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

Techniques d’enquête

Techniques d’enquête
47-2

Date de diffusion : le 6 janvier 2022

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Programme des services de dépôt

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2022

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

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête

Nº 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Décembre 2021

•

Volume 47

•

Numéro 2

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

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

COMITÉ DE DIRECTION

Président E. Rancourt

Anciens présidents C. Julien (2013-2018)
J. Kovar (2009-2013)
D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Membres J.-F. Beaumont
S. Fortier (Gestionnaire de la production)
D. Haziza
J. Keenan
W. Yung

COMITÉ DE RÉDACTION

Rédacteur en chef J.-F. Beaumont, *Statistique Canada*

Anciens rédacteurs en chef W. Yung (2016-2020)
M.A. Hidiroglou (2010-2015)
J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Rédacteurs associés

- J.M. Brick, *Westat Inc.*
- S. Cai, *Carleton University*
- P.J. Cantwell, *U.S. Census Bureau*
- G. Chauvet, *École nationale de la statistique et de l'analyse de l'information*
- S. Chen, *University of Oklahoma Health Sciences Center*
- J. Chipperfield, *Australian Bureau of Statistics*
- J. Dever, *RTI International*
- J.L. Eltinge, *U.S. Bureau of Labor Statistics*
- W.A. Fuller, *Iowa State University*
- D. Haziza, *University of Ottawa*
- M.A. Hidiroglou, *Statistique Canada*
- B. Hulliger, *University of Applied and Arts Sciences Northwestern, Switzerland*
- D. Judkins, *ABT Associates Inc Bethesda*
- J.K. Kim, *Iowa State University*
- P.S. Kott, *RTI International*
- P. Lahiri, *University of Maryland*
- E. Lesage, *L'Institut national de la statistique et des études économiques*
- K. McConville, *Reed College*
- I. Molina, *Universidad Carlos III de Madrid*
- J. Opsomer, *Westat Inc*
- D. Pfeffermann, *University of Southampton*
- J.N.K. Rao, *Carleton University*
- L.-P. Rivest, *Université Laval*
- F.J. Scheuren, *National Opinion Research Center*
- P.L.d.N. Silva, *Escola Nacional de Ciências Estatísticas*
- P. Smith, *University of Southampton*
- D. Steel, *University of Wollongong*
- M. Torabi, *University of Manitoba*
- D. Toth, *U.S. Bureau of Labor Statistics*
- J. van den Brakel, *Statistics Netherlands*
- C. Wu, *University of Waterloo*
- W. Yung, *Statistique Canada*
- L.-C. Zhang, *University of Southampton*

Rédacteurs adjoints C. Bocci, K. Bosa, C. Boulet, S. Matthews, C.O. Nambeu et Y. You, *Statistique Canada*

POLITIQUE DE RÉDACTION

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

Présentation de textes pour la revue

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

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 47, numéro 2, décembre 2021

Table des matières

Article sollicité Waksberg

Sharon L. Lohr
Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples.....247

Articles réguliers

Jean D. Opsomer et Andreea L. Erciulescu
Estimation de la variance par répliques après calage fondé sur l'échantillon.....287

Éric Lesage, Jean-François Beaumont et Cynthia Bocci
Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu du modèle de Fay-Herriot.....303

Abel Dasylva et Arthur Goussanou
Estimation des faux négatifs attribuables à la création des pochettes dans le couplage d'enregistrements.....325

Pascal Bessonneau, Gwennaëlle Brilhaut, Guillaume Chauvet et Cédric Garcia
Estimation de la variance par le bootstrap avec remise pour les enquêtes auprès des ménages Principes, exemples et mise en œuvre.....339

Communications brèves

Phillip S. Kott et Dan Liao
Utilisation d'un autre estimateur jackknife de la variance aux fins du calage des poids pour tenir compte de la non-réponse totale dans une enquête complexe.....377

Yong You
Estimation sur petits domaines à l'aide du modèle au niveau de domaine de Fay-Herriot avec lissage et modélisation de variance d'échantillonnage.....389

Omar De La Riva Torres, Gonzalo Pérez-de-la-Cruz et Guillermina Eslava-Gómez
Évaluer la couverture des intervalles de confiance en cas de non-réponse. Étude de cas sur la moyenne et les quantiles de revenu dans certaines municipalités de l'Enquête intercensitaire mexicaine de 2015.....401

Remerciements.....413
Annonces.....415
Autres revues.....417

Techniques d'enquête, décembre 2021
Vol. 47, nº2, pp. 245-246
Statistique Canada, nº 12-001-X au catalogue

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied en 2001 une série annuelle d'articles sollicités en l'honneur de Joseph Waksberg, en reconnaissance de ses contributions exceptionnelles à la statistique et méthodologie d'enquête. Chaque année, un éminent statisticien d'enquête est choisi par un comité de sélection composé de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*. Le statisticien sélectionné est invité à rédiger un article pour *Techniques d'enquête* qui passe en revue l'évolution et l'état actuel d'un thème important du domaine de la statistique et méthodologie d'enquête. L'article reflète le mélange de théorie et de pratique caractéristique des travaux de Joseph Waksberg. Le lauréat du prix Waksberg est également invité à présenter la communication sollicitée Waksberg, habituellement au Symposium de Statistique Canada, et reçoit une prime en argent.

Veuillez consulter la section annonces à la fin de la revue pour des informations sur le processus de nomination et de sélection du prix Waksberg 2023.

Ce numéro de *Techniques d'enquête* commence par le 21e article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé de Bob Fay (président), Jean Opsomer, Jack Gambino et Elizabeth Stuart, d'avoir choisi Sharon Lohr comme auteure de l'article du prix Waksberg de 2021.

Communication sollicitée pour le prix Waksberg 2021

Auteure : Sharon Lohr

Sharon Lohr est l'auteure de nombreuses publications sur l'échantillonnage, la planification des expériences et les méthodes statistiques dans les domaines de l'éducation, des politiques publiques, du droit et de la criminalité. Elle a publié de multiples articles dans des revues de statistiques et est l'auteure des livres *Sampling: Design and Analysis*, qui en est à sa troisième édition, et *Measuring Crime: Behind the Statistics*. Anciennement professeure émérite de statistiques à l'Arizona State University et vice-présidente de Westat, elle est aujourd'hui consultante en statistiques et se consacre à la recherche et à la rédaction dans ce domaine.

Sharon est *fellow* de l'American Statistical Association et membre élue de l'Institut international de statistique. Elle est actuellement membre du Comité des statistiques nationales des National Academies of Sciences, Engineering, and Medicine des États-Unis. Elle a été la première lauréate du prix Gertrude M. Cox pour ses contributions à la pratique statistique, et a été choisie pour donner la conférence Morris Hansen à la Washington Statistical Society et la conférence Deming aux Joint Statistical Meetings.

Gagnants du prix Waksberg et leurs articles sollicités depuis 2001

2022 Roderick **Little**, Sujet de l'article à déterminer, (prévu pour le vol. 48, 2).

2021 Sharon **Lohr**, « Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples ». *Techniques d'enquête*, vol. 47, 2, 247-285.

2020 Roger **Tourangeau**, « Science et gestion d'enquête ». *Techniques d'enquête*, vol. 47, 1, 3-32.

2019 Chris **Skinner.**

2018 Jean-Claude **Deville**, « De la pratique à la théorie : l'exemple du calage à poids bornés ». 10ème Colloque francophone sur les sondages, Université Lumière Lyon 2.

2017 Donald **Rubin**, « Le calage conditionnel et le sage statisticien ». *Techniques d'enquête*, vol. 45, 2, 199-210.

2016 Don **Dillman**, « Inciter les participants aux enquêtes à mode mixte à répondre sur le Web : les promesses et les défis ». *Techniques d'enquête*, vol. 43, 1, 3-34.

2015 Robert **Groves**, « Vers un cadre de qualité pour les mélanges de données conçues et de données organiques ». Recueil : *Symposium 2016, Croissance de l'information statistique : défis et bénéfices.*

2014 Constance **Citro**, « Des modes multiples pour les enquêtes à des sources de données multiples pour les estimations ». *Techniques d'enquête*, vol. 40, 2, 151-181.

2013 Ken **Brewer**, « Trois controverses dans l'histoire de l'échantillonnage ». *Techniques d'enquête*, vol. 39, 2, 275-289.

2012 Lars **Lyberg**, « La qualité des enquêtes ». *Techniques d'enquête*, vol. 38, 2, 115-142.

2011 Danny **Pfeffermann**, « Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ? Comment le résoudre ? ». *Techniques d'enquête*, vol. 37, 2, 123-146.

2010 Ivan **Fellegi**, « L'organisation de la méthodologie statistique et de la recherche méthodologique dans les bureaux nationaux de la statistique ». *Techniques d'enquête*, vol. 36, 2, 131-139.

2009 Graham **Kalton**, « Méthodes de suréchantillonnage des sous-populations rares dans les enquêtes sociales ». *Techniques d'enquête*, vol. 35, 2, 133-152.

2008 Mary **Thompson**, « Enquêtes internationales : motifs et méthodologies ». *Techniques d'enquête*, vol. 34, 2, 145-157.

2007 Carl-Erik **Särndal**, « La méthode de calage dans la théorie et la pratique des enquêtes ». *Techniques d'enquête*, vol. 33, 2, 113-135.

2006 Alastair **Scott**, « Études cas-témoins basées sur la population ». *Techniques d'enquête*, vol. 32, 2, 137-147.

2005 J.N.K. **Rao**, « Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage ». *Techniques d'enquête*, vol. 31, 2, 127-151.

2004 Norman **Bradburn**, « Comprendre le processus de question et réponse ». *Techniques d'enquête*, vol. 30, 1, 5-16.

2003 David **Holt**, « Enjeux méthodologiques de l'élaboration et de l'utilisation d'indicateurs statistiques pour des fins de comparaisons internationales ». *Techniques d'enquête*, vol. 29, 1, 5-19.

2002 Wayne **Fuller**, « Estimation par régression appliquée à l'échantillonnage ». *Techniques d'enquête*, vol. 28, 1, 5-25.

2001 Gad **Nathan**, « Méthodes de téléenquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir ». *Techniques d'enquête*, vol. 27, 1, 7-34.

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 247-285
Statistique Canada, nº 12-001-X au catalogue

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

Sharon L. Lohr[1]

Résumé

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

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

1. Introduction

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

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

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

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

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

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

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

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

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

2. Structure et hypothèses classiques des enquêtes à bases multiples

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

2.1 NSAF (National Survey of America's Families)

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

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

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

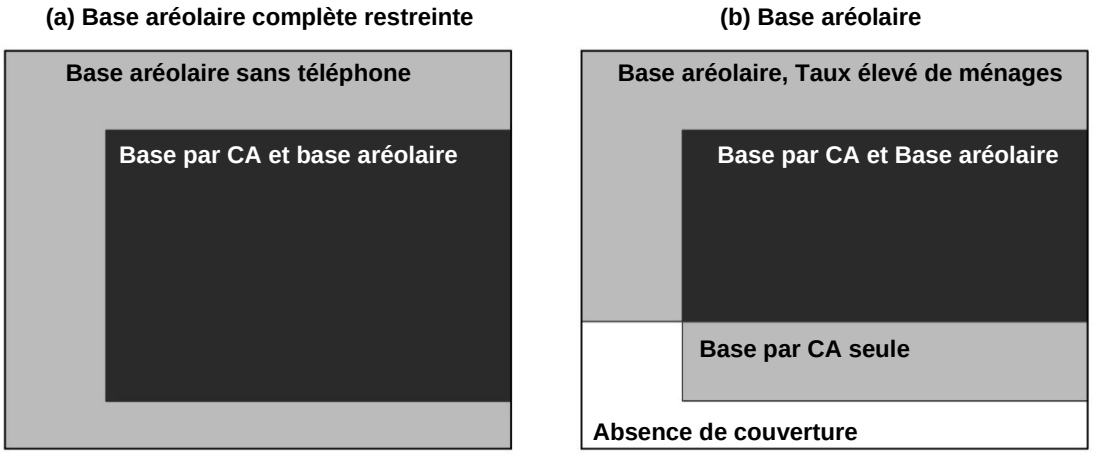
Ainsi, un échantillon de la base aréolaire donnerait une couverture élevée, mais entraînerait aussi des coûts élevés inacceptables. Une enquête par composition aléatoire coûterait moins, mais entraînerait un

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

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

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

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



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

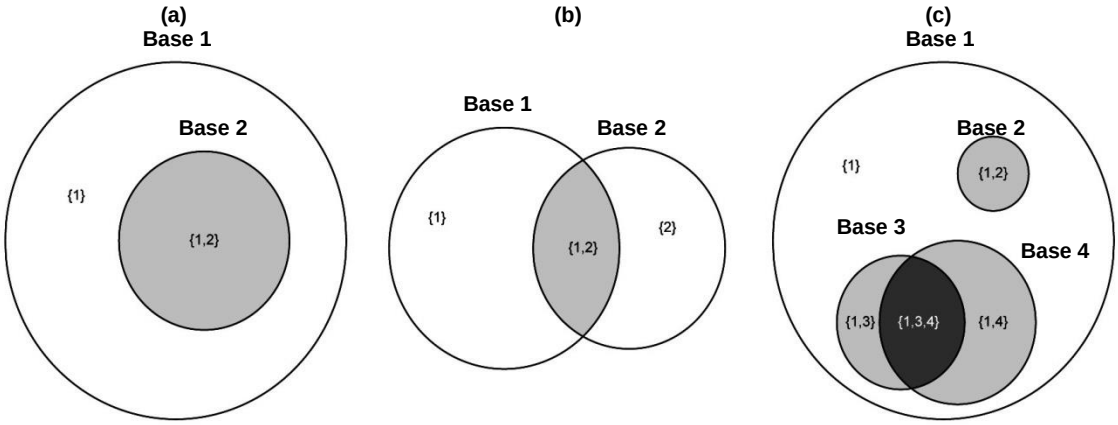
ménages ayant le téléphone ont été échantillonnés uniquement au moyen de la base de sondage par CA; les ménages de l'échantillon par CA sans enfants et situés au-dessus de 200 % du seuil de pauvreté ont été sous-échantillonnés. Comme une enquête de présélection a été utilisée, l'échantillon combiné des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

2.2 Notation et hypothèses pour les enquêtes à bases multiples

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3. Estimation dans les enquêtes classiques à bases multiples

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

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

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

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

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

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

3.1 Estimateur composite de Hartley

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2 Ajustements de la pondération pour la multiplicité

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

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

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

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

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

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

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

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

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

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

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

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

3.3. Calage

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

et

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

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

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

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

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

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

4. Enquêtes à bases multiples et intégration de données

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

4.1 Estimation sur petits domaines

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

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

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

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

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

4.2 Imputation massive et appariement d'échantillons

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.3 Imputation et NSAF

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

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

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

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

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

4.4 Classification erronée de domaine

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

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

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

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

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

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

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

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

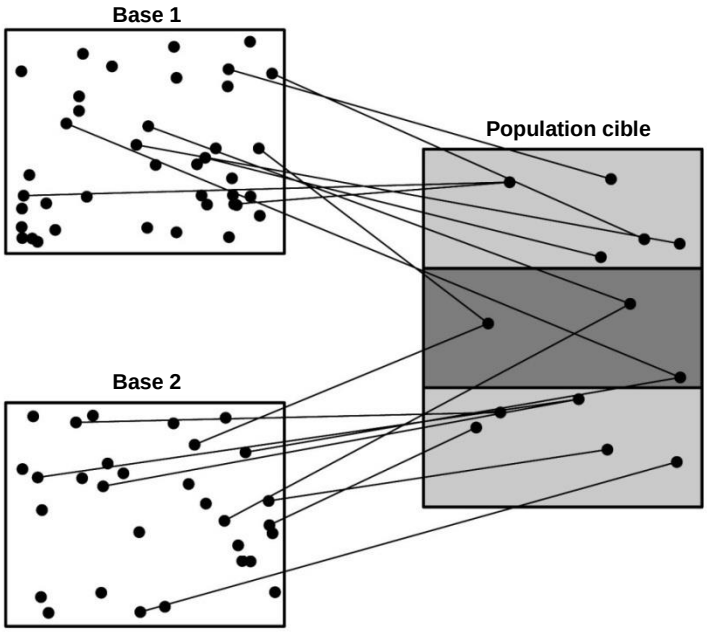
Dever (2018) a utilisé l'appariement d'échantillons pour évaluer le chevauchement de la base pour un échantillon probabiliste S_1, tiré d'une base de sondage fondée sur l'adresse, et un échantillon non probabiliste S_2 recruté à partir de sites de médias sociaux. Elle s'est intéressée au pourcentage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

4.5 Sondage indirect et estimation de la saisie-ressaisie

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

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

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



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

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

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

où

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

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

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

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

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

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

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

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

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

5. Conception de systèmes de collecte de données

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

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

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

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

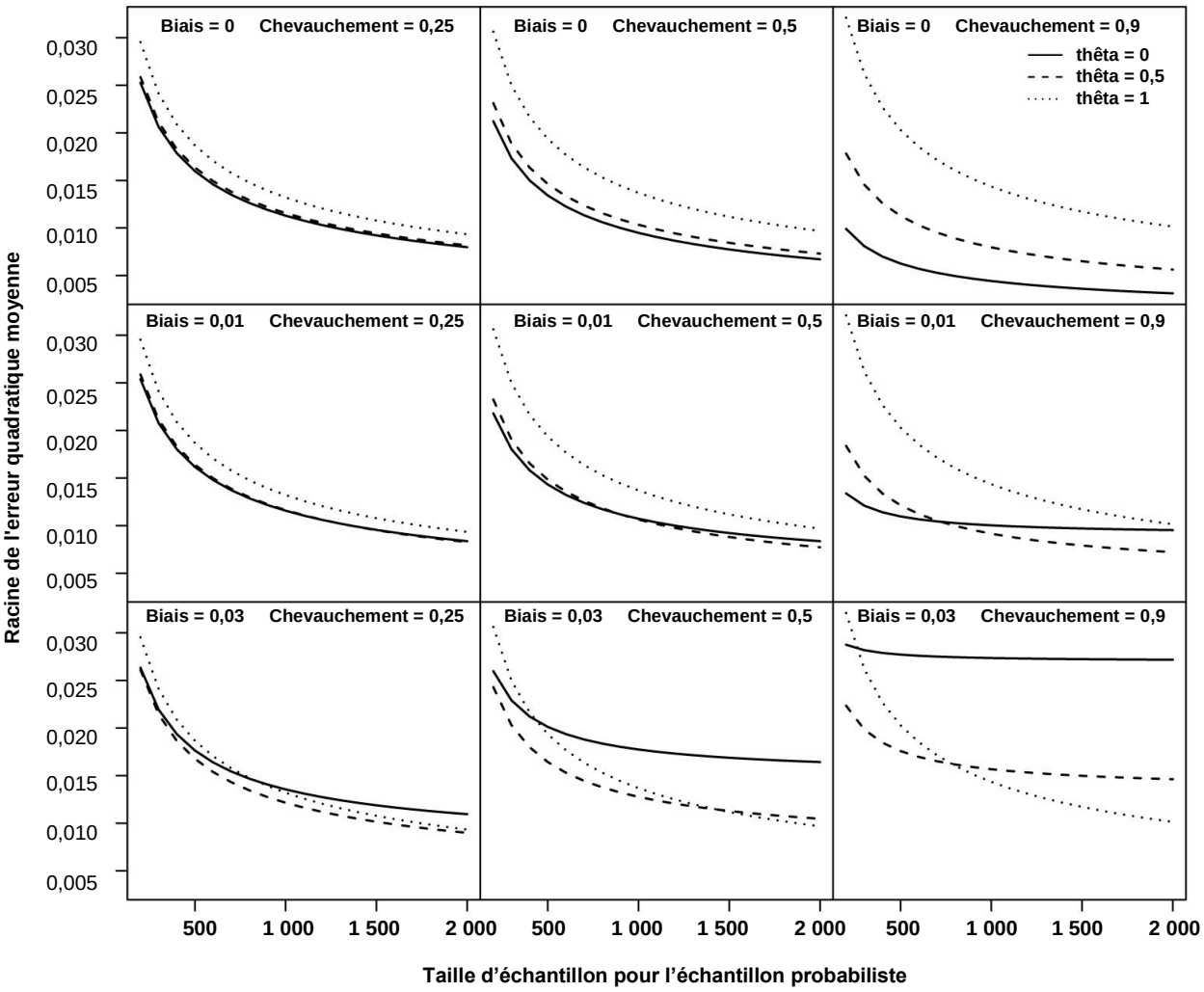
Un léger biais peut avoir un effet semblable pour la situation examinée à la section 3.4, où un recensement est tiré de la base de sondage 2 incomplète et un échantillon probabiliste de grande qualité est tiré de la base 1 complète. Les graphiques de la figure 5.1 montrent la racine de l'erreur quadratique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

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

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Remerciements

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

Bibliographie

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 287-301
Statistique Canada, nº 12-001-X au catalogue

Estimation de la variance par répliques après calage fondé sur l'échantillon

Jean D. Opsomer et Andreea L. Erciulescu[1]

Résumé

Le calage fondé sur l'échantillon se produit quand les poids d'une enquête sont calés pour contrôler les totaux aléatoires, au lieu de représenter les totaux fixes au niveau de la population. Les totaux de contrôle peuvent être estimés à partir de différentes phases de la même enquête ou d'une autre enquête. En cas de calage fondé sur l'échantillon, pour que l'estimation de la variance soit valide, il est nécessaire de tenir compte de la contribution de l'erreur due à l'estimation des totaux de contrôle. Nous proposons une nouvelle méthode d'estimation de la variance qui utilise directement les poids de rééchantillonnage de deux enquêtes, dont une sert à fournir des totaux de contrôle pour le calage des autres poids d'enquête. Aucune restriction n'est établie quant à la nature des deux méthodes de rééchantillonnage et il n'est pas nécessaire de calculer d'estimation de la variance-covariance, ce qui simplifie la mise en œuvre pratique de la méthode proposée. Nous fournissons la description générale de la méthode utilisée pour les enquêtes comportant deux méthodes de rééchantillonnage arbitraire avec un nombre de répliques différent. Il est démontré que l'estimateur de la variance obtenu est convergent pour la variance asymptotique de l'estimateur calé, quand le calage est effectué au moyen de l'estimation par la régression ou la méthode itérative du quotient (*raking*). La méthode est illustrée dans une application réelle, dans laquelle il faut harmoniser la composition démographique de deux enquêtes pour améliorer la comparabilité des estimations de l'enquête.

Mots-clés : Pêche; enquêtes sur la chasse et l'observation de la faune; méthode itérative du quotient; estimation par la régression; construction de réplique.

1. Introduction

Les méthodes d'estimation de la variance pour les enquêtes complexes comprennent des méthodes par linéarisation et rééchantillonnage. Les méthodes de rééchantillonnage présentent plusieurs avantages pratiques, notamment : les ajustements de poids multiples, comme les ajustements pour compenser la non-réponse et le calage, sont facilement intégrés aux estimations; il n'est pas nécessaire de publier l'information du plan détaillée dans les ensembles de données publics; et les utilisateurs des données peuvent facilement obtenir des estimations de la variance pour de grandes catégories d'estimateurs sans besoin de dérivations. De nombreuses méthodes de rééchantillonnage ont cours, le choix approprié de la méthode étant dicté par le plan de sondage et les objectifs d'estimation de l'enquête. Nous renvoyons à Wolter (2007) pour un aperçu des types de méthodes de rééchantillonnage de l'estimation de la variance.

Dans le présent article, nous traitons de la façon d'intégrer le calage dans l'estimation de la variance par répliques, quand les totaux de contrôle du calage sont eux-mêmes aléatoires et que leur variance est également estimée par une méthode de rééchantillonnage. Nous (les auteurs) avons rencontré ce problème alors que nous travaillions sur deux enquêtes – portant sur le même sujet et la même population cible – pour lesquelles nous avions été chargés de produire un ensemble unifié d'estimations.

La première enquête est la *National Survey of Fishing, Hunting, and Wildlife-Associated Recreation* de 2016 (FHWAR, Enquête nationale sur la pêche, la chasse et les loisirs associés à la faune). Cette enquête,

1. Jean D. Opsomer, Westat, Inc., 1600 Research Blvd., Rockville, MD, États-Unis, 20850. Courriel : JeanOpsomer@westat.com; Andreea L. Erciulescu, 1600 Research Blvd., Rockville, MD, États-Unis, 20850. Courriel : AndreeaErciulescu@westat.com.

menée par le Bureau du recensement des États-Unis, utilisait la méthode des répliques des différences successives (RDS), qui est une variante de la méthode par répliques répétées équilibrées (RRE). Proposée pour la première fois par Fay et Train (1995), la RRE est fréquemment utilisée dans les enquêtes du Bureau du recensement des États-Unis. La deuxième enquête est la *50-state Survey of FHWAR* de 2016 (Enquête sur les 50 États de FHWAR), réalisée par le Rockville Institute, la filiale sans but lucratif de Westat. Cette enquête utilisait la méthode du jackknife avec suppression de groupe (DAGJK) comme méthode de rééchantillonnage (Kott, 2001).

Les deux enquêtes FHWAR de 2016 ont été réalisées sur le terrain simultanément à l'aide de différents modes de collecte des données, précisément pour permettre la comparaison entre les deux et le rapprochement subséquent des estimations. L'enquête nationale utilisait une combinaison de collecte de données par téléphone et en personne et avait une taille d'échantillon suffisant à produire des estimations au niveau des divisions de recensement. L'enquête sur les 50 États était réalisée par la poste et, comme son nom l'indique, la taille de l'échantillon suffisait à produire des estimations au niveau de l'État. Toutefois, ces différences de mode, ainsi que d'autres différences, y compris d'autres aspects de la mise en œuvre de l'enquête, les stratégies de sous-échantillonnage et les méthodes d'estimation, ont conduit à des différences importantes et souvent statistiquement significatives dans les estimations, avec des estimations généralement plus élevées dans l'Enquête sur les 50 États que dans l'Enquête nationale. Pour en savoir plus sur les deux enquêtes FHWAR, voir Fish and Wildlife Service and Census Bureau (2018) et Rockville Institute (2018).

Comme nous l'avons indiqué plus haut, nous étions chargés d'élaborer une approche par calage pour « harmoniser » les estimations des deux enquêtes, dans le sens qu'il fallait produire des estimations au niveau des États à partir de l'enquête sur les 50 États qui soient compatibles avec celles obtenues par l'enquête nationale. Cela devait permettre de comparer les estimations à l'échelle des États de 2016 à celles des itérations antérieures de l'enquête nationale, qui a été menée depuis 1955 et dont les résultats d'enquête sont directement comparables depuis 1991. L'une des principales étapes du rapprochement des estimations consistait à caler la composition démographique de l'enquête sur les 50 États sur celle de l'enquête nationale, sachant que cette dernière était considérée comme l'étalon-or dans cette application. À cette fin, un ensemble d'estimations démographiques de l'enquête nationale ont été utilisées comme totaux de contrôle pour le calage de l'enquête sur les 50 États. Étant donné que ces totaux de contrôle sont eux-mêmes des estimations, il était toutefois nécessaire de s'assurer que leur variabilité est traduite dans les estimations de la variance des estimations calées de l'enquête sur les 50 États. Il s'agit d'une application *du calage fondé sur l'échantillon* (calage sur les totaux de contrôle aléatoires). Le calage fondé sur l'échantillon est habituellement observé dans les enquêtes à plusieurs phases, où les échantillons et les méthodes d'estimation peuvent être coordonnés. Dans la configuration actuelle, les deux enquêtes sont indépendantes et ont deux ensembles de répliques créés au moyen de méthodes de rééchantillonnage différentes.

Il existe peu de publications traitant de la façon de tenir compte du calage fondé sur l'échantillon dans l'estimation de la variance par répliques. Fuller (1998) a mis au point un estimateur de la variance par répliques pour les échantillons à deux phases, dans lequel les estimations de la phase deux sont calées sur les totaux de contrôle de la phase un. Dans cette approche, les répliques de la phase deux sont modifiées par des ajustements dérivés de la décomposition spectrale de la matrice de variance-covariance estimée de la phase un des totaux de contrôle. Dever et Valliant (2010) et Dever et Valliant (2016) ont étudié le calage des poids sur des totaux de contrôle estimés dans un scénario où une enquête (de référence) est utilisée pour caler une autre enquête (analytique), ce qui est plus étroitement lié à notre configuration. Dans le second article, leurs études par simulations ont été élaborées pour un estimateur par la régression généralisée, et les méthodes d'estimation de la variance par répliques par linéarisation et jackknife ont été comparées. Pour la réplique jackknife, les auteurs ont comparé les performances de l'ajustement de Fuller (1998) et deux ajustements basés sur des tirages d'une distribution normale multivariée, l'un utilisant la matrice de variance-covariance complète des totaux de contrôle, et l'autre utilisant seulement la diagonale de cette matrice. Cette dernière approche avait été proposée par Nadimpalli, Judkins et Chu (2004), mais aucune justification théorique n'avait été donnée. La méthode a été motivée par la prise en compte de la distribution asymptotique des totaux de contrôle estimés, qui est ensuite utilisée pour générer des versions « synthétiques » de ces estimations qui sont utilisées comme totaux de contrôle de rééchantillonnage.

Dans le présent article, nous décrivons une méthode visant à modifier les répliques de l'enquête qu'il faut caler en utilisant directement les répliques de l'enquête de contrôle. Nous montrons comment cette méthode peut être utilisée même si les méthodes de rééchantillonnage ou le nombre de répliques diffèrent entre les deux enquêtes. Il est intéressant de noter que Kott (2005) a mentionné brièvement une approche similaire qui utilise également les répliques directement, dans le cas particulier de deux enquêtes utilisant la méthode du jackknife avec suppression de groupe (DAGJK) avec le même nombre de répliques. Contrairement aux méthodes de Fuller (1998) et de Nadimpalli et coll. (2004), ces approches n'exigent pas le calcul explicite de la matrice de variance-covariance de l'enquête de contrôle, ce qui simplifie grandement la mise en œuvre en pratique. De plus, elles emploient des totaux calés valides, contrairement aux méthodes s'appuyant sur des tirages d'une distribution normale qui peuvent entraîner des totaux calés instables ou irréalisables.

De façon plus générale, les méthodes d'harmonisation des estimations de deux enquêtes peuvent être considérées comme une application de l'*intégration des données statistiques* (IDS), (Lahiri, 2020), qui consiste à utiliser un ensemble de méthodes pour combiner plusieurs sources de données en vue de créer des estimations améliorées ou nouvelles par rapport à ce qui peut être obtenu à partir des ensembles de données séparés. Bien qu'ils n'aient pas employé le terme IDS, Lohr et Raghunathan (2017) donnent un aperçu des outils de pointe permettant d'exécuter la plupart des activités d'IDS couramment rencontrées. Dans une application type de l'IDS, l'objectif est d'obtenir une combinaison optimale de l'information de plusieurs sources de données, ce qui implique presque toujours de créer un estimateur différent de ceux qui sont obtenus à partir de sources séparées. Les méthodes pour y parvenir peuvent être fondées sur le

plan de sondage, comme dans l'estimation à partir de bases de sondage multiples (Lohr et Rao, 2006) et l'estimation composite par régression (Merkouris, 2004), ou fondée sur un modèle (par exemple, Raghunathan, Xie, Schenker, Parsons, Davis, Dodd et Feuer, 2007). Le calage fondé sur un échantillon appartient à la catégorie fondée sur le plan de sondage, mais il vise également à reproduire exactement les estimations de l'une des sources de données.

Le reste de l'article est structuré comme suit. La méthode proposée est élaborée selon une configuration d'estimation par la régression à la section 2. La méthode itérative du quotient est une autre méthode de calage courante; étant donné qu'elle est utilisée pour les deux enquêtes d'intérêt, nous étendons les résultats à cette configuration dans la section 3. Dans la section 4, nous illustrons à la fois la méthode de Fuller (1998) et la méthode proposée au moyen des données des deux enquêtes FHWAR de 2016. La section 5 présente des conclusions générales.

2. Calage par régression fondé sur l'échantillon

Nous examinons une enquête sur une population U avec un échantillon s, des poids w_i, des variables cibles y_i. Pour un estimateur d'enquête donné $\hat{\theta}$ construit au moyen des poids w_i, l'inférence est effectuée par rééchantillonnage, mise en œuvre par la fourniture de R ensembles de poids de rééchantillonnage $w_i^{(r)}, r = 1, \ldots, R,$ et la formule d'estimation de la variance

$$\hat{V}(\hat{\theta}) = A\sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2, \tag{2.1}$$

où l'on calcule les $\hat{\theta}^{(r)}$ de la même manière que $\hat{\theta}$ mais en remplaçant w_i par $w_i^{(r)}$. La constante A dépend de la méthode de rééchantillonnage. Par souci de simplicité, nous nous concentrons sur ce qui suit de l'estimateur de Horvitz-Thompson de $t_y = \sum_U y_i$, noté $\hat{t}_y = \sum_s y_i/\pi_i$. Dans ce cas, de nombreuses méthodes de rééchantillonnage de forme (2.1) donnent un estimateur convergent par rapport au plan de $\mathrm{Var}(\hat{t}_y)$. Nous appellerons cette enquête « enquête primaire ».

Nous souhaitons créer des poids ajustés w_i^* qui sont calés sur un ensemble de totaux de contrôle d'une enquête secondaire de U avec un échantillon s_C, des poids w_{Ci}. Un estimateur de cette enquête est noté $\hat{\theta}_C$. Pour la deuxième enquête, un estimateur de la variance par répliques est également fourni,

$$\hat{V}_C(\hat{\theta}_C) = A_C\sum_{r=1}^{R_C} (\hat{\theta}_C^{(r)} - \hat{\theta}_C)^2,$$

avec des poids de rééchantillonnage $w_{Ci}^{(r)}, r = 1, \ldots, R_C$ et une constante propre au rééchantillonnage A_C. Les variables de contrôle seront notées $\mathbf{x}_i$, avec des totaux estimés $\hat{\mathbf{t}}_{Cx}$. En utilisant l'estimation par la régression comme cadre de calage, l'estimateur ajusté est

$$\hat{t}_{y,\mathrm{reg}} = \hat{t}_y + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x)^T \hat{\boldsymbol{\beta}} = \sum_s w_i^* y_i \tag{2.2}$$

où $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s^T \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{W}_s \mathbf{Y}_s$ avec $\mathbf{X}_s$ une matrice dont la i^{e} ligne est égale à $\mathbf{x}_i^T$, $\mathbf{W}_s$ une matrice diagonale avec la i^{e} entrée w_i et $\mathbf{Y}_s$ un vecteur contenant $y_i, i \in s.$ Ainsi, on peut écrire les poids calés comme étant

$$w_i^* = w_i \left(1 + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x)^T (\mathbf{X}_s^T \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{x}_i\right). \tag{2.3}$$

Nous notons que la post-stratification est un cas spécial d'estimation par la régression, voir Särndal, Swensson et Wretman (1992, chapitre 7.6).

Pour obtenir un estimateur de la variance, nous suivons la méthode classique de linéarisation pour les estimateurs par la régression par rapport au plan de sondage (voir par exemple, Särndal et coll., 1992, chapitre 5.5). Dans des conditions de régularité faible (comme la convergence par rapport au plan des estimateurs de Horvitz-Thompson et l'invertibilité des matrices requises), la version linéarisée de l'estimateur par la régression (2.2) est égale à l'estimateur par la différence,

$$\hat{t}_{y,\,\text{diff}} = \hat{t}_y + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x)^T \boldsymbol{\beta}_N = \hat{\mathbf{t}}_{Cx}^T \boldsymbol{\beta}_N + \hat{t}_e \tag{2.4}$$

où $\boldsymbol{\beta}_N = (\mathbf{X}_U^T \mathbf{X}_U)^{-1} \mathbf{X}_U^T \mathbf{Y}_U$ est la cible de population de $\hat{\boldsymbol{\beta}}$ et $\hat{t}_e = \sum_s w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_N).$ La variance de $\hat{t}_{y,\,\text{diff}}$ est égale à

$$\text{Var}(\hat{t}_{y,\text{diff}}) = \text{Var}(\hat{t}_e) + \boldsymbol{\beta}_N^T \,\text{Var}(\hat{\mathbf{t}}_{Cx})\, \boldsymbol{\beta}_N, \tag{2.5}$$

puisque les deux enquêtes sont indépendantes. Cette « variance linéarisée » est la variance de la distribution asymptotique de l'estimateur par la régression (2.2). Dans l'expression (2.5), le premier terme de la variance peut être estimé au moyen des répliques de l'enquête primaire, et la variance-covariance des totaux de contrôle du second terme peut être estimée au moyen des répliques de l'enquête secondaire. Par conséquent, l'estimateur par substitution (*plug-in*) de la variance

$$\tilde{V}(\hat{t}_{y,\text{reg}}) = V(\hat{t}_{\hat{e}}) + \hat{\boldsymbol{\beta}}^T \hat{V}_C(\hat{\mathbf{t}}_{Cx})\, \hat{\boldsymbol{\beta}}, \tag{2.6}$$

où $\hat{t}_{\hat{e}} = \sum_s w_i (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}),$ peut être utilisé pour une inférence asymptotique valide de $\hat{t}_{y,\text{reg}}$.

Cependant, il est souvent peu pratique de conserver les deux ensembles de données et les ensembles de répliques associées aux fins de l'estimation de la variance. Dans le contexte du calage d'enquête, l'organisme chargé de créer les poids ajustés pour l'enquête primaire préférerait souvent continuer de fournir son ensemble de données inchangé, sauf pour ce qui est des nouveaux poids calés et des poids de rééchantillonnage associés, afin que les utilisateurs de données puissent effectuer leurs analyses au moyen d'outils d'enquête classiques. C'est pourquoi il est intéressant de créer un seul ensemble de répliques pour l'enquête primaire, qui peut servir à estimer la variance, tout en tenant compte du fait que les totaux de contrôle sont eux-mêmes estimés à partir d'une enquête différente.

Nous proposons donc de construire de nouvelles répliques pour l'enquête primaire afin d'estimer (2.5). Supposons pour le moment que $R_C \le R.$ À partir des poids de rééchantillonnage $w_i^{(r)}$ pour l'estimateur de

la variance de l'enquête primaire, on obtient un estimateur de la variance par répliques du premier terme dans (2.6) en utilisant les poids de rééchantillonnage calés.

$$w_{1i}^{*(r)} = w_i^{(r)} \left(1 + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x^{(r)})^T (\mathbf{X}_s^T \mathbf{W}_s^{(r)} \mathbf{X}_s)^{-1} \mathbf{x}_i \right). \tag{2.7}$$

On obtient ces poids de rééchantillonnage en répétant le calage pour chacun des poids de rééchantillonnage $w_i^{(r)}$ et ils conduisent à une estimation de la variance convergente pour les estimateurs par la régression, comme nous l'avons vu pour le cas général dans Fuller (2009, chapitre 4). Voir aussi Valliant (1993) pour le cas particulier de la post-stratification.

On peut modifier les poids de rééchantillonnage $w_{1i}^{*(r)}$ pour saisir le deuxième terme dans (2.6) de la façon suivante :

$$w_i^{*(r)} = w_{1i}^{*(r)} + a_r w_i^{(r)} (\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})^T (\mathbf{X}_s^T \mathbf{W}_s^{(r)} \mathbf{X}_s)^{-1} \mathbf{x}_i, \tag{2.8}$$

avec les constantes a_r qui seront définies plus en détail ci-dessous. En combinant (2.7) et (2.8), les poids de rééchantillonnage obtenus sont

$$w_i^{*(r)} = w_i^{(r)} \left(1 + (\hat{\mathbf{t}}_{Cx} + a_r (\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx}) - \hat{\mathbf{t}}_x^{(r)})^T (\mathbf{X}_s^T \mathbf{W}_s^{(r)} \mathbf{X}_s)^{-1} \mathbf{x}_i \right). \tag{2.9}$$

On obtient de nouveau ces poids en appliquant le même calage que pour les poids d'origine à chacune des répliques, mais avec les totaux de contrôle de répliques $\hat{\mathbf{t}}_{Cx}^{*(r)} = \hat{\mathbf{t}}_{Cx} + a_r (\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})$. Les estimations par répliques qui en résultent sont

$$\hat{t}_{y,\text{reg}}^{(r)} = \sum_s w_i^{*(r)} y_i = \hat{t}_y^{(r)} + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x^{(r)})^T \hat{\boldsymbol{\beta}}^{(r)} + a_r (\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})^T \hat{\boldsymbol{\beta}}^{(r)}.$$

On choisit les constantes a_r de façon à tenir compte de la différence entre les méthodes de rééchantillonnage primaire et de contrôle, en particulier entre R_C et R et A_C et A, en supposant que

$$a_r = \begin{cases} \sqrt{\dfrac{A_C}{A}} & r = 1, \ldots, R_C \\ 0 & r = R_C + 1, \ldots, R. \end{cases} \tag{2.10}$$

Cela signifie que pour $r > R_C$, les poids de rééchantillonnage $w_i^{*(r)} = w_{1i}^{*(r)}$ dans (2.8), c'est-à-dire que les totaux de contrôle non ajustés sont utilisés pour caler les poids de rééchantillonnage. Aux seules fins de simplification de la notation, les valeurs a_r sont écrites avec les premières valeurs R_C non nulles. L'attribution des répliques de l'enquête de contrôle à celles de l'enquête primaire devrait être randomisée, afin de s'assurer que les estimateurs et les estimateurs par répliques des deux enquêtes restent indépendants, quelles que soient les méthodes de rééchantillonnage.

En utilisant les poids de rééchantillonnage (2.9) avec des constantes (2.10), l'estimateur de la variance par répliques (2.1) devient

$$\hat{V}(\hat{t}_{y,\mathrm{reg}}) = A\sum_{r=1}^{R}(\hat{t}_{y,\mathrm{reg}}^{(r)} - \hat{t}_{y,\mathrm{reg}})^2, \tag{2.11}$$

Si l'on ignore les termes d'un ordre plus petit ainsi que ceux avec $a_r = 0$, cela est à peu près égal à

$$\begin{aligned}\hat{V}(\hat{t}_{y,\mathrm{reg}}) &\approx A\sum_{r=1}^{R}(\hat{t}_{\hat{e}}^{(r)} - \hat{t}_{\hat{e}})^2 + \hat{\boldsymbol{\beta}}^T A_C \sum_{r=1}^{R_C}(\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})(\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})^T \hat{\boldsymbol{\beta}} \\ &+ A\sum_{r=1}^{R} a_r(\hat{t}_{\hat{e}}^{(r)} - \hat{t}_{\hat{e}})(\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})^T \hat{\boldsymbol{\beta}}.\end{aligned}$$

Le terme croisé est également d'un ordre plus petit en raison de l'indépendance des deux enquêtes et du fait que $\sum_{r=1}^{R}\hat{t}_{\hat{e}}^{(r)}/R \approx \hat{t}_{\hat{e}}$ et $\sum_{r=1}^{R_C}\hat{\mathbf{t}}_{Cx}^{(r)}/R_C \approx \hat{\mathbf{t}}_{Cx}$. Par conséquent, l'estimateur de la variance par répliques (2.11) hérite de la convergence par rapport au plan des méthodes de rééchantillonnage originales pour les deux enquêtes et est convergent par rapport au plan pour la variance linéarisée (2.5).

Enfin, discutons du cas où $R_C > R$. On peut facilement étendre l'approche décrite ci-dessus à ce cas en répétant les R répliques des K fois de l'enquête primaire, de sorte que $R_C \le KR$ avec K étant le plus petit nombre entier positif pour lequel cette inégalité est satisfaite. L'estimateur de la variance par répliques qui en résulte a la même forme que (2.1), mais R y est remplacé par KR et A y est remplacé par A/K. Ensuite, la méthode présentée plus haut s'applique directement à ce nouvel estimateur de la variance par répliques pour l'enquête primaire. Par exemple, si $R = 120$ et $R_C = 150$, chaque réplique de l'enquête primaire sera répétée $K = 2$ fois, ce qui donnera 240 répliques pour l'enquête primaire, dont 150 seront modifiées.

3. Calage par la méthode itérative du quotient fondé sur l'échantillon

Dans l'application aux deux enquêtes FHWAR de 2016, nous avons utilisé la méthode itérative du quotient plutôt que l'estimation par la régression pour le calage. Comme pour l'estimation par la régression, l'objectif est de créer de nouveaux contrôles par la méthode itérative du quotient pour chacune des répliques, de sorte que l'estimateur de la variance par répliques pour l'enquête principale tienne compte de la variabilité des totaux de contrôle de l'enquête secondaire. Les résultats ci-dessus de l'estimation par la régression ne s'appliquent pas directement, mais nous pouvons appliquer le même raisonnement que dans Deville et Särndal (1992) pour montrer qu'ils continuent de se vérifier pour la méthode itérative du quotient. Au lieu de nous fier à ce résultat général, nous le calculerons ici explicitement pour montrer comment obtenir les totaux de contrôle ajustés pour les répliques.

Par souci de simplicité, nous décrivons ici le cas où nous contrôlons les comptes marginaux dans les domaines définis par les niveaux de 2 variables catégoriques, notées a et b, ayant respectivement K et L niveaux. Dans l'enquête primaire, les comptes estimés dans les domaines définis par les intersections des deux variables sont $\hat{N}_{a_k b_l} = \sum_s w_i \delta_i(a_k, b_l)$, avec $\delta_i(a_k, b_l) = 1$ si l'élément i se trouve dans le domaine défini par l'intersection de a_k et b_l et 0 autrement. Les comptes marginaux estimés sont définis

de façon analogue, $\hat{N}_{a_k} = \sum_s w_i \delta_i(a_k, \cdot)$ et $\hat{N}_{b_l} = \sum_s w_i \delta_i(\cdot, b_l)$. Nous notons $\boldsymbol{\delta}_i = (\delta_i(a_1), \ldots, \delta_i(a_K), \delta_i(b_1), \ldots, \delta_i(b_L))^T$ le vecteur d'indicateurs pour les domaines marginaux de l'élément i. Les comptes marginaux estimés de l'enquête primaire sont $\hat{\mathbf{N}} = \sum_s w_i \boldsymbol{\delta}_i = (\hat{N}_{a_1}, \ldots, \hat{N}_{a_K}, \hat{N}_{b_1}, \ldots, \hat{N}_{b_L})^T$ et les totaux de contrôle correspondants de l'enquête secondaire sont $\hat{\mathbf{N}}_C = \sum_{s_C} w_{Ci} \boldsymbol{\delta}_i = (\hat{N}_{Ca_1}, \ldots, \hat{N}_{Ca_K}, \hat{N}_{Cb_1}, \ldots, \hat{N}_{Cb_L})^T$. En utilisant l'algorithme de ratissage croisé classique de Deming et Stephan (1940) jusqu'à la convergence, nous pouvons écrire les poids traités par la méthode itérative du quotient pour l'enquête primaire ainsi

$$\begin{aligned} w_i^* &= w_i \exp(\hat{u}_{a_k} + \hat{u}_{b_l}) \quad \text{pour} \quad \delta_i(a_k, b_l) = 1 \\ &= w_i \exp(\hat{\mathbf{u}}^T \boldsymbol{\delta}_i) \end{aligned} \tag{3.1}$$

où $\hat{\mathbf{u}} = (u_{a_1}, \ldots, u_{a_K}, u_{b_1}, \ldots, u_{b_L})^T$ est une solution au système de $K + L$ équations

$$\begin{aligned} \sum_{l=1}^{L} \hat{N}_{a_k b_l} \exp(u_{a_k} + u_{b_l}) &= \hat{N}_{Ca_k} \quad (k = 1, \ldots, K) \\ \sum_{k=1}^{K} \hat{N}_{a_k b_l} \exp(u_{a_k} + u_{b_l}) &= \hat{N}_{Cb_l} \quad (l = 1, \ldots, L). \end{aligned} \tag{3.2}$$

La solution à ces équations n'est pas unique, de sorte que l'une des inconnues peut être paramétrée à 0 et qu'une équation peut être supprimée. Cela n'affecte pas les valeurs de $\exp(u_{a_k} + u_{b_l})$, et nous allons définir $v_{b_L} = 0$ et supprimer la dernière équation dans ce qui suit.

Il n'y a pas d'expression explicite pour la solution à (3.2), mais on peut l'établir par approximation en utilisant un argument de linéarisation. Dans le cadre asymptotique habituel de l'enquête qui assure la convergence par rapport au plan des estimateurs de Horvitz-Thompson, les valeurs $\hat{u}_{a_k}$ et $\hat{u}_{b_l}$ convergent vers 0 à mesure que la taille des échantillons des deux enquêtes augmente, de sorte que la dilatation autour de 0 est valide. En le faisant pour les équations dans (3.2), nous calculons par approximation l'ensemble réduit de $K + L - 1$ équations par

$$\begin{aligned} \sum_{l=1}^{L} \hat{N}_{a_k b_l} (1 + u_{a_k} + u_{b_l} + o_p(1)) &= \hat{N}_{Ca_k} \quad (k = 1, \ldots, K) \\ \sum_{k=1}^{K} \hat{N}_{a_k b_l} (1 + u_{a_k} + u_{b_l} + o_p(1)) &= \hat{N}_{Cb_l} \quad (l = 1, \ldots, L-1), \end{aligned}$$

qui peut être réécrit comme étant

$$\begin{aligned} \hat{N}_{a_k} u_{a_k} + \sum_{l=1}^{L} \hat{N}_{a_k b_l} u_{b_l} &= (\hat{N}_{Ca_k} - \hat{N}_{a_k})(1 + o_p(1)) \quad (k = 1, \ldots, K) \\ \hat{N}_{b_l} u_{b_l} + \sum_{k=1}^{K} \hat{N}_{a_k b_l} u_{a_k} &= (\hat{N}_{Cb_l} - \hat{N}_{b_l})(1 + o_p(1)) \quad (l = 1, \ldots, L-1). \end{aligned} \tag{3.3}$$

Si l'on ignore les restes d'ordre plus petit, la solution à ce système d'équations linéaires peut être écrite sous la forme $\hat{\mathbf{u}} = \hat{\mathbf{J}}^{-1}(\hat{\mathbf{N}}_C - \hat{\mathbf{N}})$, où $\hat{\mathbf{J}}$ est une matrice symétrique $(K + L - 1) \times (K + L - 1)$ contenant les comptes de domaine estimés facilement obtenus à partir du premier membre de (3.3). Notons que le $\hat{\mathbf{u}}$

qu'on obtient n'est pas un estimateur linéaire, car dans la linéarisation nous avons conditionné sur $\hat{\mathbf{N}}$. Enfin, en insérant $\hat{\mathbf{u}}$ dans l'expression (3.1) et en réalisant une nouvelle linéarisation, nous obtenons

$$w_i^* \approx w_i \left(1 + (\hat{\mathbf{N}}_C - \hat{\mathbf{N}})^T \hat{\mathbf{J}}^{-1} \boldsymbol{\delta}_i\right) \tag{3.4}$$

et l'estimateur après l'exécution de la méthode itérative du quotient est

$$\hat{t}_{y,\text{rak}} = \sum_s w_i^* y_i \approx \sum_s w_i y_i + (\hat{\mathbf{N}}_C - \hat{\mathbf{N}})^T \hat{\mathbf{J}}^{-1} \boldsymbol{\delta}_s^T \mathbf{W}_s \mathbf{Y}_s. \tag{3.5}$$

Notons que la taille de la variable de contrôle $\boldsymbol{\delta}_i$ et des estimations associées est maintenant $K + L - 1$, mais nous conservons la notation précédente à des fins de simplicité. Dans (3.5), $\hat{\mathbf{J}}^{-1} \boldsymbol{\delta}_s^T \mathbf{W}_s \mathbf{Y}_s$ correspond à $\hat{\boldsymbol{\beta}}$ dans l'estimateur par la régression (2.2). Cette équivalence asymptotique entre l'estimateur par la méthode itérative du quotient et l'estimateur par la régression avec les mêmes totaux de contrôle a été établie par Deville et Särndal (1992). En particulier, ils fournissent les conditions suffisantes dans lesquelles la variance asymptotique de l'estimateur par la régression équivalent peut également être utilisée pour l'inférence de l'estimateur par la méthode itérative du quotient. On peut alors construire un estimateur de la variance ayant la forme (2.6), avec $\hat{t}_{\hat{e}} = \sum_s w_i (y_i - \boldsymbol{\delta}_i^T \hat{\mathbf{J}}^{-1} \boldsymbol{\delta}_s^T \mathbf{W}_s \mathbf{Y}_s)$.

Nous considérons maintenant la construction de poids de rééchantillonnage pour l'enquête primaire estimant la variance asymptotique de l'estimateur par la méthode itérative du quotient. Comme auparavant, nous construisons de nouveaux totaux de contrôle de répliques $\hat{\mathbf{N}}_C + a_r (\hat{\mathbf{N}}_C^{(r)} - \hat{\mathbf{N}}_C)$ au moyen des estimations par répliques de l'enquête secondaire. Chacun des ensembles de poids de rééchantillonnage $w_i^{(r)}$ de l'enquête primaire est ajusté par la méthode itérative du quotient sur son ensemble correspondant de totaux de contrôle, afin d'obtenir les $w_i^{*(r)}$ et les estimations par répliques par la méthode itérative du quotient. En utilisant l'approximation dans (3.5) pour chaque réplique, nous obtenons que l'estimateur de la variance par répliques résultant est convergent pour la variance asymptotique de l'estimateur par la méthode itérative du quotient.

