire (et l'on invitait les personnes à répondre sur Internet) et assuré un suivi en personne auprès des non-répondants. Les bases de sondage améliorées fondées sur l'adresse pour les recensements ont une couverture presque complète, et l'utilisation de méthodes de collecte des données moins coûteuses pour les premiers contacts permet de réduire les coûts.

De nombreuses enquêtes dont la collecte des données est réalisée par des organisations privées s'appuient maintenant sur des panels en ligne, dans lesquels les échantillons pour les enquêtes individuelles sont prélevés à partir d'un groupe de personnes qui ont été recrutées en tant que participants potentiels à l'enquête. Les membres du groupe peuvent être recrutés à l'aide de méthodologies probabilistes (par exemple, par l'intermédiaire d'une enquête par la poste menée en fonction d'adresses choisies aléatoirement) ou peuvent former un échantillon non probabiliste (par exemple, si les membres du groupe sont recrutés par l'intermédiaire de publicités en ligne). En 2019, dans plus de 80 % des sondages d'opinion aux États-Unis, on a utilisé des panels non probabilistes en ligne (Kennedy, Hatley, Lau, Mercer, Keeter, Ferno et Asare-Marfo, 2021).

Ces innovations dans la collecte des données ont mené à un grand nombre de recherches empiriques pour déterminer l'incidence de la méthode, ou celle de l'utilisation de plusieurs méthodes, sur la qualité des enquêtes (Couper, 2017; De Leeuw, 2018; Olson, Smyth, Horwitz, Keeter, Lesser, Marken, Mathiowetz, McCarthy, O'Brien, Opsomer, Steiger, Sterrett, Su, Suzer-Gurtekin, Turakhia et Wagner, 2021). Par exemple, Kennedy, Mercer et Lau (2024) ont comparé les caractéristiques de population estimées à partir des répondants de six échantillons en ligne avec des données repères issues d'enquêtes de grande qualité menées par le gouvernement américain. Pour trois des échantillons en ligne, le recrutement était volontaire; pour les trois autres, les membres du panel étaient recrutés à partir d'un échantillon probabiliste fondé sur l'adresse postale et ils devaient ensuite participer à des enquêtes en ligne. Bien que les taux de réponse pour les échantillons probabilistes aient été de moins de 10 %, l'erreur absolue moyenne après le calage était deux fois plus élevée pour les échantillons volontaires que pour les échantillons probabilistes; le nombre d'erreurs était particulièrement élevé pour les personnes de 18 à 29 ans, les personnes qui ont déclaré être hispaniques et les questions liées à la réception de prestations du gouvernement. Kennedy et coll. (2024) ont laissé supposer que certains participants aux enquêtes volontaires pourraient avoir tendance à dire « oui » quelle que soit la question posée et ont recommandé que plus de recherches soient menées sur l'exactitude des renseignements démographiques utilisés pour la pondération que les répondants aux enquêtes volontaires fournissent.

Un autre domaine émergent est le potentiel des méthodes d'intelligence artificielle pour la collecte et le traitement des données (Yung, Karkimaa, Scannapieco, Barcarolli, Zardetto, Sanchez, Braaksma, Buelens et Burger, 2018; Commission économique des Nations Unies pour l'Europe, 2021). Beck, Dumpert et Feuerhake (2018) ont présenté un compte rendu de 136 projets menés dans les organismes nationaux de

statistique qui faisaient appel à des méthodes d'apprentissage automatique. Les méthodes étaient la plupart du temps utilisées pour la classification, le repérage des enregistrements en double ou des valeurs aberrantes, et l'imputation des éléments manquants.

Nous donnons trois exemples pour illustrer les types de problèmes de traitement des données qui peuvent être réglés à l'aide des méthodes d'intelligence artificielle. Hibben, Smith, Hoppe, Ryan, Rogers, Scanlon et Miller (2022) ont entraîné un modèle de traitement du langage naturel pour interpréter les réponses à des questions ouvertes et différencier les réponses valides des non-réponses et des réponses inintelligibles.

Roberson (2021) a abordé un problème de fardeau de réponse dans le Recensement économique des États-Unis de 2017, qui comportait une section du questionnaire longue et coûteuse en temps sur les biens et les services fournis. Roberson a tenté de voir si les entreprises pouvaient aussi fournir les codes universels des produits, qui étaient déjà dans leurs bases de données, au Bureau du recensement. Une méthode d'apprentissage automatique supervisée qui classait les produits en fonction des codes universels des produits a permis d'obtenir une exactitude de plus de 90 %.

Finalement, il existe une vaste littérature croissante sur les méthodes d'apprentissage automatique pour la vérification et l'imputation des données. À titre d'exemple, McClellan, Mitchell, Anderson et Zuvekas (2023) ont découvert que la forêt aléatoire et les combinaisons de méthodes d'apprentissage automatique avaient une meilleure exactitude que la régression linéaire pour imputer les éléments de dépense manquants dans le Medical Expenditure Panel Survey aux États-Unis. Lin et Tsai (2020) et Dagdoug, Goga et Haziza (2023a) ont examiné et comparé des méthodes d'apprentissage automatique pour imputer des éléments manquants des répondants aux enquêtes.

Bien entendu, les méthodes d'apprentissage automatique ne peuvent pas régler tous les problèmes, et il est important de tester leur exactitude avant leur mise en œuvre. Par exemple, Ikudo, Lane, Staudt et Weinberg (2019) ont examiné la possibilité d'utiliser des modèles de forêt aléatoire pour classer des professions à partir de dossiers des ressources humaines d'établissements universitaires, mais ont découvert que les titres de poste dans les données des ressources humaines ne sont pas cohérents d'un établissement à l'autre et qu'il n'y a pas assez de renseignements pour permettre un classement exact. Un autre enjeu important est l'équité en matière de données, qui fait en sorte que les algorithmes ne désavantagent pas systématiquement certains sous-groupes de la population. Ce sujet sera abordé plus à fond à la section 6.

2.2 Plan d'enquête

Les articles sur les plans d'enquête au cours des 25 premières années de *Techniques d'enquête* étaient principalement répartis en trois types : la description de la conception d'une enquête en particulier, la mise en œuvre d'un échantillonnage avec probabilité proportionnelle à la taille et la répartition d'observations dans des plans d'échantillonnage stratifiées, à deux phases et à plusieurs degrés. La répartition continue d'être un thème pour les articles sur les plans d'enquête, mais de plus en plus de recherches menées au cours des 25 dernières années ont porté essentiellement sur des sujets tels que l'anticipation des erreurs non dues

à l'échantillonnage à l'étape de la conception de l'enquête, l'échantillonnage équilibré et l'échantillonnage adaptatif.

Anticiper les erreurs non dues à l'échantillonnage à l'étape de la conception de l'enquête. Les organismes de statistique à l'échelle mondiale ont fait de grands pas pour améliorer la couverture des bases de sondage, étudier l'erreur de mesure et réduire les erreurs de codage et de traitement des données. L'utilisation accrue d'expérimentations conçues à une fin précise permet aux spécialistes de la recherche sur les enquêtes d'étudier les effets des méthodes et des questions, des méthodes de sélection au sein des ménages, des techniques de minimisation de la non-réponse, de la variabilité entre les intervieweurs et de nombreux autres aspects de la qualité des enquêtes (Lavrakas, Traugott, Kennedy, Holbrook, de Leeuw et West, 2019). Les nouvelles technologies et sources de données peuvent être utilisées pour déterminer les unités qui pourraient être manquantes dans les bases de sondage, ou pour déterminer les sous-ensembles aux fins de suréchantillonnage (Datta, Ugarte et Resnick, 2020). Par exemple, Hyman, Sartore et Young (2022) ont utilisé le moissonnage Web pour trouver les fermes alimentaires locales qui pourraient ne pas figurer dans une liste des fermes, et Daas et van der Doef (2020) ont employé un modèle textuel pour dresser la liste des entreprises innovatrices en étudiant le texte sur leurs sites Web.

Les taux de réponse dans de nombreux pays ont toutefois chuté depuis l'an 2000, et bien que les modèles de non-réponse puissent réduire le biais, rien ne garantit qu'il est supprimé (voir la section 3.2). Mercer, Lau et Kennedy (2018) ont découvert que la qualité de l'information auxiliaire disponible avait plus d'incidence sur la réduction du biais que le modèle de non-réponse particulier utilisé. Par conséquent, la mise au point de bases de sondage ayant une riche information auxiliaire sur les unités de population permet d'avoir des modèles de non-réponse améliorés ainsi que des plans d'échantillonnage plus efficaces. Pour les enquêtes sur des personnes, ces bases pourraient provenir de registres de la population ou d'initiatives telles que le projet de bases de sondage du Bureau du recensement des États-Unis, dans le cadre duquel on fusionne les données de bases géospatiales et démographiques et de bases d'entreprises et d'emplois (National Academies of Sciences, Engineering, and Medicine, 2023, chapitre 4). La technologie de télédétection fournit des renseignements détaillés pour les bases de sondage dans les études en agriculture, en foresterie et sur l'environnement. Le passage des enquêtes par téléphone à l'échantillonnage fondé sur l'adresse augmente aussi l'information auxiliaire disponible pour le plan d'échantillonnage et les ajustements pour la non-réponse parce que les caractéristiques du quartier sont connues pour chaque unité de logement échantillonnée.

Il y a eu une utilisation accrue de modèles d'apprentissage automatique pour anticiper les erreurs non dues à l'échantillonnage lors de la conception d'enquêtes (Buskirk et Kirchner, 2020). Par exemple, Oberski et DeCastellarnau (2021) ont utilisé la forêt aléatoire pour prédire la variance de l'erreur de mesure des questions d'enquête à partir de caractéristiques clés codées pour chaque question. Kern, Weiß et Kolb (2021) ont utilisé des modèles d'apprentissage automatique pour prédire la non-réponse dans les futures vagues d'une enquête par panel, ce qui a permis de cibler à l'avance les non-répondants probables.

Échantillonnage équilibré. Un échantillon aléatoire simple sera certainement représentatif sur le plan de l'espérance, mais un échantillon particulier, surtout s'il est petit, peut produire des statistiques pour certaines caractéristiques qui sont loin des valeurs de population. L'échantillonnage équilibré tire profit de l'information auxiliaire $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})^T$ présente dans la base de sondage pour chaque unité de population i et restreint l'ensemble d'échantillons qui peuvent être choisis à ceux pour lesquels les totaux de population estimés pour ces variables, $\hat{\mathbf{X}}$, équivalent aux totaux de population connus $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i$. L'échantillonnage stratifié est un cas spécial dans lequel $\mathbf{x}$ est le vecteur des indicateurs de strate. Il pourrait être impossible d'obtenir un plan de sondage qui permet de procéder à une stratification selon le classement recoupé du sexe, de l'âge, de la race, de la région, du niveau de scolarité, du revenu et d'autres variables, mais l'échantillonnage équilibré permet de sélectionner un échantillon qui correspond aux moments ou aux proportions de catégorie de chaque variable séparément.

L'échantillonnage équilibré offre beaucoup plus de souplesse que la stratification lors de la sélection d'un échantillon à partir d'une base de sondage contenant de l'information auxiliaire riche; c'est la raison pour laquelle il est de plus en plus employé dans les pays qui ont des registres de la population. On peut le voir comme une façon de caler à l'avance l'enquête en fonction des variables d'équilibrage; l'estimation par régression généralisée a la même relation avec l'échantillonnage équilibré que la stratification *a posteriori* avec l'échantillonnage stratifié.

