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

Techniques d’enquête

Techniques d’enquête
47-1

Date de diffusion : le 24 juin 2021

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 STATCAN.infostats-infostats.STATCAN@canada.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 2021

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

Juin 2021

•

Volume 47

•

Numéro 1

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

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

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*
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, *Université de Montréal*
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.N.D. 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*
A. Zaslavsky, *Harvard University*
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@canada.ca).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 47, numéro 1, juin 2021

Table des matières

Article sollicité Waksberg

Roger Tourangeau
Science et gestion d'enquête 3

Articles réguliers

Shu Yang, Jae Kwang Kim et Youngdeok Hwang
Intégration de données d'enquêtes probabilistes et de mégadonnées aux fins d'inférence de population finie au moyen d'une imputation massive 33

Sixia Chen, Yichuan Zhao et Yuke Wang
La méthode de vraisemblance empirique fondée sur l'échantillon sous un plan de sondage complexe avec réponses brouillées 65

Edgar Bueno et Dan Hedlin
Méthode de détermination d'une stratégie d'échantillonnage efficace et robuste en cas d'incertitude du modèle 83

Seongmi Choi, Balgobin Nandram et Dalho Kim
Inférence bayésienne prédictive des proportions dans de petits domaines avec biais de sélection 99

Marius Stefan et Michael A. Hidiroglou
Estimation sur petits domaines réconciliée sous le modèle de base au niveau de l'unité lorsque les taux d'échantillonnage sont non négligeables 133

Jan A. van den Brakel et Harm-Jan Boonstra
Estimation des discontinuités de domaine au moyen de modèles de Fay-Herriot hiérarchiques bayésiens 163

Aejeong Jo, Balgobin Nandram et Dal Ho Kim
Le regroupement bayésien aux fins d'analyse des données catégoriques sur petits domaines 205

Communication brève

Sixia Chen, David Haziza et Alexander Stubblefield
Note sur l'imputation multirobuste par appariement selon la moyenne prédictive réalisée avec des données d'enquête complexes 231

Autres revues 241

Techniques d'enquête, juin 2021
Vol. 47, nº 1, pp. 1-2
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.

Ce numéro de *Techniques d'enquête* commence par le 20e article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé de Michael A. Hidiroglou (président), Bob Fay, Jean Opsomer et Elizabeth Stuart, d'avoir choisi Roger Tourangeau comme auteur de l'article du prix Waksberg de 2020.

Communication sollicitée pour le prix Waksberg 2020

Auteur : Roger Tourangeau

Roger Tourangeau est un ancien vice-président de Westat, où il a codirigé l'unité de méthodologie. Avant de joindre Westat, il a été directeur du *Joint Program in Survey Methodology* à l'Université du Maryland et professeur de recherche à l'*Institute for Social Research* de l'Université du Michigan. Il est l'auteur principal de deux ouvrages : *The Psychology of Survey Response* (coécrit avec Lance Rips et Kenneth Rasinski), gagnant du prix du livre 2006 de l'*American Association for Public Opinion Research* (AAPOR), et *The Science of Web Surveys* (coécrit avec Frederick Conrad et Mick Couper). Il a publié plus de 90 articles sur des sujets traitant des méthodes d'enquête. Ses travaux ont été financés par la *National Science Foundation*, le *National Institute on Drug Abuse* et le *National Institute of Child Health and Human Development*. Tourangeau a reçu de nombreux prix et distinctions. Il a été nommé membre de l'*American Statistical Association* (ASA) en 1999, a reçu le prix Helen Dinerman (la plus haute distinction décernée par l'Association mondiale de recherche sur l'opinion publique – WAPOR) en 2002 et a reçu le prix AAPOR Innovators (avec Thomas Jabine, MironStraf et Judy Tanur) en 2006. Il a présidé la section sur les méthodes de recherche par sondage de l'ASA en 2006, a cofondé le *Journal of Survey Statistics and Methodology* en 2012, a été président de l'AAPOR de 2016 à 2017, a été le conférencier principal à la Conférence Morris Hansen (parrainée par la Washington Statistical Society) de 2018, et a été désigné conférencier de la Conférence Sirken 2020 par l'AAPOR et l'ASA.

Gagnants du prix Waksberg et leurs articles sollicités depuis 2001

2021 Sharon **Lohr**, Sujet de l'article à déterminer, (prévu pour le vol. 47, 2)

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, juin 2021
Vol. 47, nº 1, pp. 3-32
Statistique Canada, nº 12-001-X au catalogue

Science et gestion d'enquête

Roger Tourangeau[1]

Résumé

Il est maintenant possible de gérer des enquêtes à l'aide de modèles statistiques et d'autres outils applicables en temps réel. Le présent article porte sur trois nouveautés nées des tentatives d'adoption d'une démarche plus scientifique dans la gestion du travail sur le terrain : (1) l'utilisation de plans réactifs et adaptatifs aux fins de réduction du biais de non-réponse, d'autres sources d'erreur ou des coûts; (2) la définition de l'itinéraire optimal des déplacements des intervieweurs de façon à réduire les coûts; et (3) la rétroaction rapide aux intervieweurs aux fins de diminution des erreurs de mesure. L'article commence en examinant les expériences et les études par simulations portant sur l'efficacité des plans adaptatifs et réactifs. Ces études suggèrent à penser que ces plans peuvent produire des gains modestes dans la représentativité des échantillons d'enquête ou des économies de coûts modestes, mais qu'ils sont aussi susceptibles d'entraîner des répercussions négatives. La section suivante de l'article étudie les efforts déployés pour fournir aux intervieweurs un itinéraire recommandé à suivre pendant leur prochain déplacement sur le terrain, l'objectif étant de faire correspondre davantage le travail des intervieweurs sur le terrain avec les priorités de recherche tout en réduisant les temps de déplacement. Notons toutefois qu'une étude mettant à l'essai cette stratégie a révélé que, souvent, les intervieweurs ignorent les instructions. Ensuite, l'article décrit les tentatives de donner une rétroaction rapide aux intervieweurs, à partir des enregistrements automatiques de leurs interviews. Souvent, les intervieweurs lisent les questions d'une manière qui influence les réponses des répondants. La correction de ce type de problèmes a rapidement apporté des améliorations notables de la qualité des données. Toutes les méthodes visent à remplacer le jugement des intervieweurs, des superviseurs sur le terrain et des gestionnaires d'enquête par des modèles statistiques et des constatations scientifiques.

Mots-clés : Gestion d'enquête; plan réactif; plan adaptif; itinéaire optimal.

1. Introduction

De nos jours, il n'est pas facile de mener une enquête en raison du double problème de l'augmentation des coûts et de la baisse des taux de réponse (par exemple Tourangeau, 2017; Williams et Brick, 2018). Aux États-Unis, ces deux tendances se dessinent nettement depuis les années 1970 (Atrostic, Bates, Burt et Silberstein, 2001; Steeh, Kirgis, Cannon et Dewitt, 2001), mais elles semblent s'être accélérées depuis une dizaine d'années environ. Des tendances identiques se vérifient dans tous les pays développés (de Leeuw et de Heer, 2002). On peut dire que les spécialistes de la recherche sur les enquêtes ne savent pas vraiment ce qui leur est arrivé (mais voir Brick et Williams (2013) pour une étude réfléchie des causes possibles de ces tendances). Il apparaît néanmoins clair que les personnes disposées à répondre à une enquête sont de moins en moins nombreuses. La tendance à la baisse des taux de réponse s'explique principalement par une résistance croissante aux enquêtes dans le grand public.

En partie en réponse à cette crise sectorielle mondiale, les chercheurs ont examiné de plus près l'effet de la baisse des taux de réponse sur l'exactitude des estimations d'enquête et ont également proposé diverses mesures visant à contrer la baisse des taux de réponse. Par exemple, de plus en plus d'enquêtes ont commencé à offrir des incitatifs, à utiliser des lettres d'avis et à augmenter le nombre de tentatives de contact.

1. Roger Tourangeau, 1601, Third Avenue, Apt. 19AW, New York, NY 10128. Courriel : RTourang@gmail.com.

Une autre tendance a consisté à utiliser une série de méthodes visant à améliorer la *gestion* des enquêtes afin de réduire le risque d'erreur, les coûts de la collecte des données, ou les deux. Dans la section 2, nous commençons l'article en passant en revue ces efforts, généralement nommés plans *réactifs* et *adaptatifs*. Dans la section 3, nous examinons une autre méthode de réduction des coûts et d'augmentation de l'efficacité des enquêtes en personne. Cette méthode d'*itinéraire optimal* consiste pour les gestionnaires d'enquête à donner aux intervieweurs sur le terrain des instructions détaillées sur les cas qu'ils doivent essayer d'interviewer et sur l'itinéraire de leur prochaine visite sur le terrain. Dans la section 4, nous examinons une autre nouveauté susceptible d'améliorer le rendement des intervieweurs avec l'interview enregistrée assistée par ordinateur, ou *IEAO* (Hicks, Edwards, Tourangeau, McBride, Harris-Kojetin et Moss, 2010). L'IEAO permet au personnel du bureau central d'entendre la manière dont les intervieweurs posent les questions sur le terrain et de corriger à mi-parcours leur rendement. Des recherches ont montré que les intervieweurs sur le terrain s'écartent plus souvent du scénario que les intervieweurs par téléphone (Schaeffer, Dykema et Maynard, 2010; West et Blom, 2017), probablement parce qu'on peut surveiller les intervieweurs travaillant par téléphone et leur donner des rétroactions en temps réel. Dans cette quatrième section, nous décrivons deux expériences dans lesquelles le personnel du bureau central a fourni une *rétroaction rapide* aux intervieweurs sur le terrain, soit dans les deux ou trois jours qui suivent l'interview. Ces techniques ont en commun de remplacer le jugement des intervieweurs et du personnel sur le terrain par les prescriptions fondées sur des données probantes des gestionnaires d'enquête, c'est-à-dire qu'elles tentent de remplacer l'art de la gestion par la science de la gestion. Finalement, la section 5 présente quelques conclusions.

2. Plan de collecte de données adaptatif et réactif

Les plans adaptatifs et réactifs désignent une famille de méthodes destinées à adapter le travail sur le terrain afin de réduire le biais, la variance ou le coût (voir les revues de la littérature dans Chun, Heeringa et Schouten (2018); Schouten, Peytchev et Wagner (2017); et Tourangeau, Brick, Lohr et Li (2017)). Au moyen des plans réactifs, les chercheurs utilisent plusieurs phases de la collecte des données pour réduire les coûts ou les erreurs d'enquête. Les plans adaptatifs utilisent différentes formes de priorisation des cas, d'adaptation et de règles pour arrêter la collecte de données afin d'atteindre des objectifs semblables.

Groves et Heeringa (2006) ont lancé ce mouvement par cette description des plans réactifs :

> Les plans réactifs sont organisés en phases de plan. Une phase de plan est une période de la collecte des données pendant laquelle le même ensemble de base de sondage, de mode de collecte des données, de plan de sondage, de protocoles de recrutement et de conditions de mesure existe. Par exemple, une enquête peut commencer par une tentative de questionnaire par la poste à la première phase, suivie d'une phase d'interview téléphonique sur les non-répondants de la première phase, puis d'une troisième phase

> d'interviews en personne. Il faut noter que cette utilisation du terme « phase » comprend plus de caractéristiques du plan que simplement celles du plan de sondage, qui sont communément impliquées dans le terme « échantillonnage à plusieurs phases ». (Pp. 440-441).

Bien sûr, l'American Community Survey (Enquête sur les collectivités américaines) utilisait un plan à trois phases (envoi par la poste, suivi d'un suivi téléphonique, après lequel vient un suivi en personne avec un sous-échantillon des cas restants) exactement comme celui que Groves et Heeringa ont décrit plusieurs années avant que ces auteurs ne qualifient ces méthodes de « plans réactifs » (U.S. Census Bureau, 2014).

Groves et Heeringa citent plusieurs enquêtes utilisant des plans réactifs, mais ils s'intéressent principalement au cycle 6 de la *National Survey of Family Growth* (NSFG, Enquête nationale sur la croissance des familles). La plupart des enquêtes qu'ils abordent, y compris le cycle 6 de la NSFG, appliquent un échantillonnage à deux phases (ce qui signifie qu'elles sélectionnent le sous-échantillon des non-répondants restants à un temps donné de la période sur le terrain et restreignent le suivi à ce sous-échantillon) et offraient des incitatifs importants ou apportaient d'autres modifications à la collecte des données pour les répondants de la dernière phase. L'innovation réelle de la NSFG ne résidait pas dans l'utilisation de plusieurs phases d'échantillonnage (qui existait environ depuis Hansen et Hurwitz (1946)) ni de plusieurs modes de collecte de données (en effet, tous les cas de la NSFG sont interviewés en personne), mais dans l'application de paradonnées et d'une modélisation de la propension en temps réel afin d'orienter le travail sur le terrain. Le sous-échantillonnage des non-répondants au cycle 6 de la NSFG était fondé sur des modèles de propension qui étaient mis à jour fréquemment et qui intégraient des renseignements recueillis lors de contacts antérieurs avec le cas de l'échantillon. Dans la phase finale du cycle 6 de la NSFG, la collecte des données a été limitée à certaines zones d'échantillonnage, soit les zones comptant un plus grand nombre de cas actifs et les zones aux propensions estimées relativement élevées qui étaient plus susceptibles d'être conservées dans les travaux de suivi ultérieurs sur le terrain.

Les plans réactifs et les plans à phases multiples plus classiques se distinguent aussi, du moins conceptuellement, par la notion de capacité de phase. Groves et Heeringa soutiennent qu'une phase donnée de la collecte de données approche une limite dans sa capacité de modifier les estimations de l'enquête (et de réduire les biais). Une fois que cette limite de capacité est atteinte, il peut être nécessaire de modifier le protocole pour améliorer la représentativité de l'échantillon et réduire le biais. Idéalement, les dernières phases de la collecte des données font intervenir différents types de répondants des phases précédentes, réduisant ainsi les biais de non-réponse restants. Certaines personnes peuvent être plus portées à répondre par la poste qu'à une interview en personne. Des incitatifs plus importants peuvent aider à recruter des personnes non intéressées par le sujet (Groves, Singer et Corning, 2000). Au mieux, les différentes phases de la collecte de données sont complémentaires et, ensemble, forment un échantillon plus représentatif que chacune des phases séparément.

2.1 Priorisation des cas et stratégies connexes

Le cycle 6 de la NSFG est un premier exemple d'une stratégie connue sous le nom de *priorisation des cas* qui attribue délibérément plus d'efforts à certains cas de l'échantillon qu'à d'autres. Bien entendu, les gestionnaires d'enquête ont toujours accordé la priorité à certains cas par rapport à d'autres. On demande aux intervieweurs de s'assurer de respecter les rendez-vous, par exemple, ou de mettre de côté les cas de refus « non catégoriques » pendant un certain temps. Ce qui diffère dans les utilisations récentes de la priorisation des cas est qu'elles ne se fondent pas sur la disposition d'un cas, mais sur des modèles de la propension à répondre du cas. Au cycle 6 de la NSFG, un sous-échantillon probabiliste de cas a été conservé pour des travaux ultérieurs, les probabilités d'échantillonnage de la deuxième phase étant en partie fondées sur les propensions prédites des cas restants. Les efforts ultérieurs ont été explicites dans leur utilisation des propensions à répondre aux fins d'orientation du travail sur le terrain.

Selon les cas auxquels la priorité est accordée, la priorisation des cas peut servir différents objectifs. Par exemple, le fait de concentrer le travail sur le terrain sur les cas ayant une propension à répondre élevée peut maximiser la taille de l'échantillon final ou réduire les coûts par cas. Beaumont, Bocci et Haziza (2014) distinguent trois objectifs potentiels de ce type de plans :

1) minimiser la variance;
2) minimiser le biais de non-réponse ou sa mesure approximative, comme le déséquilibre de l'échantillon (Särndal, 2011; voir aussi Schouten, Cobben et Bethlehem, 2009);
3) maximiser les taux de réponse.

Les premier et troisième objectifs sont liés, car la maximisation des taux de réponse tend à produire des échantillons plus grands et, par conséquent, à réduire les variances de l'échantillon. Bien que certains chercheurs aient commencé à examiner l'utilisation de ces plans aux fins de réduction des erreurs de mesure (Calinescu, Bhulai et Schouten, 2013), la plupart des efforts déployés à ce jour cherchaient à réduire le biais de non-réponse ou les coûts.

Au cycle 6 de la NSFG, l'objectif statistique n'est pas tout à fait clair. Le fait de suréchantillonner les zones comportant les plus grands nombres de cas restants et celles caractérisées par une propension plus élevée tendrait à maximiser la taille de l'échantillon final et à réduire les coûts par cas. Dans le même ordre d'idées, Groves, Benson, Mosher, Rosenbaum, Granda, Axinn, Lepkowski et Chandra (2005) ont fait remarquer que « cette option de plan mettait beaucoup l'accent sur la rentabilité du… plan de la phase [finale] pour produire des interviews, et non sur la minimisation des erreurs-types de l'ensemble de données qui en résulte ». Cependant, Groves et coll. (2005) ont également avancé que la phase finale de la collecte de données visait à produire un échantillon « plus représentatif » (page 38) en modifiant le protocole de collecte de données pour attirer les membres de l'échantillon qui n'avaient pas répondu auparavant. Il reste que le fait de cibler les zones comprenant plus de cas avec des propensions à répondre estimées élevées – c'est-à-dire des cas qu'on prévoit comme étant plus faciles à obtenir – pourrait en fait

exacerber les problèmes de représentativité en ajoutant des répondants semblables à ceux qui ont répondu auparavant.

La plupart des applications ultérieures de la priorisation des cas ont adopté l'approche opposée : elles ont essayé d'égaliser la propension globale à répondre en concentrant les efforts sur le terrain sur les cas les plus difficiles. Pour comprendre pourquoi cette stratégie est raisonnable, il est utile de se pencher sur les mathématiques du biais de non-réponse.

2.2 Facteurs influant sur le biais de non-réponse

Dans une perspective stochastique (par exemple Bethlehem, 1988), le biais de l'estimateur non corrigé d'une moyenne ou d'une proportion $(\hat{\bar{y}})$ peut être exprimé comme suit :

$$\text{Biais}(\hat{\bar{y}}) \approx \frac{\sigma_\phi \sigma_y \rho_{\phi,y}}{\bar{\phi}}, \tag{2.1}$$

où $\bar{\phi}$ et σ_ϕ sont la moyenne et l'écart-type des propensions à répondre, σ_y est l'écart-type d'une variable d'enquête, et $\rho_{\phi,y}$ est la corrélation entre les propensions à répondre et cette variable d'enquête. Comme (2.1) le montre clairement, à la fois le taux de réponse global $(\bar{\phi})$ et la variation des taux de réponse (σ_ϕ) jouent un rôle dans le biais, c'est pourquoi il semble raisonnable d'essayer de maximiser les taux de réponse (par exemple en priorisant les cas relativement faciles) ou d'égaliser les propensions à répondre (en priorisant les cas les plus difficiles).

Comme plusieurs chercheurs l'ont souligné, le biais de non-réponse est une propriété d'une estimation d'enquête et non d'une enquête, et, comme (2.1) le montre explicitement, deux propriétés au niveau des variables influent aussi sur le biais : la corrélation entre la variable d'enquête et les propensions à répondre $(\rho_{\phi,y})$ et la variabilité de la variable d'enquête (σ_y). Ces deux propriétés varient d'une variable d'enquête à l'autre. Étant donné que deux des ingrédients de l'expression du biais sont des facteurs au niveau de l'étude et que deux autres sont au niveau de la variable, la question se pose de savoir quelle proportion de la variation du biais de non-réponse se trouve entre les différentes enquêtes et quelle proportion au sein des enquêtes.

Brick et moi-même (Brick et Tourangeau, 2017) avons cherché à traiter la question en analysant de nouveau les données d'une étude réalisée par Groves et Peytcheva (2008). Ils ont examiné 959 estimations du biais de non-réponse provenant de 59 études. Huit cent quatre de ces estimations du biais comportaient des proportions. Presque toutes les autres étaient des moyennes. (Parce que quatre des estimations nous semblaient problématiques, nous les avons retirées de notre nouvelle analyse.) Comme Groves et Peytcheva, nous avons examiné la statistique du biais relatif absolu (biaisrel absolu), ou la différence absolue entre l'estimation du répondant et l'estimation de l'échantillon complet divisée par l'estimation de l'échantillon complet :

$$R_i = \frac{|\theta_{ri} - \theta_{ni}|}{\theta_{ni}}, \tag{2.2}$$

où R_i est le biaisrel absolu pour la statistique i, θ_{ri} est la valeur estimée pour cette statistique en fonction des répondants, et θ_{ni} est l'estimation correspondante de l'échantillon complet. Le biaisrel absolu est utile car il met toutes les estimations du biais à la même métrique que le pourcentage de décalage de l'estimateur. Dans cette nouvelle analyse, nous avons également examiné les différences absolues (le numérateur dans (2.2)) pour les proportions estimées.

Le tableau 2.1 présente plusieurs statistiques de la réanalyse. Par exemple, nous avons calculé la corrélation entre les estimations de biais individuelles et les taux de réponse au niveau de l'étude; ces résultats sont présentés dans le panneau supérieur du tableau. Les trois panneaux centraux du tableau montrent ce qui se passe quand le biais moyen de l'étude est utilisé à la place des estimations de biais individuelles. Certaines corrélations fondées sur les moyennes au niveau de l'étude sont considérablement plus élevées que celles fondées sur les estimations individuelles, particulièrement quand les données sont pondérées par le nombre d'estimations de chaque étude (les valeurs r de 0,40 à 0,55). Les deux panneaux inférieurs du tableau montrent que le biais de non-réponse comporte une composante importante au niveau de l'étude. Par exemple, les estimations R^2 provenant d'une analyse de la variance unidirectionnelle indiquent que la composante interétude représente de 21 à 40 % de la variation globale des estimations du biais de non-réponse. Les résultats des modèles multiniveaux donnent des conclusions semblables. Cette composante interétude du biais reflète vraisemblablement deux variables principales, soit la propension moyenne à répondre (reflétée dans le taux de réponse global) et la variation des propensions à répondre entre répondants.

Tableau 2.1
Relation entre les taux de réponse et les mesures du biais au niveau de l'estimation et de l'étude

	Toutes les statistiques	Proportions seulement
Corrélations au niveau de l'estimation		
Taux de réponse et biaisrel absolu	-0,191(n = 955)	-0,256 (n = 802)
Taux de réponse et différence absolue	-	-0,323 (n = 802)
Corrélations au niveau de l'étude non pondérées		
Taux de réponse et biaisrel absolu moyen	-0,255 (n = 57)	-0,315 (n = 43)
Taux de réponse et différence absolue moyenne	-	-0,246 (n = 43)
Corrélations au niveau de l'étude pondérées par le nombre d'estimations		
Taux de réponse et biaisrel absolu moyen	-0,402(n = 57)	-0,552 (n = 43)
Taux de réponse et différence absolue moyenne	-	-0,508 (n = 43)
Corrélations au niveau de l'étude pondérées par la taille d'échantillon moyenne		
Taux de réponse et biaisrel absolu moyen	-0,413 (n = 57)	-0,247 (n = 43)
Taux de réponse et différence absolue moyenne	-	-0,208 (n = 43)
Coefficients de corrélation intraclasse au niveau de l'estimation à partir du modèle multiniveau		
Biaisrel absolus	0,164 (n = 955)	0,161 (n = 802)
Différences absolues	-	0,509 (n = 802)
au niveau de l'estimation R^2 à partir d'une analyse de la variance unidirectionnelle		
Biaisrel absolus	0,221(n = 955)	0,211 (n = 802)
Différences absolues	-	0,395 (n = 802)

Les résultats du tableau 2.1 sont importants parce que les plans de sondage réactifs et adaptatifs fonctionnent principalement au niveau de l'étude. Par exemple, en général, la priorisation des cas augmente les propensions globales à répondre ou réduit la variation des propensions. Or il s'agit des deux

variables principales au niveau de l'étude influant sur le niveau du biais de non-réponse. De plus, si un plan réussit à réduire la variation globale des propensions à répondre, cela aura tendance à atténuer les corrélations entre les propensions et les variables d'enquête à tous les niveaux. À l'extrême, en l'absence de variation dans les propensions à répondre, la corrélation avec toutes les variables de l'enquête sera nulle et il n'y aura pas de biais de non-réponse. Les résultats du tableau 2.1 semblent contredire l'opinion selon laquelle les taux de réponse n'ont pas d'importance. Les taux de non-réponse sont manifestement une mesure approximative imparfaite du biais de non-réponse, mais ils constituent un prédicteur important du niveau moyen de biais dans les estimations d'une enquête.

2.3 Évaluations expérimentales des plans réactifs et adaptatifs

Dans quelle mesure les plans réactifs et adaptatifs atteignent-ils leurs objectifs ? D'entrée de jeu, il me faut souligner que nos attentes ne doivent pas être trop élevées. Comme cela a été dit dans un article précédent (Tourangeau et coll., 2017, page 208), ces plans « représentent une tentative de faire plus avec moins ou, tout du moins, de faire le plus possible avec moins » dans un environnement de plus en plus défavorable aux enquêtes. À ce jour, les études ont utilisé quatre stratégies élémentaires pour atteindre un ou plusieurs objectifs statistiques : les plans à plusieurs phases (comme celui décrit par Groves et Heeringa, 2006), d'autres types de priorisation des cas (où différents niveaux d'effort sont attribués à différents cas), des stratégies de contact adaptatives (qui modifient le calendrier des tentatives de contact en fonction de modèles de propension pour maximiser les chances de contact), et la personnalisation du travail sur le terrain ou du mode de collecte des données en fonction de ce qu'on sait sur les cas avant qu'ils ne soient contactés. Je passerai brièvement en revue certains des principaux efforts déployés aux fins d'évaluation de chaque approche.

Plans à plusieurs phases et priorisation des cas. Peytchev, Baxter et Carley-Baxter (2009) décrivent une autre étude qui, comme le cycle 6 de la NSFG, utilise un plan à plusieurs phases. Ils ont effectué une étude téléphonique en deux phases. La deuxième phase utilisait un questionnaire nettement plus court et offrait un plus grand incitatif que la première. Les cas ont reçu jusqu'à 20 appels pendant la phase 1. Certains cas en ont reçu encore plus. Dans l'ensemble, cette phase a produit un taux de réponse de 28,5 %. Pendant la phase 2, les chercheurs ont sous-échantillonné les non-répondants restants, raccourci le questionnaire de 30 à 14 minutes, offert un incitatif prépayé de 5 $ et un incitatif conditionnel de 20 $. (La phase 1 offrait seulement des incitatifs conditionnels.) La phase 2 a produit un taux de réponse de 9,8 % (soit 35,5 % dans l'ensemble). L'évaluation du plan était fondée sur deux ensembles de comparaisons : Peytchev et ses collègues ont comparé les premiers répondants et les répondants tardifs de la phase 1 et ils ont comparé les répondants de la phase 1 à ceux de la phase 2. Ils ont estimé qu'il était peu probable que les répondants tardifs (interviewés après au moins six tentatives d'appel) de la phase 1 diffèrent sur les variables clés de l'étude – victimisation criminelle déclarée de diverses sortes – des premiers répondants (interviewés après cinq tentatives au plus), parce qu'ils ont été recrutés selon le même protocole. Les résultats indiquent que l'ajout des répondants tardifs de la phase 1 n'a pas significativement changé les

estimations. En revanche, les auteurs pensaient que les répondants de la phase 2 étaient susceptibles de différer des répondants de la phase 1, parce que les changements de protocole attireraient différents types de répondants. Ce type d'argument était quelque peu étayé pour les hommes. Les répondants de sexe masculin de la phase 1 étaient plus susceptibles de déclarer des victimisations que les répondants de sexe masculin de la phase 2, avec des différences significatives sur quatre des six taux de victimisation. Toutefois, il y avait moins de données prouvant que le changement de protocole à la phase 2 avait une incidence sur les estimations pour les répondantes. De plus, même dans l'échantillon de la phase 1, on constate des différences entre les cas de répondants de sexe masculin qui n'ont jamais refusé de répondre et ceux qui ont répondu après avoir refusé. À l'instar des répondants de sexe masculin de la phase 2, les taux de victimisation de la phase 1 chez les hommes qui avaient d'abord refusé de répondre étaient également significativement plus faibles pour quatre des six estimations clés. Cela indiquerait que les protocoles de conversion des refus ont changé la composition de l'échantillon de la phase 1 et qu'ils n'ont pas seulement servi à attirer plus de répondants du même type.

Peytchev, Riley, Rosen, Murphy et Lindblad (2010) décrivent une étude qui a adapté dès le départ le protocole de collecte des données pour différents groupes de cas. Leur étude comportait un sondage par panel et les propensions à répondre pour chaque cas ont été estimées à l'aide de l'information du cycle précédent. Les cas présentant une faible propension à répondre prédite ont été attribués aléatoirement à un traitement expérimental ou témoin. Pendant la plus grande partie de la période de collecte des données, les intervieweurs recevaient une prime de 10 $ pour chaque interview terminée avec l'un des cas témoins, et 20 $ pour chaque interview terminée avec l'un des cas expérimentaux. (Pendant la phase 1, aucune prime n'était proposée pour les interviews témoins et une prime de 10 $ était accordée pour les interviews expérimentales.) On constate peu de différence dans les taux de réponse finaux pour les deux groupes de cas (89,8 % pour les cas témoins comparativement à 90,8 % pour les cas expérimentaux) et dans le nombre moyen de tentatives de contact par cas (5,0 pour les cas témoins comparativement à 4,9 pour les cas expérimentaux). Bien que la variance des propensions à répondre estimées soit plus faible chez les cas expérimentaux, les biais de non-réponse estimés (fondés sur les corrélations entre les variables de l'enquête et les propensions à répondre ajustées) étaient plus élevés.

Une autre série d'expériences illustre des difficultés pratiques posées par la priorisation des cas. Wagner, West, Kirgis, Lepkowski, Axinn et Kruger Ndiaye (2012; voir aussi Lepkowski, Mosher, Groves, West, Wagner et Gu (2013)) ont mené 16 expériences au cours du cycle 7 de la NSFG, qui a permis de prélever 20 échantillons trimestriels. Les expériences ont examiné l'efficacité de « l'attribution d'un sous-ensemble aléatoire de cas actifs ayant des caractéristiques particulières pour que les intervieweurs leur accordent une plus grande priorité… Le premier objectif de ces expériences était de déterminer si les intervieweurs répondraient à une demande de priorisation de cas particuliers » (Wagner et coll., 2012, page 482). Dans seulement sept des 16 expériences, les cas prioritaires ont reçu significativement plus d'appels que les cas témoins et, deux fois seulement, cela a entraîné une augmentation importante des taux de réponse des cas prioritaires. D'autres expériences ont cherché à déplacer les efforts des intervieweurs

de la NSFG qui, au lieu d'essayer de terminer les interviews principales, devaient essayer de réaliser des présélections pendant une semaine de la période sur le terrain. Cette intervention a donné lieu à un plus grand nombre d'appels de présélection qu'au cours des semaines précédentes ou ultérieures, mais l'incidence sur le nombre de présélections *terminées* variait d'un trimestre à l'autre. Dans les deux cas, les efforts de priorisation des cas du cycle 7 de la NSFG ont eu des effets sur les actions des intervieweurs, mais moins sur les résultats d'enquête voulus, comme les taux de réponse.

Statistique Canada a également commencé à mettre en œuvre des plans réactifs pour ses enquêtes par ITAO et a mené deux expériences visant à évaluer ces plans. Les deux expériences ont utilisé trois phases de collecte de données avec priorisation des cas dans l'une des phases (Laflamme et Karaganis, 2010; Laflamme et St-Jean, 2011). À la phase 1, les cas ont été classés en fonction de la propension à répondre; à la phase 2, les cas ont été attribués aléatoirement soit à la condition de collecte réactive (dans laquelle des priorités sont attribuées aux cas et les cas hautement prioritaires reçoivent plus d'appels) soit à la condition de témoin; et à la phase 3, tous les cas restants ont reçu le même traitement. Pendant la phase 2, les cas prioritaires du groupe de collecte réactive étaient apparemment ceux qui avaient de propensions à répondre prédites élevées. L'objectif de la phase 3 était d'égaliser les propensions à répondre dans tous les sous-groupes clés. Encore une fois, les résultats indiquent des effets modestes. Fondamentalement, les taux de réponse globaux n'ont pas été touchés par la priorisation des cas. Dans une enquête, les taux de réponse étaient de 74,0 % pour le groupe témoin par rapport à 74,1 % pour le groupe de collecte réactive. Dans l'autre enquête, le groupe témoin avait un taux de réponse légèrement plus élevé (73,0 % par rapport à 72,8 %). Cela est un peu surprenant étant donné que la collecte réactive ciblait les cas les plus faciles de la phase 2. De plus, ni le nouveau plan d'échantillonnage à trois phases ni le protocole de collecte réactive n'ont eu d'effets manifestes sur la représentativité des échantillons, mais il se peut qu'ils aient réduit le nombre d'heures de travail des intervieweurs (voir le tableau 2.2 dans Laflamme et St-Jean (2011)). Or, la réduction des coûts sans réduction de la représentativité peut représenter un progrès valable, quoique modeste.

Stratégies de contact adaptatives. Les gestionnaires d'enquête peuvent-ils améliorer le taux de contact avec les membres de l'échantillon en modélisant le meilleur moment pour communiquer avec eux ? Bien que de nombreux articles aient étudié les moments optimaux pour communiquer avec les membres d'un échantillon d'enquête, peu d'entre eux ont vérifié si ces horaires d'appels « optimaux » produisent des gains empiriques. Wagner (2013) est une exception. Il rapporte cinq expériences qui ont utilisé des modèles pour prédire si un ménage d'un échantillon donné sera contacté lors de la prochaine tentative d'appel dans chacun des quatre « créneaux » d'appel (par exemple du mardi au jeudi de 16 h à 21 h). Des modèles semblables ont été utilisés dans les enquêtes téléphoniques (*Survey of Consumer Attitudes* ou SCA, Enquête sur les attitudes des consommateurs) et en personne (cycle 7 de la NSFG). Les modèles ont servi à déterminer le meilleur créneau d'appel (celui qui présente la probabilité de contact la plus élevées) pour chaque ménage de l'échantillon. Dans les groupes expérimentaux, les cas ont été placés

en haut de la liste d'appel pour ce créneau (dans la SCA) ou les intervieweurs sur le terrain ont reçu ce créneau comme moment recommandé pour communiquer avec le ménage (dans la NSFG).

Trois des expériences concernaient la SCA. Dans la première, la proportion d'appels produisant un contact était plus élevée pour les cas expérimentaux que pour les témoins (12,0 % par rapport à 9,9 %), mais la stratégie semblait contre-productive pour les cas qui avaient initialement refusé, avec des taux de contact *plus faibles* chez les refus initiaux dans le groupe expérimental. Une deuxième expérience variait le créneau d'appel pour les cas expérimentaux après un refus initial, mais cette stratégie a réduit la proportion globale d'appels produisant un contact. L'expérience finale de la SCA a tout de même constaté que le taux de contact pour les appels après conversion de refus était plus faible dans le groupe expérimental que dans le groupe témoin. Les résultats concernant la NSFG étaient également quelque peu décevants. Les intervieweurs sur le terrain ont apparemment ignoré les créneaux d'appel recommandés; seulement 23,6 % des cas expérimentaux ont été contactés dans le créneau recommandé (comparativement à 23,0 % dans le groupe témoin). Nous avons eu une expérience semblable quand nous avons souhaité que les intervieweurs suivent un itinéraire optimal dans leurs déplacements sur le terrain (voir la section 3.1 plus loin).

Personnalisation du travail sur le terrain. Luiten et Schouten (2013) rapportent une expérience qui a personnalisé l'approche de collecte de données selon différents sous-groupes dans l'Enquête néerlandaise sur les sentiments des consommateurs (ENSC). L'objectif était d'égaliser les propensions à répondre de tous les sous-groupes. L'ENSC consiste en enquêtes transversales répétées et, d'après les cycles précédents, Luiten et Schouten ont ajusté des modèles de propension à la coopération et au contact à partir des caractéristiques démographiques des membres de l'échantillon. Ces variables étaient disponibles pour l'ensemble de l'échantillon dans le registre de la population. La collecte des données a été réalisée en deux phases. Dans la première phase, les cas présentant la plus faible propension à la coopération ont reçu un questionnaire par la poste, ceux ayant la propension estimée la plus élevée ont été invités à répondre à un sondage sur le Web, et ceux entre les deux ont eu le choix entre un sondage par courrier et sur le Web. La deuxième phase consistait à assurer par téléphone le suivi des non-répondants. Les cas des différents quartiles de propension au contact ont été attribués à des horaires d'appel différents. Ceux ayant la propension de contact la plus élevée ont été interviewés plus tard dans la période sur le terrain et appelés dans la journée; ceux dans le second quartile supérieur ont été appelés deux fois la nuit, puis selon un calendrier qui alternait les appels de jour et de nuit; et ceux dans les deux quartiles inférieurs de propension de contact ont été appelés tous les jours à chaque quart de travail. Enfin, les meilleurs intervieweurs par téléphone ont été affectés aux cas ayant les propensions à la coopération estimées les plus faibles et les pires intervieweurs par téléphone ont été affectés aux cas ayant les propensions à la coopération estimées les plus élevées. Le groupe témoin de l'expérience était l'ENSC ordinaire, qui est une enquête par ITAO seulement.

Bien que le groupe de travail adaptatif sur le terrain n'ait eu qu'un taux de réponse légèrement supérieur à celui de l'ENSC ordinaire (63,8 % par rapport à 62,8 %, soit une différence non significative), la représentativité de l'échantillon expérimental, telle qu'elle est mesurée par l'indicateur R, était

significativement plus élevée que celle de l'échantillon témoin. (L'indicateur R, introduit par Schouten, Cobben et Bethlehem (2009), est basé sur la variation des propensions à répondre estimées. Un nombre plus élevé indique une variation moindre et, par conséquent, un échantillon plus représentatif.) Le tableau 2.2 ci-dessous montre que le travail adaptatif sur le terrain a réduit la variation des taux de contact et celle des taux de coopération. Dans les quartiles de propension au contact, les taux de contact variaient de 84,2 % à 96,9 % dans l'ENSC ordinaire; dans l'échantillon expérimental, la fourchette allait de 87,1 % à 95,3 %. Le plan adaptatif réduisait également la variation des taux de coopération. Cependant, les coûts du plan adaptatif étaient légèrement plus élevés que ceux de l'ENSC et le taux de coopération global était significativement *plus faible* dans l'échantillon expérimental. Malheureusement, comme le montre la présente étude, la réduction de la variabilité des propensions à répondre signifie souvent que moins d'efforts sont déployés pour l'obtention des cas les plus faciles, ce qui peut réduire le taux de réponse global.

Tableau 2.2
Taux de contact et de coopération, par groupes de quartile de propension

	Taux de contact	
Quartile de propension de contact	**Échantillon expérimental**	**Échantillon témoin**
Propension de contact la plus faible	87,1	84,2
Deuxième propension de contact la plus faible	96,6	94,5
Deuxième propension de contact la plus élevée	93,7	95,7
Propension de contact la plus élevée	95,3	96,9
	Taux de coopération	
Quartile de propension à la coopération	**Échantillon expérimental**	**Échantillon témoin**
Propension à la coopération la plus faible	65,1	62,7
Deuxième propension à la coopération la plus faible	71,4	68,4
Deuxième propension à la coopération la plus élevée	72,8	75,3
Propension à la coopération la plus élevée	74,7	79,2

Source : Tourangeau et coll. (2017); données de Luiten et Schouten (2013).

2.4 Études par simulations

Outre les expériences abordées dans la section précédente, trois autres études ont utilisé des simulations pour examiner les propriétés des plans de sondage adaptatifs et réactifs.

Règles d'arrêt. Lundquist et Särndal (2013) ont utilisé les données de l'Enquête suédoise sur les conditions de vie (ESCV) de 2009 pour étudier l'effet de différentes « règles d'arrêt », soit les règles déterminant qu'il faut mettre fin à la collecte des données. L'ESCV suit une stratégie de collecte des données en deux phases, avec jusqu'à 20 tentatives de contact téléphonique pendant la première phase de la collecte de données, suivies de dix autres au cours de la deuxième phase. Les auteurs observent que le fait de continuer à suivre le même protocole de collecte des données « produira très peu de changements dans les estimations au-delà d'un certain « point de stabilité » atteint assez tôt dans la collecte des données » (page 561). Cela ressemble beaucoup à la notion de « capacité de phase » de Groves et Heeringa (2006), soit le moment où un protocole de collecte de données commence à produire des rendements décroissants (ou quasi nuls). Sturgis, Williams, Brunton-Smith et Moore (2017) présentent des résultats semblant indiquer que ce point de stabilité serait atteint assez tôt pendant la période de terrain. Ils

ont examiné les estimations obtenues à partir de 541 questions provenant de six enquêtes menées en personne au Royaume-Uni. Ils ont constaté que les proportions espérées n'étaient, en moyenne, qu'à 1,6 % de l'estimation finale après une seule tentative de contact et qu'elles étaient erronées de seulement 0,4 % après cinq tentatives. Ces résultats montreraient que, sous l'angle de la réduction du biais, beaucoup d'efforts sur le terrain sont gaspillés.

Lundquist et Särndal ont montré que le biais de non-réponse estimé (fondé sur trois variables disponibles à la fois pour les répondants et les non-répondants provenant du registre de la population suédoise) dans l'ESCV était le plus faible après cinq à dix tentatives d'appel, mais qu'au-delà, il empirait progressivement. La deuxième phase de la collecte des données, qui accroît le taux de réponse de 60,4 % à 67,4 %, a empiré les biais de non-réponse pour deux des trois variables du registre. Ils ont examiné trois solutions de rechange au maintien du même protocole jusqu'à 30 tentatives. Ils ont divisé l'échantillon en huit sous-groupes en fonction de l'éducation, de la propriété et de l'origine nationale. Dans la première solution, les taux de réponse de chacun des huit sous-groupes seraient vérifiés à l'appel nº 12 de la phase initiale de la collecte des données et de nouveau à l'appel nº 2 de la deuxième phase; la collecte des données prendrait fin pour les sous-groupes ayant des taux de réponse de 65 % ou plus à ces points. Cette stratégie aurait produit un taux de réponse plus faible (63,9 %) que le protocole actuel, mais l'échantillon serait plus étroitement aligné sur la population selon huit caractéristiques démographiques. La deuxième solution qu'ils ont examiné aurait mis fin à la collecte de données pour un sous-groupe dès que son taux de réponse aurait atteint 60 %, et la troisième solution dès que le taux de réponse du sous-groupe aurait atteint 50 %. La stratégie des 50 % aurait produit l'échantillon le plus équilibré de tous et elle aurait réduit le nombre total de tentatives d'appel de plus du tiers. En partie, cette stratégie fonctionnerait aussi bien parce qu'elle aurait réduit les taux de réponse dans les sous-groupes à propension élevée, de sorte qu'ils seraient plus près de ceux des sous-groupes à faible propension. Comme dans l'étude de Peytchev et ses collaborateurs (Peytchev et coll., 2009), poursuivre au moyen du même protocole de collecte de données ne semblait guère améliorer la représentativité de l'échantillon, et il se pourrait que cela l'ait en fait réduite.

Dans une volonté du même ordre, en 2017, la *Medical Expenditure Panel Survey* (MEPS ou Enquête sur les dépenses médicales par panel) a utilisé une règle d'arrêt fondée sur un modèle de propension. La MEPS est une étude avec renouvellement de panel. Chaque année, un nouveau panel d'environ 10 000 adresses est sélectionné parmi un échantillon de ménages qui ont répondu à la *National Health Interview Survey* (Enquête nationale sur la santé réalisée par interviews) l'année précédente. Il est demandé aux ménages de l'échantillon de réaliser deux interviews de la MEPS au cours de la première année, deux au cours de la deuxième année et une cinquième au cours de la troisième. Le sondage est continu et des interviews sont réalisées tout au long de l'année. Les règles d'arrêt ont été appliquées en deux étapes pendant le premier semestre de 2017 : d'abord aux cas du troisième cycle (un début relativement en douceur, puisque la plupart des interviews du troisième cycle ont été organisées par téléphone et que la plupart des répondants étaient coopératifs, étant donné qu'ils avaient déjà participé deux fois), puis aux

cas du premier cycle. Les intervieweurs hésitent souvent à se conformer aux directives leur demandant de cesser de communiquer avec un cas après un nombre précis de tentatives. L'approche de la MEPS consistait à supprimer les cas à faible propension ayant un trop grand nombre de tentatives, généralement six, des tâches attribuées à l'intervieweur et à demander à un superviseur de les examiner. Le superviseur pouvait réintégrer un cas dans les tâches de l'intervieweur s'il y avait une raison de croire que le cas pourrait être terminé, mais la plupart du temps, ces cas étaient fermés (Hubbard, 2018). Dans l'ensemble, la mise en œuvre de la règle d'arrêt a réduit le nombre de tentatives en personne de 8 500, ce qui a permis d'importantes économies sur les coûts de travail sur le terrain.

Différentes stratégies de priorisation des cas. Dans un article ultérieur, Särndal et Lundquist (2014b) ont simulé les effets de deux méthodes d'égalisation des propensions à répondre entre les cas, à l'aide des données de l'Enquête suédoise sur les conditions de vie et de l'Enquête suédoise sur les préférences partisanes. Dans la première méthode (la méthode du *seuil*), aucune autre tentative de suivi n'est effectuée pour les cas dont la propension à répondre a atteint un certain seuil (inférieur au taux de réponse cible global). Cette méthode ressemble aux stratégies examinées dans leur article précédent (Lundquist et Särndal, 2013). Dans l'autre méthode (la méthode des *proportions égales*), à différents moments de la période de terrain (par exemple après trois, six ou neuf tentatives d'appel), la portion de l'échantillon ayant la propension à répondre la plus élevée est mise de côté et le travail sur le terrain se poursuit seulement pour les cas restants. Dans les deux enquêtes, les deux méthodes d'égalisation des propensions à répondre ont réduit la distance entre les répondants et l'échantillon complet sur un ensemble de variables auxiliaires, comparativement à la poursuite du travail de terrain sur tous les non-répondants restants, comme cela avait été fait dans les enquêtes réelles. L'étude est aussi parvenue à la conclusion que le calage de l'échantillon au moyen des variables auxiliaires a éliminé une partie du biais de non-réponse, mais que ce biais a été encore plus réduit quand l'ensemble des répondants était plus étroitement aligné sur la population au commencement. Cette constatation est importante, car les variables disponibles pour ajuster les modèles de propension sont également celles qui sont disponibles pour les corrections postérieures à l'enquête, et on ne sait pas s'il est plus efficace d'égaliser les taux de réponse (ou les propensions à répondre) pendant la collecte des données plutôt que de simplement corriger les poids des cas par la suite. Särndal et Lundquist (2014b) trouvent des gains pour les deux méthodes.

Beaumont, Bocci et Haziza (2014) présentent une autre étude par simulations qui examine l'effet de la priorisation des cas. Ils ont comparé quatre stratégies : (1) *effort constant* (pas de priorisation des cas); (2) *effort optimal* (en réduisant les appels aux membres des groupes s'approchant de leur taux de réponse cible); (3) *égalisation des taux de réponse* entre groupes (en concentrant les appels sur les groupes à faible propension à répondre); et (4) *maximisation du taux de réponse global* (en concentrant les appels sur les groupes à propension élevée). Les simulations de Beaumont et de ses collaborateurs ont supposé trois scénarios différents : propensions à répondre uniformes, propensions à répondre uniformes au sein des groupes, et propensions à répondre fortement $(r = 0{,}67)$ corrélées avec la variable d'intérêt de l'enquête. (De plus, la simulation supposait que l'échantillon se composait de trois sous-groupes, que les appels

donnant lieu à une interview coûtaient 25 fois plus cher que ceux qui n'y parvenaient pas, que les appels à un cas étaient plafonnés à 25, et que le budget de collecte des données de l'enquête était fixe.)

Les simulations ont permis d'étayer trois grandes conclusions. Premièrement, quand les propensions à répondre sont constantes globalement ou constantes au sein de chaque groupe, toutes les stratégies d'effort produisent des estimations sans biais, mais quand les propensions sont fortement liées à la variable d'enquête, elles produisent toutes un biais. Deuxièmement, ni l'indicateur R ni le taux de non-réponse n'étaient un bon indicateur du biais de non-réponse ou de la variance due à la non-réponse. Enfin, quand les propensions à répondre sont connues, la stratégie de l'effort optimal produit une erreur quadratique moyenne un peu plus faible que les autres stratégies (voir le tableau 2.2 dans Beaumont et coll. (2014)) et la stratégie visant à maximiser les taux de réponse a produit les pires résultats. La stratégie d'effort optimal ressemble aux approches étudiées par Lundquist et Särndal (2013). Bien entendu, en pratique, une difficulté se pose : les propensions à répondre ne sont pas connues dans le cas d'enquêtes réelles et elles peuvent ne pas être estimées avec exactitude à partir des variables auxiliaires disponibles.

2.5 Sommaire

Le tableau 2.3 résume les résultats des études expérimentales et par simulations. En général, ils montrent à quel point il est difficile d'augmenter les taux de réponse dans le contexte actuel. Par exemple, seules deux des 16 expériences décrites par Wagner et ses collaborateurs ont augmenté significativement les taux de réponse dans la NSFG (Wagner et coll., 2012). Certaines études (par exemple Luiten et Schouten, 2013) démontrent des réductions de la variation des taux de réponse entre les sous-groupes de l'échantillon, bien que dans une étude (Peytchev et coll., 2010), cette diminution apparente de la variation des propensions à répondre estimées semblait *augmenter* le biais de non-réponse plutôt que de le réduire. Laflamme et St-Jean (2011) ont rapporté qu'un plan réactif réduisait les coûts comparativement au protocole standard, mais Luiten et Schouten (2013) ont indiqué qu'un plan adaptatif augmentait les coûts par cas. Dans toutes les études (y compris le cycle 6 de la NSFG), il semblerait que les plans de sondage réactifs et adaptatifs produisent certains gains en matière de représentativité de l'échantillon, mais qu'ils ont peu d'effet sur les taux de réponse globaux ou les coûts globaux.

Plusieurs études non expérimentales arrivent à des conclusions similaires. Ces études comparent les estimations finales d'une enquête à celles qui auraient été obtenues sans la phase finale de la collecte des données, lors de laquelle un changement majeur a été apporté au protocole de collecte des données. Par exemple, Groves et ses collaborateurs (Groves et coll., 2005) ont montré que la phase finale de la collecte des données du cycle 6 de la NSFG, qui fait remonter le taux de réponse global de 64 % à 80 %, réduit également la variation des taux de réponse entre les sous-groupes (voir aussi Axinn, Link et Groves, 2011). Ces résultats sont semblables aux résultats expérimentaux rapportés par Peytchev, Baxter et Carley-Baxter (2009), qui ont constaté que des changements importants dans le protocole (des incitatifs plus grands et un questionnaire plus court) modifiaient les estimations de l'étude, du moins pour les hommes. Toutefois, ces changements étaient généralement faibles, puisqu'ils étaient inférieurs à deux points de pourcentage.