4. Application

Nous revenons au problème de calage qui s'est posé quand nous avons voulu combler l'écart entre les deux enquêtes FHWAR de 2016. Dans les deux enquêtes, la population est définie comme étant les personnes âgées de 16 ans et plus vivant dans des ménages américains. Les principales sources de données de cette application sont les fichiers de données au niveau de l'enregistrement, contenant les poids et les poids de rééchantillonnage des deux enquêtes. Au moyen de ces ensembles de données, nous avons effectué une première analyse et déterminé les écarts dans les données démographiques, que nous avons ajustés par méthode itérative du quotient fondée sur l'échantillon. Les totaux de population estimés construits au moyen des données au niveau des enregistrements de l'enquête nationale ont été considérés comme des contrôles aléatoires, pour les tableaux croisés des divisions de recensement (neuf catégories) et chacune des variables démographiques suivantes :

- résidence : deux catégories correspondant à la classification urbaine et rurale;
- âge : huit catégories correspondant aux tranches d'âge 16-17; 18-24; 25-34; 35-44; 45-54; 55-64; 65-74 et 75+;
- sexe : deux catégories correspondant aux hommes et aux femmes;
- race-ethnicité : quatre catégories correspondant aux personnes hispaniques, aux personnes blanches non hispaniques, aux personnes afro-américaines non hispaniques et à toutes les autres personnes non hispaniques;
- revenu annuel : neuf catégories correspondant à des fourchettes de revenu de -20 000 $; de 20 000 $ à 29 999 $; de 30 000 $ à 39 999 $; de 40 000 $ à 49 999 $; de 50 000 $ à 74 999 $; de 75 000 $ à 99 999 $; de 100 000 $ à 149 999 $; de 150 000 $ et plus, et non déclaré.

Pour l'application examinée dans l'article, nous utilisons le fichier à grande diffusion de l'enquête sur les 50 États, qui ne contient pas d'information sur le revenu. Par conséquent, nous illustrons la méthode proposée dans une configuration légèrement simplifiée, en utilisant les tableaux croisés des divisions de recensement ainsi que la résidence, l'âge, le sexe et la race et l'origine ethnique comme dimensions de la méthode itérative du quotient. Nous avons mis en œuvre à la fois la méthode de Fuller (1998) et la méthode de calage proposée décrite à la section 2 au moyen des fichiers de données à grande diffusion disponibles pour les deux enquêtes. À des fins de comparaison, nous présentons également les résultats du calage sans ajustement des estimations de la variance pour les contrôles aléatoires, que l'on appelle ci-dessous la méthode « naïve » parce qu'elle ignore la variabilité des contrôles dans les estimations de la variance. Afin de comparer les méthodes d'estimation de la variance pour les variables d'enquête qui ne sont pas des variables de contrôle, nous montrerons également des estimations pour les domaines définis par des tableaux croisés de la résidence et du sexe.

Bien que les méthodes de rééchantillonnage des deux enquêtes FHWAR soient différentes, les deux utilisent des répliques $R = R_C = 160$. En ce qui concerne l'expression (2.1), la constante de rééchantillonnage pour la méthode DAGJK de l'enquête des 50 États est $A = 159/160$ et la constante correspondante pour la méthode RDS de l'enquête nationale est $A_C = 4/160$, toutes deux disponibles dans les documents de leur enquête respective. Par conséquent, les constantes d'ajustement de rééchantillonnage a_r dans (2.10) sont égales à $2/\sqrt{159}$ pour $r = 1, \ldots, R$.

Les estimations que nous prendrons en considération étant toutes des comptes de domaine estimés, nous définissons alors la variable cible $y_i = I_{\{i \in U_d\}}$ pour un domaine d'intérêt U_d. Pour les 144 domaines définis par les dimensions de la méthode itérative du quotient, nous écrivons les comptes de domaine estimés comme étant $\hat{t}_k = \sum_s w_i I_{\{i \in U_k\}}, k = 1, \ldots, 144$. De même, étant donné que les totaux de contrôle sont des comptes de domaine estimés à partir de l'enquête nationale, le vecteur de la variable auxiliaire est $\mathbf{x}_i = \mathbf{I}_i$, un vecteur de longueur 144 contenant les indicateurs pour l'inclusion du répondant i dans les

domaines de contrôle $U_k, k = 1,\ldots,144,$ et supposons $\hat{t}_{C,k} = \sum_{s_C} w_{Ci} I_{\{i \in U_k\}}$. Nous notons $\hat{\mathbf{t}}_C = (\hat{t}_{C,1},\ldots,\hat{t}_{C,144})^T$ le vecteur des totaux de contrôle et les totaux de contrôle des répliques ajustées sont $\hat{\mathbf{t}}_C^{*(r)} = \hat{\mathbf{t}}_C + 2\big/\sqrt{159}\,(\hat{\mathbf{t}}_C^{(r)} - \hat{\mathbf{t}}_C)$.

Afin de mettre en œuvre la méthode de Fuller (1998), nous avons estimé la matrice de variance-covariance des totaux de contrôle $\hat{V}_C(\hat{\mathbf{t}}_C)$ au moyen des poids de rééchantillonnage de l'enquête nationale. La décomposition spectrale de cette matrice a donné un ensemble de 144 vecteurs propres $\mathbf{q}_i$ et valeurs propres associées λ_i, pour $i = 1,\ldots,144.$ D'après Fuller (1998), nous obtenons un ensemble de 144 vecteurs $\mathbf{v}_i$ satisfaisant

$$\hat{V}_C(\hat{t}_C) = \sum_{i=1}^{144} \mathbf{v}_i' \mathbf{v}_i,$$

où $\mathbf{v}_i = \sqrt{\lambda_i}\,\mathbf{q}_i$, pour $i = 1,\ldots,144.$ Enfin, les contrôles de rééchantillonnage ajustés sont $\hat{\mathbf{t}}_C^{*(r)} = \hat{\mathbf{t}}_C + \frac{\sqrt{160}}{2}\mathbf{v}_r$ pour $r = 1,\ldots,144$ et $\hat{\mathbf{t}}_C^{*(r)} = \hat{\mathbf{t}}_C$ pour $r = 145,\ldots,160.$ Cela révèle un inconvénient de la méthode de Fuller (1998), car bien que notre approche perturbe les totaux de contrôle des 160 répliques, cette méthode ne perturbe qu'une fraction d'entre elles dans ce cas. De plus, 30 des 144 valeurs propres étaient presque nulles, et 18 d'entre elles étaient inférieures à zéro en raison d'une erreur de virgule flottante. Nous avons tronqué les 18 valeurs propres négatives à zéro, et laissé le reste inchangé. Par conséquent, dans la mesure où toutes les répliques ne contribuent pas aux estimations de la variance pour certaines estimations d'enquête (par exemple pour les totaux de domaine), le calage fondé sur l'échantillon est susceptible d'être traduit de façon imparfaite dans les estimations de la variance. En général, nous nous attendons à ce qu'un plus grand nombre de répliques soient perturbées au moyen de notre approche, puisque la matrice estimée de la variance-covariance des totaux de contrôle ne peut être estimée de façon fiable que si sa dimension est adéquatement plus petite que R_C.

Les tableaux 4.1 et 4.2 montrent les estimations et les erreurs-types, respectivement, pour les domaines définis selon la résidence et le sexe, avant et après le calage. Les quatre premières lignes contiennent les résultats des totaux marginaux pour les variables de la méthode itérative du quotient, qui sont calées exactement, tandis que les quatre dernières sont des totaux qui correspondent à l'intersection des dimensions de la méthode itérative du quotient et ne sont par conséquent pas calées exactement.

Les deux enquêtes sont représentatives de la même population cible, mais les estimations et les erreurs-types associées diffèrent, reflétant à la fois la variabilité de l'échantillonnage et les méthodes de calage différentes appliquées par les deux organismes d'enquête. Comme le confirme le tableau 4.1, une fois que l'enquête sur les 50 États est traitée par la méthode itérative du quotient dans l'enquête nationale, les totaux estimés pour les domaines définis comme des domaines de calage exact correspondent exactement entre les deux enquêtes. Pour les domaines définis par les tableaux croisés de la résidence et du sexe, les estimations traitées par la méthode itérative du quotient pour l'enquête sur les 50 États sont proches, mais ne sont pas identiques à celles de l'enquête nationale.

Tableau 4.1
Estimations de la population avant et après le calage, arrondies à l'entier le plus proche, après mise à l'échelle par 10^3

Domaine		Avant calage		Après calage
		50 États	Nationale	
Résidence :	Milieu urbain	203 445	208 695	208 695
	Milieu rural	51 511	45 991	45 991
Sexe :	Masculin	128 276	121 775	121 775
	Féminin	126 680	132 911	132 911
Milieu rural :	Masculin	99 547	98 511	98 089
	Féminin	103 898	110 184	110 607
Milieu urbain :	Masculin	28 729	23 264	23 686
	Féminin	22 782	22 727	22 305

Le tableau 4.2 montre les erreurs types obtenues par les deux méthodes de réplication avec les totaux de contrôle ajustés et par la méthode naïve, qui ne tient pas compte du caractère aléatoire des totaux de contrôle. En raison de leur construction, la méthode d'ajustement fondée sur des répliques et la méthode de Fuller (1998) donnent des estimations de la variance identiques pour les domaines utilisés dans le calage. Ces estimations de la variance sont aussi égales à celles de l'enquête de contrôle dans ce cas. Cela traduit le fait que la composante de la variance correspondant au premier terme de (2.5) est paramétrée à 0 pour les totaux de contrôle, tandis que la composante de la variance pour le deuxième terme est exactement égale à l'estimation de la variance de l'enquête de contrôle dans le cas de la méthode itérative du quotient. Comme cette composante de la variance est ignorée dans la méthode naïve, les estimations de la variance sont égales à zéro. Pour les totaux estimés des domaines définis comme le tableau croisé de la résidence et du sexe, les estimations de la variance des deux méthodes ne sont pas identiques, mais proches (à 8 % l'une de l'autre), ce qui reflète le fait que les deux méthodes sont convergentes pour la variance asymptotique (2.5). Les estimations de la variance selon la méthode naïve sont plus petites que les estimations de la variance selon les deux autres méthodes de calage, ce qui donne un résultat manifestement incorrect, faute de prise en compte de la variance dans les totaux de contrôle aléatoires. Pour les autres variables, on s'attend toujours à ce que la variance soit sous-estimée selon la méthode naïve, parce que le deuxième terme de la variance asymptotique (2.5) est ignoré.

Tableau 4.2
Erreurs-types des estimations de la population avant et après le calage, arrondies à l'entier le plus proche, après mise à l'échelle par 10^3

Domaine		Avant calage		Après calage		
		50 États	Nationale	Naïve	De Fuller	Proposée
Résidence :	Milieu urbain	1 922	2 664	0	2 664	2 664
	Milieu rural	1 922	2 598	0	2 598	2 598
Sexe :	Masculin	2 117	1 074	0	1 074	1 074
	Féminin	2 117	1 112	0	1 112	1 112
Milieu rural :	Masculin	2 118	1 399	853	1 658	1 533
	Féminin	2 514	1 797	853	1 964	1 970
Milieu urbain :	Masculin	1 595	1 449	853	1 709	1 641
	Féminin	979	1 271	853	1 470	1 547

5. Conclusions

Nous avons proposé une méthode pour tenir compte du calage fondé sur l'échantillon dans les estimations de la variance. La méthode s'applique aux situations dans lesquelles l'enquête qui est calée et l'enquête qui fournit les contrôles de calage utilisent l'estimation de la variance par répliques, comme cela est souvent le cas dans les enquêtes à grande échelle réalisées par l'État. Les méthodes de rééchantillonnage dans chaque cas sont arbitraires, dans la mesure où elles sont toutes deux valides pour leurs enquêtes. Nous avons décrit la méthode pour les cas de calage avec estimation par la régression (post-stratification comprise) et la méthode itérative du quotient, deux méthodes courantes dans la pratique, et nous pensons qu'elle devrait fonctionner de la même façon pour d'autres types d'estimation, comme les estimateurs de calage généraux de Deville et Särndal (1992).

La principale solution de rechange à la méthode proposée est celle de Fuller (1998). Par rapport à cette méthode, la nôtre présente un avantage important, à savoir qu'elle n'exige pas de calculer la matrice de variance-covariance estimée des totaux de contrôle, ce qui simplifie sa mise en œuvre. Dans une application typique où le nombre de totaux de contrôle est inférieur au nombre de répliques, un autre avantage potentiel de la méthode proposée est que les perturbations seront appliquées à une plus grande fraction des répliques. Cela réduit le risque de calculer des estimations de la variance par répliques qui ne reflètent pas entièrement la variabilité des totaux de contrôle. Cela peut se produire par exemple quand un sous-ensemble seulement des répliques contribue à l'estimation de la variance d'une moyenne de domaine. Si ces répliques sont pour la plupart non perturbées, l'estimation de la variance qui en résulte peut sous-estimer la variance. Il serait judicieux d'examiner de manière plus approfondie les performances de la méthode proposée en cas de nombre de répliques différent dans les deux enquêtes.

Bibliographie

Deming, W., et Stephan, F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

Dever, J.A., et Valliant, R. (2010). Une comparaison des estimateurs de la variance pour la poststratification en fonction de totaux de contrôle estimés. *Techniques d'enquête*, 36, 1, 49-61. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2010001/article/11251-fra.pdf.

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

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

Fay, R.E., et Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandrie, Virginie, 154-159.

Fish and Wildlife Service and Census Bureau (2018). *2016 National Survey of Fishing, Hunting, and Wildlife-Related Recreation*. Rapport de la méthodologie.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Kott, P.S. (2005). Using the delete-a-group jackknife variance estimator in an economic survey of US farms. Dans *Proceedings of the 55th Session of the World Statistics Congress*, Sydney. International Statistical Institute.

Lahiri, P. (2020). Preface to special issue on statistical data integration. *Statistics in Transition*, 21, III-VI.

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

Lohr, S., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475), 1019-1030.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Nadimpalli, V., Judkins, D. et Chu, A. (2004). Survey calibration to CPS household statistics. Dans *Proceedings of the Section on Survey Research Method*s, American Statistical Association, Alexandrie, Virginie, 4090-4094.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W. et Feuer, E.J. (2007). Combining information from two surveys toestimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, 102, 474-486.

Rockville Institute (2018). *2016 50-State Survey of Fishing, Hunting, and Wildlife-Related Recreation*. Rapport de la méthodologie.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

Wolter, K.M. (2007). *Introduction to Variance Estimation* (2 Éd.). New York: Springer-Verlag Inc.

Techniques d'enquête, décembre 2021
Vol. 47, n°2, pp. 303-323
Statistique Canada, n° 12-001-X au catalogue

Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu du modèle de Fay-Herriot

Éric Lesage, Jean-François Beaumont et Cynthia Bocci[1]

Résumé

Le modèle de Fay-Herriot est souvent utilisé pour obtenir des estimations sur petits domaines. Ces estimations sont généralement plus efficaces que les estimations directes classiques. Afin d'évaluer les gains d'efficacité obtenus par les méthodes d'estimation sur petits domaines, on produit généralement des estimations de l'erreur quadratique moyenne fondée sur le modèle. Cependant, ces estimations ne permettent pas de tenir compte de toute la spécificité d'un domaine en particulier car elles font disparaître l'effet local en prenant une espérance par rapport au modèle. Une alternative consiste à estimer l'erreur quadratique moyenne fondée sur le plan de sondage des estimateurs sur petits domaines. Cette dernière est souvent plus attrayante du point de vue des utilisateurs. Il est cependant connu que les estimateurs de l'erreur quadratique moyenne fondée sur le plan de sondage peuvent être très instables, particulièrement pour les domaines qui contiennent peu d'unités échantillonnées. Dans cet article, nous proposons deux diagnostics locaux qui ont pour objectif de faire un choix entre le meilleur prédicteur empirique et l'estimateur direct pour un domaine en particulier. Nous trouvons d'abord un intervalle de valeurs de l'effet local tel que le meilleur prédicteur est plus efficace sous le plan que l'estimateur direct. Ensuite, nous considérons deux approches différentes pour évaluer s'il est plausible que l'effet local se trouve dans cet intervalle. Nous examinons nos diagnostics au moyen d'une étude par simulation. Nos résultats préliminaires semblent prometteurs quant à l'utilité de ces diagnostics pour choisir entre le meilleur prédicteur empirique et l'estimateur direct.

Mots-clés : Meilleur prédicteur empirique; erreur quadratique moyenne fondée sur le plan; erreur quadratique moyenne fondée sur le modèle; diagnostic local; effet local; modèle de Fay-Herriot.

1. Introduction

Les gouvernements ont besoin d'information socio-économique à des niveaux de détail de plus en plus fins. Les instituts de statistique nationaux sont donc amenés à produire des statistiques pour des sous-populations qui n'avaient pas été identifiées au moment de la définition des objectifs de précision de l'enquête ou du moins qui ne pouvaient pas être pris en compte. Il en résulte que, pour ces sous-populations, le nombre d'unités enquêtées peut être trop faible pour garantir une bonne précision des estimateurs directs classiques de la théorie des sondages comme l'estimateur de Horvitz-Thompson ou l'estimateur par calage. Ce type de sous-population, où l'échantillon est de taille insuffisante, est appelé un petit domaine. Pour pallier le manque de précision des estimateurs directs sur les petits domaines, on peut avoir recours à des estimateurs indirects, ou estimateurs sur petits domaines, qui s'appuient sur des modèles, comme par exemple le modèle de Fay-Herriot (Fay et Herriot, 1979). Le meilleur prédicteur empirique, aussi appelé l'estimateur EB (pour *Empirical Best* ou *Empirical Bayes*), est un estimateur sur petits domaines fréquemment utilisé en pratique.

Les méthodes d'estimation sur petits domaines utilisent des modèles statistiques pour tirer le meilleur parti de l'information provenant de l'enquête et de l'information provenant de sources de données auxiliaires. Le modèle de Fay-Herriot est un modèle linéaire qui permet de décomposer le paramètre

1. Éric Lesage, Insee, Direction régionale, 35, place du Colombier - CS 94439 - 35044 Rennes Cedex. Courriel : eric.lesage@insee.fr; Jean-François Beaumont, Statistique Canada, 100 Pré Tunney, Immeuble R.H. Coats, Ottawa, Ontario, Canada, K1A 0T6. Courriel : jean-francois.beaumont@statcan.gc.ca; Cynthia Bocci, Statistique Canada, 100 Pré Tunney, Immeuble R.H. Coats, Ottawa, Ontario, Canada, K1A 0T6. Courriel : cynthia.bocci@statcan.gc.ca.

d'intérêt d'un domaine en deux termes : le premier terme est l'effet expliqué par le modèle et le second terme est l'erreur du modèle qu'on peut interpréter comme un effet local inexpliqué et inconnu.

On peut utiliser des outils statistiques classiques pour juger de la validité du modèle de Fay-Herriot tels que, par exemple, des graphiques de résidus du modèle. Ces outils donnent toutefois peu d'indications sur l'efficacité d'une estimation indirecte pour un domaine en particulier. L'erreur quadratique moyenne (EQM) fondée sur le modèle de Fay-Herriot d'un estimateur indirect peut être considérée comme un indicateur local de qualité puisqu'elle varie d'un domaine à l'autre. L'EQM fondée sur le modèle prend en compte l'effet expliqué par le modèle mais fait disparaître l'effet local inexpliqué en prenant une espérance par rapport au modèle.

L'EQM fondée sur le plan de sondage est une alternative à l'EQM fondée sur le modèle qui ne fait pas disparaître l'effet local inexpliqué. Cependant, les estimateurs sans biais de l'EQM fondée sur le plan d'estimateurs sur petits domaines ont tendance à être très instables, particulièrement pour les domaines qui contiennent peu d'unités échantillonnées (Rivest et Belmonte, 2000; Rao, Rubin-Bleuer et Estevao, 2018; et Pfeffermann et Ben-Hur, 2019). Pour contourner ce problème, on a suggéré dans la littérature de prendre une moyenne sur plusieurs domaines de l'EQM fondée sur le plan (Rao et Molina, 2015; et Pfeffermann et Ben-Hur, 2019) comme mesure de qualité. Cependant, de nombreux utilisateurs des statistiques publiques ne s'intéressent qu'à leur seul domaine et ne sont pas preneurs d'un critère global de qualité pour juger de l'efficacité des estimations pour leur domaine d'intérêt. C'est d'autant plus le cas lorsqu'ils sont convaincus que leur domaine est très spécifique et que cette spécificité ne se retrouve pas dans le terme explicatif du modèle mais plutôt dans le terme d'erreur, i.e. l'effet local inexpliqué.

Pour faire face au problème de l'instabilité des estimateurs sans biais de l'EQM fondée sur le plan, Rao, Rubin-Bleuer et Estevao (2018) ont proposé un estimateur composite qu'ils évaluent au moyen d'une étude par simulations. Leur estimateur composite consiste à prendre une moyenne pondérée d'un estimateur de l'EQM fondée sur le modèle et d'un estimateur de l'EQM fondée sur le plan. Ils obtiennent ainsi plus de stabilité au prix d'une augmentation du biais. Pfeffermann et Ben-Hur (2019) proposent également une méthode pour estimer l'EQM fondée sur le plan d'un estimateur sur petits domaines. La méthode est plutôt complexe et repose principalement sur le choix d'un modèle approprié. Elle n'est donc pas entièrement fondée sur le plan de sondage. Outre ces tentatives d'estimer l'EQM fondée sur le plan, il n'existe pas, à notre connaissance, de diagnostics locaux qui permettent de déterminer si l'estimation sur petits domaines est préférable à l'estimation directe pour un domaine en particulier.

Dans cet article, on propose une approche différente qui a pour objectif de comparer l'efficacité sous le plan des estimateurs EB et direct. On procède en deux temps. Dans un premier temps on détermine l'intervalle des valeurs de l'effet local inexpliqué qui garantissent que l'EQM fondée sur le plan du meilleur prédicteur, ou estimateur B (pour *Best* ou *Bayes*), soit inférieure à l'EQM fondée sur le plan de l'estimateur direct. Dans un second temps, on cherche à évaluer s'il est plausible que l'effet local inexpliqué se situe dans cet intervalle. A cette fin, on propose deux diagnostics : un premier qui s'appuie sur la distribution conditionnelle aux estimations directes de l'effet local inexpliqué et un second qui s'appuie sur un test d'hypothèse sur l'effet local inexpliqué, réalisé par rapport au plan de sondage. Nous

avons trouvé que, selon l'importance du résidu normalisé du modèle et d'un facteur associé à la précision de l'estimation directe, il est possible de détecter si les estimateurs B ou EB sont susceptibles d'avoir une EQM fondée sur le plan plus petite que celle de l'estimateur direct.

À la section 2, on présente le modèle de Fay-Herriot et on rappelle comment est construit le meilleur prédicteur (estimateur B) du paramètre d'intérêt de la population. À la section 3, on dérive les EQM fondées sur le modèle et fondées sur le plan de l'estimateur direct et du meilleur prédicteur. À la section 4, on décrit les deux diagnostics proposés. À la section 5, on explique comment estimer les paramètres du modèle et ainsi obtenir le meilleur prédicteur empirique (estimateur EB) et les estimateurs des diagnostics. À la section 6, on présente les résultats d'une étude par simulation à partir de données auxiliaires réelles. On termine par une brève conclusion à la section 7.

2. Le modèle de Fay-Herriot et le meilleur prédicteur

On considère une population finie U de taille N et un échantillon s de taille n tiré dans U selon un plan de sondage $p(s)$. La population U est partitionnée en m domaines qui ne se chevauchent pas. Les domaines sont repérés par un indice i qui prend les valeurs de là m. La population du domaine i, de taille N_i, est notée U_i. L'échantillon du domaine i est noté s_i et sa taille est n_i. On s'intéresse à estimer m paramètres de la population finie, $\theta_i, i = 1, \ldots, m,$ associés aux m domaines. Le paramètre θ_i est habituellement un total, une moyenne ou un ratio pour le domaine i. On dispose d'information auxiliaire sous la forme de vecteurs, $\mathbf{z}_i$, disponibles pour tous les domaines $i = 1, \ldots, m.$ On note $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1,\ldots,m}$, l'ensemble contenant les m vecteurs auxiliaires. On note également Ω, l'ensemble de toutes les variables utilisées pour faire les inférences excluant les variables indicatrices d'appartenance à l'échantillon s; Ω inclut entre autres $\mathcal{Z}$ et $\theta_i, i = 1, \ldots, m.$ Ainsi, on notera l'espérance par rapport au plan de sondage d'une variable aléatoire, disons A, par $\mathbb{E}(A \mid \Omega)$.

On pose un modèle de liaison qui nous permet de décomposer les paramètres d'intérêt θ_i de la façon suivante :

$$\theta_i = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i + b_i v_i, \qquad i = 1, \ldots, m, \tag{2.1}$$

où $\boldsymbol{\beta}$ est un vecteur de paramètres du modèle de même dimension que $\mathbf{z}_i$, les b_i sont des facteurs fixes qui permettent de tenir compte de l'hétérosédasticité dans le modèle et les v_i sont des termes d'erreur qui suivent la loi normale : $v_i \mid \mathcal{Z} \sim \mathcal{N}(0, \sigma_v^2)$, où σ_v^2 est un paramètre du modèle. En pratique, on pose souvent $b_i = 1$ mais il est plus naturel de poser $b_i = N_i$ quand θ_i est un total. Le terme $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i$ est l'effet connu ou expliqué par le modèle du paramètre de population finie θ_i alors que $b_i v_i$ est l'effet inconnu ou inexpliqué que l'on appelle effet local inexpliqué de θ_i ou encore simplement effet local de θ_i.

L'estimateur direct de θ_i est noté $\hat{\theta}_i$. Il est habituellement obtenu en pondérant chaque unité de l'échantillon s_i par un poids de sondage. Le poids de sondage d'une unité peut tout simplement être l'inverse de sa probabilité de sélection dans l'échantillon s ou un poids de calage. On note e_i l'erreur d'échantillonnage :

$$e_i = \hat{\theta}_i - \theta_i. \tag{2.2}$$

On considèrera par la suite que l'estimateur direct est sans biais sous le plan de sondage, c'est-à-dire que $\mathbb{E}(\hat{\theta}_i \mid \Omega) = \theta_i$ ou encore que $\mathbb{E}(e_i \mid \Omega) = 0.$ Cette hypothèse n'est pas toujours satisfaite en pratique, par exemple lorsqu'on utilise des poids de calage, mais on fera l'hypothèse habituelle que le biais reste négligeable. On suppose également que l'estimateur direct $\hat{\theta}_i,$ et ainsi l'erreur $e_i,$ suit une loi normale. Tel que discuté dans Rao et Molina (2015, page 77), l'hypothèse de normalité des erreurs e_i est possiblement moins forte que celle de la normalité des erreurs v_i en raison de l'effet du théorème central limite sur les $\hat{\theta}_i.$ Bien sûr, cet effet est moins important pour les plus petits domaines. Sous ces hypothèses, on a : $e_i \mid \Omega \sim \mathcal{N}(0, \psi_i),$ où $\psi_i = \mathbb{V}(\hat{\theta}_i \mid \Omega)$ est la variance sous le plan de sondage de $\hat{\theta}_i.$ La taille d'échantillon n_i peut être très faible, ce qui peut conduire à une mauvaise précision de l'estimateur direct $\hat{\theta}_i.$ Ce problème est à l'origine de la recherche sur l'estimation sur petits domaines.

En combinant le modèle (2.1) et l'expression (2.2), on obtient le modèle combiné, encore appelé modèle de Fay-Herriot :

$$\hat{\theta}_i = \boldsymbol{\beta}^\top \mathbf{z}_i + b_i\, v_i + e_i. \tag{2.3}$$

En notant que v_i est fixe sous le plan de sondage, on peut facilement montrer que $\mathbb{V}\left(b_i v_i + e_i \mid \mathcal{Z}\right) = b_i^2 \sigma_v^2 + \tilde{\psi}_i,$ où $\tilde{\psi}_i = \mathbb{E}\left(\psi_i \mid \mathcal{Z}\right)$ est la variance lissée (voir la remarque à la fin de cette section). On note ε_i l'erreur normalisée du modèle combiné :

$$\varepsilon_i = \frac{\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i}{\sqrt{b_i^2 \sigma_v^2 + \tilde{\psi}_i}}. \tag{2.4}$$

L'observation des estimations directes $\hat{\theta}_i$ fournit de l'information sur les $\theta_i.$ On peut obtenir les distributions conditionnelles des θ_i (Rao et Molina, 2015, chapitre 9, pages 271-272) :

$$\theta_i \mid \mathcal{Z}, \hat{\theta}_i \sim \mathcal{N}\left\{\boldsymbol{\beta}^\top \mathbf{z}_i + \gamma_i(\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i), (1-\gamma_i)\, b_i^2 \sigma_v^2\right\}, \tag{2.5}$$

où $\gamma_i = \frac{b_i^2 \sigma_v^2}{b_i^2 \sigma_v^2 + \tilde{\psi}_i}.$

Le meilleur prédicteur de θ_i conditionnellement à l'observation de $\hat{\theta}_i$ (Rao et Molina, 2015) est alors donné par :

$$\hat{\theta}_i^B = \mathbb{E}(\theta_i \mid \mathcal{Z}, \hat{\theta}_i) = \gamma_i\, \hat{\theta}_i + (1-\gamma_i)\, \boldsymbol{\beta}^\top \mathbf{z}_i. \tag{2.6}$$

On utilisera dans la suite de cet article le nom d'estimateur B pour le meilleur prédicteur $\hat{\theta}_i^B.$

Aux sections 3 et 4, on développe la théorie en supposant que $\boldsymbol{\beta}$, σ_v^2 et $\tilde{\psi}_i$ sont connus. À la section 5, on discute de l'estimation de ces trois quantités ce qui permet d'obtenir une version empirique du meilleur prédicteur et de nos diagnostics.

Remarque : Dans la littérature sur l'estimation sur petits domaines, on développe habituellement la théorie en supposant que ψ_i est fixe. Par conséquent, on fait implicitement l'hypothèse que $\tilde{\psi}_i = \psi_i.$ Lorsqu'on fait l'inférence sous le modèle de Fay-Herriot, on ne peut s'attendre à ce que ψ_i soit fixe. Par exemple, considérons le cas où θ_i est une proportion dans le domaine i et où on utilise un plan stratifié aléatoire simple avec remise dont les strates coïncident avec les domaines. L'estimateur direct $\hat{\theta}_i$ est simplement la

proportion échantillonnale dans le domaine i et il est bien connu que sa variance est donnée par $\psi_i = n_i^{-1}\theta_i(1-\theta_i)$. Dans ce cas, il est évident que ψ_i est aléatoire puisqu'il dépend de θ_i. On peut également facilement montrer que $\tilde{\psi}_i = n_i^{-1}\left(\boldsymbol{\beta}^{\top}\mathbf{z}_i(1-\boldsymbol{\beta}^{\top}\mathbf{z}_i) - b_i^2\sigma_v^2\right) \neq \psi_i$ sauf si $v_i = \sigma_v = 0$. Dans le reste de cet article, on développe toute la théorie en faisant l'hypothèse habituelle que $\tilde{\psi}_i = \psi_i$. En pratique, ces deux variances sont inconnues et doivent être estimées. À la section 5, on discute de l'estimation de $\tilde{\psi}_i$ au moyen d'un modèle de lissage. On peut facilement montrer que si on dispose d'un estimateur, $\hat{\tilde{\psi}}_i$, sans biais sous le modèle pour $\tilde{\psi}_i$, c'est-à-dire que $\mathbb{E}(\hat{\tilde{\psi}}_i \mid \mathcal{Z}) = \tilde{\psi}_i$, alors cet estimateur est également sans biais sous le modèle pour ψ_i, c'est-à-dire que $\mathbb{E}(\hat{\tilde{\psi}}_i - \psi_i \mid \mathcal{Z}) = 0$. L'inverse est aussi vrai : un estimateur sans biais sous le modèle pour ψ_i sera également sans biais sous le modèle pour $\tilde{\psi}_i$. Par conséquent, même si $\tilde{\psi}_i \neq \psi_i$, on peut estimer ces deux variances par le même estimateur. Cela suggère que l'hypothèse $\tilde{\psi}_i = \psi_i$ n'est peut-être pas si critique en pratique.

3. Les erreurs quadratiques moyennes de l'estimateur direct et de l'estimateur B

Afin d'apprécier l'efficacité de l'estimateur B donné à l'équation (2.6), on se tourne habituellement vers un calcul d'erreur quadratique moyenne. On peut envisager deux possibilités naturelles : soit prendre l'EQM par rapport au plan de sondage, soit prendre l'EQM par rapport au modèle combiné (2.3).

L'EQM fondée sur le modèle de l'estimateur direct $\hat{\theta}_i$ vaut :

$$\mathrm{EQM}_m(\hat{\theta}_i) = \mathbb{E}\left\{(\hat{\theta}_i - \theta_i)^2 \mid \mathcal{Z}\right\} = \tilde{\psi}_i$$

et l'EQM fondée sur le modèle de l'estimateur B vaut :

$$\mathrm{EQM}_m(\hat{\theta}_i^B) = \mathbb{E}\left\{(\hat{\theta}_i^B - \theta_i)^2 \mid \mathcal{Z}\right\} = \gamma_i\tilde{\psi}_i.$$

L'estimateur B est donc toujours plus efficace que l'estimateur direct avec une inférence fondée sur le modèle. Cette propriété résulte de la construction même de l'estimateur B. Par contre, et cela est une question légitime, on peut se demander si l'estimateur B est toujours plus efficace que l'estimateur direct avec une inférence fondée sur le plan de sondage.

Les erreurs quadratiques moyennes fondées sur le plan de sondage des estimateurs direct et B pour le domaine i valent :

$$\begin{aligned}\mathrm{EQM}_p(\hat{\theta}_i) = \mathbb{E}\left\{(\hat{\theta}_i - \theta_i)^2 \mid \Omega\right\} &= \psi_i \\ &= \tilde{\psi}_i \end{aligned} \tag{3.1}$$

et

$$\begin{aligned}\mathrm{EQM}_p(\hat{\theta}_i^B) = \mathbb{E}\left\{(\hat{\theta}_i^B - \theta_i)^2 \mid \Omega\right\} &= \gamma_i^2\psi_i + (1-\gamma_i)^2\, b_i^2 v_i^2 \\ &= \gamma_i\tilde{\psi}_i + (1-\gamma_i)^2 b_i^2(v_i^2 - \sigma_v^2). \end{aligned} \tag{3.2}$$

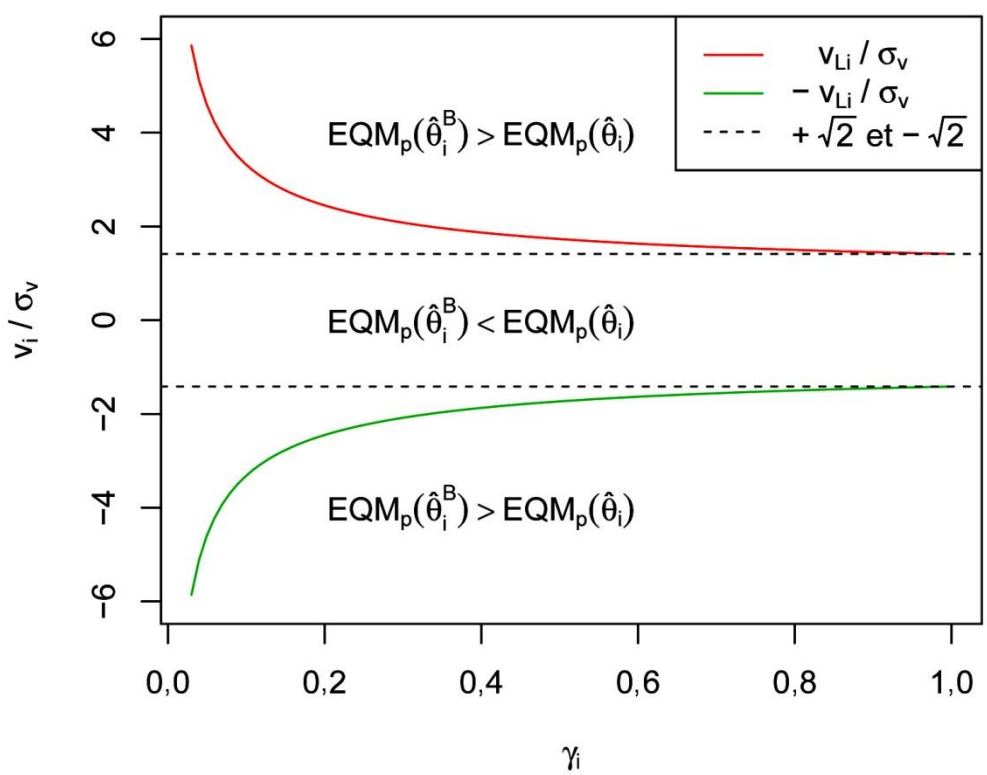
Notons que la deuxième égalité de (3.1) et (3.2) vient de l'hypothèse $\tilde{\psi}_i = \psi_i$. On observe que $\mathrm{EQM}_p\left(\hat{\theta}_i^B\right)$ peut être très différente de $\mathrm{EQM}_m\left(\hat{\theta}_i^B\right)$ quand la valeur inconnue v_i^2 est loin de σ_v^2. Par conséquent, pour un domaine avec une grande valeur de v_i^2, $\mathrm{EQM}_m\left(\hat{\theta}_i^B\right)$ pourrait être significativement plus petite que $\mathrm{EQM}_p\left(\hat{\theta}_i^B\right)$ et mener à une conclusion erronée sur l'efficacité relative de l'estimateur B par rapport à l'estimateur direct.

En remarquant que $\gamma_i \tilde{\psi}_i = (1-\gamma_i) b_i^2 \sigma_v^2$, on peut montrer que $\mathrm{EQM}_p\left(\hat{\theta}_i^B\right) \leq \mathrm{EQM}_p\left(\hat{\theta}_i\right)$ si et seulement si

$$v_i \in \left[-v_{L,i}; v_{L,i}\right],$$

où $v_{L,i} = \sigma_v \sqrt{\frac{1+\gamma_i}{\gamma_i}}$. La figure 3.1 donne les valeurs limites $v_{L,i}/\sigma_v$ et $-v_{L,i}/\sigma_v$ en fonction de γ_i. On note que lorsque $|v_i| \leq \sigma_v \sqrt{2}$, $\mathrm{EQM}_p\left(\hat{\theta}_i^B\right) \leq \mathrm{EQM}_p\left(\hat{\theta}_i\right)$ quelque soit la valeur de γ_i. On note également que l'estimateur direct peut devenir plus intéressant que l'estimateur B pour les domaines où l'effet local est important, particulièrement lorsque γ_i n'est pas petit. Mais comment savoir si l'effet local est important ou non pour un domaine i donné ? C'est l'objet de la section suivante où nous présentons deux diagnostics.

Figure 3.1 Valeurs limites de l'effet local normalisé par σ_v en fonction de γ_i.



4. Deux diagnostics pour évaluer la performance locale de l'estimateur B

4.1 Une approche conditionnelle à l'observation de $\hat{\theta}_i$

A partir de l'expression (2.5) de la section 2 et en notant que $\gamma_i\left(\hat{\theta}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i\right) = b_i\sigma_v\sqrt{\gamma_i}\varepsilon_i$, on obtient la distribution conditionnelle de v_i :

$$v_i \,|\, \mathcal{Z}, \hat{\theta}_i \sim \mathcal{N}\left(\sigma_v\sqrt{\gamma_i}\varepsilon_i, (1-\gamma_i)\sigma_v^2\right).$$

Le conditionnement sur $\hat{\theta}_i$ permet d'avoir une meilleure idée des valeurs possibles de v_i. En particulier, lorsque γ_i prend une valeur strictement supérieure à 0, la distribution conditionnelle de v_i peut s'écarter nettement de sa distribution non conditionnelle : $v_i \,|\, \mathcal{Z} \sim \mathcal{N}(0, \sigma_v^2)$.

Le premier diagnostic est défini comme la probabilité conditionnelle :

$$\begin{aligned} D_{1i} &= \mathrm{Prob}\left(\mathrm{EQM}_p(\hat{\theta}_i^B) \le \mathrm{EQM}_p(\hat{\theta}_i) \,\middle|\, \mathcal{Z}, \hat{\theta}_i\right) \\ &= \mathrm{Prob}\left(-v_{L,i} \le v_i \le v_{L,i} \,\middle|\, \mathcal{Z}, \hat{\theta}_i\right). \end{aligned} \tag{4.1}$$

Ce diagnostic peut s'écrire en fonction de γ_i et de l'erreur normalisée (2.4) :

$$\begin{aligned} D_{1i} = D_{1i}\left(\gamma_i, |\varepsilon_i|\right) &= \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\varepsilon_i| + \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\} \\ &\quad - \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\varepsilon_i| - \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\}, \end{aligned} \tag{4.2}$$

où $\Phi(\cdot)$ est la fonction de répartition de la loi normale centrée réduite. La preuve du résultat (4.2) est donnée à l'annexe A.

Quand ce diagnostic prend des valeurs proches de 0, on peut conclure que $|v_i|$ est vraisemblablement plus grand que $v_{L,i}$ et que l'estimateur direct est préférable à l'estimateur B. Pour obtenir une règle de décision associée à ce diagnostic, il est nécessaire de choisir un seuil en dessous duquel on décide de prendre l'estimateur direct et au dessus duquel on prend l'estimateur B. Un seuil à 50 % semble assez naturel. Une autre idée est d'adopter une approche empirique et de repérer une rupture dans la distribution des valeurs du diagnostic D_{1i} pour les m domaines.

Ce diagnostic n'est pas entièrement fondé sur le plan de sondage car il fait intervenir la distribution conditionnelle $v_i \,|\, \mathcal{Z}, \hat{\theta}_i$. Il est donc nécessaire de valider au mieux possible le modèle de Fay-Herriot avant de l'utiliser. Il n'est malheureusement pas possible de valider séparément les hypothèses sur v_i et e_i parce que les valeurs des paramètres $\theta_i, i = 1, \ldots, m,$ ne sont pas observées. On peut toutefois valider le modèle combiné de Fay-Herriot (2.3) au moyen des résidus du modèle (voir, par exemple, Hidiroglou, Beaumont et Yung, 2019). Ces résidus sont obtenus en remplaçant les quantités inconnues dans l'erreur

normalisée (2.4) par leurs estimations (voir section 5). Un graphique des résidus en fonction des valeurs prédites du modèle est souvent suggéré pour valider l'hypothèse de linéarité du modèle. L'hypothèse de normalité de l'erreur $b_i v_i + e_i$ peut être vérifiée par un graphique Q-Q des résidus ou des tests de normalité tels que le test de Shapiro-Wilk. Pour être conservateur, au cas où le modèle ne serait pas complètement satisfaisant, un seuil à 75 % peut être approprié.

Le diagnostic de la section suivante est fondé entièrement sur le plan de sondage. Il ne dépend donc pas de la validité du modèle de liaison. En ce sens, il est considéré plus robuste que le diagnostic (4.2). Il repose toutefois sur les hypothèses sur les erreurs d'échantillonnage e_i, discutées à la section 2, notamment l'hypothèse de normalité des e_i.

4.2 Utilisation d'un test d'hypothèse fondé sur le plan de sondage portant sur le paramètre v_i

Dans le cadre d'une inférence fondée sur le plan de sondage, v_i est fixe et l'erreur normalisée (2.4) suit la loi :

$$\varepsilon_i \,|\, \Omega \sim \mathcal{N}\left(v_i \frac{\sqrt{\gamma_i}}{\sigma_v}, (1-\gamma_i) \right). \tag{4.3}$$

Nous avons une observation unique de cette variable aléatoire. Nous l'utilisons pour tester si $|v_i|$ est plus grand que $v_{L,i}$. On pose le test :

$$\mathbf{H_0} : |v_i| = v_{L,i} \quad \text{contre} \quad \mathbf{H_1} : |v_i| > v_{L,i}.$$

On choisit d'utiliser $|\varepsilon_i|$ comme statistique de test. On s'attend à ce que $|\varepsilon_i|$ prenne des valeurs plus faibles sous H_0 que sous H_1. Soient $\varepsilon_{\text{obs},i}$, la valeur observée de la statistique ε_i et $P_i(v_i) = \text{Prob}\left(|\varepsilon_i| > |\varepsilon_{\text{obs},i}| \,|\, \Omega; v_i \right)$. La valeur-$p$ du test est définie comme étant la probabillité que la statistique $|\varepsilon_i|$ soit plus grande que la valeur observée $|\varepsilon_{\text{obs},i}|$ sous l'hypothèse nulle. On montre à l'annexe B que la valeur-p vaut :

$$P_i\left(v_{L,i}\right) = P_i\left(-v_{L,i}\right) = \Phi\left(-\tau_i\right) + \Phi\left(-\tau_i - 2\frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}} \right),$$

où

$$\tau_i = \frac{|\varepsilon_{\text{obs},i}| - \sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}.$$

Dans la mesure où le second terme est souvent négligeable par rapport au premier terme, en particulier quand $\tau_i > 0$ ou γ_i est grand, on prend comme second diagnostic :

$$D_{2i} = D_{2i}\left(\gamma_i, \left|\varepsilon_{\text{obs},i}\right|\right) = \Phi\left(\frac{\sqrt{1+\gamma_i} - \left|\varepsilon_{\text{obs},i}\right|}{\sqrt{1-\gamma_i}}\right). \tag{4.4}$$

Le second diagnostic s'interprète de la façon suivante. Quand D_{2i} est faible, on peut supposer que $|v_i|$ est vraisemblablement plus grand que $v_{L,i}$ et on préfère alors l'estimateur direct à l'estimateur B. Pour le choix du seuil de décision, on peut s'inspirer des valeurs utilisées habituellement comme niveaux pour les tests (par exemple, 5 % ou 10 %). Avec ces valeurs, compte tenu du sens du test, on favorise l'estimateur B. Comme pour le diagnostic précédent, on peut déterminer la valeur du seuil en repérant une rupture dans la distribution des valeurs du diagnostic D_{2i} pour les m domaines.

4.3 Quelques propriétés des diagnostics 1 et 2

Dans cette section, nous étudions le comportement des fonctions $D_{1i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ et $D_{2i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ pour des cas limites de γ_i et $|\varepsilon_i|$ et nous notons leurs similarités et différences.

Cas 1 : $0 < \gamma_i < 1$ est fixe et $|\varepsilon_i| \to \infty$.

À partir des équations (4.2) et (4.4), on peut montrer que pour $|\varepsilon_i| > 0$, les deux fonctions $D_{1i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ et $D_{2i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ décroissent à mesure que $|\varepsilon_i|$ croît. Autrement dit, la dérivée de ces fonctions par rapport à $|\varepsilon_i|$ est négative. De plus, la limite quand $|\varepsilon_i| \to \infty$ de ces deux fonctions tend vers 0. Pour une valeur de $|\varepsilon_i|$ suffisamment élevée, les deux diagnostics vont donc privilégier l'estimateur direct.

Cas 2 : $0 < \gamma_i < 1$ est fixe et $|\varepsilon_i| = 0$.

À partir de l'équation (4.2), on observe que

$$D_{1i}\left(\gamma_i, 0\right) = \Phi\left(\sqrt{\frac{1+\gamma_i}{\gamma_i\left(1-\gamma_i\right)}}\right) - \Phi\left(-\sqrt{\frac{1+\gamma_i}{\gamma_i\left(1-\gamma_i\right)}}\right).$$

On peut montrer que $D_{1i}\left(\gamma_i, 0\right)$ est minimisé quand $\gamma_i = -1 + \sqrt{2}$. Par conséquent, $D_{1i}\left(\gamma_i, 0\right) \geq D_{1i}\left(-1+\sqrt{2}, 0\right) =$ 0,98. Puisque cette valeur est près de 1, le diagnostic 1 conduit à choisir l'estimateur B dans ce cas si on prend un seuil à 0,50 ou même à 0,75.

À partir de l'équation (4.4), on obtient :

$$D_{2i}\left(\gamma_i, 0\right) = \Phi\left(\sqrt{\frac{1+\gamma_i}{1-\gamma_i}}\right).$$

On peut montrer que, pour $0 \leq \gamma_i < 1$, la fonction $D_{2i}\left(\gamma_i, 0\right)$ est minimisée quand $\gamma_i = 0$. Ainsi, $D_{2i}\left(\gamma_i, 0\right) \geq D_{2i}\left(0, 0\right) =$ 0,84. Avec un seuil plus petit que 0,50, le diagnostic 2 conduit à la même décision que le diagnostic 1 dans ce cas, c'est-à-dire celle de choisir l'estimateur B.

Cas 3 : $|\varepsilon_i| < \sqrt{2}$ est fixe et $\gamma_i \to 1$.

Les deux fonctions $D_{1i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ et $D_{2i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ tendent vers 1 dans ce cas. Les diagnostics 1 et 2 conduisent donc à choisir l'estimateur B.

Cas 4 : $|\varepsilon_i| > \sqrt{2}$ est fixe et $\gamma_i \to 1$.

Les deux fonctions $D_{1i}(\gamma_i, |\varepsilon_i|)$ et $D_{2i}(\gamma_i, |\varepsilon_i|)$ tendent vers 0 dans ce cas. Les diagnostics 1 et 2 conduisent ici à choisir l'estimateur direct.

Cas 5 : $|\varepsilon_i|$ est fixe et $\gamma_i \to 0$.

La fonction $D_{1i}(\gamma_i, |\varepsilon_i|)$ tend vers 1 pour n'importe quelle valeur fixe de $|\varepsilon_i|$. Le diagnostic 1 privilégie donc l'estimateur B pour de petites valeurs de γ_i.

On note que $D_{2i}(0, |\varepsilon_i|) = \Phi(1 - |\varepsilon_i|)$. Par conséquent, contrairement au diagnostic 1, le diagnostic 2 conduira à choisir l'estimateur direct si $|\varepsilon_i|$ est suffisamment grand même quand γ_i est infiniment près de 0. Par exemple, avec un seuil de décision à 0,05 et $\gamma_i = 0$, le diagnostic 2 favorise l'estimateur direct quand $|\varepsilon_i| > 1 - \Phi^{-1}(0{,}05) =$ 2,64.

Dans les quatre premiers cas ci-dessus, les deux diagnostics conduisent à la même décision. On observe une différence seulement dans le cas 5 où $\gamma_i \to 0$. On s'attend donc à ce que le diagnostic 2 choisisse plus souvent l'estimateur direct que le diagnostic 1 pour de petites valeurs de γ_i. Considérons, par exemple, un seuil à 0,5 pour le diagnostic 1 et à 0,05 pour le diagnostic 2. Pour un seuil à 0,5, on peut montrer que le diagnostic 1 conduit à choisir l'estimateur direct dès que $|\varepsilon_i|$ est supérieur à une valeur approximativement égale à $\frac{\sqrt{1+\gamma_i}}{\gamma_i}$, c'est-à-dire dès que $|\varepsilon_i| \gtrsim \frac{\sqrt{1+\gamma_i}}{\gamma_i}$. Quant au diagnostic 2, pour un seuil à 0,05, il conduit à choisir l'estimateur direct dès que $|\varepsilon_i| > \sqrt{1+\gamma_i} - \sqrt{1-\gamma_i}\,\Phi^{-1}(0{,}05)$. Pour γ_i = 0,01, le diagnostic 1 choisit donc l'estimateur direct quand $|\varepsilon_i| \gtrsim$ 100,5 alors que le diagnostic 2 choisit l'estimateur direct dès que $|\varepsilon_i| >$ 2,64. L'écart se rétrécit à mesure que γ_i augmente. Par exemple, pour γ_i = 0,2, le diagnostic 1 choisit l'estimateur direct quand $|\varepsilon_i| \gtrsim$ 5,48 et le diagnostic 2 choisit l'estimateur direct quand $|\varepsilon_i| >$ 2,57. La discussion ci-dessus semble suggérer que le diagnostic 2 choisit plus souvent l'estimateur direct que le diagnostic 1. Il existe cependant des cas où le diagnostic 1 choisit l'estimateur direct contrairement au diagnostic 2. Ces cas surviennent généralement pour des valeurs assez grandes de γ_i. Par exemple, pour γ_i = 0,8, le diagnostic 1 choisit l'estimateur direct quand $|\varepsilon_i| \gtrsim$ 1,68 alors que le diagnostic 2 choisit l'estimateur direct seulement quand $|\varepsilon_i| >$ 2,08.

5. Version empirique de l'estimateur B et des diagnostics

On a développé la théorie en supposant les paramètres $\boldsymbol{\beta}$, σ_v^2 et $\tilde{\psi}_i$ connus. En pratique, ces quantités sont inconnues et on ne peut utiliser le meilleur prédicteur $\hat{\theta}_i^B$. On peut les remplacer par des estimateurs $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_v^2$ et $\hat{\tilde{\psi}}_i$ pour ainsi obtenir le meilleur prédicteur empirique (estimateur EB) :

$$\hat{\theta}_i^{\mathrm{EB}} = \hat{\gamma}_i\,\hat{\theta}_i + (1 - \hat{\gamma}_i)\,\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{z}_i,$$

où $\hat{\gamma}_i = \frac{b_i^2\hat{\sigma}_v^2}{b_i^2\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}$.