Un échantillon équilibré peut être choisi délibérément (comme dans Royall et Herson, 1973) ou de façon probabiliste (comme dans l'échantillonnage aléatoire stratifié). Un échantillon équilibré choisi à l'aide de méthodes d'échantillonnage probabiliste combine les avantages d'efficacité de l'équilibrage avec la robustesse du modèle que l'échantillonnage probabiliste permet. Fuller (2009) a proposé une procédure d'échantillonnage probabiliste réjectif pour garantir un équilibrage approximatif en ce sens qu'un échantillon probabiliste sélectionné est rejeté à moins que la moyenne pondérée d'un vecteur auxiliaire $\mathbf{x}$ se situe à une distance précise de la moyenne de population correspondante. L'utilisation de l'échantillonnage équilibré probabiliste a grandement augmenté depuis la mise au point de la méthode du cube pour calculer des plans de sondage approximativement équilibrés ayant des probabilités d'inclusion fixes (Deville et Tillé, 2004; Leuenberger, Eustache, Jauslin et Tillé, 2022) et la mise au point d'algorithmes pour calculer les variances des estimations à partir d'échantillons équilibrés (Deville et Tillé, 2005; Breidt et Chauvet, 2011; Kim et Wu, 2013; Tillé, 2020).

Falorsi et Righi (2008, 2015) ont employé une méthode d'échantillonnage équilibré lorsque des estimations sur petits domaines sont souhaitées pour un certain nombre de différents types de domaines et qu'il est impossible d'utiliser le classement recoupé des différentes partitions. La répartition est conçue pour garantir que les estimations directes (ou indirectes, au besoin) sur petits domaines permettent d'obtenir une erreur quadratique moyenne prédéterminée pour chaque domaine d'intérêt. Chauvet, Deville et Haziza (2011) ont appliqué les principes de l'échantillonnage équilibré à l'imputation en échantillonnant de manière

aléatoire les donneurs pour l'imputation par enregistrement donneur tout en respectant un ensemble de contraintes.

Les plans d'échantillonnage spatialement équilibrés sont souvent souhaitables pour les enquêtes sur l'écologie ou la foresterie. Ils forcent l'échantillon à être réparti sur toute la région à l'étude, ce qui réduit les corrélations spatiales parmi les unités dans l'échantillon. Tillé (2020, chapitre 8) et Robertson et Price (2024) ont résumé les méthodes pour sélectionner des échantillons probabilistes spatialement équilibrés, ainsi que les mesures pour évaluer l'équilibre spatial. Grafström et Ringvall (2013) ont dérivé des plans de sondage pour atteindre l'équilibre spatial ainsi que l'équilibre sur les variables auxiliaires obtenues au moyen de la télédétection.

Plans d'enquête adaptatifs. Dans l'échantillonnage probabiliste classique, le plan d'échantillonnage est fixe avant le début de la collecte des données – les strates sont déterminées, les unités primaires d'échantillonnage (UPE) sont sélectionnées selon les probabilités préalablement précisées, et les données sont recueillies selon le protocole prédéterminé. Dans les plans d'enquête adaptatifs, les renseignements provenant des unités mesurées au début du processus sont utilisés pour modifier (adapter) les efforts de collecte des données ultérieurs.

Certains plans d'enquête adaptatifs donnent lieu à des échantillons probabilistes – par exemple, dans l'échantillonnage en grappes adaptatif, on commence par un petit échantillon probabiliste d'UPE et l'on modifie ensuite les probabilités de sélection pour les UPE subséquentes en fonction des données déjà recueillies (voir Thompson, 2017, pour obtenir un examen). Toutefois, quand l'échantillon initial n'est pas choisi de manière probabiliste, l'échantillon final provenant d'une procédure adaptative n'est pas un échantillon probabiliste. Par exemple, l'échantillonnage déterminé selon les répondants est devenu populaire au cours des récentes années en tant que méthode de collecte des données auprès d'une population rare ou difficile à déterminer comme les immigrants récents ou les groupes minoritaires (Heckathorn et Cameron, 2017; Gile, Beaudry, Handcock et Ott, 2018). On demande aux membres d'un échantillon de commodité tiré d'une population difficile à déterminer de nommer d'autres membres de la population pour les inclure dans l'échantillon; on demande par la suite à ces derniers de proposer d'autres membres de la population et ainsi de suite jusqu'à ce que la taille souhaitée de l'échantillon soit atteinte. Comme l'échantillon de départ est formé de membres de la population faciles à recruter, un échantillon déterminé selon les répondants est un échantillon de commodité, et de solides hypothèses de modélisation sont requises pour tirer des inférences sur la population.

Une autre forme de plan de sondage adaptatif tente d'améliorer les mesures de la qualité en modifiant les protocoles de collecte des données pour les unités qui n'ont pas encore répondu à l'enquête. On pourrait accorder la priorité à certains cas pour augmenter le taux de réponse global, cibler des sous-populations à faible propension, réduire la variance des ajustements de pondération pour la non-réponse ou contrôler les coûts. Coffey, Damineni, Eltinge, Mathur, Varela et Zotti (2024) ont donné des exemples dans lesquels l'information auxiliaire tirée de sources de données externes pourrait éclairer les décisions en matière de collecte de données.

Groves et Heeringa (2006) ont inventé l'expression « responsive survey design » (plan d'enquête réactif) pour une procédure qui permet de diviser la période de collecte des données en étapes et d'utiliser l'information accumulée des étapes précédentes (par exemple, le nombre d'appels ou les observations notées par les intervieweurs) pour modifier les protocoles de collecte des données pour les étapes ultérieures. L'idée du plan de collecte réactif n'est pas nouvelle – la méthode à deux étapes pour le suivi de la non-réponse de Hansen et Hurwitz (1946) est utilisée pour de nombreuses enquêtes, y compris l'American Community Survey – mais au cours des dernières années, on a assisté à une augmentation des recherches dans lesquelles le recrutement de l'enquête est adapté aux non-répondants individuels (Schouten, Peytchev et Wagner, 2018; Zhang et Wagner, 2024). Tourangeau, Brick, Lohr et Li (2017) ont indiqué que les hausses du taux de réponse, les réductions du biais et les réductions de la variabilité des propensions de réponse estimées étaient modestes dans les applications de plans d'enquête réactifs qu'ils avaient examinées. Cependant, la plupart de ces applications avaient des bases de sondage comptant peu de renseignements, et il est possible que les méthodes de plan d'enquête réactif fonctionnent mieux au moyen de renseignements riches sur la base de sondage. Kaputa, Thompson et Beck (2019) ont insisté sur l'importance d'évaluer les stratégies de plan réactif à l'aide d'expérimentations conçues à cette fin.

3. Estimation

3.1 Pondération et utilisation de l'information auxiliaire

La plupart des enquêtes s'appuient sur la pondération et le calage pour produire des estimations des caractéristiques de population. La pondération tient compte de probabilités de sélection inégales dans le plan d'échantillonnage et elle est aussi couramment utilisée pour les probabilités d'autosélection estimées à partir de la non-réponse ou dans des échantillons non probabilistes. Un avantage de la pondération est que les estimations sont cohérentes sur le plan interne les unes avec les autres – par exemple, les totaux de population estimés pour subdiviser la somme des sous-ensembles en fonction du total estimé pour la population dans son ensemble – car toutes les estimations sont calculées en utilisant le même ensemble de poids. Le calage permet d'ajuster les poids d'échantillonnage de sorte que les totaux de population estimés pour les variables auxiliaires correspondent aux totaux de population connus pour ces variables. Brick (2013) et Haziza et Beaumont (2017) ont examiné les travaux de recherche sur les méthodes de pondération et Särndal (2007) a examiné l'approche par calage.

Pour de nombreuses enquêtes, les totaux de contrôle utilisés pour le calage proviennent d'autres enquêtes et ont eux-mêmes des erreurs d'échantillonnage et des erreurs non dues à l'échantillonnage. Dever et Valliant (2016) ont dérivé des propriétés statistiques d'estimateurs calés selon les quantités dérivées d'une autre enquête, et avec Opsomer et Erciulescu (2021), ils ont proposé des estimateurs de variance qui saisissent la variabilité des deux sources. Haziza et Lesage (2016) ont conclu qu'une approche à deux étapes à l'égard des ajustements de pondération – c'est-à-dire que l'on ajuste d'abord les poids selon les propensions de

réponse estimées et que l'on cale ensuite les poids selon les totaux de contrôle de la population – fournit plus de robustesse aux mécanismes de non-réponse présumés qu'une approche à une seule étape où l'on procède à un calage uniquement. La procédure à deux étapes permet de supprimer le biais si le modèle de propension de réponse ou le modèle de calage est approprié (voir la section 3.2 pour obtenir les modèles de non-réponse).

Dans certaines situations, un grand nombre de variables auxiliaires pourraient potentiellement être utilisées aux fins de calage. L'utilisation de toutes les variables dans le modèle de calage pourrait donner lieu à une instabilité en raison des variances gonflées ou de l'augmentation des poids. Une solution potentielle est de restreindre l'ensemble de variables utilisé dans le calage afin d'éviter la multicolinéarité. Cardot, Goga et Shehzad (2017) ont proposé de réduire la dimension des variables de calage au moyen de l'analyse en composantes principales. Il est aussi possible de réduire la multicolinéarité en utilisant les critères d'optimisation pénalisée de type régression ridge (Rao et Singh, 1997, 2009; Beaumont et Bocci, 2008; Goga, 2024).

D'autres spécialistes ont employé des méthodes d'apprentissage automatique pour sélectionner des variables et déterminer la forme des ajustements de calage. Les spécialistes des enquêtes ont longtemps utilisé des méthodes d'arbre de décision telles que l'algorithme de détection automatique d'interactions du khi carré (CHAID) (Kass, 1980) pour sélectionner des variables et former des cellules d'ajustement pour la non-réponse (voir, par exemple, Rizzo, Kalton et Brick, 1996). Depuis, un certain nombre de spécialistes ont profité d'autres avancées dans l'apprentissage automatique pour améliorer les ajustements pour la non-réponse dans les enquêtes. Par exemple, Phipps et Toth (2012) ont construit des arbres de régression pour étudier les tendances de non-réponse dans l'Occupational Employment Statistics Survey. Lohr, Hsu et Montaquila (2015) ont examiné le rendement de divers arbres de régression et de la forêt aléatoire, qui donne une approximation des prédictions à partir d'un grand nombre d'arbres, pour estimer les propensions de réponse. McConville et Toth (2019) ont établi les propriétés asymptotiques d'estimateurs stratifiés *a posteriori* où les strates *a posteriori* sont choisies par un modèle d'arbre de régression. Opsomer et Riddles (2025) ont élargi l'approche CHAID pour tenir compte du plan d'enquête en incorporant une correction Rao-Scott dans le critère de séparation.

Breidt et Opsomer (2017) ont examiné les approches assistées par un modèle pour estimer les moyennes et les totaux de population. Une variable d'étude y est prédite par une fonction $m(\mathbf{x})$ des variables auxiliaires $\mathbf{x}$. Le total de population est alors estimé par l'estimateur par différence

$$\hat{Y}_d = \sum_{i=1}^{N} \hat{m}(\mathbf{x}_i) + \sum_{i \in \mathcal{S}} d_i \left[y_i - \hat{m}(\mathbf{x}_i) \right], \tag{3.1}$$

où $\mathcal{S}$ désigne l'ensemble d'unités échantillonnées, $d_i = 1/\pi_i = 1/P(i \in \mathcal{S})$ est le poids de sondage pour l'unité i et $\hat{m}$ permet d'estimer m. L'estimateur est asymptotiquement sans biais dans des conditions de régularité, car le second terme permet de corriger le biais potentiel dans le modèle de prédiction. Wu et

Sitter (2001), proposant d'utiliser une fonction générale m pour le calage, ont utilisé un estimateur semblable à (3.1), mais où $\left[\sum_{i=1}^{N}\hat{m}(\mathbf{x}_i)-\sum_{i\in\mathcal{S}}d_i\hat{m}(\mathbf{x}_i)\right]$ est multiplié par un coefficient de régression.