Tableau 2.3
Caractéristiques et résultats de l'étude sélectionnés, par étude

Étude expérimentale	Objectif statistique	Intervention	Résultats
Peytchev et coll. (2010)	Égaliser les propensions à répondre	Prime pour les intervieweurs qui terminent des cas hautement prioritaires	• Variance des propensions à répondre plus faible dans le groupe expérimental • Taux de réponse supérieur de 1,5 % dans le groupe expérimental • Biais estimé *plus élevé* dans le groupe expérimental
Wagner et coll. (2012)	Augmenter les taux de réponse, améliorer la représentativité	Priorisation des cas Semaine de présélection	• Augmentation significative du nombre d'appels aux cas prioritaires dans 7 des 16 expériences • Taux de réponse significativement accru dans deux expériences • Augmentation du nombre d'appels de présélection
Laflamme et St-Jean (2011)	Augmenter les taux de réponse (phase 2), égaliser les propensions à répondre (phase 3)	Catégorisation et priorisation des cas	• Variance des propensions à répondre plus faible dans le groupe expérimental • Taux de réponse supérieur de 1,5 % dans le groupe expérimental
Wagner (2013)	Augmenter le taux de contact par appel	Modèles utilisés pour attribuer les cas au créneau d'appel optimal	ENSC • Amélioration du taux de contact (12,0 par rapport à 9,9 %) • Pas de changement du taux de réponse NSFG • Les intervieweurs n'ont pas suivi la recommandation de créneau d'appel
Luiten et Schouten (2013)	Égaliser les propensions à répondre	Le mode initial (courrier contre Web) variait selon le quartile de propension Les cas difficiles sont attribués aux meilleurs intervieweurs par téléphone, les cas les plus faciles aux pires intervieweurs téléphoniques	• Taux de coopération plus faible dans le groupe adaptatif • Amélioration significative de l'indicateur R dans le groupe adaptatif • Réduction de la variation des taux de contact et de coopération dans le groupe adaptatif • Pas de différence significative dans les coûts ni dans les taux de réponse
Étude par simulations	**Objectif statistique**	**Intervention**	**Résultats**
Lundquist et Särndal (2013)	Augmenter l'équilibre de l'échantillon, réduire le biais de non-réponse	Arrêt de la collecte de données pour un sous-groupe une fois le taux cible atteint pour ce sous-groupe	• Le seuil de taux de réponse le plus bas produit l'équilibre le plus élevé • Le seuil le plus bas produit également le biais de non-réponse le plus faible (sur trois variables du registre)
Särndal et Lundquist (2014a, b)	Augmenter l'équilibre de l'échantillon, réduire le biais de non-réponse	12 règles d'arrêt	• Le seuil de taux de réponse le plus bas produit encore l'équilibre le plus élevé • Le seuil le plus bas produit également le biais de non-réponse le plus faible pour trois variables du registre • L'équilibre dans la collecte des données et le calage réduisent tous deux le biais de non-réponse
Beaumont, Bocci et Haziza (2014)	Effort optimal, égalisation des taux de réponse, maximisation du taux de réponse global	Quatre stratégies de priorisation des cas (effort constant, réduction de l'effort pour les groupes qui approchent du taux de réponse cible, priorisation des cas à faible propension, priorisation des cas à propension élevée)	• Avec une propension à répondre uniforme, les quatre stratégies produisent des estimations sans biais • Quand les propensions à répondre sont fortement liées aux variables d'enquête, toutes les stratégies produisent des estimations biaisées • Avec des propensions connues, la stratégie optimale produit la meilleure racine de l'erreur quadratique moyenne (REQM); la maximisation du taux de réponse produit la pire REQM

3. Itinéraire optimal

La priorisation des cas, comme les plans adaptatifs en général, est un exemple d'intervention dans le processus de collecte des données visant à réduire les erreurs, les coûts, ou les deux. Dans les deux sections qui suivent, nous examinons deux autres interventions conçues pour améliorer les résultats de la collecte de données : l'itinéraire optimal et la rétroaction rapide aux intervieweurs. La première utilise une variation de la priorisation des cas; la seconde se concentre sur la réduction de l'erreur de mesure.

Prioriser les cas à valeur élevée. L'un des problèmes des études sur la priorisation des cas est qu'elles utilisent toutes les propensions à répondre estimées comme base de la priorisation. La propension à répondre peut se révéler un résumé utile des variables utilisées dans la modélisation de la propension, mais il se peut qu'elle ne reflète pas pleinement les priorités des chercheurs. Dans un article précédent (Tourangeau et coll., 2017), nous avons proposé une autre base de la priorisation des cas. Dans notre plan, les cas qui reçoivent la priorité la plus élevée doivent être ceux présentant le ratio valeur anticipée sur coût anticipé le plus élevé :

$$B_i = \frac{\hat{\rho}_i W_i V_i}{C_i} \tag{3.1}$$

où B_i représente le ratio avantages-coûts du cas i; le numérateur est le produit de la propension à répondre estimée du cas $(\hat{\rho}_i)$, son poids (W_i), et d'une mesure quelconque de sa valeur pour la recherche (V_i); et le dénominateur représente le coût probable de l'achèvement du cas (C_i). Par exemple, la valeur attribuée à un membre de sous-groupe rare peut être supérieure à celle attribuée à un membre de grand groupe. Ou encore, la valeur d'un cas peut être l'estimation de son effet sur la réduction de la distance entre l'échantillon actuel et un vecteur de repères démographiques. Parce qu'il inclut la propension estimée et le poids dans le numérateur, le plan de (3.1) peut donner la priorité aux cas « faciles » ou donner une priorité moindre aux cas des sous-groupes suréchantillonnés, qui auraient des poids plus faibles. Par conséquent, beaucoup dépend de la façon dont la valeur d'un cas est évaluée. Nous avons essayé d'appliquer une version de (3.1) à l'essai pilote d'une stratégie que nous appelons *itinéraire optimal*.

Essai pilote. Notre essai pilote a été effectué dans le cadre de l'étude sur la fiabilité et la viabilité de la *Population Assessment of Tobacco and Health* (PATH ou Évaluation de la population concernant le tabac et la santé) (étude PATH-RV; Tourangeau, Yan, Sun, Hyland et Stanton, 2019), conçue pour évaluer la fiabilité et la validité des réponses aux questionnaires de la vague 4 de l'étude PATH. (L'étude PATH est une grande étude longitudinale sur le tabagisme, et les commanditaires de l'étude voulaient s'assurer que les questions donnaient des réponses fiables.) Dans l'étude PATH-RV, un échantillon de 524 répondants a répondu aux questions de PATH à deux reprises, à environ deux semaines d'intervalle. Deux questionnaires étaient proposés : un pour les adultes (âgés de 18 ans et plus) et un pour les jeunes (de 12 à 17 ans). Compte tenu des objectifs de l'étude de fiabilité, nous avons jugé que les cas de jeunes avaient deux fois plus de valeur que les cas d'adultes (parce que les jeunes étaient plus rares et qu'il était plus

difficile de les interviewer que les adultes) et que la valeur d'une deuxième interview était 1,5 fois plus élevée que celle d'une première interview. Ainsi, une première interview d'adulte valait 1; une deuxième interview d'adulte valait 1,5; une première interview de jeune valait 2; et une deuxième interview de jeune valait 3. Nous avons utilisé ces valeurs à la place de V_i dans (3.1). Puisque l'échantillon de l'étude PATH-RV était presque à probabilité égale, nous avons ignoré les poids. Cependant, nous avons inclus l'estimation de la probabilité que le cas coopère à la tentative de contact suivante. Nous avons également mis au point un programme qui calculait un itinéraire optimal pour communiquer avec un ensemble de cas un jour donné, en partie pour minimiser les déplacements et le temps de l'intervieweur, c'est-à-dire pour minimiser C_i dans (3.1).

Le système que nous avons élaboré comportait deux composantes. La première a estimé la probabilité que chaque cas restant coopère à la tentative de contact suivante. Les modèles de cette première composante ont utilisé de l'information sociodémographique au niveau du groupe d'îlots de la Base de données de planification du recensement et l'historique des tentatives de contact antérieures chaque fois qu'au moins une tentative de contact était disponible. Pour ce qui est des cas sans tentatives de contact antérieures, nous avons utilisé un modèle de régression logistique afin d'estimer la propension à la coopération. Pour les cas avec tentatives de contact antérieures, nous avons utilisé un modèle de régression à risques proportionnels de Cox. La deuxième composante était un système d'itinéraire qui examinait l'information au niveau du répondant et produisait l'échéancier d'un intervieweur pour une journée donnée. La composition de l'ensemble de cas attribués pour la journée s'expliquait par la valeur anticipée des cas. La géolocalisation de tous les cas de l'échantillon a permis d'estimer le temps de déplacement entre chaque paire de cas pour une heure donnée de la journée. Le système d'itinéraire incluait comme données d'entrée les tâches réalisables pour un cas donné (par exemple le système ne prévoyait pas de deuxième interview avant la réalisation de la première interview), ainsi que l'emplacement géographique du cas, la durée estimée de la tâche, la valeur du cas et la propension à répondre. Il calculait ensuite le parcours le plus court ayant la valeur espérée la plus élevée possible et il sélectionnait un ensemble de tâches qui pourraient convenir à la journée de travail d'un intervieweur donné. L'itinéraire fourni aux intervieweurs comprenait la séquence des cas et des tâches que nous demandions aux intervieweurs de suivre. Il tenait compte des rendez-vous et était conçu de façon à ce que l'intervieweur puisse y arriver à l'heure.

Le plan expérimental. Nous avons réalisé une expérience qui comparait le rendement des intervieweurs entre les jours de « traitement » pour lesquels nous leur avions donné la liste des cas qu'ils devaient contacter en essayant de respecter l'itinéraire proposé et les jours « témoins » pour lesquels nous ne leur avions pas donné d'instructions particulières, ni sur le choix des cas ni sur la méthode de travail. La collecte de données de l'expérience s'est déroulée entre octobre et décembre 2017.

Avant le début de la collecte de données, les intervieweurs ont choisi au moins six jours pendant lesquels ils travailleraient uniquement à l'étude PATH-RV. Nous avons ensuite attribué aléatoirement trois de ces jours au volet témoin de l'expérience et trois au volet de traitement. Les jours témoins, nous

avons envoyé aux intervieweurs un courriel dans lequel nous leur demandions de « faire preuve de jugement sur la façon de communiquer » avec les cas de leur volume de travail. Les jours de traitement, nous leur avons envoyé un courriel contenant la liste des cas sur lesquels nous voulions qu'ils travaillent et l'itinéraire qu'ils devaient suivre. Il était demandé aux intervieweurs de suivre nos recommandations « dans la mesure du possible ». Pendant les séances de formation et dans le courriel accompagnant l'itinéraire choisi, nous avons déconseillé les dérogations aux instructions, mais les avons autorisées si les intervieweurs les jugeaient nécessaires pour tenir compte d'imprévus, par exemple un accident de la route.

Cinquante-trois intervieweurs ont participé à l'expérience. Des changements apportés aux jours où les intervieweurs ont travaillé à notre étude ainsi que l'épuisement du bassin de cas ouverts dans les derniers jours de l'étude ont entraîné une réduction du nombre de jours de traitement et de jours témoins réalisables par les intervieweurs. Nous avons finalement obtenu 220 observations.

Conformité et efficacité des intervieweurs. Les intervieweurs ont-ils suivi les instructions envoyées dans le courriel concernant les jours de traitement ? Eh bien, la réponse est : parfois. Nous avons constaté un chevauchement moyen de 62 % entre les cas que nous avons recommandés pour un jour de traitement donné et les cas sur lesquels les intervieweurs travaillaient ce jour-là. Il est particulièrement frappant qu'en moyenne, nous ayons obtenu un chevauchement de 52 % entre les cas sélectionnés par notre modèle et les cas sélectionnés par les intervieweurs les jours témoins, alors que nous ne leur avions donné aucune instruction. Cette faible différence entre les jours de traitement et les jours témoins s'explique en partie par le nombre limité de cas traitables en une journée donnée. C'est pourquoi les décisions prises par les intervieweurs de leur propre chef étaient souvent proches de ce que nous pensions être optimal, ce qui plafonne vers le bas l'effet possible du traitement.

Notons toutefois que les intervieweurs ont modérément respecté les instructions. Un essai mené par le Bureau du recensement a donné des résultats similaires. Il a été réalisé dans huit zones de Philadelphie, en Pennsylvanie (Walejko et Miller, 2015). Dans certaines zones, sept cas hautement prioritaires étaient attribués aux intervieweurs chaque jour. Ces cas hautement prioritaires étaient ceux jugés les plus susceptibles d'être interviewés lors de la prochaine tentative de contact, selon des modèles de propension à répondre. Comme dans notre expérience, la conformité des intervieweurs aux instructions posait problème. Comme l'écrivent Walejko et Miller (2015) : « La capacité des modèles de propension à répondre à déterminer les cas prometteurs pour un contact quotidien demeure toutefois incertaine après l'essai pilote étant donné que les intervieweurs n'ont pas travaillé scrupuleusement sur les cas prioritaires. »

Notre étude a-t-elle révélé que le traitement par itinéraire optimal améliorait l'efficacité de l'intervieweur ? Nous avons examiné cinq résultats d'intérêt :

1) le nombre de milles terrestres parcourus par les intervieweurs;
2) les heures qu'ils ont passées à réaliser leur travail;
3) le nombre de contacts par interview terminée;
4) le nombre de cas terminés;
5) la valeur moyenne des cas terminés.

Les deux premières variables correspondent à l'effet du traitement sur les coûts de la collecte. Nous voulions aussi déterminer si notre système d'itinéraire réduisait le nombre de tentatives de contact nécessaires pour terminer un cas, c'est-à-dire si l'itinéraire augmentait la productivité des intervieweurs. De même, nous avons examiné si le traitement augmentait le nombre de cas terminés et si ces derniers avaient des valeurs plus élevées en moyenne les jours de traitement comparativement aux jours témoins. Nos analyses des effets du traitement sont présentées dans le tableau 3.1. Les modèles comprennent des effets aléatoires pour chaque intervieweur et regroupent l'effet du traitement pour tous les intervieweurs. Les deux volets supérieurs montrent les estimations des effets de l'ordonnée à l'origine et du traitement selon un modèle en intention de traiter (qui ignore si les intervieweurs ont réellement suivi nos instructions), et le deuxième volet comprend une mesure de la conformité des intervieweurs aux instructions. Aucune des mesures des résultats ne montre que le traitement a un effet significatif, bien qu'il y ait eu un effet principal significatif de la conformité pour ce qui est des milles et des contacts : les intervieweurs ont parcouru moins de milles et ont contacté moins de cas quand ils suivaient nos instructions (qu'elles leur aient été transmises ou non). Bien qu'il y ait eu un traitement par effet d'interaction de la conformité sur les contacts, l'effet net des traitements semble avoir été négatif.

Tableau 3.1
Estimations des ordonnées à l'origine et des effets (et erreurs-types), par résultat et modèle

	Milles	Heures	Contacts	Cas terminés	Valeur
Intention de traiter					
Ordonnée à l'origine	76,8 (7,0)	5,26 (0,37)	5,40 (0,53)	0,81 (0,17)	1,73 (0,26)
Traitement	8,06 (5,6)	0,04 (0,27)	-0,28 (0,48)	-0,15 (0,16)	-0,24 (0,30)
Incorporation de la conformité					
Ordonnée à l'origine	89,2 (8,99)	5,69 (0,46)	7,66 (0,62)	0,98 (0,21)	1,84 (0,40)
Traitement	0,30 (9,87)	-0,30 (0,48)	-1,35 (0,79)	0,15 (0,28)	0,25 (0,53)
Conformité	-28,9 (13,4)	-1,01 (0,65)	-5,32 (1,06)	-0,03 (0,37)	-0,28 (0,70)
Traitement x conformité	20,5 (17,4)	0,86 (0,85)	3,07 (1,38)	-0,55 (0,48)	-0,87 (0,92)

Note : Résultats fondés sur 53 intervieweurs et 220 observations au total.

Réaction des intervieweurs. Les séances de compte rendu avec les intervieweurs ont révélé certaines des raisons du degré relativement faible de conformité à nos recommandations. Bien que les intervieweurs aient généralement une opinion positive au sujet du système d'itinéraire, ils ont émis plusieurs réserves. Le comportement des intervieweurs traduit que leur objectif est d'obtenir des interviews terminées, mais selon leur hypothèse implicite que toutes les interviews terminées sont de valeur égale. Or notre système d'itinéraire correspondait à une définition précise de la valeur espérée d'un cas ainsi que l'estimation de son coût. C'est pourquoi il omettait parfois des cas qui se trouvaient près des ménages de l'itinéraire recommandé. Les intervieweurs ont indiqué qu'ils auraient mieux compris les décisions du système automatique s'ils avaient eu à leur disposition la liste des priorités ou la notation des cas en fonction de leur valeur; cela leur aurait également permis d'intégrer ces valeurs dans leur propre planification de leur journée de travail. De plus, les intervieweurs désapprouvaient parfois les itinéraires proposés en raison de

circonstances qui ne pouvaient être observées par notre système d'itinéraire. Dans toute stratégie de plan adaptatif (ou, de façon plus générale, dans tout système de planification), il existe un risque de manquer certains renseignements utiles, qui peut nuire à la conformité.

Les comptes rendus ont également attiré l'attention sur certaines des hypothèses intégrées du modèle. Par exemple, nous avions établi un créneau unique pour tous les intervieweurs comme période de travail la plus probable. Cela nous permettait de tenir compte de la circulation routière quotidienne dans nos recommandations. Cependant, parfois, il aurait été préférable de choisir un autre parcours que l'itinéraire recommandé, à un autre moment de la journée pendant lequel la circulation est moindre ou plus dense. Tous les intervieweurs qui ont participé à l'expérience étaient des intervieweurs ayant une grande expérience du terrain, et certains ont déclaré que, selon eux, les instructions détaillées d'itinéraire équivalaient à ignorer leurs capacités et leur expérience. Ils estimaient que le système pourrait être un bon outil pour les intervieweurs débutants, mais que le leur proposer indiquait un manque de confiance de la part des gestionnaires d'enquête. Enfin, ils ont tous déclaré qu'une des raisons pour lesquelles ils travaillaient comme intervieweurs sur le terrain était la possibilité de planifier eux-mêmes leur propre journée de travail. Indubitablement, plusieurs de ces facteurs ont joué un rôle dans le maigre succès des tentatives de Wagner et de ses collaborateurs (voir la section 2.3) et de l'essai du recensement qui visaient à modifier le comportement des intervieweurs.

Malgré ces obstacles à la conformité, des recherches ont montré que les intervieweurs sont sensibles aux incitatifs. Tourangeau, Kreuter et Eckman (2012) ont montré que les intervieweurs d'une étude téléphonique effectuaient plus de présélections quand ils recevaient une prime par présélection terminée, et qu'ils menaient plus d'interviews principales quand ils recevaient une prime par interview principale terminée. Des incitatifs similaires pourraient encourager les intervieweurs à traiter les cas hautement prioritaires ou à réduire leur temps de déplacement. Par exemple, les intervieweurs pourraient recevoir une petite prime pour chaque cas hautement prioritaire contacté. De toute évidence, il nous faut trouver les moyens de faire respecter nos instructions d'interviews si nous souhaitons que nos interventions aient un impact.

Autres études sur les déplacements des intervieweurs. Plus récemment, Wagner et Olson (2018) ont effectué une analyse approfondie des déplacements des intervieweurs dans le cadre de deux enquêtes en personne, soit la *National Survey of Family Growth* (NSFG) et la *Health and Retirement Survey* (HRS ou Enquête sur la santé et la retraite). Les deux enquêtes comportent des échantillons probabilistes aréolaires nationaux et le Centre de recherche sur les sondages de l'University of Michigan a réalisé le travail sur le terrain pour les deux enquêtes. Les populations cibles des enquêtes sont différentes : de 15 à 44 ans dans la NSFG et de 51 à 56 ans dans la HRS. Les auteurs ont examiné la distance parcourue par les intervieweurs et le nombre de zones de l'échantillon visitées pendant chaque journée de travail. Dans les deux études, les intervieweurs ont visité, en moyenne, environ deux régions par journée de travail, mais ils ont parcouru en moyenne environ 30 milles de plus dans la NSFG que dans la HRS (85,4 par rapport à 53,4). Wagner et Olson ont constaté qu'un plus grand nombre de zones visitées était associé à un plus

grand nombre de tentatives de contact, mais à un nombre inférieur de contacts et d'interviews terminées (voir le tableau 4.1). Bien que leur étude soit une étude d'observation et non une expérience, ses résultats sont conformes aux résultats de notre étude pilote : un plus grand nombre de déplacements semble réduire le nombre de contacts effectués et d'interviews réalisées. Toutefois, la direction de la causalité dans cette conclusion est assez ambiguë. Il se pourrait que le temps de déplacement réduise le temps restant aux intervieweurs pour contacter et interviewer les cas de l'échantillon, mais il se pourrait aussi que les intervieweurs continuent de travailler quand leurs tentatives de contact sont infructueuses et se rendent par conséquent dans d'autres zones d'échantillonnage.

4. Rétroaction rapide des IEAO

Les intervieweurs peuvent contribuer de plusieurs manières à l'erreur totale d'une estimation d'enquête, ce qui a une incidence sur les erreurs de couverture, dues la non-réponse et de mesure (Schaeffer, Dykema et Maynard, 2010; West et Blom, 2017). Les interactions entre ces différentes sources d'erreurs liées aux intervieweurs peuvent être complexes. Par exemple, il peut y avoir un compromis entre les erreurs de couverture et celles de non-réponse (Tourangeau et coll., 2012); dans notre étude, les intervieweurs ayant les taux de réponse les plus élevés ont également trouvé les ménages les moins admissibles. Dans une série d'articles, West et ses collaborateurs (West et Olson, 2010; West, Kreuter et Jaenichen, 2013; West, Conrad, Kreuter et Mittereder, 2018) ont montré que différents intervieweurs peuvent obtenir différentes réponses en raison des différences dans les répondants qu'ils recrutent (par exemple certains intervieweurs peuvent avoir plus de facilité à recruter des répondants plus âgés), mais aussi en raison des différences dans leurs niveaux d'erreur de mesure. Comme toute personne ayant déjà écouté des enregistrements d'IEAO peut le confirmer, les intervieweurs ne s'en tiennent pas toujours au scénario et leurs improvisations suscitent parfois des réponses de piètre qualité.

Étude pilote. Après avoir écouté les enregistrements d'intervieweurs sur le terrain dans le cadre de l'essai sur le terrain d'une grande étude nationale, nous avons conçu une expérience visant à vérifier l'hypothèse selon laquelle le fait de fournir aux intervieweurs une rétroaction en temps opportun sur leur lecture des questions améliorerait la qualité des réponses obtenues. (À la demande du client, nous ne divulguons pas le titre de l'étude.) Cette enquête en particulier a été un bon banc d'essai pour évaluer les effets de la rétroaction rapide parce que les intervieweurs devaient faire répondre à un questionnaire de présélection court une personne du ménage, avant que des questions semblables soient posées à chaque membre de l'échantillon par interview audio autoadministrée assistée par ordinateur (IAAAO). Par conséquent, nous pouvions comparer les données de présélection recueillies par chaque intervieweur à un « étalon de référence » pour plusieurs des éléments clés de l'enquête. Bien entendu, les données de l'IAAAO ne sont pas exemptes d'erreurs, mais nous les avons considérées comme moins sujettes à l'erreur que les données de la présélection pour deux raisons : chaque personne y répondait pour elle-même, tandis que le questionnaire était présenté à une seule personne du ménage; et les questions étaient

autoadministrées plutôt que posées par un intervieweur; or l'autoadministration était susceptible de réduire le biais de désirabilité sociale dans les réponses.

L'expérience comprenait 291 intervieweurs. La moitié d'entre eux ont reçu une rétroaction rapide tandis que l'autre moitié a été affectée au groupe témoin. Un questionnaire de présélection sur cinq effectués par les intervieweurs du groupe de rétroaction rapide était codé par IEAO aux fins de détermination des écarts par rapport aux interviews normalisées. La figure 4.1 montre les questions auxquelles les codeurs répondent pour chaque interview de présélection. Après le codage d'une présélection, les intervieweurs (et leurs superviseurs) recevaient un rapport contenant leur rendement et un lien vers les enregistrements des questions. Pour leur première présélection codée, les intervieweurs ont reçu pour instruction de planifier une séance de rétroaction avec un « mentor » du bureau central qui, après examiné des résultats, fournit des conseils à des fins d'amélioration. Pour leur deuxième présélection codée, les intervieweurs recevaient seulement le rapport et un lien vers les enregistrements. Pour les présélections suivantes, il était demandé aux intervieweurs de prévoir une séance de rétroaction avec leur mentor seulement si le codage avait relevé des problèmes. Sinon, ils recevaient simplement le rapport écrit.

Figure 4.1 Questions de codage pour l'étude pilote de rétroaction rapide. Les questions ont été répétées à chaque membre du ménage.

Q1. Dans quelle mesure entendez-vous clairement l'intervieweur sur cet enregistrement ? [HEARINT]

- o Très clairement (4)
- o Assez clairement (3)
- o Pas très clairement (2)
- o Impossible d'entendre l'intervieweur (1)

Q2. Dans quelle mesure entendez-vous clairement le répondant sur cet enregistrement ? [HEARRESP]

- o Très clairement (4)
- o Assez clairement (3)
- o Pas très clairement (2)
- o Impossible d'entendre l'intervieweur (1)

Q3. L'intervieweur a-t-il lu la question exactement comme elle est formulée ? [EXWORD]

- o Oui (1)
- o Non (2)

Q4. [SI NON À Q3] De quelle manière l'intervieweur a-t-il modifié le libellé de la question ? Sélectionnez toutes les réponses pertinentes

- o N'a pas lu le texte de préambule ou d'introduction avant la question [NOINTRO]
- o N'a pas lu « Regardez cette photo s'il vous plaît » [NOPIC]
- o N'a pas lu « Regardez cette liste s'il vous plaît » [NOLIST]
- o N'a pas lu tous les noms de marque ni tous les exemples de produits [NONAMES]
- o N'a pas lu correctement les choix de réponse [NORESP]
- o N'a pas lu « choisissez toutes les réponses qui s'appliquent » [NOCHOOSE]
- o Omission, ajout ou modification de mots dans la question [NOREADOTH]

Q5. L'intervieweur a-t-il entré correctement la réponse du répondant ? [ENTERANS]

- o Oui (1)
- o Non (2)

L'expérience a été menée de mai à août 2014; 1 729 répondants ont été interviewés par le groupe de rétroaction et 1 717 par le groupe témoin.

Pour évaluer les effets de la rétroaction rapide, nous avons comparé trois variables dérivées des éléments de préselection aux variables dérivées correspondantes des interviews IAAAO. En principe, les deux types de variables devraient correspondre. Le tableau 4.1 montre la proportion de répondants des groupes de traitement et témoin qui ont été classés de la même façon dans les données de la présélection et les données des IAAAO. Pour les trois variables, le taux d'appariement était beaucoup plus élevé chez les répondants interviewés par les intervieweurs qui ont reçu une rétroaction. (Nous avons utilisé le test F de Rao-Scott qui tenait compte de la corrélation intra-grappe de l'échantillon par zones. Les trois valeurs F étaient significatives à $p < 0{,}01$.) Les valeurs du kappa mesurant la concordance corrigée du hasard entre les réponses de présélection et celles des IAAAO sont aussi nettement plus élevées pour les intervieweurs du groupe de rétroaction rapide.

Tableau 4.1
Concordance entre les réponses de présélection et celles de l'IAAAO, selon la condition et la variable

	Rétroaction rapide	Témoin	Valeur F de Rao-Scott (1 et 230 df)
Composite			
Concordance à 100 %	93 %	88 %	15,5***
Kappa	0,85	0,76	
Variable 1			
Concordance à 100 %	95 %	92 %	8,8**
Kappa	0,89	0,83	
Variable 2			
Concordance à 100 %	89 %	85 %	7,7**
Kappa	0,76	0,69	

** $p < 0{,}01$;

*** $p < 0{,}001$.

Note : La variable composite était une variable sommaire dérivée des variables 1 et 2.

L'étude MEPS. En se fondant sur la réussite de cette première étude, Edwards, Sun et Hubbard (2019) ont entrepris une reproduction. En 2018, la *Medical Expenditure Panel survey* (MEPS ou Enquête sur les dépenses médicales par panel) a réalisé une mise à niveau importante du système d'IPAO et a simplifié certaines sections du questionnaire. Deux séries de questions étaient particulièrement intéressantes parce qu'elles étaient posées dans toutes les interviews, toujours enregistrées dans le système d'IEAO (presque tous les répondants consentent à être enregistrés) et étaient essentielles à la production de données concernant l'utilisation et le coût des services de soins de soins santé, qui sont des statistiques clés de la MEPS. Il s'agissait des questions concernant l'utilisation de calendriers ou d'autres dossiers de soins médicaux pendant l'interview et des « questions d'approfondissement sur le fournisseur », des questions filtres qui incitent le répondant à se rappeler des services de différents types de fournisseurs de soins médicaux. La série de questions sur les calendriers demandait si divers dossiers étaient disponibles au cours de l'interview (par exemple un calendrier comportant des indications de visites médicales, des

relevés d'assurance, etc.) et quelle personne du ménage était associée à chaque type d'enregistrement. La zone de saisie de l'IPAO pour ces éléments est une grille, dans laquelle chaque membre du ménage est indiqué sur une ligne et chaque type de dossier est un en-tête de colonne. Les intervieweurs peuvent entrer les réponses dans l'ordre qu'ils choisissent, par personne ou par type de dossier. L'objectif était d'encourager les répondants à apporter des dossiers pour tous les membres de la famille dans l'interview et de structurer les questions de façon à ce que les dossiers puissent être intégrés à l'interview sans que l'ordre importe. Les questions d'approfondissement sur le fournisseur consistent en 15 questions sur plusieurs types de fournisseurs de soins de santé. Elles ont été réorganisées dans la mise à niveau technique pour que l'intervieweur commence par les trois fournisseurs représentant les dépenses les plus élevées.

Les enregistrements audio de la série des calendriers et de la série des questions d'approfondissement sur le fournisseur ont été examinés par deux codeurs de comportements. Le système de codage permettait aux codeurs d'appeler des intervieweurs ou des questions en particulier. Les codeurs évaluaient la qualité globale de l'interview et de chaque cas de demande de calendriers et de questions d'approfondissement du fournisseur. Le taux de concordance entre codeurs était de 0,82. Une rétroaction verbale et écrite était fournie rapidement à l'intervieweur (idéalement dans les 72 heures après l'interview). L'interview suivante réalisée par l'intervieweur était également codée, de sorte que chaque intervieweur avait une paire d'interviews dans l'ensemble de données, une interview juste avant la rétroaction et une juste après. Étant donné que le processus a été mis en œuvre à la fin de l'automne, seul un sous-ensemble d'intervieweurs (122) de la MEPS était disponible pour participer à l'étude, ce qui a donné 244 interviews dans l'ensemble de données. Des données sur l'interaction de la rétroaction ont également été saisies (notamment sur l'approbation de la rétroaction ou sa non-approbation par l'intervieweur ou s'il demandait des précisions). Encore une fois, nous nous attendions à ce que le comportement de l'intervieweur soit plus conforme au protocole de l'étude après la rétroaction, tant en matière de qualité globale de l'interview que pour chaque série de questions.

Le tableau 4.2 montre les résultats de la rétroaction rapide pour chaque série de questions. Les intervieweurs ont conservé le sens des questions, mais ils n'ont pas suivi exactement le protocole dans la majorité des cas $(n = 5\,259)$, avant et après la rétroaction. Néanmoins, le comportement de questionnement conforme au protocole est passé de 33,4 % avant la rétroaction à 43,4 % après la rétroaction; et la non-conservation du sens de la question a diminué pour passer de 9,8 % avant la rétroaction à 3,7 % après la rétroaction. Un test F qui tenait compte de la corrélation intra-grappe des observations par l'intervieweur a révélé une différence globale significative entre le comportement de l'intervieweur avant et après la rétroaction, tant dans l'ensemble $(F\,(2\,118) = 3{,}86;\ p < 0{,}05)$ que pour les questions d'approfondissement sur le fournisseur $(F\,(2\,118) = 5{,}71;\ p < 0{,}01)$. Les différences pour la série des calendriers allaient dans le même sens, mais elles n'étaient pas statistiquement significatives. Ces résultats, comme ceux de l'étude pilote, indiquent que le fait de fournir rapidement une rétroaction aux intervieweurs peut entraîner des améliorations marquées dans la façon dont ils posent les questions.

Tableau 4.2
Comportements des intervieweurs, avant et après la rétroaction, par série de questions

Comportement de l'intervieweur	Série des calendriers		Questions d'approfondissement sur le fournisseur		Les deux séries	
	Avant	Après	Avant	Après	Avant	Après
A suivi exactement le protocole	18,6 %	27,9 %	43,3 %	51,7 %	33,4 %	43,4 %
A conservé le sens, mais n'a pas suivi le protocole exactement	68,9 %	65,1 %	48,7 %	46,4 %	56,8 %	52,9 %
N'a pas conservé le sens	12,5 %	7,0 %	8,0 %	2,0 %	9,8 %	3,7 %
Total	100,0 %	100,0 %	100,0 %	100,0 %	100,0 %	100,0 %
n	1 240	759	1 832	1 428	3 072	2 187

5. Conclusions

Les trois principales méthodes étudiées ici ont connu un succès mitigé. Quelles leçons pouvons-nous tirer des efforts cherchant à substituer les données à l'intuition dans la gestion des enquêtes ?

La documentation sur les plans adaptatifs et réactifs conduit à plusieurs conclusions. Premièrement, il est important de clarifier les objectifs statistiques du plan au début de l'enquête et de surveiller les mesures de la qualité liées à ces objectifs. En effet, différentes stratégies servent des objectifs différents. Par exemple, l'égalisation des propensions à répondre peut réduire le biais de non-réponse au prix d'une réduction de la taille d'échantillon et d'une augmentation de la variance due à l'échantillonnage. Il est essentiel de tenir compte de ces compromis. Deuxièmement, le taux de réponse global et la variation des propensions à répondre contribuent tous deux au biais de non-réponse moyen. Par conséquent, aucun indicateur ne donne d'image complète du risque d'erreur dans une enquête, et les gestionnaires d'enquête doivent surveiller plusieurs indicateurs, notamment les changements dans un ensemble d'estimations clés de l'enquête. Les progrès réalisés dans la conception du « tableau de bord » (Mohadjer et Edwards, 2018) permettent au personnel du bureau central et aux superviseurs sur le terrain de surveiller plus facilement un grand nombre d'indicateurs concernant le déroulement du travail sur le terrain. Troisièmement, le fait de simplement continuer à utiliser un protocole de collecte de données en particulier pourrait ne pas modifier les estimations de beaucoup (Sturgis et coll., 2017) et, dans certains cas, pourrait diminuer la représentativité de l'échantillon (Lundquist et Särndal, 2013; Särndal et Lundquist, 2014). Selon un protocole donné de collecte de données, les répondants recrutés tardivement dans la période de travail sur le terrain ne diffèrent probablement pas beaucoup de ceux recrutés plus tôt. L'échantillon continuera de surreprésenter les cas présentant des propensions plus élevées selon ce protocole. Pour modifier la composition des répondants, et ainsi améliorer la représentativité globale de l'échantillon, il faut parfois modifier considérablement le protocole de collecte des données, par exemple en proposant des incitatifs nettement plus grands, en changeant le mode de collecte des données ou en présentant un questionnaire nettement plus court. Ces stratégies comportant toutes des inconvénients, on peut en conclure que, parfois, la meilleure stratégie consiste simplement à cesser les efforts en imposant des règles d'arrêt. Continuer à chercher à traiter des cas ayant une très faible propension à répondre est une formule qui accroît les coûts sans vraiment améliorer les propriétés statistiques des estimations finales.

La littérature sur les plans réactifs et adaptatifs ainsi que l'étude sur la priorisation des cas et les itinéraires optimaux dont il est question à la section 3 plus haut indiquent que la résistance des intervieweurs est un des facteurs limitant l'efficacité des interventions du bureau central sur le travail sur le terrain. Il nous faut plus de recherches sur la façon d'améliorer la conformité des intervieweurs et sur l'effet d'une surveillance plus étroite (ou d'incitatifs plus grands) sur le respect de la part des intervieweurs des changements souhaités au protocole. À cet égard, les études s'intéressant aux rétroactions données rapidement aux intervieweurs sont encourageantes. Les deux études examinées à la section 4 indiquent que les intervieweurs recevant une rétroaction en temps opportun sur la façon de poser les questions accomplissent un meilleur travail, ce qui réduit le niveau d'erreur de mesure dans les réponses qu'ils obtiennent.

En tout cas, une chose est certaine. Dans un climat de plus en plus difficile pour les enquêtes, les efforts cherchant à améliorer la gestion des enquêtes et à appliquer le plus possible des données scientifiques à cet égard se poursuivront.

Remerciements

L'article a été rédigé en réponse à la réception du prix Waksberg. Je remercie Brad Edwards, Gonzalo Rivero et Tammy Cook pour leurs précieuses suggestions sur l'article, Aaron Maitland et Gonzalo Rivero de leur aide dans la conception et la réalisation de certaines des études citées, et Statistique Canada de m'avoir décerné ce prix.

Bibliographie

Atrostic, B.K., Bates, N., Burt, G. et Silberstein, A. (2001). Nonresponse in U.S. government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics*, 17, 209-226.

Axinn, W.G., Link, C.E. et Groves, R.M. (2011). Responsive survey design, demographic data collection, and models of demographic behavior. *Demography*, 48, 1127-1149.

Beaumont, J.-F., Bocci, C. et Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-621.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Brick, J.M., et Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33, 735-752.

Brick, J.M., et Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, 645, 36-59.

Calinescu, M., Bhulai, S. et Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, 226, 115-121.

Chun, A.Y., Heeringa, S.G. et Schouten, B. (2018). Responsive and adaptive design for survey optimization. *Journal of Official Statistics*, 34, 581-597.

de Leeuw, E.D., et de Heer W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. Dans *Survey Nonresponse*, (Éds., R.M. Groves, D. Dillman, J.L. Eltinge et R.J.A. Little), New York: John Wiley & Sons, Inc, 41-54.

Edwards, B., Sun, H. et Hubbard, R. (2019). Behavior change techniques for reducing interviewer contributions to total survey error. Manuscrit non publié.

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

Groves, R.M., et Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, 72, 167-189.

Groves, R.M., Singer, E. et Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64, 299-308.

Groves, R.M., Benson, G., Mosher, W.D., Rosenbaum, J., Granda, P., Axinn, W., Lepkowski, J. et Chandra, A. (2005). *Plan and Operation of Cycle 6 of the National Survey of Family Growth*. Hyattsville: National Center for Health Statistics.

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

Hicks, W.D., Edwards, B., Tourangeau, K., McBride, B., Harris-Kojetin, L.D. et Moss, A.J. (2010). Using CARI tools to understand measurement error. *Public Opinion Quarterly*, 74, 985-1003.

Hubbard, R. (2018). That wasn't part of the plan! Reducing effort through stopping rules to place CAPI cases on hold and work plans to set them free. Article présenté à la conférence annuelle de l'American Association for Public Opinion Research, Denver, Colorado, 17 mai 2018.

Laflamme, F., et Karaganis, M. (2010). Development and implementation of responsive design for CATI surveys at Statistics Canada. Communication présentée à l'European Quality Conference, Helsinki, Finlande.

Laflamme, F., et St-Jean, H. (2011). Highlights and lessons from the first two pilots of responsive collection design for CATI surveys. Dans *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, 1617-1628.

Lepkowski, J.M., Mosher, W.D., Groves, R.M., West, B.T., Wagner, J. et Gu, H. (2013). *Responsive Design, Weighting, and Variance Estimation in the 2006–2010 National Survey of Family Growth*. Vital and Health Statistics, Séries 2, No. 158. Hyattsville, MD: National Center for Health Statistics.

Luiten, A., et Schouten, B. (2013). Tailored fieldwork design to increase representative household survey response: An experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society A*, 176, 169-189.

Lundquist, P., et Särndal, C.-E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29, 557-582.

Mohadjer, L., et Edwards, B. (2018). Paradata and dashboards in PIAAC. *Quality Assurance in Education*, 26, 263-277.

Peytchev, A., Baxter, R.K., et Carley-Baxter, L.R. (2009). Not all survey effort is equal: Reduction of nonresponse bias and nonresponse error. *Public Opinion Quarterly*, 73, 785-806.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. et Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, 4, 21-29.

Särndal, C.-E. (2011). The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27, 1-21.

Särndal, C.-E., et Lundquist, P. (2014a). Balancing the response and adjusting estimates for nonresponse bias: Complementary activities. *Journal de la Société de Statistique*, 155, 28-50.

Särndal, C.-E., et Lundquist, P. (2014b). Accuracy in estimation with nonresponse: A function of the degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, 2, 361-387.

Schaeffer, N.C., Dykema, J. et Maynard, D.W. (2010). Interviews and interviewing. Dans *Handbook of Survey Research*, Bingley, Royaume-Uni, Emerald Group Publishing, 437-470.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 107-121. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10887-fra.pdf.

Schouten, B., Peytchev, A. et Wagner, J. (2017). *Adaptive Survey Design*, Boca Raton, FL: CRC Press.

Steeh, C., Kirgis, N., Cannon, B. et DeWitt, J. (2001). Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century. *Journal of Official Statistics*, 17, 227-247.

Sturgis, P., Williams, J., Brunton-Smith, I. et Moore, J. (2017). Fieldwork effort, response rate, and the distribution of survey outcomes: A multilevel meta-analysis. *Public Opinion Quarterly*, 81, 523-542.

Tourangeau, R. (2017). Presidential address: Paradoxes of nonresponse. *Public Opinion Quarterly*, 81, 803-814.

Tourangeau, R., Kreuter, F. et Eckman, S. (2012). Motivated underreporting in screening surveys. *Public Opinion Quarterly*, 76, 453-469.

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

Tourangeau, R., Yan, T., Sun, H., Hyland, A. et Stanton, C.A. (2019). Population Assessment of Tobacco and Health (PATH) reliability and validity study: Selected reliability and validity estimates. *Tobacco Control*, 28, 663-668.

U.S. Census Bureau (2014). *American Community Survey Design and Methodology*. Téléchargé le 19 février 2016 à partir de https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, 7, 45-55.

Wagner, J., et Olson, K. (2018). An analysis of interviewer travel and field outcomes in two field surveys. *Journal of Official Statistics*, 34, 211-237.

Wagner, J., West, B.T., Kirgis, N., Lepkowski, J.M., Axinn, W.G. et Kruger Ndiaye, S. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, 28, 477-499.

Walejko, G., et Miller, P. (2015). The 2013 census test: Piloting methods to reduce 2020 Census costs. *Survey Practice*, 8.

West, B.T., et Blom, A.G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5, 175-211.

West, B.T., et Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 1004-1026.

West, B.T., Kreuter, F. et Jaenichen, U. (2013). Interviewer effects in face-to-face surveys: A function of sampling, measurement error or nonresponse? *Journal of Official Statistics*, 29, 277-297.

West, B.T., Conrad, F.G., Kreuter, F. et Mittereder, F. (2018). Nonresponse and measurement error variance among interviewers in standardized and conversational interviewing. *Journal of Survey Statistics and Methodology*, 6, 335-359.

Williams, D., et Brick, J.M. (2018). Trends in U.S. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, 6, 186-211.

Techniques d'enquête, juin 2021
Vol. 47, nº 1, pp. 33-64
Statistique Canada, nº 12-001-X au catalogue

Intégration de données d'enquêtes probabilistes et de mégadonnées aux fins d'inférence de population finie au moyen d'une imputation massive

Shu Yang, Jae Kwang Kim et Youngdeok Hwang[1]

Résumé

À l'ère des mégadonnées, on dispose d'un nombre croissant de sources de données pour les analyses statistiques. Comme exemple important d'inférence de population finie, nous examinons une approche d'imputation pour la combinaison de données tirées d'une enquête probabiliste et de mégadonnées. Nous nous concentrons sur une situation où la variable à l'étude est observée dans les mégadonnées seulement, mais les autres variables auxiliaires sont couramment observées dans les deux sources de données. Contrairement à l'imputation habituellement utilisée pour l'analyse des données manquantes, nous créons des valeurs imputées pour toutes les unités de l'échantillon probabiliste. Une imputation massive de ce type est intéressante dans le contexte de l'intégration des données d'enquête (Kim et Rao, 2012). Nous étendons l'imputation massive comme outil d'intégration des données d'enquête et des mégadonnées ne provenant pas d'enquêtes. Nous présentons les méthodes d'imputation massive et leurs propriétés statistiques. De plus, l'estimateur d'appariement de Rivers (2007) est traité comme cas particulier. L'estimation de la variance au moyen de données obtenues par imputation massive est abordée. Les résultats de la simulation montrent que les estimateurs proposés donnent de meilleurs résultats que les estimateurs concurrents en matière de robustesse et d'efficacité.

Mots-clés : Pondération par calage; fusion de données; modèle additif généralisé; appariement; imputation par le plus proche voisin; poststratification.

1. Introduction

Dans l'inférence de population finie, l'échantillonnage probabiliste est la norme de référence pour l'obtention d'un échantillon représentatif de la population cible. Comme la probabilité de sélection est connue, l'inférence subséquente à partir d'un échantillon probabiliste est souvent fondée sur le plan et respecte la façon dont les données ont été recueillies; voir les discussions dans les manuels de Särndal, Swensson et Wretman (2003), Cochran (2007), Fuller (2009). Cependant, les programmes d'enquête à grande échelle se heurtent continuellement à des exigences accrues malgré des ressources réduites. Ils doivent notamment répondre à des demandes d'estimations pour des domaines avec une petite taille d'échantillon et des demandes d'estimations plus rapides. Simultanément, les compressions budgétaires des programmes entraînent des réductions de la taille des échantillons, et la diminution des taux de réponse fait du biais de non-réponse une préoccupation importante. Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau (2013) traitent des difficultés actuelles concernant l'utilisation d'échantillons probabilistes aux fins d'inférences de population finie.

Afin de remédier à ces difficultés, les organismes statistiques sont de plus en plus poussés à utiliser des mégadonnées qui sont des sources certes pratiques, mais non contrôlées, comme des données satellitaires

1. Shu Yang, Département de statistique, North Carolina State University, 2 311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695, États-Unis. Courriel : syang24@ncsu.edu; Jae Kwang Kim, Département de statistique, 1 208 Snedecor Hall, Iowa State University, Ames, IA 50011, États-Unis; Youngdeok Hwang, Paul H. Chook Department of Information Systems and Statistics, Baruch College, City University of New York, New York, NY 10010, États-Unis.

(McRoberts, Tomppo et Næsset, 2010), des données de capteurs mobiles (Palmer, Espenshade, Bartumeus, Chung, Ozgencil et Li, 2013) et des panels d'enquête sur le Web (Tourangeau, Conrad et Couper, 2013). Couper (2013), Citro (2014), Tam et Clarke (2015) et Pfeffermann, Eltinge et Brown (2015) expliquent les résultats prometteurs de l'exploitation des mégadonnées dans le calcul de statistiques officielles et d'enquête, mais soulèvent aussi de nombreuses interrogations concernant les sources de mégadonnées. Bien que ces sources de données fournissent rapidement des données pour un grand nombre de variables et d'éléments de population, il s'agit d'échantillons non probabilistes qui, souvent, ne représentent pas la population cible d'intérêt en raison des biais de sélection inhérents. Tam et Kim (2018) traitent également de défis d'ordre éthique posés par les mégadonnées pour les statisticiens officiels et discutent de certaines méthodes préliminaires de correction du biais de sélection des mégadonnées. Voir Keiding et Louis (2016), Elliott et Valliant (2017), Buelens, Burger et van den Brakel (2018) et Beaumont (2020) pour lire de récents examens des difficultés liées à l'utilisation d'échantillons non probabilistes dans les inférences.

Pour utiliser les sources de données modernes de façon statistiquement défendable, il est important d'élaborer des outils statistiques d'intégration des données permettant de combiner un échantillon probabiliste avec des mégadonnées non probabilistes. L'intégration des données aux fins d'inférence de population finie ressemble au problème de la combinaison d'études cliniques randomisées et d'études épidémiologiques non randomisées aux fins d'inférence causale des effets de traitement (Keiding et Louis, 2016). Plus particulièrement, nous nous sommes intéressés à l'élaboration d'une intégration des données dans un contexte où la variable de l'étude est observée dans les mégadonnées seulement, mais que d'autres variables sont couramment observées dans les deux groupes de données. Pour ce cas, les statisticiens d'enquête et les biostatisticiens ont fourni différentes méthodes permettant de combiner l'information provenant de sources de données multiples. Lohr et Raghunathan (2017), Yang et Kim (2020) et Rao (2020) présentent un examen des méthodes statistiques d'intégration des données aux fins d'inférence de population finie. Les méthodes actuelles d'intégration des données peuvent être classées en trois catégories, comme suit.

Le premier type est ce qu'on appelle ajustement du score de propension (Rosenbaum et Rubin, 1983). Dans cette méthode, la probabilité qu'une unité soit sélectionnée dans le grand échantillon, qu'on appelle score de propension, est modélisée et estimée pour toutes les unités de l'échantillon de mégadonnées. Les ajustements suivants, comme la stratification ou la pondération du score de propension, peuvent servir à tenir compte des biais de sélection; voir, par exemple, Lee et Valliant (2009), Valliant et Dever (2011), Elliott et Valliant (2017). Stuart, Bradshaw et Leaf (2015), Stuart, Cole, Bradshaw et Leaf (2011), Buchanan, Hudgens, Cole, Mollan, Sax, Daar, Adimora, Eron et Mugavero (2018) utilisent la pondération du score de propension pour généraliser les résultats des essais randomisés à une population cible. O'Muircheartaigh et Hedges (2014) proposent d'utiliser une stratification de score de propension pour

analyser une expérience sociale non randomisée. L'un des inconvénients notables des méthodes se fondant sur le score de propension est qu'elles dépendent d'un modèle du score de propension explicite et sont biaisées en cas de spécification erronée du modèle (Kang et Schafer, 2007).

Le deuxième type utilise le calage (Deville et Särndal, 1992; Kott, 2006; Dong, Yang, Wang, Zeng et Cai, 2020). Cette technique peut servir à calibrer l'information auxiliaire de l'échantillon de mégadonnées avec celle de l'échantillon probabiliste, de sorte qu'après le calage, l'échantillon de mégadonnées soit semblable à la population cible (DiSogra, Cobb, Chan et Dennis, 2011). Comme le calage ne nécessite pas de modélisation paramétrique, il est intéressant pour les spécialistes des enquêtes. Toutefois, cette méthode nécessite que l'information (par exemple les moments) des variables auxiliaires de la population soit connue ou, à tout le moins, qu'elle puisse être estimée à partir d'un échantillon probabiliste.