Dans ce qui suit, on commence d'abord par discuter de l'estimation de $\boldsymbol{\beta}$ en supposant σ_v^2 et $\tilde{\boldsymbol{\psi}} = (\tilde{\psi}_1, \ldots, \tilde{\psi}_m)^{\mathrm{T}}$ connus. On obtient ainsi l'estimateur $\tilde{\boldsymbol{\beta}}(\sigma_v^2, \tilde{\boldsymbol{\psi}})$ de $\boldsymbol{\beta}$. Ensuite, on discute de l'estimation de σ_v^2 en supposant que $\tilde{\boldsymbol{\psi}}$ est connu et on obtient l'estimateur $\tilde{\sigma}_v^2(\tilde{\boldsymbol{\psi}})$ de σ_v^2. Finalement, on discute de l'estimation des variances lissées $\tilde{\psi}_i, i = 1, \ldots, m$. On note $\hat{\tilde{\psi}}_i, i = 1, \ldots, m$, les estimateurs

obtenus et $\hat{\tilde{\boldsymbol{\psi}}} = \left(\hat{\tilde{\psi}}_1, \ldots, \hat{\tilde{\psi}}_m\right)^\top$. En pratique, il faut d'abord commencer par estimer les variances lissées et ensuite on calcule successivement $\hat{\sigma}_v^2 = \tilde{\sigma}_v^2\left(\hat{\tilde{\boldsymbol{\psi}}}\right)$ et $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2, \hat{\tilde{\boldsymbol{\psi}}}\right)$, les estimations de σ_v^2 et $\boldsymbol{\beta}$.

En supposant σ_v^2 et $\tilde{\boldsymbol{\psi}} = \left(\tilde{\psi}_1, \ldots, \tilde{\psi}_m\right)^\top$ connus, l'estimation de $\boldsymbol{\beta}$ peut se faire en utilisant la méthode des moindres carrés généralisés qui est équivalente à la méthode du maximum de vraisemblance sous l'hypothèse d'indépendance et de normalité des erreurs $b_i v_i + e_i$. On obtient :

$$\tilde{\boldsymbol{\beta}}\left(\sigma_v^2, \tilde{\boldsymbol{\psi}}\right) = \left(\sum_{i=1}^m \frac{\mathbf{z}_i \mathbf{z}_i^\top}{b_i^2 \sigma_v^2 + \tilde{\psi}_i}\right)^{-1} \sum_{i=1}^m \frac{\mathbf{z}_i \hat{\theta}_i}{b_i^2 \sigma_v^2 + \tilde{\psi}_i}.$$

Différentes méthodes existent pour estimer σ_v^2. Par exemple, on peut utiliser la méthode des moments de Fay et Herriot (1979) ou encore la méthode du maximum de vraisemblance ou du maximum de vraisemblance restreint. Cette dernière est la plus fréquemment utilisée en pratique. Toutes ces méthodes consistent à résoudre itérativement une équation d'estimation de la forme $g\left(\sigma_v^2, \tilde{\boldsymbol{\psi}}\right) = 0$, où la fonction g dépend de la méthode. On note l'estimateur qui en résulte par $\tilde{\sigma}_v^2\left(\tilde{\boldsymbol{\psi}}\right)$. Rao et Molina (2015, Chapitres 5 et 6) fournissent plus de détails sur l'estimation de $\boldsymbol{\beta}$ et σ_v^2 et sur les propriétés des estimateurs telles que la convergence sous le modèle.

Avant d'estimer σ_v^2 et $\boldsymbol{\beta}$ par $\hat{\sigma}_v^2 = \tilde{\sigma}_v^2\left(\hat{\tilde{\boldsymbol{\psi}}}\right)$ et $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2, \hat{\tilde{\boldsymbol{\psi}}}\right)$, il faut d'abord estimer la variance lissée $\tilde{\psi}_i = \mathbb{E}\left(\psi_i \,|\, \mathcal{Z}\right), i = 1, \ldots, m.$ On suppose qu'on dispose d'un estimateur sans biais sous le plan, $\hat{\psi}_i$, c'est-à-dire que $\mathbb{E}\left(\hat{\psi}_i \,|\, \Omega\right) = \psi_i$. Sous cette hypothèse, on observe que $\mathbb{E}\left(\hat{\psi}_i \,|\, \mathcal{Z}\right) = \tilde{\psi}_i$. L'estimateur $\hat{\psi}_i$ est donc sans biais pour la variance lissée $\tilde{\psi}_i$ mais peut être très instable quand n_i est petite. En général, on préfère plutôt modéliser $\hat{\psi}_i$ en fonction de $\mathbf{z}_i$ pour obtenir un estimateur moins variable. Le modèle de lissage suivant est fréquemment utilisé en pratique :

$$\log\left(\hat{\psi}_i\right) = \boldsymbol{\alpha}^\top \mathbf{x}_i + \eta_i,$$

où $\mathbf{x}_i$ est une fonction de $\mathbf{z}_i$, $\boldsymbol{\alpha}$ est un vecteur de paramètres du modèle et $\eta_i, i = 1, \ldots, m,$ sont des erreurs indépendantes et identiquement distribuées avec une moyenne égale à 0 et une variance égale à σ_η^2. On peut facilement montrer que

$$\tilde{\psi}_i = \mathbb{E}\left(\hat{\psi}_i \,|\, \mathcal{Z}\right) = \exp\left(\boldsymbol{\alpha}^\top \mathbf{x}_i\right) \Delta,$$

où $\Delta = \mathbb{E}\left\{\exp\left(\eta\right)\right\}$ et η est une variable aléatoire qui suit la même distribution que le terme d'erreur du modèle de lissage ci-dessus. On estime $\boldsymbol{\alpha}$ par un estimateur convergent sous le modèle, $\hat{\boldsymbol{\alpha}}$, en utilisant la méthode des moindres carrés. Hidiroglou, Beaumont et Yung (2019) suggèrent d'estimer Δ par un estimateur convergent sous le modèle, $\hat{\Delta}$, en utilisant une méthode des moments. L'estimateur de la variance lissée s'écrit :

$$\hat{\tilde{\psi}}_i = \exp\left(\hat{\boldsymbol{\alpha}}^\top \mathbf{x}_i\right) \hat{\Delta},$$

où

$$\hat{\Delta} = \frac{\sum_{i=1}^m \hat{\psi}_i}{\sum_{i=1}^m \exp\left(\hat{\boldsymbol{\alpha}}^\top \mathbf{x}_i\right)}.$$

On peut s'attendre à ce que l'EQM fondée sur le plan de l'estimateur EB,

$$\mathrm{EQM}_p(\hat{\theta}_i^{\mathrm{EB}}) = \mathbb{E}\left\{(\hat{\theta}_i^{\mathrm{EB}} - \theta_i)^2 \,\middle|\, \Omega\right\},$$

soit supérieure à l'EQM fondée sur le plan de l'estimateur B donné à l'équation (3.2). Tel que mentionné ci-dessus, les estimateurs des paramètres $\boldsymbol{\beta}$, σ_v^2, $\boldsymbol{\alpha}$ et Δ sont convergents sous le modèle à mesure que m augmente pourvu que certaines conditions de régularité soient satisfaites. On note également que l'erreur quadratique moyenne fondée sur le plan de l'estimateur B (voir l'équation 3.2) ne dépend pas de m. Par conséquent, on peut s'attendre à ce que l'augmentation de l'erreur quadratique moyenne résultant de l'estimation de ces paramètres soit modeste quand le nombre de domaines est grand. Cela suggère que la dérivation de la borne $v_{L,i}$ sera peu affectée par l'estimation des paramètres $\boldsymbol{\beta}$, σ_v^2, $\boldsymbol{\alpha}$ et Δ si m est grand. Ainsi, nos deux diagnostics (4.2) et (4.4) devraient rester pertinents même si c'est l'estimateur EB qui est utilisé plutôt que l'estimateur B. Il faut cependant remplacer γ_i par $\hat{\gamma}_i$ et ε_i par

$$\hat{\varepsilon}_i = \frac{\hat{\theta}_i - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{z}_i}{\sqrt{b_i^2\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}}$$

dans les expressions (4.2) et (4.4) pour être en mesure de calculer ces diagnostics avec des données réelles. On obtient ainsi $\hat{D}_{1i}$, l'estimateur de D_{1i}, et $\hat{D}_{2i}$, l'estimateur de D_{2i}.

6. Étude par simulation

On réalise une étude par simulation pour évaluer l'efficacité de $\hat{D}_{1i}$ et $\hat{D}_{2i}$ à détecter lequel des estimateurs direct et EB est préférable. On s'intéresse à $m =$ 140 domaines qui représentent des villes canadiennes. Le vecteur de variables auxiliaires est : $\mathbf{z}_i^{\mathrm{T}} = (1, z_{1i})$. La variable auxiliaire z_{1i} est obtenue à partir de fichiers administratifs et est définie comme le ratio du nombre de bénéficiaires d'assurance-emploi dans la ville i sur le nombre de personnes de plus de 15 ans dans la ville i. La taille de l'échantillon dans la ville i, n_i, est obtenue au moyen de l'Enquête canadienne sur la population active (EPA). Parmi les 140 domaines, 2 ont une taille d'échantillon plus petite que 10, 10 ont une taille d'échantillon plus petite que 30, 40 ont une taille d'échantillon plus petite que 60 et 68 ont une taille d'échantillon plus petite que 100, soit tout près de 50 % des domaines. Pour ces 68 domaines, les coefficients de variation estimés du taux de chômage de l'EPA sont la plupart du temps trop élevés pour pouvoir publier les estimations directes du taux de chômage; des techniques d'estimation sur petits domaines sont donc requises pour ces domaines. À l'opposé, il y a également 17 domaines parmi les 140 qui ont une taille d'échantillon plus grande que 1 000 et pour lesquels l'estimation directe du taux de chômage est fiable.

À partir des valeurs réelles de n_i et $\mathbf{z}_i$, on simule le paramètre de population θ_i pour les m domaines. Le paramètre θ_i peut être interprété comme la proportion de personnes sans emploi dans le domaine i. On génère θ_i selon la loi béta de moyenne $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i$ et de variance σ_v^2, où $\sigma_v^2 = 7{,}58 \times 10^{-5}$ et $\boldsymbol{\beta}^{\mathrm{T}} = (0{,}0484;$ $0{,}95)$. Ces valeurs de $\boldsymbol{\beta}$ et σ_v^2 ont été choisies à partir de données réelles. On pose $b_i = 1, i = 1, \ldots, m.$

Ensuite, on change manuellement les valeurs de θ_i pour 4 domaines (villes) de manière à avoir un effet local v_i égal à $5\sigma_v$. On a choisi des domaines avec des tailles d'échantillon différentes : 10, 100, 501 et 3 773. Dans le reste de cette section, on identifiera la plus petite de ces 4 villes par la ville 1 $(n_i = 10)$, la deuxième plus petite par la ville 2 $(n_i = 100)$, la deuxième plus grande par la ville 3 $(n_i = 501)$ et la plus grande par la ville 4 (n_i = 3 773).

On considère un plan de sondage stratifié aléatoire simple avec remise dont les strates coïncident avec les domaines. L'estimateur direct $\hat{\theta}_i$ de θ_i est simplement la proportion de personnes échantillonnées dans le domaine i qui ont la caractéristique d'intérêt (par exemple, être sans emploi). Sous un tel plan simple, il est facile de constater que l'estimateur direct peut être généré comme suit : $\hat{\theta}_i = n_i^{-1}$ Binomiale (n_i, θ_i). Il n'est donc pas nécessaire de créer la population de personnes dans le domaine i pour générer $\hat{\theta}_i$. On a procédé de cette façon dans la simulation. La variance sous le plan de $\hat{\theta}_i$ est donnée par $\psi_i = n_i^{-1}\theta_i(1-\theta_i)$ et son estimateur par $\hat{\psi}_i = (n_i - 1)^{-1}\hat{\theta}_i(1-\hat{\theta}_i)$. On estime la variance lissée $\tilde{\psi}_i$ en utilisant le modèle de lissage de la section 5 avec $\mathbf{x}_i^{\mathrm{T}} = \left(1, \log(z_{1i}), \log(1-z_{1i}), \log(n_i)\right)$.

Afin de simuler un scénario réaliste, nous n'avons pas respecté exactement toutes les hypothèses du modèle de Fay-Herriot. Entre autres, les erreurs v_i et e_i ne sont pas distribuées selon des lois normales. Nous avons opté pour une loi béta pour générer θ_i; l'hypothèse de normalité des v_i n'est donc pas satisfaite bien que la déviation de la loi normale ne soit pas drastique dans notre simulation. Les estimations $\hat{\theta}_i$ ont été générées selon une loi binomiale qui peut être approchée par une loi normale pour les domaines qui ont une grande valeur de n_i. La relation entre les estimations $\hat{\theta}_i$ simulées et les $\mathbf{z}_i$ ressemble à celle qu'on observe avec les estimations réelles de l'EPA. De plus, notre scénario de simulation est tel que l'hypothèse $\tilde{\psi}_i = \psi_i$ n'est pas respectée puisque, pour ce plan simple,

$$\tilde{\psi}_i = n_i^{-1}\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i(1-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i) - b_i^2\sigma_v^2\right)$$

(voir la remarque de la section 2). On note toutefois que le coefficient de corrélation entre $\tilde{\psi}_i$ et ψ_i est égal à 0,98 ce qui indique que la déviation de l'hypothèse $\tilde{\psi}_i = \psi_i$ est modeste. Tel que mentionné au paragraphe précédent, on utilise le modèle de lissage de la section 5 pour estimer $\tilde{\psi}_i$. Cela permet de rester dans un cadre réaliste où le modèle de lissage postulé est différent du vrai modèle qui a été utilisé pour générer les estimations $\hat{\psi}_i$.

On effectue une simulation fondée sur le plan de sondage, c'est-à-dire que les paramètres de population $\theta_i, i = 1, \ldots, m,$ ne sont générés qu'une seule fois. On répète K = 10 000 fois l'échantillonnage. Pour chaque itération $k, k = 1, \ldots, K,$ on génère une estimation directe $\hat{\theta}_i(k)$ et on calcule une estimation de la variance lissée $\hat{\tilde{\psi}}_i(k)$ tel que décrit ci-dessus. On calcule ensuite l'estimation EB :

$$\hat{\theta}_i^{\mathrm{EB}}(k) = \hat{\gamma}_i(k)\,\hat{\theta}_i(k) + \left(1 - \hat{\gamma}_i(k)\right)\hat{\boldsymbol{\beta}}(k)^{\mathrm{T}}\mathbf{z}_i,$$

où $\hat{\gamma}_i(k) = \frac{\hat{\sigma}_v^2(k)}{\hat{\sigma}_v^2(k) + \hat{\tilde{\psi}}_i(k)}$ et $\hat{\boldsymbol{\beta}}(k)$ et $\hat{\sigma}_v^2(k)$ sont calculés tels que décrits à la section 5. On utilise la méthode des moindres carrés généralisés pour obtenir $\hat{\boldsymbol{\beta}}(k)$ et la méthode du maximum de vraisemblance restreint

pour $\hat{\sigma}_v^2(k)$. Les calculs des estimations ont été réalisés avec le système d'estimation sur petits domaines de Statistique Canada (Hidiroglou, Beaumont et Yung, 2019).

Pour chaque itération, on calcule également les résidus normalisés $\hat{\varepsilon}_i(k)$ et les diagnostics $\hat{D}_{1i}(k)$ et $\hat{D}_{2i}(k)$ pour les m domaines. On enregistre si on a préféré l'estimateur direct à l'estimateur EB pour chacun des deux diagnostics. On utilise pour ce faire des seuils de décision. En deçà des seuils, on prend l'estimateur direct. Pour le diagnostic 1, on a retenu des seuils de 50 % et 75 % et pour le diagnostic 2, des seuils de 5 % et 25 %.

À partir des quantités précédentes, calculées pour chacune des 10 000 itérations, on calcule les moyennes Monte Carlo des diagnostics 1 et 2 pour les m domaines : $\bar{\hat{D}}_{1i}$ et $\bar{\hat{D}}_{2i}$. On calcule aussi le taux de sélection de l'estimateur direct pour chacun des deux diagnostics, c'est-à-dire le pourcentage de fois où un diagnostic donné a conduit à choisir l'estimateur direct.

On calcule l'approximation Monte Carlo de $\mathrm{EQM}_p\left(\hat{\theta}_i^{\mathrm{EB}}\right)$:

$$\mathrm{EQM}_{\mathrm{MC}}(\hat{\theta}_i^{\mathrm{EB}}) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_i^{\mathrm{EB}}(k) - \theta_i\right)^2.$$

A partir de cette dernière quantité, on calcule l'efficacité relative de l'estimateur EB :

$$\frac{\mathrm{EQM}_{\mathrm{MC}}(\hat{\theta}_i^{\mathrm{EB}}) - \psi_i}{\psi_i}. \tag{6.1}$$

Ce ratio est positif lorsque l'estimateur EB est moins efficace que l'estimateur direct sous le plan de sondage. Un diagnostic est potentiellement performant s'il est corrélé négativement à ce ratio.

Les figures 6.1 et 6.2 présentent les moyennes Monte Carlo des diagnostics 1 et 2 respectivement en fonction de l'efficacité relative de l'estimateur EB définie à l'équation (6.1). Les quatre villes dont les valeurs de v_i ont été modifiées, les villes 1 à 4, sont colorées en *violet*, *orange*, *vert* et *rouge*. Dans la légende, on a indiqué la taille de l'échantillon dans ces villes. Les valeurs du paramètre γ_i pour les villes 1 à 4 valent respectivement : 0,01; 0,08; 0,35 et 0,81. Toutes les autres villes sont colorées en bleu.

On observe d'abord, sur les figures 6.1 et 6.2, que l'estimateur EB est plus efficace que l'estimateur direct pour la ville 1 (en violet) puisque cette ville se retrouve à gauche de la ligne verticale (efficacité relative négative) et cela en dépit du fort effet local. L'explication de ce phénomène se trouve à la figure 3.1. On y constate que l'étendue des valeurs de v_i pour lesquelles l'estimateur B est plus efficace que l'estimateur direct augmente à mesure que γ_i diminue. Puisque γ_i est petit pour la ville 1 $(\gamma_i = 0{,}01)$, il n'est pas surprenant d'observer une efficacité relative négative malgré un effet local prononcé. Pour la ville 2 (en orange), l'estimateur direct est légèrement plus efficace que l'estimateur EB. Par contre, pour les villes 3 (en vert) et 4 (en rouge), l'estimateur direct est nettement plus efficace. On note également qu'il y a 5 villes pour lesquelles l'estimateur direct est plus efficace que l'estimateur EB : les villes 2 à 4 de même que deux autres villes dont les valeurs de v_i ont été générées aléatoirement et n'ont pas été modifiées manuellement. Une de ces villes a la plus petite valeur de v_i et l'autre a la plus

grande valeur de v_i après celle des 4 villes modifiées manuellement. Ces deux villes ont également de grandes valeurs de γ_i (0,62 et 0,49).

Les figures 6.1 et 6.2 indiquent que nos deux diagnostics semblent assez performants à déceler les cas où l'estimateur direct est plus efficace que l'estimateur EB sauf pour la ville 2 $(n_i = 100)$ où la moyenne Monte Carlo du diagnostic 1 est très élevée à 0,97. C'est toutefois un domaine pour lequel choisir l'estimateur le moins efficace n'est pas dramatique puisqu'il y a très peu de différence entre les efficacités des deux estimateurs. Outre ce cas spécifique, le diagnostic 1 semble avoir un meilleur comportement que le diagnostic 2. La moyenne Monte Carlo du diagnostic 1 est tout près de 1 quand l'estimateur EB est significativement plus efficace que l'estimateur direct, diminue lentement quand les efficacités des deux estimateurs se rapprochent et devient petite quand l'estimateur direct est significativement plus efficace que l'estimateur EB. On n'observe pas tout-à-fait le même comportement pour le diagnostic 2. La moyenne Monte Carlo du diagnostic 2 est petite quand l'estimateur direct est significativement plus efficace que l'estimateur EB mais elle n'est pas près de 1 quand l'estimateur EB est nettement plus efficace que l'estimateur direct. Surtout, elle semble augmenter quand les efficacités des deux estimateurs se rapprochent ce qui est contre-intuitif.

Figure 6.1 Moyenne Monte Carlo des estimations du diagnostic 1 pour les 140 domaines en fonction de l'efficacité relative de l'estimateur EB.

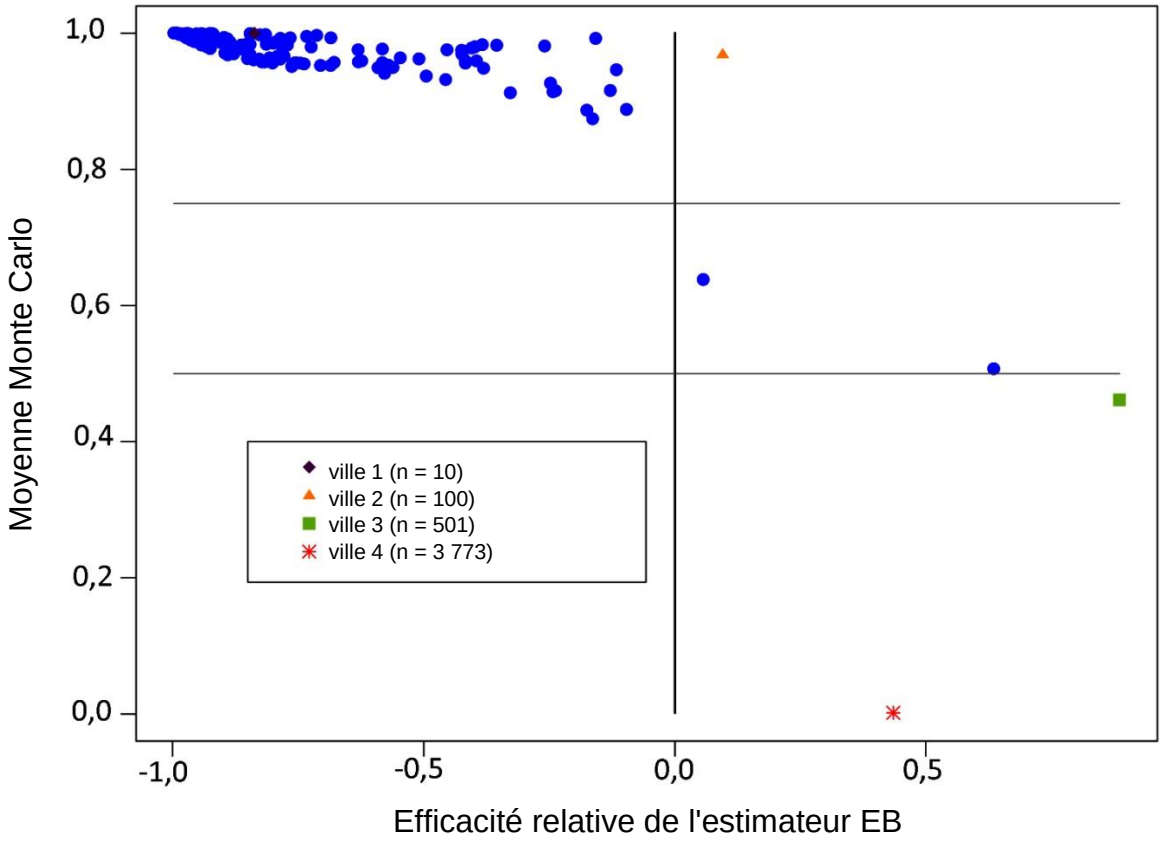
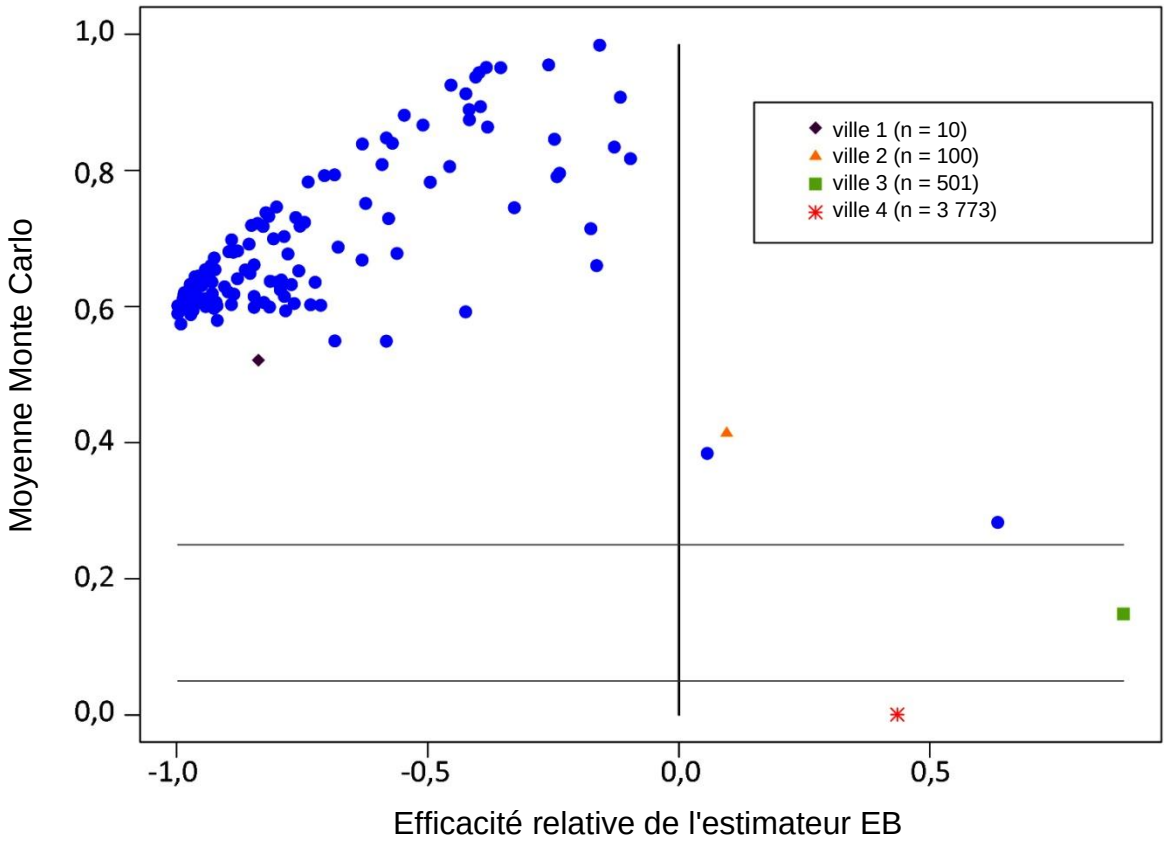
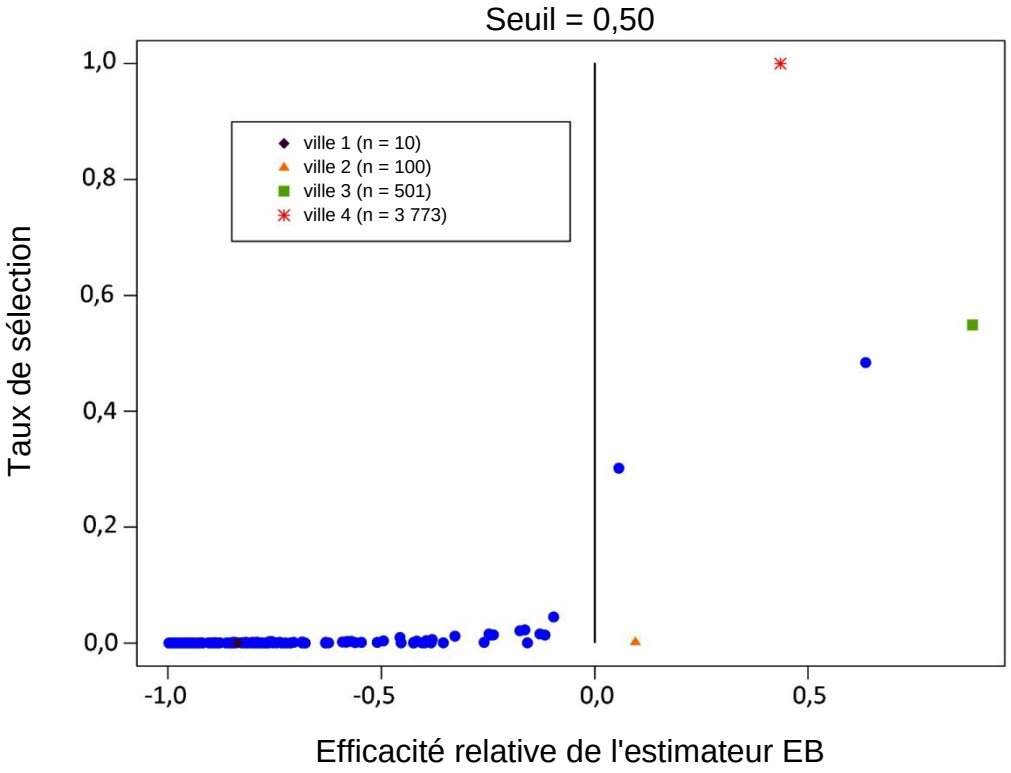


Figure 6.2 Moyenne Monte Carlo des estimations du diagnostic 2 pour les 140 domaines en fonction de l'efficacité relative de l'estimateur EB.



Les figures 6.3 et 6.4 montrent le taux de sélection de l'estimateur direct sur les 10 000 répétitions pour les diagnostics 1 et 2. On peut tirer des conclusions similaires à celles obtenues en analysant les figures 6.1 et 6.2. Comme on s'y attend, les seuils de 75 % pour le diagnostic 1 et de 25 % pour le diagnostic 2 permettent de mieux détecter les cas où l'estimateur direct est plus efficace que l'estimateur EB mais ces seuils conduisent également à choisir un peu trop souvent l'estimateur direct alors qu'il était moins efficace que l'estimateur EB. Cette constatation est particulièrement notable pour le diagnostic 2. On peut réduire cette erreur en réduisant les seuils mais on détectera alors moins souvent les situations où l'estimateur direct était plus efficace que l'estimateur EB. Tel que noté précédemment, le diagnostic 1 semble avoir un comportement plus désirable que le diagnostic 2, peu importe les seuils choisis, avec des taux de sélection de l'estimateur direct très petits quand l'estimateur direct est nettement moins efficace que l'estimateur EB. Cela semble montrer les limites d'une approche entièrement fondée sur le plan de sondage, telle que celle présentée à la section 4.2, pour faire face au défi posé par l'observation de petites tailles d'échantillon dans les domaines.

Figure 6.3 Taux de sélection de l'estimateur direct pour le diagnostic 1.



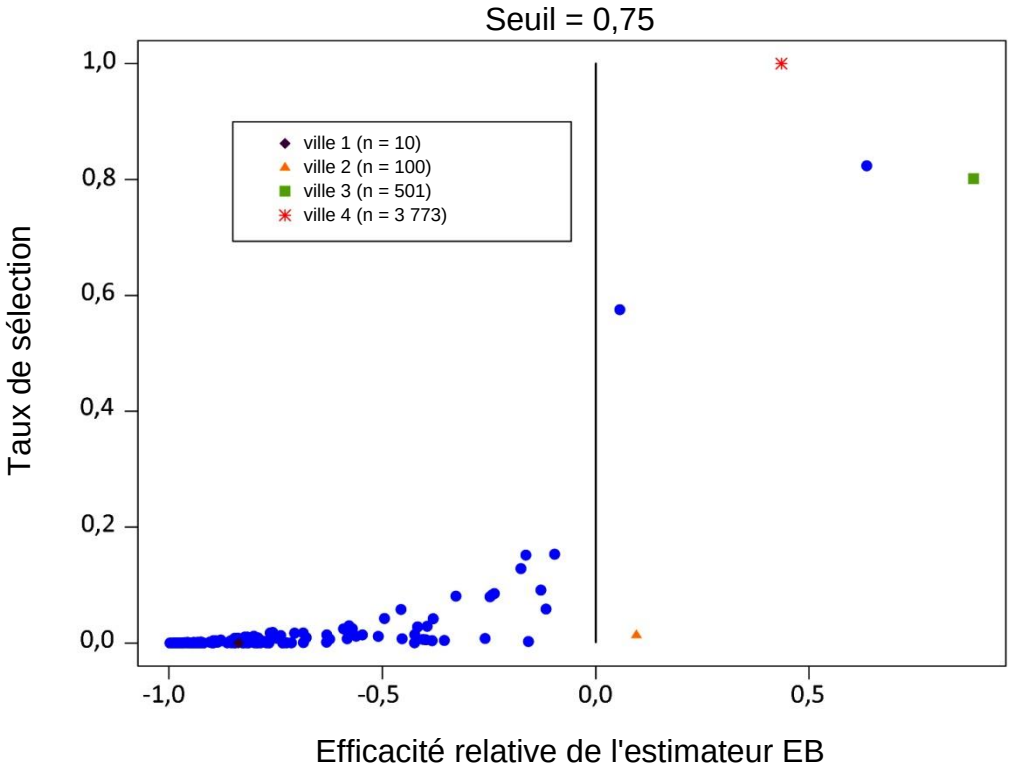
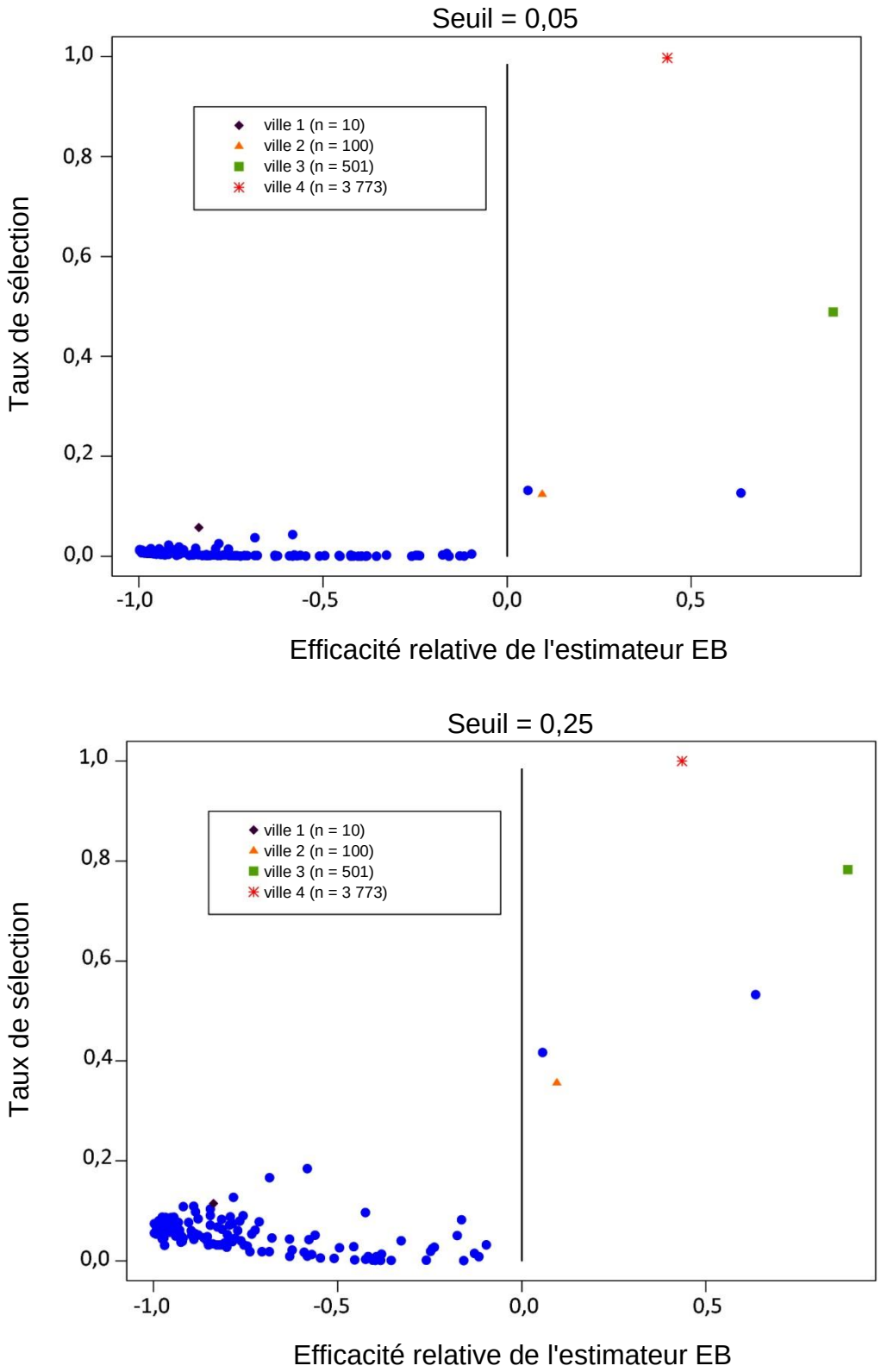


Figure 6.4 Taux de sélection de l'estimateur direct pour le diagnostic 2.



7. Conclusion

Les utilisateurs d'estimations sur petits domaines ne s'intéressent en général qu'à un unique domaine. Ils désirent donc avoir un indicateur de qualité qui s'applique à leur domaine et non un indicateur global. L'EQM fondée sur le plan de sondage des estimateurs sur petits domaines est un indicateur de qualité conceptuellement attrayant puisqu'il conditionne sur l'effet local inexpliqué du modèle. Il est cependant connu que les estimateurs sans biais de l'EQM fondée sur le plan de sondage sont généralement instables quand la taille de l'échantillon dans le domaine est petite. Pour contourner ce problème, nous avons proposé deux diagnostics qui ont pour but de déceler les domaines pour lesquels l'EQM fondée sur le plan de sondage de l'estimateur direct est plus petite que celle de l'estimateur EB. Nos résultats de simulation semblent prometteurs et permettent d'envisager la mise au point d'un indicateur utile pour choisir entre l'estimateur direct et l'estimateur EB pour un domaine en particulier. Dans la recherche future, il serait intéressant d'évaluer l'efficacité d'un estimateur hybride qui s'appuierait sur ces diagnostics.

Annexe

A. Preuve de l'équivalence entre les équations (4.1) et (4.2)

À partir de l'équation (4.1) et de la distribution conditionnelle de v_i donnée à la section 4.1, on a :

$$\begin{aligned} D_{1i} &= \text{Prob}\left(-v_{L,i} \le v_i \le v_{L,i} \middle| \mathcal{Z}, \hat{\theta}_i\right) \\ &= \text{Prob}\left(\frac{-v_{L,i} - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \le \frac{v_i - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \le \frac{v_{L,i} - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \middle| \mathcal{Z}, \hat{\theta}_i\right). \end{aligned}$$

En remplaçant $v_{L,i}$ par $\sigma_v\sqrt{\frac{1+\gamma_i}{\gamma_i}}$, on obtient :

$$\begin{aligned} D_{1i} &= \text{Prob}\left(\frac{-\sqrt{(1+\gamma_i)/\gamma_i} - \sqrt{\gamma_i}\varepsilon_i}{\sqrt{1-\gamma_i}} \le \frac{v_i - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \le \frac{\sqrt{(1+\gamma_i)/\gamma_i} - \sqrt{\gamma_i}\varepsilon_i}{\sqrt{1-\gamma_i}} \middle| \mathcal{Z}, \hat{\theta}_i\right) \\ &= \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(-\varepsilon_i + \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\} - \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(-\varepsilon_i - \sqrt{\frac{1+\gamma_i}{\gamma_i}}\right)\right\}. \end{aligned}$$

Puisque pour n'importe quelle valeur t, on a $\Phi(t) = 1 - \Phi(-t)$ alors

$$D_{1i} = \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(\varepsilon_i + \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\} - \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(\varepsilon_i - \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\}.$$

On remarque que D_{1i} est une fonction symétrique de ε_i autour de 0, c'est-à-dire que $D_{1i}(\varepsilon_i) = D_{1i}(-\varepsilon_i)$. Par conséquent, on peut ré-écrire D_{1i} comme à l'équation (4.2) :

$$D_{1i} = \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\varepsilon_i| + \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\} - \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\varepsilon_i| - \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\}.$$

B. Valeur-p associée à la statistique de test $|\varepsilon_i|$

Rappelons d'abord que $P_i(v_i) = \text{Prob}\left(|\varepsilon_i| > |\varepsilon_{\text{obs},i}| \,\middle|\, \Omega; v_i\right)$. On définit la valeur-$p$ comme le maximum de $P_i(v_{L,i})$ et $P_i(-v_{L,i})$. Puisque que $\tau_i = \frac{|\varepsilon_{\text{obs},i}| - \sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}$, on peut écrire :

$$\begin{aligned} P_i(v_i) &= \text{Prob}\left(|\varepsilon_i| > \sqrt{1+\gamma_i} + \sqrt{1-\gamma_i}\,\tau_i \,\middle|\, \Omega; v_i\right) \\ &= \text{Prob}\left(\varepsilon_i > \sqrt{1+\gamma_i} + \sqrt{1-\gamma_i}\,\tau_i \,\middle|\, \Omega; v_i\right) \\ &\quad + \text{Prob}\left(\varepsilon_i < -\sqrt{1+\gamma_i} - \sqrt{1-\gamma_i}\,\tau_i \,\middle|\, \Omega; v_i\right). \end{aligned}$$

En utilisant la loi du résidu normalisé (4.3), on obtient :

$$\begin{aligned} P_i(v_i) &= \text{Prob}\left(\frac{\varepsilon_i - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} > \frac{\sqrt{1+\gamma_i} - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} + \tau_i \,\middle|\, \Omega; v_i\right) \\ &\quad + \text{Prob}\left(\frac{\varepsilon_i - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} < \frac{-\sqrt{1+\gamma_i} - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} - \tau_i \,\middle|\, \Omega; v_i\right) \\ &= \Phi\left(\frac{-\sqrt{1+\gamma_i} + v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} - \tau_i\right) \\ &\quad + \Phi\left(\frac{-\sqrt{1+\gamma_i} - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} - \tau_i\right). \end{aligned}$$

En utilisant l'expression $v_{L,i} = \sigma_v\sqrt{\frac{1+\gamma_i}{\gamma_i}}$, on a :

$$\begin{aligned} P_i(v_i) &= \Phi\left(-\tau_i + \frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}\left[\frac{v_i}{v_{L,i}} - 1\right]\right) \\ &\quad + \Phi\left(-\tau_i + \frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}\left[-\frac{v_i}{v_{L,i}} - 1\right]\right). \end{aligned}$$

Sous l'hypothèse nulle H_0, $v_i = v_{L,i}$ ou $v_i = -v_{L,i}$ et dans les deux cas l'équation ci-dessus se réduit à :

$$P_i(v_{L,i}) = P_i(-v_{L,i}) = \Phi(-\tau_i) + \Phi\left(-\tau_i - 2\frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}\right).$$

On va maintenant montrer que si on rejette H_0 (avec un seuil plus petit que 0,5 tel que 0,1) alors on rejetterait encore plus fortement l'hypothèse nulle $H_0^*: |v_i| = v_i^*$ pour n'importe quelle valeur $0 \le v_i^* < v_{L,i}$. Premièrement, si $\tau_i \le 0$, c'est-à-dire que $|\varepsilon_{\text{obs},i}| \le \sqrt{(1+\gamma_i)}$, on observe que $P_i(v_{L,i}) = P_i(-v_{L,i}) \ge 0{,}5$ et on ne rejette jamais l'hypothèse H_0. Deuxièmement, si $\tau_i > 0$, on peut facilement montrer que la fonction $P_i(v_i)$ est croissante en v_i sur l'intervalle $[0, v_{L,i}]$. On note également que c'est une fonction de v_i qui est symétrique autour de l'abscisse $v_i = 0$ puisque $P_i(v_i) = P_i(-v_i)$. Par conséquent, $P_i(v_i)$ est

décroissante sur l'intervalle $\left[-v_{L,i}, 0\right]$, est minimale en $v_i = 0$ et maximale en $v_i = v_{L,i}$ et $v_i = -v_{L,i}$. Lorsque $|v_i| < v_{L,i}$, on a donc :

$$P_i\left(v_i\right) < P_i\left(v_{L,i}\right) = \Phi\left(-\tau_i\right) + \Phi\left(-\tau_i - 2\frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}\right).$$

Bibliographie

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

Hidiroglou, M.A., Beaumont, J.-F. et Yung, W. (2019). Élaboration d'un système d'estimation sur petits domaines à Statistique Canada. *Techniques d'enquête*, 45, 1, 107-133. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019001/article/00009-fra.pdf.

Pfeffermann, D., et Ben-Hur, D. (2019). Estimation of randomisation mean square error in small area estimation. *Revue Internationale de Statistique*, 87, S1, S31-S49.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Rao, J.N.K., Rubin-Bleuer, S. et Estevao, V.M. (2018). Mesure de l'incertitude associée aux estimateurs pour petits domaines basés sur un modèle. *Techniques d'enquête*, 44, 2, 163-180. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018002/article/54958-fra.pdf.

Rivest, L.-P., et Belmonte, E. (2000). Une erreur quadratique moyenne conditionnelle des estimateurs régionaux. *Techniques d'enquête*, 26, 1, 79-90. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2000001/article/5179-fra.pdf.

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 325-338
Statistique Canada, nº 12-001-X au catalogue

Estimation des faux négatifs attribuables à la création des pochettes dans le couplage d'enregistrements

Abel Dasylva et Arthur Goussanou[1]

Résumé

Dans le couplage d'ensembles de données massifs, on a recours aux pochettes pour sélectionner un sous-ensemble gérable de paires d'enregistrements quitte à perdre quelques paires appariées. Cette perte tient une grande place dans l'erreur de couplage globale, parce que les décisions relatives aux pochettes se prennent tôt dans le processus sans qu'on puisse les réviser par la suite. Mesurer le rôle que joue cette perte demeure un grand défi si on considère la nécessité de modéliser toutes les paires dans le produit cartésien des sources, et non seulement celles qui répondent aux critères des pochettes. Malheureusement, les modèles antérieurs d'erreur ne nous aident guère parce qu'ils ne respectent normalement pas cette exigence. Il sera question ici d'un nouveau modèle de mélange fini, qui ne demande ni vérifications manuelles, ni données d'entraînement, ni hypothèse d'indépendance conditionnelle des variables de couplage. Il s'applique dans le cadre d'une procédure de pochettes typique dans le couplage d'un fichier avec un registre ou un recensement exhaustif lorsque ces deux sources sont exemptes d'enregistrements en double.

Mots-clés : Indexation; ensembles massifs de données; résolution d'entités; intégration des données; apprentissage automatique; classification.

1. Introduction

Le couplage d'enregistrements vise à repérer les enregistrements d'une même personne dans un ou plusieurs fichiers (Fellegi et Sunter, 1969; Christen, 2012; Statistique Canada, 2017a). Il diffère de l'appariement statistique, qui est une méthode d'imputation où on se met à la recherche d'enregistrements de personnes qui se ressemblent (D'Orazio, Di Zio et Scanu, 2006). Il est aujourd'hui une importante méthode d'intégration des données et la création des pochettes en constitue une étape d'importance. Créer des pochettes est sélectionner un sous-ensemble gérable de paires d'enregistrements, qui contient la plupart des paires appariées, c'est-à-dire des paires dont les enregistrements proviennent d'une même personne. Fellegi et Sunter (1969, section 3.4) en donnent une définition abstraite comme la sélection d'un sous-ensemble du produit cartésien de deux sources de données. Herzog, Scheuren et Winkler (2007, page 123, deuxième paragraphe) y vont d'une définition semblable : [traduction] « La création des pochettes est un mode de réduction du nombre de paires d'enregistrements à examiner. » Christen (2012, page 28, troisième paragraphe) parle plutôt d'indexation avec la même signification : [traduction] « Pour réduire le nombre sans doute très grand de paires d'enregistrements à mettre en comparaison, on recourt communément à des techniques d'indexation […] Ces techniques éliminent les paires d'enregistrements qui n'ont guère de chances de correspondre à des paires appariées. » Nous nous servons du terme « création de pochettes » ici pour désigner ce processus qui devient essentiel lors du couplage d'ensembles de données massifs qui comprennent des millions d'enregistrements. En effet, le produit cartésien est juste trop grand. Grâce aux pochettes, on trouve un juste milieu entre les ressources de calcul et de mémoire,

1. Abel Dasylva, Statistique Canada, 100 promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6. Courriel : abel.dasylva@statcan.gc.ca; Arthur Goussanou, Statistique Canada, 100 promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6. Courriel : arthur.goussanou@statcan.gc.ca.

d'une part, et la perte de quelques paires appariées, d'autre part. Ces paires appariées correspondent à des faux négatifs et représentent une part importante de l'erreur de couplage globale, ne serait-ce que parce que les décisions relatives aux pochettes se prennent habituellement tôt dans le processus de couplage sans possibilité de les modifier par la suite. Il reste qu'on ne dispose là-dessus que de rares données empiriques, puisque ces faux négatifs ne sont jamais déclarés à quelques exceptions près, notamment dans le repérage des doublons dans une base de sondage comme cette opération est décrite par Herzog et coll. (2007, section 12.3). Dans ce rare cas où la base de sondage comprenait 176 000 enregistrements d'entreprises, on a estimé le nombre de paires appariées à 3 219, dont 3 050 avaient été détectées par les critères de pochettes pour une proportion de faux négatifs de $(3\,219 - 3\,050)/3\,219 = 5{,}25\,\%$, phénomène non négligeable si on le compare aux proportions de faux négatifs déclarées dans diverses études de couplage recensées par Bohensky (2016). De nos jours, il est tentant de minimiser les faux négatifs dûs aux pochettes en relâchant les critères de pochettes autant que le permettent les ressources de calcul. Après tout, ces ressources sont déjà considérables et croissent sans cesse en cette ère de mégadonnées. Une conséquence peu souhaitable peut en être que les paramètres de couplage probabiliste deviennent impossibles à estimer parce que la proportion de paires appariées est trop faible (Winkler, 2016, section 2.2.3.2). Ainsi, la question des faux négatifs dûs aux pochettes continue à se poser, quelles que soient les ressources de calcul disponibles. Leur estimation a toutefois présenté tout un défi, parce qu'on se doit de tenir compte de toutes les paires du produit cartésien des deux sources, et non des seules paires répondant aux critères des pochettes. À cet égard, la plupart des modèles antérieurs d'erreur n'aident guère, puisqu'ils ne respectent pas cette exigence; voir notamment à ce sujet Fellegi et Sunter (1969), Armstrong et Mayda (1993), Thibaudeau (1993), Winkler (1993), Belin et Rubin (1995), Sariyar, Borg et Pommerening (2011), Daggy, Xu, Hui, Gamache et Grannis (2013) et Chipperfield, Hansen et Rossiter (2018). Herzog et coll. (2007, chapitre 12.5) ont décrit une technique de capture-recapture qui n'a pas cet inconvénient, mais qui demeure peu pratique parce qu'exigeant des vérifications manuelles et l'indépendance conditionnelle pour un certain nombre de variables de pochettes.

Nous décrirons une nouvelle solution où ces deux exigences disparaissent. Il s'agit d'étendre le modèle de Blakely et Salmond (2002) aux situations où les enregistrements sont hétérogènes et où la population finie sous-jacente est nombreuse. Cette solution est d'abord conçue dans le cas idéal où deux sources exemptes de doublons sont liées, soit un fichier avec un registre ou un recensement exhaustif, de sorte que la décision de conserver une paire dans les pochettes dépende uniquement de ses deux enregistrements constitutifs comme pour les procédures typiques de création de pochettes (voir Christen, 2012, chapitre 4.1). Elle est intéressante dans un cadre pratique où les deux sources ont peu de doublons et où le recensement est quasiment exhaustif. Des exemples en sont le couplage entre les données fiscales et le recensement canadien pour le remplacement de questions sur le revenu (Statistique Canada, 2017b), ou encore une étude par cohorte avec couplage entre des données de mortalité et un recensement (Blakely et Salmond, 2002).

Notre exposé est structuré de la manière suivante : la section 2 présente les hypothèses, la notation et la terminologie; la section 3 indique en quoi la distribution des voisins livre une importante information sur les erreurs; les sections 4, 5 et 6 décrivent respectivement le modèle proposé de mélange fini, la procédure d'espérance-maximisation et l'étude empirique; enfin, la section 7 expose les conclusions et les travaux futurs.

2. Définitions, notation et hypothèses

Enregistrements appariés : Dans un couplage d'enregistrements comme pour d'autres problèmes de classification en mode automatique, une nette distinction doit s'établir entre la nature des entités à classer (deux enregistrements sont-ils réellement de la même entité ?) et les décisions à prendre (les enregistrements sont-ils *réputés* appartenir à la même entité ?) d'après les observations qui se font sur ces entités (quel est le degré d'accord entre les enregistrements ?). On ne s'entend toutefois pas sur les termes à employer pour désigner ces concepts clés, car le couplage d'enregistrements est une activité multidisciplinaire au croisement de la statistique, de l'épidémiologie et de l'informatique. Au premier paragraphe de leur résumé, Fellegi et Sunter (1969) écrivent en ce sens : [traduction] « Un modèle mathématique vise à créer un cadre théorique pour une solution informatique au problème consistant à reconnaître les enregistrements de deux fichiers qui représentent les mêmes personnes, objets ou événements (et qui sont donc dits « appariés »). » Le point est donc de savoir si deux enregistrements sont de la même entité. Dans leur livre (2007, page 83, dernier paragraphe), Herzog, Scheuren et Winkler utilisent le terme « appariement réel » pour le même concept. Dans la documentation spécialisée en informatique, le terme « appariement » a cependant un sens tout à fait différent. Il vise la décision de classification, dont Christen (2012) donne le meilleur exemple dans son livre intitulé « Data matching ». Dans son ouvrage, Newcombe (1988, page 105, deuxième paragraphe) déplore aussi le manque de consensus sur le sens à donner au terme « appariement » lorsqu'il écrit : [traduction] « Ce terme est utilisé diversement dans la documentation sur le couplage d'enregistrements. Ici toutefois, il n'a aucun sens technique particulier, désignant seulement un appariement d'enregistrements en fonction d'une certaine similitude (ou dissimilitude) indiquée. »

Dans les lignes qui suivent, le sens du terme sera celui que définissent Fellegi et Sunter (1969) pour les enregistrements d'une même entité (personne, entreprise, ménage, etc.). Il s'emploie aussi à propos d'une paire lorsque les enregistrements constitutifs sont appariés. Deux enregistrements sont dits *non appariés* s'ils proviennent d'entités différentes.