Montanari et Ranalli (2005) et Breidt et Opsomer (2017) ont présenté les propriétés statistiques de l'estimateur dans (3.1) au moyen d'un polynôme local, d'une fonction spline, d'un réseau neuronal et d'autres modèles non paramétriques utilisés pour $m(\mathbf{x})$. McConville, Breidt, Lee et Moisen (2017) et Chen, Valliant et Elliott (2018) ont sélectionné des variables de régression en utilisant la méthode LASSO, et Dagdoug, Goga et Haziza (2023b) ont dérivé des propriétés asymptotiques quand le modèle est choisi en utilisant la forêt aléatoire. Lundy et Rao (2022), dans une étude par simulation des méthodes d'apprentissage automatique pour le calage, ont découvert que les arbres de régression et la méthode LASSO avaient un rendement similaire dans le cas de covariables catégoriques donnant lieu aux interactions et aux effets principaux, et les deux méthodes traitaient automatiquement la multicolinéarité parmi les variables auxiliaires. La majeure partie de la recherche sur l'estimation assistée par un modèle a été effectuée dans un contexte de réponse complète, mais les résultats ont aussi été appliqués au calage pour traiter la non-réponse (voir la section 3.2).

Bien que ces fonctions plus générales puissent donner lieu à une plus petite variance pour le total de population estimé $\hat{Y}_d$, il y a des avantages à utiliser une fonction de régression linéaire pour m. Puisque $\hat{m}(\mathbf{x})=\mathbf{x}^T\hat{\mathbf{B}}=\mathbf{x}^T\left(\sum_{k\in\mathcal{S}}d_k\mathbf{x}_k\mathbf{x}_k^T\right)^{-1}\sum_{k\in\mathcal{S}}d_k\mathbf{x}_k y_k$ est une fonction linéaire de y, l'estimateur par la régression généralisée peut être calculé comme suit : $\sum_{i\in\mathcal{S}}d_i g_i y_i$, où $g_i=\left[1+(\mathbf{X}-\hat{\mathbf{X}})^T\left(\sum_{k\in\mathcal{S}}d_k\mathbf{x}_k\mathbf{x}_k^T\right)^{-1}\mathbf{x}_i\right]$. Le calage au moyen d'un modèle linéaire permet une cohérence interne en utilisant le même ensemble de poids ajustés, $w_i=d_i g_i$, pour toutes les estimations. De plus, il n'est pas nécessaire de connaître $\mathbf{x}_i$ pour chaque unité de population – $\mathbf{x}_i$ doit être connu pour $i\in\mathcal{S}$, mais l'estimateur dépend de l'information auxiliaire à propos des unités non échantillonnées uniquement par l'entremise du total de population $\mathbf{X}$.

Quand une fonction de prédiction non linéaire est utilisée dans (3.1), le calage est effectué en ce qui a trait au total de population des valeurs prédites, $\sum_{i=1}^{N}\hat{m}(\mathbf{x}_i)$, plutôt qu'en ce qui a trait à $\mathbf{X}$. Pour la plupart des fonctions m, $\mathbf{x}_i$ doit être connu pour toutes les unités échantillonnées et non échantillonnées pour pouvoir calculer $\sum_{i=1}^{N}\hat{m}(\mathbf{x}_i)$, contrairement à la régression linéaire. L'utilisation des fonctions non linéaires peut aussi donner lieu à de multiples ensembles de poids quand les modèles sont ajustés pour différentes variables à l'étude. Par exemple, en ayant une fonction d'arbre de régression pour m, l'estimateur dans (3.1) peut être exprimé sous la forme $\sum_{i\in\mathcal{S}}w_i y_i$ pour un ensemble de poids, mais le poids w_i dépend du nombre d'unités de population qui se situent dans le nœud terminal qui contient l'unité i et aura une valeur différente pour une autre variable à l'étude. Il est possible d'assurer une cohérence d'une estimation à l'autre en adaptant le modèle à une variable à l'étude et en utilisant ces poids pour toutes les estimations, mais cela pourrait réduire l'efficacité pour d'autres variables à l'étude. Montanari et Ranalli (2009) ont proposé d'obtenir un seul ensemble de poids w_i qui satisfont les multiples contraintes de calage $\sum_{i\in\mathcal{S}}w_i\mathbf{u}_i=\sum_{i=1}^{N}\mathbf{u}_i$, où le vecteur $\mathbf{u}_i=(\hat{m}_1(\mathbf{x}_i),\ldots,\hat{m}_p(\mathbf{x}_i),\mathbf{z}_i^T)^T$ consiste en des prédictions de modèle distinctes pour chacune

des P variables clés à l'étude ainsi qu'en un sous-ensemble $\mathbf{z}$ des variables auxiliaires $\mathbf{x}$. Ils ont proposé d'utiliser les estimateurs par la régression de type ridge pour réduire la variabilité des poids et respecter les contraintes d'intervalle.

3.2 Modèles de non-réponse

Puisque les taux de réponse pour les enquêtes-ménages ont diminué, la recherche sur les modèles de pondération, de calage et d'imputation pour les données d'enquête manquantes s'est intensifiée. Bon nombre des méthodes peuvent être appliquées aux échantillons non probabilistes ainsi qu'aux échantillons probabilistes ayant des cas de non-réponse.

Les méthodes d'ajustement de la pondération pour la non-réponse tentent d'estimer la probabilité ϕ_i que l'unité i réponde à l'enquête, ce que l'on appelle la propension de réponse. Si l'échantillon initial est choisi de façon probabiliste, le total de population est alors estimé par $\hat{Y}_\phi = \sum_{i\in\mathcal{R}} y_i \big/ \left(\pi_i \hat{\phi}_i\right)$, où $\mathcal{R}$ est l'ensemble de répondants à l'enquête et $\hat{\phi}_i$ est un estimateur de ϕ_i. Kim et Kim (2007) ont dérivé la valeur et la variance attendues du total de population estimé quand $\hat{\phi}_i$ est estimé en utilisant un modèle de régression logistique. L'estimateur $\hat{Y}_\phi$ est approximativement sans biais quand toutes les propensions de réponse dépassent une constante fixe $k > 0$ et sont modélisées avec exactitude par la fonction présumée. Il s'agit là de solides hypothèses, car de nombreuses enquêtes ont des non-répondants irréductibles dont les propensions de réponse pourraient être considérées comme nulles et, en pratique, le mécanisme de réponse pourrait avoir une forme fonctionnelle différente ou pourrait dépendre de variables non mesurées.

Les récentes avancées dans la modélisation et l'imputation de la non-réponse sont trop nombreuses pour les décrire dans le présent article, et nous proposons au lecteur de consulter Kim et Shao (2022) pour obtenir un examen sur l'imputation multiple et l'imputation fractionnaire, les données manquantes non ignorables et les données en grappe. Deux avancées concernent toutefois d'autres thèmes du présent article. La première, décrite aux sections 2.1 et 3.1, est l'utilisation accrue des modèles d'apprentissage automatique pour l'imputation et les ajustements de poids pour la non-réponse. Ferri-García et Rueda (2020) ont examiné les méthodes d'apprentissage automatique pour estimer les propensions de réponse dans les enquêtes en ligne.

La deuxième avancée, qui concerne l'examen de l'estimation assistée par un modèle de la section précédente, est la recherche sur la double robustesse pour les modèles de non-réponse. Dans une telle approche, comme elle a été présentée par Kott (1994) pour les données d'enquête, deux modèles sont mis au point. Le premier modèle prédit la propension de répondre à l'enquête; le second modèle (d'imputation) prédit y pour les unités manquantes à partir de l'information auxiliaire, qui est désignée par $\hat{m}(\mathbf{x})$. Un estimateur doublement robuste du total de population a une forme analogue à (3.1) :

$$\hat{Y}_{\mathrm{DR}} = \sum_{i\in\mathcal{S}} \frac{\hat{m}(\mathbf{x}_i)}{\pi_i} + \sum_{i\in\mathcal{R}} \frac{y_i - \hat{m}(\mathbf{x}_i)}{\pi_i \hat{\phi}_i}. \tag{3.2}$$

Si $\mathbf{x}$ est connu pour chaque unité de population, $\sum_{i=1}^{N} \hat{m}(\mathbf{x}_i)$ peut être substitué au premier terme dans (3.2). L'estimateur est approximativement sans biais si le modèle de propension ϕ ou le modèle d'imputation m

est correctement précisé, d'où le nom de double robustesse. Toutefois, si les deux modèles sont mal précisés, $\hat{Y}_{\mathrm{DR}}$ peut encore comporter un biais important; Kang et Schafer (2007, page 523) ont conclu que dans au moins certains contextes, deux modèles incorrects ne sont pas mieux qu'un. Chen et Haziza (2017, 2023) et Wu (2023) ont décrit de récents travaux sur les estimateurs doublement robustes et les estimateurs multi-robustes en présence de non-réponse.

3.3 Estimation de la variance

Quand le premier numéro de *Techniques d'enquête* a été publié, en 1975, la plupart des estimations de la variance étaient calculées en utilisant des formules pour le plan d'échantillonnage approprié et les méthodes de la série de Taylor. Les méthodes des groupes aléatoires, la méthode du jackknife et les répliques répétées équilibrées avaient été introduites pour au moins certains plans d'enquête avant 1975 (Mahalanobis, 1939; Durbin, 1959; McCarthy, 1966), mais un manque de puissance de calcul et de logiciels limitait leur utilisation en pratique. Au moment où Rao (1999) a rédigé son examen, une théorie asymptotique rigoureuse avait été élaborée pour les méthodes d'estimation de la variance par répliques (voir, par exemple, Krewski et Rao, 1981; Rao et Wu, 1985) et plusieurs organismes de statistique utilisaient la méthode du jackknife ou les répliques répétées équilibrées pour l'estimation de la variance.

L'estimation de la variance par répliques est maintenant la norme pour de nombreuses enquêtes. Cela a été facilité par la disponibilité généralisée des logiciels pour construire des poids de rééchantillonnage et calculer les variances. Les méthodes des répliques comportent de nombreux avantages, y compris la capacité de calculer les variances pour plusieurs types de statistiques sans avoir à dériver des formules de variance distinctes. Elles intègrent également les effets des ajustements de la pondération sur les estimations de la variance avec remise (qui estiment aussi les variances sans remise quand les fractions d'échantillonnage du premier degré sont petites), car les mêmes étapes d'ajustement pour la non-réponse sont effectuées sur les poids complets et sur chaque ensemble de poids de rééchantillonnage. On a vu, au cours des 25 dernières années, une utilisation accrue de la méthode bootstrap avec rééchelonnement des poids (re-scaling bootstrap) (Rao et Wu, 1988; Rao, Wu et Yue, 1992) en raison de sa souplesse sur le plan des tailles d'échantillon et du nombre d'estimations bootstrap; Mashreghi, Haziza et Léger (2016) ont examiné les récentes avancées relatives à la méthode bootstrap au moyen d'échantillons probabilistes, et Beaumont et Émond (2022) ont proposé une méthode bootstrap qui peut être utilisée avec de petites fractions d'échantillonnage du premier degré.

Les méthodes des répliques telles que la méthode du jackknife et la méthode bootstrap sont habituellement utilisées quand toutes les estimations ponctuelles sont calculées en utilisant un seul ensemble de poids. Les poids de rééchantillonnage sont dérivés de ce seul ensemble selon la méthode des répliques particulière. Quand les estimations dépendent de multiples ensembles de poids, ou ont des sources d'aléa supplémentaires au plan d'enquête et aux ajustements pour la non-réponse, les méthodes des répliques standards doivent être modifiées pour prendre en compte toutes les sources de variabilité. Certaines estimations assistées par un modèle ou doublement robustes dépendent des prédictions de y ainsi que d'un

vecteur de poids final. Wu (2022) a résumé les méthodes d'estimation de la variance qui peuvent être utilisées pour les estimations doublement robustes.