Le troisième type est l'imputation massive, dans laquelle les valeurs imputées sont créées pour les éléments entiers de l'échantillon probabiliste. Dans l'imputation habituelle aux fins d'analyse des données manquantes, les répondants de l'échantillon fournissent un ensemble de données d'entraînement servant à l'élaboration d'un modèle d'imputation. Dans l'imputation massive, un échantillon indépendant de mégadonnées est utilisé comme ensemble de données d'entraînement, et l'imputation est appliquée à toutes les unités de l'échantillon probabiliste. Bien que l'idée d'incorporer l'information provenant de mégadonnées par une imputation massive soit très naturelle, la littérature sur l'imputation massive elle-même est rare. Breidt, McVey et Fuller (1996) traitent de l'imputation massive pour échantillonnage à deux phases. Rivers (2007) propose une méthode d'imputation massive utilisant l'imputation par le plus proche voisin, mais la théorie n'est pas totalement développée. Kim et Rao (2012) élaborent une théorie rigoureuse d'imputation massive à l'aide de deux échantillons probabilistes indépendants. Chipperfield, Chessman et Lim (2012) discutent de l'estimation composite quand une des enquêtes fait l'objet d'une imputation massive. Bethléem (2016) discute de questions pratiques liées à l'appariement d'échantillons. Récemment, Kim et Wang (2019) ont élaboré une théorie d'imputation massive de mégadonnées à l'aide d'une approche de modèle paramétrique. Toutefois, les hypothèses du modèle paramétrique ne se vérifient pas nécessairement en pratique. Pour que l'imputation massive soit plus utile et plus pratique, les hypothèses doivent être aussi faibles que possible.

Les contributions que nous apportons dans le présent article peuvent se résumer comme suit.

1. Nous élaborons d'abord un cadre officiel d'imputation massive qui incorpore l'information de mégadonnées dans un échantillon probabiliste et nous présentons des résultats asymptotiques rigoureux pour les estimateurs d'imputation massive. Notre cadre couvre l'estimateur d'imputation par le plus proche voisin de Rivers (2007). Contrairement à Kim et Wang (2019), nous ne posons pas d'hypothèses fortes de modèle paramétrique pour l'imputation massive. C'est pourquoi la méthode proposée est intéressante pour les spécialistes des enquêtes.

2. Nous étudions également deux stratégies d'amélioration de l'estimateur d'imputation par le plus proche voisin, l'une utilisant l'imputation par les k plus proches voisins (Mack et Rosenblatt, 1979) et l'autre utilisant des modèles additifs généralisés (Wood, 2006). Dans l'imputation par les k plus proches voisins, au lieu d'utiliser un plus proche voisin, nous identifions plusieurs plus proches voisins dans l'échantillon de mégadonnées et utilisons la réponse moyenne comme valeur imputée. Cette méthode est couramment utilisée par la communauté internationale des inventaires forestiers pour combiner des observations au sol et des images obtenues à partir de capteurs à distance (McRoberts et coll., 2010). Dans l'article, nous établissons des résultats asymptotiques pour l'estimateur des k plus proches voisins. Dans la deuxième stratégie, nous étudions les techniques modernes de prédiction pour l'imputation massive au moyen de modèles souples. Nous utilisons des modèles additifs généralisés (Wood, 2006) pour connaître la relation entre le résultat et les covariables à partir des mégadonnées et créer des prévisions pour les échantillons probabilistes. Nous constatons que cette stratégie peut s'appliquer à une plus grande catégorie d'estimateurs semi-paramétriques et non paramétriques comme les modèles à indice unique, les estimateurs de Lasso (Belloni, Chernozhukov, Chetverikov et Kato, 2015) et les méthodes d'apprentissage automatique, par exemple les forêts d'arbres décisionnels (Breiman, 2001).
3. À l'aide d'une nouvelle idée de pondération par calage, nous proposons un estimateur d'imputation massive efficace et développons ses résultats asymptotiques. Le gain d'efficacité est justifié selon un cadre fondé purement sur le plan et aucune hypothèse de modèle n'est utilisée. Nous examinons un cas où il est de plus possible de déterminer l'appartenance aux mégadonnées dans tout l'échantillon probabiliste. La principale idée est que le sous-échantillon d'unités de l'échantillon A appartenant aux mégadonnées constitue un échantillon de deuxième phase de l'échantillon de mégadonnées, qui agit comme une nouvelle population. Nous calons l'information de l'échantillon de deuxième phase pour qu'elle soit identique à celle de l'échantillon agissant comme une nouvelle population. Le processus de calage améliore alors l'exactitude de l'estimateur d'imputation massive sans spécification d'hypothèses de modèle.

Le plan de l'article est structuré comme suit. Dans la section 2, nous présentons le scénario de base. Dans la section 3, nous présentons la méthodologie pour l'imputation par le plus proche voisin et établissons ses propriétés asymptotiques. À la section 4, nous examinons deux stratégies d'amélioration de l'estimateur d'imputation par le plus proche voisin, l'une utilisant l'imputation par les k plus proches voisins et l'autre utilisant des modèles additifs généralisés. À la section 5, nous proposons une technique de calage par régression pour améliorer l'efficacité des estimateurs d'imputation massive quand en plus, l'appartenance aux mégadonnées est observée dans tout l'échantillon probabiliste. Dans la section 6, nous démontrons que les estimateurs proposés sont robustes et efficaces au moyen d'études par simulations fondées sur des données artificielles et des données réelles tirées de la *Monthly Retail Trade Survey* du U.S. Census Bureau. Dans la section 7, nous présentons une étude de cas qui applique la méthode

proposée pour intégrer les données nationales de santé et les dossiers nationaux d'assurance-santé. Enfin, les conclusions de l'étude sont présentées à la section 8.

2. Scénario de base

2.1 Notation : Deux sources de données

Soit $\mathcal{F}_N = \{(\mathbf{X}_i, Y_i) : i \in U\}$ avec $U = \{1, \ldots, N\}$ qui désigne une population finie, où $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)$ est un vecteur de covariables à p dimensions, et Y_i est la variable étudiée. Nous supposons que $\mathcal{F}_N$ est un échantillon aléatoire d'un modèle de superpopulation ζ et que N est connu. Notre objectif est d'estimer le paramètre de population finie générale $\mu_g = N^{-1}\sum_{i=1}^N g(Y_i)$ pour une valeur connue $g(\cdot)$. Par exemple, si $g(Y) = Y$, $\mu_g = N^{-1}\sum_{i=1}^N Y_i$ est la moyenne de la population de Y. Si $g(Y) = \mathbf{1}(Y < c)$ pour une constante c, $\mu_g = N^{-1}\sum_{i=1}^N \mathbf{1}(Y_i < c)$ est la proportion de population de Y inférieure à c.

Supposons qu'il existe deux sources de données, l'une provenant d'un échantillon probabiliste, appelé échantillon A, et l'autre d'une source de mégadonnées, appelée échantillon B. Le tableau 2.1 illustre la structure des données observées. L'échantillon A contient des observations $\mathcal{O}_A = \{(d_i = \pi_i^{-1}, \mathbf{X}_i) : i \in A\}$ dont la taille d'échantillon est $n = |A|$, où $\pi_i = P(i \in A)$ est connu dans tout l'échantillon A, et l'échantillon B contient des observations $\mathcal{O}_B = \{(\mathbf{X}_i, Y_i) : i \in B\}$ dont la taille d'échantillon est $N_B = |B|$. Souvent, l'échantillon probabiliste contient de nombreux autres éléments, mais nous n'utilisons que ceux qui chevauchent nos mégadonnées. Bien que la source de mégadonnées ait une grande taille d'échantillon, le mécanisme d'échantillonnage est souvent inconnu, et nous ne pouvons pas calculer la probabilité d'inclusion de premier ordre pour l'estimation de Horvitz-Thompson. Les estimateurs naïfs sans ajustement pour le processus d'échantillonnage présentent des biais de sélection. Par ailleurs, bien que l'échantillon probabiliste avec poids de sondage représente la population finie, il n'observe pas la variable étudiée.

Tableau 2.1
Deux sources de données. « √ » et « ? » indiquent respectivement les données observées et non observées

		Poids de sondage $d = \pi^{-1}$	Covariable $\mathbf{X}$	Variable étudiée Y
Échantillon probabiliste	1	√	√	?
$\mathcal{O}_A$	⋮	⋮	⋮	⋮
	n	√	√	?
Échantillon de mégadonnées	1	?	√	√
$\mathcal{O}_B$	⋮	⋮	⋮	⋮
	N_B	?	√	√

L'échantillon A est un échantillon probabiliste, et l'échantillon B provient de mégadonnées, mais peut comporter des biais de sélection.

2.2 Hypothèses

Soit $f(Y|\mathbf{X})$ la fonction de densité conditionnelle de Y donnée $\mathbf{X}$ dans le modèle de superpopulation ζ. Soit $f(\mathbf{X})$ et $f(\mathbf{X}|\delta_B = 1)$ la fonction de densité de $\mathbf{X}$ respectivement dans la population finie et l'échantillon B, où δ_B est l'indicateur de sélection de l'échantillon B. Nous posons d'abord les hypothèses suivantes.

Hypothèse 1 (ignorabilité). *Conditionnellement à* $\mathbf{X}$, *la densité de* Y *dans l'échantillon B suit le modèle de superpopulation, c'est-à-dire* $f(Y|\mathbf{X}; \delta_B = 1) = f(Y|\mathbf{X})$.

Les hypothèses 1 et 2 constituent la condition d'ignorabilité forte (Rosenbaum et Rubin, 1983). Ce scénario a été utilisé auparavant par plusieurs auteurs; voir, par exemple, Rivers (2007), Vavreck et Rivers (2008). L'hypothèse 1 énonce que l'ignorabilité du mécanisme de sélection de l'échantillon B est conditionnelle aux covariables. L'hypothèse 1 implique également que $P(\delta_B = 1|\mathbf{X}, Y) = P(\delta_B = 1|\mathbf{X})$. Cette hypothèse se vérifie si l'ensemble de covariables contient tous les prédicteurs du résultat qui influent sur la possibilité d'être sélectionné dans l'échantillon B. Selon cette hypothèse, les résultats manquants dans l'échantillon A sont manquants au hasard (Rubin, 1976).

Hypothèse 2 (support commun). *Le vecteur des covariables* $\mathbf{X} \in \mathbb{R}^p$ *possède un support compact et convexe, avec une densité bornée ayant une borne inférieure strictement positive. Il existe des constantes* C_l *et* C_u *telles que* $C_l \leq f(\mathbf{X})/f(\mathbf{X}|\delta_B = 1) \leq C_u$ *presque certainement.*

L'hypothèse 2 implique que le support de $\mathbf{X}$ dans l'échantillon B est identique à celui de la population finie. Cette hypothèse peut également être formulée comme une hypothèse de positivité selon laquelle $P(\delta_B = 1|\mathbf{X}) > 0$ que pour tous les $\mathbf{X}$. L'hypothèse 2 ne se vérifie pas si certaines unités ne sont jamais incluses dans l'échantillon de mégadonnées. La plausibilité de l'hypothèse peut être jugée en fonction de la connaissance du domaine spécialisé. À des fins de diagnostic, nous pouvons examiner la distribution des scores de propension estimés ou la distribution des poids des scores de propension dans l'échantillon A. Les valeurs des scores de propension proches de zéro ou les valeurs extrêmement élevées de scores de propension indiquent la possibilité de violation de la positivité. Nous supposons que toutes les covariables sont continues. On peut traiter les variables catégoriques en définissant d'abord des classes d'imputation au moyen de la partition des catégories, puis en estimant la moyenne du résultat à l'aide de l'imputation par le plus proche voisin dans les classes d'imputation. Dans notre contexte, parce que l'échantillon B est un échantillon de mégadonnées, la taille des donneurs pour chaque catégorie d'imputation peut être raisonnablement grande.

3. Méthodologie

3.1 Imputation par le plus proche voisin

Par souci de simplicité, nous nous concentrerons sur l'estimateur de type Horvitz-Thompson, bien que notre analyse s'applique aussi à d'autres types d'estimateurs. Si Y_i est observée dans l'échantillon A, on

peut utiliser l'estimateur de Horvitz-Thompson $\hat{\mu}_{g,\mathrm{HT}} = N^{-1}\sum_{i\in A}\pi_i^{-1}g(Y_i)$. Nous considérons l'estimateur d'imputation de μ_g, donné par $\hat{\mu}_{g,I} = N^{-1}\sum_{i\in A}\pi_i^{-1}g(Y_i^*)$, où Y_i^* est une valeur imputée pour Y_i. La création de valeurs imputées pour l'ensemble des données est appelée imputation massive (Chipperfield et coll., 2012; Kim et Rao, 2012).

Pour obtenir des valeurs imputées appropriées, nous examinons l'imputation par le plus proche voisin afin de trouver l'unité d'appariement la plus proche dans l'échantillon B en fonction des valeurs $\mathbf{X}$ et d'utiliser la valeur Y correspondante de cette unité comme valeur imputée. Cette méthode a été appelée *appariement d'échantillons* par Rivers (2007). Pour étudier les propriétés théoriques, nous envisageons d'abord l'appariement avec remise au moyen d'une seule imputation; la discussion sur l'imputation par les k plus proches voisins est présentée à la section 4.

La méthode d'imputation massive par le plus proche voisin peut se décrire selon les étapes suivantes :

Étape 1. Pour chaque unité $i \in A$, trouver le plus proche voisin dans l'échantillon B ayant la distance minimale entre $\mathbf{X}_j$ et $\mathbf{X}_i$. Soit $i(1)$ l'indice de son plus proche voisin, qui satisfait $d(\mathbf{X}_{i(1)}, \mathbf{X}_i) \le d(\mathbf{X}_j, \mathbf{X}_i)$, pour $j \in B$, où $d(\mathbf{X}_i, \mathbf{X}_j)$ est une fonction de distance entre $\mathbf{X}_i$ et $\mathbf{X}_j$. En cas d'égalité, en sélectionner une valeur au hasard comme plus proche voisin. Sans perte de généralité, nous utilisons la distance euclidienne, $d(\mathbf{X}_i, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|$ où $\|\mathbf{X}\| = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{1/2}$, pour déterminer les voisins.

Étape 2. L'estimateur d'imputation par le plus proche voisin de μ_g est

$$\hat{\mu}_{g,\,\mathrm{nni}} = \frac{1}{N}\sum_{i\in A}\pi_i^{-1}g(Y_{i(1)}). \tag{3.1}$$

Remarque 1. *Notre développement théorique s'applique à une classe générale de distances* $\|\mathbf{X}\|_\Sigma = (\mathbf{X}^{\mathrm{T}}\Sigma^{-1}\mathbf{X})^{1/2}$, *où* Σ *est une matrice définie positive (Abadie et Imbens, 2006). Cette classe inclut la distance de Mahalanobis standard en établissant* Σ *comme matrice de covariance empirique de* $\mathbf{X}$. *Nous écrivons* $\Sigma = L^{\mathrm{T}}L$. *Observons que* $\|\mathbf{X}\|_\Sigma = \{(L\mathbf{X})\,L\mathbf{X}\}^{1/2} = \|L\mathbf{X}\|$. *Par conséquent, l'utilisation de* $\|\cdot\|_\Sigma$ *et* $\mathbf{X}$ *équivaut à utiliser* $\|\cdot\|$ *et* $L\mathbf{X}$. *On peut donc reporter le résultat théorique au cas avec* $\|\mathbf{X}\|_\Sigma$.

Comparativement à l'imputation fondée sur un modèle, l'imputation par le plus proche voisin présente plusieurs avantages. Premièrement, elle ne nécessite pas d'hypothèses de modèle paramétrique fortes et, par conséquent, elle est robuste en cas de spécification erronée du modèle. Deuxièmement, l'imputation par le plus proche voisin est basée sur le donneur, ce qui fait que la valeur imputée est une valeur réellement mesurée, qui sera toujours dans les limites des valeurs observées. Troisièmement, contrairement aux méthodes d'imputation par la régression, l'imputation par le plus proche voisin peut conserver la structure variance-covariance complexe des données. De plus, pour le même ensemble de données imputé, on peut estimer différents paramètres en choisissant une valeur $g(\cdot)$ raisonnable. Gardons à l'esprit que p est la dimension de $\mathbf{X}$. Le biais asymptotique de $\hat{\mu}_{g,\mathrm{nni}}$ est d'ordre $O_{\mathrm{p}}(N_B^{-1/p})$ (Abadie et Imbens, 2006), qui est négligeable quand le nombre de covariables continues est fixé à un nombre raisonnable et que la taille du bassin de donneurs appariés est très grand, comme dans notre

scénario de mégadonnées. Si $\mathbf{X}$ a une grande dimension, il faut sélectionner des variables pour que l'estimateur d'imputation par le plus proche voisin ait de bonnes propriétés statistiques. Dans ce cas, nous proposons de choisir des variables importantes associées au résultat de sorte que l'hypothèse 1 se vérifie et que l'estimation soit plus précise (Brookhart, Schneeweiss, Rothman, Glynn, Avorn et Stürmer, 2006).

3.2 Résultats asymptotiques

Pour étudier les propriétés asymptotiques de $\hat{\mu}_{g,\mathrm{nni}}$, nous imposons les conditions de régularité suivantes.

Hypothèse 3. *(i)* $f(\mathbf{X})$ *et* $\mu_g(\mathbf{X}) = E\{g(Y)\,|\,\mathbf{X}\}$ *sont continuellement différentiables de tout* $g(Y)$ *borné et continu, et (ii)* $E\{g(Y)^{\beta}\,|\,\mathbf{X}\}$ *est borné pour* $\beta = 1, 2$.

Hypothèse 4. *(i) Il existe des constantes positives* C_1 *et* C_2 *de sorte que* $C_1 \le Nn^{-1}\pi_i \le C_2$, *pour* $i = 1, \ldots, N$; *(ii) la fraction de sondage de l'échantillon A est négligeable,* $nN^{-1} = o(1)$; *et (iii) la séquence des estimateurs de Horvitz-Thompson* $\hat{\mu}_{g,\mathrm{HT}}$ *satisfaisait* $\mathrm{var}_p(\hat{\mu}_{g,\mathrm{HT}}) = O(n^{-1})$ *et* $\{\mathrm{var}_p(\hat{\mu}_{g,\mathrm{HT}})\}^{-1/2}(\hat{\mu}_{g,\mathrm{HT}} - \mu_g)\,|\,\mathcal{F}_N \to \mathcal{N}(0, 1)$ *en loi, quand* $n \to \infty$, *où* $\mathrm{var}_p(\cdot) = \mathrm{var}(\cdot\,|\,\mathcal{F}_N)$ *est la variance selon le plan de sondage de l'échantillon A.*

À des fins de clarification, la loi de probabilité distribution étayant la notation $E(\cdot)$, $\mathrm{var}(\cdot)$, $o_{\mathrm{p}}(\cdot)$ et $O_{\mathrm{p}}(\cdot)$ est la distribution conjointe du modèle de superpopulation et des processus d'échantillonnage pour les échantillons A et B. L'hypothèse 3 est une condition technique imposée par la continuité fonctionnelle et des moments finis, qui se vérifie pour les modèles courants; voir par exemple Mack (1981). L'hypothèse 4 se vérifie dans les plans de sondage généralement utilisés dans les enquêtes (Fuller, 2009; chapitre 1). Elle exige le bon comportement des poids de sondage, en ce sens qu'il n'existe pas de poids extrêmement grands dominant les autres poids. Cela se produit quand des sujets présentant certaines caractéristiques sont largement sous-représentés dans un échantillon. Des conditions suffisantes pour l'hypothèse 4 (iii) sont présentées dans le chapitre 3 de Fuller (2009).

Nous dérivons la théorie asymptotique pour $\hat{\mu}_{g,\mathrm{nni}}$ dans le théorème suivant et en ferons la démonstration dans le supplément de documentation.

Théorème 1. *Selon les hypothèses 1 à 3 et* $NN_B^{-1} = O(1)$, $\hat{\mu}_{g,\mathrm{nni}}$ *a la même distribution que* $\hat{\mu}_{g,\mathrm{HT}}$ *quand* $N_B \to \infty$. *De plus, selon l'hypothèse 4,* $\hat{\mu}_{g,\mathrm{nni}}$ *est convergent pour* μ_g, *et*

$$n^{1/2}(\hat{\mu}_{g,\mathrm{nni}} - \mu_g) \to \mathcal{N}(0, V_{\mathrm{nni}}), \tag{3.2}$$

où

$$V_{\mathrm{nni}} = \lim_{n\to\infty} \frac{n}{N^2} E\left[\mathrm{var}_p\left\{\sum_{i\in A} \pi_i^{-1} g(Y_i)\right\}\right].$$

Le théorème 1 implique que l'estimateur ponctuel type puisse être appliqué aux données imputées $\{(\mathbf{X}_i, Y_{i(1)}) : i \in A\}$ comme si les valeurs $Y_{i(1)}$ étaient observées. Soit π_{ij} la probabilité d'inclusion conjointe pour les unités i et j. Nous montrons dans la documentation supplémentaire que l'estimateur de variable direct fondé sur les données imputées

$$\hat{V}_{\text{nni}} = \frac{n}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{g(Y_{i(1)})}{\pi_i} \frac{g(Y_{j(1)})}{\pi_j}$$

est convergent pour V_{nni}.

4. Autres techniques d'imputation massive

4.1 Imputation par les k plus proches voisins

Au lieu d'utiliser une seule valeur imputée, nous considérons maintenant l'imputation fractionnelle avec k valeurs imputées pour chaque résultat manquant. L'imputation fractionnelle est conçue pour réduire la variance de l'estimateur final attribuable à l'imputation (Kalton et Kish, 1984; Kim et Fuller, 2004).

Supposons l'absence d'égalités d'appariement, soit $\mathcal{J}_k(i)$ l'ensemble de k plus proches voisins pour l'unité i

$$\mathcal{J}_k(i) = \left\{ l \in B : \sum_{j \in B} 1_{\{d(\mathbf{X}_j, \mathbf{X}_i) \le d(\mathbf{X}_l, \mathbf{X}_i)\}} \le k \right\} = \{i(1), \ldots, i(k)\}.$$

La méthode des k plus proches voisins aux fins d'imputation massive peut être décrite selon les étapes suivantes :

Étape 1. Pour toute unité $i \in A$, trouver les k plus proches voisins dans l'échantillon B, $\mathcal{J}_k(i)$. Imputer la valeur Y pour l'unité i au moyen de $\hat{\mu}_g(\mathbf{X}_i) = k^{-1} \sum_{j=1}^{k} g(Y_{i(j)})$.

Étape 2. L'estimateur d'imputation par les k plus proches voisins de μ_g est

$$\hat{\mu}_{g,\text{knn}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} \hat{\mu}_g(\mathbf{X}_i). \tag{4.1}$$

Dans la littérature sur l'estimation non paramétrique, les chercheurs ont étudié en profondeur les propriétés asymptotiques des estimateurs d'imputation par les k plus proches voisins. Voir, par exemple, Mack et Rosenblatt (1979) et Mack (1981) pour les références antérieures. Cheng (1994) établit la convergence de racine n de l'estimateur d'imputation par les k plus proches voisins de la moyenne du résultat quand le résultat est manquant. Nous dérivons la théorie asymptotique pour $\hat{\mu}_{g,\text{knn}}$ dans le contexte de l'imputation massive combinant un échantillon probabiliste et un échantillon de mégadonnées dans le théorème suivant et nous reportons sa démonstration au supplément de documentation.

Théorème 2. *Selon les hypothèses 1 à 4,* $n(k/N)^{4/p} \to 0$, $k/n \to 0$ *et* $k^2/n \to \infty$,

$$n^{1/2}\left(\hat{\mu}_{g,\mathrm{knn}} - \mu_g\right) \to \mathcal{N}(0, V_{\mathrm{knn}}), \tag{4.2}$$

où

$$V_{\mathrm{knn}} = \lim_{n\to\infty} \frac{n}{N^2}\left(E\left[\mathrm{var}_p\left\{\sum_{i\in A}\pi_i^{-1}\mu_g(\mathbf{X}_i)\right\}\right] + E\left\{\frac{1-\pi_B(\mathbf{X})}{\pi_B(\mathbf{X})}\sigma_g^2(\mathbf{X})\right\}\right),$$

et $\pi_B(\mathbf{X}) = P(\delta_B = 1|\mathbf{X})$ *et* $\sigma_g^2(\mathbf{X}) = \mathrm{var}\{g(Y)|\mathbf{X}\}$.

Si $\pi_B(\mathbf{X})$ atteint 1, V_{knn} diminue à $\lim_{n\to\infty}(n/N^2)E\left[\mathrm{var}_p\left\{\sum_{i\in A}\pi_i^{-1}\mu_g(\mathbf{X}_i)\right\}\right]$. Cela laisse penser que si le grand échantillon représente une grande fraction de la population cible, V_{knn} peut être plus petit que V_{nni}, ce qui indique que $\hat{\mu}_{g,\mathrm{knn}}$ gagne en efficacité par rapport à $\hat{\mu}_{g,\mathrm{nni}}$. Dans des échantillons finis, Beretta et Santaniello (2016) ont mené une étude par simulations pour comparer l'imputation par le plus proche voisin et l'imputation par les k plus proches voisins dans le contexte de données indépendantes et identiquement distribuées. Ils ont constaté que l'imputation par les k plus proches voisins avec une petite valeur de k donne de meilleurs résultats que l'imputation par le plus proche voisin pour ce qui est de l'erreur quadratique moyenne. D'une part, un plus grand k peut utiliser plus d'information de l'échantillon de mégadonnées et entraîner un plus grand gain d'efficacité. D'autre part, k ne peut pas être trop grand, afin que le biais de notre estimateur soit maîtrisé. En pratique, nous proposons d'utiliser des méthodes axées sur les données, comme la validation croisée, pour choisir une valeur raisonnable de k, et d'effectuer l'analyse de sensibilité en variant le choix de k.

4.2 Modèles additifs généralisés

Les méthodes d'imputation par le plus proche voisin sont non paramétriques. Par ailleurs, les modèles paramétriques, en particulier les modèles linéaires, sont sensibles aux spécifications erronées du modèle. Nous envisageons maintenant des méthodes semi-paramétriques d'imputation massive. Parmi les méthodes semi-paramétriques, les modèles additifs généralisés (Hastie et Tibshirani, 1990) sont souples en ce qui concerne la spécification du modèle de la dépendance de Y à $\mathbf{X}$, car le modèle est spécifié seulement au moyen de fonctions lisses plutôt qu'une relation paramétrique soit supposée. Comme dans d'autres méthodes non paramétriques, les performances des modèles additifs généralisés se détériorent quand la dimension de $\mathbf{X}$ devient grande. Pour une valeur $\mathbf{X}$ ayant une dimension modérée, nous appliquons des modèles additifs généralisés pour tirer parti de la puissance prédictive de l'échantillon de mégadonnées afin de produire un modèle prédictif pour Y étant donné $\mathbf{X}$, de façon à faciliter l'imputation massive pour l'échantillon probabiliste.

Nous supposons que $g(Y_i)$ étant donné $\mathbf{X}_i$ suit une distribution de famille exponentielle et

$$h^{-1}\{\mu_g(\mathbf{X}_i)\} = f_1(X_i^1) + f_2(X_i^2) + \ldots f_p(X_i^p), \tag{4.3}$$

où $h(\cdot)$ est une fonction de lien inverse, et chaque $f_k(\cdot)$ est une fonction lisse de X^k, pour $k = 1, \ldots, p$. Le modèle (4.3) permet une spécification plutôt souple de la dépendance de Y à $\mathbf{X}$. La fonction estimée $f_k(X^k)$ peut révéler des non-linéarités possibles de la relation entre Y et X^k.

Le modèle d'ajustement présente plusieurs difficultés (4.3). Premièrement, $f_k(x)$ est un paramètre de dimensions infinies, dont l'estimation repose souvent sur une certaine approximation. Deuxièmement, nous devons décider jusqu'à quel degré $f_k(x)$ doit être lisse pour équilibrer le compromis entre la complexité du modèle et le surajustement aux données disponibles.

Pour résoudre le premier problème, on peut utiliser une méthode courante d'approximation de $f_k(x)$ au moyen de splines. Soit $B_m(x)$ les fonctions splines de base pour $m = 1, \ldots, M$ (Ruppert, Wand et Carroll, 2009). Nous calculons par approximation $f_k(x)$ au moyen de $f_k(x) = \sum_{m=1}^{M} \gamma_m^k B_m(x)$ avec des coefficients de splines γ_m^k. Cela donne l'approximation du modèle (4.3) :

$$h^{-1}\left[\hat{E}\left\{g(Y_i) \mid \mathbf{X}_i\right\}\right] = \sum_{k=1}^{p}\sum_{m=1}^{M} \gamma_m^k B_m(X_i^k). \tag{4.4}$$

Dans (4.4), une grande valeur de M permet d'augmenter la complexité du modèle et d'augmenter le risque de surajustement, tandis qu'une petite valeur de M peut donner un modèle inadéquat. Ce compromis est contrebalancé par le choix d'une valeur relativement grande de M, puis la pénalisation de la complexité du modèle à l'étape de l'estimation (Eilers et Marx, 1996). Soit le vecteur des coefficients de splines $\gamma_k^{\mathrm{T}} = (\gamma_1^k, \ldots, \gamma_m^k)$ et $\gamma^{\mathrm{T}} = (\gamma_1^{\mathrm{T}}, \ldots, \gamma_p^{\mathrm{T}})$. On obtient la valeur estimée de $\hat{\gamma}$ en maximisant la probabilité de pénalité :

$$-2l(\gamma) + \sum_{k=1}^{p} \lambda_k \gamma_k^{\mathrm{T}} S_k \gamma_k \tag{4.5}$$

où $l(\gamma)$ est la fonction de log-vraisemblance de γ, S_k est une matrice avec le $(m, l)^{\mathrm{e}}$ composant $\int B_m''(x) B_l''(x)\,\mathrm{d}x$, $\gamma_k^{\mathrm{T}} S_k \gamma_k$ régularise f_k pour qu'elle soit lisse et que son degré de lissage soit contrôlé par λ_k. Compte tenu du paramètre de lissage $\lambda^{\mathrm{T}} = (\lambda_1, \ldots, \lambda_p)$, la fonction de probabilité pénalisée dans (4.5) est optimisée par une version pénalisée de l'algorithme des moindres carrés itérativement repondéré (Nelder et Baker, 1972; McCullagh, 1984) pour obtenir $\hat{\gamma}$. En ce qui concerne le choix de λ, nous remarquons que λ contrôle le compromis entre la complexité du modèle et le surajustement, qui peut être estimé séparément des autres coefficients du modèle au moyen d'une validation croisée généralisée ou estimé simultanément au moyen d'une estimation par le maximum de vraisemblance restreinte (Wood, 2006). En pratique, les performances du modèle ne sont pas sensibles au choix du nombre de fonctions de base tant que le nombre de fonctions de base est élevé par rapport à la taille de l'échantillon dans la spécification. En revanche, l'estimation du paramètre de lissage est essentielle au contrôle de la complexité du modèle.

Une fois le modèle ajusté, nous pouvons créer une valeur imputée pour chaque élément i de l'échantillon A comme suit :

$$\hat{\mu}_{g,\mathrm{GAM}}(\mathbf{X}_i) = h\left\{\hat{f}_1(X_i^1) + \hat{f}_2(X_i^2) + \ldots \hat{f}_p(X_i^p)\right\},$$

où $\hat{f}_k(x) = \sum_{m=1}^{M} \hat{\gamma}_m^k B_m(x)$ pour $k = 1, \ldots, p$. L'estimateur d'imputation massive fondé sur le modèle additif généralisé est

$$\hat{\mu}_{g,\,\mathrm{GAM}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} \hat{\mu}_{g,\mathrm{GAM}}(\mathbf{X}_i).$$

Étant donné que, dans notre contexte, la taille de l'échantillon B est beaucoup plus grande que celle de l'échantillon A, l'erreur d'estimation dans le modèle d'imputation peut être négligeable comparativement à la variabilité d'échantillonnage de $\hat{\mu}_{g,\mathrm{GAM}}$.

En conclusion du paragraphe, il est intéressant de commenter l'hypothèse des effets additifs de $\mathbf{X}$ dans le modèle (4.3). Cette hypothèse est peut-être plutôt forte. Pour assouplir l'hypothèse d'additivité, nous pouvons étendre le modèle (4.3) pour inclure les interactions au moyen de la base du produit tensoriel. Par exemple, nous pouvons inclure une surface d'interaction bivariée $f_{12}(X^1, X^2) = \sum_{m=1}^{M} \sum_{l=1}^{L} \gamma_{ml} B_m(X^1) B_l(X^2)$. Quand on emploie la base du produit tensoriel, il faut faire attention à la fonction de pénalité afin d'obtenir des degrés de liberté efficaces et appropriés pour le lisseur. Ce sujet été amplement étudié dans la littérature; voir par exemple Wood (2006).

5. Calage par régression

En pratique, en particulier pour les organismes gouvernementaux, la méthode par le plus proche voisin (unique) peut être privilégiée en raison de sa simplicité de mise en œuvre et de stockage des données. Nous envisageons maintenant une autre stratégie afin d'améliorer l'efficacité de $\hat{\mu}_{g,\mathrm{nni}}$ quand, en plus, l'appartenance à l'échantillon B peut être déterminée dans tout l'échantillon A avec l'indicateur δ_B. Dans certaines situations, nous pouvons obtenir δ_B en appariant les membres à l'échantillon B (c'est-à-dire couplage de données). Nous nous intéressons au scénario idéal sans erreurs de couplage. La principale idée est que le sous-échantillon d'unités de l'échantillon A avec $\delta_B = 1$ constitue un échantillon de deuxième phase de l'échantillon B, où l'échantillon B agit comme une nouvelle population. Un calage par régression type nécessite que toutes les variables de calage soient observées dans l'échantillon A et l'échantillon B, ce qui exclut la possibilité d'utiliser Y comme variable de calage en raison du manque de données de résultat provenant de l'échantillon B. Un des avantages des imputations massives réside dans le fait qu'on peut exploiter les résultats imputés pour faciliter le calage de Y.

Soit $\mathbf{h}(\delta_B, \mathbf{X}, Y)$ une fonction multidimensionnelle de δ_B, $\delta_B \mathbf{X}$ et $\delta_B Y$, par exemple $\mathbf{h}(\delta_B, \mathbf{X}, Y) = (\delta_B, 1 - \delta_B, \delta_B \mathbf{X}, \delta_B Y)^{\mathrm{T}}$. Pour des raisons de simplicité de la notation, nous utilisons $\mathbf{h}_i$ pour désigner $\mathbf{h}(\delta_{Bi}, \mathbf{X}_i, Y_i)$ et $\mathbf{h}_i^*$ pour désigner $\mathbf{h}(\delta_{Bi}, \mathbf{X}_i, Y_{i(1)})$. Nous pouvons calculer la quantité de population $\mathbf{H} = N^{-1} \sum_{i=1}^{N} \mathbf{h}_i$ à partir de l'échantillon B. Cette information permet la pondération par calage typique des échantillonnages d'enquête avec des totaux marginaux connus. Dans l'échantillon A, nous traitons les valeurs imputées comme des valeurs observées, et l'estimateur pondéré par le plan de $\mathbf{H}$ est $\hat{\mathbf{H}}_A = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{h}_i^*$. En général, $\hat{\mathbf{H}}_A$ n'est pas égal à $\mathbf{H}$. Nous pouvons utiliser l'information connue $\mathbf{H}$ pour améliorer l'efficacité de $\hat{\mu}_{g,\mathrm{nni}}$.

Cela permet de proposer la stratégie de calage suivante. Nous modifions les poids de sondage d'origine $\{d_i : i \in A\}$ dans $\hat{\mu}_{g,\mathrm{nni}}$ en un nouvel ensemble de poids $\{\omega_i : i \in A\}$ en réduisant au minimum une fonction de distance

$$\sum_{i \in A} G(\omega_i, d_i) = \sum_{i \in A} d_i \left(\frac{\omega_i}{d_i} - 1 \right)^2, \tag{5.1}$$

sous les contraintes de calage $N^{-1} \sum_{i \in A} \omega_i \mathbf{h}_i^* = \mathbf{H}$. Au moyen du multiplicateur de Lagrange, la solution au problème de minimisation des contraintes est

$$\omega_i = d_i + \left(N \times \mathbf{H} - \sum_{k \in A} d_k \mathbf{h}_k^* \right)^{\mathrm{T}} \left(\sum_{k \in A} d_k \mathbf{h}_k^* \mathbf{h}_k^{*\mathrm{T}} \right)^{-1} d_i \mathbf{h}_i^*,$$

pour $i \in A$. Les poids obtenus $\{\omega_i : i \in A\}$ peuvent être appelés poids de régression généralisés.

L'estimateur proposé qui utilise le nouvel ensemble de poids est

$$\hat{\mu}_{g,\mathrm{RC}} = \frac{1}{N} \sum_{i \in A} \omega_i g(Y_{i(1)}), \tag{5.2}$$

qui est l'équivalent asymptotique d'un estimateur par la régression généralisée (Park et Fuller, 2012). À la suite de Yang et Ding (2020), on peut montrer que $\hat{\mu}_{g,\mathrm{RC}}$ est l'estimateur optimal dans la classe de $\left\{ \hat{\mu}_{g,\mathrm{nni}} + \left(N \times \mathbf{H} - \sum_{k \in A} d_k \mathbf{h}_k^* \right)^{\mathrm{T}} \boldsymbol{\gamma} : \boldsymbol{\gamma} \in \mathbb{R}^{\dim(\mathbf{h})} \right\}$.

Nous dérivons la théorie asymptotique pour $\hat{\mu}_{g,\mathrm{RC}}$ dans le théorème suivant et en ferons la démonstration dans le supplément de documentation.

Théorème 3. *Selon les hypothèses 1 à 4,*

$$n^{1/2} \left(\hat{\mu}_{g,\mathrm{RC}} - \mu_g \right) \to \mathcal{N}(0, V_{\mathrm{RC}}), \tag{5.3}$$

en loi, quand $n \to \infty$, *où*

$$V_{\mathrm{RC}} = \lim_{n \to \infty} \frac{n}{N^2} E\left(\mathrm{var}_p \left[\sum_{i \in A} \pi_i^{-1} \left\{ g(Y_i) - \mathbf{h}_i^{\mathrm{T}} \boldsymbol{\beta}_N \right\} \right] \right),$$

et $\boldsymbol{\beta}_N = \left(\sum_{i=1}^N \mathbf{h}_i \mathbf{h}_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^N \mathbf{h}_i g(Y_i)$.

L'estimateur calé $\hat{\mu}_{g,\mathrm{RC}}$ améliore l'efficacité de $\hat{\mu}_{g,\mathrm{nni}}$ dans le sens que V_{RC} est au plus aussi grand que V_{nni} donné dans le théorème 1. Si $\mathbf{h}_i$ explique une proportion de la variabilité de $g(Y_i)$, V_{RC} est strictement inférieure à V_{nni} et le gain d'efficacité ne nécessite pas d'hypothèse de modèle paramétrique.

Remarque 2 (Choix des fonctions de distance). *On peut envisager différentes fonctions de distance dans (5.1). Si nous choisissons* $G(\omega_i, d_i) = -d_i \log(\omega_i / d_i)$, *cela mène à une estimation par la vraisemblance empirique (Newey et Smith, 2004). Si nous choisissons la fonction de distance de Kullback-Leibler* $G(\omega_i, d_i) = \omega_i \log(d_i / \omega_i)$, *cela nous conduit à une estimation par inclinaison exponentielle (Kitamura et Stutzer, 1997; Imbens, Johnson et Spady, 1998; Schennach, 2007; Dong et coll., 2020). Sous des conditions faibles, ces procédures fournissent un ensemble de poids qui est asymptotiquement équivalent à l'ensemble de poids de régression (Deville et Särndal, 1992; Breidt et Opsomer, 2017).*

Pour l'estimation de la variance, par le théorème (3), nous construisons un estimateur de variance convergent pour $\hat{\mu}_{g,\mathrm{RC}}$ quand $\hat{V}_{\mathrm{RC}}/n$, où

$$\hat{V}_{\mathrm{RC}} = \frac{n}{N^2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{\hat{e}_i}{\pi_i}\frac{\hat{e}_j}{\pi_j},$$

avec $\hat{e}_i = g\left(Y_{i(1)}\right) - \mathbf{h}_i^{*\mathrm{T}}\hat{\boldsymbol{\beta}}$, et

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{h}_i^{*}\mathbf{h}_i^{*\mathrm{T}}\right)^{-1}\begin{pmatrix}\sum_{i=1}^{N}\delta_{Bi}g\left(Y_i\right)\\ \sum_{i\in A}\pi_i^{-1}\left(1-\delta_{Bi}\right)g\left(Y_{i(1)}\right)\\ \sum_{i=1}^{N}\delta_{Bi}\mathbf{X}_i g\left(Y_i\right)\\ \sum_{i=1}^{N}\delta_{Bi}Y_i g\left(Y_i\right)\end{pmatrix}.$$

6. Expériences empiriques

Dans la présente section, nous évaluons les performances de l'échantillon fini de l'estimateur proposé au moyen d'études par simulations, l'une basée sur des données artificielles avec échantillonnage aléatoire simple, et l'autre basée sur un fichier de population synthétique provenant d'un échantillon mensuel unique de la *Monthly Retail Trade Survey* du U.S. Cencus Bureau avec échantillonnage stratifié.

6.1 Exemple de Kim-Wang

Nous utilisons l'exemple de simulation donné par Kim et Wang (2019) pour comparer différents estimateurs. Nous générons les données selon le mécanisme suivant. Nous générons d'abord une population finie $\mathcal{F}_N = \{\mathbf{X}_i = (X_{1i}, X_{2i}), \mathbf{Y}_i = (Y_{1i}, Y_{2i}) : i = 1, \ldots, N\}$ de taille $N = 1\,000\,000$, où Y_{1i} est un résultat continu et Y_{2i} est un résultat binaire. Dans la population finie, nous sélectionnons un échantillon B de mégadonnées où l'indicateur d'inclusion $\delta_{Bi} \sim \mathrm{Ber}\left(p_i\right)$ avec p_i la probabilité d'inclusion pour l'unité i ayant une taille d'échantillon d'environ 700 000. Nous obtenons un échantillon A représentatif de taille $n = 1\,000$ au moyen d'un échantillonnage aléatoire simple. Les paramètres d'intérêt sont la moyenne de population $N^{-1}\sum_{i=1}^{N}\mathbf{Y}_i$ et la moyenne de la population conditionnelle de Y_1 étant donné $Y_2 = 1$.

Pour générer la population finie, nous posons des modèles linéaires

$$Y_{1i} = 1 + X_{1i} + X_{2i} + \alpha_i + \varepsilon_i, \tag{6.1}$$
$$P\left(Y_{2i} = 1 \mid X_{1i}, X_{2i}; \alpha_i\right) = \mathrm{logit}\left(1 + X_{1i} + X_{2i} + \alpha_i\right),$$

et des modèles non linéaires

$$Y_{1i} = 0{,}5\left(X_{1i} - 1{,}5\right)^2 + X_{2i}^2 + \alpha_i + \varepsilon_i, \tag{6.2}$$
$$P\left(Y_{2i} = 1 \mid X_{1i}, X_{2i}; \alpha_i\right) = \mathrm{logit}\left\{0{,}5\left(X_{1i} - 1{,}5\right)^2 + X_{2i}^2 + \alpha_i\right\},$$

où $X_{1i} \sim \mathcal{N}(1, 1)$, $X_{2i} \sim \text{Exp}(1)$, $\alpha_i \sim \mathcal{N}(0, 1)$, $\varepsilon_i \sim \mathcal{N}(0, 1)$, et X_{1i}, X_{2i}, α_i et ε_i mutuellement indépendants. Les variables α_i induisent la dépendance de Y_{1i} et Y_{2i} et la correction de X_{1i} et X_{2i}. Pour la probabilité d'inclusion des mégadonnées, nous envisageons aussi un modèle logistique linéaire

$$\text{logit}(p_i) = X_{2i}, \tag{6.3}$$

et un modèle logistique non linéaire

$$\text{logit}(p_i) = -3 + (X_{1i} - 1{,}5)^2 + (X_{2i} - 2)^2. \tag{6.4}$$

Nous examinons les combinaisons suivantes : I. (6.1) et (6.3); II. (6.1) et (6.4); II. (6.2) et (6.3); et IV. (6.2) et (6.4) pour les mécanismes de génération de données. Par conséquent, le scénario de la simulation est un plan factoriel 2×2 avec deux niveaux dans chaque facteur.

Chen, Li et Wu (2020) proposent l'estimateur de pondération par l'inverse du score de propension au moyen de la probabilité estimée de sélection dans l'échantillon B et l'estimateur doublement robuste qui incorpore en outre un modèle de régression des résultats. Afin d'évaluer la robustesse et l'efficacité, nous comparons les estimateurs suivants :

1. $\hat{\mu}_{\text{HT}}$, l'estimateur de Horvitz–Thompson en supposant que les valeurs Y_i sont observées dans l'échantillon A à des fins de comparaison des données repères;
2. $\hat{\mu}_{\text{ipw}}$, l'estimateur de pondération par l'inverse du score de propension,

   $$\hat{\mu}_{\text{ipw}} = \frac{1}{N} \sum_{i \in B} \frac{1}{p_i(\hat{\eta})} Y_i,$$

   où $p_i(\eta) = P(\delta_{Bi} = 1 | X_{2i}; \eta)$ est un modèle de régression logistique avec le prédicteur linéaire X_{2i} avec un paramètre inconnu η, et $\hat{\eta}$ est un estimateur de η obtenu par la maximisation de la fonction de vraisemblance modifiée de η (Chen et coll., 2019) à partir des échantillons A et B;
3. $\hat{\mu}_{\text{dr}}$, l'estimateur doublement robuste de Chen et coll. (2019),

   $$\hat{\mu}_{\text{dr}} = \frac{1}{N} \sum_{i \in B} \frac{1}{p_i(\hat{\eta})} \left(Y_i - \mathbf{X}_i^{\text{T}} \hat{\boldsymbol{\beta}}\right) + \frac{1}{n} \sum_{i \in A} \mathbf{X}_i^{\text{T}} \hat{\boldsymbol{\beta}},$$

   où $\hat{\boldsymbol{\beta}}$ est les coefficients de régression estimés au moyen de (6.1) comme modèle de régression des résultats de travail à partir de l'échantillon B;
4. $\hat{\mu}_{\text{nni}}$, l'estimateur d'imputation par le plus proche voisin;
5. $\hat{\mu}_{\text{knn}}$, l'estimateur par les k plus proches voisins avec $k = 5$;
6. $\hat{\mu}_{\text{GAM}}$, l'estimateur d'imputation du modèle additif généralisé;
7. $\hat{\mu}_{\text{RC}}$, l'estimateur par calage par régression fondé sur $\hat{\mu}_{\text{nni}}$ avec les variables de calage $\mathbf{H}(\delta_B, \mathbf{X}, Y) = (\delta_B, 1 - \delta_B, \delta_B \mathbf{X}, \delta_B Y)^{\text{T}}$.

Tous les résultats de la simulation sont fondés sur 1 000 exécutions par la méthode Monte Carlo. Le tableau 6.1 résume les résultats de la simulation avec des biais, des erreurs-types et des taux de couverture d'intervalles de confiance de 95 % utilisant la normalité asymptotique des estimateurs ponctuels. Les

observations suivantes peuvent être faites à la lecture du tableau 6.1. $\hat{\mu}_{\text{ipw}}$ a des biais importants en cas de spécification erronée du score de propension. $\hat{\mu}_{\text{dr}}$ gagne en robustesse par rapport à $\hat{\mu}_{\text{ipw}}$ si l'un des modèles de régression des résultats ou le score de propension est correctement spécifié. Toutefois, si la spécification des deux modèles est erronée, $\hat{\mu}_{\text{dr}}$ présente un biais plus important. $\hat{\mu}_{\text{nni}}$ présente de faibles biais dans les quatre scénarios, ce qui démontre sa robustesse. Il est important de souligner que les performances de $\hat{\mu}_{\text{nni}}$ se rapprochent de celles de $\hat{\mu}_{\text{HT}}$ pour ce qui est des erreurs types et des taux de couverture, ce qui est conforme à notre théorie présentée dans le théorème 1. De plus, comme le prédisent nos résultats théoriques, $\hat{\mu}_{\text{knn}}$ améliore l'efficacité de $\hat{\mu}_{\text{nni}}$. De plus, $\hat{\mu}_{\text{GAM}}$ montre une robustesse en raison de la souplesse de la spécification du modèle. L'estimateur par calage par régression $\hat{\mu}_{\text{RC}}$ présente de faibles biais dans tous les scénarios et montre donc une robustesse par rapport aux spécifications du modèle pour le score d'échantillonnage et le résultat. De plus, ses erreurs-types sont plus petites que celles de $\hat{\mu}_{\text{nni}}$ et $\hat{\mu}_{\text{knn}}$. Les taux de couverture se rapprochent tous du niveau nominal.