Population finie et sources de données : Dans l'examen de ce problème, considérons une grande population finie de N individus avec un processus d'enregistrement tel que les enregistrements d'individus différents sont indépendants les uns des autres, tout comme les erreurs d'enregistrement. Soit m la taille du fichier que l'on suppose être une variable aléatoire avec $m \leq N$ et $m \to \infty$ lorsque $N \to \infty$

$(m = O(N))$, par exemple). Soit V l'ensemble de valeurs possibles d'enregistrements dans l'une ou l'autre des sources de données, et soit v_i l'enregistrement i du fichier, où $v_i \in V$ par définition. Pour simplifier, posons que V est fini même s'il est habituellement très grand. Simplifions davantage en posant que les deux sources de données sont réellement exemptes de doublons et que le registre est exhaustif. En d'autres termes, chaque enregistrement du fichier correspond à un seul enregistrement du même individu dans le registre. Chaque enregistrement est enfin jugé complet, c'est-à-dire sans valeurs manquantes.

Stratégies de pochettes : Dans le couplage de deux grandes sources de données, la création de pochettes permet d'éliminer la vaste majorité des paires d'enregistrements d'individus différents, tout en conservant les autres paires et en économisant de précieuses ressources de calcul. Toutefois, certaines paires d'enregistrements d'un même individu se perdent inévitablement en cours de processus. Christen (2012, chapitre 4.4) a passé en revue un éventail de procédures de création de pochettes, dont la stratégie la plus simple consistant à sélectionner une paire si les enregistrements s'accordent parfaitement selon une seule clé. C'est une procédure qui est souvent posée au départ dans les études publiées sur l'analyse de données couplées (Chambers et Kim, 2016; Han et Lahiri, 2018). On sélectionne ainsi un sous-ensemble de paires en fonction d'une union de produits cartésiens dans des poststrates disjointes appelées blocs. Dans la pratique, on peut affiner cette méthode en conservant une paire si les enregistrements s'accordent parfaitement pour au moins une clé parmi plusieurs. Dans ce cas, le sous-ensemble de paires sélectionnées n'est plus l'union de produits cartésiens dans des poststrates disjointes. Dans ce qui suit, nous ne regarderons pas de tels détails, mais verrons quelle est notre capacité d'estimer fidèlement la perte attribuable à la procédure de création des pochettes dans un couplage entre un fichier et un registre ou recensement lorsque ces deux sources ont peu de doublons et que le registre ou le recensement est à peu près exhaustif. Le couplage de données fiscales avec le recensement canadien (Statistique Canada, 2017b) et une étude par cohorte dans un couplage entre des données de mortalité et un recensement (Blakely et Salmond, 2002) en sont de parfaits exemples.

Dans la suite de cet exposé, nous posons que la décision de conserver une paire dépend seulement des enregistrements constitutifs, c'est-à-dire que la décision de pochettes est l'équivalent d'une application mathématique de $V \times V$ dans $\{0,1\}$. On a là une grande catégorie de procédures de pochettes comprenant les procédures typiques dans ce domaine (Christen, 2012, section 4.4). On se trouve néanmoins à exclure les stratégies de pochettes comportant une certaine forme de partitionnement, comme le mode dit de « partitionnement de canopée » (Christen, 2012, section 4.8).

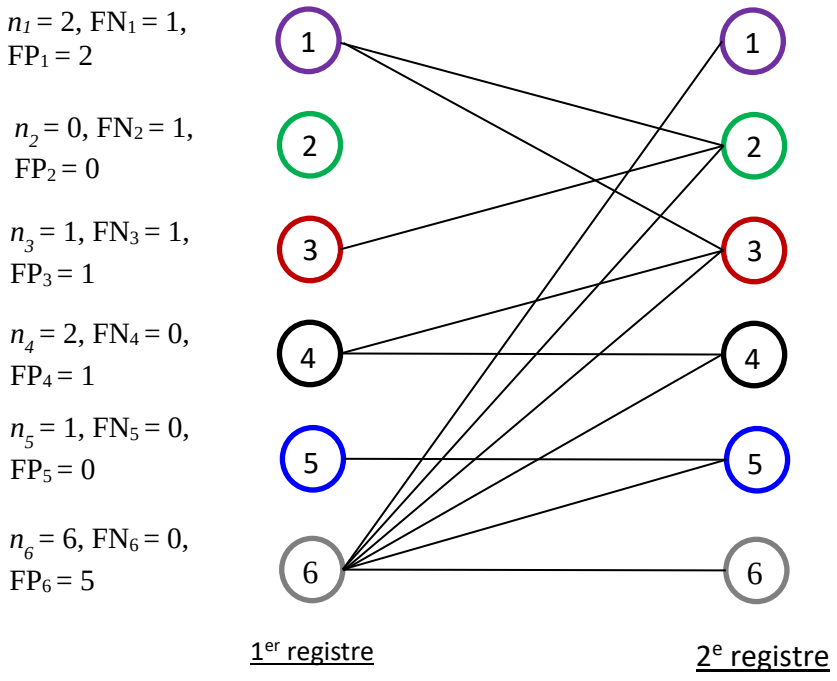
Erreurs : Dans l'application de critères de pochettes, deux types d'erreurs sont possibles sous forme de *faux négatifs* et de *faux positifs*. Il y a faux négatif si une paire appariée est rejetée par les critères applicables. Il y a faux positif si une paire non appariée est acceptée par ces mêmes critères. Ces erreurs se mesurent par le *taux de faux négatifs* (TFN) et le *taux de faux positifs* (TFP), où le premier est la proportion des paires appariées qui sont rejetées et, le second est la proportion des paires non appariées qui sont acceptées.

Lorsqu'on conçoit des critères de pochettes, on peut réduire au minimum le taux de faux positifs, tout en gardant le taux de faux négatifs sous une valeur seuil (1 %, par exemple). Comme il y a habituellement beaucoup plus de paires non appariées qu'appariées dans les pochettes, l'idée est en gros de minimiser le nombre de paires dans les pochettes, tout en gardant le taux de perte de paires appariées sous le seuil établi. Bien sûr, la mise en œuvre d'une telle stratégie exige une estimation fidèle des deux taux d'erreur en question. Le taux de faux positifs est souvent bien plus facile à estimer que le taux de faux négatifs. Soit B le nombre total de paires acceptées par les critères de pochettes. Comme le taux de faux positifs n'est pas inférieur à $(B-m)/m(N-1)$ ni supérieur à $B/m(N-1)$, il est bien approché par B/mN si $B >> m$. Cet estimateur est lié au *rapport de réduction* défini comme $1-B/mN$ (Christen, 2012, chapitre 7.3). Il est bien plus difficile d'estimer les faux négatifs. Heureusement, le concept de *voisin* jette un bon éclairage sur la question.

3. Voisins et erreurs

Lorsqu'on examine les erreurs possibles, on a intérêt à regarder combien d'enregistrements d'un registre forment une paire acceptée (par les critères de pochettes) avec un enregistrement donné du fichier. Dans ce qui suit, ces enregistrements sont appelés *voisins* de l'enregistrement du fichier et leur nombre est désigné par n_i pour l'enregistrement i du fichier. La distribution empirique de n_i nous renseigne largement sur les erreurs parce que, dans le cadre établi, chaque enregistrement du fichier aurait exactement un voisin si la stratégie de pochettes était exempte d'erreurs, c'est-à-dire sans faux négatifs ni faux positifs. À noter que le respect de cette condition n'implique pas l'absence d'erreurs. Considérons, par exemple, la situation illustrée à la figure 3.1 où les deux sources sont des registres d'une population de $N=6$ individus et où l'individu i a l'enregistrement i dans chacun des registres, c'est-à-dire que la paire d'enregistrements (i,i) est appariée pour $i=1,\ldots,6$. Avant même de regarder les n_i, nous savons que le nombre de faux négatifs est 0 ou 1, et le nombre de faux positifs, de 0 à 5 pour chaque enregistrement dans le premier registre. Toutefois, les n_i nous éclairent davantage sur les erreurs. Avec $n_i=0$ (avec l'enregistrement 2, par exemple), nous savons avec certitude qu'il existe un faux négatif, mais aucun faux positif. Avec $n_i=1$, deux cas sont possibles : dans le premier (comme avec l'enregistrement 5), le voisin est l'enregistrement apparié sans erreurs et, dans le second (comme avec l'enregistrement 3), le voisin est un enregistrement non apparié avec deux erreurs, soit un faux négatif et un faux positif. Bref, lorsque $n_i=1$, le nombre de faux négatifs est 0 ou 1, et il en va de même du nombre de faux positifs. Rien ne nous est connu d'autre sur les faux négatifs; nous savions en effet que leur nombre était 0 ou 1 avant de regarder n_i. En revanche, nous en savons beaucoup plus sur les faux positifs, puisque nous savions qu'ils se trouvaient dans un intervalle plus étendu (0 à 5) avant de regarder n_i. Cette observation confirme la facilité relative avec laquelle les faux positifs peuvent être estimés.

Figure 3.1 Deux registres avec six individus.



Le tableau 3.1 résume le lien général entre le nombre de voisins et les erreurs de couplage pour un certain enregistrement dans le cadre présent où chaque enregistrement du fichier est apparié avec exactement un seul enregistrement du registre.

Tableau 3.1
Voisins et erreurs

Voisins (n_i)	Faux négatifs	Faux positifs	Information complète sur les erreurs (oui/non)
0	1	0	Oui
1	0 ou 1	0 ou 1	Non
$[2, N-1]$	0 ou 1	$n_i - 1$ ou n_i	Non
N	0	$N-1$	Oui

Le tableau qui précède démontre clairement que le nombre de voisins nous renseigne beaucoup sur les erreurs, notamment dans le cas $n_i = 0$ et dans le cas plus improbable $n_i = N$, où l'information est complète. Là où la décision de pochettes concernant deux enregistrements ne dépend d'aucun autre enregistrement, avec une forte probabilité lorsque m et N sont élevés, un certain nombre d'enregistrements du fichier seront inévitablement sans voisin. En théorie, nous pourrions concevoir les critères de pochettes de sorte que n_i soit positif pour chaque enregistrement du fichier, mais ce serait aller

à l'encontre de l'hypothèse formulée à la section 2 d'absence de dépendance de la décision de pochettes concernant deux enregistrements à l'égard de tout autre enregistrement. Selon cette hypothèse, le nombre de voisins nous renseigne utilement sur les erreurs, mais une certaine incertitude subsiste dans le cas $1 \le n_i \le N-1$ où un modèle statistique est nécessaire à la prévision des erreurs en fonction de n_i.

4. Modèle de mélange fini

Le couplage de deux sources est intéressant s'il s'agit d'une possibilité viable, même lorsque N est très grand. Pour traduire l'essentiel dans de telles situations, nous posons les deux conditions suivantes en matière de régularité :

a. deux enregistrements appariés sont voisins avec une probabilité bornée loin de 0, quel que soit N;
b. deux enregistrements non appariés sont des voisins *accidentels* avec une probabilité de $O(1/N)$.

Ces hypothèses impliquent que chaque enregistrement a un nombre espéré de voisins qui est borné et que $O(N)$ paires (plutôt que $O(mN)$ paires et même $O(N^2)$ paires si $m = O(N)$) sont sélectionnées par les critères de pochettes. Une autre implication est que les variables de couplage donnent assez d'information pour permettre de reconnaître les enregistrements appariés avec une probabilité de succès qui est bornée loin de zéro indépendamment de la taille de la population. Une dernière implication avec ces hypothèses est l'existence d'une distribution asymptotique particulière pour le nombre de voisins n_i. Soit $n_i = n_{i|M} + n_{i|U}$, où $n_{i|M}$ est le nombre de voisins appariés et $n_{i|U}$ le nombre de voisins non appariés. À noter que ces dernières variables ne sont pas directement observées sauf si $n_i = 0$ ou $n_i = N$ (voir le tableau 3.1). Elles sont conditionnellement indépendantes étant donné v_i de sorte que $n_{i|M} \,|\, v_i \sim \text{Bernoulli}\,(p(v_i))$, $n_{i|U} \,|\, v_i \sim \text{Binomial}\,(N-1, \lambda(v_i)/(N-1))$, si un enregistrement non apparié est un voisin avec la probabilité $\lambda(v_i)/(N-1)$ indépendamment des autres enregistrements non appariés. Là où les fonctions $p(.)$ et $\lambda(.)$ ne dépendent pas de N et où N est élevé, nous avons $n_{i|U} \,|\, v_i \dot{\sim} \text{Poisson}\,(\lambda(v_i))$ (Billingsley, 1995), où $\dot{\sim}$ signifie « approximativement distribué comme ». Dans ce cas, $n_i \,|\, v_i \dot{\sim} \text{Bernoulli}\,(p(v_i)) * \text{Poisson}\,(\lambda(v_i))$, où $*$ est l'opérateur de convolution. À noter qu'en général, les fonctions $p(.)$ et $\lambda(.)$ sont des paramètres inconnus de haute dimension. Pour simplifier, posons également que $(p(.), \lambda(.))$ est représenté (bien approché) par une fonction constante par morceaux avec G niveaux, ce qui nous donne le modèle demélange fini $n_i \sim \sum_{g=1}^{G} \alpha_g\,(\text{Bernoulli}\,(p_g) * \text{Poisson}\,(\lambda_g))$ qui se vérifie approximativement. Avec G fixe, les paramètres inconnus du modèle sont donnés par le vecteur $\psi = \left[(\alpha_g, p_g, \lambda_g)\right]_{1 \le g \le G}$ qui peut être estimé par la procédure d'espérance-maximisation (EM) à la prochaine section.

Le lien entre les taux d'erreur et les paramètres du modèle s'établit si on note d'abord que les définitions du TFN et du TFP impliquent ce qui suit :

$$\begin{aligned} \text{TFN} &= \frac{1}{m}\sum_{i=1}^{m}(1-n_{i|M}), \\ (N-1)\,\text{TFP} &= \frac{1}{m}\sum_{i=1}^{m} n_{i|U}. \end{aligned} \tag{4.1}$$

Si $m = N$ presque sûrement, les équations qui précèdent impliquent que

$$\begin{aligned} E[\text{TFN}] &= 1 - E[n_{i|M}], \\ &= 1 - E[p(v_i)], \\ (N-1)\,E[\text{TFP}] &= E[n_{i|U}] \\ &= E[\lambda(v_i)], \end{aligned} \tag{4.2}$$

où $E[p(v_i)] = \sum_{g=1}^{G} \alpha_g p_g$ et $E[\lambda(v_i)] = \sum_{g=1}^{G} \alpha_g \lambda_g$ avec le modèle de mélange fini. Si m est aléatoire de sorte que

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m} n_{i|M} &\xrightarrow{p} E[n_{i|M}], \\ \frac{1}{m}\sum_{i=1}^{m} n_{i|U} &\xrightarrow{p} E[n_{i|U}], \end{aligned} \tag{4.3}$$

et si $N \to \infty$, les taux d'erreur et les paramètres du modèle sont liés de la manière suivante :

$$\begin{aligned} \text{TFN} &\xrightarrow{p} 1 - E[p(v_i)], \\ (N-1)\,\text{TFP} &\xrightarrow{p} \quad E[\lambda(v_i)]. \end{aligned} \tag{4.4}$$

5. Procédure d'estimation

Nous pouvons estimer les paramètres du modèle en maximisant la vraisemblance *composite* (Varin, Reid et Firth, 2011) de l'échantillon $n_1, \ldots, n_m$. Dans ce qui suit, nous emploierons le terme « vraisemblance » tout court pour désigner la vraisemblance composite. Dans l'application de la procédure EM (Dempster, Laird et Rubin, 1977), il est commode de calculer d'abord les équations de maximum de vraisemblance pour les *données complètes* de la manière suivante : $n_{i|M}$, $n_{i|U}$ et $(c_{i1}, \ldots, c_{iG})$ pour chaque i; c_{ig} est l'indicateur d'appartenance de l'enregistrement i à la classe g.

Avec un peu d'algèbre, voici les équations de maximum de vraisemblance pour les données complètes :

$$\begin{aligned} \hat{p}_g &= \frac{\sum_{i=1}^{m} c_{ig} n_{i|M}}{\sum_{i=1}^{m} c_{ig}} \\ \hat{\lambda}_g &= \frac{\sum_{i=1}^{m} c_{ig} n_{i|U}}{\sum_{i=1}^{m} c_{ig}} \\ \hat{\alpha}_g &= \frac{1}{m}\sum_{i=1}^{m} c_{ig}. \end{aligned} \tag{5.1}$$

Par conséquent, les équations applicables aux données observées (les n_i) sont les suivantes :

$$\begin{aligned}
\hat{p}_g &= \frac{\sum_{i=1}^{m} E[c_{ig} n_{i|M} \mid n_i;\psi]}{\sum_{i=1}^{m} E[c_{ig} \mid n_i;\psi]} \\
\hat{\lambda}_g &= \frac{\sum_{i=1}^{m} E[c_{ig} n_{i|U} \mid n_i;\psi]}{\sum_{i=1}^{m} E[c_{ig} \mid n_i;\psi]} \\
\hat{\alpha}_g &= \frac{1}{m}\sum_{i=1}^{m} E[c_{ig} \mid n_i;\psi].
\end{aligned} \tag{5.2}$$

La procédure EM alterne entre l'étape M à l'équation (5.2) et les équations de l'étape E à l'annexe A.

La procédure en question peut produire des estimateurs ponctuels convergents même si elle traite l'échantillon $n_1, \ldots, n_m$ comme s'il était indépendant et identiquement distribué. Il peut cependant en résulter un certain biais lors de l'estimation de la variance et des valeurs critiques de test d'hypothèse.

6. Étude empirique

L'étude empirique est fondée sur les données de dotation du Système de ressourcement de la fonction publique (SRFP) au service des demandeurs d'emploi dans la fonction publique fédérale au Canada. Un utilisateur peut ouvrir autant de comptes et solliciter autant d'emplois qu'il le désire avec le même compte; chaque compte est lié à une adresse électronique distincte. Dans l'exécution de son mandat, la Commission de la fonction publique doit recenser tous les comptes d'un même demandeur. La chose est difficile, parce qu'il n'y a pas d'identificateur unique, sauf pour une minorité de demandeurs. Pour la plupart des enregistrements, le couplage doit plutôt se faire avec le prénom, le nom et une date partielle de naissance qui peuvent être tirés de tous les enregistrements. La date de naissance partielle comprend le jour et le mois et le dernier chiffre de l'année de naissance.

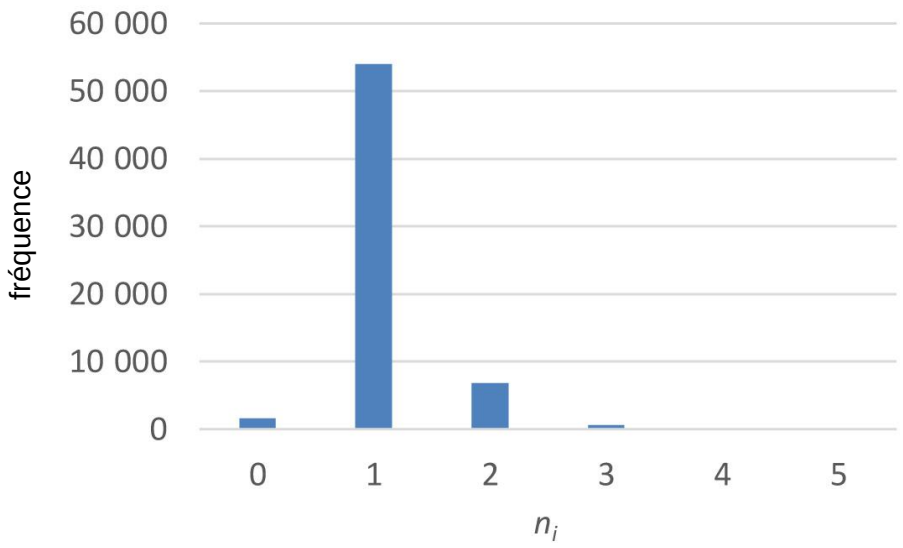
L'étude empirique fait appel à un sous-ensemble de 126 330 enregistrements prélevés sur les données SRFP depuis 2006. Les critères de sélection sont les suivants :

- identificateur unique non manquant;
- prénom, nom et date de naissance partielle non manquants;
- deux enregistrements pour chaque valeur sélectionnée de l'identificateur unique.

Les enregistrements sélectionnés correspondent à 63 155 valeurs distinctes de l'identificateur et à autant d'individus à raison de deux enregistrements appariés par individu. Ces enregistrements sont répartis entre deux registres complets et exempts de doublons qui sont liés et soumis aux critères de pochettes qui suivent, et ce, sans l'identificateur unique. Une paire est sélectionnée si la date de naissance partielle est la même, tout comme le code SOUNDEX (Herzog et coll., 2007, chapitre 11) pour le prénom ou le nom. Les taux d'erreur espérés s'estiment avec le modèle et se comparent aux valeurs réelles basées sur l'identificateur unique.

L'histogramme de la figure 6.1 indique que, dans leur vaste majorité, les enregistrements ont exactement un seul voisin. Néanmoins, 1 659 enregistrements sont sans voisin et cinq en ont cinq, soit le maximum de voisins pour tout enregistrement.

Figure 6.1 Histogramme du nombre de voisins.



Le tableau 6.1 croise les enregistrements selon le nombre de voisins et les erreurs de couplage conformément au tableau 3.1.

Tableau 6.1
Nombre de voisins et d'erreurs

Voisins (n_i)	Faux négatifs	Faux positifs	Fréquence
0	1	0	1 659
1	1	1	116
1	0	0	53 835
2	1	2	8
2	0	1	6 867
3	1	3	1
3	0	2	602
4	0	3	62
5	0	4	5

La matrice de confusion est la suivante :

Tableau 6.2
Matrice de confusion

	Lien	Non-lien	Total
Paires appariées	61 371	1 784	63 155
Paires non appariées	8 412	3,99E9	3,99E9
Total	69 783	3,988E9	3,989E9

Dans cette matrice, $\text{TFN} = 1\,784 / 63\,155 = 0{,}0282$ et $\text{TFP} = 8\,412 / 3{,}99\text{E}9 = 2{,}11\text{E}-6$. Les deux mesures peuvent être considérées comme les estimateurs $\hat{E}[\text{TFN}]$ et $\hat{E}[\text{TFP}]$ des espérances correspondantes respectives. Comme le taux de faux négatifs est la sommation de variables aléatoires indépendantes et identiquement distribuées, sa variance peut s'estimer par

$$\text{vâr}(\text{TFN}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (1 - n_{i|M} - \text{TFN})^2,$$

en fonction des variables latentes $n_{1|M}, \ldots, n_{m|M}$, qui ne sont pas directement observées dans la pratique. Ainsi, la variance estimée du TFN est $\text{vâr}(\text{TFN}) = 4{,}35\text{E}-7$. L'erreur-type estimée est donc $\hat{\text{S}}\text{E}(\hat{E}[\text{TFN}]) = 6{,}6\text{E}-4$ pour l'estimateur $\hat{E}[\text{TFN}]$ et l'intervalle normal de confiance à 95 % est $\hat{E}[\text{TFN}] \mp z_{\alpha/2} \hat{\text{S}}\text{E}(\hat{E}[\text{TFN}]) = (2{,}82\text{E}-2 \mp 1{,}3\text{E}-3)$ pour le TFN espéré, où $\alpha = 0{,}05$ et $z_{\alpha/2} = 1{,}96$. L'intervalle de confiance à 99 % correspondant est $(2{,}82\text{E}-2 \mp 1{,}71\text{E}-3)$. Il est plus difficile d'estimer la variance du TFP, puisque ce taux fait intervenir une statistique U du second ordre (Hoeffding, 1948; Lee, 1990). En fait, le tableau 6.1 ne donne pas suffisamment d'information pour l'estimation de cette statistique. Il est tout autant difficile d'estimer la variance des estimateurs par modèle parce que les n_i sont corrélés. Toutes les estimations ponctuelles figurent au tableau 6.3 où la première ligne donne les valeurs réelles du TFN et du TFP.

Tableau 6.3
Estimations ponctuelles

		$\hat{E}$**[TFN]**	$\hat{E}$**[TFP]**
Identificateur unique		0,0282	2,11E-6
Modèle	G = 1	0,0301	2,14E-6
	G = 2	0,0298	2,13E-6
	G = 3	0,0303	2,14E-6

Les résultats indiquent que les estimations par le modèle sont très proches des valeurs réelles du TFN et du TFP si on emploie une, deux ou trois classes. Pour le taux de faux négatifs, l'erreur relative est $100 \times |0{,}0303 - 0{,}0282| / 0{,}0282 = 7{,}45\,\%$, et, pour le taux de faux positifs, $100 \times |2{,}11 - 2{,}14| / 2{,}11 = 1{,}42\,\%$. La faiblesse des erreurs relatives est encourageante pour ce qui est de la précision des estimateurs proposés, bien que les estimations par le modèle du TFN espéré se situent quelque peu en dehors de l'intervalle de confiance à 95 %. Il reste que l'estimation se trouve bel et bien dans l'intervalle de confiance à 99 % lorsque deux classes sont utilisées. Le choix de deux classes paraît optimal, puisque l'estimation résultante accuse l'erreur relative la plus basse par rapport à la valeur réelle du TFN.

7. Conclusions et travaux futurs

Nous avons proposé un nouveau modèle de mélange fini pour l'estimation des faux négatifs dans une procédure de création de pochettes typique lors du couplage entre un fichier et un registre ou un recensement exhaustif et là où les deux sources sont exemptes de doublons. Une étude empirique avec des données sociales donne des résultats encourageants. Il faudra cependant poursuivre les travaux pour étudier les questions d'estimation de variance et d'inférence statistique sur le nombre de classes. Il faudra également pousser le traitement du sous-dénombrement et des doublons.

Avertissement

Cet article expose les opinions des auteurs qui ne sont pas nécessairement celles de Statistique Canada. Il décrit des méthodes théoriques qui pourraient ne pas correspondre à celles qu'emploie actuellement l'organisme.

Remerciements

Les auteurs expriment leur gratitude envers le Dr. Jonnagada Rao pour son éclairage et envers la Commission de la fonction publique du Canada pour l'accès aux données.

Annexe A

Pour l'étape E, les équations sont les suivantes :

$$
\begin{aligned}
P(n_i \mid c_{ig}=1) &= I(n_i=0)(1-p_g)\,e^{-\lambda_g} + I(n_i>0)\left(p_g+(1-p_g)\frac{\lambda_g}{n_i}\right)\frac{e^{-\lambda_g}\lambda_g^{n_i-1}}{(n_i-1)!} \\
P(c_{ig}=1 \mid n_i) &= \frac{\alpha_g P(n_i \mid c_{ig}=1)}{\sum_{g'=1}^{G}\alpha_{g'}P(n_i \mid c_{ig'}=1)} \\
P(n_{i|M}=1 \mid n_i, c_{ig}=1) &= \frac{p_g n_i}{p_g n_i + (1-p_g)\lambda_g} \\
P(n_{i|U}=n_i \mid n_i, c_{ig}=1) &= I(n_i=0) + I(n_i>0)\frac{(1-p_g)\lambda_g}{p_g n_i + (1-p_g)\lambda_g} \\
P(n_{i|U}=n_i-1 \mid n_i, c_{ig}=1) &= \frac{p_g n_i}{p_g n_i + (1-p_g)\lambda_g}
\end{aligned}
$$

et

$$
\begin{aligned}
E\,[c_{ig} n_{i|M} \,|\, n_i] &= P\,(c_{ig} = 1 \,|\, n_i)\, P\,(n_{i|M} = 1 \,|\, n_i, c_{ig} = 1) \\
E\,[c_{ig} n_{i|U} \,|\, n_i] &= P\,(c_{ig} = 1 \,|\, n_i)\, E\,[n_{i|U} \,|\, n_i, c_{ig} = 1] \\
E\,[n_{i|U} \,|\, n_i, c_{ig} = 1] &= \left(\frac{p_g\,(n_i - 1) + (1 - p_g)\,\lambda_g}{p_g n_i + (1 - p_g)\,\lambda_g} \right) n_i .
\end{aligned}
$$

Bibliographie

Armstrong, M., et Mayda, J. (1993). Estimation modéliste des taux d'erreur liés au couplage d'enregistrements. *Techniques d'enquête*, 19, 2, 147-158. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993002/article/14459-fra.pdf.

Belin, T., et Rubin, D. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Billingsley, P. (1995). *Probability and Measure*, Wiley.

Blakely, T., et Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predicted value. *Journal of Epidemiology*, 31, 1246-1252.

Bohensky, M. (2016). Bias in data linkage studies. Dans *Methodological Developments in Data Linkage*, (Éds., K. Harron, H. Goldstein et C. Dibben), Chichester: Wiley, 63-82.

Chambers, R., et Kim, G. (2016). Secondary analysis of linked data. Dans *Methodological Developments in Data Linkage*, (Éds., K. Harron, H. Goldstein et C. Dibben), Chichester: Wiley, 83-108.

Chipperfield, J., Hansen, N. et Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *Revue Internationale de Statistique*, 86, 219-236.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection*, Springer.

Daggy, J., Xu, H., Hui, S., Gamache, R. et Grannis, S. (2013). A practical approach for incorporating dependence among fields in probabilistic record linkage. *BMC Medical Informatics and Decision Making*, 13, 1-8.

Dempster, A., Laird, N. et Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39, 1-38.

D'Orazio, M., Di Zio, M. et Scanu, M. (2006). *Statistical Matching, Theory and Practice*, Wiley.

Fellegi, I., et Sunter, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Han, Y., et Lahiri, P. (2018). Statistical analysis with linked data. *Revue Internationale de Statistique*, 87, numéro spécial, 139-157, 1C19 doi:10.1111/insr.12295.

Herzog, T., Scheuren, F. et Winkler, W. (2007). *Data Quality and Record Linkage Techniques*, Springer.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Statistics*, 19, 293-325.

Lee, A.J. (1990). *U-Statistics: Theory and Practice*, Marcel Dekker.

Newcombe, H. (1988). *Handbook of Record Linkage*, Oxford University Press.

Sariyar, M., Borg, A. et Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44, 648-654.

Statistique Canada (2017a). *Modèle du processus d'un projet de couplage d'enregistrements*, nº 12-605-X au catalogue.

Statistique Canada (2017b). *Guide de référence sur le revenu, Recensement de la population, 2016*, nº 98-500-X2016004 au catalogue.

Thibaudeau, Y. (1993). Le pouvoir discriminant des strructures de dépendance dans le couplage d'enregistrements. *Techniques d'enquête*, 19, 1, 35-43. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993001/article/14477-fra.pdf.

Varin, C., Reid, N. et Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5-42.

Winkler, W. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. Dans *Proceedings of the 1993 Joint Statistical Meetings*, American Statistical Association, 274-279.

Winkler, W.E. (2016). Probabilistic linkage. Dans *Methodological Developments in Data Linkage*, (Éds., K. Harron, H. Goldstein et C. Dibben), Chichester: Wiley, 7-35.

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 339-375
Statistique Canada, nº 12-001-X au catalogue

Estimation de la variance par le bootstrap avec remise pour les enquêtes auprès des ménages Principes, exemples et mise en œuvre

Pascal Bessonneau, Gwennaëlle Brilhaut, Guillaume Chauvet et Cédric Garcia[1]

Résumé

L'estimation de la variance est un problème difficile dans les enquêtes, car plusieurs facteurs non négligeables contribuent à l'erreur d'enquête totale, notamment l'échantillonnage et la non-réponse totale. Initialement conçue pour saisir la variance des statistiques non triviales à partir de données indépendantes et identiquement distribuées, la méthode bootstrap a depuis été adaptée de diverses façons pour tenir compte des éléments ou facteurs propres à l'enquête. Dans l'article, nous examinons l'une de ces variantes, le bootstrap avec remise. Nous considérons les enquêtes auprès des ménages, avec ou sans sous-échantillonnage de personnes. Nous rendons explicites les estimateurs de la variance que le bootstrap avec remise vise à reproduire. Nous expliquons comment le bootstrap peut servir à tenir compte de l'effet de l'échantillonnage, du traitement de la non-réponse et du calage sur l'erreur d'enquête totale. Par souci de clarté, les méthodes proposées sont illustrées au moyen d'un exemple traité en fil rouge. Elles sont évaluées dans le cadre d'une étude par simulations et appliquées au Panel Politique de la Ville (PPV) français. Deux macros SAS pour exécuter les méthodes bootstrap sont également élaborées.

Mots-clés : Bootstrap; calage; estimation de variance; non-réponse totale.

1. Introduction

L'estimation de la variance est un problème difficile dans les enquêtes. Les poids finaux utilisés à l'étape de l'estimation comprennent plusieurs traitements statistiques, notamment la correction de la non-réponse totale et le calage, dont l'effet sur la variance doit être évalué. Le bootstrap est un instrument utile, qui permet de créer les poids dits bootstrap publiés avec l'ensemble de données de l'enquête. Ces poids peuvent servir à calculer de façon répétée la version bootstrap du paramètre d'intérêt, ce qui donne un estimateur de la variance ou un intervalle de confiance basés sur des simulations. L'intérêt pour les utilisateurs est le fait qu'aucune information autre que les poids bootstrap n'est requise pour l'estimation de la variance. En particulier, il n'est pas nécessaire de décrire de façon exhaustive le plan de sondage initial et le processus d'estimation, ce qui serait le cas dans le cadre d'une approche analytique où l'estimateur de la variance doit être mis au point. Ainsi, un même ensemble de poids bootstrap sert à obtenir l'estimation de la variance, que les paramètres d'intérêt soient des totaux, des médianes ou des coefficients de régression. Même si l'on dispose de la description complète du plan de sondage et du processus d'estimation, l'approche analytique pose des problèmes pour des paramètres importants pour lesquels l'estimation de la variance par linéarisation n'est pas simple; voir par exemple Shao (1994) pour les L-statistiques, et Shao et Rao (1993) pour les proportions de faible revenu.

La littérature sur le bootstrap dans l'échantillonnage d'enquête est abondante; on trouve par exemple des revues détaillées dans Rao et Wu (1988), Rao, Wu et Yue (1992), Shao et Tu (1995, chapitre 6), Davison et Hinkley (1997, section 3.7), Davison et Sardy (2007), Chauvet (2007) et Mashreghi, Haziza et

1. Pascal Bessonneau, Ined; Gwennaëlle Brilhaut, Ined; Guillaume Chauvet, Ensai (Irmar), Campus de Ker Lann, Bruz, France. Courriel : guillaume.chauvet@ensai.fr; Cédric Garcia, Université Gustave Eiffel.

Léger (2016). L'une de ces techniques est le dit *rescaled bootstrap* (bootstrap rééchelonné) proposé par Rao et Wu (1988), qui peut se résumer comme suit. Premièrement, à l'intérieur de chaque échantillon au premier degré S_h de taille n_h sélectionné dans la strate h, un échantillon aléatoire simple avec remise de taille m_h est sélectionné, ce qui donne les poids bootstrap initiaux. Ensuite, ces poids peuvent être rééchelonnés de façon à reproduire un estimateur de la variance sans biais pour l'estimation d'un total (cas linéaire). Comme l'expliquent Rao et Wu (1988), le bootstrap rééchelonné peut être appliqué à divers plans de sondage, y compris l'échantillonnage à deux degrés et l'échantillonnage avec ou sans remise au premier degré. Toutefois, il n'est pas facile de tenir compte de certaines caractéristiques pratiques des enquêtes, comme le traitement de la non-réponse totale. Cette question est examinée par Yeo, Mantel et Liu (1999) et Girard (2009). Un sujet connexe est traité dans Kim, Navarro et Fuller (2006), qui se penchent sur l'estimation de la variance par répliques pour un échantillonnage à deux phases.

L'application du bootstrap Rao-Wu dans le cas particulier où les tailles de rééchantillonnage sont $m_h = n_h - 1$ donne ce qu'on appelle bootstrap des unités primaires d'échantillonnage (UPE) ou bootstrap avec remise (McCarthy et Snowden, 1985). Le bootstrap avec remise est assez simple à mettre en œuvre, notamment parce qu'il suffit de rééchantillonner les unités primaires d'échantillonnage, et non les unités finales. La prise en compte du traitement de la non-réponse et du calage est assez naturelle, comme l'explique le présent article. Une des propriétés importantes de toute méthode bootstrap consiste à reproduire (au moins approximativement) un estimateur de la variance connu dans le cas linéaire, que nous appelons estimateur repère de la variance. Pour le bootstrap avec remise, il est possible d'énoncer précisément cet estimateur repère de la variance à toute étape de la méthode, ce qui est utile afin de comprendre le fonctionnement de la méthode pour évaluer l'erreur d'enquête totale. Le bootstrap avec remise donne une estimation prudente de la variance, en ce sens que la variance d'échantillonnage au premier degré est surestimée si les plans de sondage utilisés à l'intérieur des strates au premier degré sont plus efficaces que l'échantillonnage multinomial, ce que nous supposons vrai dans le présent article. Il s'agit donc d'une méthode prudente de production des intervalles de confiance. Le biais positif de l'estimateur de la variance bootstrap devrait être négligeable lorsque les taux de sondage au premier degré à l'intérieur des strates sont négligeables, ce qui est souvent le cas dans les enquêtes téléphoniques. De plus, si l'enquête est répétée au fil du temps, il est probable que la contribution de la variance due à l'échantillonnage au premier degré s'estompe, tandis que la variance attribuable à l'attrition et à la non-réponse totale augmente.

Notre article, qui porte sur le bootstrap avec remise, se veut axé sur les utilisateurs. C'est pourquoi nous ne proposons pas de modifications particulières du bootstrap avec remise. Nous expliquons plutôt comment appliquer cette méthode bootstrap pour tenir compte de l'échantillonnage, du traitement de la non-réponse et du calage et, ce faisant, quel est l'estimateur de la variance que nous cherchons à reproduire lors de l'estimation d'un total. Nous donnons des exemples en fil rouge pour illustrer comment les poids bootstrap sont calculés dans des cas simples. Deux macros SAS mettant en œuvre les méthodes bootstrap proposées sont présentées, évaluées au moyen d'une étude par simulations et illustrées sur un ensemble réel de données d'enquête, tirées du Panel Politique de la Ville.

Pour simplifier la présentation, notre terminologie est celle des enquêtes auprès des ménages, qui sont également la motivation première de notre article. Nous examinons deux cas : premièrement, quand un échantillon de ménages seulement est sélectionné et deuxièmement, quand un sous-échantillon de personnes est sélectionné dans les ménages sélectionnés. Malgré cette terminologie particulière, notre démarche est générale et peut être appliquée à toute autre situation où l'enquête est effectuée par sondage à un degré (premier cas) ou par sondage à deux degrés (deuxième cas).

Nous nous intéressons plus particulièrement aux enquêtes téléphoniques auprès des ménages, largement utilisées à l'Institut national d'études démographiques (Ined) français au cours des dernières décennies. À l'origine, un échantillon de numéros de téléphone était sélectionné dans un registre de numéros de téléphone fixes, et plus récemment, les numéros de téléphone utilisés dans l'enquête sont générés de façon aléatoire pour tenir compte des ménages non couverts dans les registres (numéros de téléphone non répertoriés et numéros de téléphone cellulaire). À la deuxième étape, des personnes sont sélectionnées au sein des ménages, au moyen de méthodes de sélection classiques (par exemple individu Kish). Les sondages téléphoniques ont prouvé leur efficacité, en particulier pour des sujets sensibles comme la sexualité, la violence ou les dépendances. Parmi les exemples d'enquêtes réalisées par l'Ined, citons l'enquête nationale sur les violences faites aux femmes en France (Enveff) en 2000, l'enquête Violences et rapports de genre en 2015 et 2018 (Virage et Virage Dom, respectivement), ou l'enquête nationale sur le contexte de la sexualité en France en 2006. Le même protocole sera probablement utilisé dans un proche avenir pour des enquêtes aux sujets similaires, comme l'enquête sur la sexualité des jeunes adultes ou celle sur le contrôle des naissances, qui doivent commencer entre 2021 et 2023.

L'article est organisé comme suit. À la section 2, nous définissons nos principales notations et nous considérons l'estimation d'un total en tenant compte de l'échantillonnage, de la non-réponse totale et du calage. Nous traitons à la section 2.1 la situation où un échantillon de ménages seulement est sélectionné (scénario à un degré) et à la section 2.2 le cas où des personnes sont sous-échantillonnées au sein des ménages (scénario à deux degrés). La méthode bootstrap de base est décrite à la section 3 : le scénario à un degré est examiné aux sections 3.1 et 3.2, et le scénario à deux degrés est examiné aux sections 3.3 et 3.4. Nous expliquons à la section 3.5 comment on peut appliquer la procédure bootstrap élémentaire pour obtenir un estimateur de la variance ou un intervalle de confiance. Les méthodes bootstrap proposées sont évaluées à la section 4 au moyen d'une étude par simulations. À la section 5, nous illustrons les méthodes au moyen d'un échantillon de ménages et d'individus du Panel Politique de la Ville français. Des conclusions sont données à la section 6. Les estimateurs repères de la variance pour l'échantillon de personnes sont présentés à l'annexe A. Le programme SAS qui a servi à réaliser l'estimation de la variance bootstrap est présenté aux annexes B et C. Ces programmes SAS peuvent être mis à disposition par l'auteur correspondant sur demande.

2. Notation et estimation

Dans la présente section, nous définissons nos principales notations et décrivons le processus d'échantillonnage et d'estimation. Nous examinons d'abord à la section 2.1 le cas où un échantillon de

ménages seulement est sélectionné, et nous décrivons le processus d'estimation qui comprend traitement de la non-réponse totale et calage. Nous indiquons dans chaque cas l'estimateur repère de la variance considéré, c'est-à-dire l'estimateur de la variance que nous cherchons à reproduire pour l'estimation d'un total au moyen de la méthode bootstrap proposée à la section 3. Le cas où des personnes sont sous-échantillonnées dans les ménages est traité à la section 2.2. Les estimateurs repères de la variance pour ce deuxième cas sont donnés à l'annexe A.

2.1 Scénario avec échantillon de ménages seulement

Nous considérons une estimation pour une population U_{hh} de ménages. Nous désignons par y_k la valeur prise par une variable d'intérêt pour le ménage k. Nous nous intéressons à l'estimation du total

$$Y_{hh} = \sum_{k \in U_{hh}} y_k. \tag{2.1}$$

2.1.1 Plan de sondage

Nous supposons qu'un échantillon S_{hh} est sélectionné dans U_{hh} au moyen d'un plan de sondage à un degré stratifié. La population U_{hh} est divisée en H strates $U_{hh}^1, \ldots, U_{hh}^H$, les échantillons $S_{hh}^1, \ldots, S_{hh}^H$ y sont sélectionnés indépendamment, et l'échantillon S_{hh} est l'union de ces échantillons. Soit π_k la probabilité d'inclusion d'un ménage donné k. Le poids de sondage est

$$d_k = \frac{1}{\pi_k}. \tag{2.2}$$

En cas de réponse complète, l'estimateur de Y_{hh} est

$$\hat{Y}_{hh} = \sum_{k \in S_{hh}} d_k y_k. \tag{2.3}$$

Nous considérons comme un estimateur repère de la variance

$$v_{\text{mult}}(\hat{Y}_{hh}) = \sum_{h=1}^{H} \left[\frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k y_k - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} y_{k'} \right)^2 \right], \tag{2.4}$$

avec n_h la taille de l'échantillon S_{hh}^h. Cet estimateur de la variance est sans biais si les échantillons sont sélectionnés dans les strates par échantillonnage multinomial (Tillé, 2011, section 5.4), aussi appelé échantillonnage avec remise. Il est conservatif si les plans d'échantillonnage utilisés dans les strates sont plus efficaces que l'échantillonnage multinomial (Särndal, Swensson et Wretman, 1992, section 4.6), ce que nous supposons vrai dans le reste de l'article. Le biais positif de cet estimateur de la variance devrait être négligeable quand les taux de sondage à l'intérieur des strates sont eux-mêmes négligeables, ce qui est souvent le cas dans les enquêtes téléphoniques. Les résultats de notre étude par simulations en témoignent, voir la section 4.

2.1.2 Traitement de la non-réponse

En pratique, l'échantillon S_{hh} est sujet à la non-réponse totale, ce qui mène à l'observation d'un sous-échantillon de répondants seulement. Nous désignons par r_k l'indicateur de réponse d'un ménage k et par p_k la probabilité de réponse du ménage k. Nous supposons que les ménages répondent indépendamment les uns des autres. De plus, nous supposons que la non-réponse totale est traitée par la méthode des groupes de réponse homogènes (GRH), populaire en pratique (par exemple Brick, 2013; Juillard et Chauvet, 2018). Dans ce cadre, on suppose que l'échantillon S_{hh} peut être divisé en C GRH notés $S_{1,hh},\ldots,S_{C,hh}$, de sorte que la probabilité de réponse p_k soit constante à l'intérieur d'un GRH.

Pour $c=1,\ldots,C$, p_c désigne la probabilité de réponse commune à l'intérieur du GRH $S_{c,hh}$. Elle est estimée par

$$\hat{p}_c = \frac{\sum_{k\in S_{c,hh}} \theta_k r_k}{\sum_{k\in S_{c,hh}} \theta_k}, \tag{2.5}$$

avec θ_k un certain poids donné au ménage k. Le choix $\theta_k = 1$ conduit à estimer p_c par le taux de réponse non pondéré dans le GRH. Le choix $\theta_k = d_k$ conduit à estimer p_c par le taux de réponse dans le GRH, pondéré par les poids d'échantillonnage (par exemple Kott, 2012).

La prise en compte des probabilités de réponse estimées donne les poids corrigés pour la non-réponse

$$d_{rk} = \frac{d_k}{\hat{p}_{c(k)}}, \tag{2.6}$$

avec $c(k)$ le GRH du ménage k. L'estimateur de Y_{hh} ajusté pour tenir compte de la non-réponse est

$$\hat{Y}_{r,hh} = \sum_{k\in S_{r,hh}} d_{rk} y_k. \tag{2.7}$$

Construit à partir de l'estimateur de la variance multinomial de (2.4) et de la linéarisation des estimateurs repondérés pour tenir compte de la non-réponse totale (Kim et Kim, 2007, section 2), notre estimateur repère de la variance est

$$v_{\text{mult}}(\hat{Y}_{r,hh}) = \sum_{h=1}^{H}\left[\frac{n_h}{n_h-1}\sum_{k\in S_{hh}^h}\left(d_k u_{1k} - \frac{1}{n_h}\sum_{k'\in S_{hh}^h} d_{k'} u_{1k'}\right)^2\right], \tag{2.8}$$

avec

$$u_{1k} = \theta_k \pi_k \bar{y}_{rc(k)} + \frac{r_k}{\hat{p}_{c(k)}}\left\{y_k - \theta_k \pi_k \bar{y}_{rc(k)}\right\},$$

et

$$\bar{y}_{rc} = \frac{\sum_{k \in S_{c,hh}} d_k r_k y_k}{\sum_{k \in S_{c,hh}} \theta_k r_k}.$$

Il s'agit d'un estimateur conservatif pour la variance asymptotique de $\hat{Y}_{r,hh}$. Une hypothèse essentielle à cet égard est que les indicateurs de réponse r_k sont mutuellement indépendants.

2.1.3 Calage

Enfin, les poids ajustés pour tenir compte de la non-réponse sont calés sur les totaux auxiliaires connus dans la population. Par souci de simplicité, nous décrivons seulement l'estimateur par la régression généralisée (GREG, Särndal et coll., 1992, chapitre 6). Soit x_k le vecteur des variables de calage au niveau du ménage, et X_{hh} le total sur la population U_{hh}. Pour l'échantillon $S_{r,hh}$, cela donne les poids linéaires calés

$$w_k = d_{rk}\left(1 + x_k^{\mathrm{T}} \lambda_{hh}\right),$$

avec

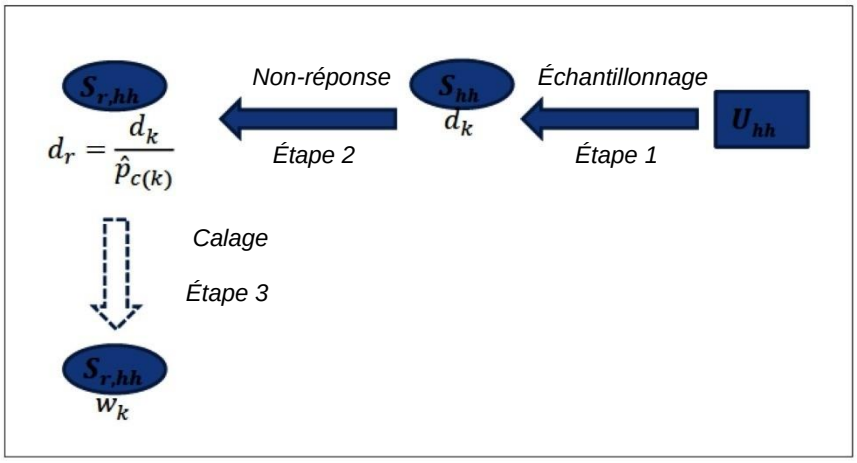
$$\lambda_{hh} = \left(\sum_{k \in S_{r,hh}} d_{rk} x_k x_k^{\mathrm{T}}\right)^{-1} \left(X_{hh} - \hat{X}_{r,hh}\right), \tag{2.9}$$

et où $\hat{X}_{r,hh}$ est l'estimateur de X_{hh}, obtenu en insérant x_k dans (2.7). L'estimateur calé est

$$\hat{Y}_{\mathrm{cal},hh} = \sum_{k \in S_{r,hh}} w_k y_k. \tag{2.10}$$

Les étapes d'échantillonnage et d'estimation sont résumées dans la figure 2.1.

Figure 2.1 Étapes d'échantillonnage et d'estimation pour un échantillon de ménage.



Construit à partir de la linéarisation des estimateurs repondérés pour tenir compte de la non-réponse totale et du calage (Kim et Kim, 2007, section 5), notre estimateur repère de la variance est

$$v_{\text{mult}}(\hat{Y}_{\text{cal},hh}) = \sum_{h=1}^{H}\left[\frac{n_h}{n_h-1}\sum_{k\in S_{hh}^h}\left(d_k u_{2k} - \frac{1}{n_h}\sum_{k'\in S_{hh}^h} d_{k'}u_{2k'}\right)^2\right], \tag{2.11}$$

avec

$$u_{2k} = \theta_k \pi_k \bar{e}_{rc(k)} + \frac{r_k}{\hat{p}_{c(k)}}\left\{e_k - \theta_k \pi_k \bar{e}_{rc(k)}\right\},$$

et

$$\bar{e}_{rc} = \frac{\sum_{k\in S_{c,hh}} d_k r_k e_k}{\sum_{k\in S_{c,hh}} \theta_k r_k},$$

où nous supposons que

$$e_k = y_k - \hat{B}_{r,hh}^{\mathrm{T}} x_k \quad \text{avec} \quad \hat{B}_{r,hh} = \left(\sum_{k\in S_{r,hh}} d_{rk} x_k x_k^{\mathrm{T}}\right)^{-1} \sum_{k\in S_{r,hh}} d_{rk} x_k y_k \tag{2.12}$$

désigne les résidus de régression estimés de la variable d'intérêt sur les variables de calage. Il s'agit d'un estimateur conservatif pour la variance asymptotique de $\hat{Y}_{\text{cal},hh}$.

2.1.4 Calcul des poids des ménages dans un exemple

Voici un petit exemple pour fixer les idées. Examinons une population U_{hh} composée de $N_{hh} = 100$ ménages. Nous supposons sans perte de généralité qu'une seule strate est utilisée et qu'un échantillon de $n_{hh} = 10$ ménages est sélectionné.