À la suite des premiers travaux de Kott et Stukel (1997), il y a eu un certain nombre d'avancées dans l'utilisation des méthodes des répliques pour calculer les variances pour les échantillons à deux phases. Les procédures décrites dans Kim et Yu (2011) consistent à créer des ensembles distincts de répliques jackknife pour les vecteurs de poids de phase 1 et de phase 2. Opsomer, Breidt, White et Li (2016) ont proposé une estimation de la variance au moyen de répliques des différences successives dans les échantillons à deux phases, et Beaumont, Béliveau et Haziza (2015) ont fait un lien entre une proposition d'estimateur de variance simplifié pour des plans à deux phases et les méthodes des répliques.

Une autre situation dans laquelle l'estimation dépend de plus d'un seul vecteur de poids se produit quand il faut adapter des modèles multiniveaux aux données d'enquête. Dans la méthode de vraisemblance composite proposée par Yi, Rao et Li (2016), les estimations des composantes de variance dépendent des probabilités d'inclusion des UPE du premier degré et des probabilités d'inclusion conjointes pour les paires d'éléments au sein des UPE. Lumley et Huang (2024) ont mis au point une méthode bootstrap qui permet d'estimer la variance avec remise en utilisant des poids de rééchantillonnage pour les paires d'observations.

En plus des nombreuses avancées dans l'estimation de la variance par répliques, il y a aussi eu des avancées dans les méthodes de linéarisation. L'approche traditionnelle exprime le paramètre d'intérêt en tant que fonction d'un vecteur des totaux de population $\mathbf{Y}$ et donne une approximation de la variance en utilisant des dérivées de la fonction, qui peuvent être définies explicitement ou implicitement, en ce qui a trait aux composantes de $\mathbf{Y}$ (Binder, 1983). Si $\theta = g(\mathbf{Y})$ et $\hat{\theta} = g(\hat{\mathbf{Y}})$, il est alors possible de définir une variable linéarisée

$$u_i = \left(\frac{\partial g(\mathbf{a})}{\partial \mathbf{a}} \right)^T \Bigg|_{\mathbf{a}=\mathbf{Y}} \mathbf{y}_i, \tag{3.3}$$

et, dans des conditions de régularité, $V(\hat{\theta})$ est estimé approximativement par la variance de $\hat{U} = \sum_{i \in \mathcal{S}} d_i u_i$, l'estimation d'enquête de $U = \sum_{i=1}^{N} u_i$. Binder (1996) a recommandé une variante de (3.3) dans laquelle la dérivée est évaluée à $\hat{\mathbf{Y}}$. Demnati et Rao (2004) ont noté que $\hat{\theta}$ peut aussi être exprimé en tant que fonction $f(\mathbf{d})$ du vecteur des poids de sondage et ont écrit $\hat{\theta} \approx \sum_{i \in \mathcal{S}} d_i z_i$, où $z_i = \left(\frac{\partial f(\mathbf{b})}{\partial b_k} \right) \Big|_{\mathbf{b}=\mathbf{1}}$. Graf (2011) a linéarisé l'estimateur de θ en prenant des dérivées relativement aux indicateurs d'inclusion de l'échantillon et en évaluant les dérivées aux probabilités d'inclusion.

4. Estimation sur petits domaines

Les enquêtes-échantillons sont habituellement conçues pour fournir des estimations fiables des totaux et des moyennes pour la population cible ainsi que pour des sous-populations (domaines) ayant des échantillons suffisamment grands. Ces estimations s'appuient sur des données propres au domaine seulement et sont appelées « estimateurs directs ». Cependant, la demande en statistiques fiables pour des régions locales

(appelées petits domaines) a grandement augmenté au cours des quelque 30 dernières années. Les estimateurs directs pour les petits domaines ne sont pas fiables en raison des petites tailles d'échantillon au sein des domaines ou ne sont pas concevables en raison de l'absence de tailles d'échantillon propres à un domaine. Il est donc nécessaire d'utiliser les données-échantillons dans des domaines connexes à l'aide de modèles de couplage implicites ou explicites. Quand le premier numéro de *Techniques d'enquête* est apparu, en juin 1975, seules quelques applications de méthodes d'estimation sur petits domaines fondées sur des modèles implicites avaient été mises en œuvre. La plupart de ces méthodes étaient basées sur des estimateurs synthétiques qui, par exemple, présumaient que les moyennes des domaines dans une strate *a posteriori* sont approximativement égales à la moyenne de la strate (National Center for Health Statistics, 1968). L'estimation synthétique est fondée sur de solides modèles implicites alors que les modèles explicites tiennent compte des erreurs de modélisation et permettent une vérification et une validation des modèles.

Les modèles explicites au niveau du domaine, à compter des travaux d'avant-garde de Fay et Herriot (1979), ont donné lieu à de nombreuses extensions du modèle au niveau du domaine de Fay-Herriot (FH) et à diverses applications pour estimer les moyennes de domaine. Les modèles au niveau de l'unité ont aussi reçu beaucoup d'attention, à compter des travaux de Battese, Harter et Fuller (1988). Le livre de Rao (2003) a fourni la théorie et les méthodes fondamentales pour les modèles au niveau du domaine et au niveau de l'unité. Étant donné les nombreuses extensions utiles des modèles de base au niveau du domaine et au niveau de l'unité se trouvant dans la littérature après 2003, une deuxième édition du livre, écrite par Rao et Molina (2015), a fourni un compte rendu exhaustif des avancées en date de 2014. De nombreuses nouvelles extensions ont vu le jour dans la littérature depuis 2015. Un récent livre de Morales, Esteban, Pérez et Hobza (2021) a fourni des renseignements sur les avancées mathématiques dans l'estimation sur petits domaines au moyen d'un code R pour les applications. Molina et Rao (2023), Sugasawa et Kubokawa (2020) et Ghosh (2020) ont retracé l'histoire de l'estimation sur petits domaines au cours des 50 dernières années et abordé certains sujets actuels liés à l'estimation sur petits domaines fondée sur un modèle.

Le modèle de FH de base au niveau du domaine pour estimer les moyennes de domaine consiste en un modèle d'échantillonnage sur les estimateurs directs et un modèle connexe sur les moyennes de domaine couplant les moyennes de domaine θ_i à l'aide de covariables au niveau du domaine $\mathbf{z}_i$. Les estimateurs directs sont obtenus à partir de données au niveau de l'unité propres au domaine qui tiennent compte du plan d'enquête. Le modèle d'échantillonnage est donné par $\hat{\theta}_i = \theta_i + e_i, i = 1,\ldots,m,$ où m est le nombre de domaines et les erreurs d'échantillonnage e_i sont présumées indépendantes avec une moyenne de 0 et une variance connue ψ_i. Par souci de simplicité, nous présumons que tous les domaines dans la population sont échantillonnés. Le modèle de couplage est donné par $\theta_i = \mathbf{z}_i^T\boldsymbol{\beta} + v_i,$ où $\boldsymbol{\beta}$ est le vecteur des paramètres du modèle et v_i est un effet aléatoire de domaine qui a une moyenne nulle et une variance commune $\sigma_v^2,$ et les effets du domaine sont présumés indépendants. Puisque les deux modèles concordent, un modèle combiné, appelé « modèle de FH », est donné par $\hat{\theta}_i = \mathbf{z}_i^T\boldsymbol{\beta} + v_i + e_i,$ qui est un cas spécial d'un modèle linéaire mixte. Un meilleur prédicteur linéaire sans biais empirique (MPLSBE) de $\theta_i,$ sans hypothèses de

distribution, peut être obtenu. Le MPLSBE peut être exprimé comme une combinaison pondérée de l'estimateur direct et d'un estimateur synthétique de type régression. On accorde plus de poids à l'estimateur synthétique quand la variance échantillonnale est grande par rapport à la variance de l'effet aléatoire de domaine. D'autres estimateurs comprennent notamment les meilleurs prédicteurs empiriques (MPE) et les prédicteurs hiérarchiques bayésiens (PHB), selon des hypothèses paramétriques pour les MPE et les PHB et, en plus, des distributions *a priori* sur les paramètres de modèle $\boldsymbol{\beta}$ et σ_v^2 pour les PHB. Les méthodes des MPE et des PHB peuvent traiter plus de cas généraux et la méthode des PHB fournit des inférences exactes, contrairement au MPLSBE et aux MPE, qui nécessitent de se tourner vers des théories asymptotiques pour estimer l'erreur quadratique moyenne de prédiction (EQMP) des estimateurs. Rao (1999) a décrit des méthodes de linéarisation pour estimer l'EQMP et, par la suite, des méthodes du jackknife et des méthodes bootstrap paramétriques pour estimer l'EQMP ont été proposées; voir Rao et Molina (2015), section 6.2.

Les modèles au niveau du domaine ont l'avantage de tenir compte du plan d'enquête dans le modèle d'échantillonnage et de s'appuyer sur les covariables au niveau du domaine dans le modèle de couplage. Les covariables au niveau du domaine sont plus facilement accessibles que les covariables au niveau de l'unité. En raison de ces avantages, le modèle de FH de base a été élargi dans plusieurs directions différentes et pour aborder de nombreux enjeux pratiques. Il s'agit notamment des suivants : 1) les covariables au niveau du domaine sujettes à des erreurs d'échantillonnage ou à des erreurs de mesure; 2) les variances échantillonnales inconnues; 3) le modèle de couplage mal précisé; 4) l'étalonnage des estimateurs au niveau du domaine avec un estimateur direct fiable à un niveau agrégé; 5) les effets de domaine incertains concernant l'absence d'effet de domaine au-delà des covariables pour plusieurs domaines; voir Rao et Molina (2015), section 6.4. Ghosh, Ghosh, Maples et Tang (2022) ont utilisé des distributions *a priori* globales et locales pour l'estimation hiérarchique bayésienne afin de répondre à l'enjeu 5.

Les extensions des modèles de FH de base au niveau du domaine comprennent : 1) l'utilisation de données transversales et de données tirées de séries chronologiques pour emprunter de la puissance sur le plan transversal et avec le temps; 2) les modèles spatiaux impliquant une corrélation spatiale parmi les effets aléatoires de domaine; 3) les modèles de FH bivariés; 4) l'estimation des proportions de domaine pour une variable binaire ou une variable polytomique en utilisant une formule de modèle linéaire mixte logistique; 5) les modèles de sous-domaine ou les modèles doubles tels que la modélisation des comtés (sous-domaines) au sein des États (domaines); voir Rao et Molina (2015), chapitre 8. Les modèles spatiaux (extension 2) sont pertinents quand les moyennes de domaine des domaines rapprochés géographiquement affichent une variation spatiale et que les covariables disponibles au niveau du domaine ne permettent pas d'expliquer la variation spatiale. En outre, si certains domaines ne sont pas échantillonnés, le modèle de FH traditionnel s'appuie alors sur des estimations synthétiques fondées uniquement sur les covariables associées à ces domaines. En revanche, les modèles spatiaux peuvent donner lieu à de meilleures estimations en modifiant les estimations synthétiques au moyen du modèle spatial. Chung et Datta (2020) ont utilisé des modèles spatiaux bayésiens

et l'approche hiérarchique bayésienne pour estimer les moyennes des domaines échantillonnés et non échantillonnés.