Tableau 6.1
Résultats de la simulation : biais, erreur-type et taux de couverture d'intervalles de confiance de 95 % selon quatre scénarios fondés sur 1 000 échantillons de Monte Carlo. MR : modèle de résultats; SP : modèle de score de propension (tous les chiffres du tableau sont les résultats numériques multipliés par 100)

MR SP	Scénario I linéaire linéaire			Scénario II linéaire non linéaire			Scénario III non linéaire linéaire			Scénario IV non linéaire non linéaire		
	Biais	**ET**	**TC**	**Biais**	**ET**	**TC**	**Biais**	**ET**	**TC**	**Biais**	**ET**	**TC**
					Moyenne de la population de Y_1							
$\hat{\mu}_{\text{HT}}$	0,2	6,5	96,0	-0,2	6,4	94,5	0,61	15,2	95,7	-0,5	15,6	93,5
$\hat{\mu}_{\text{ipw}}$	-0,1	1,6	95,3	22,2	35,8	97,5	-0,1	4,2	95,3	432,7	284,5	75,6
$\hat{\mu}_{\text{dr}}$	0,0	4,6	94,5	0,0	4,3	96,5	0,5	14,2	95,2	229,8	168,8	35,8
$\hat{\mu}_{\text{nni}}$	0,2	6,5	95,1	-0,3	6,4	94,7	0,7	15,2	94,6	-0,6	15,6	93,7
$\hat{\mu}_{\text{knn}}$	0,2	4,9	96,1	-0,3	4,9	95,6	0,5	14,5	94,6	-0,6	14,9	93,8
$\hat{\mu}_{\text{GAM}}$	0,1	4,5	95,7	-0,2	4,5	96,0	0,5	14,3	94,9	-0,6	14,8	93,4
$\hat{\mu}_{\text{RC}}$	0,0	3,2	95,5	-0,2	4,1	95,3	-0,1	4,8	95,0	0,1	6,7	95,5
					Moyenne de la population de Y_2							
$\hat{\mu}_{\text{HT}}$	-0,0	1,5	96,2	-0,0	1,6	95,1	-0,1	1,6	95,2	0,1	1,6	94,4
$\hat{\mu}_{\text{ipw}}$	0,0	0,2	95,0	-12,1	3,1	0,0	-0,0	0,3	95,4	3,0	1,8	94,7
$\hat{\mu}_{\text{dr}}$	-0,0	0,9	95,0	-1,1	1,8	68,6	0,0	0,4	94,9	-2,9	2,2	59,8
$\hat{\mu}_{\text{nni}}$	0,0	1,4	95,3	-0,0	1,6	95,3	-0,1	1,6	94,6	0,1	1,6	95,3
$\hat{\mu}_{\text{knn}}$	0,0	1,0	95,8	-0,0	1,1	95,8	-0,0	1,0	95,2	0,0	0,9	96,1
$\hat{\mu}_{\text{GAM}}$	-0,0	0,9	95,3	-0,0	0,9	94,8	-0,0	0,8	96,2	0,0	0,8	94,5
$\hat{\mu}_{\text{RC}}$	0,0	1,2	95,5	-0,1	1,4	94,2	-0,0	1,4	94,1	0,1	1,5	95,6
					Moyenne conditionnelle de Y_1 étant donné $Y_2 = 1$							
$\hat{\mu}_{\text{HT}}$	0,0	7,3	95,1	-0,3	7,2	95,2	0,2	9,3	95,3	-0,1	9,8	94,1
$\hat{\mu}_{\text{ipw}}$	-0,1	1,6	95,2	-9,1	10,3	69,8	-0,1	4,3	95,0	534,2	329,8	65,3
$\hat{\mu}_{\text{dr}}$	0,1	4,7	95,6	2,5	4,6	93,2	9,8	18,0	93,1	452,0	465,4	65,6
$\hat{\mu}_{\text{nni}}$	-0,0	7,3	95,0	-0,3	7,3	95,3	0,1	9,2	95,4	-2,2	9,5	95,2
$\hat{\mu}_{\text{knn}}$	-0,1	4,7	96,8	-0,3	4,6	96,5	0,1	6,0	94,8	0,0	6,4	93,6
$\hat{\mu}_{\text{GAM}}$	0,0	4,8	94,2	-0,3	4,5	96,0	-0,1	6,5	95,5	-0,6	6,8	94,8
$\hat{\mu}_{\text{RC}}$	-0,0	3,9	94,8	-0,2	5,0	96,0	-0,2	5,4	95,1	-0,1	5,4	96,7

6.2 Monthly Retail Trade Survey des États-Unis

Pour démontrer la pertinence pratique de notre analyse, nous examinons la *Monthly Retail Trade Survey* de 2014 réalisée par le U.S. Census Bureau (Mulry, Oliver et Kaputa, 2014). Cette enquête se

fonde sur des indicateurs économiques dont les estimations mensuelles sont entrées dans les estimations du produit intérieur brut. Dans le cadre de l'enquête, un échantillon d'environ 12 000 commerces de détail est sélectionné chaque mois; des employés rémunérés sont chargés de recueillir des données sur les ventes et les stocks. Elle emploie un échantillon stratifié à un degré avec une stratification fondée sur une grande industrie, qui est ensuite substratifié par les ventes annuelles estimées qu'on appelle variable de taille.

À des fins de simulation, nous utilisons les données simulées de la *Monthly Retail Trade Survey* de 2014 pour proposer le modèle générateur de données et les vraies valeurs des paramètres (https://ww2.amstat.org/meetings/ices/2016/contests.cfm). Nous générons une population finie de $N =$ 812 765 commerces de détail avec 16 strates ayant un identificateur de strate h, des ventes Y, des stocks $\mathbf{X}$ et une variable de taille Z à l'échelle logarithmique. Le tableau 6.2 présente quelques statistiques sommaires. Nous générons les données d'inventaire de $X_{hi} \sim N\left(\mu_{X,h}, \sigma^2_{X,h}\right)$ pour $i = 1, \ldots, N_h$ et $h = 1, \ldots, 16$, et les données de vente à partir d'un modèle linéaire

$$Y_{hi} = \beta_0 + X_{hi} + \varepsilon_{hi}, \tag{6.5}$$

et un modèle non linéaire

$$Y_{hi} = \beta_0 + 0{,}5X^2_{hi} + \varepsilon_{hi}, \tag{6.6}$$

où $\varepsilon_{hi} \sim \mathcal{N}(0; 0{,}25)$. Dans (6.5) et (6.6), nous spécifions différentes valeurs de β_0 afin que le paramètre d'intérêt, $\mu = N^{-1}\sum_{h=1}^{16}\sum_{i=1}^{N_h} Y_{hi}$, corresponde à la moyenne de la population réelle de 12,73.

Tableau 6.2
Taille de la strate, répartition de l'échantillon, moyenne et erreur type des données de stock à l'échelle logarithmique, extraites de l'ensemble de données simulé de la *Monthly Retail Trade Survey* de 2014

Strate h	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
N_h	366	20	2 015	4 646	7 402	700	12 837	17 080	29 808	2 400	41 343	57 518	83 465	95 244	115 028	342 893
n_h	37	5	34	57	74	7	103	115	116	12	184	196	218	200	220	336
$\mu_{X,h}$	16,8	16,7	16,6	16,4	16,1	15,6	16,0	15,7	15,6	15,5	15,4	15,1	14,8	14,5	13,9	11,5
$\sigma_{X,h}$	1,1	0,8	0,4	0,3	0,4	0,6	0,4	0,4	0,4	0,3	0,4	0,4	0,3	0,7	0,5	1,1
$\mu_{Z,h}$	5,9	2,3	5,8	6,3	6,6	4,2	6,9	7,0	7,4	4,8	7,5	7,6	7,7	7,6	7,7	8,1

Nous générons également un échantillon de mégadonnées $\mathcal{S}_B$ où l'indicateur d'inclusion $\delta_{hi} \sim \mathrm{Ber}(p_{hi})$ avec la probabilité d'inclusion p_{hi} pour l'unité i dans la strate h. En pratique, l'échantillon de mégadonnées est souvent disponible auprès d'entreprises commerciales électroniques qui surveillent les stocks et les ventes des commerces de détail. Pour la probabilité d'inclusion des mégadonnées, soit $Z_{hi} \sim N\left(\mu_{Z,h}, \sigma^2_{Z,h}\right)$, pour $i = 1, \ldots, N_h$ et $h = 1, \ldots, 16$. Considérons un modèle linéaire logistique

$$\mathrm{logit}(p_{hi}) = \alpha_0 + Z_{hi}, \tag{6.7}$$

et un modèle logistique non linéaire

$$\mathrm{logit}(p_{hi}) = \alpha_0 + X_{hi} + Z^2_{hi}, \tag{6.8}$$

où nous spécifions différentes valeurs de α_0 afin que la probabilité d'inclusion moyenne soit d'environ 30 %. Enfin, nous générons un échantillon représentatif $\mathcal{S}_A$ par échantillonnage stratifié avec échantillonnage aléatoire simple sans remise dans les strates; voir dans le tableau 6.2 la répartition de l'échantillon.

Nous considérons les sept estimateurs de la section 6.1 adoptés pour l'échantillonnage stratifié. Dans chaque ensemble de données obtenues par imputation massive, nous appliquons l'estimateur ponctuel et l'estimateur de la variance suivants : $\hat{\mu} = N^{-1}\sum_{h=1}^{H} N_h \bar{y}_{n_h}$ et $\bar{y}_{n_h}$ est la moyenne de l'échantillon de y dans la h^{e} strate, $\hat{V}(\hat{\mu}) = N^{-2}\sum_{h=1}^{H} N_h^2 (1 - n_h / N_h) s_{n_h}^2 \big/ n_h$ avec $s_{n_h}^2 = (n_h - 1)^{-1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_{n_h})^2$.

Le tableau 6.3 résume les résultats de la simulation. Une discussion semblable à celle de la section 6.1 s'applique. $\hat{\mu}_{\text{ipw}}$ est sensible à la spécification erronée du modèle de sélection, tandis que $\hat{\mu}_{\text{dr}}$ comporte une caractéristique de double robustesse, qui repose toujours sur le fait qu'au moins un modèle soit spécifié correctement. L'imputation massive fondée sur l'imputation par le plus proche voisin, l'imputation par les k plus proches voisins et le modèle additif généralisé se montrent performants parce qu'ils tirent parti de la représentativité de l'échantillon d'enquête et de la puissance prédictive de l'échantillon de mégadonnées. En outre, si l'appartenance aux mégadonnées est connue dans l'ensemble des données d'enquête, l'estimateur par calage par régression gagne en efficacité tout en conservant sa robustesse par rapport à la spécification erronée du modèle.

Tableau 6.3
Résultats de la simulation : biais, erreur-type et taux de couverture d'intervalles de confiance de 95 % selon quatre scénarios fondés sur 1 000 exécutions par la méthode Monte Carlo pour la Monthly Retail Trade Survey de 2014. MR : modèle de résultats; SP : modèle de score de propension (tous les chiffres du tableau sont les résultats numériques multipliés par 100)

MR SP	Scénario I linéaire linéaire			Scénario II linéaire non linéaire			Scénario III non linéaire linéaire			Scénario IV non linéaire non linéaire		
	Biais	ET	TC	Biais	ET	TC	Biais	ET	TC	Biais	ET	TC
$\hat{\mu}_{\text{HT}}$	0,0	3,0	95,0	0,0	3,0	95,0	1,1	31,5	95,0	1,1	31,5	95,0
$\hat{\mu}_{\text{ipw}}$	-0,6	5,8	96,6	-55,5	1,7	0,0	-7,3	76,2	96,6	-735,8	22,3	0,0
$\hat{\mu}_{\text{dr}}$	-0,3	2,7	94,4	-0,2	2,7	94,0	-3,3	34,6	93,8	-52,3	33,2	65,0
$\hat{\mu}_{\text{nni}}$	0,1	3,1	94,5	-0,1	3,1	94,6	1,1	31,5	95,3	-0,3	31,7	94,6
$\hat{\mu}_{\text{knn}}$	0,1	2,7	94,4	-0,2	2,7	94,3	1,0	31,4	94,9	-2,3	31,4	94,1
$\hat{\mu}_{\text{GAM}}$	0,1	2,7	94,9	0,1	2,7	94,9	1,1	31,6	94,9	-2,5	31,4	94,2
$\hat{\mu}_{\text{RC}}$	0,1	2,9	94,1	-0,1	2,6	95,1	0,6	30,7	94,6	-0,5	26,9	95,0

7. Application sur données réelles

7.1 Description des données

Pour démontrer l'utilisation pratique de la méthode proposée, nous l'appliquons aux données de l'enquête *Korea National Health and Nutrition Examination Survey* (KNHANES) et aux mégadonnées du *National Health Insurance Sharing Service* (NHISS). La KNHANES, une enquête nationale sur l'état de

santé et nutritionnel de la population coréenne, menée annuellement depuis 1998 par les centres nationaux de prévention et de lutte contre les maladies de la Corée. Il s'agit d'une enquête transversale et représentative à l'échelle nationale, fondée sur un échantillon de 10 000 répondants. Elle comporte trois volets, soit une interview sur la santé, un examen de santé et une enquête sur la nutrition, qui permettent de recueillir des renseignements sur le statut socioéconomique, les comportements liés à la santé, la qualité de vie et l'utilisation des soins de santé, ainsi que des mesures anthropométriques et des profils biochimiques et cliniques concernant les maladies non transmissibles et les apports nutritionnels. Pour en savoir plus sur la KNHANES, voir Kweon, Kim, jin Jang, Kim, Kim, Choi, Chun, Khang et Oh (2014). L'ensemble de données utilisé dans la présente étude compte 4 929 échantillons.

Les mégadonnées du NHISS fournissent des renseignements sur la santé recueillies dans le cadre du National Health Screening Program (NHSP) de la Corée du Sud. Le NHSP a été lancé dans le but d'améliorer la santé générale des citoyens sud-coréens et de prévenir les maladies chroniques coûteuses. Tous les bénéficiaires sont admissibles à un dépistage annuel ou bisannuel, selon leur profil démographique ou leur situation professionnelle. Les éléments de dépistage sont énumérés précisément dans les normes de mise en œuvre; ils comprennent plusieurs analyses sanguines et le dépistage du cancer. Le nombre total de bénéficiaires admissibles est d'environ 16 millions, et autour de 75 % d'entre eux ont participé au dépistage. Les données utilisées dans la présente étude constituent le sous-ensemble correspondant aux résultats des analyses sanguines associées au syndrome métabolique du programme de 2014. Les variables de cet ensemble de données sont des données démographiques comme le sexe et l'âge, et des mesures cliniques comme le taux total de glycérides (mg/dL), le cholestérol total (mg/dL), le cholestérol à lipoprotéines de haute densité (HDL, mg/dL) et le diagnostic médical de l'anémie. L'ensemble de données est mis à la disposition du public après l'anonymisation et la sélection aléatoire de 1 million d'observations (National Health Insurance Data Sharing Service, 2014). Il est aussi possible d'acheter des données plus complètes avec un abonnement payant et un examen par un groupe d'experts.

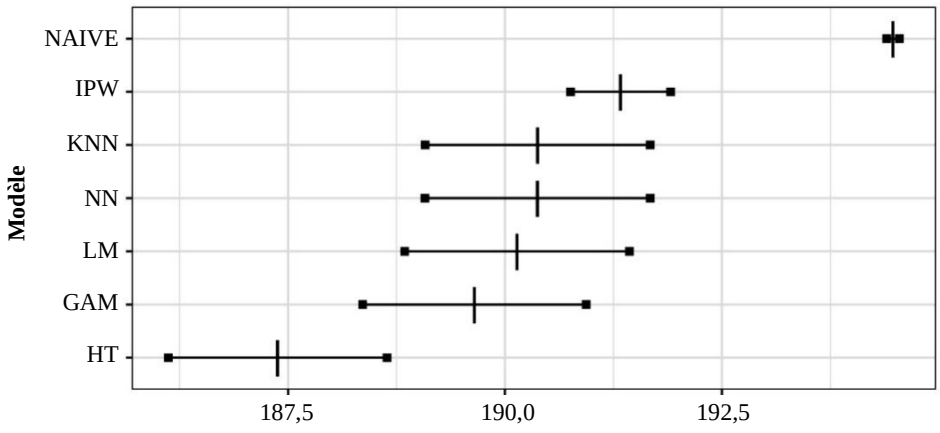
7.2 Analyse et résultats

Pour appliquer la méthode d'imputation massive proposée, nous supposons que le cholestérol total n'est pas disponible dans les données de l'enquête KNHANES, et nous utilisons les mégadonnées du NHSP afin d'effectuer l'imputation massive pour la variable du cholestérol total. Nous utilisons les valeurs réelles de l'enquête KNHANES pour calculer des données repères afin de pouvoir valider l'efficacité de la méthode proposée. Nous considérons six estimateurs différents :

- HT : l'estimateur de Horvitz-Thompson basé sur les données de l'échantillon A. Cette méthode est utilisée à des fins d'analyse comparative;
- NN : l'estimateur d'imputation par le plus proche voisin (*nearest neighbor*);
- kNN : l'estimateur d'imputation par les k plus proches voisins avec $k = 5$;
- GAM : estimateur d'imputation du modèle additif généralisé (*generalized additive model*);
- LM : estimateur d'imputation du modèle de régression linéaire (*linear model*) utilisant le sexe, le groupe d'âge, le cholestérol HDL et les taux de glycérides totaux comme covariables;
- IPW : estimateur de pondération par l'inverse du score de propension (*inverse propensity score weighting estimator*);
- NAIVE : l'estimateur naïf utilisant l'échantillon B sans traitement.

La quantité de HDL influe sur le cholestérol total, parce que le HDL est un des composants du cholestérol total, et on sait que ce dernier est également modifié par le sexe et l'âge. À moins que l'échantillon B ne provienne d'une sous-population particulière comme le groupe des patients atteints de sténose aortique, nous pouvons supposer que la relation entre le cholestérol total et d'autres variables demeure identique. Par conséquent, l'ignorabilité se vérifie. De plus, les covariables sont toutes des mesures médicales ou biologiques, ce qui signifie qu'elles devraient rester dans la même fourchette pour les échantillons A et B. L'estimateur de la variance pour chaque estimateur est calculé, et l'intervalle de confiance de Wald à 95 % pour μ_g est obtenu au moyen de la normalité asymptotique. La figure 7.1 illustre les intervalles : l'estimation de la moyenne de la population selon chaque méthode est présentée sous forme de barre verticale. L'intervalle obtenu à partir de HT peut être considéré comme un repère. On voit que tous les estimateurs produisent des intervalles légèrement surestimés comparativement à celui de HT. Cela s'explique par le biais inhérent du taux de cholestérol total dans les données du NHSP; les valeurs moyennes de l'échantillon du taux de cholestérol total des données du NHSP sont de 7 points ou 3,7 % supérieures à celles calculées au moyen de l'estimateur HT à partir des données de la KNHANES, comme le montre le résultat naïf. On constate que toutes les méthodes proposées réduisent considérablement ce biais pour que l'estimateur se rapproche de l'estimateur de HT, ce qui démontre l'avantage que présentent les méthodes proposées. L'estimateur IPW a produit une estimation relativement mauvaise comparativement aux autres méthodes, probablement en raison soit de la spécification erronée du modèle de régression logistique soit d'un écart considérable entre les tailles d'échantillon des données de la KNHANES (4 929) et du NHSP (1 million). Nous avons également essayé l'estimateur doublement robuste, mais nous ne l'avons pas inclus ici, car l'effet du IPW est très marginal en raison du nombre limité de variables auxiliaires disponibles.

Figure 7.1 Intervalle de confiance à 95 % estimé pour le taux de cholestérol total.



Pour mieux comprendre les propriétés prédictives des méthodes d'imputation massive, nous avons calculé la REQM, le biais moyen et la corrélation des valeurs imputées en comparant les valeurs imputées et les valeurs réelles de l'enquête. Puisque nous pouvons observer les valeurs réelles de l'enquête KNHANES, nous pouvons calculer les mesures de la qualité de la prédiction. Le tableau 7.1 est un tableau résumé, où nous avons comparé les résultats aux niveaux individuels et aux niveaux moyens par sous-groupe, divisés par groupe d'âge et sexe. Pour les niveaux des sous-groupes, nous obtenons d'abord les

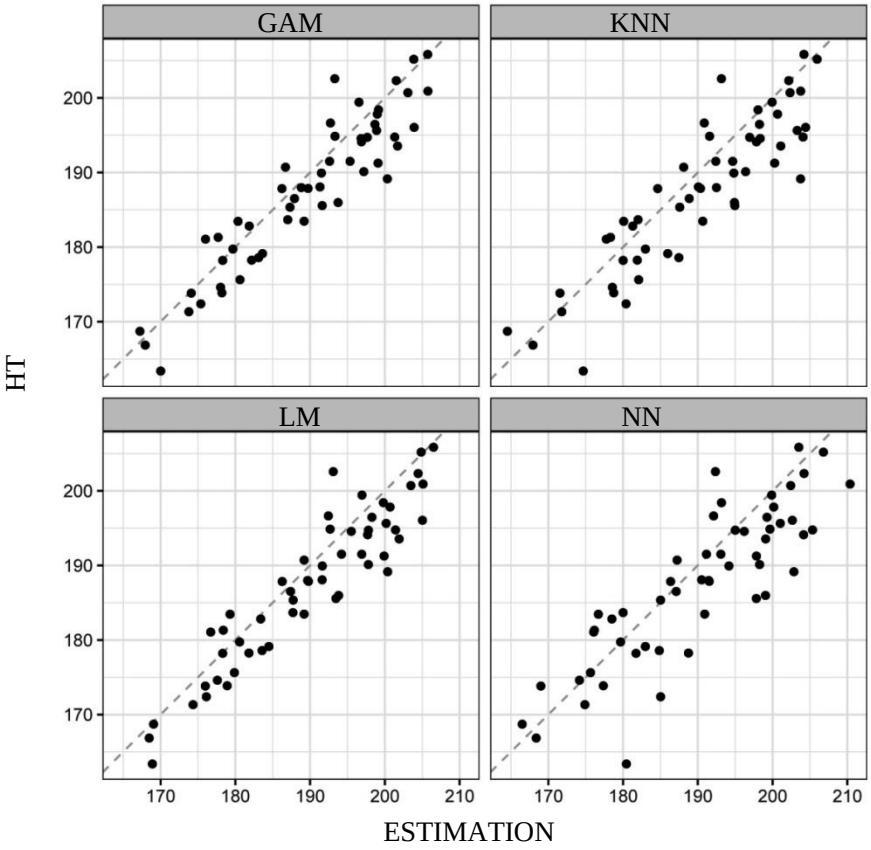
estimations de la moyenne des sous-groupes, puis nous calculons les statistiques agrégées pour différents groupes. On constate que le GAM donne de meilleurs résultats que les autres méthodes en matière de REQM et de corrélation. Dans l'ensemble, la méthode d'imputation massive donne des résultats raisonnables pour le niveau des sous-groupes, comme l'illustre la figure 7.2.

Ces mesures de la qualité nécessitent une valeur prédite pour l'échantillon A, c'est pourquoi les estimateurs IPW sont exclus de la comparaison. L'estimation des moyennes de la population et des sous-groupes au moyen de l'échantillon B peut donner un résultat très biaisé : dans le cas des données du NHSP, la différence entre la moyenne des données du NHSP et l'estimateur de HT à partir des données de la KNHANES est d'environ 7,09, ou 3,7 %.

Tableau 7.1
Comparaison des méthodes d'imputation

	Méthode	REQM	Biais	Corr.
Individuel	NN	43,94	2,87	0,26
	KNN	32,62	2,86	0,42
	GAM	29,15	2,13	0,54
	LM	30,35	2,59	0,48
Moyennes des groupes	NN	6,33	2,68	0,85
	KNN	5,44	2,70	0,90
	GAM	4,33	2,03	0,93
	LM	4,57	2,52	0,93

Figure 7.2 Comparaison des estimations de HT et des estimations à l'aide de l'imputation massive pour la moyenne des sous-groupes.



8. Discussion

L'imputation massive est une technique importante pour l'intégration des données d'enquête. Quand l'ensemble de données d'entraînement pour l'imputation est obtenu à partir d'un échantillon probabiliste, la théorie de Kim et Rao (2012) peut être appliquée directement. Si l'ensemble de données d'entraînement est un échantillon non probabiliste de très grande taille, nous avons montré dans l'article que plusieurs méthodes non paramétriques peuvent servir à l'imputation massive et que l'erreur d'estimation dans le modèle d'imputation peut être ignorée sans risque, sous l'hypothèse que le mécanisme d'échantillonnage des données d'entraînement est manquant au hasard dans le sens de Rubin (1976). Si l'on croit que le mécanisme d'échantillonnage n'est pas manquant au hasard, les techniques d'imputation peuvent être développées selon des hypothèses structurelles fortes pour le mécanisme d'échantillonnage (par exemple Riddles, Kim et Im, 2016; Morikawa et Kim, 2018) ou le modèle de résultat (par exemple Yang, Zeng et Wang, 2020). De plus, quand l'ensemble de données d'entraînement a une structure hiérarchique, des modèles multiniveaux peuvent servir à développer l'imputation massive. Cela est étroitement lié à l'estimation pour petits domaines au niveau de l'unité dans l'échantillonnage d'enquête (Rao et Molina, 2015).

L'estimateur d'imputation massive n'est pas nécessairement efficace. À la section 5, nous avons décrit une méthode d'utilisation de la pondération par calage comme outil d'intégration efficace des données en présence de mégadonnées. La pondération par calage nécessite un appariement correct entre deux sources de données, comme l'ont montré Kim et Tam (2020). De plus, si la fraction des mégadonnées dans la population finie n'est pas importante, le gain d'efficacité est limité. On pourrait plutôt améliorer l'efficacité en conjuguant l'estimateur d'imputation massive et l'estimateur de pondération par l'inverse du score de propension dans les mégadonnées (Yang, Kim et Song, 2020). Il sera toutefois difficile de spécifier correctement le modèle de score de propension. Ces sujets devront faire l'objet de futures recherches.

Remerciements

Nous remercions les deux examinateurs anonymes et le rédacteur adjoint pour leurs commentaires très constructifs. Shu Yang est partiellement soutenu par la subvention DMS 1811245 de la NSF et la subvention 1R01AG066883 du NIA. Jae Kwang Kim est partiellement soutenu par la subvention MMS 1733572 de la NSF et la *Iowa Agriculture and Home Economics Experiment Station*, Ames, Iowa.

Annexe

A.1 Démonstration du théorème 1

Pour tout $\mathbf{X}_i = \mathbf{x}$ dans l'échantillon A, nous montrons que $\mathbf{X}_{i(1)}$ converge vers $\mathbf{x}$ en probabilité quand $N_B \to \infty$. Pour tout $\varepsilon > 0$, nous montrons que

$$P\left\{d\left(\mathbf{X}_{i(1)}, \mathbf{x}\right) > \varepsilon\right\} = P\left\{d\left(\mathbf{X}_j, \mathbf{x}\right) > \varepsilon, \forall j \in B\right\} \tag{A.1}$$

converge vers zéro, et alors $\mathbf{X}_{i(1)}$ converge vers $\mathbf{x}$ en probabilité quand $N_B \to \infty$, où la probabilité est induite par le processus d'échantillonnage de l'échantillon B de taille N_B. Nous le démontrons par contradiction. Supposons que pour une valeur $\varepsilon > 0$, $P\left\{d\left(\mathbf{X}_{i(1)}, \mathbf{x}\right) > \varepsilon\right\}$ ne converge pas vers zéro quand $N_B \to \infty$. Définissons la région $\mathcal{R}_\varepsilon = \{\mathbf{X} : d(\mathbf{X}, \mathbf{x}) \le \varepsilon\}$. Ensuite, nous devons avoir $f(\mathbf{X} \mid \delta_B = 1) = 0$ pour $\mathbf{X} \in \mathcal{R}_\varepsilon$; sinon, $\tilde{\mathbf{X}} \in \mathcal{R}_\varepsilon$ existe avec une probabilité positive dans l'échantillon B quand $N_B \to \infty$, et par conséquent $P\left\{d\left(\mathbf{X}_{i(1)}, \mathbf{x}\right) > \varepsilon\right\} = 0$ quand $N_B \to \infty$. Mais l'affirmation selon laquelle $f(\mathbf{X} \mid \delta_B = 1) = 0$ pour $\mathbf{X} \in \mathcal{R}_\varepsilon$ implique que $\mathcal{R}_\varepsilon$ est une région sans chevauchement de la distribution de $\mathbf{X}$ entre l'échantillon A (et également la population) et l'échantillon B, ce qui contredit l'hypothèse 2.

Soit $\mathbf{X}_i = \mathbf{x}$ dans l'échantillon A, pour toute valeur continue et bornée $g(y)$,

$$\begin{aligned}
E\left\{g\left(Y_{i(1)}\right) \middle| \mathbf{X}_i = \mathbf{x}, i \in A\right\} &= E\left[E\left\{g\left(Y_{i(1)}\right) \middle| \mathbf{X}_{i(1)}, \mathbf{X}_i = \mathbf{x}, i \in A\right\} \middle| \mathbf{X}_i = \mathbf{x}, i \in A\right] \\
&= E\left[E\{g\left(Y_{i(1)}\right) \middle| \mathbf{X}_{i(1)}\} \middle| \mathbf{X}_i = \mathbf{x}, i \in A\right] \\
&= E\left\{\mu_g\left(\mathbf{X}_{i(1)}\right) \middle| \mathbf{X}_i = \mathbf{x}, i \in A\right\} \to E\left\{\mu_g\left(\mathbf{X}_i\right) \middle| \mathbf{X}_i = \mathbf{x}, i \in A\right\} \\
&= E\left\{g\left(Y_i\right) \middle| \mathbf{X}_i = \mathbf{x}, i \in A\right\},
\end{aligned}$$

en probabilité quand $N_B \to \infty$, où $\to$ découle du fait que $\mu_g(\mathbf{x})$ est bornée et continue. Puis, par le lemme porte-manteau (Klenke, 2006), $Y_{i(1)} \to Y_i \mid (\mathbf{X}_i = \mathbf{x}, i \in A)$ en loi quand $N_B \to \infty$. Selon l'hypothèse 1, $g\left(Y_{i(1)}\right) \mid (\mathbf{X}_i, i \in A) \to \mu_g(\mathbf{X}_i) + e_g^*(\mathbf{X}_i)$ en loi quand $N_B \to \infty$, où $e_g^*(\mathbf{X}_i)$ a la même distribution que $\{g(Y_i) \mid (\mathbf{X}_i, i \in A)\} - \mu_g(\mathbf{X}_i)$.

Nous montrons maintenant que pour $i \ne j \in A$, $e_g^*(\mathbf{X}_i)$ et $e_g^*(\mathbf{X}_j)$ sont conditionnellement indépendants, compte tenu des données $\mathcal{O}_A$. Il suffit de montrer que $P\{i(1) = j(1)\} \to 0$ quand $N_B \to \infty$. En d'autres termes, la même unité ne peut pas être appariée pour l'unité i et l'unité j avec la probabilité 1. Cela peut être montré au moyen de (A.1) avec $\varepsilon = \min_{i \ne j \in A} \| \mathbf{X}_i - \mathbf{X}_j \|$.

Par conséquent, conditionnellement aux données $\mathcal{O}_A$, nous avons

$$\hat{\mu}_{g,\mathrm{nni}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} g\left(Y_{i(1)}\right) \to \frac{1}{N} \sum_{i \in A} \pi_i^{-1} g\left(Y_i\right) = \hat{\mu}_{g,\mathrm{HT}}$$

en loi quand $N_B \to \infty$. Le théorème 1 est ainsi démontré.

Supposons que

$$\tilde{V}_{\mathrm{nni}} = \frac{n}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{g\left(Y_i\right)}{\pi_i} \frac{g\left(Y_j\right)}{\pi_j}. \tag{A.2}$$

Alors, $\tilde{V}_{\mathrm{nni}}$ est convergent pour V_{nni}.

Comme dans l'argument ci-dessus, pour $i, j \in A$, conditionnellement aux données $\mathcal{O}_A$, $g\left(Y_{i(1)}\right) g\left(Y_{j(1)}\right) \to g\left(Y_i\right) g\left(Y_j\right)$ quand $N_B \to \infty$. Par conséquent, conditionnement aux données $\mathcal{O}_A$,

$$\hat{V}_{\mathrm{nni}} = \frac{n}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{g\left(Y_{i(1)}\right)}{\pi_i} \frac{g\left(Y_{j(1)}\right)}{\pi_j} \to \tilde{V}_{\mathrm{nni}}, \tag{A.3}$$

en loi quand $N_B \to \infty$. En conjuguant (A.2) et (A.3), $\hat{V}_{\text{nni}}$ converge pour V_{nni}.

A.2 Démonstration du théorème 2

Pour étudier les propriétés asymptotiques de $\hat{\mu}_{g,\text{knn}}$, nous exprimons de nouveau

$$\hat{\mu}_g(\mathbf{x}) = \frac{\sum_{j\in B} K_{R_\mathbf{x}}(\mathbf{x}-\mathbf{X}_j)\, g(Y_j)}{\sum_{j\in B} K_{R_\mathbf{x}}(\mathbf{x}-\mathbf{X}_j)},$$

où

$$K_h(u) = \frac{1}{h^p} K\left(\frac{u}{h}\right), \quad K(u) = 0{,}5 I(\|u\| \le 1),$$

et la largeur de bande $h = R_\mathbf{x}$ est la distance aléatoire entre $\mathbf{x}$ et la plus éloignée parmi les k plus proches voisins. Par conséquent, $\hat{\mu}_{g,\text{knn}}$ peut être considéré comme un estimateur par noyau incorporant une largeur de bande basée sur les données.

La littérature a amplement étudié les propriétés asymptotiques de l'estimateur d'imputation par les k plus proches voisins. Le résultat indiqué dans le lemme suivant sur l'imputation par les k plus proches voisins est extrait de Mack (1981).

Lemme 1. *Selon les hypothèses 1 à 3,*

$$N^{-1}\sum_{j=1}^{N} \delta_{B,j} K_{R_\mathbf{x}}(\mathbf{x}-\mathbf{X}_j)\, g(Y_j) = f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x}) + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}. \tag{A.4}$$

Nous exprimons maintenant

$$\hat{\mu}_{g,\text{knn}} = \frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1}\delta_{A,i}\mu_g(\mathbf{X}_i) + \frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1}\delta_{A,i}\left\{\hat{\mu}_g(\mathbf{X}_i) - \mu_g(\mathbf{X}_i)\right\}.$$

Soit $T_N = N^{-1}\sum_{i=1}^{N} \pi_i^{-1}\delta_{A,i}\left\{\hat{\mu}_g(\mathbf{X}_i) - \mu_g(\mathbf{X}_i)\right\}$. Pour étudier les propriétés pour T_N, nous observons d'abord $\hat{\mu}_g(\mathbf{x})$, qui peut être exprimé comme suit :

$$\hat{\mu}_g(\mathbf{x}) = \frac{h_N(\mathbf{x})}{f_N(\mathbf{x})},$$

où $h_N(\mathbf{x}) \equiv N^{-1}\sum_{j=1}^{N} \delta_{B,j} K_{R_\mathbf{x}}(\mathbf{x}-\mathbf{X}_j)\, g(Y_j)$ et $f_N(\mathbf{x}) \equiv N^{-1}\sum_{j=1}^{N} \delta_{B,j} K_{R_\mathbf{x}}(\mathbf{x}-\mathbf{X}_j)$. Au moyen du résultat du lemme 1, nous obtenons

$$h_N(\mathbf{x}) = f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x}) + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}$$

$$f_N(\mathbf{x}) = f(\mathbf{x})\pi_B(\mathbf{x}) + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.$$

Maintenant, au moyen d'un développement en séries de Taylor, nous obtenons

$$\begin{aligned}\hat{\mu}_g(\mathbf{x}) - \mu_g(\mathbf{x}) &= \frac{h_N(\mathbf{x})}{f_N(\mathbf{x})} - \mu_g(\mathbf{x}) \\ &= \frac{1}{f(\mathbf{x})\pi_B(\mathbf{x})}\{h_N(\mathbf{x}) - f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x})\} \\ &\quad - \frac{f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x})}{\{f(\mathbf{x})\pi_B(\mathbf{x})\}^2}\{f_N(\mathbf{x}) - f(\mathbf{x})\pi_B(\mathbf{x})\} + O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\} \\ &= \frac{1}{f(\mathbf{x})\pi_B(\mathbf{x})}\{h_N(\mathbf{x}) - f_N(\mathbf{x})\mu_g(\mathbf{x})\} + O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.\end{aligned}$$

Nous obtenons alors

$$T_N = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\delta_{A,i}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}\sum_{j=1}^{N}\delta_{B,j}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\} + O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.$$

Selon l'hypothèse du théorème 2, il est facile de calculer que $(k/N)^{2/p} + 1/k = o(n^{-1/2})$, et, par conséquent,

$$T_N = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\delta_{A,i}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}\sum_{j=1}^{N}\delta_{B,j}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\} + o_{\mathrm{p}}(n^{-1/2}).$$

Nous exprimons ensuite T_N sous forme de U-statistiques (van der Vaart, 2000; chapitre 12) :

$$T_N = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}h(\mathbf{Z}_i, \mathbf{Z}_j) + o_{\mathrm{p}}(n^{-1/2}),$$

où $\mathbf{Z}_i = (\mathbf{X}_i, Y_i, \delta_{A,i}, \delta_{B,i})$ et

$$\begin{aligned}h(\mathbf{Z}_i, \mathbf{Z}_j) &= \frac{1}{2}\left[\frac{\delta_{A,i}\delta_{B,j}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\}\right. \\ &\quad \left. + \frac{\delta_{A,j}\delta_{B,i}}{\pi_j}\frac{1}{f(\mathbf{X}_j)\pi_B(\mathbf{X}_j)}K_{R_{\mathbf{X}_j}}(\mathbf{X}_j - \mathbf{X}_i)\{g(Y_i) - \mu_g(\mathbf{X}_j)\}\right] \\ &\equiv \frac{1}{2}(\zeta_{ij} + \zeta_{ji}).\end{aligned}$$

Maintenant, au moyen du lemme 1, nous obtenons

$$\begin{aligned}E(\zeta_{ij}\,|\,\mathbf{Z}_i) &= E\left[\frac{\delta_{A,i}\delta_{B,j}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\}\,\middle|\,\mathbf{Z}_i\right] \\ &= \frac{\delta_{A,i}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}E\left[\delta_{B,j}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\}\,|\,\mathbf{Z}_i\right] \\ &= O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\},\end{aligned}$$

et

$$
\begin{aligned}
E\left(\zeta_{ji} \,\middle|\, \mathbf{Z}_i\right) &= E\left[\frac{\delta_{A,j}\delta_{B,i}}{\pi_j}\frac{1}{f\left(\mathbf{X}_j\right)\pi_B\left(\mathbf{X}_j\right)}K_{R_{\mathbf{X}_j}}\left(\mathbf{X}_j - \mathbf{X}_i\right)\left\{g\left(Y_i\right) - \mu_g\left(\mathbf{X}_j\right)\right\} \,\middle|\, \mathbf{Z}_i\right] \\
&= \delta_{B,i}E\left(E\left[\frac{\delta_{A,j}}{\pi_j}\frac{1}{f\left(\mathbf{X}_j\right)\pi_B\left(\mathbf{X}_j\right)}K_{R_{\mathbf{X}_j}}\left(\mathbf{X}_j - \mathbf{X}_i\right)\left\{g\left(Y_i\right) - \mu_g\left(\mathbf{X}_j\right)\right\} \,\middle|\, R_{\mathbf{X}_j}, \mathbf{Z}_i\right] \,\middle|\, \mathbf{Z}_i\right) \\
&= \frac{\delta_{B,i}}{\pi_B\left(\mathbf{X}_i\right)}\left\{g\left(Y_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} + O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.
\end{aligned}
$$

Alors, selon la théorie des U-statistiques, nous obtenons

$$
\begin{aligned}
T_N &= \frac{2}{N}\sum_{i=1}^{N} E\left\{h\left(\mathbf{Z}_i, \mathbf{Z}_j\right) \,\middle|\, \mathbf{Z}_i\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_{B,i}}{\pi_B\left(\mathbf{X}_i\right)}\left\{g\left(Y_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right).
\end{aligned}
$$

La combinaison des résultats ci-dessus donne

$$
\begin{aligned}
\hat{\mu}_{g,\mathrm{knn}} - \mu_g &= \frac{1}{N}\sum_{i=1}^{N}\left\{\pi_i^{-1}\delta_{A,i}\mu_g\left(\mathbf{X}_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\delta_{B,i}}{\pi_B\left(\mathbf{X}_i\right)} - 1\right\}\left\{g\left(Y_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right). \qquad \text{(A.5)}
\end{aligned}
$$

Ensuite, les résultats asymptotiques du théorème 2 sont conformes aux hypothèses 1 à 4 et (A.5).

A.3 Démonstration du théorème 3

La convergence et la normalité asymptotique de $n^{1/2}\hat{\mu}_{g,\mathrm{nni}}$ sont conformes aux arguments standard présentés dans les hypothèses 1 à 4. Le reste sert à montrer que la variance asymptotique de $n^{1/2}\hat{\mu}_{g,\mathrm{nni}}$ est V_{nni}.

Au moyen de la fonction de distance $G\left(\omega_i, d_i\right) = d_i\left(\omega_i / d_i - 1\right)^2$ dans (5.1), l'estimation de la distance minimale mène à une estimation par la régression généralisée (Park et Fuller, 2012). Alors, nous exprimons

$$
\begin{aligned}
n^{1/2}\hat{\mu}_g &= \frac{n^{1/2}}{N}\sum_{i\in A}\omega_i g\left(Y_{i(1)}\right) \\
&= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}g\left(Y_{i(1)}\right) - \frac{n^{1/2}}{N}\left(\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N - \sum_{i=1}^{N}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N\right) + o_{\mathrm{p}}\left(n^{-1/2}\right). \qquad \text{(A.6)}
\end{aligned}
$$

Comme dans l'argument de la démonstration du théorème 1, nous écrivons

$$
\begin{aligned}
n^{1/2}\hat{\mu}_g &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}g\left(Y_{i(1)}\right) - \frac{n^{1/2}}{N}\left(\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N - \sum_{i=1}^{N}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N\right) + o_{\mathrm{p}}\left(n^{-1/2}\right) \\
&= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}\left\{g\left(Y_{i(1)}\right) - \mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N\right\} + \frac{n^{1/2}}{N}\sum_{i=1}^{N}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N + o_{\mathrm{p}}\left(n^{-1/2}\right). \qquad \text{(A.7)}
\end{aligned}
$$

Il est simple de montrer que la variance du second terme dans (A.7) est négligeable étant donné $nN^{-1} = o(1)$. Conformément aux arguments de la démonstration des théorèmes 1 et 2, $g(Y_{i(1)})$ et $\mathbf{h}_i^*$ ont la distribution asymptotique quand $g(Y_i)$ et $\mathbf{h}_i$ compte tenu des données $\mathcal{O}_A$ de l'échantillon A, respectivement. Par conséquent, la variance asymptotique de $n^{1/2}\hat{\mu}_g$ est

$$V_{\mathrm{RC}} = \lim_{n\to\infty} \operatorname{var}\left[\frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}\left\{g(Y_i) - \mathbf{h}_i^{\mathrm{T}}\boldsymbol{\beta}_N\right\}\right].$$

Bibliographie

Abadie, A., et Imbens, G.W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74, 235-267.

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

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

Belloni, A., Chernozhukov, V., Chetverikov, D. et Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186, 345-366.

Beretta, L., et Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16, 198-208.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? *Social Science Computer Review*, 34, 59-77.

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

Breidt, F.J., McVey, A. et Fuller, W.A. (1996). Two-phase estimation by imputation. *Journal of the Indian Society of Agricultural Statistics*, 49, 79-90.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J. et Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163, 1149-1156.

Buchanan, A.L., Hudgens, M.G., Cole, S.R., Mollan, K.R., Sax, P.E., Daar, E.S., Adimora, A.A., Eron, J.J. et Mugavero, M.J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society, Series A, (Statistics in Society)*, 181, 1193-1209.

Buelens, B., Burger, J. et van den Brakel, J.A. (2018). Comparing inference methods for nonprobability samples. *Revue Internationale de Statistique*, 86, 322-343.

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

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of American Statistical Association*, 89, 81-87.

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.

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.

Cochran, W.G. (2007). *Sampling Techniques*, New York: John Wiley & Sons, Inc.

Couper, M.P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7, 145-156.

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

DiSogra, C., Cobb, C., Chan, E. et Dennis, J.M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings (JSM), Survey Research Methods*, 4501-4515.

Dong, L., Yang, S., Wang, X., Zeng, D. et Cai, J. (2020). Integrative analysis of randomized clinical trials with real world evidence studies, *arXiv preprint arXiv:2003.01242*.

Eilers, P.H., et Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-102.

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

Fuller, W.A. (2009). *Sampling Statistics*, Wiley, Hoboken.

Hastie, T., et Tibshirani, R. (1990). *Generalized Additive Models*, NY: Chapman and Hall, Inc.

Imbens, G., Johnson, P. et Spady, R.H. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, 66, 333-357.

Kalton, G., et Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics-Theory and Methods*, 13, 1919-1939.

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

Keiding, N., et Louis, T.A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 179, 319-376.

Kim, J.K., et Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

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. (2020). Data integration by combining big data and survey sample data for finite population inference. *Revue Internationale de Statistique*. Accepté (accessible à l'adresse https://doi.org/10.1111/insr.12434).

Kim, J.K., et Wang, Z. (2019). Sampling techniques for big data analysis. *Revue Internationale de Statistique*, 87, 177-191.

Kitamura, Y., et Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65, 861-874.

Klenke, A. (2006). *Probability Theory*, Springer-Verlag: Heidelberg.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2006002/article/9547-fra.pdf.

Kweon, S., Kim, Y., jin Jang, M., Kim, Y., Kim, K., Choi, S., Chun, C., Khang, Y.-H. et Oh, K. (2014). Data resource profile: The Korea National Health and Nutrition Examination Survey (KNHANES). *International Journal of Epidemiology*, 43, 69-77.

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.

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

Mack, Y.-P. (1981). Local properties of k-NN regression estimates. *SIAM Journal on Algebraic Discrete Methods*, 2, 311-323.

Mack, Y., et Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9, 1-15.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16, 285-292.

McRoberts, R.E., Tomppo, E.O. et Næsset, E. (2010). Advances and emerging issues in national forest inventories. *Scandinavian Journal of Forest Research*, 25, 368-381.

Morikawa, K., et Kim, J.K. (2018). A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse. *Statistics & Probability Letters*, 140, 1-6.

Mulry, M.H., Oliver, B.E. et Kaputa, S.J. (2014). Detecting and treating verified influential values in a Monthly Retail Trade Survey. *Journal of Official Statistics*, 30, 721-747.

National Health Insurance Data Sharing Service (2014). National health screening data. https://nhiss.nhis.or.kr/bd/ab/bdabf006cv.do, [Consulté le : 2019-07-11].

Nelder, J.A., et Baker, R.J. (1972). *Generalized Linear Models*, New York: John Wiley & Sons, Inc.

Newey, W.K., et Smith, R.J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72, 219-255.

O'Muircheartaigh, C., et Hedges, L.V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 63, 195-210.

Palmer, J.R., Espenshade, T.J., Bartumeus, F., Chung, C.Y., Ozgencil, N.E. et Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. *Demography*, 50, 1105-1128.

Park, M., et Fuller, W.A. (2012). Generalized regression estimators. *Encyclopedia of Environmetrics*, 2, 1162-1166.

Pfeffermann, D., Eltinge, J.L. et Brown, L.D. (2015). Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, 3, 425-483.

Rao, J.N.K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*. pp. DOI 10.1007/s13571–020–00227–w.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*, New York: John Wiley & sons, Inc.

Riddles, M.K., Kim, J.K. et Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, 4, 215-245.

Rivers, D. (2007). Sampling for web surveys. *Joint Statistical Meetings*.

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

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Ruppert, D., Wand, M.P. et Carroll, R.J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics*, 3, 1193-1256.

Särndal, C.-E., Swensson, B. et Wretman, J. (2003). *Model-Assisted Survey Sampling*. Springer Science & Business Media, New York: Springer-Verlag.

Schennach, S.M. (2007). Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics*, 35, 634-672.

Stuart, E.A., Bradshaw, C.P. et Leaf, P.J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16, 475-485.

Stuart, E.A., Cole, S.R., Bradshaw, C.P. et Leaf, P.J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 174, 369-386.

Tam, S.-M., et Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *Revue Internationale de Statistique*, 83, 436-448.

Tam, S.-M., et Kim, J.-K. (2018). Big data ethics and selection-bias: An official statistician's perspective. *Statistical Journal of the IAOS*, 34, 577-588.

Tourangeau, R., Conrad, F.G. et Couper, M.P. (2013). *The Science of Web Surveys*, New York: Oxford University Press.

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

van der Vaart, A.W. (2000). *Asymptotic Statistics*, Cambridge University Press, Cambridge, MA.

Vavreck, L., et Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, 18, 355-366.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC.

Yang, S., et Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115, 1540-1554.

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 Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high-dimensional data. *Journal of the Royal Statistical Society, Series B, (Statistical Methodology)*, 82, 445-465.

Yang, S., Zeng, D. et Wang, X. (2020). Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding, *arXiv preprint arXiv:2007.12922*.

Techniques d'enquête, juin 2021
Vol. 47, nº 1, pp. 65-81
Statistique Canada, nº 12-001-X au catalogue

La méthode de vraisemblance empirique fondée sur l'échantillon sous un plan de sondage complexe avec réponses brouillées

Sixia Chen, Yichuan Zhao et Yuke Wang[1]

Résumé

L'utilisation de réponses brouillées est un moyen efficace d'effectuer le contrôle de la divulgation statistique. Les réponses brouillées peuvent être produites au moyen d'un procédé aléatoire contrôlé. Dans le présent article, nous proposons d'utiliser la méthode de la vraisemblance empirique fondée sur l'échantillon pour effectuer des inférences statistiques dans le cadre d'un plan de sondage complexe avec réponses brouillées. Plus précisément, nous proposons d'utiliser un intervalle de confiance de type Wilk pour l'inférence statistique. La méthode proposée peut servir d'outil général pour l'inférence en présence de fichiers de données d'enquête confidentielles à grande diffusion. Les propriétés asymptotiques sont produites, et l'étude par simulations limitée confirme la validité de la théorie. De plus, nous appliquons la méthode proposée à des situations réelles.

Mots-clés : Vraisemblance empirique; réponses brouillées; contrôle de la divulgation statistique; données d'enquête.

1. Introduction

Il a été démontré que la technique d'échantillonnage d'enquête est un des moyens les plus efficaces de collecte de renseignements représentatifs pour la population d'intérêt étudiée sous-jacente; voir Kish (1965) et Cochran (1977), entre autres. Cette approche est courante en pratique pour obtenir des données importantes sur la santé, l'économie sociale et l'opinion publique. Toutefois, la collecte de données au moyen d'un plan d'échantillonnage complexe sans contrôle minutieux de la divulgation statistique peut entraîner un taux de réponse faible et des erreurs de mesure importantes (Hundepool, Domingo-Ferrer, Franconi, Giessing, Nordholt, Spicer et Wolf, 2012). Le contrôle de la divulgation statistique (CDS) est défini comme une des étapes nécessaires à la publication des fichiers à grande diffusion par des organismes comme le *US Census Bureau*. Par exemple, Krenzke, Li, Freedman, Judkins, Hubble, Roisman et Larsen (2011) ont généré des produits de données sur le transport à partir de l'*American Community Survey* (ACS, Enquête sur les collectivités américaines) qui sont conformes aux règles de la divulgation. Gouweleeuw, Kooiman, Willenborg et Wolf (1998) ont discuté de la protection des données statistiques à *Statistics Netherlands*.

Le CDS repose sur l'idée de générer une certaine perturbation à partir du fichier de données brutes original de telle sorte que le risque d'identification des personnes soit minuscule, et que l'utilité du fichier de données perturbées soit élevée. Il existe de nombreuses méthodes de CDS, comme le groupement des données, la suppression de variables, la permutation des données (Fienberg et McIntyre, 2005), le remplacement séquentiel multivarié fondé sur un modèle paramétrique (Raghunathan, Lepkowski, van Hoewyk et Solenberger, 2001) et les réponses brouillées ou méthodes fondées sur la réponse aléatoire

1. Sixia Chen, Département de biostatistique et d'épidémiologie, University of Oklahoma, Oklahoma City, OK 73104, États-Unis. Courriel : sixia-chen@ouhsc.edu; Yichuan Zhao, Département de mathématiques et statistique, Georgia State University, Atlanta, GA 30303, États-Unis; Yuke Wang, Département de mathématiques et statistique, Georgia State University, Atlanta, GA 30303, États-Unis.

(Horvitz, Shah et Simmons, 1967; Fox et Tracy, 1986). Pour en savoir plus sur ces méthodes, voir Hundepool et coll. (2012).