L'échantillon est $S = \{A, B, \ldots, J\}$. Les probabilités d'inclusion des unités sélectionnées sont (supposons)

$$\pi_A = \pi_B = \pi_C = \pi_D = \frac{1}{4} \quad \text{et} \quad \pi_E = \pi_F = \pi_G = \pi_H = \pi_I = \pi_J = \frac{1}{16}, \tag{2.13}$$

ce qui donne les poids de sondage

$$d_A = d_B = d_C = d_D = 4 \quad \text{et} \quad d_E = d_F = d_G = d_H = d_I = d_J = 16. \tag{2.14}$$

Parmi les 10 ménages sélectionnés, 7 seulement sont sondés en raison de la non-réponse. On en tient compte au moyen de la méthode des GRH, avec deux groupes : les unités A, B, F et J dans le premier, et les unités C, D, E, G, H et I dans le second. Les unités B, C et G sont des non-répondants. Dans chaque GRH, nous calculons les probabilités de réponse estimées, pondérées par les poids de sondage $(\theta_k = d_k)$. Cela donne

$$\hat{p}_1 = \frac{\sum_{k \in S_{1,hh}} d_k r_k}{\sum_{k \in S_{1,hh}} d_k} = \frac{d_A + d_F + d_J}{d_A + d_B + d_F + d_J} = \frac{9}{10},$$

$$\hat{p}_2 = \frac{d_D + d_E + d_H + d_I}{d_C + d_D + d_E + d_G + d_H + d_I} = \frac{13}{18}. \tag{2.15}$$

On obtient les poids tenant compte de la non-réponse pour les répondants en divisant les poids d'échantillonnage par les probabilités de réponse estimées. Cela donne les poids

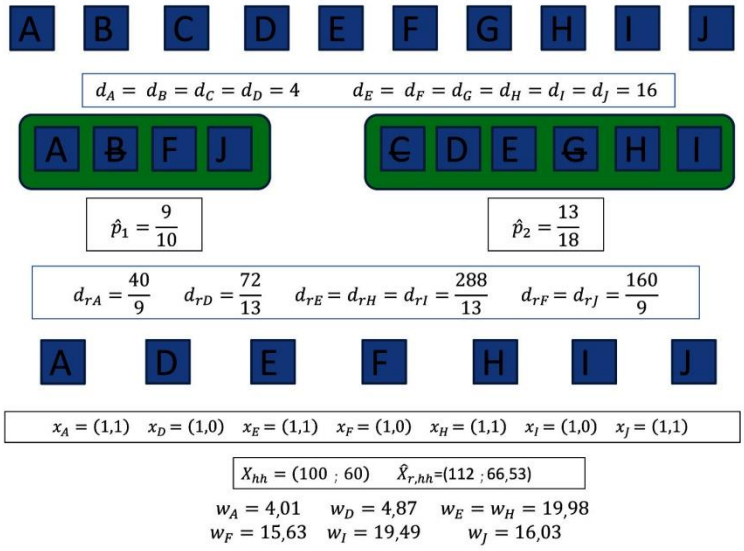
$$d_{rA} = \frac{40}{9} \quad d_{rD} = \frac{72}{13} \quad d_{rE} = d_{rH} = d_{rI} = \frac{288}{13} \quad d_{rF} = d_{rJ} = \frac{160}{9}. \tag{2.16}$$

Enfin, les poids sont calés pour qu'ils permettent de reproduire exactement la taille de la population $N_{hh} = 100$ et à un total auxiliaire $X_{1,hh} = 60$. Notons qu'en utilisant l'échantillon de répondants, nous obtenons $\hat{N}_{r,hh} = 112$ et $\hat{X}_{1r,hh} = 66{,}53$. Les poids calés sont

$$\begin{aligned} &w_A = 4{,}01; \quad w_D = 4{,}87; \quad w_E = w_H = 19{,}98; \\ &w_F = 15{,}63; \quad w_I = 19{,}49; \quad w_J = 16{,}03. \end{aligned} \tag{2.17}$$

Les étapes d'échantillonnage et d'estimation sont résumées dans la figure 2.2.

Figure 2.2 Étapes d'estimation de la pondération des ménages.



2.2 Scénario avec échantillon de ménages et de personnes

Nous nous intéressons à la population U_{ind} de personnes associées à la population U_{hh} de ménages considérée à la section 2.1. Si nous désignons par y_l la valeur prise par une variable d'intérêt pour la personne l, le paramètre d'intérêt est

$$Y_{\text{ind}} = \sum_{l \in U_{\text{ind}}} y_l. \tag{2.18}$$

2.2.1 Plan de sondage

Dans tout ménage échantillonné $k \in S_{hh}$, un sous-échantillon $S_{\text{ind},k}$ de personnes est sélectionné, et l'échantillon S_{ind} est l'union de ces échantillons. Nous désignons par $\pi_{l|k}$ la probabilité d'inclusion conditionnelle de la personne l à l'intérieur du ménage k. Le poids de sondage conditionnel de l est

$$d_{l|k} = \frac{1}{\pi_{l|k}} \quad \text{pour tout} \quad l \in k, \tag{2.19}$$

et le poids de sondage non conditionnel est

$$d_l = d_{l|k} \times d_k \quad \text{pour tout} \quad l \in k. \tag{2.20}$$

En cas de réponse complète, l'estimateur de Y_{ind} est

$$\hat{Y}_{\text{ind}} = \sum_{k \in S_{hh}} d_k \sum_{l \in S_{\text{ind},k}} d_{l|k} y_l = \sum_{k \in S_{\text{ind}}} d_l y_l. \tag{2.21}$$

On obtient l'estimateur repère de la variance pour $\hat{Y}_{\text{ind}}$ à partir de (2.4), en remplaçant y_k par

$$\hat{y}_k = \sum_{l \in S_{\text{ind},k}} d_{l|k} y_l. \tag{2.22}$$

2.2.2 Traitement de la non-réponse

En tenant compte de la non-réponse des ménages, les poids des personnes sont

$$d_{rl} = d_{rk(l)} d_{l|k(l)} \quad \text{avec } k(l) \text{ le ménage contenant } l, \tag{2.23}$$

avec d_{rk} le poids du ménage k corrigé pour tenir compte de la non-réponse totale (voir l'équation (2.6)) et $d_{l|k}$ le poids d'échantillonnage conditionnel de la personne l dans le ménage k (voir l'équation (2.19)). Soit

$$S_{r,\text{ind}} = \bigcup_{k \in S_{r,hh}} S_{\text{ind},k} \tag{2.24}$$

l'ensemble de toutes les personnes échantillonnées à l'intérieur des ménages répondants.

Les personnes dans $S_{r,\text{ind}}$ sont elles-mêmes sujettes à la non-réponse, bien qu'habituellement, on s'attende à ce que ce soit dans une moindre mesure. Cela conduit à l'observation d'un sous-échantillon de répondants $S_{rr,\text{ind}}$ seulement. Nous désignons par r_l l'indicateur de réponse et par p_l la probabilité de réponse de la personne l. Nous supposons que les personnes répondent indépendamment les unes des autres. De plus, nous supposons que cette non-réponse est traitée au moyen de la méthode des GRH :

l'échantillon $S_{r,\text{ind}}$ peut être divisé en D GRH notés $S_{r1,\text{ind}},\ldots,S_{rD,\text{ind}}$ de telle sorte que la probabilité de réponse p_l soit constante à l'intérieur d'un GRH.

Pour $d=1,\ldots,D$, p_d désigne la probabilité de réponse commune à l'intérieur du GRH $S_{rd,\text{ind}}$. Elle est estimée par

$$\hat{p}_d = \frac{\sum_{l\in S_{rd,\text{ind}}} \theta_l r_l}{\sum_{l\in S_{rd,\text{ind}}} \theta_l}, \tag{2.25}$$

avec θ_l un certain poids donné à la personne l. Le choix $\theta_l=1$ conduit à estimer p_d par le taux de réponse non pondéré dans le GRH. Le choix $\theta_l=d_l$ conduit à estimer p_d par le taux de réponse dans le GRH, pondéré par les poids d'échantillonnage de la personne. Le choix $\theta_l=d_{rl}$ conduit à estimer p_d par le taux de réponse dans le GRH, pondéré par les poids d'échantillonnage de la personne corrigés pour tenir compte de la non-réponse totale du ménage. Nous comparons ces différents choix dans l'étude par simulations présentée à la section 4.

La prise en compte des probabilités de réponse estimées donne les poids individuels corrigés pour tenir compte de la non-réponse de la personne et du ménage.

$$d_{rrl} = \frac{d_{rl}}{\hat{p}_{d(l)}} \quad \text{avec } d(l) \text{ le ménage contenant } l. \tag{2.26}$$

L'estimateur de Y_{ind} ajusté pour tenir compte de la non-réponse de la personne et du ménage est

$$\hat{Y}_{rr,\text{ind}} = \sum_{l\in S_{rr,\text{ind}}} d_{rrl} y_l. \tag{2.27}$$

2.2.3 Calage

Soit z_l le vecteur des variables de calage au niveau de la personne, et Z_{ind} le total sur la population U_{ind}. Pour l'échantillon $S_{rr,\text{ind}}$, cela donne les poids linéaires calés

$$w_l = d_{rrl}\left(1 + z_l^{\mathsf{T}}\lambda_{\text{ind}}\right),$$

avec

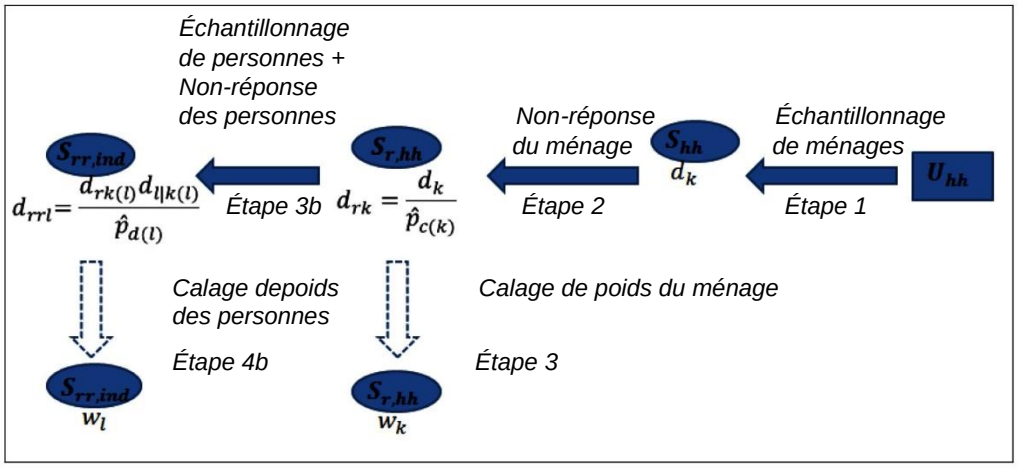
$$\lambda_{\text{ind}} = \left(\sum_{l\in S_{rr,\text{ind}}} d_{rrl} z_l z_l^{\mathsf{T}}\right)^{-1}\left(Z_{\text{ind}} - \hat{Z}_{rr,\text{ind}}\right), \tag{2.28}$$

et où $\hat{Z}_{rr,\text{ind}}$ est l'estimateur de Z_{ind}, obtenu en insérant z_l dans (2.27). L'estimateur calé est

$$\hat{Y}_{\text{cal,ind}} = \sum_{l\in S_{rr,\text{ind}}} w_l y_l. \tag{2.29}$$

Les étapes d'échantillonnage et d'estimation sont résumées dans la figure 2.3.

Figure 2.3 Étapes d'échantillonnage et d'estimation pour un échantillon de ménages avec sous-échantillonnage de personnes.



2.2.4 Calcul des poids des personnes dans un exemple

Nous poursuivons avec l'exemple présenté à la section 2.1.4. Rappelons que l'échantillon des ménages répondants est $S_{r,hh} = \{A, D, E, F, H, I, J\}$. L'ensemble de toutes les personnes à l'intérieur des ménages répondants se présente comme suit (supposons) :

$$\underbrace{(i_1, i_2, i_3)}_{A} \quad \underbrace{(i_4)}_{D} \quad \underbrace{(i_5, i_6)}_{E} \quad \underbrace{(i_7, i_8, i_9)}_{F} \quad \underbrace{(i_{10}, i_{11})}_{H} \quad \underbrace{(i_{12})}_{I} \quad \underbrace{(i_{13})}_{J}. \tag{2.30}$$

Nous supposons que le plan de sondage consiste à sélectionner exactement une personne à l'intérieur de chaque ménage. L'ensemble $S_{r,\text{ind}}$ de toutes les personnes échantillonnées à l'intérieur des ménages répondants est

$$S_{r,\text{ind}} = \{i_1, i_4, i_6, i_8, i_{11}, i_{12}, i_{13}\}. \tag{2.31}$$

À partir des équations (2.23) et (2.16), les poids des personnes corrigés pour tenir compte de la non-réponse des ménages sont par conséquent

$$d_{r1} = \frac{40}{3},\ d_{r4} = \frac{72}{13},\ d_{r6} = \frac{576}{13},\ d_{r8} = \frac{160}{3},\ d_{r11} = \frac{576}{13},\ d_{r12} = \frac{288}{13},\quad d_{r13} = \frac{160}{9}. \tag{2.32}$$

Parmi ces sept personnes sélectionnées, quatre seulement sont sondées en raison de la non-réponse, prise en compte au moyen de la méthode des groupes de réponse homogènes (GRH). Nous supposons qu'il y a deux GRH : les unités i_1, i_6, i_8 et i_{11} dans le premier, et les unités i_4, i_{12} et i_{13} dans le second. Les unités i_4, i_{11} et i_{13} sont des non-répondants. À l'intérieur de chaque GRH, nous calculons les probabilités de réponse estimées non pondérées $(\theta_l = 1)$. Cela donne

$$\hat{p}_1 = \frac{\sum_{l \in S_{r1,\text{ind}}} r_l}{\sum_{l \in S_{r1,\text{ind}}} 1} = \frac{3}{4},$$

$$\hat{p}_2 = \frac{\sum_{l \in S_{r2,\text{ind}}} r_l}{\sum_{l \in S_{r2,\text{ind}}} 1} = \frac{1}{3}. \tag{2.33}$$

On obtient les poids tenant compte de la non-réponse des personnes ou des ménages pour les répondants en divisant les poids de (2.32) par les probabilités de réponse estimées. Cela donne les poids

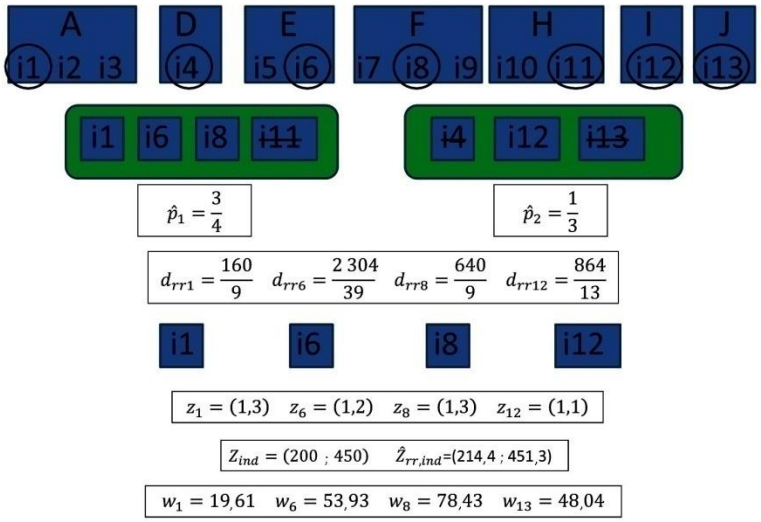
$$d_{rr1} = \frac{160}{9}, \quad d_{rr6} = \frac{2\,304}{39}, \quad d_{rr8} = \frac{640}{9}, \quad d_{rr12} = \frac{864}{13}. \tag{2.34}$$

Enfin, les poids sont calés pour qu'ils permettent de reproduire exactement la taille de la population $N_{\text{ind}} = 200$ et le total auxiliaire $Z_{1,\text{ind}} = 450$. Notons qu'en utilisant l'échantillon de répondants, nous obtenons $\hat{N}_{r,\text{ind}} = 214{,}4$ et $\hat{Z}_{1r,\text{ind}} = 451{,}3$. Les poids calés sont

$$w_1 = 19{,}61; \quad w_6 = 53{,}93; \quad w_8 = 78{,}43; \quad w_{13} = 48{,}04. \tag{2.35}$$

Les étapes d'échantillonnage et d'estimation sont résumées dans la figure 2.4.

Figure 2.4 Étapes d'estimation de la pondération des personnes.



3. Estimation de la variance bootstrap

Nous commençons à la section 3.1 par la description de l'étape élémentaire de la méthode bootstrap quand on sélectionne seulement un échantillon de ménages. Nous l'illustrons dans la section 3.2 sur l'exemple présenté à la section 2.1.4. La méthode bootstrap en cas d'échantillonnage de personnes à l'intérieur des ménages est décrite à la section 3.3, et elle est illustrée à la section 3.4. Dans la section 3.5,

nous expliquons comment l'étape élémentaire de la méthode bootstrap proposée sert à effectuer l'estimation de la variance et produire des intervalles de confiance.

3.1 Étape élémentaire du bootstrap pour les ménages

Au moyen du bootstrap avec remise, nous tirons d'abord à l'intérieur de l'échantillon initial S_{hh}^{h} sélectionné dans la strate U_{hh}^{h} un rééchantillon avec remise S_{hh*}^{h} de $n_h - 1$ ménages, avec probabilités égales. Notons que le rééchantillonnage est effectué sur l'unité d'échantillonnage (un ménage) plutôt que sur l'unité finale d'observation (une personne), ce qui est essentiel pour saisir correctement la variance due à l'échantillonnage. En particulier, cette méthode bootstrap permet de saisir la variance due à l'échantillonnage du second degré (sélection des personnes) sans rééchantillonner les unités finales dans le processus de bootstrap. Pour tout $k \in S_{hh}^{h}$, nous définissons le facteur d'ajustement de repondération

$$G_k = \frac{n_h}{n_h - 1} \times m_k, \tag{3.1}$$

avec m_k le nombre de fois que le ménage k est sélectionné dans le rééchantillon S_{hh*}^{h}, également appelé la multiplicité. Il faut noter qu'une unité $k \in S_{hh}^{h}$ peut ne pas apparaître dans le rééchantillon, auquel cas cette unité a une multiplicité nulle; un exemple est donné à la section 3.2. Les facteurs d'ajustement de la repondération G_k sont utilisés pour obtenir les poids bootstrap qui tiennent compte du plan d'échantillonnage, de la non-réponse totale et du calage, comme le décrit l'algorithme 1. Les étapes font référence à la figure 2.1. Le rééchantillonnage présenté dans l'algorithme 1 est ensuite répété B fois indépendamment pour l'estimation de la variance ou pour produire un intervalle de confiance, voir l'algorithme 3 à la section 3.5.

Algorithme 1. Calcul des poids bootstrap des ménages tenant compte de la non-réponse et du calage

- Étape 1 : nous tenons compte de l'échantillonnage des ménages en calculant, pour tout $k \in S_{hh}$, le poids d'échantillonnage bootstrap

$$d_{k*} = G_k d_k. \tag{3.2}$$

La version bootstrap de l'estimateur en présence de réponse complète donnée dans (2.3) est

$$\hat{Y}_{hh*} = \sum_{k \in S_{hh}} d_{k*} y_k. \tag{3.3}$$

- Étape 2 : nous tenons compte de la non-réponse totale des ménages en calculant les probabilités estimées bootstrap à l'intérieur des GRH.

$$\hat{p}_{c*} = \frac{\sum_{k \in S_{c,hh}} G_k \theta_k r_k}{\sum_{k \in S_{c,hh}} G_k \theta_k}, \tag{3.4}$$

et nous calculons les poids bootstrap corrigés pour tenir compte de la non-réponse

$$d_{rk*} = \frac{d_{k*}}{\hat{p}_{c(k)*}}, \tag{3.5}$$

avec $c(k)$ le GRH contenant le ménage k. La version bootstrap de l'estimateur corrigé pour tenir compte de la non-réponse totale donnée dans (2.7) est

$$\hat{Y}_{r,hh*} = \sum_{k \in S_{r,hh}} d_{rk*} y_k. \tag{3.6}$$

- Étape 3 : nous tenons compte du calage en calant les poids d_{rk*} sur les totaux X_{hh}. Cela donne les poids bootstrap calés

$$w_{k*} = d_{rk*}\left(1 + x_k^{\mathrm{T}} \lambda_{hh*}\right), \tag{3.7}$$

avec

$$\lambda_{hh*} = \left(\sum_{k \in S_{r,hh}} d_{rk*} x_k x_k^{\mathrm{T}}\right)^{-1} \left(X_{hh} - \hat{X}_{r,hh*}\right)$$

et

$$\hat{X}_{r,hh*} = \sum_{k \in S_{r,hh}} d_{rk*} x_k.$$

La version bootstrap de l'estimateur calé donnée dans (2.10) est

$$\hat{Y}_{\mathrm{cal},hh*} = \sum_{k \in S_{r,hh}} w_{k*} y_k. \tag{3.8}$$

Le traitement de la non-réponse totale dans le processus bootstrap mérite quelques explications. Premièrement, notre approche est conditionnelle aux indicateurs de réponse r_k. Contrairement aux indicateurs d'appartenance de l'échantillon qui sont traités par bootstrap à l'étape 1 de l'algorithme 1, les indicateurs de réponse demeurent fixes dans le processus bootstrap. Cela est dû au fait que nous cherchons à reproduire un estimateur de la variance qui considère l'échantillon S_{hh} comme étant sélectionné avec remise, et que dans ce cas, il n'est pas nécessaire d'appliquer la technique bootstrap aux r_k. Deuxièmement, la prise en compte de la non-réponse à l'étape 2 de l'algorithme 1 est réalisée conditionnellement sur les GRH : nous n'appliquons pas de bootstrap au processus menant à la construction des GRH (sur le sujet, voir par exemple Girard, 2009; Haziza et Beaumont, 2017). Enfin, le bootstrap des probabilités de réponse tel qu'il est décrit dans l'équation (3.4) tient compte de l'estimation des probabilités de réponse p_c. En d'autres termes, nous utilisons dans chaque rééchantillonnage les

mêmes GRH que ceux déterminés à partir de l'échantillon, mais les ajustements pour tenir compte de la non-réponse dans les GRH sont basés sur le contenu du rééchantillonnage. Cela est illustré dans l'exemple présenté à la section 3.2. Si nous n'appliquons pas de bootstrap aux probabilités de réponse et que nous insérons directement dans l'équation (3.5) les probabilités estimées à l'origine $\hat{p}_c$, alors les probabilités de réponse sont traitées comme si elles étaient connues, ce qui entraîne habituellement une surestimation de la variance (Beaumont, 2005; Kim et Kim, 2007).

Discutons maintenant de l'estimation de la variance bootstrap pour les estimateurs calés, comme cela est fait à l'étape 3 de l'algorithme 1 où l'étape de calage est réalisée sur le total réel de la population X_{hh}. Si l'on suit le principe bootstrap selon lequel l'échantillon S_{hh} est à l'échantillon bootstrap S_{hh*} ce que la population U_{hh} est à l'échantillon S_{hh}, il pourrait sembler plus intuitif de caler plutôt les totaux estimés $\hat{X}_{hh}$ obtenus en insérant x_k dans l'équation (2.3). Les deux démarches semblent valides pour ce qui est de l'estimation de la variance bootstrap pour l'estimateur calé $\hat{Y}_{\text{cal},hh}$, mais les variables de calage x_k peuvent être sujettes à la non-réponse sur l'échantillon S_{hh}, ce qui rend l'estimateur $\hat{X}_{hh}$ impossible à calculer, alors que le total X_{hh} est connu à partir d'une source extérieure.

3.2 Exemple de calcul des poids bootstrap des ménages

Nous poursuivons avec l'exemple présenté à la section 2.1.4. On réalise le bootstrap en sélectionnant d'abord un rééchantillon de $n_{hh} - 1 = 9$ ménages, avec remise et probabilités égales, parmi les ménages initialement échantillonnés. Dans cet exemple, nous supposons que le ménage A est sélectionné trois fois, que le ménage G est sélectionné deux fois et que les ménages D, E, H et I sont sélectionnés une fois. Au moyen de l'équation (3.2), on obtient les poids d'échantillonnage bootstrap

$$d_{A*} = \frac{40}{3} \quad d_{D*} = \frac{40}{9} \quad d_{E*} = d_{H*} = d_{I*} = \frac{160}{9} \quad d_{G*} = \frac{320}{9}. \tag{3.9}$$

Les poids d'échantillonnage bootstrap sont corrigés pour tenir compte de la non-réponse de la même façon que dans la correction initiale de la non-réponse, au moyen des mêmes GRH et des probabilités estimées pondérées. Dans ce cas, le premier GRH contient uniquement l'unité A qui est un répondant, de sorte que $\hat{p}_{1*} = 1$. Le second GRH contient D, E, G (non-répondant), H et I. Cela donne

$$\hat{p}_{2*} = \frac{d_{D*} + d_{E*} + d_{H*} + d_{I*}}{d_{D*} + d_{E*} + d_{G*} + d_{H*} + d_{I*}} = \frac{13}{21}, \tag{3.10}$$

et les poids bootstrap corrigés pour tenir compte de la non-réponse

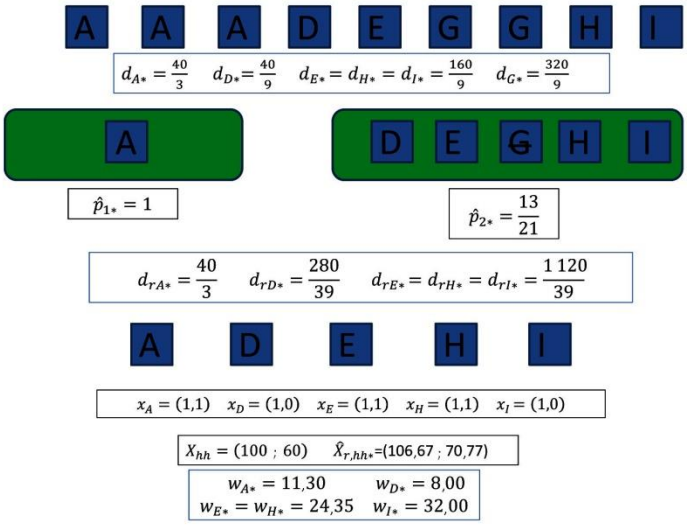
$$d_{rA*} = \frac{40}{3} \quad d_{rD*} = \frac{280}{39} \quad d_{rE*} = d_{rH*} = d_{rI*} = \frac{1\,120}{39}. \tag{3.11}$$

Enfin, les poids sont calés pour qu'ils soient appariés à la taille de la population $N_{hh} = 100$ et au total auxiliaire $X_{1,hh} = 60$. Cela donne les poids bootstrap calés

$$w_{A*} = 11{,}30 \quad w_{D*} = 8{,}00 \quad w_{E*} = w_{H*} = 24{,}35 \quad w_{I*} = 32{,}00. \tag{3.12}$$

Le calcul des poids bootstrap est résumé à la figure 3.1.

Figure 3.1 Calcul des poids bootstrap des ménages.



3.3 Calcul des poids bootstrap pour les personnes

Le calcul des poids bootstrap tenant compte du plan d'échantillonnage, de la non-réponse des ménages et des personnes et du calage est décrit dans l'algorithme 2. Les étapes font référence à la figure 2.3. En plus des étapes de bootstrap de l'algorithme 1, notons que l'algorithme 2 implique le calcul bootstrap des probabilités de réponse individuelles uniquement. Ajoutons que le sous-échantillonnage des personnes à l'intérieur des ménages n'a pas besoin d'être traité par bootstrap, comme nous l'indiquons à la section 3.1.

Algorithme 2. Calcul des poids individuels bootstrap tenant compte de la non-réponse des ménages, de la non-réponse des personnes et du calage

- Exécuter les étapes 1 et 2 de l'algorithme 1. Les poids bootstrap des ménages corrigés pour tenir compte de la non-réponse sont d_{rk}^{*}, selon l'équation (3.5).
- Étape 3b : nous tenons d'abord compte de l'échantillonnage des personnes en calculant les poids bootstrap individuels corrigés pour tenir compte de la non-réponse totale du ménage.

$$d_{rl*} = d_{rk(l)*} d_{l|k(l)} \quad \text{avec } k(l) \text{ le ménage contenant } l. \tag{3.13}$$

Nous tenons ensuite compte de la non-réponse totale des personnes. Nous calculons les probabilités estimées bootstrap à l'intérieur des GRH.

$$\hat{p}_{d*} = \frac{\sum_{l \in S_{rd,\text{ind}}} G_{k(l)} \theta_l r_l}{\sum_{l \in S_{rd,\text{ind}}} G_{k(l)} \theta_l}. \tag{3.14}$$

Nous calculons les poids bootstrap des personnes avec correction pour tenir compte de la non-réponse du ménage ou d'une personne, à savoir :

$$d_{rrl*} = \frac{d_{rl*}}{\hat{p}_{d(l)*}}, \tag{3.15}$$

avec $d(l)$ le GRH contenant la personne l. La version bootstrap de l'estimateur corrigé pour tenir compte de la non-réponse totale donnée dans (2.27) est

$$\hat{Y}_{rr,\text{ind}*} = \sum_{l \in S_{rr,\text{ind}}} d_{rrl*} y_l. \tag{3.16}$$

- Étape 4b : nous tenons compte du calage en calant les poids d_{rrl*} sur les totaux Z_{ind}. Cela donne les poids bootstrap calés

$$w_{l*} = d_{rrl*}\left(1 + z_l^{\mathrm{T}} \lambda_{\text{ind}*}\right), \tag{3.17}$$

avec

$$\lambda_{\text{ind}*} = \left(\sum_{k \in S_{rr,\text{ind}}} d_{rrl*} z_l z_l^{\mathrm{T}}\right)^{-1} \left(Z_{\text{ind}} - \hat{Z}_{rr,\text{ind}*}\right)$$

et

$$\hat{Z}_{rr,\text{ind}*} = \sum_{l \in S_{rr,\text{ind}}} d_{rrl*} z_l.$$

La version bootstrap de l'estimateur calé donnée dans (2.29) est

$$\hat{Y}_{\text{cal,ind}*} = \sum_{l \in S_{rr,\text{ind}}} w_{l*} y_l. \tag{3.18}$$

3.4 Exemple de calcul des poids bootstrap des personnes

Nous poursuivons avec l'exemple présenté à la section 3.2. L'échantillon bootstrap de ménages est constitué de A (trois fois), G (deux fois), et D, E, H et I (une fois). En raison de la non-réponse des ménages, nous observons A, D, E, H et I seulement. À partir de (2.30), on obtient l'échantillon bootstrap de personnes

$$S_{r,\text{ind}*} = \{i_1, i_4, i_6, i_{11}, i_{12}\}. \tag{3.19}$$

Les poids bootstrap des ménages corrigés pour tenir compte de la non-réponse totale sont donnés dans l'équation (3.11). À partir de l'équation (3.13), les poids bootstrap des personnes ajustés pour tenir compte de la non-réponse des ménages sont

$$d_{r1*} = 40 \quad d_{r4*} = \frac{280}{39} \quad d_{r6*} = \frac{2\ 240}{39} \quad d_{r11*} = \frac{2\ 240}{39} \quad d_{r12*} = \frac{1\ 120}{39}. \tag{3.20}$$

Ces poids bootstrap sont corrigés pour tenir compte de la non-réponse des personnes de la même façon que dans la correction initiale de la non-réponse individuelle, au moyen des mêmes GRH et des probabilités estimées non pondérées. Cependant, nous devons tenir compte dans ces probabilités de la multiplicité m_k et du facteur d'ajustement de la repondération G_k, voir l'équation (3.1). Dans notre cas, le premier GRH contient les personnes i_1, i_6 et i_{11}, et i_{11} est un non-répondant. La personne i_1 appartient au ménage A, qui a été sélectionné trois fois $(m_A = 3)$ dans l'échantillon bootstrap. La personne i_6 appartient au ménage E, et la personne i_{11} appartient au ménage H, qui ont tous deux été sélectionnés une fois dans l'échantillon bootstrap $(m_E = m_H = 1)$. Le calcul est semblable pour le second GRH et donne

$$\hat{p}_{1*} = \frac{G_A + G_E}{G_A + G_E + G_H} = \frac{4}{5},$$
$$\hat{p}_{2*} = \frac{G_I}{G_D + G_I} = \frac{1}{2}, \tag{3.21}$$

et les poids bootstrap des personnes corrigés pour tenir compte de la non-réponse du ménage ou de la personne sont donnés par

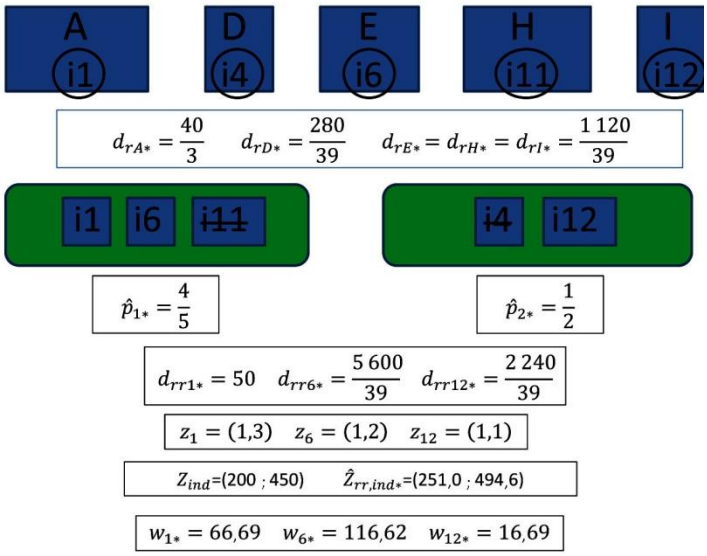
$$d_{rr1*} = 50 \quad d_{r6*} = \frac{5\ 600}{39} \quad d_{r12*} = \frac{2\ 240}{39}. \tag{3.22}$$

Enfin, les poids sont calés pour qu'ils permettent de reproduire exactement la taille de la population $N_{\text{ind}} = 200$ et le total auxiliaire $Z_{1,\text{ind}} = 450$. Cela donne les poids bootstrap calés

$$w_{1*} = 66{,}69 \quad w_{6*} = 116{,}62 \quad w_{12*} = 16{,}69. \tag{3.23}$$

Le calcul des poids bootstrap des personnes est résumé à la figure 3.2.

Figure 3.2 Calcul des poids bootstrap des personnes.



3.5 Estimation de la variance bootstrap et intervalles de confiance

Dans la présente section, nous nous intéressons aux paramètres qui peuvent être écrits comme des fonctions lisses de totaux. Nous expliquons comment l'étape élémentaire de la méthode bootstrap proposée sert à effectuer l'estimation de la variance et produire des intervalles de confiance. Par souci de concision, nous nous concentrons sur les paramètres définis sur la population de ménages U_{hh}. Le traitement des paramètres d'intérêt dans la population de personnes U_{ind} est semblable.

Supposons que y_k est un vecteur de taille q de variables d'intérêt, et que nous nous intéressons à un paramètre $\beta_{hh} = f(Y_{hh})$ utilisant une fonction connue et lisse $f : \mathbb{R}^q \to \mathbb{R}$. En cas de réponse complète, l'estimateur par substitution de β_{hh} est

$$\hat{\beta}_{hh} = f(\hat{Y}_{hh}), \tag{3.24}$$

voir, par exemple, Deville (1999). En cas de non-réponse totale au niveau du ménage, l'estimateur de β_{hh} corrigé pour tenir compte de la non-réponse totale est

$$\hat{\beta}_{r,hh} = f(\hat{Y}_{r,hh}), \tag{3.25}$$

et l'estimateur calé de β_{hh} est

$$\hat{\beta}_{\text{cal},hh} = f(\hat{Y}_{\text{cal},hh}). \tag{3.26}$$

Dans chaque cas, on obtient un estimateur de la variance bootstrap en appliquant un grand nombre de fois (disons B) l'étape de base de la méthode bootstrap dans l'algorithme 1, puis en calculant la dispersion des estimateurs bootstrap. Cela est résumé dans l'algorithme 3.

Algorithme 3. Estimation de la variance bootstrap pour l'estimation de la population des ménages

1. Répéter B fois la procédure bootstrap décrite dans l'algorithme 1. Soit $\hat{Y}^b_{hh*}$, $\hat{Y}^b_{r,hh*}$ et $\hat{Y}^b_{\text{cal},hh*}$ les estimateurs bootstrap des totaux calculés sur le b-ème échantillon. De plus, désignons par $\hat{\beta}^b_{hh*}$, $\hat{\beta}^b_{r,hh*}$ et $\hat{\beta}^b_{\text{cal},hh*}$ les estimateurs bootstrap associés à β_{hh}.
2. L'estimateur de la variance bootstrap pour $\hat{\beta}_{hh}$ est

$$\hat{V}_{\text{boot}}(\hat{\beta}_{hh}) = \frac{1}{B-1}\sum_{b=1}^{B}\left\{\hat{\beta}^b_{hh*} - \frac{1}{B}\sum_{b'=1}^{B}\hat{\beta}^{b'}_{hh*}\right\}^2, \tag{3.27}$$

et de même pour $\hat{\beta}_{r,hh}$ et $\hat{\beta}_{\text{cal},hh}$.

L'estimateur de la variance bootstrap peut servir à calculer un intervalle de confiance reposant sur la normalité avec un niveau ciblé $1-2\alpha$. Par exemple, l'intervalle de confiance quand on utilise l'estimateur en présence de réponse complète $\hat{\beta}_{hh}$ est

$$\mathrm{IC}_{\mathrm{nor}}(\beta_{hh}) = \left[\hat{\beta}_{hh} \pm u_{1-\alpha}\left\{\hat{V}_{\mathrm{boot}}(\hat{\beta}_{hh})\right\}^{0,5}\right], \tag{3.28}$$

avec $u_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la distribution normale standard. On s'attend à ce que cet intervalle de confiance soit conservatif, puisque la méthode bootstrap proposée l'est.

Nous examinons aussi les intervalles de confiance bootstrap (aussi dits élémentaires) dites du percentile et du percentile inverse. Ils peuvent être calculés directement à partir des poids bootstrap et sont par conséquent intéressants du point de vue des utilisateurs des données, contrairement aux méthodes nécessitant une grande puissance de calcul comme le bootstrap t (par exemple Davison et Hinkley, 1997; Shao et Tu, 1995). Pour $\hat{\beta}_{hh}$, l'intervalle de confiance percentile est obtenu au moyen de la distribution de $\hat{\beta}_{hh*}$ comme approximation de la distribution de $\hat{\beta}_{hh}$. Cette méthode utilise les estimations bootstrap ordonnées $\hat{\beta}_{hh*}^{(1)}, \ldots, \hat{\beta}_{hh*}^{(B)}$ pour former l'intervalle de confiance

$$\mathrm{IC}_{\mathrm{per}}(\beta_{hh}) = \left[\hat{\beta}_{hh*}^{(L)},\ \hat{\beta}_{hh*}^{(U)}\right], \tag{3.29}$$

avec le niveau ciblé $1-2\alpha$, où $L = \alpha B$ et $U = (1-\alpha)B$. On obtient l'intervalle de confiance percentile inverse en considérant la distribution de $(\hat{\beta}_{hh*} - \hat{\beta}_{hh})$ comme une approximation de la distribution de $(\hat{\beta}_{hh} - \beta_{hh})$. Cela donne l'intervalle de confiance

$$\mathrm{IC}_{\mathrm{rev}}(\beta_{hh}) = \left[2\hat{\beta}_{hh} - \hat{\beta}_{hh*}^{(U)},\ 2\hat{\beta}_{hh} - \hat{\beta}_{hh*}^{(L)}\right]. \tag{3.30}$$

Les propriétés de l'estimateur de la variance bootstrap et des trois intervalles de confiance sont évaluées dans l'étude par simulations effectuée à la section 4 pour l'estimation d'un total.

Le choix du nombre B de rééchantillonnages constitue un problème pratique important. Girard (2009) propose d'envisager plusieurs tailles de rééchantillonnage possibles (par exemple en augmentant B par un incrément de 100) et de représenter graphiquement les estimateurs de la variance bootstrap en fonction de B. La valeur pour laquelle cet estimateur de la variance commence à se stabiliser est alors retenue. Il s'agit d'une méthode simple, mais qui peut nécessiter une solution de compromis si différentes variables d'intérêt donnent différentes valeurs de stabilisation. Beaumont et Patak (2012) proposent de choisir B de telle sorte qu'avec une forte probabilité, la longueur de l'intervalle de confiance bootstrap donnée dans (3.28) soit proche de la longueur de l'intervalle de confiance obtenu au moyen d'un estimateur de la variance analytique. En supposant que, conditionnellement à l'échantillon initial, l'estimateur bootstrap normalisé du total est normalement distribué, ils établissent que la valeur B peut être déterminée à partir de la distribution d'une variable du chi deux (Beaumont et Patak, 2012, équation 10). Il est intéressant d'observer que la valeur obtenue ne dépend pas de la variable d'intérêt. À partir de ces résultats, ils proposent d'utiliser une valeur B qui ne soit pas inférieure à 750 et une valeur plus grande si l'hypothèse de normalité de l'estimateur bootstrap peut ne pas se vérifier. Nous avons utilisé B = 1 000 dans l'étude par simulations présentée dans la section suivante. Pour les enquêtes devant répondre à plusieurs besoins analytiques – allant de paramètres de population simples à complexes et à diverses tailles de domaine – la

sélection d'au moins 1 000 répliques est la norme compte tenu des ressources informatiques disponibles à l'heure actuelle.

4. Étude par simulations

Afin d'évaluer la méthode bootstrap proposée, nous avons mené une étude par simulations sur une population artificielle. Nous générons d'abord une population U_{hh} contenant N_{hh} = 100 000 ménages, avec quatre variables auxiliaires $x_1, \ldots, x_4$ générées à partir d'une distribution gamma avec les paramètres de forme et d'échelle 2 et 5. À l'intérieur de la population, nous générons trois variables d'intérêt $y_1, \ldots, y_3$ conformément aux modèles suivants :

$$\begin{aligned} y_{1k} &= 10 + x_{1k} + x_{2k} + \sigma_\varepsilon \varepsilon_k, \\ y_{2k} &= 10 + x_{1k} + x_{3k} + \sigma_\varepsilon \varepsilon_k, \\ y_{3k} &= 10 + x_{3k} + x_{4k} + \sigma_\varepsilon \varepsilon_k, \end{aligned} \tag{4.1}$$

où ε_k est généré selon une distribution normale standard. Nous utilisons $\sigma_\varepsilon = 10$, ce qui donne un coefficient de détermination d'environ 0,50 pour chaque modèle. Les variables auxiliaires $1, x_{1k}, x_{2k}$ sont utilisées comme variables de calage au niveau du ménage dans cette étude par simulations. Les trois variables d'intérêt correspondent donc à des cas où le modèle de calage est correctement spécifié (y_1), en partie correctement spécifié (y_2), ou incorrectement spécifié (y_3). La population U_{hh} est répartie aléatoirement dans cinq groupes de réponse homogènes (GRH) de tailles égales. La probabilité de réponse p_c dans le GRH c est égale à 0,5 pour le premier groupe, à 0,6 pour le deuxième groupe, ..., et à 0,9 pour le cinquième groupe, ce qui donne un taux de réponse moyen de 70 % pour les ménages.

À l'intérieur de chaque ménage k, nous générons N_k individus, où $N_k - 1$ est généré selon une distribution de Poisson avec le paramètre 1, ce qui donne un nombre moyen de 2 individus par ménage. Dans la population correspondante U_{ind}, nous générons quatre variables auxiliaires $z_1, \ldots, z_4$ avec les paramètres de forme et d'échelle 2 et 0,5. De plus, nous générons trois variables d'intérêt y_4, y_5, y_6 selon les modèles suivants

$$\begin{aligned} y_{4l} &= 5 + 0{,}5z_{1l} + 0{,}5z_{2l} + \sigma_\eta \eta_l, \\ y_{5l} &= 5 + 0{,}5z_{1l} + 0{,}5z_{3l} + \sigma_\eta \eta_l, \\ y_{6l} &= 5 + 0{,}5z_{3l} + 0{,}5z_{4l} + \sigma_\eta \eta_l, \end{aligned} \tag{4.2}$$

où η_l est généré selon une distribution normale standard. Nous utilisons σ_η = 0,4, ce qui donne un coefficient de détermination d'environ 0,6 pour chaque modèle. Les variables auxiliaires $1, z_{1l}, z_{2l}$ sont utilisées comme variables de calage au niveau de la personne dans cette étude par simulations. Les trois variables d'intérêt correspondent donc à un cas où le modèle de calage est correctement spécifié (y_4), en partie correctement spécifié (y_5), ou incorrectement spécifié (y_6).

La population U_{ind} est divisée en cinq GRH comme suit. Les personnes qui sont seules dans leur ménage forment un GRH distinct, avec une probabilité de réponse de 1. La justification de ce choix est

que, dans ce cas, la personne équivaut en quelque sorte à son ménage et que la non-réponse est modélisée au niveau du ménage. Parmi le reste des personnes vivant dans un ménage k comprenant $N_k = 2$ personnes ou plus, les variables z_1 et z_2 servent à former quatre GRH de taille approximativement égale. La probabilité de réponse p_d varie de 0,80 à 0,95 dans les quatre GRH restants. Cela donne un taux de réponse global d'environ 90 % pour les personnes.

Dans la population U_{hh}, nous sélectionnons un échantillon S_{hh} de n_{hh} = 1 000 ménages par échantillonnage aléatoire simple sans remise. Notons que comme le taux d'échantillonnage est faible (1 %), l'échantillonnage aléatoire simple n'est pas très différent qu'il soit avec ou sans remise, et le biais des estimateurs de la variance bootstrap devrait être faible dans cette configuration. La non-réponse est générée selon le modèle des GRH pour les ménages, ce qui donne un échantillon $S_{r,hh}$ de ménages répondants. Les probabilités de réponse estimées $\hat{p}_c$ sont obtenues à partir de l'équation (2.5), avec un poids égal $\theta_k = 1$. Dans chaque $k \in S_{r,hh}$, un individu Kish est sélectionné aléatoirement avec des probabilités égales, ce qui donne un échantillon de personnes $S_{r,\text{ind}}$. Dans $S_{r,\text{ind}}$, la non-réponse est générée selon le modèle des GRH pour les individus, ce qui donne un échantillon $S_{rr,\text{ind}}$ de personnes répondantes. Les probabilités de réponse estimées $\hat{p}_d$ sont obtenues à partir de l'équation (2.25), de trois manières possibles : soit avec des poids égaux $\theta_l = 1$, soit avec les poids d'échantillonnage $\theta_l = d_l$, soit avec les poids des personnes corrigés pour tenir compte de la non-réponse du ménage $\theta_l = d_{rl}$.

Les étapes d'échantillonnage et de non-réponse sont répétées R = 1 000 fois. Sur chaque échantillon S_{hh}, nous calculons l'estimateur en présence de réponse complète donné dans (2.3), et sur chaque échantillon $S_{r,hh}$, nous calculons l'estimateur ajusté pour tenir compte de la non-réponse $\hat{Y}_{r,hh}$ donné dans (2.7) et l'estimateur $\hat{Y}_{\text{cal},hh}$ donné dans (2.10) avec l'ensemble des variables de calage $x_k = (1, x_{1k}, x_{2k})^{\top}$. Sur chaque échantillon $S_{rr,\text{ind}}$, nous calculons l'estimateur ajusté pour tenir compte de la non-réponse $\hat{Y}_{rr,\text{ind}}$ donné dans (2.27) et l'estimateur $\hat{Y}_{\text{cal,ind}}$ donné dans (2.29) avec l'ensemble des variables de calage $z_l = (1, z_{1l}, z_{2l})^{\top}$. Pour ces cinq estimateurs, nous calculons la racine de l'erreur quadratique moyenne normalisée

$$\text{REQMN}(\hat{Y}) = 100 \times \frac{\sqrt{\text{EQM}(\hat{Y})}}{Y}, \tag{4.3}$$

avec $\text{EQM}(\hat{Y})$ une approximation basée sur la simulation de l'erreur quadratique moyenne de $\hat{Y}$, obtenue à partir de l'exécution indépendante de 10 000 simulations.

Pour ces cinq estimateurs, nous calculons aussi les estimateurs de la variance bootstrap obtenus par l'application de l'algorithme 3 avec B = 1 000. Pour mesurer le biais d'un estimateur de la variance $v(\hat{Y})$, nous utilisons le biais relatif Monte Carlo en pourcentage

$$\text{BR}\left\{v(\hat{Y})\right\} = 100 \times \frac{R^{-1}\sum_{c=1}^{R} v_c(\hat{Y}_c) - \text{EQM}(\hat{Y})}{\text{EQM}(\hat{Y})}, \tag{4.4}$$

où $v_c(\hat{Y}_c)$ représente l'estimateur de la variance dans le c-ème échantillon. Comme mesure de stabilité de $v(\hat{Y})$, nous utilisons la stabilité relative

$$\mathrm{SR}\left\{v(\hat{Y})\right\} = 100 \times \frac{\left[R^{-1}\sum_{c=1}^{R}\left\{v_c(\hat{Y}_c) - \mathrm{EQM}(\hat{Y})\right\}^2\right]^{1/2}}{\mathrm{EQM}(\hat{Y})}. \tag{4.5}$$

De plus, nous calculons les taux de couverture de l'intervalle de confiance associé au bootstrap par la méthode du percentile, la méthode du percentile inverse et à l'intervalle de confiance reposant sur la normalité, avec un taux d'erreur nominal unitaléral de 2,5 % dans chaque queue de distribution.

Les résultats sont présentés au tableau 4.1 pour l'estimation de la population des ménages. La racine de l'erreur quadratique moyenne normalisée de l'estimateur calé $\hat{Y}_{\mathrm{cal},hh}$ est plus petite quand les variables de calage sont explicatives pour la variable d'intérêt, comme cela était attendu. Nous observons un biais légèrement positif de l'estimateur de la variance bootstrap pour l'estimateur en présence de réponse complète $\hat{Y}_{hh}$, mais presque aucun biais pour les estimateurs repondérés $\hat{Y}_{r,hh}$ et $\hat{Y}_{\mathrm{cal},hh}$. L'estimateur de la variance bootstrap est un peu moins stable avec les estimateurs repondérés, ce qui est probablement attribuable à la variabilité supplémentaire associée à la correction de la non-réponse totale. En ce qui concerne les intervalles de confiance, nous constatons que les taux de couverture sont correctement respectés dans tous les cas et pour les trois méthodes étudiées.

Nous passons maintenant au résultat sur la population de personnes, qui est présenté au tableau 4.2. Nous observons que le biais relatif de l'estimateur de la variance bootstrap est très petit dans tous les cas. Le choix des poids θ_k utilisés dans l'estimation des probabilités de réponse ne semble pas avoir d'effet sur la racine de l'erreur quadratique moyenne normalisée des estimateurs, mais l'utilisation des poids $\theta_l = d_{rl}$ ajustés pour tenir compte de la non-réponse des ménages donne des estimateurs de la variance légèrement plus stables pour $\hat{Y}_{rr,\mathrm{ind}}$. Les taux de couverture sont à peu près respectés dans tous les cas.

Tableau 4.1
Coefficient de variation de l'estimateur du total, Biais relatif et Stabilité relative de l'estimateur de la variance bootstrap et Taux d'erreur nominaux unilatéraux pour l'estimation de la variance bootstrap par la méthode des centiles et du bootstrap de base pour 3 variables de la population des ménages

					Bootstrap par la méthode des centiles			Bootstrap de base			Reposant sur la normalité		
		REQMN	**BR**	**SR**	**L**	**U**	**L+U**	**L**	**U**	**L+U**	**L**	**U**	**L+U**
$\hat{Y}_{hh}$	y_1	1,47	2,48	7,2	2,2	3,1	5,3	2,1	3,3	5,4	2,2	3,2	5,4
	y_2	1,48	0,73	6,6	2,6	3,3	5,9	2,7	3,4	6,1	2,6	3,2	5,8
	y_3	1,48	1,11	6,6	2,6	2,7	5,3	2,7	3,0	5,7	2,4	2,7	5,1
$\hat{Y}_{r,hh}$	y_1	1,82	0,42	8,7	2,4	2,4	4,8	2,3	2,7	5,0	2,3	2,6	4,9
	y_2	1,83	-0,76	8,2	2,7	2,8	5,5	2,5	3,0	5,5	2,2	2,7	4,9
	y_3	1,82	0,72	8,4	2,8	2,1	4,9	2,8	2,2	5,0	2,8	1,9	4,7
$\hat{Y}_{\mathrm{cal},hh}$	y_1	1,29	1,27	8,3	2,4	2,7	5,1	2,8	2,8	5,6	2,8	2,7	5,5
	y_2	1,58	-0,55	8,2	2,5	3,5	6,0	2,8	3,9	6,7	2,8	3,6	6,4
	y_3	1,82	0,49	8,4	2,9	1,8	4,7	3,0	2,2	5,2	2,9	2,0	4,9

Tableau 4.2
Coefficient de variation de l'estimateur du total, Biais relatif et Stabilité relative de l'estimateur de la variance bootstrap et Taux d'erreur nominaux unilatéraux pour l'estimation de la variance bootstrap par la méthode des centiles et du bootstrap de base pour 3 variables de la population des personnes

					Bootstrap par la méthode des centiles			Bootstrap de base			Reposant sur la normalité		
		REQMN	BR	SR	L	U	L+U	L	U	L+U	L	U	L+U
		Poids égaux $\theta_l = 1$											
$\hat{Y}_{rr,\text{ind}}$	y_4	2,01	0,31	9,6	2,0	3,2	5,2	1,9	3,3	5,2	1,9	3,0	4,9
	y_5	2,02	-0,17	9,6	2,4	3,4	5,8	2,2	3,7	5,9	2,3	3,5	5,8
	y_6	2,02	-0,24	9,6	2,2	3,3	5,5	2,0	3,7	5,7	2,0	3,2	5,2
$\hat{Y}_{\text{cal,ind}}$	y_4	0,29	1,72	10,8	2,1	2,4	4,5	2,1	2,3	4,4	2,1	2,2	4,3
	y_5	0,39	1,04	11,3	2,3	2,5	4,8	2,3	2,5	4,8	2,2	2,4	4,6
	y_6	0,47	1,90	11,2	2,8	2,1	4,9	2,2	2,5	4,7	2,3	2,0	4,3
		Poids d'échantillonnage $\theta_l = d_l$											
$\hat{Y}_{rr,\text{ind}}$	y_4	2,00	-0,08	9,5	1,8	3,8	5,6	1,7	3,8	5,5	1,7	3,4	5,1
	y_5	2,00	0,14	9,4	1,9	3,3	5,2	2,2	3,5	5,7	1,8	3,5	5,3
	y_6	1,99	0,61	9,3	1,7	3,2	4,9	1,7	3,4	5,1	1,7	3,2	4,9
$\hat{Y}_{\text{cal,ind}}$	y_4	0,29	-0,57	10,3	2,9	2,4	5,3	3,3	2,2	5,5	3,0	2,3	5,3
	y_5	0,39	0,40	11,6	2,4	3,2	5,6	2,7	3,3	6,0	2,3	3,2	5,5
	y_6	0,47	-0,05	11,2	2,3	2,2	4,5	1,8	2,3	4,1	1,8	2,3	4,1
		Poids corrigés pour tenir compte de la non-réponse des ménages $\theta_l = d_{rl}$											
$\hat{Y}_{rr,\text{ind}}$	y_4	1,99	-0,71	8,9	2,5	2,3	4,8	2,6	2,7	5,3	2,5	2,4	4,9
	y_5	1,99	-0,82	8,9	3,1	2,2	5,3	2,9	2,5	5,4	2,5	2,2	4,7
	y_6	1,99	-0,26	9,1	3,1	2,3	5,4	3,0	3,0	6,0	2,9	2,5	5,4
$\hat{Y}_{\text{cal,ind}}$	y_4	0,29	1,70	10,6	2,7	3,4	6,1	2,6	3,3	5,9	2,5	3,3	5,8
	y_5	0,39	1,38	11,3	2,1	2,7	4,8	2,2	3,0	5,2	1,7	3,0	4,7
	y_6	0,47	0,61	10,9	2,5	2,8	5,3	2,3	3,0	5,3	2,3	2,8	5,1

5. Application sur le Panel Politique de la Ville français

Dans cette section, nous présentons une illustration de la méthodologie proposée sur un panel français concernant la politique de la ville. Le plan de sondage et les étapes d'estimation pour l'échantillon des ménages sont brièvement décrits à la section 5.1, et trois intervalles de confiance bootstrap possibles sont calculés. La macro SAS élaborée en vue de mettre en œuvre la méthodologie proposée pour l'échantillonnage à un degré est donnée à l'annexe B, accompagnée d'un petit exemple. Les étapes supplémentaires d'échantillonnage et d'estimation pour l'échantillon de personnes sont décrites à la section 5.2, et trois intervalles de confiance bootstrap possibles sont calculés. La macro SAS élaborée en vue de mettre en œuvre la méthodologie proposée pour l'échantillonnage à deux degrés est donnée à l'annexe C, accompagnée d'un petit exemple.