Pour ce qui est des modèles au niveau de l'unité, encore une fois, nous présumons que tous les domaines sont échantillonnés par souci de simplicité. Les données de la population au niveau de l'unité sont données par $\{(y_{ij}, \mathbf{x}_{ij}), j = 1,\ldots,N_i, i = 1,\ldots,m\}$ et le modèle de population est un modèle de régression à erreurs emboîtées $y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij}, j = 1,\ldots,N_i, i = 1,\ldots,m.$ Dans ce cas-ci, i et j désignent le domaine et l'unité au sein du domaine, y_{ij} et $\mathbf{x}_{ij}$ désignent une variable d'intérêt et un vecteur de covariables connexes, v_i est un effet aléatoire de domaine que l'on présume indépendant de l'erreur d'unité e_{ij}, et les deux sont indépendants et répartis de manière identique avec une moyenne de zéro et des variances communes σ_v^2 et σ_e^2, respectivement. Les tailles d'échantillon des domaines sont désignées par n_i et l'on présume que l'échantillon obéit au modèle de population, ce qui indique un échantillonnage non informatif ou l'absence de biais de sélection au sein des domaines. Dans les conditions indiquées ci-dessus, les estimateurs du MPLSBE des moyennes de domaine peuvent être obtenus sans hypothèses de normalité en tant que combinaisons pondérées d'un estimateur par la régression sur l'échantillon et d'un estimateur synthétique de type régression. Nous présumons qu'il n'y a aucune erreur dans le fait de coupler y_{ij} au vecteur de covariables connexes $\mathbf{x}_{ij}$. Des travaux d'envergure ont été effectués sur l'estimation de l'EQMP du MPLSBE; voir Rao et Molina (2015), chapitre 7.

Fay (2018) a affirmé que lorsque les données au niveau de l'unité sont disponibles, la moyenne pondérée de l'échantillon dans le modèle au niveau du domaine devrait être remplacée par l'estimateur par la régression de l'échantillon, et il a démontré que le MPLSBE obtenu pour le modèle au niveau du domaine est comparable sur le plan de l'efficacité au MPLSBE avec le modèle au niveau de l'unité. Par contre, l'utilisation de sommes pondérées comme estimateur direct dans le modèle au niveau du domaine peut entraîner une perte d'efficacité considérable comparativement au MPLSBE avec le modèle au niveau de l'unité (Hidiroglou et You, 2016). Les enjeux pratiques abordés comprennent notamment l'estimation robuste des moyennes de domaine en présence de valeurs aberrantes, l'estimation du pseudo-MPLSBE fondée sur les poids du plan de sondage et l'erreur de spécification du modèle; voir Rao et Molina (2015), section 7.6. Sun, Berg et Zhu (2024) ont examiné les travaux sur les modèles au niveau de l'unité multivariés et proposé une extension dans laquelle les variables à l'étude peuvent être continues ou discrètes. Pfeffermann et Sverchkov (2007) et Verret, Rao et Hidiroglou (2015) ont proposé des modèles pour traiter l'échantillonnage informatif au sein des domaines. Verret et coll. (2015) ont aussi montré que le pseudo-MPLSBE traite très bien l'échantillonnage informatif.

Les modèles au niveau de l'unité permettent l'estimation de paramètres complexes tels que les indicateurs de pauvreté de petits domaines utilisés pour construire des cartes géographiques de la pauvreté. La Banque mondiale utilise le taux de pauvreté, l'écart de pauvreté et la gravité de la pauvreté comme indicateurs de la pauvreté. La meilleure estimation empirique et l'estimation hiérarchique bayésienne des indicateurs de pauvreté avec les modèles au niveau de l'unité ont été étudiées; voir Guadarrama, Molina et Rao (2016) et Rao et Molina (2015), sections 9.4 et 10.7.3.

Les modèles au niveau du domaine et de l'unité ont tous deux été étendus pour s'appuyer sur les fonctions autres que la régression linéaire pour le modèle de couplage. Opsomer, Claeskens, Ranalli, Kauermann et Breidt (2008) et Rao, Sinha et Dumitrescu (2014) ont utilisé la régression spline pénalisée pour permettre aux données de déterminer la forme de la courbe dans les modèles au niveau de l'unité. Lohr et Mendez (2009), Krennmair et Schmid (2022) et Beaumont, Bocci, Bosa et Sombo (2024) ont étudié l'utilisation de la forêt aléatoire dans les modèles sur petits domaines.

5. Intégration des données et échantillons non probabilistes

Bon nombre des sujets mentionnés dans les sections précédentes mettent en cause l'intégration de données provenant de plusieurs sources. Par exemple, le calage et l'estimation sur petits domaines s'appuient sur des données provenant de sources externes pour les totaux de contrôle ou en tant que données d'entrée du modèle. Certaines méthodes d'imputation reposent sur des sources de données externes pour les valeurs à imputer. Au cours des dernières années, alors que le coût des enquêtes a augmenté, on a observé une intensification de la recherche sur l'intégration d'ensembles de données administratives, de données de capteurs, de données provenant des réseaux sociaux et de sources de données autres que des données d'enquête pour produire des statistiques. De nombreux organismes de statistique se tournent vers des statistiques provenant de sources multiples pour profiter des ensembles de données recueillies à des fins autres que statistiques (De Waal, van Delden et Scholtus, 2020). La recherche sur l'utilisation de multiples sources de données pour produire des statistiques est examinée dans Lohr et Raghunathan (2017), Thompson (2019), Zhang et Chambers (2019), Beaumont (2020), Yang et Kim (2020), Rao (2021) et Cabrera-Álvarez (2022).

Bon nombre des premiers articles de *Techniques d'enquête* sur l'intégration des données avaient trait au couplage des enregistrements provenant de multiples ensembles de données. Le couplage d'enregistrements continue d'être une méthode principale d'intégration des données – il peut être utilisé pour élargir l'ensemble de variables dans un ensemble de données, imputer des valeurs manquantes en couplant un non-répondant à l'enquête à des dossiers administratifs ou former un ensemble de données enchaîné d'entités uniques en détectant les enregistrements dans les multiples sources qui appartiennent à la même entité. Asher, Resnick, Brite, Brackbill et Cone (2020), Winkler (2021) et Binette et Steorts (2022) ont examiné les récents travaux de recherche sur les méthodes de couplage d'enregistrements; Han et Lahiri (2019) ont étudié l'analyse statistique au moyen de données couplées et Reiter (2021) a résumé les approches qui peuvent tenir compte de l'incertitude dans les couplages.

Un couplage d'enregistrements exact exige que toutes les sources de données contiennent des renseignements détaillés qui peuvent être utilisés pour déterminer les enregistrements appartenant à la même entité. Dans la présente section, nous nous concentrons sur d'autres méthodes qui ont été proposées pour combiner les renseignements obtenus d'enquêtes ayant d'autres sources et, en particulier, sur les dernières avancées qui n'exigent pas le couplage d'unités individuelles.

Il y a de nombreux avantages à utiliser des sources de données autres que des données d'enquête pour augmenter (ou parfois remplacer) les données d'enquête. L'utilisation d'un ensemble de données (par exemple, des dossiers administratifs) qui sont déjà recueillies à d'autres fins peut réduire les coûts et le fardeau des enquêtes. D'autres sources de données peuvent aussi fournir des renseignements sur les membres de la population qui pourraient ne pas se trouver dans un échantillon probabiliste, comme les non-répondants déterminés ou les personnes hors du champ de l'enquête. Un vaste ensemble de données administratives peut permettre de fournir des renseignements détaillés pour de petites sous-populations qui ne sont pas accessibles au moyen d'une enquête.

La principale préoccupation liée à l'utilisation d'échantillons non probabilistes aux fins d'inférence pour une population est un biais potentiel attribuable à :

- l'autosélection. Les dossiers administratifs tels que les dossiers d'impôt sur le revenu peuvent avoir une population bien définie, mais d'autres types d'enquêtes, comme les sondages en ligne à participation volontaire ou les données des médias sociaux, sont sujets à l'autosélection. La probabilité qu'une unité de population participe dans l'échantillon est généralement inconnue.
- le sous-dénombrement. Les dossiers fiscaux manquent de renseignements sur les personnes qui ne sont pas tenues de produire des déclarations de revenus, les dossiers médicaux électroniques manquent de données sur les personnes qui n'utilisent pas les systèmes de soins de santé et les bases de données liées aux cartes de crédit manquent de renseignements sur les opérations au comptant. Pour les enquêtes à participation volontaire, il y a des personnes dans la population qui n'auront pas l'occasion d'y participer ou qui n'y participeront pas, quelle que soit la situation – ces personnes pourraient être considérées comme ayant une probabilité de participation de 0.
- l'erreur de mesure. Il se pourrait que les concepts mesurés dans les données administratives ou les données autres que les données d'enquête ne soient pas exactement ce que le statisticien souhaite étudier. L'erreur de mesure peut découler d'effets de la formulation des questions, d'erreurs d'enregistrement ou même de fausses déclarations délibérées dans certaines sources (comme cela a été examiné dans Kennedy et coll., 2024).

Keiding et Louis (2018) ont abordé ces enjeux dans le contexte des études épidémiologiques. Bien entendu, le biais attribuable à la non-réponse, au sous-dénombrement et à l'erreur de mesure est aussi préoccupant pour les échantillons probabilistes, et les méthodes statistiques pour la réduction des biais au moyen de multiples sources de données se sont inspirées de celles qui avaient été mises au point pour les échantillons probabilistes.

Le problème de sous-dénombrement peut parfois être réglé en employant deux sources de données ou plus dans une approche à bases de sondage multiples. Le sous-dénombrement provenant d'une seule source peut alors être partiellement compensé par l'autre source. La combinaison des données exige toutefois de connaître le chevauchement dans les deux univers représentés par les échantillons – en d'autres mots, même si les analystes n'ont pas besoin des données d'identification détaillées requises aux fins de couplage

d'enregistrements, ils doivent savoir si une unité échantillonnée à partir d'une seule source de données se trouve aussi dans l'univers pour l'autre source de données. Lohr (2021) s'est penchée sur les extensions d'approches à bases de sondage multiples qui pourraient être prises en compte pour fusionner les échantillons probabilistes et non probabilistes.

Lorsque l'on présume que les multiples sources proviennent de la même population ou que le chevauchement est inconnu, les modèles hiérarchiques peuvent être utilisés pour combiner leurs renseignements. Lohr et Raghunathan (2017) ont examiné les méthodes d'intégration des modèles hiérarchiques jusqu'en 2016. Dans bon nombre de ces approches, qui sont liées aux méthodes statistiques utilisées pour la méta-analyse, les termes à effets aléatoires modélisent l'hétérogénéité dans l'ensemble des sources de données. On présume qu'une moyenne de population (ou, dans certains cas, de domaine) de la source de données j, $\hat{\mu}_j$, suit une distribution $N(\mu+\Delta_j,\tau_j)$; si la source j est une enquête probabiliste de grande qualité, on présume que le biais Δ_j est nul. Cahoy et Sedransk (2023) ont permis aux moyennes d'enquête $\hat{\mu}_j$ d'être réunies en grappes, et la forme de la mise en grappes était estimée à partir des données, de sorte que la distribution *a posteriori* de μ dépendait plus fortement des enquêtes dans la même grappe que l'enquête probabiliste de grande qualité. Wiśniowski, Sakshaug, Perez Ruiz et Blom (2020) ont proposé une méthode bayésienne pour réduire les variances des estimations faites à partir d'un petit échantillon probabiliste en dérivant les distributions *a posteriori* pour les paramètres du modèle à partir d'un plus grand échantillon non probabiliste parallèle. Nandram et Rao (2024) ont obtenu des estimations sur petits domaines à l'aide d'un modèle bayésien qui intégrait un échantillon probabiliste à un échantillon non probabiliste.

L'approche à bases de sondage multiples et l'approche de modélisation hiérarchique mentionnées ci-dessus exigent que la variable à l'étude y soit mesurée dans toutes les sources de données. Dans de nombreuses situations pertinentes, toutefois, aucun échantillon probabiliste qui mesure y n'est disponible. Wu (2022) et Valliant (2024) ont examiné deux approches principales qui ont été adoptées pour obtenir des estimations du total de population Y pour une variable y mesurée seulement dans un échantillon non probabiliste (de commodité) $\mathcal{S}_C$. Les deux approches laissent supposer l'existence d'un échantillon probabiliste de grande qualité ou d'un recensement $\mathcal{S}_P$ qui peut être utilisé pour ajuster le biais de sélection. Bien que y ne soit pas mesurée dans $\mathcal{S}_P$, $\mathcal{S}_C$ et $\mathcal{S}_P$ mesurent tous deux un ensemble commun de variables auxiliaires $\mathbf{x}$.