L'inférence après le CDS est un problème à la fois important et complexe. Toute analyse statistique qui ne tient pas compte du CDS donne lieu à une estimation de la variance biaisée (Raghunathan, Reiter et Rubin, 2003). Raghunathan et coll. (2003) ont proposé d'utiliser la procédure d'imputation multiple (IM) pour générer des fichiers de données perturbées, et d'effectuer l'inférence au moyen de la formule d'estimateur de la variance de Rubin. Cependant, la plupart des organismes cherchent à produire un seul fichier à grande diffusion, plutôt que de nombreux fichiers, et la validité de l'IM dépend de la condition de compatibilité bien connue de Meng (1994). Il se peut que cette condition ne se vérifie pas dans un plan d'échantillonnage informatif (Kim et Yang, 2017). Comparativement à d'autres approches, la méthode fondée sur les réponses brouillées est très facile à mettre en œuvre et représente un bon compromis entre risque et utilité. De plus, elle permet d'établir une inférence statistique valide pour la plupart des plans de sondage complexes. Warner (1965) a d'abord proposé d'utiliser un procédé de randomisation, comme un jeu de cartes, pour estimer la proportion de caractéristiques sensibles, comme les avortements provoqués, les médicaments utilisés, etc. Tracy et Mangat (1996) ont présenté une revue complète des méthodes fondées sur la réponse aléatoire. Citons le modèle multiplicatif examiné par Eichhorn et Hayre (1983) comme méthode de réponse aléatoire efficace (technique des réponses brouillées). Bar-Lev, Bobovitch et Boukai (2004) ont proposé une version améliorée de ce modèle. Saha (2011) a présenté une technique optionnelle de réponse aléatoire brouillée pour les enquêtes pratiques. Plus récemment, Singh et Kim (2011) ont proposé d'utiliser un estimateur par la vraisemblance pseudo-empirique avec un plan d'échantillonnage aléatoire simple sans remise (EASSR) selon ce modèle. Toutefois, ils ont seulement examiné une estimation ponctuelle selon le plan EASSR, et la méthode qu'ils exposent pourrait ne pas fonctionner pour d'autres plans d'échantillonnage, comme un plan avec probabilité proportionnelle à la taille.

L'approche de la vraisemblance empirique a été proposée par Hartley et Rao (1968) et étudiée par Owen (1988, 2001) et Qin et Lawless (1994) dans des contextes statistiques classiques. Dans des contextes d'enquête complexes, Wu et Rao (2006) ont considéré une approche fondée sur la vraisemblance pseudo-empirique. Chen et Kim (2014) ont proposé des méthodes par la vraisemblance empirique fondée sur la population et l'échantillon qui sont plus efficaces que la méthode de vraisemblance pseudo-empirique en cas de plans à grande entropie. Berger and Torres (2016), Berger (2018a, 2018b) a étendu la méthode fondée sur la vraisemblance empirique fondée sur l'échantillon proposée par Chen et Kim (2014) à un contexte plus général. Dans le présent article, nous ne considérons que les plans d'échantillonnage à un degré, qui comprennent l'échantillonnage de Poisson et les plans d'échantillonnage stratifié à probabilité proportionnelle à la taille. La démarche que nous proposons peut être généralisée à un plan à plusieurs degrés au moyen de la méthode traitée dans Berger (2018b). Dans les enquêtes avec plan à plusieurs degrés, l'une des difficultés est que nous devons spécifier les conditions des probabilités d'inclusion et tenir compte de la corrélation des observations dans une même grappe aux différents degrés. Nous prenons également en compte l'estimation des intervalles au moyen de la méthode

de vraisemblance empirique fondée sur l'échantillon examinée dans Chen et Kim (2014). Après une estimation constante du facteur d'échelle, le rapport de vraisemblance pseudo-empirique corrigé converge vers une loi du khi carré standard, qui peut servir à construire l'intervalle de confiance. L'information auxiliaire agrégée externe, comme la taille de la population selon l'âge, le genre et la race, peut être naturellement incorporée à la méthode que nous présentons en vue d'améliorer l'efficacité des estimateurs proposés. Notre méthode est pratique et peut être utilisée avec la plupart des fichiers de données d'enquête à grande diffusion, comme ceux de la *National Health and Nutrition Examination Survey* (NHANES, Enquête nationale sur la santé et la nutrition des États-Unis), de la *National Health Interview Survey* (NHIS, Enquête nationale sur la santé réalisée par interviews des États-Unis) et du *Behavioral Risk Factor Surveillance System* (BRFSS, Système de surveillance des facteurs de risque comportementaux des États-Unis).

L'article est organisé comme suit. Les notations élémentaires, les questions de recherche et l'estimateur de Hájek sont présentés à la section 2. La section 3 traite de la méthode proposée de vraisemblance empirique fondée sur l'échantillon. Une étude par simulations est présentée dans la section 4. Nous appliquons les méthodes proposées aux données de la *National Health and Nutrition Examination Survey* (NHANES) de 2015-2016 dans la section 5. À la section 6, nous présentons les conclusions de l'article. Tous les détails techniques sont en annexe.

2. Préliminaires

Supposons que la population finie $\mathcal{F}_N = (X_i, Y_i, i = 1, \ldots, N)$ est générée à partir d'un modèle de super-population inconnu, où Y_i est une variable étudiée et X_i est une covariable. Pour simplifier la présentation, étant donné $\mathcal{F}_N$, on suppose qu'un échantillon aléatoire A est sélectionné à partir d'un plan de sondage non stratifié à un degré. Soit I_i l'indicateur d'échantillonnage pour l'unité i tel que $I_i = 1$ si l'unité i est sélectionnée et 0 sinon. Notons les probabilités d'inclusion de premier et de second ordre comme étant $\pi_i = E(I_i)$ et $\pi_{ij} = E(I_i I_j)$ pour $i, j = 1, \ldots, N$. Ensuite, le poids d'échantillonnage peut être écrit comme étant $d_i = \pi_i^{-1}$ et la taille de l'échantillon est $n = \sum_{i=1}^{N} I_i$. Supposons que le paramètre d'intérêt est $\theta_N = N^{-1}\sum_{i=1}^{N} Y_i$. Pour des questions de confidentialité, nous prévoyons utiliser les réponses brouillées Z_i de Y_i de sorte que $Z_i = Y_i S_i$ avec la probabilité $1 - p$ et $Z_i = Y_i$ avec la probabilité p, où $E(S_i) = a$ et $V(S_i) = b^2$ avec des valeurs connues p, a et b^2. Barlev et coll. (2004) et Singh et Kim (2011) ont examiné des modèles similaires. Au lieu d'observer directement Y_i, nous observons seulement les réponses brouillées Z_i dans le fichier de données. L'estimateur de Hájek présenté dans Hájek (1971) et Fuller (2009) est souvent utilisé dans l'analyse des données d'enquête. Dans certaines conditions de régularité, on peut démontrer que l'estimateur de type Hájek (HJ) suivant est convergent :

$$\hat{\theta}_{\mathrm{HJ}} = \frac{1}{\hat{N}} \sum_{i \in A} d_i Y_i^*, \tag{2.1}$$

où $Y_i^* = Z_i \{(1-p)\,a + p\}^{-1}$ et $\hat{N} = \sum_{i \in A} d_i$ étant donné que $E(\hat{N}) = N$ et

$$E\left(\sum_{i \in A} d_i Y_i^*\right) = \sum_{i=1}^{N} \{E(Y_i^*)\} = \sum_{i=1}^{N} \left[\{E(Y_i S_i)\,(1-p) + Y_i p\}\,\{(1-p)\,a + p\}^{-1}\right] = \sum_{i=1}^{N} Y_i.$$

Les propriétés asymptotiques de $\hat{\theta}_{\mathrm{HJ}}$ sont décrites dans le théorème 1 suivant, et la démonstration schématique se trouve à l'annexe B.

Théorème 1. *Dans les conditions de régularité énoncées à l'annexe A, $\hat{\theta}_{\mathrm{HJ}}$ possède la dilatation asymptotique suivante :*

$$\hat{\theta}_{\mathrm{HJ}} = \theta_N + \frac{1}{N} \sum_{i \in A} d_i \left(Y_i^* - \theta_N\right) + o_p\left(n^{-1/2}\right), \tag{2.2}$$

et

$$V_{\mathrm{HJ}}^{-1/2}\left(\hat{\theta}_{\mathrm{HJ}} - \theta_N\right) \to^d N(0, 1), \tag{2.3}$$

quand $n, N \to \infty$ avec

$$V_{\mathrm{HJ}} = V_1 + V_2, \tag{2.4}$$

où

$$V_1 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \left(Y_i - \theta_N\right)\left(Y_j - \theta_N\right),$$

et

$$V_2 = \frac{(1-p)\left\{p(a-1)^2 + b^2\right\}}{\{(1-p)\,a + p\}^2} \frac{1}{N^2} \sum_{i=1}^{N} Y_i^2.$$

Notons que V_1 est la variabilité sous le plan de l'estimateur de Hájek pour la moyenne de la population θ_N au moyen des valeurs vraies, et V_2 est la variabilité supplémentaire générée par l'utilisation de réponses brouillées. Selon le théorème 1, l'estimateur convergent de V_{HJ} peut s'écrire comme suit :

$$\hat{V}_{\mathrm{HJ}} = \frac{1}{\hat{N}^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{Y_i^* - \hat{\theta}_{\mathrm{HJ}}}{\pi_i} \frac{Y_j^* - \hat{\theta}_{\mathrm{HJ}}}{\pi_j} + \frac{(1-p)\left\{b^2 + p(a-1)^2\right\}}{\left(b^2 + a^2\right)(1-p) + p} \frac{1}{\hat{N}^2} \sum_{i \in A} d_i Y_i^{*2}.$$

Quand $n/N = o(1)$, le second terme ci-dessus peut être ignoré en toute sécurité. Nous pouvons alors employer un estimateur convergent par rapport au plan de sondage classique avec la variable transformée Y_i^*. Dans la prochaine section, nous proposerons d'utiliser la méthode de la vraisemblance pseudo-empirique pour construire à la fois un estimateur ponctuel et l'intervalle de confiance en présence d'une information auxiliaire agrégée.

3. Méthode proposée

L'information agrégée au niveau de la population se trouve souvent dans un recensement ou de grandes enquêtes, comme l'*American Community Survey* (ACS). Par exemple, nous pouvons connaître les chiffres de population à l'échelle nationale selon le genre, la race, le niveau de scolarité ou le niveau de revenu. L'incorporation de ces renseignements dans une estimation réduit souvent l'erreur de couverture et accroît l'efficacité des estimateurs. Dans la présente section, nous proposons d'utiliser la méthode de vraisemblance empirique fondée sur l'échantillon (VEE) proposée par Chen et Kim (2014) pour effectuer simultanément l'estimation des points et des intervalles. Supposons qu'une moyenne de population $\bar{X}_N = N^{-1}\sum_{i=1}^{N} X_i$ est connue grâce à des ressources externes. On peut alors obtenir l'estimateur VEE en maximisant la fonction de log-vraisemblance empirique fondée sur l'échantillon suivante :

$$l_s = \sum_{i \in A} \log(w_i), \tag{3.1}$$

soumise aux contraintes

$$\sum_{i \in A} w_i = 1, \quad \sum_{i \in A} w_i \pi_i^{-1}(X_i - \bar{X}_N) = 0, \quad w_i \geq 0, \tag{3.2}$$

et

$$\sum_{i \in A} w_i \pi_i^{-1}(Y_i^* - \theta) = 0. \tag{3.3}$$

En maximisant la fonction objective (3.1) soumise aux contraintes de (3.2), le poids de l'estimateur VEE peut s'écrire comme suit :

$$\hat{w}_i = \frac{1}{n} \frac{1}{1 + \hat{\lambda}\pi_i^{-1}(X_i - \bar{X}_N)},$$

avec $\hat{\lambda}$ comme multiplicateur de Lagrange, et on peut l'obtenir en résolvant la seconde contrainte de (3.2). Ensuite, selon (3.3), l'estimateur VEE de θ_N peut s'écrire comme suit :

$$\hat{\theta}_{\text{VEE}} = \frac{\sum_{i \in A} \hat{w}_i \pi_i^{-1} Y_i^*}{\sum_{i \in A} \hat{w}_i \pi_i^{-1}}.$$

Le théorème 2 suivant contient les propriétés asymptotiques de l'estimateur VEE proposé $\hat{\theta}_{\text{VEE}}$. La démonstration schématique se trouve à l'annexe C.

Théorème 2. *Dans les conditions de régularité énoncées à l'annexe A, $\hat{\theta}_{\text{VEE}}$ possède la dilatation asymptotique suivante :*

$$\hat{\theta}_{\text{VEE}} = \theta_N + \frac{1}{N}\sum_{i \in A} d_i (Y_i^* - \theta_N) - B\frac{1}{N}\sum_{i \in A} d_i (X_i - \bar{X}_N) + o_p(n^{-1/2}), \tag{3.4}$$

où

$$B = \left\{ \frac{1}{N} \sum_{i=1}^{N} \pi_i^{-1} (Y_i - \theta_N) (X_i - \bar{X}_N) \right\} \left\{ \frac{1}{N} \sum_{i=1}^{N} \pi_i^{-1} (X_i - \bar{X}_N) (X_i - \bar{X}_N)^{\mathrm{T}} \right\}^{-1}$$

et

$$V_{\mathrm{VEE}}^{-1/2} \left(\hat{\theta}_{\mathrm{VEE}} - \theta_N \right) \to^d N(0, 1),$$

quand $n, N \to \infty$ *avec*

$$V_{\mathrm{VEE}} = V_1^* + V_2,$$

où V_2 *est défini dans le théorème 1 et*

$$V_1^* = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \eta_i \eta_j,$$

avec $\eta_i = Y_i - \theta_N - B(X_i - \bar{X}_N)$.

Notons que V_1^* est la variabilité sous le plan de l'estimateur par la régression optimal qui est inférieure à V_1 définie dans le théorème 1. Fuller et Isaki (1981), Montanari (1987) et Rao (1994) ont traité de l'estimateur par la régression optimal. Selon le théorème 2, l'estimateur convergent de V_{VEE} peut s'écrire comme suit :

$$\hat{V}_{\mathrm{VEE}} = \frac{1}{\hat{N}^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\hat{\eta}_i}{\pi_i} \frac{\hat{\eta}_j}{\pi_j} + \frac{(1-p) \left\{ b^2 + p(a-1)^2 \right\}}{(b^2 + a^2)(1-p) + p} \frac{1}{\hat{N}^2} \sum_{i \in A} d_i Y_i^{*2},$$

où $\hat{\eta}_i = Y_i^* - \hat{\theta}_{\mathrm{VEE}} - \hat{B}(X_i - \bar{X}_N)$ avec

$$\hat{B} = \left\{ \sum_{i \in A} d_i^2 \left(Y_i^* - \hat{\theta}_{\mathrm{VEE}} \right) (X_i - \bar{X}_N) \right\} \left\{ \sum_{i \in A} d_i^2 (X_i - \bar{X}_N) (X_i - \bar{X}_N)^{\mathrm{T}} \right\}^{-1}.$$

Quand $n/N = o(1)$, le second terme de $\hat{V}_{\mathrm{VEE}}$ peut être ignoré. Selon le plan d'échantillonnage aléatoire simple (EAS), on peut montrer que $\hat{\theta}_{\mathrm{VEE}}$ est asymptotiquement équivalent à l'estimateur par la régression bien connu ci-dessous :

$$\hat{\theta}_{\mathrm{REG}} = \frac{1}{\hat{N}} \sum_{i \in A} d_i Y_i^* - \hat{B}_R \left(\frac{1}{\hat{N}} \sum_{i \in A} d_i X_i - \bar{X}_N \right), \tag{3.5}$$

où

$$\hat{B}_R = \left\{ \sum_{i \in A} d_i \left(Y_i^* - \hat{\theta}_{\mathrm{HJ}} \right) (X_i - \bar{X}_N) \right\} \left\{ \sum_{i \in A} d_i (X_i - \bar{X}_N) (X_i - \bar{X}_N)^{\mathrm{T}} \right\}^{-1}.$$

Cependant, dans un plan général, $\hat{\theta}_{\mathrm{VEE}}$ est différent de $\hat{\theta}_{\mathrm{REG}}$. Dans le plan d'échantillonnage de Poisson, on peut montrer que $\hat{\theta}_{\mathrm{VEE}}$ est plus efficace que $\hat{\theta}_{\mathrm{REG}}$. Le théorème 1 et le théorème 2 peuvent servir à construire un intervalle de confiance de type Wald pour θ_N. Le théorème 3 suivant peut servir à construire un intervalle de confiance de type Wilk. La démonstration schématique du théorème 3 se trouve à l'annexe D.

Théorème 3. *Définissons* $R_n(\theta_N) = 2\left\{l_s\left(\hat{\theta}_{\mathrm{VEE}}\right) - l_s(\theta_N)\right\}$, *où* $l_s(\theta)$ *est défini dans (3.1) avec* w_i *satisfaisant (3.2) et (3.3). Alors, dans les conditions de régularité énumérées à l'annexe A, quand* $n, N \to \infty$, $c_1 c_2^{-1} R_n(\theta_N) \to^d \chi_1^2$, *où* $c_1 = N^{-2}\sum_{i=1}^N \pi_i^{-1}\eta_i^{*2}$ *avec* $\eta_i^* = Y_i^* - \theta_N - B\left(X_i - \bar{X}_N\right)$ *et* $c_2 = V_{\mathrm{VEE}}$.

L'estimateur de c_1 et c_2 peut s'écrire comme étant

$$\hat{c}_1 = \hat{N}^{-2}\sum_{i\in A}\pi_i^{-2}\left\{Y_i^* - \hat{\theta}_{\mathrm{VEE}} - \hat{B}\left(X_i - \bar{X}_N\right)\right\}^2,$$

et $\hat{c}_2 = \hat{V}_{\mathrm{VEE}}$. Le théorème 3 peut servir à construire un intervalle de confiance de type Wilk pour θ_N.

4. Étude par simulations

Dans l'étude par simulations, nous considérons la population finie (X_i, Y_i), $i = 1, 2, \ldots, N$ pour $N = 10\,000$. X_i est distribué uniformément sur $[0, 1]$ et $Y_i = m(X_i) + \varepsilon_i$ avec $\varepsilon_i \sim N(0; 0{,}01)$. Quatre fonctions $m(x)$ sont énoncées ci-dessous :

(A). $m_1(x) = 2 + 2(x - 0{,}5)$,

(B). $m_2(x) = 2 + 2(x - 0{,}5)^2$,

(C). $m_3(x) = 2 + 2(x - 0{,}5) + \exp\left(-200(x - 0{,}5)^2\right)$,

(D). $m_4(x) = 2 + 2(x - 0{,}5)\Delta(x < 0{,}6) + 0{,}6\Delta(x \geq 0{,}6)$, où $\Delta(B)$ est la fonction de l'indicateur binaire pour la condition B de sorte que $\Delta(B) = 1$ si la condition B est satisfaite et 0 sinon.

Nous avons généré $B = 5\,000$ échantillons de Monte Carlo à partir de l'échantillonnage de Poisson avec des probabilités d'inclusion $\pi_i = nk_i \Big/ \sum_{j=1}^N k_j$, où la variable de taille $k_j = \max\left(0{,}5Y_j + 2, 1\right) + u_j$ avec $u_j \sim \chi^2(1)$. Nous avons examiné les tailles d'échantillon $n = 40$, 50, 100 et 200. Pour chaque échantillon de Monte Carlo, les réponses brouillées Z_i ont été générées avec $p = 0{,}6$ et $S_i \sim N(1{,}5; 0{,}2/1{,}5)$. Supposons que nous observons seulement X_i et Z_i dans l'échantillon. Les performances de l'estimateur HJ et de l'estimateur VEE proposé ont été comparées à la moyenne de la population estimée de Y, qui est $\theta_0 = E(Y)$. Les résultats sont indiqués dans le tableau 4.1.

Nous avons calculé le biais Monte Carlo $\mathrm{BMC} = B^{-1}\sum_{b=1}^B \hat{\theta}_b - \theta_0$, l'erreur-type Monte Carlo $\mathrm{ETMC} = \left\{B^{-1}\sum_{b=1}^B\left(\hat{\theta}_b - \bar{\theta}\right)^2\right\}^{1/2}$ avec $\bar{\theta} = B^{-1}\sum_{b=1}^B \hat{\theta}_b$ et l'erreur quadratique moyenne Monte Carlo $\mathrm{EQMMC} = \left\{B^{-1}\sum_{b=1}^B\left(\hat{\theta}_b - \theta_0\right)^2\right\}^{1/2}$. Pour l'estimation de la variance, nous avons calculé le taux de couverture, la longueur moyenne des estimations d'intervalle et le pourcentage de biais relatif des estimateurs de la variance $\mathrm{BR} = 100 \times \left[\left(B^{-1}\sum_{b=1}^B \hat{V}_b\right)\left\{B^{-1}\sum_{b=1}^B\left(\hat{\theta}_b - \bar{\theta}\right)^2\right\}^{-1} - 1\right]$. Les résultats de la simulation sont présentés dans le tableau 4.1.

Tableau 4.1
Résultats de simulation du biais Monte Carlo (BMC), de l'erreur-type Monte Carlo (ETMC) et de l'erreur quadratique moyenne Monte Carlo (EQMMC), du taux de couverture, de la longueur moyenne des intervalles de confiance de 95 % et du biais relatif (BR) pour l'estimateur de Hájek (HJ) et l'estimateur par la vraisemblance empirique fondée sur l'échantillon (VEE)

Scénario		BMC		ETMC		EQMMC		Taux de couverture		Longueur moyenne		BR	
Modèle	***n***	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**
m_1	40	0,0035	0,0005	0,123	0,076	0,015	0,006	0,936	0,940	0,470	0,283	-0,027	-0,075
	50	0,0026	0,0006	0,110	0,069	0,012	0,005	0,939	0,941	0,420	0,255	-0,024	-0,078
	100	0,0009	0,0003	0,077	0,048	0,006	0,002	0,946	0,950	0,300	0,183	0,007	-0,000
	200	0,0006	-0,0002	0,054	0,033	0,003	0,001	0,944	0,954	0,211	0,130	-0,010	0,000
m_2	40	0,0006	0,0007	0,083	0,085	0,007	0,007	0,937	0,937	0,319	0,314	-0,020	-0,098
	50	-0,0004	-0,0008	0,074	0,075	0,005	0,006	0,939	0,944	0,286	0,283	-0,014	-0,066
	100	-0,0002	-0,0001	0,053	0,053	0,003	0,003	0,941	0,947	0,203	0,203	-0,036	-0,057
	200	-0,0007	-0,0006	0,037	0,037	0,001	0,001	0,945	0,949	0,144	0,144	0,002	-0,013
m_3	40	0,0022	0,0011	0,138	0,091	0,019	0,008	0,926	0,939	0,512	0,344	-0,081	-0,068
	50	0,0056	0,0028	0,119	0,081	0,014	0,007	0,941	0,942	0,460	0,312	-0,018	-0,045
	100	0,0011	0,0003	0,084	0,058	0,007	0,003	0,945	0,943	0,327	0,222	-0,011	-0,053
	200	-0,0002	-0,0006	0,059	0,041	0,003	0,002	0,950	0,952	0,230	0,157	-0,010	-0,028
m_4	40	0,0040	0,0012	0,119	0,080	0,014	0,006	0,938	0,937	0,460	0,296	-0,007	-0,089
	50	0,0008	0,0002	0,107	0,071	0,012	0,005	0,943	0,943	0,413	0,267	-0,020	-0,069
	100	0,0007	0,0006	0,075	0,049	0,006	0,002	0,942	0,945	0,293	0,190	-0,013	-0,036
	200	-0,0003	-0,0002	0,053	0,034	0,003	0,001	0,946	0,957	0,206	0,135	-0,018	0,029

Pour les modèles m_1, m_3 et m_4, VEE présente un biais Monte Carlo, une erreur-type Monte Carlo et une erreur quadratique moyenne Monte Carlo plus petits, en particulier pour les petites tailles d'échantillon (n = 40 ou 50). Pour le modèle m_2, les deux méthodes ont des performances comparables. Pour les quatre modèles, nous avons constaté que, dans la plupart des cas (14 sur 16), les estimateurs VEE donnaient un taux de couverture supérieur ou égal à celui de l'estimateur HJ, tandis que la longueur moyenne de l'intervalle de confiance était plus courte comparativement à celle obtenue au moyen de l'estimateur HJ. Les deux méthodes ont produit de petits biais relatifs des estimateurs de la variance. Dans l'ensemble, l'estimateur VEE proposé surpassait l'estimateur HJ dans la plupart des cas.

Pour tester la sensibilité de la méthode proposée, nous avons ajouté du bruit, W_i, à la simulation dans les configurations actuelles de l'étude par simulations. Ensuite, $Y_i = m(\alpha X_i + (1-\alpha)W_i) + \varepsilon_i$ avec α = 0; 0,1; 0,3; 0,5; 0,7; 0,9; 1, $X_i \sim \text{Uniforme}(0; 1)$, $W_i \sim N(0; 1)$, et $\varepsilon_i \sim N(0; 0{,}01)$. Supposons que nous observons seulement X_i et Z_i (la réponse brouillée de Y_i) dans l'échantillon et comparons de nouveau l'estimateur HJ et l'estimateur VEE. Les résultats sont présentés dans les tableaux 4.2 et 4.3. Nous avons constaté que quand α diminue, les taux de couverture de l'estimateur VEE sont plus petits que ceux de l'estimateur HJ, et que la longueur moyenne de l'IC pour l'estimateur VEE n'est pas plus courte que celle de l'estimateur HJ. Par conséquent, l'estimateur VEE est plus performant que l'estimateur HJ, à condition que la majeure partie de l'information soit contenue dans la covariable actuelle.

Tableau 4.2
Résultats de simulation de l'estimateur de Hájek (HJ) et de l'estimateur par la vraisemblance empirique fondée sur l'échantillon (VEE) après l'ajout de bruit

Scénario		$\alpha = 0$				$\alpha = 0{,}1$				$\alpha = 0{,}3$			
		Taux de couverture		Longueur moyenne		Taux de couverture		Longueur moyenne		Taux de couverture		Longueur moyenne	
Modèle	***n***	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**
m_1	40	0,924	0,903	1,419	1,368	0,926	0,911	1,289	1,251	0,938	0,928	1,045	1,022
	50	0,926	0,915	1,292	1,256	0,928	0,920	1,146	1,125	0,937	0,930	0,958	0,938
	100	0,940	0,935	0,927	0,927	0,941	0,935	0,839	0,838	0,948	0,943	0,679	0,668
	200	0,949	0,941	0,651	0,657	0,942	0,943	0,589	0,593	0,948	0,948	0,478	0,469
m_2	40	0,942	0,943	1,872	1,909	0,929	0,930	1,328	1,358	0,933	0,933	1,455	1,458
	50	0,935	0,937	1,704	1,732	0,933	0,937	1,181	1,206	0,931	0,935	1,327	1,325
	100	0,941	0,947	1,191	1,202	0,942	0,949	0,843	0,854	0,945	0,948	0,931	0,927
	200	0,949	0,952	0,841	0,845	0,949	0,955	0,593	0,597	0,948	0,948	0,645	0,640
m_3	40	0,917	0,899	1,438	1,382	0,925	0,906	1,313	1,273	0,933	0,922	1,044	1,020
	50	0,922	0,908	1,297	1,264	0,928	0,916	1,154	1,131	0,939	0,935	0,927	0,911
	100	0,937	0,928	0,960	0,958	0,941	0,935	0,838	0,838	0,940	0,938	0,660	0,654
	200	0,940	0,940	0,674	0,679	0,945	0,944	0,615	0,619	0,945	0,941	0,474	0,467
m_4	40	0,903	0,885	1,226	1,167	0,912	0,894	0,994	0,947	0,927	0,909	0,518	0,511
	50	0,921	0,912	1,093	1,057	0,917	0,912	0,902	0,870	0,928	0,918	0,460	0,457
	100	0,931	0,925	0,805	0,802	0,936	0,935	0,646	0,644	0,935	0,931	0,337	0,338
	200	0,941	0,939	0,581	0,585	0,936	0,939	0,460	0,462	0,945	0,946	0,236	0,237

Tableau 4.3
Résultats de simulation de l'estimateur de Hájek (HJ) et de l'estimateur par la vraisemblance empirique fondée sur l'échantillon (VEE) après l'ajout de bruit

Scénario		$\alpha = 0{,}5$				$\alpha = 0{,}7$				$\alpha = 0{,}9$			
		Taux de couverture		Longueur moyenne		Taux de couverture		Longueur moyenne		Taux de couverture		Longueur moyenne	
Modèle	***n***	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**
m_1	40	0,934	0,933	1,002	0,934	0,933	0,935	1,091	0,959	0,937	0,940	1,292	1,096
	50	0,935	0,936	0,902	0,841	0,939	0,935	0,979	0,862	0,936	0,938	1,156	0,986
	100	0,947	0,948	0,635	0,596	0,944	0,949	0,697	0,616	0,946	0,948	0,820	0,705
	200	0,951	0,949	0,451	0,421	0,947	0,945	0,493	0,437	0,951	0,951	0,579	0,500
m_2	40	0,933	0,936	2,371	2,139	0,938	0,934	3,418	2,469	0,933	0,942	5,095	2,980
	50	0,940	0,941	2,148	1,938	0,940	0,937	3,057	2,210	0,945	0,944	4,583	2,687
	100	0,939	0,941	1,493	1,345	0,948	0,946	2,196	1,588	0,948	0,951	3,223	1,916
	200	0,942	0,942	1,054	0,938	0,944	0,947	1,545	1,113	0,949	0,947	2,264	1,356
m_3	40	0,939	0,935	1,004	0,940	0,935	0,937	1,101	0,970	0,939	0,947	1,288	1,093
	50	0,937	0,935	0,890	0,832	0,938	0,942	0,978	0,864	0,936	0,940	1,152	0,982
	100	0,946	0,945	0,635	0,595	0,951	0,952	0,698	0,616	0,948	0,952	0,821	0,706
	200	0,949	0,950	0,450	0,420	0,943	0,948	0,493	0,437	0,952	0,952	0,579	0,500
m_4	40	0,937	0,942	0,365	0,358	0,936	0,941	0,362	0,354	0,932	0,938	0,362	0,354
	50	0,935	0,939	0,326	0,322	0,939	0,943	0,325	0,320	0,938	0,947	0,324	0,320
	100	0,941	0,948	0,232	0,230	0,948	0,953	0,230	0,229	0,941	0,946	0,231	0,229
	200	0,947	0,948	0,165	0,164	0,942	0,944	0,163	0,163	0,949	0,951	0,163	0,163

5. Application à des données réelles

Dans la présente section, nous avons appliqué la démarche proposée à la *National Health and Nutrition Examination Survey* (NHANES, Enquête nationale sur la santé et la nutrition) de 2015-2016 pour évaluer ses propriétés en pratique. Le NHANES fournit régulièrement de l'information sur la santé et la nutrition de la population civile résidant hors établissement des États-Unis. Elle emploie un plan d'échantillonnage probabiliste complexe à plusieurs degrés fondé sur des données recuillies au moyen d'une enquête menée en personne. (Voir https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2015 pour en savoir plus.) La taille de l'échantillon de la NHANES de 2015-2016 est d'environ 9 000 personnes. Nous avons traité l'échantillon original de l'enquête NHANES comme une population finie et avons sélectionné un échantillon au moyen d'un plan d'échantillonnage aléatoire simple avec des tailles d'échantillon (n) de 30, 40, 50, 100 et 200, respectivement. Supposons que nos paramètres d'intérêt comprennent la tension artérielle systolique, la tension artérielle diastolique, le cholestérol HDL et le cholestérol total. Nous avons créé des réponses brouillées pour ces paramètres en utilisant $p = 0{,}6$; $a = 1{,}5$ et $b^2 = 0{,}2/1{,}5$. De plus, l'indice de masse corporelle (IMC) a été sélectionné comme covariable dans le processus d'estimation, puisque l'IMC est corrélé à ces variables de l'étude.

Nous avons comparé les performances de deux méthodes, HJ et VEE, pour ce qui est des estimations ponctuelles et des estimations d'intervalles (tableau 5.1). Les estimations ponctuelles obtenues au moyen des deux méthodes étaient similaires, et elles étaient proches des paramètres de population définis (120,47; 66,17; 54,43 et 180,25 pour la tension artérielle systolique, la tension artérielle diastolique, le cholestérol HDL et le cholestérol total), surtout pour les échantillons de grande taille (tableau 5.1). Concernant la tension artérielle systolique, la tension artérielle diastolique et le cholestérol total, les intervalles produits par VEE sont légèrement asymétriques à droite comparativement aux résultats produits par HJ quand les échantillons sont de petite taille. Toutefois, quand la taille des échantillons augmente, les résultats des deux méthodes sont similaires. Concernant le cholestérol HDL, les résultats sont comparables. Les résultats obtenus dans cette application ont permis de vérifier la validité de la méthode proposée d'estimation par la vraisemblance empirique fondée sur l'échantillon.

Tableau 5.1
Estimations ponctuelles et IC de 95 % aux fins d'estimation des moyennes de différents résultats à l'aide des résultats des réponses brouillées et de l'IMC provenant des données de la NHANES

	Tension artérielle systolique en mm Hg		Tension artérielle diastolique en mm Hg		Cholestérol HDL en mg/dL		Cholestérol total en mg/dL	
n	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**	**HJ**	**VEE**
30	124,5	124,5	67,7	69,4	57,9	57,6	187,0	188,3
	(112,3; 136,8)	(113,5; 139,6)	(61,5; 73,8)	(63,9; 75,2)	(50,3; 65,5)	(50,8; 65,9)	(160,0; 214,0)	(166,6; 225,5)
40	125,6	125,5	70,2	70,2	52,0	51,2	178,7	178,1
	(115,4; 135,8)	(116,5; 136,1)	(64,6; 75,8)	(64,9; 76,1)	(48,0; 56,0)	(47,3; 55,8)	(160,6; 196,8)	(162,1; 199,0)
50	118,3	116,9	67,1	67,1	57,1	56,8	173,7	173,3
	(110,2; 126,4)	(109,0; 126,1)	(60,9; 73,3)	(61,4; 73,8)	(50,8; 63,4)	(51,3; 63,2)	(160,2; 187,1)	(161,2; 187,8)
100	120,8	120,5	70,0	69,7	52,3	52,4	173,1	172,8
	(115,1; 126,5)	(115,1; 126,3)	(65,9; 74,0)	(65,9; 73,6)	(48,9; 55,7)	(49,2; 55,9)	(163,5; 182,7)	(164,0; 183,2)
200	124,1	123,9	67,6	67,5	54,0	53,8	181,4	181,5
	(119,4; 128,9)	(119,4; 128,8)	(64,9; 70,3)	(64,8; 70,3)	(51,1; 56,8)	(51,3; 56,5)	(172,7; 190,1)	(173,3; 190,9)

6. Conclusion

Dans l'article, nous avons proposé une méthode de vraisemblance empirique fondée sur l'échantillon (VEE) utilisant des réponses brouillées pour protéger la confidentialité des données d'enquête complexes. La méthode de VEE proposée est facile à mettre en œuvre en pratique et peut servir d'instrument général pour le contrôle de la divulgation statistique. La méthode proposée repose sur l'idée de remplacer les vraies valeurs par certaines valeurs brouillées au moyen d'un procédé aléatoire, puis d'appliquer l'approche de VEE à des valeurs brouillées pour obtenir l'estimation ponctuelle. Cependant, l'estimation de la variance et l'estimation de l'intervalle de confiance diffèrent de cela, en raison du traitement des valeurs brouillées comme des valeurs vraies, puisque nous devons incorporer le caractère aléatoire causé par le procédé aléatoire dans l'inférence statistique. Ces propriétés théoriques ont été étudiées et vérifiées au moyen d'une étude par simulations et d'applications sur des données réelles. L'estimateur VEE surpasse les approches classiques, comme l'estimateur HJ, en améliorant les taux de couverture et en réduisant la longueur des intervalles de confiance. Chen et Kim (2014) ont comparé les IC de type Wald et les IC de type Wilk dans des études par simulations en utilisant la méthode de VEE. En général, les intervalles de confiance de type Wilk présentent de meilleures propriétés de couverture que les intervalles de confiance de type Wald, pour ce qui est des taux de couverture. Nous nous attendons à obtenir des résultats semblables au moyen des méthodes proposées ici. Dans le cadre de futurs travaux, nous élargirons l'approche proposée pour estimer des paramètres plus généraux, comme les quantiles de population et les fonctions de distribution. Les outils informatiques statistiques correspondants, comme le module R, seront également élaborés.

Remerciements

Les travaux du Dr. Sixia Chen ont été partiellement appuyés par les *Oklahoma Shared Clinical and Translational Resources* (U54GM104938) avec un *Institutional Development Award* (IDeA) du *National Institute of General Medical Sciences*. Le contenu est la seule responsabilité des auteurs et ne représente pas nécessairement les points de vue officiels des *National Institues of Health.* Les travaux de Yichuan Zhao sont appuyés par une subvention de la *National Security Agency* (H98230-12-1-0209) et une subvention de la *National Science Foundation* (DMS-1613176).

Annexe

A. Conditions de régularité

Nous présentons ci-dessous les conditions de régularité nécessaires à la démonstration des théorèmes 1 à 3 :

(C1). $c_1 < \pi_i N n^{-1} < c_2$ pour $i = 1, 2, \ldots, N$ avec $0 < c_1 < c_2$.

(C2). $n^{1/2}\left(N^{-1}\sum_{i=1}^{N} I_i \pi_i^{-1} Y_i - N^{-1}\sum_{i=1}^{N} Y_i\right) \to^d N(0, V_1)$ quand $n \to \infty$ et $N \to \infty$, où $V_1 = nN^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i \pi_j) d_i Y_i d_j Y_j$.

(C3). $n^{1/2}\left(N^{-1}\sum_{i=1}^{N} I_i \pi_i^{-1} X_i - \frac{1}{N}\sum_{i=1}^{N} X_i\right) \to^d N(0, V_2)$ quand $n \to \infty$ et $N \to \infty$, où $V_2 = nN^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i \pi_j) d_i X_i d_j X_j^{\top}$.

(C4). $N^{-1}\sum_{i=1}^{N}|Y_i|^4$ et $N^{-1}\sum_{i=1}^{N}\|X_i\|^4$ sont bornés.

(C5). $\max_{i\in A}|Y_i| = o_p(n^{1/2})$ et $\max_{i\in A}\|X_i\| = o_p(n^{1/2})$.

B. Démonstration schématique du théorème 1

$\hat{\theta}_{\mathrm{HJ}}$ peut s'écrire comme la solution de l'équation d'estimation $\hat{U}_{\mathrm{HJ}}(\theta) = 0$, où

$$\hat{U}_{\mathrm{HJ}}(\theta) = \frac{1}{N}\sum_{i\in A} d_i (Y_i^* - \theta).$$

Sous les hypothèses que $\hat{U}_{\mathrm{HJ}}(\theta)$ converge vers $U_{\mathrm{HJ}}(\theta) = N^{-1}\sum_{i=1}^{N}(Y_i - \theta)$ uniformément, $E(Y^2) < \infty$, et puisque $U_{\mathrm{HJ}}(\theta_N) = 0$, on peut montrer que $\hat{\theta}_{\mathrm{HJ}} \to^p \theta_N$. Au moyen d'un développement en séries de Taylor,

$$0 = \hat{U}_{\mathrm{HJ}}(\hat{\theta}_{\mathrm{HJ}}) = \hat{U}_{\mathrm{HJ}}(\theta_N) + \frac{\partial \hat{U}_{\mathrm{HJ}}(\theta_N)}{\partial \theta}(\hat{\theta}_{\mathrm{HJ}} - \theta_N) + o_p(n^{-1/2}).$$

Avec un peu d'algèbre, on peut montrer que

$$\hat{\theta}_{\mathrm{HJ}} = \theta_N + \frac{1}{N}\sum_{i\in A} d_i (Y_i^* - \theta_N) + o_p(n^{-1/2}).$$

Parce que

$$E(Y_i^*) = Y_i, \quad V(Y_i^*) = \frac{(1-p)\left\{b^2 + p(a-1)^2\right\}}{\{(1-p)a + p\}^2} Y_i^2, \tag{B.1}$$

$$V\left\{\frac{1}{N}\sum_{i\in A} d_i (Y_i^* - \theta_N)\right\} = E\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}(Y_i^* - \theta_N)(Y_j^* - \theta_N)\right\}$$
$$+ V\left\{\frac{1}{N}\sum_{i=1}^{N}(Y_i^* - \theta_N)\right\}. \tag{B.2}$$

Selon (B.1), (B.2) et après quelques manipulations algébriques, nous pouvons montrer que

$$V\left\{\frac{1}{N}\sum_{i\in A} d_i (Y_i^* - \theta_N)\right\} = V_{\mathrm{HJ}},$$

où V_{HJ} est défini dans l'équation (2.4). Dans les conditions de régularité de Fuller et Isaki (1981), on peut calculer la normalité asymptotique.

C. Démonstration schématique du théorème 2

Définissons

$$\hat{U}_1(\lambda) = \frac{1}{N}\sum_{i\in A}\frac{\pi_i^{-1}(X_i - \bar{X}_N)}{1+\lambda\pi_i^{-1}(X_i - \bar{X}_N)}$$

et

$$\hat{U}_2(\lambda, \theta) = \frac{1}{N}\sum_{i\in A}\frac{\pi_i^{-1}(Y_i^* - \theta)}{1+\lambda\pi_i^{-1}(X_i - \bar{X}_N)}.$$

Alors, $\hat{\theta}_{\text{VEE}}$ et $\hat{\lambda}$ sont les solutions de $\hat{U}_1(\lambda) = \hat{U}_2(\lambda, \theta) = 0$. Au moyen de techniques similaires à celles de Chen et Kim (2014), on peut montrer que $\hat{\lambda} = O_p(n^{-1/2})$ et $\hat{\theta}_{\text{VEE}} \to^p \theta_N$. Ensuite, au moyen d'un développement en séries de Taylor, nous obtenons

$$0 = \hat{U}_1(\hat{\lambda}) = \hat{U}_1(0) + \frac{\partial\hat{U}_1(0)}{\partial\lambda}\hat{\lambda} + o_p(n^{-1/2}), \tag{C.1}$$

et

$$0 = \hat{U}_2(\hat{\lambda}, \hat{\theta}_{\text{VEE}}) = \hat{U}_2(0, \theta_N) + \frac{\partial\hat{U}_2(0, \theta_N)}{\partial\theta}(\hat{\theta}_{\text{VEE}} - \theta_N) + \frac{\partial\hat{U}_2(0, \theta_N)}{\partial\lambda}\hat{\lambda} + o_p(n^{-1/2}). \tag{C.2}$$

Selon (C.1), (C.2) et après quelques manipulations algébriques, nous pouvons montrer que

$$\hat{\lambda} = \left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}(X_i - \bar{X}_N)(X_i - \bar{X}_N)^{\mathsf{T}}\right\}^{-1}\frac{1}{N}\sum_{i\in A}d_i(X_i - \bar{X}_N) + o_p(n^{-1/2}) \tag{C.3}$$

et

$$\hat{\theta}_{\text{VEE}} - \theta_N = \frac{1}{N}\sum_{i\in A}d_i(Y_i^* - \theta_N) - B\frac{1}{N}\sum_{i\in A}d_i(X_i - \bar{X}_N) + o_p(n^{-1/2}),$$

où B est défini dans le théorème 2. Parce que

$$\begin{aligned} V(\hat{\theta}_{\text{VEE}}) &= V\left(\frac{1}{N}\sum_{i\in A}d_i\eta_i^*\right) + o(n^{-1}) = V\left(\frac{1}{N}\sum_{i\in A}d_i\eta_i\right) + E\left\{V\left(\frac{1}{N}\sum_{i\in A}d_i\eta_i^* \,\middle|\, A\right)\right\} \\ &= V_2 + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\eta_i\eta_j + o(n^{-1}), \end{aligned}$$

où V_2 est défini dans le théorème 1, η_i est défini dans le théorème 2 et $\eta_i^* = Y_i^* - \theta_N - B(X_i - \bar{X}_N)$. Avec un peu d'algèbre, on peut montrer que

$$\hat{V}_{\text{VEE}} = V_{\text{VEE}} + o(n^{-1}),$$

avec V_{VEE} défini dans le théorème 2. De plus, selon les conditions de régularité de Fuller et Isaki (1981), nous obtenons la normalité asymptotique.

D. Démonstration schématique du théorème 3

Parce que $\hat{\lambda} = O_p\left(n^{-1/2}\right)$ et au moyen d'un développement en séries de Taylor de $\log(1+x)$ à $x = \hat{\lambda}\pi_i^{-1}\left(X_i - \bar{X}_N\right)$ et (C.3), nous obtenons

$$\begin{aligned}
l_s\left(\hat{\theta}_{\mathrm{VEE}}\right) &= \sum_{i\in A} \log\left\{\frac{1}{n}\frac{1}{1+\hat{\lambda}\pi_i^{-1}\left(X_i - \bar{X}_N\right)}\right\} \\
&= n\log\left(\frac{1}{n}\right) - \sum_{i\in A}\log\left\{1+\hat{\lambda}\pi_i^{-1}\left(X_i - \bar{X}_N\right)\right\} \\
&= n\log\left(\frac{1}{n}\right) - \sum_{i\in A}\left\{\hat{\lambda}^{\mathrm{T}}\pi_i^{-1}\left(X_i - \bar{X}_N\right) - \frac{1}{2}\hat{\lambda}^{\mathrm{T}}\pi_i^{-2}\left(X_i - \bar{X}_N\right)^{\otimes 2}\hat{\lambda}\right\} + o_p(1) \\
&= n\log\left(\frac{1}{n}\right) - \frac{N}{2}\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i - \bar{X}_N\right)^{\mathrm{T}}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}\left(X_i - \bar{X}_N\right)^{\otimes 2}\right\}^{-1} \\
&\quad \times \frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i - \bar{X}_N\right) + o_p(1),
\end{aligned} \tag{D.1}$$

avec $a^{\otimes 2} = aa^{\mathrm{T}}$. Nous envisageons maintenant de maximiser $l_s = \sum_{i\in A}\log(w_i)$, soumis aux contraintes suivantes :

$$\sum_{i\in A} w_i = 1, \qquad \sum_{i\in A} w_i\pi_i^{-1}\left(X_i - \bar{X}_N\right) = 0, \tag{D.2}$$

et

$$\sum_{i\in A} w_i\pi_i^{-1}\eta_i^* = 0, \tag{D.3}$$

où $\eta_i^* = Y_i^* - \theta_N - B\left(X_i - \bar{X}_N\right)$. Les contraintes ci-dessus sont équivalentes aux contraintes d'origine de (3.2) et (3.3). Définissons $u_i = \left(X_i^{\mathrm{T}} - \bar{X}_N^{\mathrm{T}}, \eta_i^*\right)^{\mathrm{T}}$. Par conséquent, en utilisant un argument semblable, nous obtenons

$$\begin{aligned}
l_s(\theta_N) &= \sum_{i\in A}\log\left\{\frac{1}{n}\frac{1}{1+\hat{\lambda}(\theta_N)\pi_i^{-1}u_i}\right\} \\
&= n\log\left(\frac{1}{n}\right) - \sum_{i\in A}\log\left\{1+\hat{\lambda}(\theta_N)\pi_i^{-1}u_i\right\} \\
&= n\log\left(\frac{1}{n}\right) - \sum_{i\in A}\left\{\hat{\lambda}^{\mathrm{T}}(\theta_N)\pi_i^{-1}u_i - \frac{1}{2}\hat{\lambda}^{\mathrm{T}}(\theta_N)\pi_i^{-2}u_i^{\otimes 2}\hat{\lambda}(\theta_N)\right\} + o_p(1) \\
&= n\log\left(\frac{1}{n}\right) - \frac{N}{2}\frac{1}{N}\sum_{i\in A}\pi_i^{-1}u_i^{\mathrm{T}}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}u_i^{\otimes 2}\right\}^{-1} \times \frac{1}{N}\sum_{i\in A}\pi_i^{-1}u_i + o_p(1) \\
&= n\log\left(\frac{1}{n}\right) - \frac{N}{2}\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i - \bar{X}_N\right)^{\mathrm{T}}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}\left(X_i - \bar{X}_N\right)^{\otimes 2}\right\}^{-1} \\
&\quad \times \frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i - \bar{X}_N\right) - \frac{N}{2}\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*\mathrm{T}}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}\eta_i^{*\otimes 2}\right\}^{-1} \times \frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^* + o_p(1)
\end{aligned} \tag{D.4}$$

à condition que $\sum_{i=1}^{N}\pi_i^{-1}\left(X_i - \bar{X}_N\right)\eta_i = 0$. Selon (D.1), (D.4), et après quelques manipulations algébriques, nous obtenons

$$\frac{c_1}{c_2} R_n(\theta_N) = \frac{2c_1}{c_2}\left\{l_s\left(\hat{\theta}_{\mathrm{VEE}}\right) - l_s(\theta_N)\right\} = \frac{c_1}{c_2}\left\{\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^*\right\}^2\left(\frac{1}{N^2}\sum_{i=1}^{N}\pi_i^{-1}\eta_i^{*2}\right)^{-1}$$

$$= \left\{V\left(\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^*\right)^{-1/2}\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^*\right\}^2 \to^d \chi_1^2. \qquad \text{(D.5)}$$

Cela conclut la démonstration du théorème 3.

Bibliographie

Bar-Lev, S.K., Bobovitch, E. et Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, 60, 255-260.

Berger, Y.G. (2018a). Empirical likelihood approaches in survey sampling. *The Survey Statistician*, 78, 22-31.

Berger, Y.G. (2018b). An empirical likelihood approach under cluster sampling with missing observations. *Annals of the Institute of Statistical Mathematics*, doi:10.1007/s10463-018-0681-x.

Berger, Y.G., et Torres, O.D.L.R. (2016). An empirical likelihood approach for inference under complex sampling design. *Journal of the Royal Statistical Society, Series B*, 78(2), 319-341.

Chen, S., et Kim, J.K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, 24, 335-355.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley & Sons, Inc.

Eichhorn, B.H., et Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.

Fienberg, S.E., et McIntyre, J. (2005). Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of Official Statistics*, 21, 309-323.

Fox, J.A., et Tracy, P.E. (1986). *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills, CA: Sage.

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

Fuller, W.A., et Isaki, C.T. (1981). Survey design under superpopulation models. Dans *Current Topics in Survey Sampling*, (Éds., D. Krewski, J.N.K. Rao et R. Platek). New York: Academic Press, 199-226.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R. et Wolf, P. (1998). Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14, 463-478.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling, Part one". Dans *The Foundations of Survey Sampling*, (Éds., V.P. Godambe et D.A. Sprott), Holt, Rinehart, and Winston, 236.

Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Horvitz, D.G., Shah, B.V. et Simmons, W.R. (1967). The unrelated question randomized response model. Dans *Proceedings of the Social Statistics Section*, American Statistical Association, 65-72.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K. et Wolf, P. (2012). *Statistical Disclosure Control*, Wiley Series in Survey Methodology.