5.1 Échantillon de ménages

Le Panel Politique de la Ville (PPV) est une enquête en quatre vagues, menée entre 2011 et 2014 par le secrétariat général du Comité interministériel des villes (SGCIV) de France. L'enquête visait à recueillir des informations sur la sécurité, l'emploi, la précarité, la scolarité et la santé de personnes vivant dans les

zones urbaines sensibles (ZUS). Nous nous intéressons uniquement à la vague de 2011. Un échantillon de ménages est sélectionné, et toutes les personnes vivant dans les ménages sélectionnés sont sondées en théorie.

L'échantillon des ménages est obtenu par échantillonnage à deux degrés, voir par exemple Chauvet (2015); Chauvet et Vallée (2018). Premièrement, la population des quartiers est divisée en quatre strates, et un échantillon global de $n_I = 40$ quartiers est sélectionné au moyen d'un échantillonnage à probabilités proportionnelles à la taille à l'intérieur des strates. Un échantillon de ménages est ensuite sélectionné au deuxième degré dans chaque quartier sélectionné au moyen d'un échantillonnage aléatoire simple, de sorte que les probabilités d'inclusion finale des ménages sont approximativement égales à l'intérieur des strates (plan d'échantillonnage auto-pondéré). Aux fins de l'illustration, la sélection à deux degrés des ménages n'est pas prise en compte ici, et l'échantillon des ménages est considéré comme étant directement sélectionné au moyen d'un échantillonnage aléatoire simple stratifié.

L'échantillon comprend 2 971 ménages, mais en raison de la non-réponse totale, 1 256 ménages seulement sont observés. La non-réponse est prise en compte au moyen de groupes de réponse homogènes, définis en fonction de cinq variables auxiliaires : période de construction du logement, type de logement (appartement/maison), nombre de pièces, habitation à loyer modique (oui/non) et région. En utilisant une régression logistique et la méthode des scores (par exemple Haziza et Beaumont, 2007), nous obtenons huit groupes de réponse homogènes. Les cinq variables auxiliaires ayant servi à la définition des GRH sont également utilisées pour le calage.

Nous nous intéressons à quatre variables catégorielles liées à la sécurité, à l'urbanisme et à la mobilité résidentielle. La variable y_1 donne la réputation perçue du quartier (bonne, passable, mauvaise, sans opinion). La variable y_2 indique si un membre du ménage a été témoin de trafic (jamais, rarement, parfois, sans opinion). La variable y_3 indique si des travaux routiers importants ont été réalisés dans le quartier au cours des 12 derniers mois (oui, non, sans opinion). La variable y_4 indique si le ménage a l'intention de quitter le quartier au cours des 12 prochains mois (certainement/probablement, certainement pas, probablement pas, sans opinion). Pour chaque catégorie g de chaque variable y, nous nous intéressons à la proportion

$$\beta_{g,hh} = \frac{\sum_{k \in U_{hh}} 1(y_k = g)}{N_{hh}}, \tag{5.1}$$

avec N_{hh} le nombre total de ménages. L'estimateur de β_g ajusté pour tenir compte de la non-réponse est

$$\hat{\beta}_{\mathrm{gr},hh} = \frac{\sum_{k \in S_{r,hh}} d_{rk} 1(y_k = g)}{\sum_{k \in S_{r,hh}} d_{rk}}, \tag{5.2}$$

voir l'équation (2.7). L'estimateur calé de β_g est

$$\hat{\beta}_{\mathrm{gcal},hh} = \frac{\sum_{k \in S_{r,hh}} w_k 1(y_k = g)}{\sum_{k \in S_{r,hh}} w_k}, \tag{5.3}$$

voir l'équation (2.10).

Pour chaque proportion, nous donnons l'intervalle de confiance reposant sur la normalité en utilisant l'estimateur de la variance bootstrap, la méthode du percentile ou la méthode du percentile inverse, voir la section 3.5. Nous utilisons le bootstrap avec remise présenté dans l'algorithme 1 avec $B = 1\,000$ rééchantillonnages. Les résultats avec un taux d'erreur nominal unilatéral de 2,5 % sont présentés dans le tableau 5.1. Les trois intervalles de confiance sont très similaires dans tous les cas.

Tableau 5.1
Estimation des proportions marginales avec trois intervalles de confiance pour quatre variables d'intérêt

	Réputation perçue du statut du quartier							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Bonne	Passable	Mauvaise	Sans opinion	Bonne	Passable	Mauvaise	Sans opinion
Estimation	0,217	0,225	0,531	0,027	0,217	0,224	0,532	0,027
IC Norm.	[0,194;0,241]	[0,201;0,249]	[0,503;0,559]	[0,018;0,036]	[0,193;0,240]	[0,200;0,248]	[0,504;0,560]	[0,018;0,036]
IC. Centiles	[0,195;0,241]	[0,201;0,251]	[0,504;0,558]	[0,019;0,036]	[0,193;0,240]	[0,201;0,251]	[0,505;0,560]	[0,019;0,036]
IC de base	[0,193;0,240]	[0,200;0,249]	[0,503;0,557]	[0,018;0,035]	[0,193;0,240]	[0,198;0,248]	[0,504;0,559]	[0,018;0,035]
	Témoin de trafic							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Jamais	Rarement	Parfois	Sans opinion	Jamais	Rarement	Parfois	Sans opinion
Estimation	0,599	0,065	0,155	0,181	0,606	0,065	0,156	0,173
IC Norm.	[0,571;0,627]	[0,050;0,079]	[0,135;0,175]	[0,161;0,201]	[0,581;0,632]	[0,050;0,079]	[0,135;0,176]	[0,159;0,188]
IC Centiles	[0,572;0,628]	[0,050;0,080]	[0,134;0,175]	[0,161;0,201]	[0,582;0,633]	[0,051;0,080]	[0,134;0,175]	[0,160;0,188]
IC de base	[0,570;0,626]	[0,049;0,078]	[0,136;0,176]	[0,161;0,201]	[0,579;0,630]	[0,049;0,078]	[0,136;0,177]	[0,159;0,187]
	Travaux routiers dans le quartier							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Oui	Non	Sans opinion		Oui	Non	Sans opinion	
Estimation	0,471	0,495	0,034		0,470	0,496	0,034	
IC Norm.	[0,444;0,498]	[0,468;0,523]	[0,024;0,044]		[0,443;0,496]	[0,469;0,523]	[0,024;0,045]	
IC Centiles	[0,442;0,496]	[0,469;0,524]	[0,025;0,045]		[0,440;0,495]	[0,470;0,524]	[0,025;0,045]	
IC de base	[0,445;0,500]	[0,466;0,522]	[0,023;0,043]		[0,444;0,499]	[0,468;0,522]	[0,024;0,044]	
	Intention de quitter le quartier							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Cert./Prob.	Prob. non	Cert. non	Sans opinion	Cert./Prob.	Prob. non	Cert. non	Sans opinion
Estimation	0,286	0,130	0,548	0,036	0,287	0,131	0,546	0,036
IC Norm.	[0,260;0,312]	[0,111;0,149]	[0,520;0,576]	[0,025;0,047]	[0,261;0,313]	[0,112;0,150]	[0,518;0,573]	[0,025;0,047]
IC Centiles	[0,260;0,313]	[0,111;0,149]	[0,521;0,576]	[0,026;0,047]	[0,261;0,313]	[0,113;0,151]	[0,520;0,574]	[0,026;0,048]
IC de base	[0,259;0,312]	[0,111;0,149]	[0,520;0,575]	[0,025;0,046]	[0,261;0,313]	[0,111;0,149]	[0,517;0,572]	[0,025;0,047]

5.2 Échantillon de personnes

L'échantillon des ménages répondants comprend 3 098 personnes qui sont théoriquement sondées, mais en raison de la non-réponse totale, nous observons un sous-ensemble de 2 804 répondants individuels seulement. La non-réponse est prise en compte au moyen de groupes de réponse homogènes, définis en fonction de huit variables auxiliaires : trois au niveau de la personne (sexe, âge, nationalité) et cinq au niveau du logement (période de construction du logement, type de logement, nombre de pièces, habitation à loyer modique ou non, région). En utilisant une régression logistique et la méthode des scores, nous obtenons huit groupes de réponse homogènes. Les trois variables auxiliaires de personnes ayant servi à la définition des GRH sont également utilisées pour le calage.

Nous considérons trois variables d'intérêt. La variable y_5 est quantitative et donne le nombre d'enfants. La variable y_6 indique si la personne a un ou plusieurs emplois (un, plusieurs, aucun, aucune réponse). La variable y_7 indique si la personne bénéficie d'une couverture médicale complémentaire complète (oui, non, pas de réponse). Pour la variable y_5, nous calculons l'estimateur du total ajusté pour

tenir compte de la non-réponse et l'estimateur calé donnés dans les équations (2.27) et (2.29), respectivement. Pour les deux autres variables d'intérêt et pour chaque catégorie g, nous nous intéressons à la proportion

$$\beta_{g,\text{ind}} = \frac{\sum_{l \in U_{\text{ind}}} 1(y_k = g)}{N_{\text{ind}}}, \tag{5.4}$$

avec N_{ind} le nombre total d'individus. L'estimateur de $\beta_{g,\text{ind}}$ ajusté pour tenir compte de la non-réponse est

$$\hat{\beta}_{\text{grr,ind}} = \frac{\sum_{l \in S_{rr,\text{ind}}} d_{rrl} 1(y_l = g)}{\sum_{l \in S_{rr,\text{ind}}} d_{rrl}}, \tag{5.5}$$

voir l'équation (2.27). L'estimateur calé de $\beta_{g,\text{ind}}$ est

$$\hat{\beta}_{\text{gcal,ind}} = \frac{\sum_{l \in S_{rr,\text{ind}}} w_l 1(y_l = g)}{\sum_{l \in S_{rr,\text{ind}}} w_l}, \tag{5.6}$$

voir l'équation (2.29).

Pour chaque paramètre, nous donnons l'intervalle de confiance reposant sur la normalité en utilisant l'estimateur de la variance bootstrap, la méthode du percentile et la méthode du percentile inverse. Nous utilisons le bootstrap avec remise présenté dans l'algorithme 2 avec $B = 1\,000$ rééchantillonnages. Les résultats avec un taux d'erreur nominal unilatéral de 2,5 % sont présentés au tableau 5.2. Les trois intervalles de confiance sont très similaires dans tous les cas.

Tableau 5.2
Estimation des proportions marginales avec trois intervalles de confiance pour quatre variables d'intérêt

	Nombre d'enfants							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
Estimation ($\times 10^6$)	4,40				4,39			
IC Norm.	[4,15;4,64]				[4,21;4,58]			
IC Centiles	[4,16;4,65]				[4,21;4,58]			
IC de base	[4,14;4,63]				[4,20;4,57]			
	La personne a-t-elle plusieurs emplois ?							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Un	Plusieurs	Aucun	Aucune réponse	Un	Plusieurs	Aucun	Aucune réponse
Estimation	0,304	0,016	0,372	0,308	0,305	0,016	0,372	0,307
IC Norm.	[0,286;0,323]	[0,011;0,021]	[0,352;0,392]	[0,290;0,326]	[0,285;0,325]	[0,011;0,021]	[0,350;0,394]	[0,283;0,332]
IC Centiles	[0,287;0,323]	[0,011;0,021]	[0,351;0,393]	[0,289;0,326]	[0,284;0,325]	[0,011;0,020]	[0,351;0,393]	[0,284;0,333]
IC de base	[0,286;0,322]	[0,011;0,020]	[0,351;0,393]	[0,289; 0,326]	[0,285;0,325]	[0,011;0,020]	[0,352;0,393]	[0,282;0,330]
	Couverture médicale complémentaire complète							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Oui	Non	Aucune réponse		Oui	Non	Aucune réponse	
Estimation	0,122	0,626	0,252		0,122	0,627	0,251	
IC Norm.	[0,106;0,137]	[0,603;0,650]	[0,234;0,270]		[0,105;0,138]	[0,604;0,650]	[0,227;0,275]	
IC Centiles	[0,105;0,137]	[0,603;0,651]	[0,235;0,269]		[0,105;0,138]	[0,604;0,650]	[0,230;0,276]	
IC de base	[0,106;0,138]	[0,602;0,649]	[0,235;0,269]		[0,105;0,138]	[0,605;0,651]	[0,227;0,273]	

6. Conclusion et travaux futurs

Dans le présent article, nous avons expliqué comment il est possible d'appliquer le bootstrap avec remise aux enquêtes auprès des ménages, afin de tenir compte de toute la variabilité du processus d'échantillonnage, y compris l'échantillonnage et la non-réponse, ainsi qu'à des ajustements a posteriori comme le calage. Les méthodes ont été illustrées au moyen d'un exemple simple permettant de les exposer clairement, évaluées par une étude par simulations et appliquées à un panel français sur la politique de la ville. Afin de faciliter la mise en œuvre de la méthode par les utilisateurs, nous avons élaboré deux macros SAS que l'auteur correspondant peut mettre à disposition sur demande.

Les résultats de l'étude par simulations montrent que les estimateurs de la variance bootstrap et les trois intervalles de confiance bootstrap fonctionnent correctement dans le cas d'une petite fraction de sondage. Si la fraction de sondage est plus grande, l'estimateur de la variance bootstrap est connu pour être conservatif, et l'intervalle de confiance reposant sur la normalité devrait donc l'être également. Toutefois, les propriétés de couverture des deux autres intervalles de confiance dans ce contexte demeurent floues. Il serait intéressant d'approfondir cette question.

Dans l'article, nous nous sommes concentrés sur l'application du bootstrap pour l'estimation de la variance, après que les ajustements statistiques (traitement de la non-réponse totale et calage) ont été effectués par le méthodologiste de l'enquête. Le bootstrap peut également être utilisé a priori comme outil de diagnostic servant à évaluer la pertinence d'éventuels ajustements statistiques. Par exemple, il peut être tentant d'utiliser un grand nombre de groupes de réponse homogènes (GRH) pour corriger la non-réponse totale, afin de réduire le biais de non-réponse. Cependant, cela peut entraîner une variabilité accrue des estimateurs repondérés. Le bootstrap peut servir à évaluer plusieurs ensembles possibles de GRH, par exemple en produisant des histogrammes des ajustements de non-réponse bootstrap ou des estimateurs bootstrap corrigés pour tenir compte de la non-réponse totale, afin de donner une idée de la stabilité de l'estimation avec un ensemble possible de GRH. Cela est utile pour trouver un compromis entre le biais et la variance. Cette méthode mentionnée dans Girard (2009) est une question importante qu'il faudrait approfondir.

Nous avons étudié une situation où l'enquête n'est effectuée qu'une seule fois. Si nous souhaitons effectuer des estimations longitudinales, les unités sont généralement suivies dans le temps. Si nous nous intéressons aussi aux estimations transversales réalisées à plusieurs reprises, des échantillons supplémentaires sont sélectionnés aux vagues postérieures et mélangés avec l'échantillon initial. L'estimation de la variance bootstrap dans le contexte des enquêtes longitudinales est une question très importante qui devrait faire l'objet d'autres travaux.

Remerciements

Les auteurs remercient chaleureusement le rédacteur adjoint et les deux examinateurs pour leurs nombreuses suggestions sur une version précédente de l'article, qui ont permis de considérablement l'améliorer.

Annexe

A. Estimateurs de la variance repère pour l'échantillon de personnes

Nous considérons d'abord l'estimateur $\hat{Y}_{\text{ind}}$ dans l'équation (2.21) que nous utilisons en cas de réponse complète. L'estimateur repère de la variance est

$$v_{\text{mult}}(\hat{Y}_{\text{ind}}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k \hat{y}_k - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} \hat{y}_{k'} \right)^2, \tag{A.1}$$

avec

$$\hat{y}_k = \sum_{l \in S_{\text{ind},k}} d_{l|k} y_l.$$

Nous considérons maintenant l'estimateur $\hat{Y}_{rr,\text{ind}}$ donné dans l'équation (2.27), qui est ajusté pour tenir compte de la non-réponse des ménages et des personnes. L'estimateur repère de la variance est

$$v_{\text{mult}}(\hat{Y}_{\text{ind}}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k v_{1k} - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} v_{1k'} \right)^2, \tag{A.2}$$

où

$$v_{1k} = \hat{u}_{1k} + u_{3k},$$

où la première variable linéarisée $\hat{u}_{1k}$ est semblable à celle donnée dans l'équation (2.8), tandis que la deuxième variable linéarisée u_{3k} tient compte de l'estimation des probabilités de réponse des personnes. Nous avons pour la première variable linéarisée

$$\hat{u}_{1k} = \theta_k \pi_k \bar{\hat{y}}_{rc(k)} + \frac{r_k}{\hat{p}_{c(k)}} \left\{ \hat{y}_{r,k} - \theta_k \pi_k \bar{\hat{y}}_{rc(k)} \right\},$$

et

$$\bar{\hat{y}}_{rc} = \frac{\sum_{k \in S_{c,hh}} d_k r_k \hat{y}_{r,k}}{\sum_{k \in S_{c,hh}} \theta_k r_k},$$

et

$$\hat{y}_{r,k} = \sum_{l \in S_{\text{ind},k}} \frac{d_{l|k} r_l}{\hat{p}_l} y_l, \tag{A.3}$$

et pour la deuxième variable linéarisée

$$u_{3k} = \frac{r_k}{d_k} \sum_{l \in S_{\text{ind},k}} \theta_l \left(1 - \frac{r_l}{\hat{p}_l} \right) \bar{y}_{rrd(l)}, \tag{A.4}$$

avec

$$\bar{y}_{rrd} = \frac{\sum_{l \in S_{rd,\text{ind}}} d_{rl} r_l y_l}{\sum_{l \in S_{rd,\text{ind}}} \theta_l r_l}. \tag{A.5}$$

Considérons maintenant l'estimateur calé $\hat{Y}_{\text{cal,ind}}$ donné dans l'équation (2.29). L'estimateur repère de la variance est le même que celui donné dans l'équation (A.2) pour $\hat{Y}_{rr,\text{ind}}$, en remplaçant la variable y_l par les résidus de régression estimés de la variable d'intérêt sur les variables de calage, à savoir

$$e_l = y_l - \hat{B}_{rr,\text{ind}}^{\mathrm{T}} z_l \quad \text{avec} \quad \hat{B}_{rr,\text{ind}} = \left(\sum_{l \in S_{rr,\text{ind}}} d_{rrl} z_l z_l^{\mathrm{T}} \right)^{-1} \sum_{l \in S_{rr,\text{ind}}} d_{rrl} z_l y_l. \tag{A.6}$$

B. Programme SAS pour échantillonnage à un degré

Dans cette section, nous présentons la macro SAS élaborée pour mettre en œuvre la méthodologie proposée en cas d'échantillonnage des ménages seulement (échantillonnage à un degré). Le paramétrage du programme SAS pour le calcul des poids bootstrap est présenté à la section B.1. Par souci de clarté, un petit exemple est présenté à la section B.2.

B.1 Programme de calcul des poids bootstrap

Les paramètres liés à la base de données sont :

- `BASE` : bibliothèque contenant la table SAS avec la liste des unités échantillonnées. La valeur par défaut est `BASE=WORK`.
- `ECHMEN` : table SAS contenant la liste des unités échantillonnées dans la population. Les non-répondants doivent aussi être inclus dans cette table.

Les paramètres liés au bootstrap sont :

- `ITBOOT` : nombre d'itérations bootstrap. La valeur par défaut est `ITBOOT=1000`.

Les paramètres liés aux variables nécessaires dans la table SAS sont :

- `IDMEN` : liste des variables identifiant l'unité statistique. Elles doivent être de type caractère.
- `STMEN` : liste des variables de stratification utilisées pour la sélection de l'échantillon.
- `DMEN` : poids d'échantillonnage.
- `RMEN` : indicateur de réponse (1 pour un répondant, 0 pour un non-répondant).
- `DRMEN` : poids d'échantillonnage, corrigé pour tenir compte de la non-réponse. Les valeurs sont nécessaires uniquement pour les répondants.
- `DCMEN` : poids calé. Les valeurs sont nécessaires uniquement pour les répondants.
- `GRHMEN` : liste des variables identifiant les groupes de réponse homogènes.
- `WGRHMEN` : pondération utilisée dans le calcul des probabilités de réponse à l'intérieur des GRH.
  - Avec `WGRHMEN=0`, les taux de réponse ne sont pas pondérés. Il s'agit de la valeur par défaut.
  - Avec `WGRHMEN=1`, les taux de réponse sont pondérés par les poids de sondage.

- `XMENQUANT` : liste des variables quantitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.
- `XMENQUALI` : liste des variables qualitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.

Les paramètres liés à la sortie sont :

- `SORT_MEN` : table SAS contenant les poids d'échantillonnage bootstrap `WB_D1,...,WB_D&ITBOOT` pour l'échantillon entier.
- `SORT_RMEN` : table SAS contenant les poids bootstrap `WB_N1,...,WB_N&ITBOOT` corrigés pour tenir compte de la non-réponse, et les poids bootstrap `WB_C1,...,WB_C&ITBOOT` corrigés pour tenir compte de la non-réponse et du calage, pour le sous-échantillon de répondants.

B.2 Petit exemple

Considérons l'exemple traité à la section 2.1.4. L'échantillon se présente comme suit :

```
data ech;
input idm$ stmen$ dmen rmen GRHmen$ drmen dcmen x0 x1;
cards;
A  1  4   1  aa  4.44   4.01   1  1
B  1  4   0  aa  .      .      .  .
C  1  4   0  bb  .      .      .  .
D  1  4   1  bb  5.54   4.87   1  0
E  1  16  1  bb  22.15  19.98  1  1
F  1  16  1  aa  17.78  15.63  1  0
G  1  16  0  bb  .      .      .  .
H  1  16  1  bb  22.15  19.98  1  1
I  1  16  1  bb  22.15  19.49  1  0
J  1  16  1  aa  17.78  16.03  1  1
;run;
```

Nous pouvons obtenir $B = 1\,000$ poids bootstrap comme suit. Étant donné que `WGRHMEN=1`, on suppose que quand la non-réponse totale a été initialement corrigée par la méthode des GRH, les taux de réponse à l'intérieur des GRH ont été pondérés par les poids d'échantillonnage.

```
%BOOTUP_1DEG(BASE=work,ECHMEN=ech,
          ITBOOT=1000,
          IDMEN=idm,STMEN=stmen,DMEN=dmen,
          RMEN=rmen,DRMEN=drmen,DCMEN=dcmen,GRHMEN=GRHmen,WGRHMEN=1,
          XMENQUANT=x0 x1,XMENQUALI=,
          SORT_MEN=ech_boot,SORT_RMEN=echr_boot);
```

C. Programme SAS pour l'échantillonnage à deux degrés

Dans cette section, nous présentons la macro SAS élaborée pour mettre en œuvre la méthodologie proposée en cas d'échantillonnage des ménages et un sous-échantillonnage de personnes (échantillonnage à deux degrés). Le paramétrage du programme SAS pour le calcul des poids bootstrap est présenté à la section C.1. Par souci de clarté, un petit exemple est présenté à la section C.2.

C.1 Programme de calcul des poids bootstrap

La macro SAS `%BOOTUP_2DEG` permet de calculer les poids bootstrap pour une enquête auprès des ménages avec sous-échantillonnage de personnes, et de tenir compte de la correction de la non-réponse totale au moyen de groupes de réponse homogènes, et du calage des poids, à la fois pour les ménages et les personnes.

Les paramètres avec un signe d'égalité sont obligatoires. Toutes les variables d'identification doivent être de type caractère.

Les paramètres liés à la base de données sont :

- `BASE` : bibliothèque contenant les tables SAS `ECHMEN` et `ECHIND`. La valeur par défaut est `BASE=WORK`.
- `BASESOR` : bibliothèque contenant la sortie. La valeur par défaut est `BASESOR=WORK`.
- `ECHMEN=` : table SAS contenant la liste des ménages échantillonnés dans la population. Les non-répondants au niveau ménage doivent aussi être inclus dans cette table.
- `ECHIND=` : table SAS contenant la liste des personnes échantillonnées dans tous les ménages répondants. Les personnes non répondantes doivent aussi être incluses dans cette table.

Les paramètres liés au bootstrap sont :

- `ITBOOT` : nombre d'itérations bootstrap. La valeur par défaut est `ITBOOT=1000`.

Les paramètres liés aux variables nécessaires dans la table SAS du ménage `ECHMEN` sont :

- `IDMEN=` : liste des variables identifiant le ménage. Cette variable est requise dans `ECHMEN` et `ECHIND`.
- `STMEN` : liste des variables de stratification utilisées pour la sélection de l'échantillon.
- `DMEN` : poids d'échantillonnage du ménage.
- `RMEN` : indicateur de réponse du ménage (1 pour un répondant, 0 pour un non-répondant).
- `DRMEN` : poids d'échantillonnage du ménage, corrigé pour tenir compte de la non-réponse. Les valeurs sont nécessaires uniquement pour les répondants.
- `DCMEN` : poids calé. Les valeurs sont nécessaires uniquement pour les répondants.
- `GRHMEN` : liste des variables identifiant les groupes de réponse homogènes pour les ménages.
- `WGRHMEN` : pondération utilisée dans le calcul des probabilités de réponse à l'intérieur des GRH :

- Avec WGRHMEN=0, les taux de réponse ne sont pas pondérés. Il s'agit de la valeur par défaut.
- Avec WGRHMEN=1, les taux de réponse sont pondérés par les poids de sondage DMEN.

- XMENQUANT : liste des variables quantitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.
- XMENQUALI : liste des variables qualitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.

Les paramètres liés aux variables nécessaires dans la table SAS des personnes ECHIND sont :

- ID_IND= : liste des variables identifiant la personne (variable de caractères).
- R_IND : indicateur de réponse de la personne (1 pour un répondant, 0 pour un non-répondant).
- DR_IND : poids de la personne, corrigé pour tenir compte à la fois de la non-réponse totale du ménage et de la personne. Les valeurs sont nécessaires uniquement pour les répondants.
- DC_IND : poids calé. Les valeurs sont nécessaires uniquement pour les répondants.
- PIKSACI= : probabilité d'inclusion conditionnelle de la personne à l'intérieur de son ménage.
- GRH_IND : liste des variables identifiant les groupes de réponse homogènes.
- WGRH_IND : pondération utilisée dans le calcul des probabilités de réponse à l'intérieur des GRH :
  - Avec WGRH_IND=0, les taux de réponse ne sont pas pondérés. Il s'agit de la valeur par défaut.
  - Avec WGRH_IND=1, les taux de réponse sont pondérés par les poids de sondage des personnes.
  - Avec WGRH_IND=2, les taux de réponse sont pondérés par les poids des personnes, ajustés en fonction de la non-réponse totale du ménage.
- XINDQUANT : liste des variables quantitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.
- XINDQUALI : liste des variables qualitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.

Les paramètres liés à la sortie sont :

- SORT_MEN : table SAS contenant tous les ménages échantillonnés et les poids d'échantillonnage bootstrap WB_D1,...,WB_D&ITBOOT pour l'ensemble de l'échantillon.
- SORT_RMEN : table SAS contenant tous les ménages répondants et les poids bootstrap
  - WB_N1,...,WB_N&ITBOOT corrigés pour tenir compte de la non-réponse,
  - WB_C1,...,WB_C&ITBOOT corrigés pour tenir compte de la non-réponse et du calage.
- SORT_RIND : table SAS contenant toutes les personnes répondantes à l'intérieur des ménages répondants, et les poids bootstrap
  - WB_N1,...,WB_N&ITBOOT corrigés pour tenir compte de la non-réponse du ménage,
  - WB_NN1,...,WB_NN&ITBOOT corrigés pour tenir compte à la fois de la non-réponse du ménage et de la non-réponse des personnes,
  - WB_C1,...,WB_C&ITBOOT corrigés pour tenir compte de la non-réponse et du calage.

C.2 Petit exemple

Considérons l'exemple traité à la section 2.2.4. L'échantillon des ménages et l'échantillon des personnes sont les suivants :

```
data echmen;
input idm$ stmen$ dmen rmen GRHmen$ drmen dcmen x0 x1;
cards;
A  1  4   1  aa  4.44   4.01   1  1
B  1  4   0  aa  .      .      .  .
C  1  4   0  bb  .      .      .  .
D  1  4   1  bb  5.54   4.87   1  0
E  1  16  1  bb  22.15  19.98  1  1
F  1  16  1  aa  17.78  15.63  1  0
G  1  16  0  bb  .      .      .  .
H  1  16  1  bb  22.15  19.98  1  1
I  1  16  1  bb  22.15  19.49  1  0
J  1  16  1  aa  17.78  16.03  1  1
;run;

data echind;
input idm$ idi$ piksaci dr1_ind rind GRHind$ phat_ind dr2_ind xi1 xi2
dc_ind;
cards;
A   i01   0.34   13.06   1   g1   0.75   17.41   1   3   19.61
D   i04   1.00   5.54    0   g2   0.33   .       .   .   .
E   i06   0.34   65.15   1   g1   0.75   86.86   1   2   53.93
F   i08   0.33   53.88   1   g1   0.75   71.84   1   3   78.43
H   i11   0.50   44.30   0   g1   0.75   .       .   .   .
I   i13   1.00   22.15   1   g2   0.33   67.12   1   1   48.04
J   i14   1.00   17.78   0   g2   0.33   .       .   .   .
;run;
```

Nous pouvons obtenir $B = 1\,000$ poids bootstrap comme suit. Étant donné que `WGRHMEN=1`, on suppose que quand la non-réponse totale des ménages a été initialement corrigée par la méthode des GRH, les taux de réponse à l'intérieur des GRH ont été pondérés par les poids d'échantillonnage. Étant donné que `WGRH_IND=0`, on suppose que quand la non-réponse totale des personnes a été initialement corrigée par la méthode des GRH, les taux de réponse à l'intérieur des GRH n'étaient pas pondérés.

```
%bootup_2deg(BASE=work,BASESOR=work,ECHMEN=echmen,ECHIND=echind,
  ITBOOT=1000,
  IDMEN=idm,STMEN=stmen,DMEN=dmen,RMEN=rmen,DRMEN=drmen,GRHMEN=GRHmen,W
  GRHMEN=0,
  DCMEN=dcmen,XMENQUANT=x0 x1,XMENQUALI=,
  ID_IND=idi,R_IND=rind,DR_IND=dr2_ind,PIKSACI=piksaci,GRH_IND=GRHind,W
  GRH_IND=0,
  DC_IND=dc_ind,XINDQUANT=xi1 xi2,XINDQUALI=,
  SORT_MEN=sort_men,SORT_RMEN=sort_rmen,
  SORT_RIND=sort_rind);
```

Bibliographie

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3), 445-458.

Beaumont, J.-F., et Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *Revue Internationale de Statistique*, 80(1), 127-148.

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

Chauvet, G. (2007). *Méthodes de Bootstrap en population finie*. Thèse de doctorat, University of Rennes 2.

Chauvet, G. (2015). Coupling methods for multistage sampling. *The Annals of Statistics*, 43(6), 2484-2506.

Chauvet, G., et Vallée, A.-A. (2018). Inference for two-stage sampling designs with application to a panel for urban policy. *arXiv preprint arXiv:1808.09758*.

Davison, A.C., et Hinkley, D.V. (1997). Bootstrap methods and their application, volume 1, *Cambridge Series in Statistical and Probabilistic Mathematics*.

Davison, A.C., et Sardy, S. (2007). Resamping variance estimation in surveys with missing data. *Journal of Official Statistics*, 23(3), 371-386.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 2, 219-230. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1999002/article/4882-fra.pdf.

Girard, C. (2009). The Rao-Wu rescaling bootstrap: from theory to practice. Dans *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, pages 2-4. Citeseer.

Haziza, D., et Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *Revue Internationale de Statistique*, 75(1), 25-43.

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

Juillard, H., et Chauvet, G. (2018). Estimation de la variance sous non-réponse monotone pour une enquête par panel. *Techniques d'enquête*, 44, 2, 295-317. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018002/article/54952-fra.pdf.

Kim, J.K., et Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics/La revue canadienne de statistique*, 35(4), 501-514.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Aassociation*, 101(473), 312-320.

Kott, P.S. (2012). Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes. *Techniques d'enquête*, 38, 1, 103-107. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012001/article/11689-fra.pdf.

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

McCarthy, P., et Snowden, C. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, (95), 1-23.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401), 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.

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

Shao, J. (1994). *L*-Statistics in complex survey problems. *The Annals of Statistics*, 22(2), 946-967.

Shao, J., et Rao, J. (1993). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyā: The Indian Journal of Statistics, Series B*, 393-414.

Shao, J., et Tu, D.S. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Tillé, Y. (2011). *Sampling Algorithms*. Springer.

Yeo, D., Mantel, H. et Liu, T.-P. (1999). Bootstrap variance estimation for the national population health survey. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association. Citeseer.

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 377-387
Statistique Canada, nº 12-001-X au catalogue

Utilisation d'un autre estimateur jackknife de la variance aux fins du calage des poids pour tenir compte de la non-réponse totale dans une enquête complexe

Phillip S. Kott et Dan Liao[1]

Résumé

La pondération par calage est un moyen statistiquement efficace de traiter la non-réponse totale. En supposant que le modèle (ou la sortie) de la réponse justifiant l'ajustement du poids de calage est exact, il est souvent possible de mesurer la variance des estimations de façon asymptotique et sans biais. Une des manières d'estimer la variance consiste à créer des poids de rééchantillonnage jackknife. Cependant, il arrive que la méthode classique de calcul des poids de rééchantillonnage jackknife pour les poids d'analyse calés échoue. Dans ce cas, il existe généralement une autre méthode de calcul des poids de rééchantillonnage jackknife. Cette méthode est décrite ici et appliquée à un exemple simple.

Mots-clés : Poids d'analyse; estimateur de la variance obtenu par linéarisation; estimateur jackknife de la variance avec suppression d'une UPE; poids de rééchantillonnage; asymptotiquement sans biais; modèle de réponse logistique borné.

1. Introduction

La pondération par calage est une méthode d'ajustement des poids dans la théorie de l'échantillonnage probabiliste consistant à forcer la somme pondérée de chaque variable d'un ensemble de variables d'enquête à égaler une cible déterminée. Quand cela se produit, on dit que les poids d'analyse satisfont à l'équation de calage. Plusieurs raisons conduisent à caler les poids d'analyse. La raison sur laquelle nous nous pencherons dans l'article est l'élimination du biais de sélection potentiel découlant de la non-réponse totale.

Dans la littérature sur l'échantillonnage d'enquête, il est courant de soutenir que l'ajustement du poids de calage d'un répondant estime implicitement l'inverse de sa probabilité de réponse (voir, par exemple, la section 5.1 de Fuller, 2009). Kott et Liao (2012) montrent que l'utilisation de la pondération par calage pour tenir compte de la non-réponse totale peut fournir une *double protection* contre le biais de non-réponse quand on estime un total de population. Cela signifie que si un modèle de résultats linéaire ou un modèle de sélection implicite se vérifie, l'estimateur obtenu est asymptotiquement sans biais dans un certain sens. Ils décrivent ensuite un estimateur de la variance obtenu par linéarisation pour un total estimé basé sur un échantillon stratifié à plusieurs degrés (ou à un degré) avec des poids d'analyse ajustés par calage.

Le bref traitement aux sections 2 et 3 de la pondération par calage pour la non-réponse et de l'estimation de la variance par linéarisation pour un estimateur calé d'un total de population est expliqué plus en détail dans Kott et Liao. On y trouve des démonstrations des diverses affirmations énoncées dans ces sections. Ils établissent la théorie étayant l'estimation de la variance par la méthode du jackknife.

1. Phillip S. Kott, statisticien principal de recherche, RTI International, Rockville, Maryland 20852, États-Unis. Courriel : pkott@rti.org; Dan Liao, statisticien de recherche, RTI International, Rockville, Maryland 20852, États-Unis.

En supposant un échantillon probabiliste stratifié à plusieurs degrés, un estimateur jackknife classique de la variance (avec suppression d'une UPE) crée des ensembles de poids de rééchantillonnage, un ensemble correspondant à chaque unité primaire d'échantillonnage (UPE) sélectionnée. Une UPE sélectionnée est supprimée à la fois et les poids de rééchantillonnage de ses éléments sous-échantillonnés sont fixés à zéro. Pour compenser, les poids de probabilité de rééchantillonnage des éléments restants dans la même strate que l'UPE abandonnée sont augmentés du facteur $n_h / (n_h - 1)$, où n_h est le nombre initial d'UPE sélectionnées à partir de la strate. Les poids de probabilité de rééchantillonnage sont calés d'une manière analogue aux poids d'analyse initiaux. La section 4 décrit ce jackknife et montre sa quasi-égalité avec l'estimateur de la variance obtenu par linéarisation presque sans biais pour un total de population.

L'avantage d'un jackknife avec suppression d'une UPE par rapport à un estimateur de la variance obtenu par linéarisation est qu'une fois les poids de rééchantillonnage calculés, l'estimation de la variance de la fonction lisse des totaux estimés (comme un coefficient de régression) est simple. Krewski et Rao (1981) proposent un traitement rigoureux du jackknife avec suppression d'une UPE et de ses propriétés.

Parfois, il n'y a pas de solution à l'équation de calage quand on commence par un ensemble de poids de probabilité de rééchantillonnage. La principale contribution du présent article se trouve dans le reste de la section 4, qui décrit et justifie une autre méthode de construction de poids de rééchantillonnage jackknife qui peut habituellement surmonter ce problème. Cette méthode a été introduite par Kott (2006) à d'autres fins.

La section 6 utilise une fonction d'ajustement des poids décrite à la section 5 pour illustrer la mise en œuvre de cette méthode. Elle compare ensuite favorablement les résultats de la méthode à ceux de deux méthodes concurrentes courantes. La section 7 traite d'une variante de l'autre méthode jackknife.

2. Pondération par calage

Supposons que nous avons un échantillon tiré aléatoirement S d'une population finie U. En l'absence de non-réponse (ainsi que d'erreur de couverture et d'erreur de mesure), la pondération par calage crée un ensemble de poids d'analyse, $\{w_k \mid k \in S\}$, qui ne dépend pas des valeurs d'enquête d'intérêt qui

1. sont proches des poids de probabilité inverse initiaux, $d_k = 1/\pi_k$ où π_k est la probabilité de sélection du k^{e} élément sélectionné;
2. satisfont à un ensemble d'équations de calage linéaires, une pour chaque composante de $\mathbf{z}_k$, un vecteur de variables auxiliaires avec des totaux de population connus :

$$\sum_{k \in S} w_k \mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k$$

« Proches » signifie qu'à mesure que l'échantillon croît arbitrairement, la différence entre w_k et d_k disparaît dans la probabilité. Voir un traitement plus formel de la structure asymptotique supposée dans Isaki et Fuller (1982).

La plupart des enquêtes connaissent une non-réponse totale qui échappe au contrôle des statisticiens. On est obligé de supposer, explicitement ou implicitement, qu'il existe un type de modèle d'ajustement

pour tenir compte de la non-réponse. Un modèle de résultat (aussi appelé « modèle de prédiction ») sur une variable d'enquête d'intérêt suppose habituellement que le mécanisme de réponse/non-réponse, comme le plan de sondage, est ignorable. Un modèle de réponse suppose que le mécanisme de réponse se comporte comme une phase du sous-échantillonnage de Poisson (c'est-à-dire indépendamment). La double protection signifie que si le modèle de prédiction *ou* de réponse est spécifié correctement, l'estimateur sera presque (c'est-à-dire asymptotiquement) sans biais dans un certain sens. Ici, nous supposerons un modèle de réponse correctement spécifié.

Soit R le sous-ensemble de S contenant les répondants de l'enquête (par souci de simplicité, nous ne tenons pas compte de la possibilité de non-réponse partielle). L'échantillon des répondants peut être calé selon la population entière U :

$$\sum_{k\in R} w_k \mathbf{z}_k = \sum_{k\in U} \mathbf{z}_k \tag{2.1}$$

ou selon l'échantillon initial S :

$$\sum_{k\in R} w_k \mathbf{z}_k = \sum_{k\in S} d_k \mathbf{z}_k . \tag{2.2}$$

Nous supposons un modèle de réponse dans lequel la probabilité de réponse pour chaque $k \in U$, p_k, est une fonction indépendante ayant la forme $p_k = p(\boldsymbol{\gamma}^T \mathbf{x}_k)$, où $p(.)$ est une fonction monotonique lisse, et à la fois le vecteur connu $\mathbf{x}_k$ et le vecteur de paramètre inconnu $\boldsymbol{\gamma}$ ont le même nombre de composantes que $\mathbf{z}_k$. Dans la majorité de la littérature $\mathbf{x}_k$ est égal à $\mathbf{z}_k$, mais la plus grande partie de la théorie reste vraie si ce n'est pas le cas.

Supposons un vecteur $\mathbf{g}$ tel que l'insertion de $w_k = d_k / p(\mathbf{g}^T \mathbf{x}_k)$ résout soit l'équation de calage dans (2.1) soit dans (2.2), alors $\mathbf{g}$ est un estimateur convergent pour $\boldsymbol{\gamma}$. Kott et Liao (2017) décrivent ce qu'il faut faire si le nombre de composantes dans $\mathbf{x}_k$ est plus petit que leur nombre dans $\mathbf{z}_k$.

La fonction $f(\mathbf{g}^T \mathbf{x}_k) = 1/p(\mathbf{g}^T \mathbf{x}_k)$ est appelée *fonction d'ajustement des poids*. Le théorème de la valeur moyenne nous dit que sous des conditions faibles $f(\mathbf{g}^T \mathbf{x}_k) - f(\boldsymbol{\gamma}^T \mathbf{x}_k) \approx f'(\mathbf{g}^T \mathbf{x}_k)(\mathbf{g}-\boldsymbol{\gamma})^T \mathbf{x}_k$. Par conséquent, à mesure que l'échantillon des répondants croît de façon arbitraire $f(\mathbf{g}^T \mathbf{x}_k)$ converge vers $f(\boldsymbol{\gamma}^T \mathbf{x}_k) = 1/p_k$ et $\mathbf{g}$ converge vers $\boldsymbol{\gamma}$.

3. Estimation de la variance par linéarisation

Lors du calage de l'échantillon des répondants selon l'échantillon complet avec (2.2), l'estimateur par calage pour un total de population, $t = \sum_R w_k y_k$ peut être exprimé comme suit :

$$\begin{aligned}
t &= \sum_{k\in S} d_k \mathbf{z}_k^T \mathbf{b} + \sum_{k\in R} d_k f(\mathbf{g}^T \mathbf{x}_k)(y_k - \mathbf{z}_k^T \mathbf{b}) \\
&\approx \sum_{k\in S} d_k \mathbf{z}_k^T \mathbf{b} + \sum_{k\in R} d_k f(\boldsymbol{\gamma}^T \mathbf{x}_k)(y_k - \mathbf{z}_k^T \mathbf{b}) + \sum_{k\in R} d_k f'(\mathbf{g}^T \mathbf{x}_k)\left[(\mathbf{g}-\boldsymbol{\gamma})^T \mathbf{x}_k\right](y_k - \mathbf{z}_k^T \mathbf{b}) \\
&= \sum_{k\in S} d_k \mathbf{z}_k^T \mathbf{b} + \sum_{k\in R} d_k f(\boldsymbol{\gamma}^T \mathbf{x}_k)(y_k - \mathbf{z}_k^T \mathbf{b}) + (\mathbf{g}-\boldsymbol{\gamma})^T \sum_{k\in R} d_k f'(\mathbf{g}^T \mathbf{x}_k)\mathbf{x}_k (y_k - \mathbf{z}_k^T \mathbf{b}) \\
&= \sum_{k\in S} d_k \mathbf{z}_k^T \mathbf{b} + \sum_{k\in R} d_k p_k^{-1}(y_k - \mathbf{z}_k^T \mathbf{b}),
\end{aligned} \tag{3.1}$$

où

$$\mathbf{b} = \left[\sum_{k\in R} d_k f'(\mathbf{g}^T\mathbf{x}_k)\,\mathbf{x}_k\mathbf{z}_k^T\right]^{-1} \sum_{k\in R} d_k f'(\mathbf{g}^T\mathbf{x}_k)\,\mathbf{x}_k y_k.$$

L'étape clé ici est que $\mathbf{b}$ a été défini de telle sorte que $\sum_R d_k f_k'(\mathbf{g}^T\mathbf{x}_k)\,\mathbf{x}_k(y_k - \mathbf{z}_k^T\mathbf{b}) = 0.$ Notons que $f'(.)$ dans $\mathbf{b}$ est la *dérivée* de la fonction d'ajustement de la pondération.

Soit $\mathbf{b}^*$ la limite de probabilité de $\mathbf{b}$ quand l'échantillon de répondants (d'UPE) croît arbitrairement. La variance de t selon le plan initial et le modèle de sélection équivaut presque à la variance de $\sum_S d_k q_k^*$, où

$$q_k^* = \mathbf{z}_k^T\mathbf{b}^* + p_k^{-1}(y_k - \mathbf{z}_k^T\mathbf{b}^*)\,I_k,$$

et $I_k = 1$ quand k est un répondant d'unité et 0 autrement.

Pour de nombreux plans, on peut calculer approximativement q_k^* en remplaçant $\mathbf{b}^*$ par $\mathbf{b}$ et p_k^{-1} par $f(\mathbf{g}^T\mathbf{x}_k)$, et la variance de $\sum_S d_k q_k$ est estimée selon le plan initial comme si les $q_k = \mathbf{z}_k^T\mathbf{b} + f(\mathbf{g}^T\mathbf{x}_k)(y_k - \mathbf{z}_k^T\mathbf{b})\,I_k$ étaient des constantes. Lors du calage de l'échantillon des répondants selon la population avec l'équation (2.1), $\sum_S d_k\mathbf{z}_k^T\mathbf{b}$ dans l'équation (3.1) est remplacé par $\sum_U \mathbf{z}_k^T\mathbf{b}$, ce qui ne contribue pas à la variance, donc $q_k = f(\mathbf{g}^T\mathbf{x}_k)(y_k - \mathbf{z}_k^T\mathbf{b})\,I_k$. D'une façon ou d'une autre, le remplacement de q_k^* par q_k a tendance à sous-estimer les variances avec des échantillons *finis* (le remplacement est asymptotiquement ignorable), car $e_k = (y_k - \mathbf{z}_k^T\mathbf{b})^2$ tend à être plus petit que $e_k^* = (y_k - \mathbf{z}_k^T\mathbf{b}^*)^2$.

Étant donné un échantillon probabiliste stratifié à plusieurs degrés avec n_h UPE échantillonnées parmi les H strates, soit S_{hj} le sous-échantillon d'éléments dans chaque UPE j dans la strate h. Un estimateur obtenu par linéarisation presque sans biais pour la variance de t est

$$v(t) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1}\left[\sum_{j=1}^{n_h}\left(\sum_{k\in S_{hj}} d_k q_k\right)^2 - \frac{1}{n_h}\left(\sum_{a=1}^{n_h}\sum_{\kappa\in S_{ha}} d_\kappa q_\kappa\right)^2\right], \tag{3.2}$$

où $q_k = \alpha\mathbf{z}_k^T\mathbf{b} + f(\mathbf{g}^T\mathbf{x}_k)\,e_k I_k$, et $\alpha = 1$ quand l'échantillon des répondants est calé par rapport à l'échantillon initial et 0 quand l'échantillon des répondants est calé par rapport à la population. Comme c'est souvent le cas en pratique et comme on le voit ici, l'équation (3.2) suppose qu'on perd peu en traitant la sélection d'UPE dans les strates comme si elle avait été tirée avec remise, ce qui évite la nécessité d'une correction de population finie.

4. Estimation de la variance par la méthode du jackknife

Soit S_{h+} tous les éléments échantillonnés dans la strate h. La réplique jackknife (avec suppression d'une UPE) hj^{e} classique pour un total estimé $t = \sum_R w_k y_k$ est

$$t^{(hj)} = \sum_{k \in R} w_k^{(hj)} y_k = \sum_{k \in S} d_k^{(hj)} f(g^{(hj)T} \mathbf{x}_k) I_k y_k, \tag{4.1}$$

où $d_k^{(hj)} = 0$ quand $k \in S_{hj}$

$d_k^{(hj)} = [n_h / (n_h - 1)] d_k$ quand $k \in S_{h+}$ mais $k \notin S_{hj}$

$d_k^{(hj)} = d_k$ autrement,

et $\mathbf{g}^{(hj)}$ résout l'équation de calage de rééchantillonnage :

$$\sum_{k \in R} d_k^{(hj)} f(\mathbf{g}^{(hj)T} \mathbf{x}_k) \mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k \text{ quand l'échantillon des répondants est calé selon } U, \text{ ou}$$

$$\sum_{k \in R} d_k^{(hj)} f(\mathbf{g}^{(hj)T} \mathbf{x}_k) \mathbf{z}_k = \sum_{k \in S} d_k^{(hj)} \mathbf{z}_k \text{ quand l'échantillon des répondants est calé selon } S$$

pour chaque hj. Observons que $t^{(hj)}$ est une estimation du total de la population $\sum_U y_k$ sans suppression de l'UPE hj.

L'estimateur jackknifede la variance avec suppression d'une UPE pour t est

$$\text{var}_J(t) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (t^{(hj)} - t)^2. \tag{4.2}$$

Soit $e_k^* = (y_k - \mathbf{z}_k^T \mathbf{b}^*)$. Par conséquent,

$$\begin{aligned} t^{(hj)} - t &= \sum_{k \in S} \left\{ (d_k^{(hj)} - d_k) \alpha \mathbf{z}_k^T \mathbf{b}^* + \left[d_k^{(hj)} f(\mathbf{g}^{(hj)T} \mathbf{x}_k) - d_k f(\mathbf{g}^T \mathbf{x}_k) \right] I_k e_k^* \right\} \\ &\approx \sum_{k \in S} (d_k^{(hj)} - d_k) \left[\alpha \mathbf{z}_k^T \mathbf{b}^* + f(\mathbf{g}^T \mathbf{x}_k) I_k e_k^* \right]. \end{aligned}$$

Nous pouvons alors conclure que $\text{var}_J(t)$ est à peu près égal au jackknife avec suppression d'une UPE de $\sum_S d_k q_k^*$, où $q_k^* = \alpha \mathbf{z}_k^T \mathbf{b}^* + f(\mathbf{g}^T \mathbf{x}_k) I_k e_k^*$, d'après un échantillon stratifié à plusieurs degrés avec H strates et n_h UPE dans la strate h. Avec un peu d'algèbre, on montre que le jackknife avec suppression d'une UPE pour $\sum_S d_k q_k^*$ est égal à $\text{var}(t)$ dans l'équation (3.2) avec q_k^* remplaçant q_k parce que

$$\sum_{k \in S} (d_k^{(hj)} - d_k) q_k^* = \sum_{h=1}^{H} \sum_{k \in S_{h+}} (d_k^{(hj)} - d_k) q_k^*,$$

où

$$\begin{aligned} \sum_{k \in S_{h+}} (d_k^{(hj)} - d_k) q_k^* &= \sum_{k \in S_{h+}} \frac{1}{n_h - 1} d_k q^* - \frac{n_h}{n_h - 1} \sum_{k \in S_{hj}} d_k q^* \\ &= -\frac{n_h}{n_h - 1} \left(\sum_{k \in S_{hj}} d_k q^* - \frac{1}{n_h} \sum_{k \in S_{h+}} d_k q^* \right). \end{aligned}$$

Notons que la contribution à l'estimateur jackknife de la variance pour la hj^{e} réplique provient principalement de la hj^{e} UPE.