La première approche, semblable à la pondération par l'inverse de la propension pour la non-réponse dans les échantillons probabilistes, exprime la probabilité de participation de l'échantillon de commodité $\xi_i = P(i \in \mathcal{S}_C \mid \mathbf{x}_i, y_i)$ en tant que fonction de $\mathbf{x}_i$ et de y_i, et estime Y par $\hat{Y}_1 = \sum_{i\in\mathcal{S}_C} y_i \big/ \hat{\xi}_i$, où $\hat{\xi}_i$ est un estimateur de ξ_i. La seconde approche prédit la variable à l'étude au moyen d'un modèle $y = m(\mathbf{x})$ et estime Y par $\hat{Y}_2 = \sum_{i\in\mathcal{S}_P} d_i \hat{m}(\mathbf{x}_i)$, où d_i est le poids pour l'unité i associée à l'échantillon probabiliste. Wu (2022) a aussi recommandé une approche doublement robuste qui combine l'estimation pondérée par l'inverse de la propension et l'estimation par imputation de manière semblable à (3.2) : $\hat{Y}_3 = \sum_{i\in\mathcal{S}_C} \left[y_i - \hat{m}(\mathbf{x}_i)\right] \big/ \hat{\xi}_i + \hat{Y}_2$.

De nombreuses méthodes ont été proposées pour estimer les probabilités de participation de l'échantillon de commodité ξ_i (toutes ces méthodes comportent de solides hypothèses qui sont analysées ci-dessous). Le calage ou la stratification *a posteriori* peuvent être utilisés si les totaux de population pour $\mathbf{x}$ sont connus. Dans de nombreuses situations, toutefois, les totaux de population sont connus uniquement pour un ensemble limité de variables (par exemple, les groupes selon l'âge, la race, le sexe), et une enquête probabiliste pourrait permettre de fournir un ensemble plus riche de variables auxiliaires pour réduire le biais. Chen, Li et Wu (2020) ont proposé une méthode de pseudo-vraisemblance pour estimer ξ_i en utilisant $\mathcal{S}_P$. Savitsky, Williams, Gershunskaya et Beresovsky (2023) ont estimé ξ_i en empilant $\mathcal{S}_P$ et $\mathcal{S}_C$ et en estimant la probabilité qu'une unité dans l'échantillon empilé provienne de l'échantillon de commodité; la fonction de vraisemblance dans Wang, Valliant et Li (2021) est un cas spécial quand $\mathcal{S}_P$ est un recensement de la population.

Comme à la section 3.1, le modèle m utilisé pour l'imputation dans la seconde approche peut être paramétrique (par exemple, régression linéaire ou logistique), non paramétrique ou basé sur des méthodes d'apprentissage automatique. Kaputa, Morris et Holan (2024) ont aussi incorporé la dépendance spatiale dans leur modèle hiérarchique bayésien pour estimer les ventes au détail.

Contrairement à l'inférence à partir d'échantillons probabilistes à taux de réponse complet, qui dépend principalement du plan d'échantillonnage, les inférences statistiques à partir d'échantillons non probabilistes dépendent d'hypothèses robustes souvent impossibles à vérifier. Pour la première approche, on présume en général que le mécanisme d'échantillonnage pour $\mathcal{S}_C$ est ignorable (c'est-à-dire $P[i \in \mathcal{S}_C \mid \mathbf{x}_i, y_i] = P[i \in \mathcal{S}_C \mid \mathbf{x}_i]$) et qu'il n'y a pas de sous-dénombrement (c'est-à-dire tout $\xi_i > 0$). Pour la seconde approche, on présume en général que le modèle m pour prédire y est précisé correctement et que les valeurs prédites $\hat{m}(\mathbf{x}_i)$, obtenues en adaptant le modèle à $\mathcal{S}_C$, fournissent des imputations sans biais pour les unités dans $\mathcal{S}_P$. Pour les deux approches, on présume que les variables $\mathbf{x}$ sont mesurées de la même manière dans $\mathcal{S}_C$ que dans $\mathcal{S}_P$ (par exemple, les mêmes questions sont utilisées et il n'y a pas de source d'erreur de mesure différentielle). Les erreurs-types ou les intervalles bayésiens crédibles pour les estimations, analysés dans Dever et Liao (2023), Wu (2022) et Savitsky et coll. (2023), sous-estiment l'incertitude si l'approche d'intégration des données ne supprime pas la totalité du biais.

6. Discussion

La recherche sur les enquêtes a toujours été dictée par les besoins en information, la disponibilité des données et les avancées technologiques et méthodologiques. Ces aspects ont tous changé depuis l'examen de Rao (1999). La demande accélérée de renseignements statistiques plus rapidement accessibles sur des subdivisions de la population de plus en plus précises a stimulé la recherche sur l'estimation sur petits domaines et l'intégration des données. Les taux de réponse pour de nombreuses enquêtes probabilistes auprès des ménages et des établissements ont diminué, tandis que d'autres types de données (par exemple, les données administratives, les données du secteur privé, les données satellitaires et les données de capteurs,

ainsi que les échantillons de commodité) sont devenues plus abondantes, mais ont souvent une qualité et une couverture de population inconnues. Les ordinateurs d'aujourd'hui permettent des calculs et des approches de modélisation qui étaient impensables en 1999. Les avancées technologiques ont aussi modifié la manière de recueillir et de traiter les données; de nombreuses enquêtes qui auraient été menées en personne ou par téléphone auparavant sont maintenant menées sur Internet, et les méthodes d'apprentissage automatique sont utilisées pour réviser les fichiers et imputer les données manquantes.

Nous prévoyons que les tendances qui consistent à utiliser des modèles à forte intensité de calcul (tels que ceux qui reposent sur l'apprentissage automatique) et à se fier à de multiples sources de données se poursuivront dans un proche avenir. Vu les taux de réponse en baisse et la disponibilité accrue d'autres renseignements, Beaumont (2020) s'est demandé si les enquêtes probabilistes allaient disparaître pour la production de statistiques officielles. Nous sommes d'accord avec lui en ce qui concerne sa conclusion : bien que d'autres sources de données puissent fournir de précieux renseignements, les sources de données de référence telles que les recensements de grande qualité et les enquêtes probabilistes sont encore essentielles pour faire des inférences sur la population. Presque toutes les méthodes d'estimation examinées dans le présent article sur l'inférence de population ou de domaine présupposent l'existence d'au moins une source de données qui produit des estimations sans biais pour au moins certaines variables. C'est en poursuivant la recherche sur l'amélioration de la qualité des enquêtes probabilistes – grâce à des innovations telles que celles décrites à la section 2 – que nous pourrons garantir l'existence de ces sources de données représentatives.

La majeure partie des recherches sur l'utilisation de multiples sources de données a été axée sur l'utilisation d'ensembles de données qui ont déjà été recueillies (appelées données « organiques » par Groves, 2011) et d'enquêtes probabilistes existantes. Nous constatons un besoin de mener plus de recherches sur la conception de systèmes de données qui comprennent des données organiques, des sources telles que les données satellitaires et les données de capteurs, et des enquêtes probabilistes. Une option, abordée dans Lohr et Raghunathan (2017), est de concevoir des enquêtes probabilistes pour compléter les renseignements disponibles provenant d'autres sources. Par exemple, si un ensemble de données administratives fournit des renseignements sur une partie d'une population (par exemple, les contribuables), une enquête supplémentaire pourrait être axée sur l'obtention de renseignements pour les parties de la population manquantes dans l'ensemble de données administratives. Un plan d'enquête optimal dans ce cadre permettrait d'adapter l'enquête en fonction des renseignements existants pour obtenir des estimations ayant une faible EQM pour les estimations clés.

Une faible EQM n'est toutefois pas la seule préoccupation pour un système de données intégré. D'autres aspects qualitatifs tels que la granularité et l'actualité doivent être pris en compte, et le système doit être en mesure de détecter les changements dans les sources de données organiques et y réagir de manière robuste. Par exemple, qu'arrive-t-il si une source de données change ou n'est plus disponible ? Un ensemble de données administratives pourrait être supprimé si le programme qu'il appuie est annulé, un courtier de données privé pourrait cesser de recueillir les données ou augmenter ses prix, ou une grande compagnie

d'assurance-maladie ou une grande société émettrice de cartes de crédit pourrait cesser de transmettre ses données. Il pourrait être souhaitable de concevoir le système de données pour qu'il comporte une redondance de renseignements. Une telle redondance permettrait aussi de mieux évaluer l'exactitude et l'erreur de mesure dans chaque source, et permettrait aux sources de données d'être surveillées afin de pouvoir détecter les changements de qualité. Davantage de recherches sont requises sur la manière de profiter de multiples sources de données pour évaluer les hypothèses et la qualité des estimations (De Broe, Struijs, Daas, van Delden, Burger, van den Brakel, ten Bosch, Zeelenberg et Ypma, 2021). D'autres recherches sont aussi requises sur la protection de la confidentialité pour les données mélangées, puisque les renseignements supplémentaires provenant de multiples sources pourraient exacerber les risques de divulgation (National Academies of Sciences, Engineering, and Medicine, 2024).

L'équité en matière de données est un aspect important alors que les algorithmes d'apprentissage automatique et les données organiques deviennent plus largement utilisés. Les National Academies of Sciences, Engineering, and Medicine (2023) ont décrit l'équité en matière de données comme le fait de [*traduction*] « promouvoir la collecte et l'utilisation des données dans lesquelles toutes les populations, et plus particulièrement celles qui ont été historiquement sous-représentées ou mal représentées dans les enregistrements de données, sont visibles et représentées avec exactitude » (page 3). Les ajustements pour la non-réponse, l'imputation, l'estimation assistée par un modèle, l'estimation sur petits domaines et les méthodes d'intégration des données dépendent tous de prédictions tirées des modèles, et l'exactitude de ces prédictions dépend de la qualité des données utilisées pour adapter ou entraîner le modèle. En particulier, les prédictions peuvent être médiocres pour les sous-populations qui ne sont pas bien représentées dans les données d'entraînement, en raison du biais de sélection ou de la petite taille de l'échantillon, et cela peut se traduire par des estimations inexactes pour ces sous-populations.

Une attention particulière doit être accordée aux problèmes liés à l'équité en matière de données pour les méthodes d'intégration des données décrites à la section 5. Certaines sous-populations pourraient être absentes de toutes les sources de données ou les renseignements se rapportant à elles pourraient être de moins bonne qualité pour le couplage de données ou pour déterminer l'appartenance à la base de sondage dans une approche à bases de sondage multiples. Si $\mathcal{S}_C$ comporte un grand nombre d'enregistrements de données d'une sous-population qui ne comporte qu'une poignée de membres dans $\mathcal{S}_P$, les estimations de la probabilité de participation de l'échantillon non probabiliste ξ_i pourraient être inexactes. Pour la seconde approche, si l'ensemble de données $\mathcal{S}_C$ utilisé pour adapter le modèle m ne contient que quelques membres d'un sous-groupe de population, les prédictions du modèle appliquées à ce sous-groupe pourraient alors être suspectes. L'évaluation de la qualité des prédictions est particulièrement importante pour les méthodes de prédiction plus récentes telles que celles qui reposent sur l'apprentissage automatique. Erman, Rancourt, Beaucage et Loranger (2022) ont abordé l'utilisation équitable et responsable des méthodes d'apprentissage automatique, et le Cadre pour l'utilisation des processus d'apprentissage automatique de façon responsable de Statistique Canada exigeait de mener des évaluations de la qualité des données d'entraînement, de la validité des inférences et de la justesse des algorithmes (Bosa, 2022).