Kim, J.K., et Yang, S. (2017). A note on multiple imputation under informative sampling. *Biometrika*, 104, 221-228.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Krenzke, T., Li, J., Freedman, M., Judkins, D., Hubble, D., Roisman, R. et Larsen, M. (2011). Producing transportation data products from the American Community Survey that comply with disclosure rules. Washington, DC: National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (avec discussion). *Statistical Science*, 9, 538-573.

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

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Owen, A.B. (2001). *Empirical Likelihood*. New York: Chapman and Hall.

Qin, J., et Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22, 300-325.

Raghunathan, T.E., Reiter, J.P. et Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., et Solenberger, P. (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 1, 91-103. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001001/article/5857-fra.pdf.

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

Saha, A. (2011). An optional scrambled randomized response technique for practical surveys. *Metrika*, 73, 139-149.

Singh, S., et Kim, J.M. (2011). A pseudo-empirical log-likelihood estimator using scrambled responses. *Statistics and Probability Letters*, 81, 345-351.

Tracy, D.S., et Mangat, N.S. (1996). Some developments in randomized response sampling during the last decade-a follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4, 147-158.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

Wu, C., et Rao, J.N.K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, 34, 359-375.

Techniques d'enquête, juin 2021
Vol. 47, nº 1, pp. 83-98
Statistique Canada, nº 12-001-X au catalogue

Méthode de détermination d'une stratégie d'échantillonnage efficace et robuste en cas d'incertitude du modèle

Edgar Bueno et Dan Hedlin[1]

Résumé

Nous nous penchons sur le problème du choix d'une stratégie d'échantillonnage et, tout particulièrement, d'un plan de sondage. Nous proposons une mesure du risque, dans laquelle la minimisation de la valeur oriente le choix. La méthode repose sur un modèle de superpopulation et l'incertitude entourant ses paramètres est prise en compte grâce à une distribution a priori. L'utilisation de cette méthode est illustrée au moyen d'un ensemble de données réel, qui donne des résultats satisfaisants. Comme base de référence, nous utilisons la stratégie qui couple l'estimateur par la différence à un échantillonnage avec probabilité proportionnelle à la taille, car elle est reconnue comme optimale quand le modèle de superpopulation est entièrement connu. Nous démontrons qu'y compris en cas de spécifications erronées modérées du modèle, cette stratégie n'est pas robuste et peut être surpassée par d'autres solutions.

Mots-clés : Plan de sondage; estimateur GREG; mesure du risque.

1. Introduction

Nous nous penchons sur le problème du choix d'une stratégie, en particulier d'un plan, pour l'estimation du total d'une variable étudiée dans une population finie quand un ensemble de J variables auxiliaires est disponible dans une base de sondage en liste. Nous nous intéressons principalement à l'estimation du total.

Le choix d'une stratégie d'échantillonnage comporte des paramètres inconnus à l'étape où la décision doit être prise. Après la collecte des données, on peut estimer les paramètres, mais parfois seulement selon certaines hypothèses. En pratique, à l'étape de la planification, nous utilisons souvent les données de vagues précédentes d'une enquête répétée, les variables de la base de sondage ou les données d'une enquête similaire. Il y a un risque que les données disponibles ne fournissent pas d'information fiable sur les paramètres pertinents. La méthode présentée ici comporte une mesure du risque, qui tient compte de la possibilité d'être induit en erreur par des croyances inexactes ou incorrectes concernant les valeurs des paramètres nécessaires. La mesure du risque est calculée pour les estimateurs par la différence et par la régression généralisée. Sinon, la mesure est générale. Cette mesure et la discussion sur son utilisation pratique forment le résultat principal du présent article.

L'un des objectifs de la sélection et de l'établissement de la stratégie d'échantillonnage est l'efficacité en matière de petite erreur quadratique moyenne. Toutefois, il n'y a pas une seule définition de « l'efficacité », car elle dépend de l'approche d'inférence. Dans l'approche fondée sur le plan de sondage, Godambe (1955), Lanke (1973) et Cassel, Särndal et Wretman (1977) montrent qu'il n'y a pas de meilleur estimateur linéaire uniformément, au sens qui serait le meilleur pour toutes les populations. Il n'y a pas non plus de meilleur plan. C'est pourquoi une démarche classique de définition de la stratégie a été de

1. Edgar Bueno, Département de statistique, Université de Stockholm, 106 91 Stockholm, Suède. Courriel : edgar.bueno@stat.su.se; Dan Hedlin, Département de statistique, Université de Stockholm, 106 91 Stockholm, Suède. Courriel : dan.hedlin@stat.su.se.

supposer que la population finie est la réalisation d'un certain modèle de superpopulation. La stratégie est ensuite définie de façon à minimiser la valeur espérée du modèle de l'erreur quadratique moyenne du plan, un paramètre qu'on appelle erreur quadratique moyenne anticipée. L'adjectif « anticipé » a été introduit d'abord par Isaki et Fuller (1982) pour souligner le fait qu'il s'agit d'une erreur quadratique moyenne conceptuelle, calculée avant l'échantillonnage, à partir de l'information disponible avant l'échantillonnage.

En supposant qu'un modèle de superpopulation se vérifie et que ses paramètres sont connus, plusieurs auteurs ont démontré que la stratégie optimale devrait utiliser un plan d'échantillonnage avec probabilité proportionnelle à la taille (par exemple Hájek, 1959; Cassel, Särndal et Wretman, 1976; Nedyalkova et Tillé, 2008). Toutefois, dans la pratique, il n'existe aucun consensus sur l'existence d'un modèle générateur, et encore moins sur le modèle sur lequel s'appuyer. Et même si modèle il y a, ses paramètres sont inconnus. Des données, plutôt empiriques, montrent que l'échantillonnage avec probabilité proportionnelle à la taille n'est pas robuste pour les spécifications erronées du modèle (par exemple Holmberg et Swensson, 2001). Le deuxième résultat de l'article est de fournir des preuves théoriques de ce fait.

De nombreux articles traitent de la robustesse dans le domaine de l'échantillonnage d'enquête. À titre d'exemple, Beaumont, Haziza et Ruiz-Gazen (2013) proposent un estimateur robuste qui réduit, diminue le poids des observations influentes, Royall et Herson (1973) envisagent la robustesse dans les modèles polynomiaux, Bramati (2012) et Zhai et Wiens (2015) proposent des méthodes de stratification robustes. Nous avançons des preuves théoriques de l'absence de robustesse de l'échantillonnage avec probabilité proportionnelle à la taille et proposons une méthode aidant à décider du plan de sondage.

L'article est organisé comme suit. La stratégie optimale selon le modèle de superpopulation est définie à la section 2, puis l'absence de robustesse de cette stratégie en cas de spécification erronée du modèle est étudiée à la section 3. La méthode utilisée pour faciliter le choix du plan de sondage est présentée à la section 4. À la section 5, la mesure du risque présentée à la section précédente est élargie pour pouvoir être utilisée avec l'estimateur GREG. Des illustrations numériques des résultats de l'article sont présentées à la section 6 : nous illustrons d'abord l'absence de robustesse de l'échantillonnage avec probabilité proportionnelle à la taille et la souplesse de l'estimateur GREG au moyen d'une petite étude par simulations, puis la mise en œuvre de la mesure du risque à l'aide de données d'enquête réelles. Enfin, les conclusions de l'étude sont présentées à la section 7.

2. Stratégie optimale selon le modèle de superpopulation

Soit U une population finie de taille N avec des éléments étiquetés $\{1, 2, \ldots, k, \ldots, N\}$. Soit $x_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})$ un vecteur connu des valeurs de J variables auxiliaires et y_k la valeur inconnue d'une variable étudiée associée à l'unité $k \in U$. Nous cherchons à estimer le total de y, $t_y = \sum_U y_k$.

Soit Ω l'ensemble de puissance de U. Un *échantillon* est un sous-ensemble $s \in \Omega$ et un *plan de sondage* est une loi de probabilité sur Ω, désignée par $P(S = s)$ ou simplement par $p(s)$. Soit $\pi_k = \sum_{s \ni k} p(s)$ la *probabilité d'inclusion* de k et $\pi_{kl} = \sum_{s \supset \{k,l\}} p(s)$ la *probabilité d'inclusion conjointe* de k et l. Un *plan de sondage probabiliste* est un plan de sondage tel que $\pi_k > 0$ pour tous les $k \in U$.

Un *estimateur* est une fonction réelle évaluée de l'échantillon, $\hat{t}_y = \hat{t}_y(S)$. Par *stratégie*, nous désignons le couple du plan de sondage et de l'estimateur, $\left(p(\cdot), \hat{t}_y\right)$.

Nous considérons seulement les plans de sondage probabiliste ayant une taille d'échantillon fixe. Comme point de départ pratique, nous commençons par examiner les estimateurs linéaires sans biais ayant la forme

$$\hat{t}_y = \left(\sum_U z_k - \sum_s \frac{z_k}{\pi_k}\right) + \sum_s \frac{y_k}{\pi_k} = \sum_U z_k + \sum_s \frac{e_k}{\pi_k} \tag{2.1}$$

avec z_k constantes connues arbitraires et $e_k = y_k - z_k$. Cet estimateur est appelé l'*estimateur par la différence*. L'estimateur ainsi défini est dit *calé* sur z car il satisfait $\hat{t}_z = \sum_U z_k$. À noter que si $z_k = 0$ pour tous les $k \in U$ l'estimateur se réduit à $\hat{t}_y = \sum_s y_k / \pi_k$, c'est-à-dire, à l'estimateur de Horvitz-Thompson (Horvitz et Thompson, 1952). Dans les sections suivantes, nous nous concentrons sur l'estimateur par la régression généralisé (GREG).

EQM « sous le » plan de l'estimateur par la différence est

$$\mathrm{EQM}_{\mathrm{p}}\left(\hat{t}_y\right) = \mathrm{EQM}_{\mathrm{p}}\left(\sum_s \frac{e_k}{\pi_k}\right) = \sum_U \sum_U (\pi_{kl} - \pi_k \pi_l) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}. \tag{2.2}$$

Comme il a été mentionné dans l'introduction, en raison de l'absence de stratégie optimale dans le cadre de l'approche fondée sur le plan de sondage, souvent un modèle de superpopulation, ξ_0, est proposé et nous cherchons une stratégie optimale par rapport *à l'erreur quadratique moyenne anticipée*,

$$\mathrm{EQM}_{\xi_0 \mathrm{p}}\left(\hat{t}_y\right) = \mathrm{E}_{\xi_0} \mathrm{EQM}_{\mathrm{p}}\left(\hat{t}_y\right) = \mathrm{E}_{\xi_0} \mathrm{E}_p\left(\left(\hat{t}_y - t_y\right)^2\right). \tag{2.3}$$

Nous pouvons supposer que les valeurs y sont des réalisations du modèle suivant, désigné par ξ_0,

$$Y_k = f(x_k \mid \delta_1) + \varepsilon_k$$

avec

$$\mathrm{E}_{\xi_0}(\varepsilon_k) = 0, \quad \mathrm{V}_{\xi_0}(\varepsilon_k) = \sigma_0^2 g(x_k \mid \delta_2)^2 \quad \text{et} \quad \mathrm{E}_{\xi_0}(\varepsilon_k \varepsilon_l) = 0 \ \forall\ k \neq l \tag{2.4}$$

où $\delta = (\delta_1, \delta_2)$ est un vecteur des paramètres, $f: \mathcal{R}^J \to \mathcal{R}$ et $g: \mathcal{R}^J \to \mathcal{R}^+$. L'échantillon aléatoire s et les erreurs ε_k sont supposés indépendants. À la suite de Rosén (2000), les termes $f(x_k \mid \delta_1)$ et $g(x_k \mid \delta_2) > 0$ seront appelés *tendance* et *dispersion*, respectivement. En général, le terme tendance ne doit pas être compris dans un sens temporel, mais plutôt comme le développement de valeurs y avec x.

Notons que sous ξ_0, e_k dans l'estimateur par la différence (2.1) est une variable aléatoire qui représente la distance entre la valeur de la variable étudiée et z_k, c'est-à-dire $e_k = f(x_k \mid \delta_1) + \varepsilon_k - z_k$. Par conséquent $\mathrm{E}_{\xi_0} e_k = f(x_k \mid \delta_1) - z_k$ et $\mathrm{E}_{\xi_0} e_k^2 = (f(x_k \mid \delta_1) - z_k)^2 + \sigma_0^2 g(x_k \mid \delta_2)^2$. Un peu d'algèbre permet de voir à partir de (2.2) et (2.3) que l'EQM anticipée de l'estimateur par la différence devient

$$\mathrm{EQM}_{\xi_0 \mathrm{p}}(\hat{t}_y) = \mathrm{EQM}_{\mathrm{p}}\left(\sum_s \frac{f(x_k \mid \delta_1) - z_k}{\pi_k}\right) + \sigma_0^2 \sum_U \left(\frac{1}{\pi_k} - 1\right) g(x_k \mid \delta_2)^2 \tag{2.5}$$

Nedyalkova et Tillé (2008) calculent l'EQM anticipée dans un cas plus général.

Tillé et Wilhelm (2017) donnent l'EQM anticipée de l'estimateur de Horvitz-Thompson. Le deuxième terme dans (2.5) est la borne inférieure de Godambe-Joshi (par exemple Särndal, Swensson et Wretman, 1992, page 453).

L'EQM anticipée dans (2.5) est la somme de deux termes positifs. On constate aisément que si

1. l'estimateur est calé sur $z_k = f(x_k \mid \delta_1)$ le premier terme disparaît et l'EQM anticipée est égale à la borne inférieure de Godambe-Joshi

$$\mathrm{EQM}_{\xi_0 \mathrm{p}}(\hat{t}_y) = \sigma_0^2 \sum_U \left(\frac{1}{\pi_k} - 1\right) g(x_k \mid \delta_2)^2. \tag{2.6}$$

De plus, après avoir imposé la restriction de taille d'échantillon fixe $\sum_U \pi_k = n$, si

2. le plan est tel que $\pi_k \propto g(x_k \mid \delta_2)$, noté $\pi\mathrm{ps}(\delta_2)$, le second terme est minimisé et nous obtenons

$$\mathrm{EQM}^{\mathrm{opt}}_{\xi_0 \mathrm{p}}(\hat{t}_y) = \sigma_0^2 \left(\frac{1}{n}\left(\sum_U g(x_k \mid \delta_2)\right)^2 - \sum_U g(x_k \mid \delta_2)^2\right).$$

Les conditions 1 et 2 suggèrent les rôles particuliers du plan de sondage et de l'estimateur dans la stratégie d'échantillonnage. L'estimateur devrait « expliquer » la tendance dans le sens du calage de la condition 1. Le plan devrait « expliquer » la dispersion. Une stratégie qui satisfait simultanément les conditions 1 et 2 sera dite « optimale ». Dans le même sens, tout estimateur et tout plan satisfaisant respectivement les conditions 1 et 2 seront appelés optimaux. Comme cette stratégie joue un rôle important dans notre article, nous la noterons par $\pi\mathrm{ps}(\delta_2) - \mathrm{diff}(\delta_1)$.

3. Robustesse en cas de spécification erronée du modèle

Si la population finie est une réalisation du modèle de superpopulation (2.4), et si f, g et δ sont connus, alors une stratégie optimale pourrait être définie. Dans cette section, nous étudions la robustesse de cette stratégie en cas de spécification erronée du modèle.

Nous commençons par définir comment le terme « spécification erronée » doit être compris dans l'article. Le *modèle de travail* ξ_0 reflète les croyances que le statisticien a au sujet de la relation entre les variables auxiliaires x et la variable étudiée y à l'étape du plan. Nous devons supposer qu'il existe un modèle inconnu vrai ξ. Tout écart de ξ_0 par rapport à ξ est une spécification erronée du modèle. Pour que l'analyse reste résoluble, nous nous limitons à la situation dans laquelle le modèle de travail est de la forme (2.4) et le modèle vrai, ξ, est

$$Y_k = f(x_k \mid \beta_1) + \varepsilon_k$$

avec

$$\mathrm{E}_\xi(\varepsilon_k) = 0, \quad \mathrm{V}_\xi(\varepsilon_k) = \sigma^2 g(x_k \mid \beta_2)^2 \quad \text{et} \quad \mathrm{E}_\xi(\varepsilon_k \varepsilon_l) = 0 \ \forall\ k \neq l \tag{3.1}$$

où $\beta = (\beta_1, \beta_2)$ est un vecteur des paramètres, f et g comme dans (2.4) et $\beta \neq \delta$. L'échantillon aléatoire s et les erreurs ε_k sont supposés indépendants.

Résultat 1. Si on suppose ξ_0 quand ξ est le vrai modèle de superpopulation, la valeur espérée du modèle de EQM « sous le » plan dans (2.2), avec l'estimateur par la différence satisfaisant à la condition 1 ci-dessus, devient

$$\mathrm{EQM}_{\xi\mathrm{p}}(\hat{t}_y) = \mathrm{EQM}_{\mathrm{p}}\left(\sum_s \frac{f(x_k \mid \beta_1) - f(x_k \mid \delta_1)}{\pi_k}\right) + \sigma^2 \sum_U \left(\frac{1}{\pi_k} - 1\right) g(x_k \mid \beta_2)^2 . \tag{3.2}$$

On prouve le résultat en notant que $f(x_k \mid \delta_1)$ prend le rôle de z_k dans (2.5) et en tenant compte du fait que $e_k = f(x_k \mid \beta_1) - f(x_k \mid \delta_1) + \varepsilon_k$, par conséquent, $\mathrm{E}_\xi e_k = f(x_k \mid \beta_1) - f(x_k \mid \delta_1)$ et $\mathrm{E}_\xi e_k^2 = (f(x_k \mid \beta_1) - f(x_k \mid \delta_1))^2 + \sigma^2 g(x_k \mid \beta_2)^2$. En raison de la spécification erronée du modèle, nous avons délibérément évité d'utiliser l'adjectif « anticipé » dans le résultat 1.

Au moyen du résultat 1, on voit que pour un plan satisfaisant la condition 2, nous obtenons

$$\begin{aligned}\mathrm{EQM}_{\xi,\pi\mathrm{ps}}(\hat{t}_y) = {} & \left(\frac{\sum_U g(x_k \mid \delta_2)}{n}\right)^2 \mathrm{EQM}_{\pi\mathrm{ps}}\left(\sum_s \frac{f(x_k \mid \beta_1) - f(x_k \mid \delta_1)}{g(x_k \mid \delta_2)}\right) \\ & + \sigma^2 \sum_U \left(\frac{n \sum_U g(x_k \mid \delta_2)}{g(x_k \mid \delta_2)} - 1\right) g(x_k \mid \beta_2)^2 .\end{aligned} \tag{3.3}$$

On peut maintenant voir que, même en cas de spécification erronée légère, comme celle dont il est question ici, la stratégie $\pi\mathrm{ps}(\delta_2) - \mathrm{diff}(\delta_1)$ n'est plus optimale, car son EQM (3.3) peut être supérieure à l'EQM obtenue dans d'autres plans (3.2). En particulier, la stratégie utilisant le bon modèle, c'est-à-dire $z_k = f(x_k \mid \beta_1)$ dans l'estimateur et un plan tel que $\pi_k \propto g(x_k \mid \beta_2)$, serait plus efficace que $\pi\mathrm{ps}(\delta_2) - \mathrm{diff}(\delta_1)$.

4. Orienter le choix du plan de sondage à l'aide d'une mesure du risque

Nous avons constaté à la section 3 que même une spécification erronée simple du modèle de travail peut faire que la stratégie $\pi\mathrm{ps}(\delta_2) - \mathrm{diff}(\delta_1)$ ne soit pas optimale. Par conséquent, il est risqué d'accepter un modèle donné comme étant correct sans aucun type d'évaluation. Bien que la plus grande partie de l'information nécessaire à une évaluation « objective » du modèle ne soit pas disponible à l'étape du plan, on peut atteindre un certain degré de confiance à propos des paramètres du modèle de travail, qui permet de comparer un ensemble de plans et de décider du plan à mettre en œuvre. Dans la présente section, nous proposons une méthode qui facilite le choix du plan de sondage.

L'EQM espérée du modèle (3.2) dans le résultat 1 peut être considérée comme une fonction de β et σ^2, puisque tout le reste est disponible à l'étape du plan. Pour commencer, supposons que σ^2 soit aussi connu. Nous pouvons alors écrire

$$L_{\mathrm{p}}(\beta) = \mathrm{EQM}_{\xi\mathrm{p}}(\beta \mid x, \delta, \sigma) = \mathrm{EQM}_{\mathrm{p}}\left(\sum_s \frac{f(x_k \mid \beta_1) - f(x_k \mid \delta_1)}{\pi_k}\right) + \sigma^2 \sum_U \left(\frac{1}{\pi_k} - 1\right) g(x_k \mid \beta_2)^2 .$$

Pour tout plan, $p(\cdot)$, cette fonction peut être évaluée à tout β et indique la perte subie en supposant que δ est le vrai paramètre alors qu'il s'agit en fait de β. Nous pouvons supposer une distribution a priori sur β, $h(\beta)$, et calculer le risque sous h,

$$R(\mathrm{p}) = \mathrm{E}_h\left(\mathrm{EQM}_{\xi\mathrm{p}}(\beta \mid x, \delta, \sigma)\right) = \int_\Theta h(\beta) \cdot \mathrm{EQM}_{\xi\mathrm{p}}(\beta \mid x, \delta, \sigma)\, d\beta, \tag{4.1}$$

où Θ est l'espace échantillon de β. Il faut choisir le plan présentant le plus petit risque. Notons que des méthodes d'intégration numérique (par exemple méthodes de simulation Monte Carlo) peuvent être nécessaires aux fins d'évaluation du risque (4.1).

En pratique, σ^2 est inconnu. Nous proposons deux solutions pour le traiter. La première consiste à voir maintenant la perte comme une fonction de β et σ et à calculer le risque comme ci-dessus, en supposant une distribution a priori sur la paire β et σ. La deuxième consiste à donner « une idée » de sa valeur. Cette méthode peut se servir du fait que (Démonstration en annexe)

$$\sigma^2 \approx \frac{S_{f,f}}{\bar{g}^2}\left(\frac{1}{R_{f,y}^2} - 1\right) \tag{4.2}$$

où $S_{f,f} = \sum_U \left(f(x_k \mid \beta_1) - \bar{f}\right)^2 / N$, $\bar{f} = \sum_U f(x_k \mid \beta_1)/N$, $\bar{g}^2 = \sum_U g(x_k \mid \beta_2)^2 / N$ et $R_{f,y}$ est la corrélation entre $f(x \mid \beta_1)$ et y. (Dans l'exemple 3 ci-dessous, nous donnons une expression plus commode dans un cas particulier.) Bien que $R_{f,y}$ soit inconnu, dans les enquêtes répétées, nous en avons une certaine connaissance au préalable. Dans d'autres cas, il est souvent possible d'avoir une « idée » raisonnable à ce sujet.

Il reste à commenter le choix de la distribution a priori $h(\beta)$. Le choix de la distribution et de ses paramètres est subjectif; il est défini par le statisticien. Néanmoins, il doit s'expliquer par les

connaissances disponibles sur le paramètre du modèle β. En particulier, $h(\beta)$ doit être centré autour de $\beta = \delta$. Sa variance doit montrer le degré de confiance que nous avons à l'égard du modèle de travail. Notons qu'une confiance totale sur le modèle de travail serait une densité avec toute sa masse à $\beta = \delta$, auquel cas le risque (4.1) serait minimisé par le plan πps donné par la condition 2 de la section 2.

On pourrait faire valoir que l'ajout de $h(\beta)$ est une source supplémentaire de subjectivité dans le choix du plan de sondage. La distribution a priori peut ajouter une saveur bayésienne au processus, mais on remarque que $h(\beta)$ est nécessaire seulement pour le choix du plan. Par conséquent, l'inférence est encore fondée sur le plan. De plus, s'appuyer sur un modèle supposé est également subjectif dans un choix et comporte un risque. La mesure du risque de (4.1) permet de quantifier ce risque.

5. La mesure du risque selon l'estimateur par la régression généralisée

L'estimateur par la différence (2.1) exige que δ_1 soit entièrement spécifié afin de calculer $f(x_k \mid \delta_1)$, ce qui n'est pas souhaitable d'un point de vue pratique. L'estimateur par la régression généralisée (GREG) est une solution de rechange qui permet d'estimer la totalité ou une partie des composantes de δ_1 contre le coût de l'introduction d'un petit biais. Dans la présente section, nous adaptons le matériau des sections 2 à 4 aux stratégies utilisant l'estimateur GREG.

Nous définissons l'estimateur par la régression généralisée de façon plus générale que dans Särndal et coll. (1992) comme suit. Soit a_k $(k = 1, \ldots, N)$ un poids défini par le statisticien et $\delta_1 = (\delta_1^*, \delta_1^{**})$ où δ_1^* est fixé et δ_1^{**} doit être estimé. Soit aussi

$$\hat{\delta}_{1s}^{**} = \operatorname{argmin}_{\delta_1^{**}} \sum_s \frac{(y_k - f(x_k \mid \delta_1))^2}{a_k \pi_k}$$

et $\hat{\delta}_{1s} = (\delta_1^*, \hat{\delta}_{1s}^{**})$. L'estimateur GREG est

$$\hat{t}_{\text{greg}} = \left(\sum_U f(x_k \mid \hat{\delta}_{1s}) - \sum_s \frac{f(x_k \mid \hat{\delta}_{1s})}{\pi_k} \right) + \sum_s \frac{y_k}{\pi_k}. \tag{5.1}$$

Une approximation de EQM « sous le » plan de l'estimateur GREG prend la forme (2.2) avec $e_k = y_k - f(x_k \mid \hat{\delta}_{1U})$ où $\hat{\delta}_{1U} = (\delta_1^*, \hat{\delta}_{1U}^{**})$ et

$$\hat{\delta}_{1U}^{**} = \operatorname{argmin}_{\delta_1^{**}} \sum_U \frac{(y_k - f(x_k \mid \delta_1))^2}{a_k}.$$

Exemple 1. Examinons le cas où $f(x_k \mid \delta_1) = \delta_{1,1} x_{1k}^{\delta_{1,J+1}} + \delta_{1,2} x_{2k}^{\delta_{1,J+2}} + \ldots + \delta_{1,J} x_{Jk}^{\delta_{1,2J}}$. Soit $\delta_1^* = (\delta_{1,J+1}, \ldots, \delta_{1,2J})$, $\delta_1^{**} = (\delta_{1,1}, \ldots, \delta_{1,J})'$ et $x_k^\delta = (x_{1k}^{\delta_{1,J+1}}, \ldots, x_{Jk}^{\delta_{1,2J}})$. Dans ce cas, nous obtenons

$$\hat{\delta}_{1s}^{**} = \left(\sum_s \frac{x_k^{\delta'} x_k^\delta}{a_k \pi_k} \right)^{-1} \sum_s \frac{x_k^{\delta'} y_k}{a_k \pi_k} \quad \text{et} \quad \hat{\delta}_{1U}^{**} = \left(\sum_U \frac{x_k^{\delta'} x_k^\delta}{a_k} \right)^{-1} \sum_U \frac{x_k^{\delta'} y_k}{a_k}.$$

En supposant les exposants $\delta_1^* = (\delta_{1,J+1}, \ldots, \delta_{1,2J}) = (1, \ldots, 1)$, nous obtenons l'expression classique de l'estimateur GREG que l'on trouve dans Särndal et coll. (1992).

Exemple 2. Le cas avec une seule variable auxiliaire, c'est-à-dire $f(x_k \mid \delta_1) = \delta_{10} + \delta_{11} x_k^{\delta_{12}}$ avec $a_k = 1$, $\delta_1^* = \delta_{12}$ et $\delta_1^{**} = (\delta_{10}, \delta_{11})'$ est connu sous le nom d'estimateur par la régression. Dans ce cas, nous obtenons le résultat bien connu selon lequel EQM « sous le » plan peut être estimée par l'expression (2.2) avec $e_k = y_k - f(x_k \mid \hat{\delta}_{1U})$ où $f(x_k \mid \hat{\delta}_{1U}) = \hat{\delta}_{10} + \hat{\delta}_{11} x_k^{\delta_{12}}$ et

$$\hat{\delta}_{11} = \frac{N \sum_U x_k^{\delta_{12}} y_k - \sum_U x_k^{\delta_{12}} \sum_U y_k}{N \sum_U x_k^{2\delta_{12}} - \left(\sum_U x_k^{\delta_{12}}\right)^2} \quad \text{et} \quad \hat{\delta}_{10} = \frac{1}{N} \sum_U y_k - \hat{\delta}_{11} \frac{1}{N} \sum_U x_k^{\delta_{12}}.$$

Modèle à la spécification erronée

Examinons de nouveau le scénario dans lequel le statisticien utilise le modèle de travail (2.4), mais où le vrai modèle est celui de la forme (3.1) avec $\beta_1 = (\beta_1^*, \beta_1^{**})$, où β_1^* est la contrepartie de la composante fixe δ_1^*. Le résultat suivant énonce une condition selon laquelle le résultat 1 est valide pour l'estimateur GREG.

Résultat 2. Si on suppose ξ_0 quand ξ est le modèle de superpopulation vrai, $\hat{\delta}_{1s}^{**} \to \hat{\delta}_{1U}^{**}$ à mesure que $n \to \infty$ et $\hat{\delta}_{1U}^{**}$ convergent vers une valeur δ_1^{**} à mesure que $N \to \infty$, alors

$$\mathrm{EQM}_{\xi \mathrm{p}}(\hat{t}_{\mathrm{greg}}) \to \mathrm{EQM}_{\mathrm{p}}\left(\sum_s \frac{f(x_k \mid \beta_1) - f(x_k \mid \delta_1)}{\pi_k}\right) + \sigma^2 \sum_U \left(\frac{1}{\pi_k} - 1\right) g(x_k \mid \beta_2)^2 \qquad (5.2)$$

où $\delta_1 = (\delta_1^*, \delta_1^{**})$.

Démonstration. Notons que si $\hat{\delta}_{1s}^{**} \to \delta_{1U}^{**}$, alors $\hat{\delta}_{1s} = (\delta_1^*, \hat{\delta}_{1s}^{**}) \to (\delta_1^*, \hat{\delta}_{1U}^{**}) = \hat{\delta}_{1U}$. Alors, $f(x_k \mid \hat{\delta}_{1s}) \to f(x_k \mid \hat{\delta}_{1U})$. Ensuite, si $\hat{\delta}_{1U}^{**} \to \delta_1^{**}$, alors $\hat{\delta}_{1U} = (\delta_1^*, \hat{\delta}_{1U}^{**}) \to (\delta_1^*, \delta_1^{**}) = \delta_1$. Ainsi $f(x_k \mid \hat{\delta}_{1U}) \to f(x_k \mid \delta_1)$. Par conséquent,

$$\begin{aligned}\mathrm{EQM}_{\xi \mathrm{p}}(\hat{t}_{\mathrm{greg}}) &= \mathrm{EQM}_{\xi \mathrm{p}}\left(\left(\sum_U f(x_k \mid \hat{\delta}_{1s}) - \sum_s \frac{f(x_k \mid \hat{\delta}_{1s})}{\pi_k}\right) + \sum_s \frac{y_k}{\pi_k}\right) \\ &\to \mathrm{EQM}_{\xi \mathrm{p}}\left(\left(\sum_U f(x_k \mid \delta_1) - \sum_s \frac{f(x_k \mid \delta_1)}{\pi_k}\right) + \sum_s \frac{y_k}{\pi_k}\right),\end{aligned}$$

qui, selon le résultat 1, est (5.2).

Exemple 3 (suite de l'exemple 1). Supposons que le modèle de travail soit comme dans l'exemple 1 et que le modèle vrai soit $f(x_k \mid \beta_1) = \beta_{1,1} x_{1k}^{\beta_{1,J+1}} + \beta_{1,2} x_{2k}^{\beta_{1,J+2}} + \ldots + \beta_{1,J} x_{Jk}^{\beta_{1,2J}}$. Supposons aussi $\beta_1^* = (\beta_{1,J+1}, \ldots, \beta_{1,2J})$, $\beta_1^{**} = (\beta_{1,1}, \ldots, \beta_{1,J})'$ et $x_k^{\beta} = (x_{1k}^{\beta_{1,J+1}}, \ldots, x_{Jk}^{\beta_{1,2J}})$. Dans ce cas, $\hat{\delta}_{1U}^{**} \to A\beta_1^{**}$, où

$$A = \left(\sum_U \frac{x_k^{\delta'} x_k^{\delta}}{a_k} \right)^{-1} \sum_U \frac{x_k^{\delta'} x_k^{\beta}}{a_k},$$

et (5.2) devient

$$\mathrm{EQM}_{\xi\mathrm{p}}\left(\hat{t}_{\mathrm{greg}}\right) \rightarrow \mathrm{EQM}_{\mathrm{p}}\left(\sum_s \frac{\left(x_k^{\beta} - x_k^{\delta} A\right) \beta_1^{**}}{\pi_k} \right) + \sigma^2 \left(\sum_U \frac{g\left(x_k \mid \beta_2\right)^2}{\pi_k} - \sum_U g\left(x_k \mid \beta_2\right)^2 \right). \quad (5.3)$$

Exemple 4 (suite de l'exemple 2). Supposons que le modèle de travail soit comme dans l'exemple 2 et que le modèle vrai soit $f\left(x_k \mid \beta_1\right) = \beta_{10} + \beta_{11} x_k^{\beta_{12}}$ avec $\beta_1^* = \beta_{12}$ et $\beta_1^{**} = \left(\beta_{10}, \beta_{11}\right)'$. Il peut être démontré que (5.2) devient

$$\mathrm{EQM}_{\xi\mathrm{p}}\left(\hat{t}_{\mathrm{greg}}\right) \rightarrow \beta_{11}^2 \mathrm{EQM}_{\mathrm{p}}\left(\sum_s \frac{v_k}{\pi_k} \right) + \sigma^2 \left(\sum_U \frac{g\left(x_k \mid \beta_2\right)^2}{\pi_k} - \sum_U g\left(x_k \mid \beta_2\right)^2 \right) \quad (5.4)$$

avec

$$v_k = \left(x_k^{\beta_{12}} - \bar{x}^{\beta_{12}}\right) - \left(x_k^{\delta_{12}} - \bar{x}^{\delta_{12}}\right) \frac{S_{\beta,\delta}}{S_{\delta,\delta}}, \quad (5.5)$$

et

$$\bar{x}^{\beta_{12}} = \frac{1}{N} \sum_U x_k^{\beta_{12}} \qquad S_{\beta,\delta} = \frac{1}{N-1} \sum_U \left(x_k^{\beta_{12}} - \bar{x}^{\beta_{12}}\right)\left(x_k^{\delta_{12}} - \bar{x}^{\delta_{12}}\right)$$

$$\bar{x}^{\delta_{12}} = \frac{1}{N} \sum_U x_k^{\delta_{12}} \qquad S_{\delta,\delta} = \frac{1}{N-1} \sum_U \left(x_k^{\delta_{12}} - \bar{x}^{\delta_{12}}\right)^2.$$

Notons que (5.4) ne dépend pas de β_{10}.

Pour le cas particulier développé dans les exemples 2 et 4, où $f\left(x_k \mid \beta\right) = \beta_{10} + \beta_{11} x_k^{\beta_{12}}$ et $f\left(x_k \mid \delta\right) = \delta_{10} + \delta_{11} x_k^{\delta_{12}}$, une autre approximation de σ^2 est (Démonstration en annexe)

$$\sigma^2 \approx \beta_{11}^2 F_0 \quad \text{avec} \quad F_0 = \frac{1}{\bar{x}^{2\beta_2}} \frac{S_{1,\beta}^2}{S_{1,1}} \left(\frac{1}{R_{x,y}^2} - \frac{1}{R_{1,\beta}^2} \right) \quad (5.6)$$

où

$$\bar{x}^{2\beta_2} = \frac{1}{N} \sum_U x_k^{2\beta_2} \qquad S_{1,\beta} = \frac{1}{N} \sum_U \left(x_k - \bar{x}\right)\left(x_k^{\beta_{12}} - \bar{x}^{\beta_{12}}\right) \qquad S_{1,1} = \frac{1}{N} \sum_U \left(x_k - \bar{x}\right)^2$$

avec $\left|R_{x,y}\right| \leq \left|R_{1,\beta}\right|$ et $R_{1,\beta}$ et $R_{x,y}$ qui sont, respectivement, les coefficients de corrélation entre x et $x^{\beta_{12}}$ et entre x et y. Ce dernier est inconnu, mais il est souvent possible de le deviner convenablement.

L'approximation de σ^2 dans (5.6) est plus commode que celle de (4.2), car nous savons maintenant que (5.4) est obtenu approximativement par

$$\mathrm{EQM}_{\xi\mathrm{p}}\left(\hat{t}_{\mathrm{greg}}\right) \approx \beta_{11}^2\left[\mathrm{EQM}_{\mathrm{p}}\left(\sum_s \frac{v_k}{\pi_k}\right) + F_0\left(\sum_U \frac{g\left(x_k \mid \beta\right)^2}{\pi_k} - \sum_U g\left(x_k \mid \beta\right)^2\right)\right] \tag{5.7}$$

avec v_k donné par (5.5). Cette expression ne dépend ni de l'ordonnée à l'origine β_{01} ni du paramètre σ, et la pente β_{11} devient une constante de proportionnalité ignorable.

Mesure du risque

Comme à la section 4, l'EQM espérée du modèle asymptotique de l'estimateur GREG donnée par le résultat 2 peut être considérée comme la perte subie quand on suppose que δ est le paramètre vrai quand il s'agit en fait de β. En supposant une distribution a priori sur β, on peut calculer le risque (4.1).

6. Exemples numériques

Dans les sections 2 et 3, nous avons établi que la stratégie $\pi\mathrm{ps}\left(\delta_2\right) - \mathrm{diff}\left(\delta_1\right)$ est optimale selon un modèle de superpopulation, mais qu'elle n'est pas robuste pour les spécifications erronées du modèle. À la sous-section 6.1, nous présentons une petite étude par simulations Monte Carlo qu'on a réalisée pour illustrer ces résultats en comparant la stratégie optimale et trois solutions de rechange.

Dans les sections 4 et 5, nous avons introduit une mesure qui permet de quantifier le risque lié à la mise en œuvre d'un plan de sondage, ce qui permet d'orienter le choix du plan. À la sous-section 6.2, nous illustrons l'utilisation de la mesure du risque au moyen de données d'enquête réelles.

6.1 Étude par simulations en cas de spécification erronée du modèle

Nous comparons l'efficacité et la robustesse de quatre stratégies au moyen d'une étude par simulations. Les stratégies qui doivent être comparées sont $\pi\mathrm{ps}$ avec l'estimateur par la différence (qui est optimal quand le modèle est correct), $\pi\mathrm{ps}$ avec l'estimateur GREG (plan de sondage optimal), un échantillonnage aléatoire simple stratifié (EASS) avec l'estimateur par la différence (estimateur optimal) et un EASS avec l'estimateur GREG.

Notre mise en œuvre de $\pi\mathrm{ps}$ fait appel au $\pi\mathrm{ps}$ de Pareto (Rosén, 1997). Il existe une multitude d'autres méthodes pour tirer des échantillons $\pi\mathrm{ps}$. Néanmoins, $\pi\mathrm{ps}$ de Pareto est une méthode pratique ayant de bonnes propriétés, comme le montre, par exemple, Rosén (2000).

Notre mise en œuvre d'un EASS repose sur une stratification fondée sur un modèle (Wright, 1983). Nous considérons $H = 5$ strates dont les bornes sont définies à l'aide de la règle $\mathrm{cum}\sqrt{f}$ de Dalenius et Hodges (1959) sur $g\left(x_k \mid \delta_2\right)$ (bien décrite dans Särndal et coll. 1992, page 463) et l'échantillon est réparti à l'aide de la répartition de Neyman, $n_h \propto N_h S_{gh}$. L'utilisation de la règle $\mathrm{cum}\sqrt{f}$ peut être

sous-optimale (voir Särndal et coll., 1992, page 464), mais l'efficacité de la stratification par une variable de taille continue est assez insensible au choix exact des bornes.

Nous examinons seulement la spécification erronée de la dispersion. Le terme de tendance se présente sous la forme $f(x_k \mid \beta_1) = \beta_{10} + \beta_{11} x_k^{\beta_{12}}$ avec $\beta_{10} = 1\,000$, $\beta_{11} = 1$ et $\beta_{12} = 0{,}75$, 1 et 1,25. La dispersion vraie est $g(x_k \mid \beta_2) = x_k^{\beta_2}$ avec $\beta_2 = 0{,}5$, 0,75 et 1. La dispersion de travail est $g(x_k \mid \delta_2) = x_k^{\delta_2}$ avec $\delta_2 = 0{,}5$, 0,75 et 1.

Nous utiliserons l'estimateur par la différence (2.1) calé sur $f(x_k \mid \beta_1)$. En ce qui concerne l'estimateur GREG, nous fixerons β_{12}, tandis que les coefficients β_{10} et β_{11} seront estimés.

La simulation se présente comme suit. La taille de l'échantillon est $N = 5\,000$. Les valeurs x sont des réalisations indépendantes d'une loi Gamma de forme $\alpha = 4/100$ et d'échelle $\lambda = 1\,200$ plus une unité, tandis que y_k est une réalisation d'une loi Gamma de forme et d'échelle

$$\alpha_k = \frac{\left(\beta_{10} + \beta_{11} x_k^{\beta_{12}}\right)^2}{\sigma_0^2 x_k^{2\beta_2}} \quad \text{et} \quad \lambda_k = \frac{\sigma_0^2 x_k^{2\beta_2}}{\beta_{10} + \beta_{11} x_k^{\beta_{12}}},$$

où σ^2 a été établi de telle façon que la corrélation entre x et y soit de $\rho = 0{,}95$. EQM « sous le » plan d'un échantillon de taille $n = 500$ est ensuite calculée pour chaque stratégie. En conservant les valeurs x fixes, le processus est itéré $B = 5\,000$ fois.

Le tableau 6.1 présente les résultats de l'étude par simulations. Les trois premières colonnes indiquent les paramètres du modèle. La quatrième colonne montre l'EQM espérée du modèle (simulée) de la stratégie $\pi\text{ps} - \text{dif}$, tandis que les trois dernières colonnes montrent l'efficacité (simulée) des stratégies $\pi\text{ps} - \text{GREG}$, EAAS – dif et EAAS – GREG comparativement à $\pi\text{ps} - \text{dif}$ (en pourcentage), l'efficacité étant définie comme $\text{eff} = \text{EQM}_{\xi,\pi\text{ps}}(\hat{t}_y) / \text{EQM}_{\xi,p}(\hat{t}_y)$ où les erreurs quadratiques moyennes espérées du modèle sont obtenues approximativement par leurs homologues simulés,

$$\text{EQM}_{\xi,\text{p}}(\hat{t}_y) = \text{E}_\xi \text{EQM}_\text{p}(\hat{t}_y) \approx \frac{1}{B}\sum_{r=1}^{B} \text{EQM}_\text{p}^{(r)}(\hat{t}_y),$$

de sorte qu'une valeur de 100 indique que la stratégie est aussi efficace que $\pi\text{ps} - \text{dif}$ et que des valeurs plus petites (plus grandes) que 100 indiquent que la stratégie est moins (plus) efficace que $\pi\text{ps} - \text{dif}$.

La partie supérieure du tableau 6.1 montre un cas où le modèle de travail coïncide avec le modèle vrai. Comme prévu, la stratégie qui couple πps avec l'estimateur par la différence ($\pi\text{ps} - \text{dif}$) est toujours plus efficace que les autres stratégies. Il reste que la perte d'efficacité due à l'estimation de certains paramètres par l'estimateur GREG est négligeable. Par ailleurs, on constate une perte d'efficacité remarquable attribuable à l'utilisation de l'EASS au lieu de πps. Enfin, on voit dans (2.6) que comme l'EQM espérée pour toutes les stratégies ne dépend pas de la tendance f mais seulement de la dispersion g, l'efficacité reste constante en cas de valeur identique de δ_2, indépendamment de la valeur de β_{12}.

Tableau 6.1
Efficacité de trois stratégies en pourcentage de l'EQM espérée du modèle de πps – dif

Modèle correct						
β_{12}	β_2	δ_2	πps – dif	πps – GREG	EASS – dif	EASS – GREG
0,75	0,50	0,50	$2{,}78 \,.\, 10^5$	99,9	57,3	57,3
0,75	0,75	0,75	$4{,}82 \,.\, 10^4$	99,6	77,9	77,9
0,75	1,00	1,00	$1{,}90 \,.\, 10^4$	99,0	83,2	83,2
1,00	0,50	0,50	$7{,}64 \,.\, 10^6$	99,9	57,3	57,3
1,00	0,75	0,75	$7{,}20 \,.\, 10^5$	99,7	77,9	77,9
1,00	1,00	1,00	$2{,}14 \,.\, 10^5$	99,1	83,1	83,1
1,25	0,50	0,50	$1{,}46 \,.\, 10^8$	99,9	57,3	57,3
1,25	0,75	0,75	$7{,}85 \,.\, 10^6$	99,7	77,9	78,0
1,25	1,00	1,00	$1{,}81 \,.\, 10^6$	99,2	83,1	83,1
Spécification erronée du modèle						
β_{12}	β_2	δ_2	πps – dif	πps – GREG	EASS – dif	EASS – GREG
0,75	0,50	0,75	$3{,}98 \,.\, 10^5$	99,9	98,9	98,9
0,75	0,75	1,00	$6{,}45 \,.\, 10^4$	99,5	114,5	114,4
0,75	1,00	0,50	$4{,}73 \,.\, 10^4$	100,1	133,9	134,0
1,00	0,50	1,00	$2{,}14 \,.\, 10^7$	99,9	185,6	185,6
1,00	0,75	0,50	$1{,}03 \,.\, 10^6$	100,1	93,1	93,2
1,00	1,00	0,75	$2{,}77 \,.\, 10^5$	99,8	88,9	89,0
1,25	0,50	0,75	$2{,}09 \,.\, 10^8$	99,9	98,9	98,9
1,25	0,75	1,00	$1{,}05 \,.\, 10^7$	99,6	114,5	114,5
1,25	1,00	0,50	$4{,}50 \,.\, 10^6$	100,3	134,0	134,2

La partie inférieure du tableau 6.1 présente des comparaisons selon un modèle à la spécification erronée, en particulier une spécification erronée de la dispersion. Notons que malgré cette légère spécification erronée du modèle, πps – dif n'est plus nécessairement la meilleure stratégie, car les stratégies utilisant un EASS se sont révélées plus efficaces dans plusieurs cas. Il n'est toutefois pas évident de savoir quand un EASS est plus efficace que πps ou vice versa. La mesure du risque présentée à la section 4 peut servir à orienter le choix entre plusieurs plans. Les résultats présentés dans la section concordent avec ceux obtenus, par exemple, par Holmberg et Swensson (2001).

6.2 Choisir un plan dans une véritable enquête au moyen de la mesure du risque

Dans la présente sous-section, nous illustrons la mise en œuvre de la mesure du risque à l'aide de données provenant d'une véritable enquête. Nous voulons estimer $t_y = \sum_U y_k$ où U est l'ensemble des propriétés résidentielles de Bogotá, en Colombie, (de taille $N = 681\,276$) et y_k est la valeur de la propriété k^{e} en 2017 en pesos colombiens (COP). x_k, la zone bâtie de la propriété k^{e} en mètres carrés, est connue pour chaque $k \in U$. La variable auxiliaire x a une moyenne de 184, un écart-type de 110 et une asymétrie de 2,57. La taille souhaitée de l'échantillon est de $n = 1\,000$.

Nous supposons qu'un modèle de type ξ_0 avec $f(x_k \mid \delta_1) = \delta_{10} + \delta_{11} x_k^{\delta_{12}}$ et $g(x_k \mid \delta_2) = x_k^{\delta_2}$ décrit adéquatement l'association entre x et y. Nous prévoyons utiliser l'estimateur GREG pour estimer δ_{10} et δ_{11}, c'est-à-dire $\delta_1^{**} = (\delta_{10}, \delta_{11})$. Comme ce modèle a la forme illustrée dans l'exemple 4, l'EQM espérée du modèle peut être obtenue approximativement par l'expression (5.7).

Nous utiliserons le risque (4.1) pour faciliter le choix entre πps ou un EASS utilisant $H = 6$ strates. Nous considérons $h(\beta_{12}, \beta_2)$ comme une distribution normale bivariée sans corrélation entre β_{12} et β_2. On obtient par approximation l'intégrale au moyen du module `cubature` (Narasimhan, Johnson, Hahn, Bouvier et Kiêu, 2019) élaboré pour l'environnement logiciel statistique `R` (R Core Team, 2020).

Nous examinons deux cas ayant différents degrés de confiance à l'égard du modèle de travail.

Cas 1. Dans ce cas, on ne dispose d'aucune information sur δ_{12}, ni δ_2 ni $R_{x,y}$. Les valeurs naïves de $\delta_{12} = 1$, $\delta_2 = 1$ et $R_{x,y} = 0{,}75$ sont examinées. Afin de refléter l'incertitude, $h(\beta)$ doit avoir une grande variance. Par conséquent, nous établissons

$$\begin{bmatrix} \beta_{12} \\ \beta_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 1{,}0 \\ 1{,}0 \end{bmatrix}, \begin{bmatrix} 0{,}3295^2 & 0 \\ 0 & 0{,}3295^2 \end{bmatrix} \right).$$

La variance a été choisie de sorte que 99 % de la masse se trouve dans le cercle de rayon 1. L'évaluation de (4.1) donne $R(\pi\mathrm{ps}) = 6{,}89 \cdot 10^{15} \beta_{11}^2$ et $R(\mathrm{EASS}) = 1{,}59 \cdot 10^{15} \beta_{11}^2$, ce qui laisse penser qu'un plan stratifié doit être utilisé.

EQM « sous le » plan des deux stratégies est calculée et nous obtenons $\mathrm{EQM}_{\pi\mathrm{ps}}(\hat{t}_{\mathrm{greg}}) = 2{,}29 \cdot 10^{25}$ et l'$\mathrm{EQM}_{\mathrm{EASS}}(\hat{t}_{\mathrm{greg}}) = 1{,}36 \cdot 10^{25}$. La stratégie suggérée par (4.1) est effectivement le meilleur choix.

Cas 2. À l'aide d'un échantillon de 2010, on propose des valeurs a priori de $\delta_{12} = 1{,}9$, $\delta_2 = 2$ et $R_{x,y} = 0{,}7$. Comme l'incertitude est plus petite ici que celle du cas 1, nous établissons une variance plus petite,

$$\begin{bmatrix} \beta_{12} \\ \beta_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 1{,}9 \\ 2{,}0 \end{bmatrix}, \begin{bmatrix} 0{,}2471^2 & 0 \\ 0 & 0{,}2471^2 \end{bmatrix} \right).$$

La variance a été choisie de sorte que 99 % de la masse se trouve dans le cercle de rayon 0,75. L'évaluation de (4.1) donne $R(\pi\mathrm{ps}) = 7{,}08 \cdot 10^{22} \beta_{11}^2$ et $R(\mathrm{EASS}) = 4{,}06 \cdot 10^{18} \beta_{11}^2$, ce qui laisse penser qu'un plan stratifié doit être utilisé.