Observons que le petit biais à la baisse dans les échantillons finis causé par le remplacement de q_k par q_k^* dans $\text{var}(t)$ ne s'applique pas à $\text{var}_J(t)$ dans l'équation (4.2). Ce dernier peut avoir une légère tendance à être biaisé à la hausse dans les échantillons finis parce que $\mathbf{g}^{(hj)}$ et $\mathbf{g}$, bien qu'ils soient tous deux des estimateurs convergents de $\boldsymbol{\gamma}$, n'ont pas besoin d'être exactement égaux.

Un problème se pose parfois avec le calcul de l'estimateur jackknife de la variance $\text{var}_J(t)$ dans la pratique. Ce problème se produit quand $f(.)$ est tel que, bien qu'il y ait une valeur $\mathbf{g}$ satisfaisant l'équation de calage dans (2.1) ou (2.2), aucune valeur $\mathbf{g}^{(hj)}$ ne satisfait son analogue pour au moins une réplique jackknife hj. Quand cela se produit, on peut suivre une suggestion de Kott (2006) et calculer $w_k^{(hj)}$ dans l'équation (4.1) avec la solution suivante :

$$\tilde{w}_k^{(hj)} = d_k^{(hj)} f(g^T\mathbf{x}_k) + \left[\mathbf{c}^{(hj)} - \sum_{\kappa\in R} d_\kappa^{(hj)} f(g^T\mathbf{x}_\kappa)\mathbf{z}_\kappa^T\right]\left[\sum_{\kappa\in R} d_\kappa^{(hj)} f'(g^T\mathbf{x}_\kappa)\mathbf{x}_\kappa\mathbf{z}_\kappa^T\right]^{-1} d_k^{(hj)} f'(g^T\mathbf{x}_k)\mathbf{x}_k, \quad (4.3)$$

où $\mathbf{c}^{(hj)}$ est la cible de calage pour la hj^{e} réplique :

$$\mathbf{c}^{(hj)} = \sum_{\kappa\in U} \mathbf{z}_\kappa^T \text{ quand l'échantillon des répondants est calé selon } U,$$

et

$$\mathbf{c}^{(hj)} = \sum_{\kappa\in S} d_\kappa^{(hj)}\mathbf{z}_\kappa^T \text{ quand l'échantillon des répondants est calé selon } S.$$

Selon le plan de sondage $\sum_R \tilde{w}_k^{(hj)}\mathbf{z}_k = \mathbf{c}^{(hj)}$.

Si l'on suppose $e_k = y_k - \mathbf{z}_k^T\mathbf{b}$, on peut voir qu'avec $\tilde{t}^{(hj)} = \sum_R \tilde{w}_k^{(hj)} y_k$,

$$\begin{aligned}\tilde{t}^{(hj)} - t &= \sum_{k\in S} (d_k^{(hj)} - d_k)\left[\alpha\mathbf{z}_k^T\mathbf{b} + f(\mathbf{g}^T\mathbf{x}_k) I_k e_k\right] \\ &\approx \sum_{k\in S} (d_k^{(hj)} - d_k)\left[\alpha\mathbf{z}_k^T\mathbf{b}^* + f(\mathbf{g}^T\mathbf{x}_k) I_k e_k^*\right]\end{aligned}$$

de sorte que l'autre estimateur jackknife de la variance $\text{var}_{AJ}(t)$ calculé avec $\tilde{t}^{(hj)}$ au lieu de $t^{(hj)}$ est presque sans biais. Notons que la seule restriction possible sur le calcul de $\tilde{w}_k^{(hj)}$ est que $\sum_R d_k^{(hj)} f'(g^T\mathbf{x}_k)\,\mathbf{x}_k\mathbf{z}_k^T$ soit non singulier.

Observons que l'équation (4.3) peut être réécrite ainsi :

$$\tilde{w}_k^{(hj)} = \tilde{d}_k^{(hj)} \tilde{f}(\tilde{\mathbf{g}}^{(hj)T}\tilde{\mathbf{x}}_k) \quad (4.4)$$

où $\tilde{d}_k^{(hj)} = d_k^{(hj)} f(g^T\mathbf{x}_k)$,

$$\tilde{f}(\tilde{\mathbf{g}}^{(hj)T}\tilde{\mathbf{x}}_k) = 1 + \tilde{\mathbf{g}}^{(hj)T}\tilde{\mathbf{x}}_k,$$

$$\tilde{\mathbf{g}}^{(hj)T} = \left[\mathbf{c} - \sum_{\kappa \in R} \tilde{d}_{\kappa}^{(hj)} \mathbf{z}_{\kappa}^{T} \right] \left[\sum_{\kappa \in R} \tilde{d}_{\kappa}^{(hj)} \tilde{\mathbf{x}}_{\kappa} \mathbf{z}_{\kappa}^{T} \right]^{-1},$$

et

$$\tilde{\mathbf{x}}_k = \frac{f'(\mathbf{g}^T \mathbf{x}_k)}{f(\mathbf{g}^T \mathbf{x}_k)} \mathbf{x}_k .$$

Cette équation traite la hj^{e} réplique comme l'échantillon complet. La fonction d'ajustement des poids $\tilde{f}(.)$ est linéaire et $\tilde{f}_k$ n'est pas limité à des valeurs positives même si les f_k le sont. De plus, notons que même quand $\mathbf{x}_k = \mathbf{z}_k$, $\tilde{\mathbf{x}}_k$ n'est pas égal à $\mathbf{z}_k$, sauf si $f(.) = \exp(.)$.

5. Le modèle de réponse logistique (borné)

Jusqu'à maintenant, nous n'avons pas spéficié de fonction de réponse, $p(.) = 1/f(.)$. Envisageons maintenant un modèle de réponse logistique (ou logit) borné de forme :

$$p(\boldsymbol{\gamma}^T \mathbf{x}_k) = \frac{1 + \exp(\boldsymbol{\gamma}^T \mathbf{x}_k)/U}{L + \exp(\boldsymbol{\gamma}^T \mathbf{x}_k)}. \tag{5.1}$$

où $1 \le L < U$. Quand $L = 1$ et U est infini, il s'agit d'un modèle de réponse logistique standard, où la probabilité de réponse peut varier de 0 à 1, excluant les extrémités. Pour les valeurs finies de L et U, la probabilité de réponse bornée se situe entre $1/U$ et $1/L$. La valeur de la fonction d'ajustement varie donc de L à U. En pratique, L est généralement fixé à 1, tandis que U est souvent fixé aussi bas que l'échantillon le permet pour que l'équation de calage se vérifie.

Les procédures de calage en langage SUDAAN® (Research Triangle Institute, 2012), WTADJUST lorsque $\mathbf{x}_k = \mathbf{z}_k$ et WTADJX sinon, correspondent à une fonction équivalente d'ajustement des poids :

$$f(\mathbf{g}^T \mathbf{x}_k) = \frac{L + B \exp(A \mathbf{g}^T \mathbf{x}_k)}{1 + B \exp(A \mathbf{g}^T \mathbf{x}_k)/U}, \tag{5.2}$$

où $A = \frac{U-L}{(C-L)(U-C)}$, $B = U \frac{C-L}{U-C}$, et $L < C < U$.

Le choix de C permet de déterminer que $\mathbf{g}$ satisfait l'équation de calage, mais n'affecte pas la valeur de l'ajustement des poids en soi, $f_k = f(\mathbf{g}^T \mathbf{x}_k)$. Par conséquent, C peut être toute valeur entre L et U. Quand $L = 1$, $C = 2$ et U est infini, $\mathrm{A} = \mathrm{B} = 1$.

Un petit calcul révèle avec la fonction d'ajustement des poids dans l'équation (5.2) :

$$f_k' = f'(g^T \mathbf{x}_k) = \frac{(U - f_k)\,(f_k - L)}{(U - C)\,(C - L)},$$

qui est nécessaire pour calculer l'équation (4.3) ou (4.4). Le modèle exponentiel général dans les procédures de calage SUDAAN permet à L, C et U de varier d'un élément à l'autre, une souplesse difficile à interpréter dans la modélisation des réponses et qui n'est pas prise en compte ici.

Ce qui sera utile ici, bien que ce ne soit pas à des fins de modélisation, est la possibilité que L dans l'équation (5.2) soit 0 et que U soit infini. Lors de la résolution itérative d'une équation de calage pour $\mathbf{g}$ avec $f(\mathbf{g}^T\mathbf{x}_k) = \exp(\mathbf{g}^T\mathbf{x})$ au moyen de la méthode de Newton, les procédures de calage SUDAAN commencent par résoudre $\mathbf{g}_1$ dans l'équation de calage avec $f(\mathbf{g}_1^T\mathbf{x}_k) = 1 + \mathbf{g}_1^T\mathbf{x}_k$, ce qui est un résultat utile lors du calcul de poids de l'autre jackknife. (Les programmes fixent la première itération de l'ajustement des poids à $f_{k1} = \exp(\mathbf{g}_1^T\mathbf{x}_k)$, à partir de quoi $1 + \mathbf{g}_1^T\mathbf{x}_k$ se calcule facilement.)

6. Exemple de simulation

La population MU281 de municipalités dans Särndal, Swensson et Wretman (1992; les données de la version légèrement révisée se trouvent à l'adresse http://lib.stat.cmu.edu/datasets/mu284; une des municipalités a été accidentellement abandonnée dans cette analyse) a été augmenté d'un indicateur (RESP) pour déterminer si un élément (municipalité) répondrait s'il était échantillonné. Les probabilités de réponse par élément ont été générées à l'aide d'une fonction logistique de l'une des covariables de l'ensemble de données (le log de la population de 1975 de l'élément en milliers). La probabilité moyenne de réponse était d'environ 70 %.

Un échantillon aléatoire simple stratifié de 10 éléments toutes les 8 strates a été simulé 1 716 fois. Dans chaque échantillon simulé, les éléments avec RESP = 1 ont été traités comme des répondants, et l'échantillon des répondants a été calé selon l'échantillon complet au moyen de la fonction d'ajustement des poids de l'équation (5.2) avec une borne inférieure de 1 et une borne supérieure de 5. Dans le modèle de calage, les deux composantes de $\mathbf{x}_k$ étaient 1 et le log de la population de 1975 de l'élément en milliers; $\mathbf{z}_k$ a été réglé comme étant égal à $\mathbf{x}_k$. Dans 1 225 des 1 716 simulations, les échantillons des répondants ont été calés avec succès sur les deux composantes (c'est-à-dire qu'ils satisfaisaient l'équation de calage de 2.2) et ont produit des erreurs-types par linéarisation.

Les moyennes estimées (ratios de deux totaux estimés) et les erreurs-types (racines carrées des variances estimées) ont été calculées pour quatre variables :

P85	Population de 1985 (en milliers).
RM4T85	Recettes fiscales municipales de 1985 (en millions de couronnes).
ME84	Nombre d'employés municipaux en 1984.
REV84	Valeurs immobilières selon l'évaluation de 1984 (en millions de couronnes).

Bien que la procédure SUDAAN WTADJUST puisse calculer les erreurs-types lors de l'utilisation d'un jackknife avec suppression d'une UPE, elle ne peut pas le faire quand le calage d'une ou plusieurs répliques échoue. C'est pourquoi deux versions des erreurs-types du jackknife avec suppression d'une UPE classique ont été calculées au moyen d'une macro créée par les auteurs. Dans l'une d'elles, on utilise les poids « calés » imparfaits de la dernière itération pour les répliques ayant échoué. Dans l'autre, les

répliques dont le calage a échoué ont été abandonnées, et l'estimateur jackknife de la variance modifié suivant a été calculé :

$$\mathrm{var}_J^*(t) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h^*} \sum_{j=1}^{n_h^*} (t^{(hj)} - t)^2, \tag{6.1}$$

où n_h^* est le nombre de répliques dans la strate h qui ont été calées avec succès. Cet estimateur jackknife de la variance révisé a été proposé par Rust (1985) quand les répliques sont abandonnées au hasard, ce qui n'est pas le cas ici. Le code SUDAAN pouvant être appelé par SAS (SAS Institute Inc., 2015) utilisé dans l'analyse pour une seule simulation est disponible sur demande auprès des auteurs.

Parmi les 1 225 échantillons analysables, dans 867 simulations toutes les répliques utilisant la méthode du jackknife avec suppression d'une UPE classique avaient un calage réussi, tandis que les 358 autres simulations avaient au moins une réplique dont le calage ne réussissait pas après 50 itérations (la valeur par défaut est de 10). Le tableau 6.1 présente la moyenne des résultats dans les deux situations. Quand aucune réplique classique n'échouait, les erreurs-types de la méthode jackknife classique et de substitution sont proches (en moyenne) et légèrement supérieures à celles produites par la linéarisation comme le prédit la théorie (notons que les deux versions de jackknife classique avec suppression d'une UPE sont identiques).

Tableau 6.1
Erreurs-types fondées sur les méthodes de jackknife lors du calage de la non-réponse avec un modèle logistique borné

	Variable	Moyenne estimée	Erreur-type de la méthode de linéarisation	Erreur-type de l'autre méthode de jackknife	Erreur-type de la méthode de jackknife classique comprenant les répliques ayant échoué	Erreurs-types de la méthode de jackknife classique avec abandon des répliques ayant échoué
867 simulations sans échec de calage pour les répliques de la méthode classique	P85	22,41	2,06	2,09	2,11	2,11
	RMT85	167,70	17,31	17,72	17,86	17,86
	ME84	1 215,87	124,67	127,27	128,15	128,15
	REV84	2 425,52	212,83	217,00	219,67	219,67
358 simulations avec au moins un échec de calage pour une réplique de la méthode classique	P85	22,78	2,24	2,31	2,95	2,04
	RMT85	170,48	18,93	19,47	24,39	16,60
	ME84	1 236,75	135,94	139,65	175,99	121,31
	REV84	2 451,95	239,27	239,18	296,77	208,45

Quand le calage d'au moins une réplique échoue pour le jackknife classique avec suppression d'une UPE, les erreurs-types de l'autre méthode jackknife sont de nouveau proches des erreurs-types de la méthode de linéarisation, même si le calage échoue dans 114 de ces 358 simulations en raison d'une

(quasi) singularité dans au moins une des répliques. Il est toutefois clair que l'inclusion des répliques ayant échoué surestime l'erreur-type et leur abandon la sous-estime nettement par rapport à la linéarisation. Il semblerait que l'autre estimateur jackknife de la variance produise l'ensemble de poids de rééchantillonnage le plus utile dans cette situation.

Le tableau 6.1 compare les erreurs-types des jackknifes concurrents aux erreurs-types fondées sur la linéarisation plutôt qu'aux erreurs-types empiriques parce que la correction de la population finie a été ignorée. De plus, l'ajustement du modèle de réponse logistique borné dans les simulations n'était pas le modèle de réponse non borné servant à générer les réponses.

7. Discussion

Il y a une petite chance (environ 1,5 % dans nos simulations) que l'équation (4.4) donne des poids de rééchantillonnage négatifs. Or les procédures prédéfinies de nombreux progiciels statistiques (comme SAS) ne peuvent pas traiter les poids négatifs. Par conséquent, il se peut qu'il faille calculer les totaux estimés calculés à partir des poids de rééchantillonnage sans l'aide d'une procédure prédéfinie.

Il n'est pas nécessaire d'avoir accès à SUDAAN pour calculer des poids jackknife autres pour les estimateurs par calage. Les routines *gencalib* dans le progiciel « Sampling » de R (Tillé et Matei, 2016) peuvent réaliser le calage non seulement sous un modèle de réponse logistique borné, mais aussi selon un calage linéaire. Bien qu'il y ait des macros SAS équivalentes de WTADJUST, à notre connaissance, il n'existe actuellement aucune macro de pondération par calage SAS accessible au public utilisable quand $\tilde{\mathbf{x}}_k$ dans l'ajustement des poids (équation 4.4) n'est pas égal à $\mathbf{z}_k$. Espérons que ce sera bientôt le cas.

Remerciements

Les auteurs tiennent à remercier les rédacteurs dont les suggestions ont aidé à améliorer la qualité de cette note.

Bibliographie

Fuller, W. (2009). *Sampling Statistics*, New York: John Wiley & Sons, Inc., Hoboken.

Isaki, C., et Fuller, W. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Technique d'enquête*, 32, 2, 149-160. Article accessible a l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9547-fra.pdf.

Kott, P., et Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration weighting routine. *Survey Research Methods*, 6, 105-111.

Kott, P., et Liao, D. (2017). Calibration weighting for nonresponse that is not missing at random: Allowing more calibration than response-model variables. *Journal of Survey Statistics and Methodology*, 5, 159-174.

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

Research Triangle Institute (2012). *SUDAAN Language Manual, Release 11.0.* RTI International, Research Triangle Park, NC.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1, 381-397.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

SAS Institute Inc. (2015). SAS/STAT® 14.1 User's Guide. SAS Institute Inc., Cary, NC.

Tillé, Y., et Matei, A. (2016). Package 'sampling' disponible en ligne sur http://cran.r-project.org/web/packages/sampling/sampling.pdf.

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 389-399
Statistique Canada, nº 12-001-X au catalogue

Estimation sur petits domaines à l'aide du modèle au niveau de domaine de Fay-Herriot avec lissage et modélisation de variance d'échantillonnage

Yong You[1]

Résumé

Nous considérons ici le modèle d'estimation sur petits domaines de Fay-Herriot. Nous nous intéressons en particulier à l'incidence du lissage et de la modélisation de la variance d'échantillonnage sur les estimations par modèle. Nous présentons des méthodes permettant de lisser et de modéliser les variances d'échantillonnage et appliquons les modèles proposés à une analyse de données réelles. Nos résultats font voir qu'un lissage de variance d'échantillonnage est de nature à accroître l'efficacité et la précision de l'estimateur par modèle. Dans une modélisation de variance d'échantillonnage, les modèles hiérarchiques bayésiens de You (2016) et de Sugasawa, Tamae et Kubokawa (2017) améliorent tous aussi bien les estimations d'enquête directes.

Mots-clés : MPLSBE; méthode hiérarchique bayésienne; échantillonnage de Gibbs; modèle loglinéaire; erreur relative; échantillonnage; variance; petit domaine.

1. Introduction

L'estimation sur petits domaines est prisée et importante dans l'analyse de données d'enquête. On a largement recouru aux modèles pour établir des estimations fiables sur petits domaines. Dans la pratique, on se sert d'habitude de modèles au niveau de domaine chaque fois que des estimations d'enquête directes et des variables auxiliaires de niveau de domaine sont disponibles. Divers modèles en ce sens ont été proposés en vue d'accroître la précision des estimations d'enquête directes (voir Rao et Molina, 2015). Parmi les modèles au niveau de domaine, celui de Fay-Herriot (Fay et Herriot, 1979) est un modèle de base amplement utilisé dans l'estimation sur petits domaines. Il comporte deux composantes, à savoir un modèle d'échantillonnage pour les estimations d'enquête directes et un modèle de lien pour le paramètre du petit domaine choisi. Avec le modèle d'échantillonnage, nous posons qu'un estimateur d'enquête direct y_i est sans biais sous le plan de sondage pour le paramètre de petit domaine θ_i, de sorte que

$$y_i = \theta_i + e_i, \quad i = 1, \ldots, m, \tag{1.1}$$

où e_i est l'erreur d'échantillonnage liée à l'estimateur direct y_i et où m est le nombre de petits domaines. L'hypothèse habituelle est que les e_i sont des variables aléatoires normales indépendantes dont la moyenne est $\mathrm{E}(e_i) = 0$ et la variance d'échantillonnage, $\mathrm{Var}(e_i) = \sigma_i^2$. Avec le modèle de lien, nous posons que le paramètre de petit domaine θ_i est lié aux variables auxiliaires $x_i = (x_{i1}, \ldots, x_{ip})'$ par un modèle de régression linéaire qui est

$$\theta_i = x_i' \beta + v_i, \quad i = 1, \ldots, m, \tag{1.2}$$

1. Yong You, Centre de coopération internationale et d'innovation en méthodologie (CCIIM), Statistique Canada, Ottawa, Canada. Courriel : yong.you@statcan.gc.ca.

où $\beta = (\beta_1, \ldots, \beta_p)'$ est un vecteur $p \times 1$ de coefficients de régression et où les v_i sont des effets aléatoires propres au domaine que nous supposons indépendants et d'une distribution identique avec $\mathrm{E}(v_i) = 0$ et $\mathrm{Var}(v_i) = \sigma_v^2$. L'hypothèse de normalité est généralement formulée. Les effets aléatoires v_i et les erreurs d'échantillonnage e_i sont mutuellement indépendants. La variance de modèle σ_v^2 est inconnue et doit faire l'objet d'une estimation. La combinaison des modèles (1.1) et (1.2) nous donne un modèle mixte linéaire qui est

$$y_i = x_i' \beta + v_i + e_i, \quad i = 1, \ldots, m. \tag{1.3}$$

Le modèle (1.3) fait intervenir tant des erreurs aléatoires fondées sur le plan e_i que des effets aléatoires fondés sur le modèle v_i. Nous supposons ordinairement que, dans le cas du modèle de Fay-Herriot, la variance d'échantillonnage σ_i^2 est connue. C'est là une hypothèse très forte et, dans la pratique, nous disposerons généralement d'estimations directes sans biais des variances d'échantillonnage. S'il s'agit d'employer des estimations directes de variance d'échantillonnage, deux méthodes s'offrent dans la pratique, soit par lissage et par modélisation. Dans le premier cas, des estimations lissées de ces variances entrent dans le modèle de Fay-Herriot et sont ensuite traitées comme connues. Cette méthode exige des variables et des modèles externes comme une fonction généralisée de variance (FGV) et des effets du plan de sondage. You et Hidiroglou (2012) se sont attachés aux méthodes avec FGV et effets du plan pour un lissage des variances d'échantillonnage en présence de proportions. Nous emploierons ici un modèle FGV proposé par You et Hidiroglou (2012) en matière de lissage de variance d'échantillonnage.

Une solution autre que le lissage est le recours habituel dans la pratique à une modélisation de variance d'échantillonnage. Soit s_i^2 désignant l'estimateur direct de la variance d'échantillonnage σ_i^2. Nous considérons pour s_i^2 un modèle spécial sous la forme $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, où $d_i = n_i - 1$ et où n_i est la taille d'échantillon du i^{e} domaine. Rivest et Vandal (2002) et Wang et Fuller (2003) ont appliqué la méthode dite du meilleur prédicteur linéaire sans biais empirique (MPLSBE) pour dégager les estimations fondées sur le modèle. You et Chapman (2006) ont retenu une méthode hiérarchique bayésienne (HB) et combiné le modèle de variance d'échantillonnage $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ et le modèle de petit domaine (1.3). Le modèle ainsi obtenu tire sa puissance des estimations à la fois de petit domaine et de variance d'échantillonnage. Ainsi, la modélisation intégrée HB avec $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ a largement été utilisée dans la pratique, notamment par You (2008, 2016), Dass, Maiti, Ren et Sinha (2012), Sugasawa, Tamae et Kubokawa (2017), Ghosh, Myung et Moura (2018) et Hidiroglou, Beaumont et Yung (2019).

Nous regarderons ici les approches par lissage et par modélisation des variances d'échantillonnage. À la section 2, nous présentons la méthode MPLSBE en fonction des estimations tant lissées que directes des variances d'échantillonnage. À la section 3, nous présentons le modèle HB de Fay-Herriot et trois autres modèles HB basés sur la modélisation des variances d'échantillonnage. À la section 4, nous comparons les effets du lissage et de la modélisation de la variance d'échantillonnage par une analyse de données réelles et enfin, à la section 5, nous y allons de nos suggestions.

2. Modèle de Fay-Herriot avec le cadre MPLSBE

Si nous employons le modèle de Fay-Herriot (1.3) et posons que σ_i^2 et σ_v^2 sont connus dans ce modèle, nous obtenons le meilleur estimateur de prédiction linéaire sans biais (MPLSB) de θ_i comme $\tilde{\theta}_i = \gamma_i\, y_i + (1-\gamma_i)\, x_i' \tilde{\beta},$ où $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$ et $\tilde{\beta} = \left(\sum_{i=1}^m (\sigma_i^2 + \sigma_v^2)^{-1} x_i\, x_i'\right)^{-1} \left(\sum_{i=1}^m (\sigma_i^2 + \sigma_v^2)^{-1} x_i\, y_i\right).$ Pour estimer la composante σ_v^2 de la variance, nous devons supposer au départ que σ_i^2 est connu. Plusieurs méthodes s'offrent pour l'estimation de cette composante; nous allons employer la méthode du maximum de vraisemblance restreint (REML). Nous obtenons ainsi le MPLSBE du paramètre de petit domaine θ_i comme

$$\hat{\theta}_i = \hat{\gamma}_i y_i + (1-\hat{\gamma}_i) x_i' \hat{\beta}, \tag{2.1}$$

où $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \sigma_i^2)$ et où $\hat{\sigma}_v^2$ est l'estimateur REML. L'estimateur de l'erreur quadratique moyenne (EQM) de $\hat{\theta}_i$ nous est donné par $\text{eqm}(\hat{\theta}_i) = g_{1i} + g_{2i} + 2g_{3i}$, où $g_{1i} = \hat{\gamma}_i \sigma_i^2$ est le terme principal, où g_{2i} rend compte de la variabilité due à l'estimation du paramètre de régression β et où g_{3i} découle de l'estimation de la variance de modèle σ_v^2 (voir les détails dans Rao et Molina, 2015).

Nous pouvons utiliser l'estimation lissée ou l'estimation directe de σ_i^2 en (2.1). S'il s'agit de lisser la variance d'échantillonnage, nous appliquons un modèle de régression loglinéaire à la variance d'estimation directe s_i^2, comme le proposent You et Hidiroglou (2012). Le modèle de lissage se définit comme

$$\log(s_i^2) = \eta_0 + \eta_1 \log(n_i) + \varepsilon_i, \quad i = 1, \ldots, m, \tag{2.2}$$

où le terme d'erreur du modèle est $\varepsilon_i \sim N(0, \psi^2)$ et où ψ^2 est inconnu. Soit $\hat{\eta}_0$ et $\hat{\eta}_1$ les estimations par les moindres carrés ordinaires des coefficients de régression η_0 et η_1, et soit $\hat{\psi}^2$ la variance résiduelle estimée du modèle de régression loglinéaire (2.2). Nous pouvons obtenir un estimateur lissé de la variance d'échantillonnage σ_i^2 sous la forme suivante :

$$\tilde{\sigma}_i^2 = \exp\left(\hat{\eta}_0 + \hat{\eta}_1 \log(n_i)\right) \exp\left(\hat{\psi}^2 / 2\right).$$

Les variances d'échantillonnage en lissage $\tilde{\sigma}_i^2$ peuvent alors être employées dans l'estimateur MPLSBE (2.1) et le calcul de son EQM. C'est là une procédure courante (voir Rao et Molina, 2015).

Là où une estimation directe de la variance d'échantillonnage s_i^2 remplace sa valeur réelle σ_i^2 en (2.1), un terme supplémentaire rendant compte de l'incertitude de l'utilisation de s_i^2 est nécessaire dans l'estimateur EQM. Ce terme appelé g_{4i} est donné par $g_{4i} = 4(n_i - 1)^{-1} \hat{\sigma}_v^4 s_i^4 (\hat{\sigma}_v^2 + s_i^2)^{-3}$ (voir Rivest et Vandal (2002) et Rao et Molina (2015), page 150). En employant directement s_i^2 dans le cadre MPLSBE, on risque de surestimer la variance du modèle σ_v^2 (You, 2010; Rubin-Bleuer et You, 2016) et d'obtenir des estimations moins fidèles. Nous comparerons les estimations MPLSBE et HB en fonction des valeurs lissées et directes de variance d'échantillonnage à la section 4.

3. Modèle de Fay-Herriot avec le cadre HB en modélisation de la variance d'échantillonnage

Nous présenterons d'abord le modèle de Fay-Herriot dans un cadre hiérarchique bayésien (HB) et considérerons ensuite trois modèles de la variance d'échantillonnage. Le premier est de You et Chapman (2006) où un modèle à distribution gamma inverse est appliqué à la variance d'échantillonnage σ_i^2 avec des valeurs connues de paramètres vagues. Le deuxième modèle est de You (2016) avec un modèle loglinéaire à erreur aléatoire appliqué à σ_i^2. Le troisième est proposé par Sugasawa et coll. (2017) avec un modèle à distribution gamma inverse appliqué à σ_i^2, mais dans des réglages paramétriques différents.

Modèle HB 1 : modèle de Fay-Herriot dans un cadre HB, que nous appellerons FH-HB :

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind}\, N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind}\, N(x_i' \beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- distributions a priori uniformes pour les paramètres inconnus : $\pi(\beta) \propto 1, \;\; \pi(\sigma_v^2) \propto 1.$

À noter que, dans ce modèle, la variance d'échantillonnage σ_i^2 est censée être connue. À la place de σ_i^2, il y aura une estimation lissée de la variance d'échantillonnage $\tilde{\sigma}_i^2$ ou une estimation directe s_i^2.

Modèle HB 2 : modèle de You-Chapman (You et Chapman, 2006), que nous appellerons YCM :

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind}\, N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \,|\, \sigma_i^2 \sim \text{ind}\, \sigma_i^2 \chi_{d_i}^2, \;\; d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind}\, N(x_i' \beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i),$ où $a_i = 0{,}0001, \; b_i = 0{,}0001, \; i = 1, \ldots, m;$
- distributions a priori uniformes pour les paramètres inconnus : $\pi(\beta) \propto 1, \;\; \pi(\sigma_v^2) \propto 1.$

On peut trouver dans You et Chapman (2006) les distributions conditionnelles complètes pour la procédure d'échantillonnage de Gibbs avec les modèles FH-HB et YCM.

Modèle HB 3 : modèle loglinéaire de You (2016) pour les variances d'échantillonnage, que nous appellerons YLLM :

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind}\, N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \,|\, \sigma_i^2 \sim \text{ind}\, \sigma_i^2 \chi_{d_i}^2, \; d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind}\, N(x_i' \beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\log(\sigma_i^2) \sim N\left(\delta_1 + \delta_2 \log(n_i), \tau^2\right), \quad i = 1, \ldots, m;$

- distributions a priori uniformes pour les paramètres inconnus : $\pi(\beta) \propto 1$, $\pi(\delta_1, \delta_2) \propto 1$, $\pi(\sigma_v^2) \propto 1$, $\pi(\tau^2) \propto 1$.

Signalons que le modèle YLLM est un modèle loglinéaire appliqué à la variance d'échantillonnage σ_i^2 qui amplifie le modèle proposé par Souza, Moura et Migon (2009) en utilisant $\log(n_i)$ et en ajoutant un effet aléatoire à la partie « régression » du modèle. On trouvera en annexe les distributions conditionnelles complètes pour la procédure d'échantillonnage de Gibbs.

Modèle HB 4 : modèle de Sugasawa, Tamae et Kubokawa (2017) resserrant tant les moyennes que les variances, que nous appellerons STKM :

- $y_i \mid \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \mid \sigma_i^2 \sim \text{ind } \sigma_i^2 \chi_{d_i}^2, d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind } N(x_i'\beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i\gamma)$, où a_i et b_i sont des constantes connues, $a_i = O(1)$; $b_i = O(n_i^{-1})$;
- distributions a priori uniformes pour les paramètres inconnus : $\pi(\beta) \propto 1$, $\pi(\sigma_v^2) \propto 1$, $\pi(\gamma) \propto 1$.

Mentionnons que, dans STKM pour le modèle à distribution gamma inverse de σ_i^2, nous choisissons $a_i = 2$ et $b_i = n_i^{-1}$, ainsi que le proposent Sugasawa et coll. (2017). Ghosh et coll. (2018) ont opté pour le même paramétrage dans leur étude comparative des estimateurs HB. On peut trouver les distributions conditionnelles complètes pour STKM dans Sugasawa et coll. (2017).

À noter que, dans les modèles HB 2 à 4 qui précèdent, la modélisation khi-carré de la variance d'échantillonnage $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ fait appel à la normalité et à un échantillonnage aléatoire simple (Rivest et Vandal, 2002). Dans le cas des plans de sondage complexes, on pourrait devoir déterminer plus soigneusement les degrés de liberté d_i. Il n'y a pas de solide résultat théorique pour cette détermination (Dass et coll., 2012). Des formules peuvent se révéler utiles comme l'approximation sur erreurs de niveau d'unité non normales que proposent Wang et Fuller (2003) ou la règle par simulation de Maples, Bell et Huang (2009), mais on aura besoin dans l'un et l'autre cas de données de niveau des unités et d'une vaste étude de simulation. Une détermination prudente des degrés de liberté pourrait constituer une approximation raisonnablement utile. Ajoutons qu'une analyse bayésienne d'ajustement de modèle pourrait aussi servir à une détermination du modèle.

4. Application

Dans cette section, nous appliquons les modèles des sections 2 et 3 aux données canadiennes de l'Enquête sur la population active (EPA) et comparons les estimations MPLSBE et HB. L'EPA diffuse

chaque mois des estimations du taux de chômage pour de grands territoires comme le pays entier et les provinces aussi bien que pour de petites régions comme les régions métropolitaines de recensement (RMR) et les agglomérations de recensement (AR) de tout le Canada. Les estimations directes de l'EPA pour un certain nombre d'unités locales manquent de fiabilité, présentant des coefficients de variation (CV) très élevés à cause des petites tailles d'échantillon. Nous considérons que les estimateurs fondés sur le modèle viennent améliorer ces estimations directes. En guise d'illustration, nous appliquons le modèle de Fay-Herriot aux estimations du taux de chômage de mai 2016 au niveau des RMR-AR; nous comparons les estimations fondées sur le modèle et directes aux estimations du recensement pour mettre en contraste les effets du lissage et de la modélisation de la variance d'échantillonnage. Hidiroglou et coll. (2019) ont aussi comparé les estimations de l'EPA fondées sur le modèle aux estimations du recensement. Dans l'estimation du taux de chômage, la proportion mensuelle de prestataires locaux de l'assurance-emploi sert de variable auxiliaire dans le modèle. Dans une comparaison des estimations ponctuelles, nous calculons l'erreur relative absolue (ERA) des estimations directes et fondées sur le modèle relativement aux estimations du recensement pour chaque RMR ou AR :

$$\mathrm{ERA}_i = \left| \frac{\theta_i^{\mathrm{Census}} - \theta_i^{\mathrm{Est}}}{\theta_i^{\mathrm{Census}}} \right|,$$

où θ_i^{Est} est l'estimation directe ou par MPLSBE/HB et où $\theta_i^{\mathrm{Census}}$ est la valeur correspondante du recensement pour le taux de chômage. Il s'agit ensuite de prendre la moyenne des ERA sur l'ensemble des RMR et AR. Nous calculons les CV moyens des estimations directes et fondées sur le modèle. Nous privilégierons un modèle où les ERA et les CV sont moindres.

Nous appliquons les modèles d'abord aux 117 RMR et AR à taille d'échantillon ≥ 2, ensuite à 92 à taille d'échantillon ≥ 5 et enfin à 79 à taille d'échantillon ≥ 7. Le tableau 4.1 présente les ERA moyennes et les CV moyens correspondants (ceux-ci entre parenthèses). Dans ce tableau, les valeurs présentées de modélisation sont fonction du lissage ou de l'estimation directe de la variance d'échantillonnage.

En cas de lissage, les deux méthodes FH-MPLSBE et FH-HB améliorent nettement les estimations d'enquête directes avec les ERA et les CV sont bien moindres. Mentionnons en particulier que l'ERA et le CV sont respectivement les moindres avec FH-HB et FH-MPLSBE. Si la modélisation porte sur les 117 régions, les ERA et CV moyens sont respectivement de 0,263 et 0,329 pour l'estimateur direct de l'EPA, de 0,124 et 0,087 pour le lissage FH-MPLSBE et de 0,118 et 0,116 pour le lissage FH-HB. Le bon résultat des lissages FH-MPLSBE et FH-HB de la variance d'échantillonnage indique que la fonction généralisée de variance (2.2) est très utile dans ce cas et réussit à améliorer les estimations fondées sur le modèle.

En cas d'estimation directe de la variance d'échantillonnage, FH-MPLSBE et FH-HB ont les pires résultats de tous les modèles, et ils sont presque identiques dans ce scénario. Les trois autres modèles HB s'en tirent mieux que les modèles FH-MPLSBE et FH-HB en cas d'estimation directe. YLLM et STKM sont d'un meilleur rendement que YCM avec des ERA et des CV moindres. YLLM et STKM ont des

résultats à peu près identiques pour tous les groupes de RMR-AR; YLLM montre toujours une ERA légèrement inférieure, mais un CV légèrement supérieur à ceux de STKM. Sur l'ensemble des 117 régions, YLLM et STKM ont, par exemple, des ERA à 0,135 et 0,137 et des CV moyens à 0,123 et 0,122. YCM présente des valeurs correspondantes de 0,148 et 0,136 et FH-HB, de 0,171 et 0,221.

Tableau 4.1
Comparaison des erreurs relatives absolues (ERA) et des coefficients de variation (CV) en moyenne avec les valeurs CV entre parenthèses

RMR-AR	Estimation directe EPA	FH-MPLSBE lissage	FH-HB lissage	FH-MPLSBE estimation directe	FH-HB estimation directe	YCM estimation directe	YLLM estimation directe	STKM estimation directe
Moyenne sur les 117 RMR-AR	0,263	0,124	0,118	0,170	0,171	0,148	0,135	0,137
(taille d'échantillon ≥ 2)	(0,329)	(0,087)	(0,116)	(0,238)	(0,221)	(0,136)	(0,123)	(0,122)
Moyenne sur 92 RMR-AR	0,216	0,124	0,116	0,133	0,132	0,132	0,125	0,127
(taille d'échantillon ≥ 5)	(0,262)	(0,076)	(0,103)	(0,123)	(0,123)	(0,121)	(0,117)	(0,116)
Moyenne sur 79 RMR-AR	0,181	0,122	0,113	0,126	0,122	0,122	0,118	0,120
(taille d'échantillon ≥ 7)	(0,232)	(0,057)	(0,094)	(0,115)	(0,115)	(0,115)	(0,114)	(0,113)

Passons maintenant à une comparaison de modélisation bayésienne avec ordonnée prédictive conditionnelle (OPC) qui porte sur les quatre modèles HB pour l'estimation directe. Les OPC sont les valeurs observées de vraisemblance en fonction de la distribution prédictive sur validation croisée $f(y_i | y_{\text{obs}(i)})$. Nous calculons les valeurs OPC pour les divers points $y_{i,\text{obs}}$ de données d'observation. Une OPC plus élevée indique que $y_{i,\text{obs}}$ favorise le modèle et permet un meilleur ajustement. S'il s'agit de choisir un modèle, nous pouvons calculer le rapport des OPC entre un modèle A et un modèle B. Si ce rapport est supérieur à 1, $y_{i,\text{obs}}$ favorise le modèle A. Nous calculons les rapports OPC pour YCM/FH-HB, YLLM/FH-HB et STKM/FH-HB et comptons les fois que le rapport est supérieur à 1. Nous pouvons aussi tracer la courbe des valeurs OPC ou condenser celles-ci en prenant la moyenne des OPC estimées. Pour plus de détails sur l'ordonnée prédictive conditionnelle, voir, par exemple, Gilks, Richardson et Spiegelhalter (1996), page 153, You et Rao (2000) et Molina, Nandram et Rao (2014). Le tableau 4.2 présente les OPC moyennes et médianes pour les 117 RMR-AR, ainsi que le nombre de rapports OPC supérieurs à 1.

Tableau 4.2
Résumé des valeurs et des rapports OPC pour les 117 RMR-AR

	FH-HB estimation directe	YCM estimation directe	YLLM estimation directe	STKM estimation directe
OPC moyenne	0,1053	0,1222	0,1242	0,1238
OPC médiane	0,0976	0,1004	0,1045	0,1051
nombre de rapports OPC supérieurs à 1	-	72	78	76

Il ressort du tableau 4.2 que YCM, YLLM et STKM présentent des valeurs OPC supérieures à celles de FH-HB, indice que le modèle HB en modélisation de variance d'échantillonnage est préférable lorsque les estimations directes de cette variance sont employées. Il ressort également que YLLM et STKM sont préférables à YCM. Dans le cas des rapports OPC sur l'ensemble des 117 régions, 72 régions ou observations favorisent YCM, 78 YLLM et 76 STKM. Ainsi, plus d'observations font préférer YCM, YLLM et STKM à FH-HB; YLLM présente le plus de rapports OPC supérieurs à 1. La comparaison OPC s'accorde avec les résultats au tableau 4.1. Pour les autres méthodes de vérification et d'évaluation de modèles, voir Hidiroglou et coll. (2019).

5. Conclusion

Nous avons comparé les estimations fondées sur le modèle de Fay-Herriot dans les cas respectifs de lissage et de modélisation des variances d'échantillonnage. Comme dans Hidiroglou et coll. (2019), nos résultats indiquent que ce modèle peut grandement améliorer les estimations directes du taux de chômage de l'EPA, bien que des modèles plus complexes tels que les modèles non appariés ou les modèles de séries chronologiques puissent être employés (You (2008), par exemple). De tous les estimateurs, ce sont les modèles FH-MPLSBE et FH-HB en lissage qui donnent le meilleur résultat de réduction des ERA et des CV. En estimation directe des variances d'échantillonnage, ces mêmes modèles donnent les pires résultats. Dans la modélisation HB, YLLM et STKM s'en tirent tous deux très bien et mieux que YCM; YLLM est un peu meilleur que STKM dans l'étude. Nous proposons le modèle YLLM ou STKM en cas d'estimation directe de la variance d'échantillonnage. Autre possibilité, le lissage de variance devrait permettre dans le modèle de Fay-Herriot de surmonter la difficulté de modélisation des variances d'échantillonnage que nous évoquons à la section 3. Le lissage avec le modèle FGV en (2.2) à la section 2 peut produire un très bon résultat, comme le démontre notre étude.

Annexe

Distributions conditionnelles complètes et procédure d'échantillonnage pour YLLM

- $\left[\theta_i \,|\, y, \beta, \sigma_i^2, \sigma_v^2\right] \sim N\left(\gamma_i y_i + (1-\gamma_i)\, x_i'\beta, \gamma_i \sigma_i^2\right)$, où $\gamma_i = \sigma_v^2 / \sigma_v^2 + \sigma_i^2$, $i = 1, \ldots, m$;
- $\left[\, \beta \,|\, y, \theta, \sigma_i^2, \sigma_v^2 \,\right] \sim N_p\left(\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\left(\sum_{i=1}^m x_i \theta_i\right), \sigma_v^2\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\right)$;
- $\left[\, \sigma_v^2 \,|\, y, \theta, \beta, \sigma_i^2 \,\right] \sim \mathrm{IG}\left(\frac{m}{2} - 1, \frac{1}{2}\sum_{i=1}^m (\theta_i - x_i'\beta)^2\right)$;
- $\left[\, \sigma_i^2 \,|\, y, \theta, \beta, \sigma_v^2, \delta, \tau^2 \,\right] \propto f(\sigma_i^2) \cdot h(\sigma_i^2)$, où $f(\sigma_i^2)$ et $h(\sigma_i^2)$ sont $f(\sigma_i^2) \sim \mathrm{IG}\left(\frac{d_i+1}{2}, \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2}\right)$ et $h(\sigma_i^2) = \exp\left(-\frac{(\log(\sigma_i^2) - z_i'\delta)^2}{2\tau^2}\right)$;
- $\left[\, \delta \,|\, y, \theta, \beta, \sigma_i^2, \sigma_v^2, \tau^2 \,\right] \sim N_2\left(\left(\sum_{i=1}^m z_i z_i'\right)^{-1}\left(\sum_{i=1}^m z_i \log(\sigma_i^2)\right), \tau^2\left(\sum_{i=1}^m z_i z_i'\right)^{-1}\right)$;

- $\left[\tau^2 \,\middle|\, y, \theta, \beta, \sigma_i^2, \sigma_v^2, \delta\right] \sim \mathrm{IG}\left(\frac{m}{2}-1, \frac{1}{2}\sum_{i=1}^{m}\left(\log(\sigma_i^2) - z_i'\delta\right)^2\right)$.

Nous passons par l'étape de rejet de l'algorithme Metropolis-Hastings pour actualiser σ_i^2 :

(1) on tire σ_i^{2*} de $\mathrm{IG}\left(\frac{d_i+1}{2}, \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$;

(2) on calcule la probabilité d'acceptation $\alpha(\sigma_i^{2*}, \sigma_i^{2(k)}) = \min\left\{h(\sigma_i^{2*}) / h(\sigma_i^{2(k)}), 1\right\}$;

(3) on tire u de la distribution uniforme (0, 1); si $u < \alpha(\sigma_i^{2*}, \sigma_i^{2(k)})$, la valeur candidate σ_i^{2*} est acceptée et $\sigma_i^{2(k+1)} = \sigma_i^{2*}$; dans le cas contraire, σ_i^{2*} est rejeté et $\sigma_i^{2(k+1)} = \sigma_i^{2(k)}$.

Remerciements

J'aimerais remercier le rédacteur en chef, le rédacteur adjoint et un examinateur de leurs observations et leurs suggestions constructives ayant permis d'améliorer cet article.

Bibliographie

Dass, S.C., Maiti, T., Ren, H. et Sinha, S. (2012). Estimation des intervalles de confiance des paramètres de petit domaine avec rétrécissement des moyennes et des variances. *Techniques d'enquête*, 38, 2, 187-203. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11756-fra.pdf.

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

Ghosh, M., Myung, J. et Moura, F.A.S. (2018). Estimation bayésienne robuste sur petits domaines. *Techniques d'enquête*, 44, 1, 109-124. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018001/article/54959-fra.pdf.

Gilks, W.R., Richardson, S. et Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Hidiroglou, M.A., Beaumont, J.-F. et Yung, W. (2019). Élaboration d'un système d'estimation sur petits domaines à Statistique Canada. *Techniques d'enquête*, 45, 1, 107-133. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019001/article/00009-fra.pdf.

Maples, J., Bell, W. et Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the American community survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 5056-5067.

Molina, I, Nandram, B. et Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *Annals of Applied Statistics*, 8, 852-885.

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

Rivest, L.P., et Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, 10 au 13 juillet 2002, Ottawa, Canada.

Rubin-Bleuer, S., et You, Y. (2016). Comparaison de certains estimateurs de variance positifs pour le modèle d'estimation sur petits domaines Fay-Herriot. *Techniques d'enquête*, 42, 1, 69-93. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2016001/article/14542-fra.pdf.

Souza, D.F., Moura, F.A.S. et Migon, H.S. (2009). Prédiction de la population de petits domaines au moyen de modèles hiérarchiques. *Techniques d'enquête*, 35, 2, 203-214. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009002/article/11042-fra.pdf.

Sugasawa, S., Tamae, H. et Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150-167.

Wang, J., et Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2008). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquête*, 34, 1, 21-31. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008001/article/10614-fra.pdf.

You, Y. (2010). Small area estimation under the Fay-Herriot model using different model variance estimation methods and different input sampling variances. Document de travail de la Direction de la méthodologie, SRID-2010-003E, Statistique Canada, Ottawa, Canada.

You, Y. (2016). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models with applications. Document de travail de la Direction de la méthodologie, ICCSMD-2016-03-E, Statistique Canada, Ottawa, Canada.

You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 1, 107-114. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006001/article/9263-fra.pdf.

You, Y., and Hidiroglou, M. (2012). Sampling variance smoothing methods for small area proportion estimators. Document de travail de la Direction de la méthodologie, SRID-2012-08E, Statistique Canada, Ottawa, Canada; Présenté en tant que conférencier invité au Fields Institute Symposium on the Analysis of Survey Data and Small Area Estimation, Carleton University, Ottawa, 2012.

You, Y., et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26, 2, 197-206. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2000002/article/5537-fra.pdf.

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 401-412
Statistique Canada, nº 12-001-X au catalogue

Évaluer la couverture des intervalles de confiance en cas de non-réponse. Étude de cas sur la moyenne et les quantiles de revenu dans certaines municipalités de l'Enquête intercensitaire mexicaine de 2015

Omar De La Riva Torres, Gonzalo Pérez-de-la-Cruz et Guillermina Eslava-Gómez[1]

Résumé

L'article présente une étude comparative de trois méthodes de construction d'intervalles de confiance pour la moyenne et les quantiles à partir de données d'enquête en présence de non-réponse. Ces méthodes, à savoir la vraisemblance empirique, la linéarisation et la méthode de Woodruff (1952), ont été appliquées à des données sur le revenu tirées de l'Enquête intercensitaire mexicaine de 2015 et à des données simulées. Un modèle de propension à répondre a servi à ajuster les poids d'échantillonnage, et les performances empiriques des méthodes ont été évaluées en fonction de la couverture des intervalles de confiance au moyen d'études par simulations. Les méthodes de vraisemblance empirique et de linéarisation ont donné de bonnes performances pour la moyenne, sauf quand la variable d'intérêt avait des valeurs extrêmes. Pour les quantiles, la méthode de linéarisation s'est montrée peu performante; les méthodes de vraisemblance empirique et de Woodruff ont donné de meilleurs résultats, mais sans atteindre la couverture nominale quand la variable d'intérêt avait des valeurs à haute fréquence proches du quantile d'intérêt.

Mots-clés : Estimation de l'intervalle de confiance; vraisemblance empirique; linéarisation; données manquantes au hasard; non-réponse; échantillonnage à deux phases.

1. Introduction

L'Enquête intercensitaire mexicaine de 2015 (MIC2015) réalisée par l'Institut national de statistique, de géographie et d'informatique (*Instituto Nacional de Estadística, Geografía e Informática*, INEGI, 2015) a recueilli des renseignements à l'échelle nationale au moyen d'un plan de sondage probabiliste dans 1 643 municipalités et d'un recensement dans 814 municipalités. Dans la présente étude, nous utilisons les données du recensement correspondant à 441 municipalités de l'État de Oaxaca.

Nous nous concentrons sur la variable d'intérêt *revenu* qui présente un taux de non-réponse d'environ 22,5 %. Si l'on considère les répondants, la distribution du *revenu* présente une forte asymétrie, principalement en raison de la présence de valeurs extrêmes, et certaines valeurs ont une haute fréquence.

L'objectif de l'étude est d'évaluer le taux de couverture empirique des intervalles de confiance (IC) calculés par trois méthodes pour la moyenne de population et les quantiles de population, 0,1; 0,5 et 0,9, dans des données d'enquête en présence de non-réponse. L'échantillonnage à deux phases est utilisé avec un échantillon aléatoire sélectionné à la première phase, tandis que dans la seconde, l'échantillon est divisé en répondants et en non-répondants, compte tenu du profil de non-réponse de *revenu* dans les données du recensement. Un modèle de propension à répondre sert à ajuster les poids de non-réponse.

1. Omar De La Riva Torres, Département de mathématiques, Faculté des sciences, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. Courriel : odelariva@ciencias.unam.mx; Gonzalo Pérez-de-la-Cruz, Département de mathématiques, Faculté des sciences, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. Courriel : gonzalo.perez@ciencias.unam.mx; Guillermina Eslava-Gómez, Département de mathématiques, Faculté des sciences, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. Courriel : eslava@ciencias.unam.mx.

Pour la moyenne de la population, nous considérons l'estimateur de Hájek et deux méthodes de calcul des IC, soit la vraisemblance empirique (Berger, 2020) et la linéarisation (Särndal, Swenson et Wretman, 1992, sections 5.2 et 5.7). En ce qui concerne les quantiles de la population, nous tenons compte de l'estimateur ponctuel obtenu par interpolation de la fonction de répartition comme dans Woodruff (1952) et Graf et Tillé (2014), et de trois méthodes de calcul des IC : la probabilité empirique (Berger, 2020), la méthode de Woodruff (Woodruff, 1952) et la linéarisation (Deville, 1999). Ces méthodes sont décrites à la section 2; les résultats numériques sont présentés à la section 3 pour les données de la MIC2015 et à la section 4 pour certaines populations simulées. Des conclusions sont données à la section 5.

2. Trois méthodes d'estimation des intervalles de confiance

2.1 Estimation par échantillonnage à deux phases

Nous considérons une population finie $U = \{1, 2, \ldots, N\}$ et un échantillon probabiliste $s \subset U$ de taille fixe n, avec des probabilités d'inclusion du premier et du second ordre π_k et π_{kl}, $k, l \in U$, $k \neq l$. Soit y_k la k^{e} valeur de la variable d'intérêt y, et soit θ_0 un paramètre de population et $\hat{\theta}_0$ un estimateur de θ_0.