Finalement, toutes les méthodes abordées dans le présent article profiteront de la recherche continue sur la qualité des sources de données et sur la fourniture de mesures exactes de l'incertitude. Kalton (2019, page S27) a écrit : [*traduction*] « À moins que des mesures de la qualité facilement communicables ne soient élaborées et que l'on en fasse activement la promotion pour tous les types de plans de sondage, les utilisateurs pourraient simplement se tourner vers la méthode la moins chère et la plus rapide sans tenir compte de la qualité des estimations produites. » Les erreurs-types découlant de la théorie de l'échantillonnage probabiliste, qui sont exactes quand l'enquête présente une réponse complète et ne comporte aucune erreur de mesure, pourraient grandement sous-estimer l'incertitude à propos des estimations lorsqu'elles sont appliquées à des échantillons probabilistes à faible taux de réponse ou à des échantillons non probabilistes. De multiples sources de données – et de multiples méthodologies pour combiner les données – pourraient permettre aux statisticiens d'enquêtes de mieux comprendre les erreurs non attribuables à l'échantillonnage et de les incorporer dans les mesures de l'incertitude.

La discipline de l'échantillonnage a fait d'énormes progrès depuis le premier numéro de *Techniques d'enquête*, en 1975, et bon nombre des importantes avancées ont été présentées dans cette revue scientifique. Nous nous réjouissons de poursuivre avec 50 autres années d'innovation dans ses pages.

Remerciements

Les travaux de recherche de J.N.K. Rao ont été appuyés par une subvention de recherche accordée par le Conseil de recherches en sciences naturelles et en génie du Canada. Les auteurs remercient le rédacteur en chef et le rédacteur associé pour leurs commentaires constructifs et leurs suggestions.

Bibliographie

Asher, J., Resnick, D., Brite, J., Brackbill, R. et Cone, J. (2020). An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. *International Journal of Environmental Research and Public Health*, 17(6937), 1-16.

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401), 28-36.

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

Beaumont, J.-F., et Bocci, C. (2008). Another look at ridge calibration. *Metron*, 66(1), 5-20.

Beaumont, J.-F., Bocci, C., Bosa, K. et Sombo, S. (2024). The use of random forests in small area estimation. Communication présentée le 20 juin 2024 à l'International Conference on Establishment Statistics, Glasgow.

Beaumont, J.-F., et Émond, N. (2022). A bootstrap variance estimation method for multistage sampling and two-phase sampling when Poisson sampling is used at the second phase. *Stats*, 5(2), 339-357.

Beck, M., Dumpert, F. et Feuerhake, J. (2018). Machine learning in official statistics. *arXiv preprint*. https://arxiv.org/abs/1812.10422.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51(3), 279-292.

Binder, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases : une approche de type « recette ». *Techniques d'enquête*, 22(1), 17-22. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1996001/article/14389-fra.pdf.

Binette, O., et Steorts, R.C. (2022). (Almost) all of entity resolution. *Science Advances*, 8(12), eabi8021.

Bosa, K. (2022). Utilisation responsable de l'apprentissage automatique à Statistique Canada. https://www.statcan.gc.ca/fr/science-donnees/reseau/apprentissage-automatique.

Breidt, F.J., et Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141(1), 479-487.

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

Brick, J.M. (2011). The future of survey sampling. *Public Opinion Quarterly*, 75(5), 872-888.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3), 329-353.

Buskirk, T.D., et Kirchner, A. (2020). Why machines matter for survey and social science researchers: Exploring applications of machine learning methods for design, data collection, and analysis. Dans *Big Data Meets Survey Science: A Collection of Innovative Methods*, (Éds., C.A. Hill, P.P. Biemer, T.D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov et L.E. Lyberg), 9-62. Hoboken, New Jersey: Wiley.

Cabrera-Álvarez, P. (2022). Survey research in times of big data. *EMPIRIA: Revista de Metodología de las Ciencias Sociales*, 53, 31-51.

Cahoy, D., et Sedransk, J. (2023). Combinaison de données provenant d'enquêtes et de sources connexes. *Techniques d'enquête*, 49(1), 75-96. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023001/article/00003-fra.pdf.

Cardot, H., Goga, C. et Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27, 243-260.

Chauvet, G., Deville, J.-C. et Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98(2), 459-471.

Chen, J.K.T., Valliant, R.L. et Elliott, M.R. (2018). Calage assisté par un modèle pour des données de sondage non probabiliste en utilisant le LASSO adaptatif. *Techniques d'enquête*, 44(1), 125-155. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018001/article/54963-fra.pdf.

Chen, S., et Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 104(2), 439-453.

Chen, S., et Haziza, D. (2023). Doubly and multiply robust procedures for missing survey data. *The Survey Statistician*, 88, 75-85.

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

Chung, H.C., et Datta, G.S. (2020). Bayesian hierarchical spatial models for small area estimation. Rapport technique Statistiques nº 2020-07, U.S. Census Bureau Center for Statistical Research & Methodology. https://www.census.gov/content/dam/Census/library/working-papers/2020/adrm/RRS2020-07.pdf.

Coffey, S.M., Damineni, J., Eltinge, J., Mathur, A., Varela, K. et Zotti, A. (2024). Some open questions on multiple-source extensions of adaptive-survey design concepts and methods. *Journal of Official Statistics*, 40(1), 16-37.

Couper, M.P. (2017). New developments in survey data collection. *Annual Review of Sociology*, 43, 121-145.

Daas, P.J., et van der Doef, S. (2020). Detecting innovative companies via their website. *Statistical Journal of the IAOS*, 36(4), 1239-1251.

Dagdoug, M., Goga, C. et Haziza, D. (2023a). Imputation procedures in surveys using nonparametric and machine learning methods: An empirical comparison. *Journal of Survey Statistics and Methodology*, 11(1), 141-188.

Dagdoug, M., Goga, C. et Haziza, D. (2023b). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542), 1234-1251.

Datta, A.R., Ugarte, G. et Resnick, D. (2020). Linking survey data with commercial or administrative data for data quality assessment. Dans *Big Data Meets Survey Science: A Collection of Innovative Methods*, (Éds., C.A. Hill, P.P. Biemer, T.D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov et L.E. Lyberg), 99-129. Hoboken, New Jersey: Wiley.

De Broe, S., Struijs, P., Daas, P., van Delden, A., Burger, J., van den Brakel, J., ten Bosch, O., Zeelenberg, K. et Ypma, W. (2021). Updating the paradigm of official statistics: New quality criteria for integrating new data and methods in official statistics. *Statistical Journal of the IAOS*, 37(1), 343-360.

De Waal, T., van Delden, A. et Scholtus, S. (2020). Multi-source statistics: Basic situations and methods. *Revue Internationale de Statistique*, 88(1), 203-228.

De Leeuw, E.D. (2018). Mixed-mode: Past, present, and future. *Survey Research Methods*, 12(2), 75-89.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30(1), 17-27. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2004001/article/6991-fra.pdf.

Dever, J.A., et Liao, D. (2023). Variance estimation for probability and nonprobability establishment surveys: An overview. Dans *Advances in Business Statistics, Methods and Data Collection*, (Éds., G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, K.J. Thompson et A. van Delden), 659-683. Hoboken, New Jersey: Wiley.

Dever, J.A., et Valliant, R. (2016). General regression estimation adjusted for undercoverage and estimated control totals. *Journal of Survey Statistics and Methodology*, 4(3), 289-318.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4), 891-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 46 (3/4), 477-480.

Erman, S., Rancourt, E., Beaucage, Y. et Loranger, A. (2022). The use of data science in a national statistical office. *Harvard Data Science Review*, 4(4), 1-18.

Falorsi, P.D., et Righi, P. (2008). Une approche d'échantillonnage équilibré pour des plans de sondage à stratification multidimensionnelle pour l'estimation pour petits domaines. *Techniques d'enquête*, 34(2), 247-259. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008002/article/10763-fra.pdf.

Falorsi, P.D., et Righi, P. (2015). Cadre généralisé pour la détermination des probabilités d'inclusion optimales dans les plans de sondage à un degré pour des enquêtes à plusieurs variables et plusieurs domaines. *Techniques d'enquête*, 41(1), 225-247. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015001/article/14149-fra.pdf.

Fay, R.E. (2018). Further comparisons of unit- and area-level small area estimators. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2057-2070, Alexandrie, Virginie.

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(366a), 269-277.

Ferri-García, R., et Rueda, M.d.M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PloS One*, 15(4), e0231500.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition*, 21(4), 1-22.

Ghosh, T., Ghosh, M., Maples, J.J. et Tang, X. (2022). Multivariate global-local priors for small area estimation. *Stats*, 5(3), 673-688.

Gile, K.J., Beaudry, I.S., Handcock, M.S. et Ott, M.Q. (2018). Methods for inference from respondent-driven sampling data. *Annual Review of Statistics and its Application*, 5, 65-93.

Goga, C. (2024). From traditional to modern machine learning estimation methods for survey sampling. *The Survey Statistician*, 90, 18-29.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. Dans *Compositional Data Analysis: Theory and Applications*, (Éds., M. Graf, A. Buccianti et V. Pawlowsky-Glahn), 114-127. Hoboken, New Jersey: Wiley.

Grafström, A., et Ringvall, A.H. (2013). Improving forest field inventories by using remote sensing data in novel sampling designs. *Canadian Journal of Forest Research*, 43(11), 1015-1022.

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75 (5), 861-871.

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

Guadarrama, M., Molina, I. et Rao, J.N.K. (2016). A comparison of small area estimation methods for poverty mapping. *Statistics in Transition*, 17(1), 41-66.

Han, Y., et Lahiri, P. (2019). Statistical analysis with linked data. *Revue Internationale de Statistique*, 87(S1), S139-S157.

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

Harter, R., Battaglia, M.P. Buskirk, T.D., Dillman, D.A., English, N., Fahimi, M., Frankel, M.R., Kennel, T., McMichael, J.P., McPhee, C.B., Montaquila, J., Yancey, T. et Zukerberg, A.L. (2016). *Address-Based Sampling*. American Association of Public Opinion Research, Alexandrie, Virginie. https://aapor.org/publications-resources/reports/.

Haziza, D., et Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32(2), 206-226.

Haziza, D., et Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1), 129-145.

Heckathorn, D.D., et Cameron, C.J. (2017). Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual Review of Sociology*, 43, 101-119.

Hibben, K.C., Smith, Z., Hoppe, T., Ryan, V., Rogers, B., Scanlon, P. et Miller, K. (2022). Toward a semi-automated item nonresponse detector model for open-response data. Document présenté à la Federal Committee on Statistical Methodology Research and Policy Conference, https://www.fcsm.gov/assets/files/docs/2022-conference-docs/D1.2_Cibelli.pdf.

Hidiroglou, M.A., et You, Y. (2016). Comparaison d'estimateurs sur petits domaines au niveau de l'unité et au niveau du domaine. *Techniques d'enquête*, 42(1), 45-67. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2016001/article/14540-fra.pdf.

Hyman, M., Sartore, L. et Young, L.J. (2022). Capture–recapture estimation of characteristics of US local food farms using a web-scraped list frame. *Journal of Survey Statistics and Methodology*, 10(4), 979-1004.

Ikudo, A., Lane, J.I., Staudt, J. et Weinberg, B.A. (2019). Occupational classifications: A machine learning approach. *Journal of Economic and Social Measurement*, 44(2-3), 57-87.

Kalton, G. (2019). Developments in survey research over the past 60 years: A personal perspective. *Revue Internationale de Statistique*, 87 (S1), S10-S30.

Kang, J.D.Y., et Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22 (4), 523-539.

Kaputa, S.J., Morris, D.S. et Holan, S.H. (2024). Bayesian multisource hierarchical models with applications to the Monthly Retail Trade Survey. *Journal of Survey Statistics and Methodology*, 12(5), 1567-1589.