EQM « sous le » plan des deux stratégies est calculée et nous obtenons $\mathrm{EQM}_{\pi\mathrm{ps}}(\hat{t}_{\mathrm{greg}}) = 1{,}85 \cdot 10^{28}$ et $\mathrm{EQM}_{\mathrm{EASS}}(\hat{t}_{\mathrm{greg}}) = 1{,}91 \cdot 10^{25}$. Notons que l'utilisation de (4.1) nous a empêchés d'utiliser πps, dont l'EQM est presque mille fois plus grande que l'EQM en cas d'échantillonnage stratifié!

7. Conclusions

La stratégie couplant πps avec l'estimateur par la différence est optimale quand les paramètres du modèle de superpopulation sont connus. Étant donné que ces hypothèses sont rarement satisfaites, il a été démontré à la section 3 et illustré à la sous-section 6.1 que cette optimalité ne se vérifie plus même en cas de petites spécifications erronées du modèle.

À la section 4, nous proposons une méthode de choix de plan de sondage, qui est étendue pour être employée avec l'estimateur GREG dans la section 5. La méthode permet de prendre en compte l'incertitude des paramètres du modèle en introduisant une distribution a priori sur ceux-ci. Bien qu'on puisse avancer qu'une source de subjectivité est ajoutée par l'introduction d'une distribution a priori sur les paramètres, nous sommes d'avis qu'il est plus subjectif encore de choisir le plan sans aucun type d'évaluation des hypothèses. De plus, l'inférence reste fondée sur le plan, puisque la distribution a priori n'est utilisée que dans le choix du plan de sondage.

L'illustration de la méthode au moyen d'un ensemble de données réelles a donné des résultats satisfaisants. Ajoutons que bien que les illustrations utilisent un échantillonnage aléatoire simple stratifié, la méthode présentée dans l'article est valide pour tout plan de sondage.

Annexe

Démonstration de (4.2)

Démonstration. Les espérances suivantes sont requises dans la démonstration,

$$\mathrm{E}_\xi Y_k = \mathrm{E}_\xi\left[f(x_k\,|\,\beta_1) + \varepsilon_k\right] = f(x_k\,|\,\beta_1) \tag{A.1}$$

$$\mathrm{E}_\xi Y_k^2 = \mathrm{E}_\xi\left[\left(f(x_k\,|\,\beta_1) + \varepsilon_k\right)^2\right] = f(x_k\,|\,\beta_1)^2 + \sigma^2 g(x_k\,|\,\beta_2)^2 \tag{A.2}$$

$\mathrm{E}_\xi \bar{Y}$, $\mathrm{E}_\xi \bar{Y}^2$ et $\mathrm{E}_\xi \overline{fY}$ sont obtenus à l'aide de (A.1) et (A.2),

$$\mathrm{E}_\xi \bar{Y} = \mathrm{E}_\xi\left[\frac{1}{N}\sum_U Y_k\right] = \frac{1}{N}\sum_U \mathrm{E}_\xi Y_k = \frac{1}{N}\sum_U f(x_k\,|\,\beta_1) \equiv \bar{f} \tag{A.3}$$

$$\mathrm{E}_\xi \bar{Y}^2 = \mathrm{E}_\xi\left[\frac{1}{N}\sum_U Y_k^2\right] = \frac{1}{N}\sum_U\left(f(x_k\,|\,\beta_1)^2 + \sigma^2 g(x_k\,|\,\beta_2)^2\right) \equiv \bar{f}^2 + \sigma^2\bar{g}^2 \tag{A.4}$$

$$\mathrm{E}_\xi \overline{fY} = \mathrm{E}_\xi\left[\frac{1}{N}\sum_U f(x_k\,|\,\beta) Y_k\right] = \frac{1}{N}\sum_U f(x_k\,|\,\beta)\,\mathrm{E}_\xi Y_k = \frac{1}{N}\sum_U f(x_k\,|\,\beta)^2 = \bar{f}^2. \tag{A.5}$$

Puis, en utilisant (A.3), (A.4) et (A.5), nous obtenons

$$\mathrm{E}_\xi\left[\overline{fY} - \overline{fY}\right] = \bar{f}^2 - \bar{f}^2 = S_{f,f} \tag{A.6}$$

$$\mathrm{E}_\xi\left[\bar{Y}^2 - \bar{Y}^2\right] = \bar{f}^2 + \sigma^2\bar{g}^2 - \bar{f}^2 = S_{f,f} + \sigma^2\bar{g}^2. \tag{A.7}$$

En utilisant (A.6) et (A.7), nous obtenons une approximation du coefficient de corrélation, $R_{f,y}$,

$$R_{f,y}^2 = \frac{\left(\overline{fy} - \overline{fy}\right)^2}{\left(\bar{f}^2 - \bar{f}^2\right)\left(\bar{y}^2 - \bar{y}^2\right)} \approx \frac{\mathrm{E}_\xi^2\left[\overline{fY} - \overline{fY}\right]}{\mathrm{E}_\xi\left[\left(\bar{f}^2 - \bar{f}^2\right)\left(\bar{Y}^2 - \bar{Y}^2\right)\right]} = \frac{S_{f,f}}{S_{f,f} + \sigma^2\bar{g}^2}. \tag{A.8}$$

La résolution de (A.8) pour σ^2 permet d'obtenir (4.2), comme souhaité. La démonstration de (5.6) est analogue.

Bibliographie

Beaumont, J.-F., Haziza, D. et Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 3, 555-569.

Bramati, M. (2012). Robust Lavallée-Hidiroglou stratified sampling strategy. *Survey Research Methods*, 6, 3, 137-143.

Cassel, C.M., Särndal, C.-E. et Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 3, 615-620.

Cassel, C.M., Särndal, C.-E. et Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York: John Wiley & Sons, Inc.

Dalenius, T., et Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.

Hájek, J. (1959). Optimal strategy and other problems in probability sampling. *Casopis Pro Pestování Matematiky*, 84, 4, 387-423.

Holmberg, A., et Swensson, B. (2001). On pareto πps sampling: Reflections on unequal probability sampling strategies. *Theory of Stochastic Processes*, 7(23), 142-155.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 260, 663-685.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Lanke, J. (1973). On UMV-estimators in survey sampling. *Metrika*, 20, 196-202.

Narasimhan, B., Johnson, S., Hahn, T., Bouvier, A. et Kiêu, K. (2019). *Cubature: Adaptive Multivariate Integration Over Hypercubes.* R package version 2.0.4. https://CRAN.R-project.org/package=cubature.

Nedyalkova, D., et Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95, 3, 521-537.

R Core Team (2020). R: A language and environment for statistical computing. *The R Foundation for Statistical Computing*, Vienne, Autriche. Adresse web https://www.R-project.org/.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.

Rosén, B. (2000). *Generalized Regression Estimation and Pareto* πps. Rapport R&D 2000:5. Statistics Sweden.

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

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

Tillé, Y., et Wilhelm, M. (2017). Probability sampling designs: Principles for choice of design and balancing. *Statistical Science*, 32(2), 176-189.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

Zhai, Z., et Wiens, D. (2015). Robust model-based stratification sampling designs. *The Canadian Journal of Statistics*, 43, 4, 554-577.

Techniques d'enquête, juin 2021
Vol. 47, nº 1, pp. 99-132
Statistique Canada, nº 12-001-X au catalogue

Inférence bayésienne prédictive des proportions dans de petits domaines avec biais de sélection

Seongmi Choi, Balgobin Nandram et Dalho Kim[1]

Résumé

Dans un article précédent, nous avons élaboré un modèle pour effectuer une inférence sur des proportions de petits domaines en cas de biais de sélection dans lequel les réponses binaires et les probabilités de sélection sont corrélées. Il s'agit du modèle de sélection non ignorable homogène; une sélection non ignorable signifie que les probabilités de sélection et les réponses binaires sont corrélées. Il a été montré que le modèle de sélection non ignorable homogène donne de meilleurs résultats qu'un modèle de sélection ignorable de référence. Toutefois, l'une des limites du modèle de sélection non ignorable homogène réside dans le fait que les distributions des probabilités de sélection sont supposées identiques dans tous les domaines. C'est pourquoi nous introduisons un modèle plus général, le modèle de sélection non ignorable hétérogène, dans lequel les probabilités de sélection ne sont pas distribuées identiquement dans tous les domaines. Nous avons utilisé des méthodes de Monte Carlo par chaînes de Markov pour ajuster les trois modèles. Nous illustrons notre méthodologie et comparons nos modèles à l'aide d'un exemple sur la limitation d'activité sévère de la *U.S. National Health Interview Survey* (Enquête nationale sur la santé réalisée par interviews aux États-Unis). Nous réalisons également une étude par simulations pour démontrer que notre modèle de sélection non ignorable hétérogène est nécessaire en présence d'un biais de sélection modéré ou fort.

Mots-clés : Corrélation bisériale; algorithme de Metropolis-Hastings; modèle de sélection non ignorable; statistiques officielles; probabilités de sélection.

1. Introduction

Dans de nombreuses enquêtes-échantillons complexes, les personnes sont échantillonnées selon des probabilités de sélection différentes. Concernant les réponses binaires, si la proportion de réponses positives parmi les personnes échantillonnées diffère considérablement de celle des personnes non échantillonnées, il y a un biais de sélection. Dans certains cas, le biais de sélection résulte plan (par exemple échantillonnage avec probabilité proportionnelle à la taille (PPT)) et on peut s'occuper du biais de sélection, mais dans d'autres problèmes, ce n'est pas le cas. Par exemple, dans les échantillons non probabilistes, les probabilités de sélection sont inconnues. Dans notre problème, nous supposons que les poids d'enquête ou les probabilités de sélection peuvent nous aider à comprendre le biais de sélection. Nous cherchons à effectuer une inférence sur les proportions de population finie de petits domaines quand on dispose d'un échantillon possiblement biaisé pour chaque domaine. Choi, Nandram et Kim (2017), désignés ci-après par CNK, ont étendu le modèle de Nandram, Bhatta, Bhadra et Shen (2013), désignés ci-après par NBBS, qui étudiait un seul domaine, pour l'adapter à l'inférence dans des petits domaines. Alors que le modèle de CNK suppose que la distribution des probabilités de sélection est identique dans tous les domaines, notre nouvelle contribution consiste à supposer que les distributions des probabilités de sélection diffèrent entre domaines, mais qu'elles ont un effet en commun.

1. Seongmi Choi, Health Insurance Policy Research Institute (Institut de recherche sur les politiques d'assurance maladie), National Health Insurance Service (Service national d'assurance maladie), Wonju, Gangwon 26464, Corée du Sud. Courriel : edisil1202@gmail.com; Balgobin Nandram, Département des sciences mathématiques, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, États-Unis. Courriel : balnan@wpi.edu; Dalho Kim, Département de statistique, Kyungpook National University, Daegu 41566, Corée du Sud. Courriel : dalkim@knu.ac.kr.

Quand on effectue une inférence sur de petits domaines, on peut prendre en compte deux types de modèles. Premièrement, on peut utiliser un modèle de sélection ignorable dans lequel la variable réponse n'est pas liée aux probabilités de sélection. Ce modèle ne corrigera pas le biais de sélection et produira des estimations biaisées en l'absence d'autres covariables agissant comme les probabilités de sélection. Nous supposons que les seuls renseignements disponibles sont les réponses échantillonnées, les probabilités de sélection de l'échantillon et les identificateurs de domaine. Deuxièmement, dans un modèle de sélection non ignorable, les variables réponses sont liées aux probabilités de sélection. Dans notre étude des réponses binaires, la distribution des probabilités de sélection pour les réponses « oui » est différente de celle des réponses « non », ce qui fait que les variables réponses et les probabilités de sélection sont corrélées. Les statistiques officielles sont obtenues à partir de nombreuses enquêtes complexes, qui sont conçues de façon pratique avec un biais de sélection (un échantillonnage PPT fait habituellement partie d'un plan de sondage complexe). Cette étude est importante pour la construction de statistiques officielles à partir d'enquêtes complexes.

Il nous semble pertinent de discuter brièvement d'avancées récentes. Concernant les réponses continues, Pfeffermann (1988), Pfeffermann (1993), Sverchkov et Pfeffermann (2004) et Pfeffermann, Krieger et Rinott (1998) ont spécifié le lien entre les poids d'enquête et la variable réponse. Chambers, Dorfman et Wang (1998) ont supposé que les probabilités de sélection étaient liées aux réponses continues. Ma, Sedransk, Nandram et Chen (2018) ont publié une étude sur le même thème au sein du paradigme bayésien. Cette méthode s'applique également au calcul de la consommation totale de gaz aux États-Unis au moyen d'un échantillon PPT; voir Nandram, Choi, Shen et Burgos (2006). Chen, Elliott et Little (2010) ont effectué une inférence bayésienne prédictive au moyen d'une spline pénalisée pour un échantillonnage PPT. Toutefois, comme ces approches exigent de l'information sur les probabilités de sélection hors de l'échantillon, elles sont difficiles à utiliser. Les analystes de données secondaires n'ont pas accès aux probabilités de sélection hors de l'échantillon et, dans nos travaux, nous continuons de supposer que ces probabilités de sélection hors de l'échantillon ne sont pas disponibles. Pfeffermann et Sverchkov (2007) ont étendu les travaux de Sverchkov et de Pfeffermann (2004) afin de tenir compte des petits domaines par un échantillonnage informatif des domaines et à l'intérieur des domaines sélectionnés. À l'instar de Chen et coll. (2010), Opsomer, Glaeskens, Ranalli, Kauermann et Breidt (2008) ont utilisé une régression par spline pénalisée afin de construire un modèle pour petits domaines, qui comprend les probabilités de sélection, de façon non paramétrique. Alors que tous ces travaux portent sur une réponse continue dans de petits domaines, nous analysons des données binaires dans le présent article.

Pour traiter des problèmes de biais de sélection, nous avons choisi de corriger la partie échantillon d'un modèle de population. Autrement dit, un modèle est construit pour l'ensemble de la population, puis on le corrige pour tenir compte de l'échantillon avec les probabilités de sélection, avant d'estimer les paramètres de superpopulation. Une fois que cela est effectué pour l'échantillon, la prédiction est calculée pour l'ensemble de la population finie. Malec, Davis et Cao (1999) ont habilement décrit cette approche du biais de sélection. NBBS ont examiné la méthode de Malec et coll. (1999) et en ont fourni une analyse

bayésienne complète; Nandram (2007) a fourni une méthode d'échantillonnage de substitution dans l'esprit de Malec et coll. (1999). Plusieurs travaux dans le prolongement de cette approche conviennent aux petits domaines. Par exemple, pour analyser les réponses continues, Nandram et Choi (2010) ont inclus les probabilités de sélection dans un modèle bayésien complet non ignorable afin d'analyser les données continues de l'indice de masse corporelle, qui ont été discrétisées élégamment. Malec et coll. (1999) ont utilisé un modèle bayésien hiérarchique pour tenir compte d'un mécanisme de sélection de données binaires. Comme l'énoncent NBBS, ce modèle peut poser deux problèmes : « Premièrement, les θ_{uy} sont faiblement identifiés. Deuxièmement, les paramètres θ_{uy} ne sont jamais connus, et dans un cadre bayésien, ils doivent aussi être stochastiques. Dans le présent article, nous montrons, en une seule tentative, comment résoudre ces deux problèmes (faible identifiabilité et paramètres stochastiques) pour un échantillon biaisé tiré d'une population binaire à l'aide de l'information des poids d'enquête (ou probabilités de sélection) ». Au moyen d'un modèle de sélection non ignorable, NBBS ont montré comment corriger ces deux problèmes pour un modèle à un domaine.

Récemment, des activités bayésiennes intéressantes pour échantillonnage PPT avec réponses continues ont été réalisées. Zangeneh et Little (2015) ont utilisé un bootstrap bayésien pour modéliser les mesures de la taille (liées aux probabilités de sélection) et une régression spline des réponses conditionnelle aux mesures de la taille. Pour le même problème, Si, Pillai et Gelman (2015) ont poststratifié les probabilités de sélection et modélisé les réponses au moyen d'un processus gaussien. Ce sont de bonnes approches des analyses bayésiennes d'un échantillonnage PPT, applicables à notre problème. Toutefois, nous avons choisi une démarche différente : en effet, nous ne modélisons pas les probabilités de sélection non échantillonnées, mais nous corrigeons plutôt un modèle de population.

CNK ont étudié un cas particulier d'extension du modèle de NBBS dans le contexte des petits domaines. Ils ont supposé que les probabilités de sélection de l'échantillon ont le même support sur l'ensemble π_u^* dans les domaines, et que la distribution des probabilités de sélection étant donné la réponse binaire y_{ij} est

$$P\left(\pi_{ij} = \pi_u^* \,\middle|\, \boldsymbol{\theta},\, y_{ij} = y\right) = \theta_{uy},\ u = 1, \ldots, U,\ y = 0, 1,\ j = 1, \ldots, n_i,\ i = 1, \ldots, \ell.$$

Ce modèle suppose que les probabilités de sélection de l'échantillon ont le même support et que chaque domaine a une distribution identique pour les probabilités de sélection; on parle de modèle de sélection non ignorable homogène. Cependant, l'hypothèse d'homogénéité est forte et pourrait ne pas être vraie dans la plupart des contextes réels. En pratique, la distribution des probabilités de sélection de l'échantillon peut être très différente d'un domaine à l'autre en raison des plans de sondage. La distribution des probabilités de sélection de l'échantillon étant donné $\mathbf{y}$ peut aussi varier considérablement d'un domaine à l'autre. Dans l'article, nous examinons un modèle fondé sur une hypothèse d'hétérogénéité, selon laquelle les probabilités de sélection de l'échantillon ont des supports différents et des distributions différentes selon les domaines : on parle de modèle de sélection non ignorable hétérogène.

À partir d'une étude par simulations, CNK montrent à quel point un modèle de sélection ignorable de référence et un modèle de sélection non ignorable homogène peuvent être différents en cas de biais de sélection. On sait bien que l'information sur le plan de sondage doit être incluse quand l'échantillon n'est pas sélectionné aléatoirement. Dans le présent article, nous utilisons une étude par simulations pour évaluer les performances de notre modèle de sélection non ignorable hétérogène. Nous tirons des données du modèle de sélection non ignorable homogène et du modèle de sélection non ignorable hétérogène, respectivement, et nous ajustons les trois modèles. Ensuite, nous comparons les performances des modèles au moyen de plusieurs mesures.

Dans l'article, nous étudions le problème de l'inférence sur des proportions de population finie dans de petits domaines quand le biais de sélection est probablement différent selon le domaine. Plus précisément, nous étendons le modèle de sélection non ignorable homogène de CNK pour tenir compte de probabilités de sélection ayant des supports différents dans différents domaines. Dans la section 2, nous examinons le modèle de sélection ignorable et le modèle de sélection non ignorable homogène, qui ont été étudiés précédemment. Dans la section 3, nous montrons comment modifier un modèle bayésien de sélection ignorable pour intégrer le biais de sélection dans un cadre de petits domaines quand les probabilités de sélection de l'échantillon ont des distributions différentes selon le domaine. Nous décrivons également le mode de calcul. Dans la section 4, nous présentons les résultats d'un exemple, concernant la limitation d'activité sévère dans la *National Health Interview Survey* de 1995. Nous proposons aussi une étude par simulations pour évaluer les performances du modèle de sélection non ignorable hétérogène. La section 5 contient un résumé et des observations finales.

2. Synthèse

La présente section donne un aperçu du modèle de sélection ignorable (SI) discuté par NBBS et du modèle de sélection non ignorable homogène (SniHo) discuté par CNK. Nous supposons qu'il y a ℓ domaines et que la taille de la population du i^{e} domaine est N_i, $i = 1, \ldots, \ell$, qui sont connues. Nous prenons en considération les réponses binaires.

Les échantillons prélevés dans chaque domaine peuvent être biaisés, car la proportion de réponses positives dans les unités échantillonnées peut être différente de la proportion de réponses positives dans les unités non échantillonnées. Un échantillon de $n_i << N_i$ est prélevé dans le i^{e} domaine, et la j^{e} unité dans le i^{e} domaine est supposée avoir les probabilités de sélection π_{ij}, $j = 1, \ldots, N_i$, $i = 1, \ldots, \ell$. Comme cela est courant dans de nombreux problèmes de ce genre, les probabilités de sélection sont observées seulement pour les valeurs échantillonnées. (Seules les probabilités de sélection de l'échantillon sont présentées aux utilisateurs de données secondaires.) Les scientifiques chargés du plan de sondage corrigent les probabilités de sélection pour tenir compte des erreurs dues à la non-réponse et non dues à l'échantillonnage. Néanmoins, les probabilités de sélection constituent une partie importante des poids d'enquête et nous supposons que les probabilités de sélection sont approximativement les réciproques des poids d'enquête dans notre application.

Soit $y_{i1}, \ldots, y_{iN_i}$, $i = 1, \ldots, \ell$ les réponses binaires dans les domaines ℓ. Un échantillon biaisé S_i de taille n_i est disponible dans le i^{e} domaine avec les probabilités de sélection de l'échantillon. Notons les valeurs échantillonnées par $y_{i1}, \ldots, y_{in_i}$, $i = 1, \ldots, \ell$ et l'ensemble de valeurs non échantillonnées par $\bar{S}_i$. Une inférence est requise pour la proportion de population finie dans chaque domaine. Soit $P_i = \sum_{j=1}^{N_i} y_{ij} / N_i$ la proportion de petit domaine du i^{e} domaine et $\hat{p}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ les proportions d'échantillon correspondantes. De toute évidence, $\hat{p}_i$ peut être un estimateur biaisé de P_i, la proportion de uns dans la population finie. Dans l'analyse d'enquête fondée sur le plan, les valeurs P_i sont des quantités fixes inconnues, mais dans le paradigme bayésien, les P_i sont des variables aléatoires, qui doivent être prédites (c'est-à-dire par inférence bayésienne prédictive).

Nous examinons le modèle de SI à la section 2.1 et le modèle de SniHo à la section 2.2.

2.1 Modèle de sélection ignorable

Un modèle standard de sélection ignorable pour les variables binaires y_{ij}, $j = 1, \ldots, N_i$, $i = 1, \ldots, \ell$, est

$$y_{ij} \mid p_i \overset{\text{ind}}{\sim} \text{Bernoulli}(p_i),$$

$$p_i \mid \mu, \tau \overset{\text{iid}}{\sim} \text{Bêta}(\mu\tau, (1-\mu)\tau)$$

et

$$p(\mu, \tau) = \frac{1}{(1+\tau)^2}, \ \tau \geq 0 \, (0 < \mu < 1)$$

(c'est-à-dire les lois a priori de μ et τ sont indépendantes). Ici, la loi a priori de τ est propre et non informative.

Aux fins de simplification de la notation, supposons $\mathbf{p} = (p_1, \ldots, p_\ell)$, $\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{\ell 1}, \ldots, y_{\ell n_\ell})$ et $s_i = \sum_{j=1}^{n_i} y_{ij}$. L'inférence bayésienne prédictive de P_i peut alors être réalisée sur la base de la distribution a posteriori suivante :

$$p(y_{ij} \mid \mathbf{y}) = \int p(y_{ij} \mid p_i, \mathbf{y}) \pi(p_i \mid \mathbf{y}) dp_i = \int p(y_{ij} \mid p_i) \pi(p_i \mid \mathbf{y}) dp_i, \ j = n_i + 1, \ldots, N_i, \ i = 1, \ldots, \ell,$$

où $y_{ij} \mid p$ sont indépendants, et $\pi(p_i \mid \mathbf{y}) = \int \pi(p_i \mid \mu, \tau, \mathbf{y}) \pi(\mu, \tau \mid \mathbf{y}) d\mu d\tau$ et $p_i \mid \mu, \tau, \mathbf{y}$ sont indépendants avec $p_i \mid \mu, \tau, \mathbf{y} \sim \text{Bêta}(s_i + \mu\tau, n_i - s_i + (1-\mu)\tau)$.

2.2 Modèle de sélection non ignorable homogène

CNK ont étudié un cas particulier d'extension du modèle de NBBS dans le contexte des petits domaines. Ils ont supposé que les probabilités de sélection de l'échantillon $(\pi_{i1}, \ldots, \pi_{in_i})$ ont le même support sur l'ensemble π_u^*, $u = 1, \ldots, U$ pour $i = 1, \ldots, \ell$.

Soit $\boldsymbol{\theta} = (\theta_{10}, \ldots, \theta_{U0}, \theta_{11}, \ldots, \theta_{U1}) = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ (disons) et $\boldsymbol{\pi} = \left(\pi_{11}, \ldots, \pi_{1n_1}, \ldots, \pi_{\ell 1}, \ldots, \pi_{\ell n_\ell}\right)$, la distribution des probabilités de sélection, étant donnée la réponse binaire y_{ij}, est

$$P\left(\pi_{ij} = \pi_u^* \,|\, \boldsymbol{\theta}, y_{ij} = y\right) = \theta_{uy},\ u = 1, \ldots, U,\ y = 0, 1,\ j = 1, \ldots, n_i,\ i = 1, \ldots, \ell$$

et

$$y_{ij} \,|\, p_i \overset{\text{iid}}{\sim} \text{Bernoulli}\,(p_i),\ j = 1, \ldots, n_i,\ i = 1, \ldots, \ell.$$

Il faut noter qu'ici les valeurs θ_{uy} ne sont pas des probabilités de sélection.

De plus, les probabilités de sélection de l'échantillon ont un support et une distribution identiques dans les domaines. Cette hypothèse est toutefois forte et pourrait ne pas être vraie dans la plupart des contextes réels. En pratique, la distribution des probabilités de sélection de l'échantillon peut être très différente d'un domaine à l'autre en raison des plans de sondage.

Pour avancer, ils effectuent une transformation bijective de p_i à q_i via

$$q_i = \frac{a_1 p_i}{a_1 p_i + a_0 (1 - p_i)},\ i = 1, \ldots, \ell.$$

Ils supposent ensuite que $\mathbf{q}$, $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$ sont indépendants avec

$$q_i \,|\, \mu, \tau_1 \overset{\text{iid}}{\sim} \text{Bêta}\,(\mu\tau_1, (1 - \mu)\tau_1),\ i = 1, \ldots, \ell,$$

$$\boldsymbol{\theta}_0 \,|\, \tau_0 \sim \text{Dirichlet}\left(\boldsymbol{\theta}_0^{(0)}\tau_0\right) \quad \text{et} \quad \boldsymbol{\theta}_1 \,|\, \tau_0 \sim \text{Dirichlet}\left(\boldsymbol{\theta}_1^{(0)}\tau_0\right),$$

où $\boldsymbol{\theta}_0^{(0)}$ et $\boldsymbol{\theta}_1^{(0)}$ doivent être spécifiés.

A priori, ils supposent

$$p(\mu, \tau_0, \tau_1) = \frac{1}{(1 + \tau_0)^2} \frac{1}{(1 + \tau_1)^2},\ 0 < \mu < 1,\ \tau_0, \tau_1 \geq 0.$$

Après avoir tiré un échantillon de la priorité a posteriori $\pi(\mathbf{q}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \mu, \tau_0, \tau_1 \,|\, \mathbf{y})$, au moyen d'une retransformation de q_i à p_i,

$$p_i = \frac{a_0 q_i}{a_0 q + a_1 (1 - q_i)},$$

ils obtiennent un échantillon de la distribution a posteriori de $\pi(\mathbf{p}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \mu, \tau_0, \tau_1 \,|\, \mathbf{y})$.

Enfin, on peut tirer les valeurs de population finie entière, y_{ij}, à partir de $\text{Bernoulli}\,(p_i)$, $i = 1, \ldots, \ell$, $j = 1, \ldots, n_i$ indépendamment, au moyen de la valeur $\mathbf{p}$ estimée, et effectuer une inférence sur les proportions de petits domaines. La répartition des probabilités de sélection de l'échantillon étant donné $\mathbf{y}$ peut aussi varier considérablement d'un domaine à l'autre. C'est pourquoi, à la section 3, nous examinons un modèle ayant des supports et des distributions différents pour les probabilités de sélection de l'échantillon selon le domaine.

3. Modèle de sélection non ignorable hétérogène

Dans la section 3.1, nous décrivons le modèle de sélection non ignorable hétérogène (SniHe). Nous décrivons ensuite le mode de calcul à la section 3.2. Nous montrons une méthode d'ajustement du modèle et comment effectuer une inférence sur des proportions de petits domaines. L'inférence sur des proportions de petits domaines selon le modèle de SniHe est obtenue au moyen d'échantillons de substitution (Nandram, 2007).

3.1 Méthodologie

Nous supposons que les probabilités de sélection de l'échantillon $(\pi_{i1}, \ldots, \pi_{in_i})$ ont les supports différents sur l'ensemble $\pi_{iu}^*, u = 1, \ldots, U_i,$ pour $i = 1, \ldots, \ell$. C'est-à-dire que $\pi_{ij}, j = 1, \ldots, n_i$ ont un histogramme où les points médians des catégories sont les $\pi_{iu}^*,$ pour $i = 1, \ldots, \ell$. On suppose que ces valeurs π_{iu}^* sont connues et que les π_{ij} sont des quantités aléatoires. Aux fins de simplification de la notation, supposons $\boldsymbol{\theta}_i = (\theta_{i10}, \ldots, \theta_{iU_i 0}, \theta_{i11}, \ldots, \theta_{iU_i 1}) = (\boldsymbol{\theta}_{i0}, \boldsymbol{\theta}_{i1})$ (disons), $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\ell)$ et $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1n_1}, \ldots, \pi_{\ell 1}, \ldots, \pi_{\ell n_\ell})$. La distribution des probabilités de sélection, étant donnée la réponse binaire y_{ij}, est

$$P(\pi_{ij} = \pi_{iu}^* \,|\, \boldsymbol{\theta}, y_{ij} = y) = \theta_{iuy}, u = 1, \ldots, U_i, y = 0, 1, j = 1, \ldots, n_i, i = 1, \ldots, \ell$$

et

$$y_{ij} \,|\, p_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_i), j = 1, \ldots, N_i, i = 1, \ldots, \ell.$$

Encore une fois, il faut souligner que les valeurs θ_{iuy} ne sont pas des probabilités de sélection.

Pour tenir compte du plan de sélection de l'échantillon, nous supposons que θ_{iu0} et θ_{iu1}, pour $i = 1, \ldots, \ell,$ sont différents. Notons que nous prenons en considération l'hypothèse d'hétérogénéité pour les probabilités de sélection de l'échantillon. Nous remplaçons l'hypothèse d'homogénéité par l'hypothèse d'hétérogénéité pour les probabilités de sélection de l'échantillon du modèle de SniHo, de sorte que les probabilités de sélection de l'échantillon ont des supports différents et que les distributions des probabilités de sélection sont différentes selon le domaine.

Supposons que δ_{ij}, π_{ij} et y_{ij} désignent respectivement l'indicateur de sélection, la probabilité de sélection et la réponse binaire de la j^{e} unité dans le i^{e} petit domaine de la population. Essentiellement, NBBS postulent que les variables $(\delta_{ij}, \pi_{ij}, y_{ij})$ se trouvant dans le i^{e} petit domaine sont distribuées indépendamment avec

$$\begin{aligned} P(\delta_{ij} = \delta, \pi_{ij} = \pi_{iu}^*, y_{ij} = y \,|\, \boldsymbol{\theta}_i, p_i) &= P_1(\delta_{ij} = \delta \,|\, \pi_{ij} = \pi_{iu}^*) P_2(\pi_{ij} = \pi_{iu}^* \,|\, y_{ij} = y, \boldsymbol{\theta}_i) P_3(y_{ij} = y \,|\, p_i) \\ &= (\pi_{iu}^*)^\delta (1 - \pi_{iu}^*)^{1-\delta} \theta_{iuy} p_i^y (1 - p_i)^{1-y}, \\ &\quad \delta = 0, 1, u = 1, \ldots, U_i, y = 0, 1, j = 1, \ldots, n_i, i = 1, \ldots, \ell. \end{aligned}$$

Il y a alors une fonction de masse conjointe pour l'indicateur de sélection et l'indicateur de réponse. Par conséquent, le modèle que NBBS ont spécifié est un modèle de sélection non ignorable (c'est-à-dire que

NBBS ont supposé que le mécanisme de sélection est une sélection non aléatoire (SNA)). Comme il n'y a pas de données quand $\delta = 0$ (c'est-à-dire que π et y sont tous deux non observés), NBBS ont utilisé une fonction de masse conditionnelle

$$P\left(\pi_{ij} = \pi_{iu}^*,\, y_{ij} = y \,|\, \delta_{ij} = 1,\, \boldsymbol{\theta}_i,\, p_i\right) = \frac{\pi_{iu}^* \theta_{iuy}\, p_i^y\, (1 - p_i)^{1-y}}{\sum_{y=0}^{1} \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iuy}\, p_i^y\, (1 - p_i)^{1-y}},\ i = 1, \ldots, \ell;$$

voir la fonction de masse de probabilité dans (4) de NBBS.

Nous avons les données (π_{ij}, y_{ij}), $j = 1, \ldots, n_i$, $i = 1, \ldots, \ell$. Puisque les unités d'échantillonnage sont indépendantes, la fonction de vraisemblance est donnée par

$$\prod_{i=1}^{\ell} \prod_{j=1}^{n_i} P\left(\pi_{ij} = \pi_{iu}^*,\, Y_{ij} = y_{ij} \,|\, \delta_{ij} = 1,\, \boldsymbol{\theta}_i,\, \mathbf{p}\right) = \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \frac{\pi_{iu}^* \theta_{iuy_{ij}}\, p_i^{y_{ij}}\, (1 - p_i)^{1-y_{ij}}}{\sum_{y_{ij}=0}^{1} \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iuy_{ij}}\, p_i^{y_{ij}}\, (1 - p_i)^{1-y_{ij}}},$$

où π_{iu}^*, $u = 1, \ldots, U_i$, $i = 1, \ldots, \ell$ sont connus. La fonction de vraisemblance peut être réécrite comme suit :

$$\begin{aligned} P(\mathbf{y}, \boldsymbol{\pi} \,|\, \boldsymbol{\theta}, \mathbf{p}) &= \frac{\prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \pi_{iu}^* \theta_{iuy_{ij}} \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} p_i^{y_{ij}}\, (1 - p_i)^{1-y_{ij}}}{\prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \sum_{y_{ij}=0}^{1} \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iuy_{ij}}\, p_i^{y_{ij}}\, (1 - p_i)^{1-y_{ij}}} \\ &= \frac{\prod_{i=1}^{\ell} \prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu0})^{g_{iu0}} \prod_{i=1}^{\ell} \prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu1})^{g_{iu1}} \prod_{i=1}^{\ell} p_i^{s_i}\, (1 - p_i)^{n_i - s_i}}{\prod_{i=1}^{\ell} \left[p_i \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu1} + (1 - p_i) \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu0} \right]^{n_i}}, \end{aligned}$$

où $s_i = \sum_{j=1}^{n_i} y_{ij}$, g_{iu0} est le nombre de cellules pour la catégorie u à $y = 0$ et g_{iu1} est le nombre de cellules pour la catégorie u sous le domaine $y = 1$. Notons que $\sum_{u=1}^{U_i} g_{iu0} = n_i - s_i$, $\sum_{u=1}^{U_i} g_{iu1} = s_i$ et $\sum_{u=1}^{U_i} (g_{iu0} + g_{iu1}) = n_i$. Cette vraisemblance inclut le biais de sélection.

Soit $a_{iy} = \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iuy}$, $y = 0, 1$, $i = 1, \ldots, \ell$. Les différences entre a_{i0} et a_{i1} pour certains i indiquent qu'il y a un biais de sélection. La fonction de vraisemblance peut être exprimée comme suit :

$$P(\mathbf{y}, \boldsymbol{\pi} \,|\, \boldsymbol{\theta}, \mathbf{p}) = \prod_{i=1}^{\ell} \prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu0})^{g_{iu0}} \prod_{i=1}^{\ell} \prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu1})^{g_{iu1}} \prod_{i=1}^{\ell} \frac{p_i^{s_i}\, (1 - p_i)^{n_i - s_i}}{\left[a_{i1} p_i + a_{i0} (1 - p_i) \right]^{n_i}}.$$

Nous effectuons une transformation bijective de p_i à q_i via

$$q_i = \frac{a_{i1} p_i}{a_{i1} p_i + a_{i0} (1 - p_i)}.$$

Notons que si θ_{iuy} ne dépend pas de y, q_i et p_i sont identiques. Dans ce cas, la vraisemblance serait la même que dans le cas ignorable. Soit $\mathbf{q} = (q_1, \ldots, q_\ell)$. Alors, la fonction de vraisemblance peut être exprimée comme suit :

$$P(\mathbf{y}, \boldsymbol{\pi} \,|\, \boldsymbol{\theta}, \mathbf{q}) = \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu0})^{g_{iu0}}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu0} \right)^{n_i - s_i}} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu1})^{g_{iu1}}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu1} \right)^{s_i}} \prod_{i=1}^{\ell} q_i^{s_i}\, (1 - q_i)^{n_i - s_i}.$$

Nous supposons que $\mathbf{q}, \boldsymbol{\theta}_{i0}, \boldsymbol{\theta}_{i1}$ sont indépendants, et nous supposons

$$q_i \,|\, \mu, \tau_1 \overset{\text{iid}}{\sim} \text{Bêta}\,(\mu\tau_1, (1-\mu)\tau_1),\ i = 1, \ldots, \ell,$$

$$\boldsymbol{\theta}_{i0} \,|\, \tau_0 \overset{\text{ind}}{\sim} \text{Dirichlet}\left(\boldsymbol{\theta}_{i0}^{(0)}\tau_0\right) \quad \text{et} \quad \boldsymbol{\theta}_{i1} \,|\, \tau_0 \overset{\text{ind}}{\sim} \text{Dirichlet}\left(\boldsymbol{\theta}_{i1}^{(0)}\tau_0\right),$$

où $\boldsymbol{\theta}_{i0}^{(0)}$ et $\boldsymbol{\theta}_{i1}^{(0)}$ doivent être spécifiés. Gardons à l'esprit que $\mathbf{x}\,|\,\boldsymbol{\mu}, \tau \sim \text{Dirichlet}\,(\boldsymbol{\mu}\tau)$ a la densité $f\,(\mathbf{x}\,|\,\boldsymbol{\mu}, \tau) = \frac{\prod_{i=1}^{k} x_i^{\mu_i\tau-1}}{D(\boldsymbol{\mu}\tau)}$, $0 < x_i < 1$, $\sum_{i=1}^{k} x_i = 1$, où $D(\boldsymbol{\mu}\tau) = \prod_{i=1}^{k} \Gamma\,(\mu_i\tau)/\Gamma\,(\tau)$, $0 < \mu_i < 1$, $\sum_{i=1}^{k} \mu_i = 1$, et $\tau > 0$.

Enfin, a priori, nous supposons

$$p\,(\mu, \tau_0, \tau_1) = \frac{1}{(1+\tau_0)^2}\,\frac{1}{(1+\tau_1)^2},\ 0 < \mu < 1,\ \tau_0, \tau_1 \geq 0.$$

Bien sûr, on peut utiliser une densité suivant une loi demi-Cauchy, mais celles-ci sont très similaires. On utilise la dernière probabilité a priori pour éviter les difficultés associées aux probabilités a priori impropres de la forme $p\,(\tau) \propto 1/\tau$ (par exemple voir Gelman, 2006).

Par conséquent, la densité a priori conjointe de $\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1$ est

$$\pi\,(\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1) \propto \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{\theta_{iu0}^{(0)}\tau_0-1}}{D\left(\boldsymbol{\theta}_{i0}^{(0)}\tau_0\right)} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{\theta_{iu1}^{(0)}\tau_0-1}}{D\left(\boldsymbol{\theta}_{i1}^{(0)}\tau_0\right)}$$
$$\times \frac{\prod_{i=1}^{\ell} q_i^{\mu\tau_1-1}\,(1-q_i)^{(1-\mu)\tau_1-1}}{\left[B\,(\mu\tau_1, (1-\mu)\tau_1)\right]^{\ell}}\,\frac{1}{(1+\tau_0)^2}\,\frac{1}{(1+\tau_1)^2}.$$

Au moyen du théorème de Bayes, la densité a posteriori conjointe de $\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1$, étant donné les données $\mathbf{y}$, est

$$\pi(\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 \,|\, \mathbf{y}) \propto \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \left(\pi_{iu}^{*}\theta_{iu0}\right)^{g_{iu0}}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*}\theta_{iu0}\right)^{n_i-s_i}} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \left(\pi_{iu}^{*}\theta_{iu1}\right)^{g_{iu1}}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*}\theta_{iu1}\right)^{s_i}}$$
$$\times \prod_{i=1}^{\ell} q_i^{s_i}\,(1-q_i)^{n_i-s_i} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{\theta_{iu0}^{(0)}\tau_0-1}}{D\left(\boldsymbol{\theta}_{i0}^{(0)}\tau_0\right)} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{\theta_{iu1}^{(0)}\tau_0-1}}{D\left(\boldsymbol{\theta}_{i1}^{(0)}\tau_0\right)}$$
$$\times \frac{\prod_{i=1}^{\ell} q_i^{\mu\tau_1-1}\,(1-q_i)^{(1-\mu)\tau_1-1}}{\left[B\,(\mu\tau_1, (1-\mu)\tau_1)\right]^{\ell}}\,\frac{1}{(1+\tau_0)^2}\,\frac{1}{(1+\tau_1)^2}.$$

Pour améliorer les calculs, nous utilisons le théorème de Rao-Blackwell, qui est plus optimal, pour obtenir la densité a posteriori de $\mathbf{q}$ (donc $\mathbf{p}$). Compte tenu des données, la densité a posteriori conjointe peut être exprimée comme suit :

$$\pi(\mathbf{q}, \boldsymbol{\theta}, \mu, \tau_0, \tau_1 \,|\, \mathbf{y}) = \pi(\mathbf{q} \,|\, \boldsymbol{\theta}, \mu, \tau_0, \tau_1, \mathbf{y})\pi(\boldsymbol{\theta}, \mu, \tau_0, \tau_1 \,|\, \mathbf{y}).$$

3.2 Calculs

En éliminant par intégration $\mathbf{q}$ de la probabilité a posteriori conjointe de $\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1$ étant donné $\mathbf{y}$, nous obtenons la densité a posteriori conjointe marginale de $\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1$ étant donné $\mathbf{y}$,

$$\pi(\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 | \mathbf{y}) \propto \frac{\prod_{i=1}^{\ell} B(s_i + \mu\tau_1, n_i - s_i + (1-\mu)\tau_1)}{[B(\mu\tau_1, (1-\mu)\tau_1)]^{\ell}}$$
$$\times \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{g_{iu0} + \theta_{iu0}^{(0)}\tau_0 - 1}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*} \theta_{iu0}\right)^{n_i - s_i}} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{g_{iu1} + \theta_{iu1}^{(0)}\tau_0 - 1}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*} \theta_{iu1}\right)^{s_i}}$$
$$\times \frac{1}{\prod_{i=1}^{\ell} D(\boldsymbol{\theta}_{i0}^{(0)}\tau_0)} \frac{1}{\prod_{i=1}^{\ell} D(\boldsymbol{\theta}_{i1}^{(0)}\tau_0)} \frac{1}{(1+\tau_0)^2} \frac{1}{(1+\tau_1)^2}.$$

La densité postérieure conditionnelle de $\mathbf{q}$ est donnée par

$$\pi(\mathbf{q} | \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1, \mathbf{y}) \propto \prod_{i=1}^{\ell} q_i^{s_i + \mu\tau_1 - 1} (1 - q_i)^{n_i - s_i + (1-\mu)\tau_1 - 1},$$

que nous pouvons échantillonner directement.

Nous obtenons un échantillon $\mathbf{q}$ de la manière suivante : Nous tirons chaque élément de $(\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1)$ à partir de la densité a posteriori conditionnelle de $\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 | \mathbf{y}$ au moyen de l'algorithme de Metropolis-Hastings et d'une méthode de grille, puis nous tirons $\mathbf{q}$ de la densité a posteriori conditionnelle de $\mathbf{q} | \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1, \mathbf{y}$.

Nous avons surveillé la convergence de l'échantillonneur de Metropolis-Hastings à l'aide de tracés, de graphiques d'autocorrélation et du test de stationnarité de Geweke, ce qui a donné des résultats satisfaisants.

Les densités a posteriori conditionnelles nécessaires à l'exécution de l'échantillonneur de Metropolis-Hastings sont

$$\pi(\boldsymbol{\theta}_{i0} | \boldsymbol{\theta}_{j0}, j \neq i, \boldsymbol{\theta}_{i1}, i = 1, \ldots, \ell, \mu, \tau_0, \tau_1, \mathbf{y}) \propto \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{g_{iu0} + \theta_{iu0}^{(0)}\tau_0 - 1}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*} \theta_{iu0}\right)^{n_i - s_i}}, \; i = 1, \ldots, \ell,$$

$$\pi(\boldsymbol{\theta}_{i1} | \boldsymbol{\theta}_{j1}, j \neq i, \boldsymbol{\theta}_{i0}, i = 1, \ldots, \ell, \mu, \tau_0, \tau_1, \mathbf{y}) \propto \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{g_{iu1} + \theta_{iu1}^{(0)}\tau_0 - 1}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*} \theta_{iu1}\right)^{s_i}}, \; i = 1, \ldots, \ell,$$

$$\pi(\mu | \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \tau_0, \tau_1, \mathbf{y}) \propto \frac{\prod_{i=1}^{\ell} B(s_i + \mu\tau_1, n_i - s_i + (1-\mu)\tau_1)}{[B(\mu\tau_1, (1-\mu)\tau_1)]^{\ell}},$$

et la transformation de τ_0 et τ_1 en respectivement $\rho_0 = 1/(1+\tau_0)$ et $\rho_1 = 1/(1+\tau_1)$,

$$\pi\left(\rho_0 \,|\, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_1, \mathbf{y}\right) \propto \left[\prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{\theta_{iu0}^{(0)}\tau_0 - 1}}{D\left(\boldsymbol{\theta}_{i0}^{(0)}\tau_0\right)} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{\theta_{iu1}^{(0)}\tau_0 - 1}}{D\left(\boldsymbol{\theta}_{i1}^{(0)}\tau_0\right)}\right]_{\tau_0 = (1-\rho_0)/\rho_0},$$

$$\pi\left(\rho_1 \,|\, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \rho_0, \mathbf{y}\right) \propto \left[\frac{\prod_{i=1}^{\ell} B\left(s_i + \mu\tau_1, n_i - s_i + (1-\mu)\tau_1\right)}{\left[B\left(\mu\tau_1, (1-\mu)\tau_1\right)\right]^{\ell}}\right]_{\tau_1 = (1-\rho_1)/\rho_1}.$$

Dans la formule ci-dessus, les variables $\boldsymbol{\theta}_{i0}$, $\boldsymbol{\theta}_{i1}$ et μ sont conditionnellement indépendantes.

Nous utilisons les étapes de Metropolis pour échantillonner $\boldsymbol{\theta}_{i0}$, $i = 1, \ldots, \ell$ et $\boldsymbol{\theta}_{i1}$, $i = 1, \ldots, \ell$ à partir des distributions conditionnelles $\pi\left(\boldsymbol{\theta}_{i0} \,|\, \boldsymbol{\theta}_{j0}, j \neq i, \boldsymbol{\theta}_{i1}, i = 1, \ldots, \ell, \mu, \tau_0, \tau_1, \mathbf{y}\right)$ et $\pi(\boldsymbol{\theta}_{i1} \,|\, \boldsymbol{\theta}_{j1}, j \neq i, \boldsymbol{\theta}_{i0}, i = 1, \ldots, \ell, \mu, \tau_0, \tau_1, \mathbf{y})$, respectivement. Soit $\tau_{i0}^* = n_i - s_i + \tau_0$, $\theta_{iu0}^* = \left(g_{iu0} + \theta_{iu0}^{(0)}\tau_0\right)/\tau_{i0}^*$, $\tau_{i1}^* = s_i + \tau_0$, et $\theta_{iu1}^* = \left(g_{iu1} + \theta_{iu1}^{(0)}\tau_0\right)/\tau_{i1}^*$. Aux fins de simplification de la notation, supposons que $\boldsymbol{\theta}_{i0}^* = \left(\theta_{i10}^*, \ldots, \theta_{iU_i 0}^*\right)$ et $\boldsymbol{\theta}_{i1}^* = \left(\theta_{i11}^*, \ldots, \theta_{iU_i 1}^*\right)$. Nous choisissons des lois de Dirichlet surdispersées comme densités de proposition pour $\boldsymbol{\theta}_{i0}$ et $\boldsymbol{\theta}_{i1}$, $i = 1, \ldots, \ell$. En fait, les densités de proposition sont $\text{Dirichlet}\left(\boldsymbol{\theta}_{i0}^*\tilde{\tau}_{i0}\right)$ et $\text{Dirichlet}\left(\boldsymbol{\theta}_{i1}^*\tilde{\tau}_{i1}\right)$, où $\tilde{\tau}_{i0} = \tau_{i0}^*/\kappa_{i0}$ et $\tilde{\tau}_{i1} = \tau_{i1}^*/\kappa_{i1}$, pour tous les $i = 1, \ldots, \ell$. Notons qu'à mesure que κ_{i0} et κ_{i1} augmentent, la dispersion tend à augmenter dans la loi de Dirichlet.

En supposant que l'échantillonneur de Metropolis-Hastings est à $\boldsymbol{\theta}_{i0}^{(r)}$, alors la probabilité d'accepter $\boldsymbol{\theta}_{i0}^{(r+1)}$ est

$$U_{r,r+1}^{(0)} = \min\left\{\frac{\psi\left(\boldsymbol{\theta}_{i0}^{(r+1)}, \mathbf{y}\right)}{\psi\left(\boldsymbol{\theta}_{i0}^{(r)}, \mathbf{y}\right)}, 1\right\},$$

où $\psi\left(\boldsymbol{\theta}_{i0}, \mathbf{y}\right) = \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{\theta_{iu0}^* \tau_{i0}^* (1 - 1/\kappa_{i0})}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu0}\right)^{n_i - s_i}}$, pour tous les $i = 1, \ldots, \ell$. De plus, si nous supposons que l'échantillonneur de Metropolis-Hastings est à $\boldsymbol{\theta}_{i1}^{(r)}$, alors la probabilité d'accepter $\boldsymbol{\theta}_{i1}^{(r+1)}$ est

$$U_{r,r+1}^{(1)} = \min\left\{\frac{\psi\left(\boldsymbol{\theta}_{i1}^{(r+1)}, \mathbf{y}\right)}{\psi\left(\boldsymbol{\theta}_{i1}^{(r)}, \mathbf{y}\right)}, 1\right\},$$

où $\psi\left(\boldsymbol{\theta}_{i1}, \mathbf{y}\right) = \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{\theta_{iu1}^* \tau_{i1}^* (1 - 1/\kappa_{i1})}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu1}\right)^{s_i}}$, pour tous les $i = 1, \ldots, \ell$.