Nous supposons que la valeur y_k est disponible pour un sous-ensemble $r \subset s$ seulement. Soit ϕ_k la probabilité de réponse de l'unité k. Soit I_k une variable indicatrice de réponse telle que $I_k = 1$ pour $k \in r$ et $I_k = 0$ pour $k \in s \setminus r$. Nous supposons également qu'il y a un vecteur des variables auxiliaires $\mathbf{x}$ observées pour tous les $k \in s$. Nous formulons l'hypothèse de données manquantes au hasard (MAR pour *missing at random*) :

$$P(I_k = 1 \mid y_k, \mathbf{x}_k) = P(I_k = 1 \mid \mathbf{x}_k) = \phi_k \ \forall\, k \in U.$$

Les probabilités de réponse ϕ_k servent à ajuster les poids de sondage. Nous supposons que le plan de sondage et le mécanisme de réponse sont indépendants, comme dans Berger (2020). À partir de la théorie de l'échantillonnage à deux phases (Särndal et coll., 1992, section 9.3), les poids ajustés pour la non-réponse sont définis comme étant $\pi_k^{*-1} = 1/(\pi_k \phi_k)$ et les probabilités d'inclusion de second ordre comme étant $\pi_{kl}^* = \pi_{kl}\, \phi_k\, \phi_l, k, l \in U,\ k \neq l.$

L'intérêt réside dans l'estimation de la moyenne de population $\bar{Y} = \sum_{k \in U} y_k / N$ et du quantile de population donné par

$$Y_q = y_{(d-1)} + \frac{\left(y_{(d)} - y_{(d-1)}\right)\left[qN - N_{(d-1)}\right]}{N_{(d)} - N_{(d-1)}}, \tag{2.1}$$

où $y_{(i)}$ est la valeur de la i^{e} unité arrangée en ordre croissant, $d = \min\{l : qN < N_{(l)}, l = 1, \ldots, N\}$ et $N_{(l)} = \sum_{j \in U} I\left(y_j \leq y_{(l)}\right), l = 1, \ldots, N.$ On obtient la formule (2.1) en prenant en compte une interpolation linéaire par morceaux de la fonction de répartition par étapes $F(y) = \sum_{k \in U} I(y_k \leq y)/N,$ où $I\left(y_k \leq y\right) = 1$ quand $y_k \leq y.$

Ces paramètres de population sont estimés respectivement par

$$\hat{\bar{Y}} = \frac{\sum_{k \in r} y_k / \pi_k^*}{\sum_{k \in r} 1 / \pi_k^*}, \tag{2.2}$$

et

$$\hat{Y}_q = y_{(d-1)} + \frac{\left(y_{(d)} - y_{(d-1)}\right)\left(q\hat{N} - \hat{N}_{(d-1)}\right)}{\hat{N}_{(d)} - \hat{N}_{(d-1)}}, \tag{2.3}$$

où $d = \min\left\{l : q\hat{N} < \hat{N}_{(l)}, l = 1, \ldots, n_r\right\}$, $\hat{N} = \sum_{k \in r} 1 / \pi_k^*$, $\hat{N}_{(l)} = \sum_{k \in r} I(y_k \le y_{(l)}) / \pi_k^*$ et $n_r = \sum_{k \in s} I_k$.

Ces estimateurs et les IC décrits dans la sous-section suivante sont fondés sur l'hypothèse que les probabilités de réponse ϕ_k sont connues, contrairement à ce que l'on a dans Berger (2020) et Kim et Kim (2007). Cependant, nous utilisons $\hat{\phi}_k$ au lieu de ϕ_k dans les études par simulations, où

$$\hat{\phi}_k = \frac{\exp\left(\mathbf{x}_k^{\mathsf{T}} \hat{\beta}\right)}{1 + \exp\left(\mathbf{x}_k^{\mathsf{T}} \hat{\beta}\right)} \quad \forall \; k \in s,$$

avec $\hat{\beta}$ qu'on obtient en ajustant une régression logistique au moyen de s. Cela donne des estimateurs connus sous le nom d'estimateurs empiriques par double dilatation (Haziza et Beaumont, 2017).

2.2 Méthodes d'estimation des intervalles de confiance

2.2.1 Linéarisation

La méthode de linéarisation se fonde sur l'hypothèse selon laquelle la distribution de $\hat{\theta}_0$ est approximativement normale. Un IC pour θ_0 est

$$\left[\hat{\theta}_0 - z_{1-\alpha/2}[V(\hat{\theta}_0)]^{1/2},\; \hat{\theta}_0 + z_{1-\alpha/2}[V(\hat{\theta}_0)]^{1/2}\right], \tag{2.4}$$

où $1-\alpha$ est le niveau de confiance, aussi appelé couverture nominale; voir Särndal et coll. (1992, expression 5.2.3). En pratique, $V(\hat{\theta}_0)$ est estimé. Pour les estimateurs donnés par (2.2) et (2.3), un estimateur de la variance est donné par

$$\hat{V}(\hat{\theta}_0) = \sum_{k \in r} \sum_{l \in r} \frac{(\pi_{kl}^* - \pi_k^* \pi_l^*)}{\pi_{kl}^*} \frac{\hat{z}_k}{\pi_k^*} \frac{\hat{z}_l}{\pi_l^*}, \tag{2.5}$$

où $\hat{z}_k = \left(y_k - \hat{\bar{Y}}\right) / \hat{N}$ pour $\hat{\bar{Y}}$ (Särndal et coll., 1992, résultat 5.7.1) et $\hat{z}_k = -\left(I(y_k \le \hat{Y}_q) - q\right) / \left(f(\hat{Y}_q)\hat{N}\right)$ pour $\hat{Y}_q$ (Deville, 1999). La *fonction de densité* f a été obtenue de deux façons, a) au moyen d'un noyau gaussien comme dans Osier (2009) et b) au moyen de la technique du plus proche voisin comme dans Graf et Tillé (2014). Nous présentons les résultats relatifs à a), étant donné que la technique en b) a donné des résultats similaires.

Nous remarquons que l'utilisation de (2.5) avec $\hat{\phi}_k$ au lieu de ϕ_k peut entraîner une surestimation de la variance de l'estimateur empirique par double dilatation et des IC plus grands; voir l'expression (17) dans Kim et Kim (2007) associée aux estimateurs de $\bar{Y}$.

2.2.2 Méthode de la vraisemblance empirique

La méthode de la vraisemblance empirique suppose que θ_0 est la seule solution de l'équation d'estimation $G(\theta) = \sum_{k \in U} g_k(\theta) = 0$ pour une fonction donnée g_k. En particulier, nous utilisons :

i) $g_k(\theta) = y_k - \theta$ pour $\theta_0 = \bar{Y}$.

ii) $g_k(\theta) = \rho(y_k, \theta) - q$ pour $\theta_0 = Y_q$, où $\rho(y_k, \theta) = I(y_k < y_{(l)}) + I(y_k = y_{(l)})(\theta - y_{(l-1)}) / (y_{(l)} - y_{(l-1)})$ et $l = \min\{j : y_{(j)} > \theta\}$.

La fonction de log-vraisemblance empirique dans Berger et De La Riva Torres (2016) pour un plan de sondage à un degré sans stratification ni information auxiliaire est

$$\ell_{\max}(\theta) = \max_{m_k : k \in s} \left\{ \sum_{k \in s} \log(m_k) : m_k > 0,\ \sum_{k \in s} m_k g_k(\theta) = 0,\ \sum_{k \in s} m_k \pi_k = n \right\}, \tag{2.6}$$

où $\{m_k : k \in s\}$ satisfait aux contraintes de plan et de paramètre $\sum_{k \in s} m_k \pi_k = n$ et $\sum_{k \in s} m_k g_k(\theta) = 0$.

En présence de non-réponse, nous utilisons (2.6) en remplaçant $\sum_{k \in s} m_k g_k(\theta) = 0$ par $\sum_{k \in s} m_k I_k g_k(\theta) / \phi_k = 0$. Un IC pour θ_0 est donné par

$$\left[\min\left\{\theta : \hat{R}(\theta) \le \chi_1^2(\alpha)\right\},\ \max\left\{\theta : \hat{R}(\theta) \le \chi_1^2(\alpha)\right\} \right], \tag{2.7}$$

où $\hat{R}(\theta) = 2\left\{\ell_{\max}(\hat{\theta}_0) - \ell_{\max}(\theta)\right\}$ et $\chi_1^2(\alpha)$ est le $(1-\alpha)$-quantile de la distribution χ_1^2. L'estimateur $\hat{\theta}_0 := \operatorname{argmax}_{\{\theta\}} \ell_{\max}(\theta)$ correspond respectivement à (2.2) et (2.3).

Nous avons calculé (2.7) au moyen d'une méthode de recherche de racine, en calculant $\hat{R}(\theta)$ pour plusieurs valeurs de θ, où l'on obtient $\ell_{\max}(\theta)$ pour une valeur donnée θ par un algorithme de Newton-Raphson modifié comme dans Wu (2004).

2.2.3 Méthode de Woodruff pour les quantiles

La méthode de Woodruff (1952) est fondée sur la fonction de répartition estimée $\hat{F}(y)$. Pour un quantile Y_q, la variance de $\hat{F}(Y_q)$ peut être calculée approximativement au moyen de la méthode de linéarisation en séries de Taylor avec une variable linéarisée $z_k = \left(I(y_k \le Y_q) - q\right) / N$, tandis que la variance est estimée au moyen de (2.5) avec $\hat{z}_k = \left(I(y_k \le \hat{Y}_q) - q\right) / \hat{N}$. Si l'on suppose la normalité de $\hat{F}(Y_q)$ et que l'on utilise (2.4), il est possible de trouver un IC $[c_1, c_2]$ pour $F(Y_q)$, ce qui donne $\left[\hat{F}^{-1}(c_1),\ \hat{F}^{-1}(c_2)\right]$ pour Y_q.

3. Étude empirique fondée sur les données des populations MIC2015_Oax et MIC2015_Oax$_{\text{trunc}}$

Le recensement de la population examiné dans les travaux présentés ici comprenait 208 101 habitants avec des réponses complètes dans un vecteur $\mathbf{x}$ de six variables auxiliaires, soit l'*âge*, le *niveau de scolarité*, la *situation d'emploi*, le *genre*, la *langue autochtone* et l'*état matrimonial*. La variable d'intérêt y correspond au *revenu* mensuel.

Une régression logistique avec des interactions binaires a été ajustée aux 208 101 observations, avec une variable de réponse $I_k = 1$ si une valeur de *revenu* était donnée par la personne k, et $I_k = 0$ si elle ne l'était pas, et le vecteur $\mathbf{x}$ de six variables explicatives. Ce modèle a ensuite été appliqué à la population de 161 296 personnes, ce qui correspond à celles avec $I_k = 1$. Cela a donné un ensemble de valeurs de propension à répondre $\phi = (\phi_1, \ldots, \phi_{161\,296})$.

L'ensemble de 161 296 répondants, avec des propensions à répondre $\phi_1, \ldots, \phi_{161\,296}$, est appelé population MIC2015_Oax. La distribution du *revenu* dans cette population est très asymétrique, en partie en raison de la présence de valeurs très grandes. Si nous supprimons les 80 observations dont le *revenu* est supérieur ou égal à 50 000, nous obtenons une population tronquée appelée MIC2015_Oax$_{\text{trunc}}$, qui est également utilisée dans nos expériences.

La fonction de répartition par étapes du *revenu* n'a que 913 et 887 sauts respectivement dans chaque population, avec quelques sauts importants aux valeurs de *revenu* qui sont proches des quantiles d'intérêt. En particulier, $y =$ 2 571 représente 7,3 % de la distribution et est très proche du quantile $Y_{0,5}$; $y = 643$ (1,1 %) et $y = 857$ (4,2 %) sont proches de $Y_{0,1}$; tandis que $y =$ 6 429 (2,8 %) et $y =$ 7 000 (1,1 %) sont proches de $Y_{0,9}$.

3.1 Résultats numériques

Pour chaque population, le taux de couverture de l'IC de chaque méthode a été estimé comme suit :

1. Un échantillon aléatoire simple s de taille $n \in \{1\,000; 5\,000\}$ a été sélectionné.
2. Pour chaque unité $k \in s$ avec une propension à répondre ϕ_k, nous avons généré I_k à partir de $\text{Bernoulli}(\phi_k)$.
3. Deux cas ont été examinés : a) réponse complète et b) taux moyen de non-réponse de 22,5 %. Dans ce dernier cas, une régression logistique avec interactions binaires, avec I comme réponse et les six variables explicatives, a été sélectionnée par sélection ascendante au moyen du critère BIC. Les probabilités de réponse estimées $\hat{\phi}_k$ ont été obtenues avec le modèle sélectionné.
4. Les IC à 90 % ont été calculés au moyen de la linéarisation (Lin), de la vraisemblance empirique (VE) et de la méthode de Woodruff (W) pour les quantiles.
5. Les étapes 1 à 4 ont été répétées $M =$ 5 000 fois et le taux de couverture de chaque méthode et de chaque paramètre a été calculé comme étant la proportion des IC qui couvraient la valeur correspondante du paramètre.

Le tableau 3.1 montre les résultats pour n = 5 000. Le tableau 3.2 montre la valeur absolue du biais relatif en pourcentage, $\mathrm{BR} = 100\left(\bar{\hat{\theta}} - \theta_0\right)\Big/\theta_0$ avec $\bar{\hat{\theta}}_0 = \sum \hat{\theta}_{0i} \Big/ M$, et de la racine de l'erreur quadratique moyenne relative en pourcentage, $\mathrm{REQMR} = 100\left\{\sum (\hat{\theta}_{0i} - \theta_0)^2 \Big/ M\right\}^{1/2} \Big/ \theta_0$. La figure 3.1 présente la distribution des M = 5 000 estimations pour le scénario de non-réponse pour chaque paramètre; les distributions correspondantes pour le scénario de réponse complète sont qualitativement semblables. Les résultats pour n = 1 000 sont omis, car ils étaient semblables à ceux obtenus avec n = 5 000. À partir des tableaux 3.1 et 3.2 et de la figure 3.1, nous formulons les remarques suivantes :

a) Pour $\bar{Y}$, les méthodes Lin et VE ont des performances similaires : elles sont peu performantes (couverture aussi faible que 72,9 %) pour MIC2015_Oax, et donnent de bonnes performances pour MIC2015_Oax$_{\text{trunc}}$, atteignant le niveau nominal avec des taux d'erreur de queue et une longueur moyenne d'IC similaires. La figure 3.1 a) montre que la distribution de $\bar{\hat{Y}}$ est symétrique pour MIC2015_Oax$_{\text{trunc}}$ et hautement asymétrique pour MIC2015_Oax; cette asymétrie semble être liée aux 80 valeurs de *revenu* extrêmes non présentes dans MIC2015_Oax$_{\text{trunc}}$.

b) Pour les quantiles, la méthode Lin est peu performante avec la durée moyenne de l'IC la plus courte dans les deux scénarios, malgré la surestimation attendue de la variance dans le scénario de non-réponse. Cette méthode repose sur la normalité de $\hat{Y}_q$, mais les figures 3.1 b), c) et d) montrent que la distribution de $\hat{Y}_q$ est loin d'être symétrique et unimodale, avec des modes autour des valeurs de *revenu* à haute fréquence. En particulier pour $Y_{0,1}$, où le taux de couverture est aussi bas que 31,4 %, la distribution de $\hat{Y}_{0,1}$ est multimodale avec une proportion élevée de valeurs qui sont plus éloignées de $Y_{0,1}$ que la moitié de la longueur moyenne de l'IC. Les méthodes VE et W fonctionnent généralement bien, sauf pour $Y_{0,5}$ dans le scénario de réponse complète et pour $Y_{0,9}$ dans MIC2015_Oax. Les faibles couvertures semblent liées à la haute fréquence observée des deux valeurs de *revenu* 2 571 (7,3 %) et 6 429 (2,8 %). La première est très proche de $Y_{0,5}$ = 2 570 et certains IC de $Y_{0,5}$ sont trop étroits quand $\hat{Y}_{0,5}$ < 2 571. La seconde est plus éloignée de $Y_{0,9}$ dans MIC2015_Oax que dans MIC2015_Oax$_{\text{trunc}}$, réduisant la proportion d'IC couvrant $Y_{0,9}$ quand $\hat{Y}_{0,9} \approx$ 6 429, voir la figure 3.1 d), où $Y_{0,9}$ = 6 921 dans MIC2015_Oax et $Y_{0,9}$ = 6 856 dans MIC2015_Oax$_{\text{trunc}}$.

c) Le tableau 3.2 montre que le biais relatif (BR) est petit, inférieur à 3,3 %, pour tous les paramètres. Quand on applique un simple ajustement avec le pourcentage de non-réponse (non présenté ici), le BR est plus grand et toutes les méthodes ont de mauvaises performances. Ces résultats suggèrent que l'utilisation d'un modèle de propension aide à obtenir un BR comparable à celui du scénario de réponse complète. Pour $Y_{0,1}$, l'estimateur empirique par double dilatation est encore moins biaisé que celui associé au scénario de réponse complète; cependant, leur racine carrée de l'EQM relative (REQMR) est comparable et la plus grande parmi celles des paramètres d'intérêt, puisque la distribution des estimateurs est multimodale dans les deux scénarios, voir la figure 3.1 b).

Figure 3.1 Distribution des M = 5 000 estimations de $\bar{Y}$ dans a), et de $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$ dans b) à d) pour le scénario avec une non-réponse moyenne de 22,5 % et n = 5 000. Le graphique supérieur correspond à MIC2015_Oax et le graphique inférieur à MIC2015_Oax$_{trunc}$. Les pointillés indiquent les valeurs de population $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$.

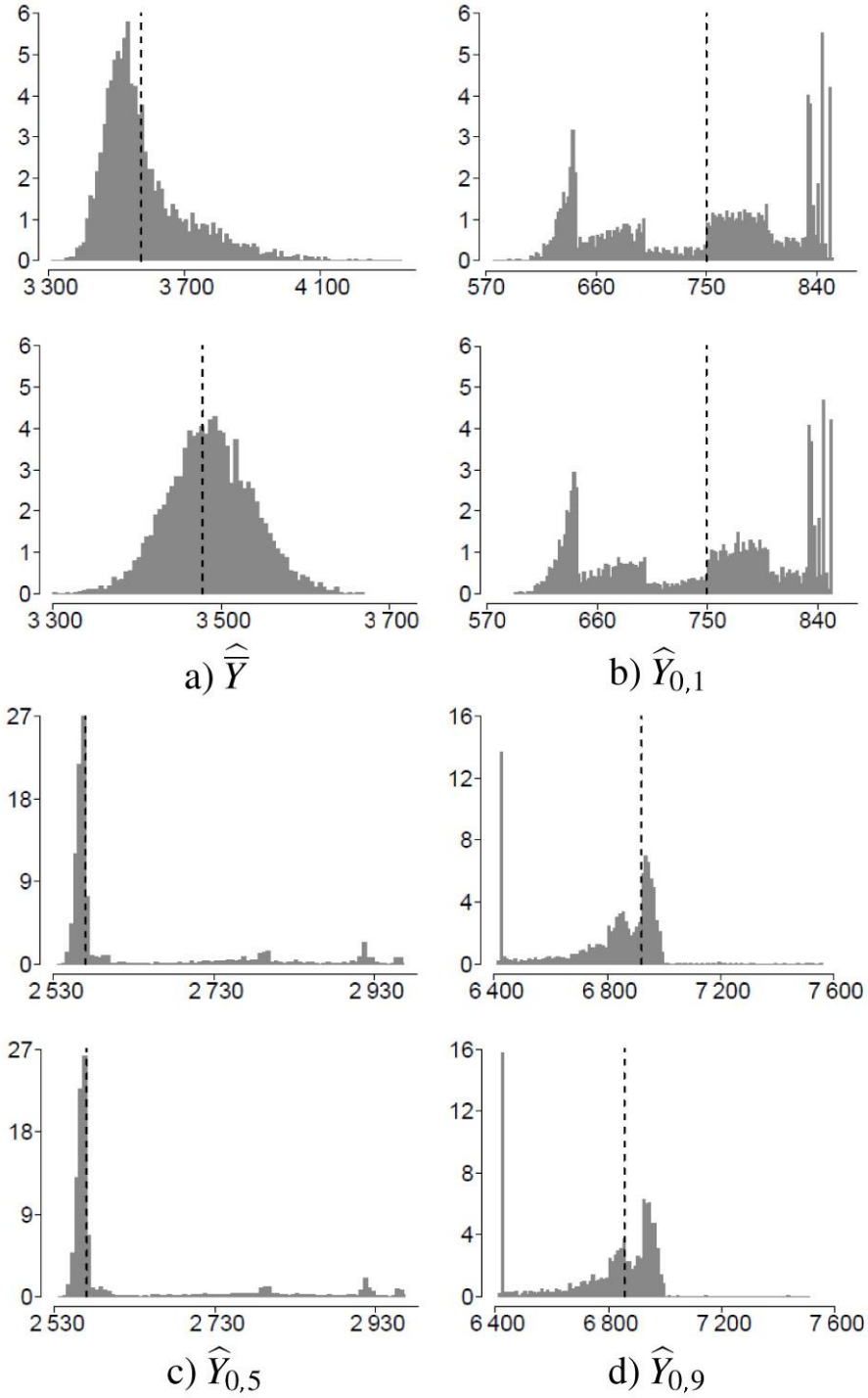


Tableau 3.1
Couvertures des IC à 90 % pour les paramètres $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$, pour y = *revenu*. Non-réponse moyenne de 22,5 % (*NR*) et réponse complète (*RC*)

Paramètre θ_0	Méthode	Couverture %		Taux d'err. de queue inférieurs en %		Taux d'err. de queue supérieurs en %		Longueur moyenne des IC	
		NR	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*
MIC2015_Oax									
$\bar{Y}$	VE	79,1*	72,9*	4,5	3,8*	16,4*	23,3*	370,8	335,4
	Lin	80,6*	73,8*	0,3*	0,1*	19,1*	26,1*	345,3	309,9
$Y_{0,1}$	VE	91,1*	90,5	6,0*	4,1*	2,9*	5,4	192,8	179,5
	W	90,6	89,8	6,2*	4,2*	3,2*	6,0*	191,2	177,1
	Lin	37,9*	35,1*	28,9*	19,2*	33,3*	45,7*	114,4	89,4
$Y_{0,5}$	VE	88,2*	82,3*	1,6*	0,7*	10,2*	17,1*	274,6	225,8
	W	88,0*	81,0*	1,7*	0,7*	10,3*	18,3*	274,4	224,1
	Lin	79,0*	88,0*	21,0*	12,0*	0,0*	0,0*	152,2	127,3
$Y_{0,9}$	VE	84,0*	83,0*	2,6*	2,7*	13,3*	14,3*	527,2	470,2
	W	86,1*	84,3*	2,5*	2,7*	11,4*	13,0*	533,8	474,2
	Lin	72,3*	73,5*	0,4*	0,1*	27,3*	26,4*	392,8	346,6
MIC2015_Oax$_{trunc}$									
$\bar{Y}$	VE	90,5	90,6	6,4*	4,4	3,0*	5,0	173,7	147,5
	Lin	90,8*	90,1	5,7*	4,2*	3,5*	5,6*	171,2	145,0
$Y_{0,1}$	VE	89,8	89,9	6,8*	4,0*	3,4*	6,1*	191,7	178,7
	W	89,1*	89,1*	7,0*	4,1*	4,0*	6,8*	190,0	176,2
	Lin	35,5*	31,4*	29,4*	21,0*	35,1*	47,6*	104,9	81,4
$Y_{0,5}$	VE	87,2*	80,4*	1,6*	0,9*	11,2*	18,7*	267,8	218,5
	W	87,1*	79,3*	1,6*	0,9*	11,3*	19,8*	267,7	216,7
	Lin	80,0*	87,4*	20,0*	12,6*	0,0*	0,0*	144,9	121,0
$Y_{0,9}$	VE	90,3	90,1	4,4	4,4*	5,3	5,5	521,3	470,8
	W	92,3*	91,9*	4,3*	4,3*	3,5*	3,8*	528,2	475,3
	Lin	75,6*	77,0*	0,1*	0,1*	24,3*	23,0*	411,7	365,5

* Couvertures et taux d'erreur de queue différant considérablement de 90 % et 5 % respectivement (Feller, 1968, page 182). p - valeur < 5 % MIC2015_Oax : $N = 161\,296$; $\rho = 0{,}08$; $\gamma = 89{,}9$; MIC2015_Oax$_{trunc}$: $N = 161\,216$; $\rho = 0{,}21$; $\gamma = 3{,}48$, où $\rho = \mathrm{corr}(y, \phi)$ et $\gamma = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3 \Big/ \left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\right]^{3/2}$.

Tableau 3.2
Biais relatif (BR) en pourcentage et racine carrée de l'erreur quadratique moyenne relative (REQMR) en pourcentage des estimateurs de $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$, basé sur 5 000 échantillons. Non-réponse moyenne de 22,5 % (*NR*) et réponse complète (*RC*)

Population	$\bar{Y}$				$Y_{0,1}$				$Y_{0,5}$				$Y_{0,9}$			
	|BR|		|REQMR|		|BR|		|REQMR|		|BR|		|REQMR|		|BR|		|REQMR|	
	NR	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*
MIC2015_Oax	0,30	0,01	3,8	3,2	0,58	3,13	10,4	9,9	1,96	0,93	4,8	3,4	1,79	1,68	3,3	3,0
MIC2015_Oax$_{trunc}$	0,27	0,02	1,5	1,3	0,71	3,23	10,5	10,0	1,87	0,96	4,8	3,5	1,05	0,95	3,0	2,8

4. Populations simulées

Afin de contrôler l'asymétrie de la distribution de y, la corrélation $\mathrm{corr}(y, \phi)$ et le pourcentage de non-réponse, nous avons simulé deux populations symétriques et deux populations asymétriques de taille $N =$ 50 000 avec une variable d'intérêt y et six variables auxiliaires $x_1, \ldots, x_6$, comme suit.

1. 50 000 échantillons aléatoires simples pour chacune des variables $x_1, \ldots, x_6$ ont été générés indépendamment d'une distribution $N(0,1)$.
2. Les probabilités de réponse ϕ_k ont été obtenues au moyen d'une régression logistique avec $\beta_1 = \cdots = \beta_6 = -0{,}3$ et β_0 choisies de sorte que la non-réponse moyenne soit égale à 24,8 %.
3. Deux paramètres ont été pris en compte pour la distribution de y :
   i) Symétrique. y_k a été généré à partir de $N(\mu, \sigma^2)$, avec $\mu = 1 + 2{,}16\ \mathrm{corr}(y, x)\sum_{j=1}^{6} x_{kj}$ et $\sigma^2 = 4{,}67*(1 - 6\mathrm{corr}^2(y, x))$, où $\mathrm{corr}(y, x) = \mathrm{corr}(y, x_j)$, j élément de $\{1, \ldots, 6\}$.
   ii) Asymétrique. $y_k = \exp(z_k)$, où z_k a été généré à partir de $N(\mu, \sigma^2)$, avec $\mu = \mathrm{corr}(z, x)\sum_{j=1}^{6} x_{kj}$ et $\sigma^2 = 1 - 6\mathrm{corr}^2(z, x)$, où $\mathrm{corr}(z, x) = \mathrm{corr}(z, x_j)$, $j \in \{1, \ldots, 6\}$.

   Les valeurs $\mathrm{corr}(y, x)$ et $\mathrm{corr}(z, x)$ ont été toutes les deux choisies de façon à ce que $\mathrm{corr}(y, \phi)$ soit à peu près égal à -0,2 ou -0,8.

4.1 Résultats numériques

Le taux de couverture des IC a été calculé comme à la section 3.1, avec $n = 500$ et au moyen d'une régression logistique sans interactions pour obtenir $\hat{\phi}_k$, $k \in s$.

Le tableau 4.1 présente les résultats pour les populations avec $\mathrm{corr}(y, \phi) = -0{,}8$. Le tableau 4.2 montre le $|\mathrm{BR}|$ et la $|\mathrm{REQMR}|$ des estimateurs. Les résultats numériques pour les populations avec $\mathrm{corr}(y, \phi) = -0{,}2$ sont omis, car ils étaient semblables à ceux obtenus pour les populations avec $\mathrm{corr}(y, \phi) = -0{,}8$. Nous faisons les observations suivantes :

a) Les méthodes VE et Lin ont des performances raisonnables similaires pour $\bar{Y}$, bien que les taux d'erreur de queue supérieurs soient plus grands que 5 % dans la population asymétrique.

b) Pour les quantiles, la méthode Lin a la couverture la plus faible et la plus élevée, soit respectivement 85,7 % et 97,9 %. Contrairement à la distribution de $\hat{Y}_q$ illustrée à la figure 3.1, la distribution de $\hat{Y}_q$ pour les populations simulées est symétrique et unimodale dans tous les cas. Les méthodes VE et W donnent de bons résultats, atteignant le niveau nominal dans tous les cas avec des taux d'erreur de queue et une longueur moyenne d'IC comparables.

c) L'ajustement de la pondération dans le scénario de non-réponse aide à obtenir un BR semblable à celui de la réponse complète.

d) Le taux de couverture de l'IC pour chaque méthode est plus élevé dans le scénario de non-réponse que dans le scénario de réponse complète, et dans certains cas, il est également plus élevé que le niveau nominal. Cela pourrait être lié au fait d'avoir traité la valeur $\hat{\phi}_k$ comme étant fixe dans le scénario de non-réponse.

Tableau 4.1
Populations simulées. Couvertures des IC à 90 % pour $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$, pour y avec $\mathrm{corr}(y, \phi) = -0{,}8$. Non-réponse moyenne de 24,8 % (*NR*) et réponse complète (*RC*)

Paramètre θ_0	Méthode	Couverture %		Taux d'err. de queue inférieurs en %		Taux d'err. de queue supérieurs en %		Longueur moyenne des IC	
		NR	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*
Asymétrique									
$\bar{Y}$	VE	90,9*	88,6*	2,2*	5,1	6,9*	6,3	0,50	0,32
	Lin	90,1	88,6*	0,5*	3,2*	9,4*	8,2*	0,48	0,31
$Y_{0,1}$	VE	90,6	89,5	4,0*	4,6	5,4	5,9	0,07	0,07
	W	90,3	89,5	3,6*	4,1*	6,2*	6,4*	0,07	0,07
	Lin	97,9*	97,4*	1,2*	1,5*	1,0*	1,2*	0,10	0,10
$Y_{0,5}$	VE	93,6*	90,5	3,0*	4,4*	3,4*	5,1	0,22	0,19
	W	93,5*	90,4	2,9*	4,4	3,6*	5,1	0,22	0,19
	Lin	92,5*	88,9*	2,9*	5,0	4,6	6,2*	0,21	0,18
$Y_{0,9}$	VE	92,7*	90,3	2,6*	4,1*	4,7	5,7*	1,35	0,91
	W	93,0*	90,4	2,7*	4,6	4,3*	5,0	1,36	0,92
	Lin	87,4*	85,7*	2,6*	4,1*	10,1*	10,2*	1,23	0,87
Symétrique									
$\bar{Y}$	VE	93,6*	90,2	3,3*	5,0	3,1*	4,8	0,40	0,32
	Lin	93,5*	89,9	3,1*	5,2	3,4*	4,9	0,39	0,32
$Y_{0,1}$	VE	91,2*	90,9*	3,6*	3,9*	5,2	5,2	0,59	0,55
	W	91,0*	90,6	3,2*	3,6*	5,8*	5,8*	0,59	0,56
	Lin	88,6*	88,2*	5,8*	5,9*	5,7*	6,0*	0,57	0,53
$Y_{0,5}$	VE	92,8*	90,1	3,3*	4,6	3,9*	5,3	0,46	0,39
	W	92,9*	90,1	3,1*	4,6	4,0*	5,3	0,46	0,39
	Lin	92,4*	90,2	3,4*	4,7	4,2*	5,2	0,46	0,39
$Y_{0,9}$	VE	92,9*	90,4	2,2*	3,6*	4,9	6,0*	0,76	0,54
	W	93,3*	90,3	2,2*	4,2*	4,5	5,4	0,77	0,54
	Lin	90,3	89,1*	2,1*	3,2*	7,6*	7,7*	0,74	0,53

* Couvertures et taux d'erreur de queue différant considérablement de 90 % et 5 % respectivement (Feller, 1968, page 182). p - valeur < 5 % Symétrique : $\gamma =$ 0,02; Asymétrique : $\gamma =$ 6,2; où $\gamma = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3 \Big/ \left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\right]^{3/2}$.

Tableau 4.2
Biais relatif (BR) en pourcentage et racine carrée de l'erreur quadratique moyenne relative (REQMR) en pourcentage des estimateurs de $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$, basé sur 5 000 échantillons. Non-réponse moyenne de 24,8 % (*NR*) et réponse complète (*RC*)

Population	$\bar{Y}$				$Y_{0,1}$				$Y_{0,5}$				$Y_{0,9}$			
	|BR|		|REQMR|		|BR|		|REQMR|		|BR|		|REQMR|		|BR|		|REQMR|	
	NR	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*
Asymétrique	0,11	0,13	9,0	5,9	0,56	0,53	7,8	7,5	0,00	0,07	6,0	5,7	0,53	0,43	9,7	7,6
Symétrique	0,12	0,09	10,3	9,5	0,98	0,91	10,2	9,7	0,33	0,35	12,7	11,8	0,43	0,34	5,5	4,3

5. Conclusions

Compte tenu de la distribution de y (*revenu*), on a observé que les mauvaises performances d'une méthode dans le scénario de réponse complète correspondaient généralement à de mauvaises performances dans le scénario de non-réponse, même si dans plusieurs cas, le taux de couverture était plus élevé dans ce dernier. Cela suggère que le fait d'avoir traité la variable $\hat{\phi}_k$ comme étant fixe a eu peu

d'effet sur les performances des méthodes par rapport à l'effet des caractéristiques de la distribution de y. Les valeurs extrêmes étaient liées à une faible couverture des IC pour la moyenne des méthodes de vraisemblance empirique et de linéarisation. La présence de valeurs à haute fréquence près d'un quantile d'intérêt avait aussi un effet sur la couverture de ses IC. Cela pourrait s'expliquer par le comportement de la fonction de répartition par étapes, où les sauts dans $F(y)$ et $\hat{F}(y)$ doivent habituellement être de petite taille pour que la méthode de Woodruff donne de bonnes performances (Lohr, 2010, page 390). En général, la méthode de linéarisation a donné des performances médiocres pour les quantiles, tandis que celles de la vraisemblance empirique et de la méthode de Woodruff étaient similaires et meilleures; ce comportement a aussi été observé dans Berger et De La Riva Torres (2016). La méthode Woodruff est simple et facile à mettre en œuvre, mais la méthode de la vraisemblance empirique présente l'avantage d'être utilisable pour des paramètres autres que les quantiles.

Remerciements

Omar De La Riva Torres était soutenu par le Programme de bourses d'études post doctorales de l'Université nationale autonome du Mexique.

Bibliographie

Berger, Y.G. (2020). An empirical likelihood approach under cluster sampling with missing observations. *Annals of the Institute of Statistical Mathematics*, 72, 91-121.

Berger, Y.G., et De La Riva Torres, O. (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 2, 319-341.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 2, 219-230. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1999002/article/4882-fra.pdf.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications: Volume I*. New York: John Wiley & Sons, Inc.

Graf, E., et Tillé, Y. (2014). Estimation de variance par linéarisation pour des indices de pauvreté et d'exclusion sociale. *Techniques d'enquête*, 40, 1, 69-88. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014001/article/14000-fra.pdf.

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

INEGI (2015). *Encuesta Intercensal 2015*. Instituto Nacional de Estadística, Geografía e Informática. https://www.inegi.org.mx/programas/intercensal/2015/.

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

Lohr, S.L. (2010). *Sampling: Design and Analysis.* Brooks/Cole.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 3, 167-195.

Särndal, C.-E., Swenson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2021.

- P.G. Andersson, *Stockholm University*
- S. Arima, *Uniersità del Salento*
- B. Baffour, *Australian National University*
- G. Barcarolli, *Independent consultant*
- E. Ben-David, *U.S. Census Bureau*
- J. Billiet, *Ku Leuven*
- C. Bocci, *Statistique Canada*
- D. Bonnery, *University of Cambridge*
- H.J. Boonstra, *Statistics Netherlands*
- C. Boulet, *Statistique Canada*
- J.M. Brick, *Westat Inc.*
- S. Cai, *Carleton University*
- P.J. Cantwell, *U.S. Census Bureau*
- I. Carrillo, *Statistique Canada*
- L. Castro-Martin, *Roma Tre University*
- S. Chatterjee, *University of Minnesota*
- G. Chauvet, *École nationale de la statistique et de l'analyse de l'information*
- Q. Chen, *Columbia University*
- S. Chen, *University of Oklahoma Health Sciences Center*
- J. Chipperfield, *Australian Bureau of Statistics*
- M. Cohen, *American Institutes for Research*
- D. da Silva, *Federal University of Rio Grande do Norte*
- S. Das, *Maastricht University*
- G. Datta, *University of Georgia*
- T. De Waal, *Centraal Bureau voor de Statistiek*
- J. Dever, *RTI International*
- P. Duchesne, *Université de Montréal*
- L. Dumitrescu, *Victoria University of Wellington*
- S. Eckman, *Research Triangle Institute*
- J.L. Eltinge, *U.S. Bureau of Labor Statistics*
- A. Erciulescu, *Westat Inc.*
- V. Estevao, *Statistique Canada*
- B. Fay, *Westat Inc.*
- W.A. Fuller, *Iowa State University*
- C. Girard, *Statistique Canada*
- D. Haziza, *University of Ottawa*
- S. Heeringa, *University of Michigan*
- M.A. Hidiroglou, *Statistique Canada*
- M. Hirose, *Kyushu University*
- N. Hoshino, *Kanazawa University - Kakuma Campus*
- B. Hulliger, *University of Applied and Arts Sciences Northwestern, Switzerland*
- K. Irimata, *U.S. Census Bureau*
- R. Jauslin, *Université de Neuchâtel*
- D. Judkins, *ABT Associates Inc Bethesda*
- A. Junior, *Senac, prospecting and evaluation management*
- A. Kaplan, *Colorado State University*
- A. Keller, *U.S. Census Bureau*
- C. Kern, *University of Mannheim*
- J.K. Kim, *Iowa State University*
- H. Kim, *University of Cincinnati*
- S. Kolenikov, *Abt Associates*
- P.S. Kott, *RTI International*
- P. Lahiri, *University of Maryland*
- T. Laitila, *Örebro University*
- D. Lee, *University of Alabama*
- S. Lee, *University of Michigan*
- É. Lesage, *L'Institut national de la statistique et des études économiques*
- D. Liao, *Research Triangle Institute*
- B. Liseo, *University of Rome La Sapienza, MEMOTEF*
- Y. Liu, *East China Normal University*
- G. Loosveldt, *KU Leuven*
- J. Maples, *U.S. Census Bureau*
- D. Marella, *Roma Tre University*
- K. McConville, *Reed College*
- T. Merkouris, *Athens University of Economics and Business Statistics*
- I. Molina, *Universidad Carlos III de Madrid*
- B. Nandram, *Worcester Polytechnic Institute*
- D. Nedyalkova, *Office fédéral de la statistique*
- A. Norberg, *Statistics Sweden*
- J. Opsomer, *Westat Inc*
- D. Pfeffermann, *University of Southampton*
- M. Ranalli, *Università degli Studi di Perugia*
- J.N.K. Rao, *Carleton University*
- J. Reiter, *Duke University*
- L.-P. Rivest, *Université Laval*
- M. Scanu, *Italian National Institute of Statistics*
- F.J. Scheuren, *National Opinion Research Center*
- M. Schonlau, *University of Waterloo*
- P.L.d.N. Silva, *Escola Nacional de Ciências Estatísticas*
- P. Smith, *University of Southampton*
- D. Steel, *University of Wollongong*
- L. Stokes, *Southern Methodist University*
- S. Sugasawa, *University of Tokyo*
- M. Templ, *Zurich University of Applied Sciences*
- N. Thuan, *Oregon Health & Science University*
- M. Torabi, *University of Manitoba*
- D. Toth, *U.S. Bureau of Labor Statistics*
- A.-A. Vallée, *Université Laval*
- S. van Buuren, *Universiteit Utrecht*
- A. Van Delden, *Statistics Netherlands*
- J. van den Brakel, *Statistics Netherlands*
- A. Veiga, *Instituto Brasileiro de Geografia e Estatistica*
- F. Verret, *Statistique Canada*
- J. Wagner, *University of Michigan*
- Z. Wang, *Wilfrid Laurier University*
- N. Wlömert, *Vienna University of Economics and Business*
- C. Wu, *University of Waterloo*
- C. Xu, *The University of Oklahoma Health Sciences Center, Biostatistics and Epidemiology*
- D. Yang, *U.S. Bureau of Labor Statistics*
- Y. You, *Statistique Canada*
- C. Yu, *Iowa State University*
- W. Yung, *Statistique Canada*
- L.-C. Zhang, *University of Southampton*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2021 : Céline Ethier et Patrick O'Leary du Centre de collaboration internationale et d'innovation en méthodologie; l'équipe de la Division de la diffusion, en particulier Chantal Chalifoux, Sophie Pilon, Eric Chénier, Kathy Charbonneau, Beth Chepita, Travis Robinson, Brianna Bucci et Karo-Lynn Audy et de même que nos partenaires de la Division des communications.

ANNONCES

Demande de candidatures pour le prix Waksberg 2023

La revue *Techniques d'enquête* a mis sur pied une série annuelle de communications sollicitées en l'honneur de Joseph Waksberg, en reconnaissance des contributions exceptionnelles qu'il a faites à la statistique et méthodologie d'enquête. Chaque année, un éminent statisticien d'enquête est choisi pour rédiger un article où il examine l'évolution et l'état actuel d'un thème important du domaine de la statistique et méthodologie d'enquête. L'article reflète le mélange de théorie et de pratique caractéristique des travaux de Joseph Waksberg.

Le lauréat du prix Waksberg recevra une prime en argent et présentera la communication sollicitée Waksberg 2023 au Symposium de Statistique Canada prévu à l'automne 2023. L'article paraîtra dans un numéro futur de *Techniques d'enquête* (prévu pour décembre 2023).

L'auteur de l'article Waksberg 2023 sera choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*. **Les candidatures doivent être envoyées par courriel avant le 28 février 2022 au président du comité, Jack Gambino, (jack.gambino@gmail.com). Les candidatures doivent inclure un Curriculum Vitae et une lettre de candidature.** Les candidatures resteront actives pour une période de 5 ans.

Membres du comité de sélection de l'article Waksberg (2021-2022)

Jack Gambino, *Statistique Canada* (Président)
Giovanna Ranalli, *University of Perugia*
Denise Silva, *Brazilian Institute of Geography and Statistics*
Kristen Olson, *University of Nebraska-Lincoln*

Présidents précédents :

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)
Elizabeth A. Martin (2010 - 2011)
Mary E. Thompson (2011 - 2012)
Steve Heeringa (2012 - 2013)
Cynthia Clark (2013 - 2014)
Louis-Paul Rivest (2014 - 2015)
Tommy Wright (2015 - 2016)
Kirk Wolter (2016 - 2017)
Danny Pfeffermann (2017 - 2018)
Michael A. Hidiroglou (2018 - 2019)
Robert E. Fay (2019 - 2020)
Jean Opsomer (2020 - 2021)

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 37, No. 2, June 2021

Special Issue on New Techniques and Technologies for Statistics

Preface
Francesca Di Iorio, Emanuele Baldacci, Dario Buono, Luca di Gennaro Splendore, Duncan Elliott, Rebecca Killick, Tiziana Laureti, Monica Pratesi and Natalie Shlomo 257

A structural Equation Model for Measuring Relative Development of Hungarian Counties in the Years 1994-2016
Klaudia Máténé Bella and Ildikó Ritzlné Kazimir.....261

Measuring and Communicating the Uncertainty in Official Economic Statistics
Gian Luigi Mazzi, James Mitchell and Florabela Carausu 289

The Evolution of the Italian Framework to Measure Well-Being
Fabio Bacchini, Barbara Baldazzi, Rita De Carli, Lorenzo Di Biagio, Miria Savioli, Maria Pia Sorvillo and Alessandra Tinto.....317

Improving Time Use Measurement with Personal Big Data Collection – The Experience of the European Big Data Hackathon 2019
Mattia Zeni, Ivano Bison, Fernando Reis, Britta Gauckler and Fausto Giunchiglia 341

A Diagnostic for Seasonality Based Upon Polynomial Roots of ARMA Models
Tucker McElroy 367

Applying Machine Learning for Automatic Product Categorization
Andrea Roberson.....395

A Product Match Adjusted R Squared Method for Defining Products with Transaction Data
Antonio G. Chessa 411

Variance Estimation after Mass Imputation Based on Combined Administrative and Survey Data
Sander Scholtus and Jacco Daalmans.....433

Assessing and Adjusting Bias Due to Mixed-Mode in Aspect of Daily Life Survey
Claudia de Vitiis, Alessio Guandalini Francesca Inglese and Marco D. Terribili.....461

Measuring the Accuracy of Aggregates Computed from a Statistical Register
Giorgio Alleva, Piero Demetrio Falorsi, Francesca Petrarca and Paolo Righi 481

A Hybrid Technique for the Multiple Imputation of Survey Data
Humera Razzak and Christian Heumann 505

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 37, No. 3, September 2021

Special Issue on Population Statistics for the 21st Century

Preface
Jakub Bijak, Johan Bryant, Elżbieta Gołata and Steve Smallwood....533

Letter to the Editors
Giampaolo Lanzieri....543

Fertility Projections in a European Context: A Survey of Current Practices among Statistical Agencies
Rebecca Folkman Gleditsch, Astri Syse and Michael J. Thomas....547

Modelling Frontier Mortality Using Bayesian Generalised Additive Models
Jason Hilton, Erengul Dodd, Jonathan J. Forster and Peter W.F. Smith....569

Probabilistic Projection of Subnational Life Expectancy
Hana Ševčíková and Adrian E. Raftery....591

Spatio-Temporal Patterns in Portuguese Regional Fertility Rates: A Bayesian Approach for Spatial Clustering of Curves
Zhen Zhang, Arnab Bhattacharjee, João Marques and Tapabrata Maiti....611

Optimal Sampling for the Population Coverage Survey of the New Italian Register Based Census
Paolo Righi, Piero Demetrio Falorsi, Stefano Daddi, Epifania Fiorello, Pierpaolo Massoli and Marco Dionisio Terribili....655

Latent Class Analysis for Estimating an Unknown Population Size – with Application to Censuses
Bernard Baffour, James J. Brown and Peter W.F. Smith....673

A General Framework for Multiple-Recapture Estimation that Incorporates Linkage Error Correction
Daan Zult, Peter-Paul de Wolf, Bart F. M. Bakker and Peter van der Heijden....699

Exploratory Assessment of the Census of Pakistan Using Demographic Analysis
Asif Wazir and Anne Goujon....719

A Simulation Study of Diagnostics for Selection Bias
Philip S. Boonstra, Roderick J.A. Little, Brady T. West, Rebecca R. Andridge and Fernanda Alvarado-Leiton....751

Fay-Herriot Model-Based Prediction Alternatives for Estimating Households with Emigrated Members
Jairo Fúquene-Patiño, César Cristancho, Mariana Ospina and Domingo Morales Gonzalez....771

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 49, No. 1, March/mars 2021

Special Issue: Special Issue on Neuroimaging Data Analysis

Issue Information

Issue Information 1

Editorial

Special Issue on Neuroimaging data analysis: Guest Editors' Introduction
Farouk Nathoo, Linglong Kong, Grace Y. Yi 4

Research Article

Statistical disease mapping for heterogeneous neuroimaging studies
Rongjie Liu, Hongtu Zhu, for the Alzheimer's Disease Neuroimaging Initiative 10

Discussions

Discussion of "Statistical disease mapping for heterogeneous neuroimaging studies"
Ben Wu, Jian Kang 35

Discussion of "Statistical disease mapping for heterogeneous neuroimaging studies"
Matthew Pietrosanu, Bei Jiang 39

Rejoinder

Rejoinder: "Statistical disease mapping for heterogeneous neuroimaging studies"
Rongjie Liu, Hongtu Zhu 43

Research Articles

A Bayesian latent spatial model for mapping the cortical signature of progression to Alzheimer's disease
Ning Dai, Hakmook Kang, Galin L. Jones, Mark B. Fiecas,
for the Alzheimer's Disease Neuroimaging Initiative 46

On statistical tests of functional connectome fingerprinting
Zeyi Wang, Haris I. Sair, Ciprian Crainiceanu, Martin Lindquist, Bennett A. Landman,
Susan Resnick, Joshua T. Vogelstein, Brian Caffo 63

A grouped beta process model for multivariate resting-state EEG microstate analysis on twins
Brian Hart, Stephen Malone, Mark Fiecas 89

Scalable spatio-temporal Bayesian analysis of high-dimensional electroencephalography data
Shariq Mohammed, Dipak K. Dey 107

A spatial Bayesian semiparametric mixture model for positive definite matrices with applications
in diffusion tensor imaging
Zhou Lan, Brian J. Reich, Dipankar Bandyopadhyay 129

Multivariate functional response low-rank regression with an application to brain imaging data
Xiucai Ding, Dengdeng Yu, Zhengwu Zhang, Dehan Kong 150

A structured brain-wide and genome-wide association study using ADNI PET images
Yanming Li, Bin Nan, Ji Zhu, for the Alzheimer's Disease Neuroimaging Initiative 182

Connectivity-informed adaptive regularization for generalized outcomes
Damian Brzyski, Marta Karas, Beau M Ances, Mario Dzemidzic, Joaquín Goñi,
Timothy W Randolph, Jaroslaw Harezlak 203

Acknowledgement

Acknowledgement of Referees' Services/Remerciements aux lecteurs critiques 228

CONTENTS TABLE DES MATIÈRES

Volume 49, No. 2, June/juin 2021

Issue Information

Issue Information 235

Research Articles

On-line partitioning of the sample space in the regional adaptive algorithm
Nicolas Grenon-Godbout, Mylène Bédard 238

Regression modelling with the tilted beta distribution: A Bayesian approach
Eugene D. Hahn 262

On variable ordination of Cholesky-based estimation for a sparse covariance matrix
Xiaoning Kang, Xinwei Deng 283

Estimation of nonparametric additive models with high order spatial autoregressive errors
Guoying Xu, Yang Bai 311

Empirical and conditional likelihoods for two-phase studies
Menglu Che, Jerald F. Lawless, Peisong Han 344

Robust estimation of mean squared prediction error in small-area estimation
Ping Wu, Jiming Jiang 362

Variable selection for proportional hazards models with high-dimensional covariates subject to measurement error
Baojiang Chen, Ao Yuan, Grace Y. Yi 397

Weighted Bayesian bootstrap for scalable posterior distributions
Michael A. Newton, Nicholas G. Polson, Jianeng Xu 421

An efficient algorithm for Elastic I-optimal design of generalized linear models
Yiou Li, Xinwei Deng 438

Efficient nonparametric estimation for skewed distributions
Cyril Favre-Martinoz, David Haziza, Jean-François Beaumont 471

Local structure graph models with higher-order dependence
Emily M. Casleton, Daniel J. Nordman, Mark S. Kaiser 497

Minimum L^q-distance estimators for non-normalized parametric models
Steffen Betsch, Bruno Ebner, Bernhard Klar 514

Sure joint feature screening in nonparametric transformation model for right censored data
Yi Liu, Jinfeng Xu, Gang Li 549

A broad class of zero-or-one inflated regression models for rates and proportions
Francisco F. Queiroz, Artur J. Lemonte 566

Sets that maximize probability and a related variational problem
Juan J. Salamanca 591

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Les auteurs désirant faire paraître un article sont invités à le soumettre en français ou en anglais via le **portail de *Techniques d'enquête* sur le site Web de ScholarOne Manuscripts** (https://mc04.manuscriptcentral.com/surveymeth). Avant de soumettre l'article, prière d'examiner un numéro récent de *Techniques d'enquête* et de noter les points ci-dessous. Les articles doivent être soumis en Word ou Latex, préférablement en Word avec MathType pour les expressions mathématiques. Une version pdf est également requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé et introduction

2.1 Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
2.2 Le dernier paragraphe de l'introduction devrait contenir une brève description de chacune des sections.

3. Rédaction

3.1 Éviter les notes au bas des pages et les abréviations.
3.2 Limiter l'utilisation d'acronymes. Si un acronyme est utilisé, il doit être défini lors de sa première utilisation.
3.3 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·), etc.
3.4 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées par un chiffre arabe à la droite si l'auteur y fait référence plus loin. Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, l'équation (4.2) est la deuxième équation importante de la section 4.
3.5 Des caractères gras devraient normalement être utilisés pour distinguer les vecteurs et les matrices des valeurs scalaires.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés avec des chiffres arabes et porter un titre aussi explicatif que possible en haut des tableaux ou des figures. Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, le tableau 3.1 est le premier tableau de la section 3.
4.2 Une description textuelle détaillée des figures pourrait être requise à des fins d'accessibilité si le message transmis par l'image n'est pas suffisamment expliqué dans le texte.

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple : Cochran (1977, page 164).
5.2 La première fois qu'une référence est citée dans le texte, le nom de chacun des auteurs doit être écrit. Pour les fois suivantes, le nom de chacun des auteurs peut être écrit à nouveau. Cependant, si la référence contient trois auteurs ou plus, les noms du deuxième auteur et des auteurs suivants peuvent être remplacés par « et coll. ».
5.3 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les articles d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

6. Communications brèves

6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots, incluant les tableaux, les figures et la bibliographie.