Kaputa, S.J., Thompson, K.J. et Beck, J.L. (2019). An embedded experiment for targeted non-response follow-up in establishment surveys. *Journal of the Royal Statistical Society Series A*, 182(4), 1371-1391.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society Series C*, 29(2), 119-127.

Keiding, N., et Louis, T.A. (2018). Web-based enrollment and other types of self-selection in surveys and studies: Consequences for generalizability. *Annual Review of Statistics and its Application*, 5, 25-47.

Kennedy, C., et Hartig, H. (2019). Response rates in telephone surveys have resumed their decline. http://www.pewresearch.org/fact-tank/2019/02/27/responserates-in-telephone-surveys-have-resumed-their-decline/.

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

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

Kern, C., Weiß, B. et Kolb, J.-P. (2021). Predicting nonresponse in future waves of a probability-based mixed-mode panel with machine learning. *Journal of Survey Statistics and Methodology*, 11(1), 100-123.

Kim, J.K., et Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35(4), 501-514.

Kim, J.K., et Shao, J. (2022). *Statistical Methods for Handling Incomplete Data, Second Edition*. Boca Raton, Floride: CRC Press.

Kim, J.K., et Wu, C. (2013). Estimation parcimonieuse et efficace de la variance par rééchantillonnage pour les enquêtes complexes. *Techniques d'enquête*, 39(1), 105-137. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013001/article/11826-fra.pdf.

Kim, J.K., et Yu, C.L. (2011). Estimation de la variance par répliques sous échantillonnage à deux phases. *Techniques d'enquête*, 37(1), 73-81. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11448-fra.pdf.

Kott, P.S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89(426), 693-696.

Kott, P.S., et Stukel, D.M. (1997). La méthode du jackknife convient-elle à un échantillon à deux phases ? *Techniques d'enquête*, 23(2), 89-98. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997002/article/3621-fra.pdf.

Krennmair, P., et Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society Series C*, 71(5), 1865-1894.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9(5), 1010-1019.

Lavrakas, P.J., Traugott, M.W., Kennedy, C., Holbrook, A.L., de Leeuw, E.D. et West, B.T. (2019). *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*. Hoboken, New Jersey: Wiley.

Leuenberger, M., Eustache, E., Jauslin, R. et Tillé, Y. (2022). Balancing a sample almost perfectly. *Statistics and Probability Letters*, 180, 109229.

Lin, W.-C., et Tsai, C.-F. (2020). Missing value imputation: A review and analysis of the literature (2006-2017). *Artificial Intelligence Review*, 53, 1487-1509.

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

Lohr, S.L., Hsu, V. et Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2071-2085. Alexandrie, Virginie.

Lohr, S.L., et Mendez, G. (2009). Regression trees with survey data. Document présenté le 6 août 2009 aux Joint Statistical Meetings, Washington, DC.

Lohr, S.L., et Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2), 293-312.

Lumley, T., et Huang, X. (2024). Linear mixed models for complex survey data: Implementing and evaluating pairwise likelihood. *Stat*, 13(1), e657.

Lundy, E.R., et Rao, J.N.K. (2022). Efficacité relative des méthodes fondées sur l'estimation par régression d'enquête assistée par un modèle : une étude par simulations. *Techniques d'enquête*, 48(1), 53-78. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022001/article/00003-fra.pdf.

Mahalanobis, P.C. (1939). A sample survey of the acreage under jute in Bengal. *Sankhyā: The Indian Journal of Statistics*, 4(4), 511-531.

Mashreghi, Z., Haziza, D. et Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1-52.

McCarthy, P.J. (1966). *Replication: An Approach to the Analysis of Data from Complex Surveys*. Washington, DC: National Center for Health Statistics.

McClellan, C., Mitchell, E., Anderson, J. et Zuvekas, S. (2023). Using machine-learning algorithms to improve imputation in the Medical Expenditure Panel Survey. *Health Services Research*, 58(2), 423-432.

McConville, K.S., Breidt, F.J., Lee, T.C. et Moisen, G.G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5(2), 131-158.

McConville, K.S., et Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2), 389-413.

Mercer, A., Lau, A. et Kennedy, C. (2018). *For Weighting Online Opt-in Samples, What Matters Most?* Washington, DC: Pew Research. https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

Molina, I., et Rao, J.N.K. (2023). Historical overview of small area estimation in the 50th birthday of the IASS. *The Survey Statistician*, 88, 23-35.

Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472), 1429-1442.

Montanari, G.E., et Ranalli, M.G. (2009). Multiple and ridge model calibration. *Proceedings of Workshop on Calibration and Estimation in Surveys*. Ottawa, ON: Statistique Canada.

Morales, D., Esteban, M.D., Pérez, A. et Hobza, T. (2021). *A Course on Small Area Estimation and Mixed Models*. Cham, Suisse: Springer.

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

National Academies of Sciences, Engineering, and Medicine (2023). *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine (2024). *Toward a 21st Century National Data Infrastructure: Managing Privacy and Confidentiality Risks with Blended Data*. Washington, DC: The National Academies Press.

National Center for Health Statistics (1968). *Synthetic State Estimates of Disability: Public Health Service Publication No. 1759*. Washington, DC: U.S. Government Printing Office.

Oberski, D.L., et DeCastellarnau, A. (2021). Predicting measurement error variance in social surveys. https://repositori.upf.edu/bitstream/handle/10230/48430/WP63_RECSM_sept2021.pdf.

Olson, K., Smyth, J.D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N.A., McCarthy, J.S., O'Brien, E., Opsomer, J.D., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z.T., Turakhia, C. et Wagner, J. (2021). Transitions from telephone surveys to self-administered and mixed-mode surveys: AAPOR task force report. *Journal of Survey Statistics and Methodology*, 9(3), 381-411.

Opsomer, J.D., Breidt, F.J., White, M. et Li, Y. (2016). Successive difference replication variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology*, 4(1), 43-70.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. et Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society Series B*, 70(1), 265-286.

Opsomer, J.D., et Erciulescu, A.L. (2021). Estimation de la variance par répliques après calage fondé sur l'échantillon. *Techniques d'enquête*, 47(2), 287-301. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00006-fra.pdf.

Opsomer, J.D., et Riddles, M.K. (2025). Approche CHAID pour les enquêtes : un outil pour construire des cellules d'ajustement pour la non-réponse dans un cadre fondé sur le plan. *Techniques d'enquête*, 51(1), 235-252. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2025001/article/00006-fra.pdf.

Pfeffermann, D., et Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102(480), 1427-1439.

Phipps, P., et Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, 6(2), 772-794.

Platek, R. (1999). Techniques d'enquête – 25 ans d'histoire. *Techniques d'enquête*, 25(2), 123-125. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1999002/article/4874-fra.pdf.

Rao, J.N.K. (1999). Some current trends in sample survey theory and methods. *Sankhyā: The Indian Journal of Statistics, Series B*, 61(1), 1-25.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rao, J.N.K. (2005). Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage. *Techniques d'enquête*, 31(2), 127-151. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005002/article/9040-fra.pdf.

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

Rao, J.N.K., et Fuller, W.A. (2017). Théorie et méthodologie des enquêtes par sondage : orientations passées, présentes et futures. *Techniques d'enquête*, 43(2), 159-176. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2017002/article/54888-fra.pdf.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation, Second Edition*. Hoboken, New Jersey: Wiley.

Rao, J.N.K., et Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 57-65. Alexandrie, Virginie.

Rao, J.N.K., et Singh, A.C. (2009). Range restricted weight calibration for survey data using ridge regression. *Pakistan Journal of Statistics*, 25(4), 371-384.

Rao, J.N.K., Sinha, S.K. et Dumitrescu, L. (2014). Robust small area estimation under semi-parametric mixed models. *Canadian Journal of Statistics*, 42(1), 126-141.

Rao, J.N.K., et Wu, C.F.J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80(391), 620-630.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

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

Reiter, J.P. (2021). Assessing uncertainty when using linked administrative records. Dans *Administrative Records for Survey Methodology*, (Éds., A.Y. Chun, M.D. Larsen, G. Durrant et J.P. Reiter), 139-153. Hoboken, New Jersey: Wiley.

Rizzo, L., Kalton, G. et Brick, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22(1), 43-53. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1996001/article/14386-fra.pdf.

Roberson, A. (2021). Applying machine learning for automatic product categorization. *Journal of Official Statistics*, 37(2), 395-410.

Robertson, B., et Price, C. (2024). One point per cluster spatially balanced sampling. *Computational Statistics and Data Analysis*, 191, 107888.

Royall, R.M., et Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68(344), 880-889.

Särndal, C.-E. (2007). La méthode de calage dans la théorie et la pratique des enquêtes. *Techniques d'enquête*, 33(2), 113-135. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2007002/article/10488-fra.pdf.

Savitsky, T.D., Williams, M.R., Gershunskaya, J. et Beresovsky, V. (2023). Methods for combining probability and nonprobability samples under unknown overlaps. *Statistics in Transition*, 24(5), 1-34.

Schouten, B., Peytchev, A. et Wagner, J. (2018). *Adaptive Survey Design*. Boca Raton, Floride: CRC Press.

Sugasawa, S., et Kubokawa, T. (2020). Small area estimation with mixed models: A review. *Japanese Journal of Statistics and Data Science*, 3, 693-720.

Sun, H., Berg, E. et Zhu, Z. (2024). Multivariate small-area estimation for mixed-type response variables with item nonresponse. *Journal of Survey Statistics and Methodology*, 12(2), 320-342.

Thompson, M.E. (2019). Combining data from new and traditional sources in population surveys. *Revue Internationale de Statistique*, 87(S1), S79-S89.

Thompson, S.K. (2017). Adaptive and network sampling for inference and interventions in changing populations. *Journal of Survey Statistics and Methodology*, 5(1), 1-21.

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations*. Hoboken, New Jersey: Wiley.

Tourangeau, R., Brick, J.M., Lohr, S.L. et Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society Series A*, 180(1), 203-223.

United Nations Economic Commission for Europe (2021). *Machine Learning for Official Statistics*. Genève: Nations Unies. https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf.

Valliant, R. (2024). Hansen Lecture 2022: The evolution of the use of models in survey sampling. *Journal of Survey Statistics and Methodology*, 12(2), 275-304.

Verret, F., Rao, J.N.K. et Hidiroglou, M.A. (2015). Estimation sur petits domaines fondée sur un modèle sous échantillonnage informatif. *Techniques d'enquête*, 41(2), 353-368. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015002/article/14248-fra.pdf.

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

Winkler, W.E. (2021). Cleaning and using administrative lists: Enhanced practices and computational algorithms for record linkage and modeling/editing/imputation. Dans *Administrative Records for Survey Methodology*, (Éds., A.Y. Chun, M.D. Larsen, G. Durrant et J.P. Reiter), 105-138. Hoboken, New Jersey: Wiley.

Wiśniowski, A., Sakshaug, J.W., Perez Ruiz, D.A. et Blom, A.G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1), 120-147.

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

Wu, C. (2023). Calibration techniques for model-based prediction and doubly robust estimation. *The Survey Statistician*, 88, 86-93.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185-193.

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.

Yi, G.Y., Rao, J.N.K. et Li, H. (2016). A weighted composite likelihood approach for analysis of survey data under two-level models. *Statistica Sinica*, 26(2), 569-587.

Yung, W., Karkimaa, J., Scannapieco, M., Barcarolli, G., Zardetto, D., Sanchez, J.A.R., Braaksma, B., Buelens, B. et Burger, J. (2018). *The Use of Machine Learning in Official Statistics*. https://unece.org/sites/default/files/2024-07/HLGMOS The use of machine learning in official statistics.pdf.

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

Zhang, S., et Wagner, J. (2024). The additional effects of adaptive survey design beyond post-survey adjustment: An experimental evaluation. *Sociological Methods & Research*, 53(3), 1350-1383.