L'étape de Metropolis s'obtient comme suit : Supposons que la chaîne de Markov est à $\boldsymbol{\theta}_{i0}^{(r)}$, un vecteur aléatoire $\boldsymbol{\theta}_{i0}^{(r+1)}$ est tiré de la densité de proposition avec une valeur κ_{i0} correctement choisie, et $U_{r,r+1}^{(0)}$ est calculé. Un écart uniforme aléatoire U dans $[0, 1]$ est tiré, et si $U \leq U_{r,r+1}^{(0)}$, alors le vecteur aléatoire $\boldsymbol{\theta}_{i0}^{(r+1)}$ est accepté, sinon la chaîne reste à $\boldsymbol{\theta}_{i0}^{(r)}$. Cet algorithme est appliqué à tous les $i = 1, \ldots, \ell$. L'étape de Metropolis est utilisée de manière similaire pour $\boldsymbol{\theta}_{i1}$ pour tous les $i = 1, \ldots, \ell$.

Nous générons μ, ρ_0 et ρ_1, au moyen de la méthode de grille; voir CNK. Après avoir tiré un échantillon de la priorité a posteriori $\pi\left(\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 \,|\, \mathbf{y}\right)$, au moyen d'une retransformation de q_i à p_i,

$$p_i = \frac{a_{i0} q_i}{a_{i0} q + a_{i1}\left(1 - q_i\right)},$$

nous pouvons tirer un échantillon de la distribution a posteriori de $\pi(\mathbf{p}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 | \mathbf{y})$.

Une fois que $\mathbf{p}$ est estimé, nous tirons les valeurs de la population finie entière, $y_{i1}, \ldots, y_{iN_i}$, indépendamment de Bernoulli(p_i), $i = 1, \ldots, \ell$. Il s'agit d'un échantillonnage de substitution (par exemple Nandram, 2007). Nous avons donc corrigé l'échantillon biaisé observé et l'avons remplacé par un échantillon de substitution pour $\mathbf{p}$ que nous avons obtenu à partir du modèle de sélection non ignorable hétérogène. Nous pouvons obtenir un échantillon de P_i en tirant $\sum_{i=1}^{N_i} y_{ij}$ de Binomial(N_i, p_i) et en divisant le résultat par N_i pour tous les $i = 1, \ldots, \ell$.

Le mécanisme de sélection ressemble au mécanisme des données manquantes. Il est donc possible d'incorporer les données manquantes dans notre cadre, ou indépendamment (c'est-à-dire en elles-mêmes), nous pouvons supposer que les données manquantes sont des « valeurs manquantes non au hasard » et utiliser un modèle de non-réponse non ignorable pour corriger un modèle de population, voir Nandram et Choi (2010).

4. Études numériques

Dans la section 4.1, nous donnons en exemple la limitation d'activité sévère. Nous présentons les résultats des modèles de SI, de SniHo et de SniHe pour les comparer. À la section 4.2, nous décrivons une étude par simulations visant à évaluer les performances des trois modèles selon deux types d'hypothèses de la distribution des probabilités de sélection de l'échantillon. Autrement dit, les données sont générées à partir du modèle de sélection non ignorable homogène ou du modèle de sélection non ignorable hétérogène, et les trois modèles sont ajustés.

4.1 Un exemple parlant

Dans notre application, nous utilisons les données de la *National Health Interview Survey* de 1995 (NHIS95, Enquête nationale sur la santé réalisée par interviews aux États-Unis). Ces données ont d'abord été utilisées par Nandram, Bai et Choi (2011) aux fins d'estimation du point de changement de la limitation d'activité. Dans leur étude, NBBS ont construit des données synthétiques provenant d'un segment de la NHIS95. Pour les adultes âgés de 30 à 80 ans, NBBS ont analysé les données sur la limitation d'activité sévère (LAS) pour un seul domaine (sans agrégation), quand $y = 1$ si un adulte est atteint de LAS et $y = 0$ s'il ne l'est pas. Au moyen des données de NBBS, CNK ont effectué une estimation pour petits domaines de la LAS. Nous utilisons les données d'une manière semblable à NBBS, mais nous ajustons un modèle plus général, ce qui représente la principale contribution du présent article.

NBBS ont formé douze domaines (petits domaines) en croisant le niveau d'études, le sexe et la race. Ils ont classé l'éducation en trois niveaux d'études (pré-collégial : L, collégial : M et post-collégial : H). Le

sexe (masculin : M et féminin : F) et la race (blanc : W et non blanc : B) ont deux niveaux chacune. Nous continuerons d'appeler ces domaines LMW, LMB, LFW, LFB, MMW, MMB, MFW, MFB, HMW, HMB, HFW, HFB (par exemple LMW : hommes blancs avec niveau d'études pré-collégial, LMB : hommes noirs avec niveau d'études pré-collégial, etc.).

La NHIS95 a employé un plan de sondage à plusieurs degrés pour tirer des échantillons dans la population des États-Unis. Par conséquent, il est nécessaire d'utiliser un poids d'enquête d'adulte pour que l'analyse soit exacte. Voir l'analyse des poids d'enquête dans NBBS et CNK. Comme NBBS et CNK, nous avons considéré le réciproque d'un poids d'enquête comme la probabilité de « sélection » de chaque adulte. Les probabilités de sélection constituent une partie majeure des poids d'enquête. Il aurait été préférable d'avoir les probabilités de sélection, mais cela est une approximation que nous faisons dans notre analyse des données. Par souci de commodité, nous avons présenté les données de nouveau dans le tableau 4.1.

Tableau 4.1
Valeurs sommaires des données sur la limitation d'activité sévère (LAS), y compris les probabilités de sélection

Domaine	n	s	p	CV	$f = n/N$	avg0	avg1	Valeur p
LMW	174	26	0,149	0,181	0,0000303	0,0004059	0,0003706	0,025*
LMB	31	10	0,323	0,260	0,0000280	0,0003179	0,0003167	0,614
LFW	200	22	0,110	0,201	0,0000325	0,0004523	0,0004381	1,00
LFB	40	9	0,225	0,293	0,0000286	0,0003152	0,0003740	0,910
MMW	760	56	0,074	0,129	0,0000227	0,0002710	0,0002901	0,006*
MMB	151	18	0,119	0,221	0,0000250	0,0002820	0,0002847	0,056
MFW	892	42	0,047	0,151	0,0000233	0,0002891	0,0002415	0,837
MFB	200	21	0,105	0,206	0,0000278	0,0003064	0,0003377	0,469
HMW	756	14	0,019	0,265	0,0000213	0,0002484	0,0001919	0,095
HMB	124	7	0,056	0,367	0,0000238	0,0002702	0,0003506	0,146
HFW	779	22	0,028	0,210	0,0000219	0,0002575	0,0002976	0,072
HFB	168	2	0,012	0,703	0,0000257	0,0002969	0,0001948	1,00

NOTE : Ici n est la taille totale de l'échantillon, s est le nombre d'adultes ayant une LAS, et $p = s/n$; $f = n/N$ est la fraction de sondage; CV = $\frac{\sqrt{(1-p)}}{\sqrt{np}}$ est le coefficient de variation; avg0 est la moyenne des probabilités de sélection pour $y = 0$ (LAS, non) et avg1 est la moyenne des probabilités de sélection pour $y = 1$ (LAS, oui); la valeur p correspond à celle du test du chi carré de l'égalité de $\theta_{iu0} = \theta_{iu1}$, $u = 1, \ldots, U_i$ $(U_i = 5)$; les domaines sont formés par croisement du niveau d'études (L, M, H), du sexe (M, F) et de la race (W, B).

Nous réduisons la taille de l'échantillon par rapport à l'ensemble de données initial afin d'accroître l'effet du modèle pour petits domaines. Pour le modèle de SniHe, nous ordonnons les probabilités de sélection de la plus petite à la plus grande $\pi_{(1)}, \ldots, \pi_{(n_i)}$ dans chaque petit domaine. Soit les quantiles $t_{i1} = \pi_{(0,20n_i)}$, $t_{i2} = \pi_{(0,40n_i)}$, $t_{i3} = \pi_{(0,60n_i)}$, $t_{i4} = \pi_{(0,80n_i)}$ et soit $t_{i0} = \pi_{(1)}$ et $t_{i5} = \pi_{(n_i)}$ pour $i = 1, \ldots, \ell$. Nous définissons $\pi^*_{iu} = (t_{i(u-1)} + t_{iu})/2$, $u = 1, \ldots, U_i$, $i = 1, \ldots, \ell$ (c'est-à-dire le point médian de chaque quantile dans chaque petit domaine). Notons que θ_{iuy} est la proportion d'unités échantillonnées dans le u^{e} quantile conditionnel à $y = 0, 1$ pour $i = 1, \ldots, \ell$. Si les θ_{iu0} sont considérablement différents

de θ_{iu1} dans certains domaines, il y a de solides preuves que les valeurs échantillonnées sont biaisées. Pour spécifier $\boldsymbol{\theta}_{i0}^{(0)}$ et $\boldsymbol{\theta}_{i1}^{(0)}$ dans les distributions a priori, nous prenons $\boldsymbol{\theta}_{iy}^{(0)} = \hat{\boldsymbol{\theta}}_{iy}$, l'estimateur par le maximum de vraisemblance de $\boldsymbol{\theta}_{iy}$, $y = 0, 1$, $i = 1, \ldots, \ell$, que nous appelons une probabilité a priori « EMV » (voir NBBS).

Pour ce qui est de nos données, le mélange est lent, donc nous tirons 120 000 échantillons et en mettons 60 000 en rodage. Ensuite, nous prenons chaque 30e itération pour obtenir un échantillon de 2 000 itérations aux fins d'inférence. Cette période de rodage est suffisamment longue pour que nous obtenions des échantillons aléatoires, et est basée sur les tracés et le test de Geweke. Nous avons suffisamment d'échantillons puisque la taille effective d'échantillon (TEE) des paramètres échantillonnés à partir de la méthode MCMC doit être inférieure ou égale à 2 000. La corrélation est non significative pour tous les paramètres. De plus, la stationnarité de notre échantillonneur est démontrée par le test de Geweke. Les résultats du diagnostic de convergence des hyperparamètres sont présentés dans le tableau 4.8 et la figure 4.2. En supposant $\boldsymbol{\kappa}_0 = (7{,}3;\ 3;\ 11;\ 4{,}8;\ 8{,}5;\ 9;\ 11{,}5;\ 10;\ 10;\ 9{,}7;\ 10{,}2;\ 11)$ et $\boldsymbol{\kappa}_1 = (3{,}2;\ 1{,}7;\ 2{,}2;\ 1{,}5;\ 5;\ 2{,}2;\ 3{,}3;\ 2{,}2;\ 2{,}5;\ 1{,}5;\ 1{,}5;\ 1{,}3)$, nous avons obtenu des taux d'acceptation de 25 % et 45 % des échantillonneurs de Metropolis-Hastings.

Les valeurs sommaires des données se trouvent dans le tableau 4.1. Les moyennes des probabilités de sélection dans chaque domaine sont pour la plupart similaires pour $y = 0$ et $y = 1$. Comme les tailles d'échantillon dans les domaines ne sont pas grandes, les fractions d'échantillonnage sont très petites. Par conséquent, l'analyse simultanée au moyen d'un modèle pour petits domaines est appropriée. Dans la colonne 4 du tableau 4.1, nous avons également présenté la proportion de personnes ayant une LAS, et nous pouvons voir que cette valeur est relativement grande chez les personnes ayant un faible niveau d'études. Nous avons comparé les nombres dans les deux ensembles de compartiments à partir des histogrammes des probabilités de sélection pour $y = 0, 1$ dans chaque domaine. En fait, il s'agit d'un test d'indépendance dans un tableau catégorique $2 \times U_i$, et nous utilisons un test du chi carré et un test exact de Fisher pour l'égalité de $\theta_{iu0} = \theta_{iu1}$, $u = 1, \ldots, U_i$ ($U_i = 5$ cellules) dans chaque domaine. Comme le montrent clairement les valeurs p présentées dans la dernière colonne du tableau 4.1, le biais de sélection devrait avoir de l'importance surtout dans le domaine MMW et peut-être dans les domaines LMW, MMB, HMW et HFW. Comme mesure du biais de sélection, nous avons également calculé la corrélation bisériale entre les valeurs y_{ij} et π_{ij} pour tous les domaines combinés et avons obtenu une valeur de 0,033 avec une valeur p de 0,03.

Dans le tableau 4.2, nous présentons les résultats de l'analyse (individuelle) sans agrégation des proportions de population finie selon le modèle de sélection ignorable et le modèle de sélection non ignorable pour chaque domaine séparément (le modèle de sélection non ignorable est celui de NBBS). Nous les comparons au moyen des moyennes a posteriori (MP), des écarts-types a posteriori (ETP) des

MP, du coefficient de variation (CV) et des intervalles de la plus haute densité a posteriori (HDP) à 95 %. Il s'agit d'une répétition de l'analyse selon le modèle de NBBS.

Dans le tableau 4.3, nous comparons les valeurs sommaires des estimateurs agrégés des proportions de la population finie dans le modèle de SI, le modèle de SniHo et le modèle de SniHe pour les 12 domaines. L'effet du biais de sélection est observé parce que les MP dans le modèle de SI et les autres modèles sont différentes. Les estimateurs selon le modèle de SniHo et le modèle de SniHe sont semblables dans certains domaines parce que le biais de sélection est plus important pour certains domaines. Les ETP selon le modèle de SniHo et le modèle de SniHe sont plus grands que selon le modèle de SI dans la plupart des domaines, mais les intervalles de HDP à 95 % se chevauchent. L'effet de l'analyse simultanée est également observé parce que les MP pour les estimateurs agrégés sont lissées par rapport aux MP pour les populations finies analysées séparément.

Les valeurs sommaires a posteriori de $\boldsymbol{\theta}_{i0}$ et $\boldsymbol{\theta}_{i1}$ pour $i = 1, \ldots, \ell$ sont présentées dans les tableaux 4.4. Les valeurs θ_{iu0} diffèrent des θ_{iu1} dans certains domaines. Cela prouve solidement la présence d'un biais de sélection. Nous présentons également les valeurs sommaires a posteriori de $\boldsymbol{\theta}_0$ et $\boldsymbol{\theta}_1$ selon le modèle de SniHo dans le tableau 4.5. Notons que dans les tableaux 4.4 et 4.5, les variables θ_{i0} et θ_{i1} ne sont pas des probabilités de sélection. Les valeurs $\boldsymbol{\theta}$ selon le modèle de SniHo semblent différer de celles selon le modèle de SniHe. Nous présentons les valeurs sommaires a posteriori de a_{i0} et a_{i1} pour $i = 1, \ldots, \ell$ dans le tableau 4.7, et a_0 et a_1 selon le modèle de SniHo dans le tableau 4.6. Ces valeurs sont très petites, mais le ratio est grand dans chaque domaine. Il s'agit également d'une forte preuve de biais de sélection, bien qu'il s'agisse de petites tailles d'échantillon dans chaque domaine.

Tableau 4.2
Comparaisons des estimateurs non agrégés des proportions de population finie selon les modèles de sélection ignorable et non ignorable pour chaque domaine

Domaine	Modèle de sélection ignorable				Modèle de sélection non ignorable			
	MP	ETP	CV	IC	MP	ETP	CV	IC
LMW	0,154	0,027	0,175	(0,098; 0,214)	0,188	0,033	0,176	(0,123; 0,265)
LMB	0,333	0,080	0,240	(0,177; 0,525)	0,336	0,083	0,247	(0,170; 0,531)
LFW	0,114	0,022	0,193	(0,068; 0,165)	0,113	0,022	0,195	(0,069; 0,165)
LFB	0,244	0,066	0,270	(0,108; 0,394)	0,215	0,064	0,298	(0,090; 0,361)
MMW	0,075	0,010	0,133	(0,055; 0,100)	0,068	0,009	0,132	(0,050; 0,086)
MMB	0,124	0,026	0,210	(0,073; 0,185)	0,114	0,024	0,211	(0,065; 0,171)
MFW	0,048	0,008	0,167	(0,032; 0,065)	0,052	0,008	0,154	(0,036; 0,072)
MFB	0,108	0,022	0,204	(0,066; 0,161)	0,104	0,021	0,202	(0,060; 0,148)
HMW	0,020	0,005	0,250	(0,008; 0,032)	0,021	0,005	0,238	(0,011; 0,034)
HMB	0,065	0,022	0,338	(0,021; 0,119)	0,046	0,018	0,391	(0,014; 0,089)
HFW	0,030	0,006	0,200	(0,017; 0,045)	0,021	0,005	0,238	(0,011; 0,032)
HFB	0,017	0,010	0,588	(0,001; 0,043)	0,021	0,012	0,571	(0,001; 0,048)

NOTE : Ici, MP est la moyenne a posteriori, ETP l'écart-type a posteriori, CV est le coefficient de variation et IC est l'intervalle de HDP à 95 %.

Tableau 4.3
Comparaisons des estimateurs agrégés des proportions de la population finie selon le modèle de sélection ignorable (SI), le modèle de sélection non ignorable homogène (SniHo) et le modèle de sélection non ignorable hétérogène (SniHe) par domaine

Domaine	Modèle	MP	ETP	CV	IC
LMW	SI	0,148	0,026	0,176	(0,094; 0,210)
	SniHo	0,177	0,030	0,169	(0,123; 0,238)
	SniHe	0,172	0,037	0,215	(0,099; 0,264)
LMB	SI	0,264	0,071	0,269	(0,109; 0,426)
	SniHo	0,310	0,082	0,265	(0,167; 0,483)
	SniHe	0,267	0,076	0,285	(0,120; 0,449)
LFW	SI	0,110	0,021	0,191	(0,062; 0,158)
	SniHo	0,134	0,026	0,194	(0,088; 0,189)
	SniHe	0,116	0,027	0,233	(0,058; 0,179)
LFB	SI	0,198	0,057	0,288	(0,083; 0,326)
	SniHo	0,237	0,064	0,270	(0,122; 0,377)
	SniHe	0,192	0,060	0,313	(0,074; 0,332)
MMW	SI	0,074	0,009	0,122	(0,054; 0,095)
	SniHo	0,091	0,011	0,121	(0,071; 0,115)
	SniHe	0,081	0,013	0,160	(0,054; 0,111)
MMB	SI	0,119	0,026	0,218	(0,067; 0,178)
	SniHo	0,144	0,030	0,208	(0,092; 0,207)
	SniHe	0,120	0,027	0,225	(0,060; 0,183)
MFW	SI	0,048	0,007	0,146	(0,032; 0,064)
	SniHo	0,059	0,009	0,153	(0,044; 0,078)
	SniHe	0,060	0,011	0,183	(0,036; 0,085)
MFB	SI	0,106	0,021	0,198	(0,062; 0,154)
	SniHo	0,128	0,026	0,203	(0,084; 0,183)
	SniHe	0,105	0,023	0,219	(0,059; 0,160)
HMW	SI	0,020	0,005	0,250	(0,010; 0,032)
	SniHo	0,025	0,006	0,240	(0,014; 0,039)
	SniHe	0,027	0,008	0,296	(0,013; 0,047)
HMB	SI	0,062	0,021	0,339	(0,019; 0,111)
	SniHo	0,075	0,025	0,333	(0,035; 0,131)
	SniHe	0,057	0,020	0,351	(0,019; 0,108)
HFW	SI	0,029	0,006	0,207	(0,017; 0,044)
	SniHo	0,037	0,008	0,216	(0,023; 0,053)
	SniHe	0,027	0,006	0,222	(0,014; 0,042)
HFB	SI	0,018	0,010	0,556	(0,001; 0,043)
	SniHo	0,022	0,012	0,545	(0,005; 0,050)
	SniHe	0,020	0,012	0,600	(0,001; 0,048)

Tableau 4.4
Comparaison de θ_{i0} et θ_{i1} au moyen de la MP, de l'ETP, de l'erreur-type numérique et de l'intervalle de HDP à 95 % selon le modèle de sélection non ignorable hétérogène

Domaine	u	θ_0				θ_1			
		MP	ETP	CV	IC	MP	ETP	CV	IC
LMW	1	0,260	0,083	0,319	(0,096; 0,447)	0,477	0,109	0,229	(0,252; 0,714)
	2	0,213	0,075	0,352	(0,073; 0,397)	0,195	0,087	0,446	(0,048; 0,401)
	3	0,214	0,070	0,327	(0,084; 0,383)	0,126	0,070	0,556	(0,008; 0,296)
	4	0,189	0,064	0,339	(0,065; 0,342)	0,079	0,052	0,658	(0,003; 0,208)
	5	0,124	0,048	0,387	(0,027; 0,235)	0,121	0,057	0,471	(0,019; 0,252)
LMB	1	0,318	0,098	0,308	(0,101; 0,536)	0,355	0,100	0,282	(0,140; 0,570)
	2	0,279	0,086	0,308	(0,058; 0,422)	0,176	0,080	0,455	(0,020; 0,361)
	3	0,209	0,083	0,397	(0,048; 0,340)	0,148	0,073	0,493	(0,021; 0,324)
	4	0,126	0,065	0,516	(0,015; 0,286)	0,229	0,083	0,362	(0,063; 0,414)
	5	0,114	0,059	0,518	(0,012; 0,254)	0,093	0,054	0,581	(0,007; 0,224)
LFW	1	0,279	0,087	0,312	(0,101; 0,473)	0,301	0,090	0,299	(0,113; 0,500)
	2	0,232	0,083	0,358	(0,075; 0,431)	0,215	0,083	0,386	(0,051; 0,410)
	3	0,194	0,075	0,387	(0,054; 0,371)	0,203	0,079	0,389	(0,056; 0,390)
	4	0,167	0,065	0,389	(0,046; 0,325)	0,191	0,071	0,372	(0,046; 0,348)
	5	0,128	0,052	0,406	(0,025; 0,245)	0,089	0,045	0,506	(0,008; 0,198)
LFB	1	0,308	0,098	0,318	(0,120; 0,538)	0,276	0,094	0,341	(0,084; 0,480)
	2	0,247	0,095	0,385	(0,058; 0,463)	0,183	0,083	0,454	(0,024; 0,377)
	3	0,192	0,085	0,443	(0,032; 0,391)	0,200	0,086	0,430	(0,044; 0,410)
	4	0,158	0,074	0,468	(0,023; 0,326)	0,202	0,081	0,401	(0,050; 0,392)
	5	0,095	0,058	0,611	(0,005; 0,239)	0,139	0,069	0,496	(0,024; 0,308)
MMW	1	0,204	0,043	0,211	(0,116; 0,304)	0,167	0,078	0,467	(0,035; 0,355)
	2	0,221	0,044	0,199	(0,133; 0,326)	0,186	0,081	0,435	(0,039; 0,380)
	3	0,213	0,043	0,202	(0,126; 0,314)	0,198	0,082	0,414	(0,043; 0,396)
	4	0,189	0,040	0,212	(0,110; 0,282)	0,359	0,094	0,262	(0,174; 0,594)
	5	0,173	0,033	0,191	(0,104; 0,248)	0,090	0,047	0,522	(0,013; 0,205)
MMB	1	0,284	0,086	0,303	(0,107; 0,472)	0,248	0,095	0,383	(0,059; 0,458)
	2	0,230	0,079	0,343	(0,088; 0,418)	0,168	0,078	0,464	(0,023; 0,347)
	3	0,196	0,068	0,347	(0,055; 0,351)	0,176	0,078	0,443	(0,034; 0,353)
	4	0,148	0,062	0,419	(0,034; 0,297)	0,339	0,096	0,283	(0,134; 0,552)
	5	0,142	0,053	0,373	(0,038; 0,264)	0,069	0,046	0,667	(0,002; 0,186)
MFW	1	0,212	0,045	0,212	(0,127; 0,323)	0,305	0,093	0,305	(0,110; 0,520)
	2	0,218	0,045	0,206	(0,122; 0,321)	0,248	0,085	0,343	(0,084; 0,446)
	3	0,207	0,043	0,208	(0,117; 0,309)	0,191	0,076	0,398	(0,042; 0,364)
	4	0,201	0,043	0,214	(0,117; 0,311)	0,186	0,073	0,392	(0,049; 0,364)
	5	0,162	0,035	0,216	(0,090; 0,246)	0,071	0,038	0,535	(0,006; 0,161)
MFB	1	0,260	0,077	0,296	(0,110; 0,427)	0,331	0,093	0,281	(0,155; 0,568)
	2	0,229	0,073	0,319	(0,071; 0,387)	0,155	0,070	0,452	(0,029; 0,327)
	3	0,203	0,068	0,335	(0,067; 0,356)	0,156	0,069	0,442	(0,032; 0,319)
	4	0,179	0,061	0,341	(0,058; 0,317)	0,189	0,072	0,381	(0,050; 0,362)
	5	0,129	0,049	0,380	(0,032; 0,242)	0,168	0,064	0,381	(0,039; 0,319)
HMW	1	0,224	0,045	0,201	(0,138; 0,333)	0,277	0,105	0,379	(0,079; 0,531)
	2	0,206	0,044	0,214	(0,119; 0,313)	0,282	0,104	0,369	(0,074; 0,513)
	3	0,193	0,044	0,228	(0,104; 0,295)	0,257	0,099	0,385	(0,071; 0,491)
	4	0,210	0,044	0,210	(0,121; 0,312)	0,162	0,079	0,488	(0,025; 0,350)
	5	0,166	0,034	0,205	(0,099; 0,247)	0,022	0,026	1,182	(0,000; 0,097)
HMB	1	0,296	0,097	0,328	(0,095; 0,504)	0,265	0,099	0,374	(0,072; 0,493)
	2	0,223	0,085	0,381	(0,062; 0,423)	0,203	0,091	0,448	(0,027; 0,417)
	3	0,204	0,078	0,382	(0,058; 0,391)	0,125	0,073	0,584	(0,004; 0,305)
	4	0,166	0,070	0,422	(0,040; 0,323)	0,266	0,091	0,342	(0,091; 0,467)
	5	0,112	0,050	0,446	(0,026; 0,236)	0,142	0,069	0,486	(0,021; 0,307)
HFW	1	0,222	0,045	0,203	(0,126; 0,322)	0,191	0,078	0,408	(0,038; 0,367)
	2	0,217	0,047	0,217	(0,125; 0,329)	0,184	0,073	0,397	(0,039; 0,356)
	3	0,207	0,044	0,213	(0,110; 0,304)	0,169	0,071	0,420	(0,030; 0,332)
	4	0,199	0,043	0,216	(0,106; 0,296)	0,260	0,081	0,312	(0,097; 0,442)
	5	0,155	0,034	0,219	(0,079; 0,229)	0,196	0,050	0,255	(0,081; 0,348)
HFB	1	0,274	0,084	0,307	(0,116; 0,470)	0,366	0,114	0,311	(0,136; 0,631)
	2	0,223	0,075	0,336	(0,072; 0,400)	0,189	0,093	0,492	(0,012; 0,400)
	3	0,198	0,071	0,359	(0,060; 0,364)	0,212	0,095	0,448	(0,018; 0,421)
	4	0,177	0,064	0,362	(0,049; 0,317)	0,178	0,088	0,494	(0,024; 0,384)
	5	0,128	0,050	0,391	(0,035; 0,246)	0,055	0,050	0,909	(0,000; 0,194)

NOTE : u indique les cinq intervalles pour les probabilités de sélection.

Tableau 4.5
Comparaison de θ_0 et θ_1 au moyen de la MP, de l'ETP, de l'erreur-type numérique et d'un intervalle de confiance à 95 % selon le modèle de sélection non ignorable homogène

u	θ_0				θ_1			
	MP	ETP	CV	IC	MP	ETP	CV	IC
1	0,384	0,011	0,029	(0,363; 0,405)	0,632	0,036	0,057	(0,559; 0,700)
2	0,222	0,008	0,036	(0,206; 0,237)	0,160	0,026	0,163	(0,114; 0,217)
3	0,189	0,007	0,037	(0,176; 0,203)	0,112	0,018	0,161	(0,080; 0,150)
4	0,153	0,006	0,039	(0,142; 0,164)	0,060	0,009	0,150	(0,045; 0,079)
5	0,053	0,002	0,038	(0,049; 0,057)	0,036	0,005	0,139	(0,027; 0,047)

NOTE : u indique les cinq intervalles pour les probabilités de sélection.

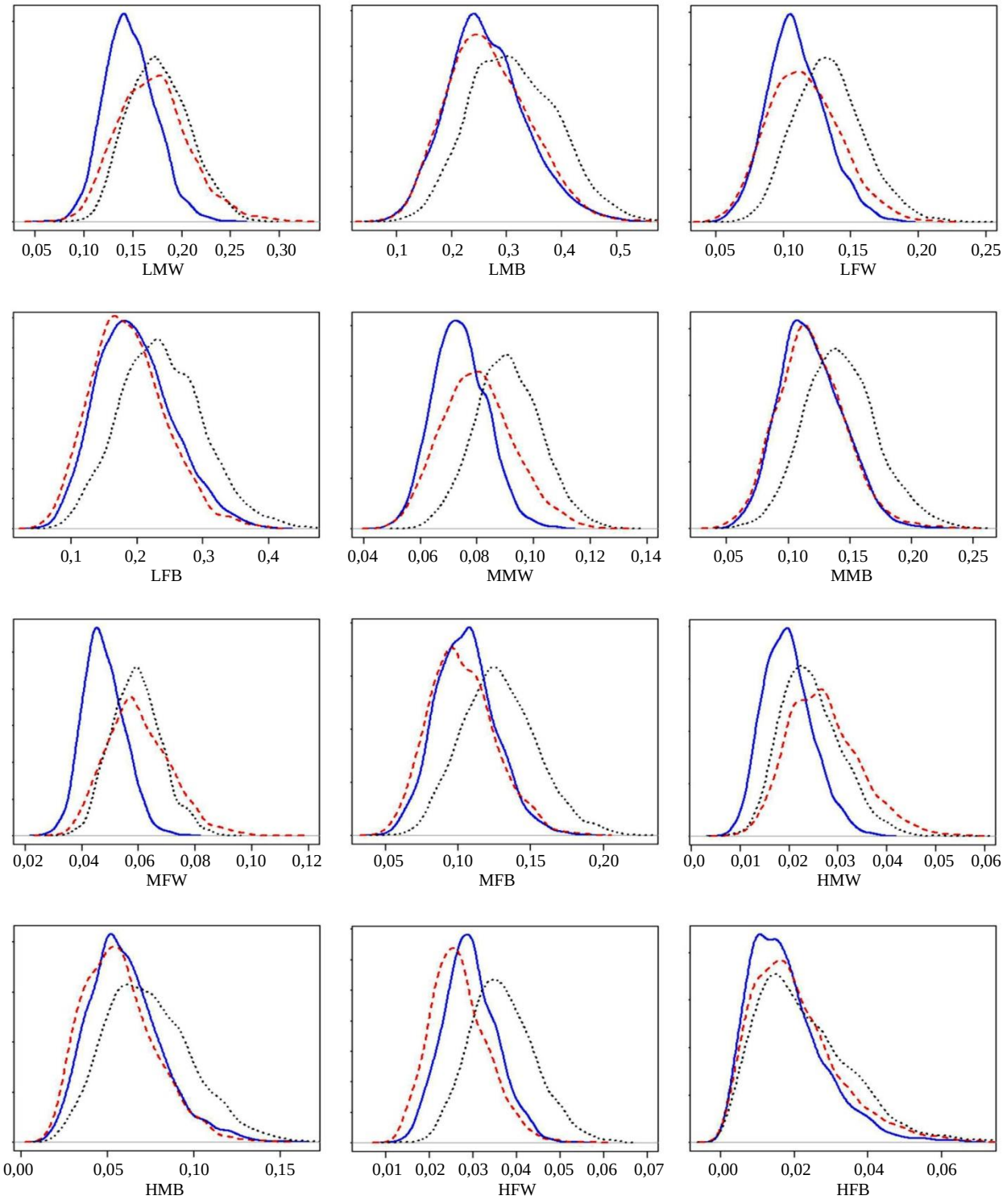
Tableau 4.6
Comparaison de a_0 et a_1 au moyen de la MP, de l'ETP, de l'erreur-type numérique et d'un intervalle de confiance à 95 % selon le modèle de sélection non ignorable homogène

a_y	MP	ETP	CV	IC
a_0	0,000222	0,000002	0,009009	(0,000218; 0,000227)
a_1	0,000177	0,000006	0,033898	(0,000164; 0,000191)

Tableau 4.7
Comparaison de a_{i0} et a_{i1} au moyen de la MP, de l'ETP, de l'erreur-type numérique et de l'intervalle de HDP à 95 % selon le modèle de sélection non ignorable hétérogène pour chaque domaine

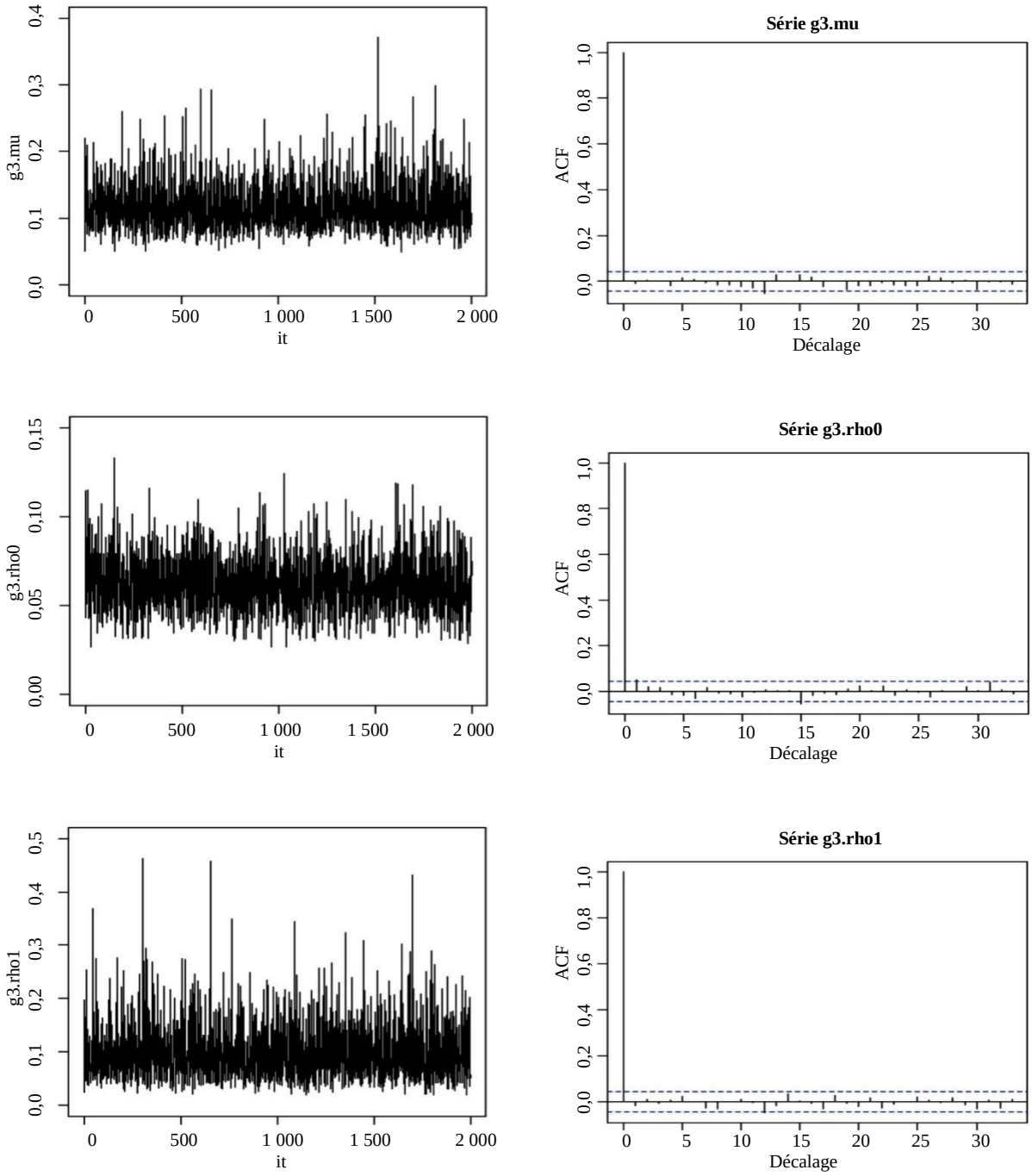
Domaine	a_{iy}	MP	ETP	CV	IC
LMW	a_0	0,000367	0,000034	0,092643	(0,000296; 0,000441)
	a_1	0,000307	0,000040	0,130293	(0,000225; 0,000400)
LMB	a_0	0,000283	0,000021	0,074205	(0,000241; 0,000332)
	a_1	0,000282	0,000020	0,070922	(0,000240; 0,000331)
LFW	a_0	0,000411	0,000042	0,102190	(0,000317; 0,000504)
	a_1	0,000391	0,000040	0,102302	(0,000307; 0,000489)
LFB	a_0	0,000294	0,000022	0,074830	(0,000247; 0,000342)
	a_1	0,000314	0,000023	0,073248	(0,000268; 0,000369)
MMW	a_0	0,000286	0,000017	0,059441	(0,000248; 0,000323)
	a_1	0,000264	0,000025	0,094697	(0,000209; 0,000327)
MMB	a_0	0,000262	0,000018	0,068702	(0,000223; 0,000304)
	a_1	0,000260	0,000018	0,069231	(0,000224; 0,000304)
MFW	a_0	0,000313	0,000023	0,073482	(0,000263; 0,000362)
	a_1	0,000247	0,000026	0,105263	(0,000194; 0,000312)
MFB	a_0	0,000289	0,000017	0,058824	(0,000253; 0,000332)
	a_1	0,000294	0,000022	0,074830	(0,000246; 0,000345)
HMW	a_0	0,000263	0,000016	0,060837	(0,000228; 0,000298)
	a_1	0,000192	0,000016	0,083333	(0,000157; 0,000234)
HMB	a_0	0,000250	0,000022	0,088000	(0,000201; 0,000299)
	a_1	0,000270	0,000027	0,100000	(0,000212; 0,000335)
HFW	a_0	0,000293	0,000021	0,071672	(0,000246; 0,000340)
	a_1	0,000324	0,000037	0,114198	(0,000248; 0,000410)
HFB	a_0	0,000278	0,000021	0,075540	(0,000234; 0,000324)
	a_1	0,000246	0,000024	0,097561	(0,000196; 0,000303)

Figure 4.1 Les densités a posteriori des proportions de population finie selon les trois modèles pour les données de LAS avec 12 domaines.



NOTE : Ici, la ligne pleine représente les densités de P selon le modèle de SI, la ligne pointillée représente les densités selon le modèle de SniHo, et la ligne en tirets représente celles selon le modèle de SniHe.

Figure 4.2 Tracé et graphique d'autocorrélation pour μ, ρ_0, ρ_1.



Nous présentons les densités a posteriori des proportions de population finie dans la figure 4.1. Ici, la ligne pleine représente la densité de P selon le modèle de SI, la ligne pointillée la représente selon le modèle de SniHo, et la ligne en tirets la représente selon le modèle de SniHe pour chaque domaine. Nous pouvons voir que les tracés du modèle de SI sont décalés en raison de l'effet du biais de sélection.

Tableau 4.8
Diagnostic de convergence de Geweke (valeur Z) et taille effective d'échantillon (TEE) pour μ, ρ_0, ρ_1

paramètre	Statistique de Geweke	Valeur p	TEE
μ	-0,4549	0,6492	2 000
ρ_0	1,7750	0,0759	1 809
ρ_1	-1,0331	0,3015	2 000

Notre exemple montre qu'il est important d'inclure une composante pour le mécanisme de sélection dans un modèle en cas d'échantillon biaisé. Sinon il y aura probablement des estimations faussées des proportions de petits domaines. Comme la distribution des probabilités de sélection de l'échantillon est un peu différente selon les domaines, l'hypothèse d'hétérogénéité est plus raisonnable pour notre exemple numérique.

4.2 Étude par simulations

Nous effectuons une étude par simulations pour évaluer l'exactitude de notre modèle. Plus précisément, nous examinons deux situations, qui sont des distributions homogènes ou hétérogènes des probabilités de sélection de l'échantillon pour montrer à quel point les modèles de SI, de SniHo et de SniHe peuvent différer. Nous montrons également qu'en cas de corrélation forte entre les réponses binaires et les probabilités de sélection, le modèle de SI peut avoir de mauvaises performances.

Pour effectuer une étude par simulation, nous générons $\ell = 12$ populations finies et la i^{e} population finie a $N_i = 1\,000$ unités et une probabilité de sélection π_{ij}, $j = 1, \ldots, 1\,000$, $i = 1, \ldots, 12$. Ensuite, des échantillons de taille $n_i = 50$ sont pris dans chaque domaine. Nous avons généré 100 ensembles de données. L'aperçu général de la collecte de données est décrit ci-après.

Étape 1. Générons $p_i \sim U(0{,}2;\ 0{,}7)$, $i = 1, \ldots, \ell$. Ces valeurs représentent des proportions vraies de petite population.

Étape 2. Posons $\tau = 100$, $f_i = n_i / N_i$, $a = 0{,}975$, $\mu_0 = af$, $\mu_1 = f/a$, $i = 1, \ldots, \ell$.

Étape 3. Générons $u \sim U(0, 1)$. Si $u \le p_i$ alors nous posons $y_{ij} = 1$; sinon, posons $y_{ij} = 0$, $j = 1, \ldots, N_i$, $i = 1, \ldots, \ell$.

Étape 4. Si $y_{ij} = 1$ alors nous générons $\pi_{ij} \sim \text{Bêta}(\mu_1\tau, (1-\mu_1)\tau)$; si $y_{ij} = 0$ alors nous générons $\pi_{ij} \sim \text{Bêta}(\mu_0\tau, (1-\mu_0)\tau)$ pour $j = 1, \ldots, N_i$, $i = 1, \ldots, \ell$.

Étape 5. Nous échantillonnons n_i unités par échantillonnage PPT systématique avec les probabilités $\frac{n_{ij}\pi_{ij}}{\sum_{i=1}^{N_i}\pi_{ij}}$.

Nous contrôlons la corrélation bisériale entre les réponses binaires et les probabilités de sélection en changeant les valeurs a à l'étape 2. Par exemple, si nous posons $a = 0{,}975$; 0,95; 0,9; 0,8; 0,7 dans le plan de simulation, alors les corrélations bisériales sont respectivement de $\rho = 0{,}05$; 0,1; 0,2; 0,4; 0,7. Nous générons des données ayant plusieurs corrélations bisériales, et les classons en trois niveaux (Faible : $\rho < 0{,}3$; Moyen : $0{,}3 \le \rho < 0{,}6$; Élevé : $\rho \ge 0{,}6$). Ils correspondent à un biais de sélection faible, moyen et fort.

Nous pouvons également contrôler les hypothèses d'hétérogénéité ou d'homogénéité en modifiant un support des probabilités de sélection de l'échantillon. Pour l'hypothèse d'homogénéité, nous générons π_{ij} de l'étape 4 à partir de supports similaires pour tous les domaines. D'autre part, nous avons établi un support de π_{ij} à partir d'un intervalle différent par domaine pour effectuer des distributions hétérogènes des probabilités de sélection de l'échantillon.

Pour comparer les performances des trois modèles, nous calculons plusieurs mesures fréquentistes. Tout d'abord, nous calculons la proportion de population finie $P_i^{(h)}$, la moyenne a posteriori $\mathrm{MP}_i^{(h)}$ et l'écart-type a posteriori $\mathrm{ETP}_i^{(h)}$, $i = 1, \ldots, \ell$, $h = 1, \ldots, 100$. Ensuite, nous calculons le biais absolu $\mathrm{BA}_i^{(h)} = \left| \mathrm{MP}_i^{(h)} - P_i^{(h)} \right|$ et la racine de l'erreur quadratique moyenne $\mathrm{REQM}_i^{(h)} = \sqrt{\mathrm{ETP}_i^{(h)2} + \mathrm{BA}_i^{(h)2}}$, $i = 1, \ldots, \ell$, $h = 1, \ldots, 100$. Au moyen de ces quantités fréquentistes, nous obtenons $\mathrm{BA}_i = \frac{1}{100}\sum_{h=1}^{100} \mathrm{BA}_i^{(h)}$ et $\mathrm{REQM}_i = \frac{1}{100}\sum_{h=1}^{100} \mathrm{REQM}_i^{(h)}$. Nous calculons également l'intervalle de la plus haute densité a posteriori (HDP) à 95 % pour chacune des 100 exécutions simulées dans chaque domaine. Ensuite, nous examinons la largeur $\left(W_i^{(h)}\right)$ et l'incidence de la HDP $\left(I_i^{(h)}\right)$. Soit $I_i^{(h)} = 1$ si l'intervalle de HDP à 95 % contient la valeur varie P_i et soit $I_i^{(h)} = 0$ sinon. Nous calculons ensuite la couverture $C_i = \sum_{h=1}^{100} I_i^{(h)} / 100$ et $W_i = \sum_{h=1}^{100} W_i^{(h)} / 100$.

La corrélation bisériale (c'est-à-dire le coefficient de corrélation de Pearson) ρ entre $\mathbf{y}$ et $\boldsymbol{\pi}$ est calculée puisque la variable $\mathbf{y}$ est binaire (voir par exemple Cox, 1974). Les valeurs sommaires fondées sur le BA, la REQM, la couverture et la largeur des intervalles de HDP à 95 % pour trois cas de corrélation selon l'hypothèse d'homogénéité sont présentées dans le tableau 4.9. En vertu de cette hypothèse, les performances du modèle de SniHo sont meilleures que celles du modèle de SniHe dans certains domaines, mais la différence est très faible. Les deux modèles sont meilleurs que le modèle de SI en ce sens qu'ils s'approchent plus de la vraie valeur P_i. En particulier, nous constatons qu'à mesure que la corrélation entre $\mathbf{y}$ et $\boldsymbol{\pi}$ augmente, il y a un plus grand écart entre le modèle de SI et les autres.

Le tableau 4.10 présente les valeurs sommaires de trois cas de corrélation selon l'hypothèse d'hétérogénéité. À partir du tableau, nous constatons que le modèle de SniHe se comporte bien, car il est plus près de la vraie valeur P_i que les autres modèles quand l'effet du biais de sélection est modéré ou fort. Il présente un biais plus petit, une erreur quadratique moyenne plus petite et une meilleure couverture.

À mesure que la corrélation bisériale augmente, il y a des disparités accrues entre le modèle de SI et le modèle de SniHe. Nous présentons quelques graphiques des distributions a posteriori des proportions de population finie pour des cas de corrélation faible, moyenne et élevée selon l'hypothèse d'homogénéité (figures 4.3 à 4.5). Il faut souligner que de grandes différences apparaissent dans les distributions des moyennes de la population finie à mesure que ρ augmente. Il n'y a pas de différences dans les densités a posteriori du modèle de SniHo et du modèle de SniHe selon l'hypothèse d'homogénéité. Les distributions a posteriori selon le modèle de SniHe s'écartent de la droite de la distribution dans le modèle de SI, leurs dispersions étant presque inchangées dans tous les domaines.

De même, nous présentons quelques graphiques des distributions a posteriori des proportions de population finie pour des cas de corrélation faible, moyenne et élevée selon l'hypothèse d'hétérogénéité (figures 4.6 à 4.8). Les densités a posteriori selon le modèle de SniHe sont différentes des autres pour tous les domaines.

De plus, nous avons sélectionné deux valeurs de $\ell = 12,\ 24,$ le nombre de domaines, pour observer les variations quand ℓ augmente. Nous calculons le BA, la REQM, la couverture et la largeur des intervalles de HDP à 95 % pour chaque domaine, puis nous établissons la moyenne de ces valeurs sur tous les domaines. Nous présentons les résultats selon les hypothèses d'homogénéité et d'hétérogénéité respectivement dans les tableaux 4.11 et 4.12. Les résultats convergent quand ℓ augmente.

L'étude par simulations montre qu'à mesure que la corrélation bisériale entre les réponses binaires et les probabilités de sélection augmente, il y a un plus grand écart entre le modèle de SI, le modèle de SniHo et le modèle de SniHe. Ainsi, plus la corrélation est forte, plus l'effet du biais de sélection est important. Le modèle de SniHe est meilleur que les autres modèles quand chaque domaine a une distribution des probabilités de sélection de l'échantillon différente selon un biais de sélection modéré ou fort.

Tableau 4.9
Comparaisons des estimations fondées sur le biais absolu (BA), la racine de l'erreur quadratique moyenne (REQM) a posteriori, la couverture (C) et la largeur (W pour *width*) des intervalles de HDP à 95 % pour les cas de corrélation faible (L pour *low*), moyenne (M) et élevée (H pour *high*) selon l'hypothèse d'homogénéité

domaine	ρ	BA			REQM			W			C		
		SI	SniHo	SniHe	SI	SniHo	SniHe	SI	SniHo	SniHe	SI	SniHo	SniHe
1	L	0,061	0,044	0,051	0,087	0,076	0,087	0,241	0,248	0,281	0,90	0,98	0,98
	M	0,134	0,046	0,057	0,149	0,075	0,092	0,259	0,238	0,293	0,50	0,98	0,97
	H	0,188	0,047	0,050	0,198	0,075	0,089	0,268	0,239	0,296	0,16	0,99	0,99
2	L	0,050	0,053	0,051	0,081	0,087	0,093	0,257	0,277	0,315	0,98	0,98	0,98
	M	0,103	0,043	0,042	0,122	0,081	0,091	0,260	0,279	0,331	0,72	0,96	1,0
	H	0,166	0,045	0,051	0,177	0,084	0,100	0,254	0,288	0,350	0,24	0,99	1,0
3	L	0,051	0,065	0,063	0,083	0,098	0,102	0,262	0,286	0,320	0,97	0,90	0,98
	M	0,084	0,064	0,058	0,107	0,099	0,103	0,258	0,295	0,342	0,81	0,95	0,98
	H	0,132	0,056	0,059	0,146	0,094	0,108	0,250	0,304	0,361	0,45	0,97	0,97
4	L	0,058	0,084	0,085	0,091	0,114	0,120	0,269	0,297	0,333	0,91	0,88	0,89
	M	0,052	0,078	0,070	0,083	0,110	0,113	0,258	0,311	0,355	0,97	0,93	0,98
	H	0,112	0,066	0,072	0,128	0,105	0,118	0,243	0,325	0,372	0,57	0,97	0,98
5	L	0,039	0,037	0,041	0,073	0,074	0,083	0,250	0,263	0,296	0,99	0,99	1,0
	M	0,118	0,037	0,043	0,135	0,074	0,088	0,261	0,261	0,315	0,64	0,98	1,0
	H	0,193	0,050	0,055	0,202	0,083	0,100	0,261	0,269	0,336	0,14	0,99	0,99
6	L	0,052	0,040	0,046	0,081	0,075	0,085	0,245	0,256	0,289	0,96	0,96	0,99
	M	0,125	0,041	0,049	0,142	0,073	0,089	0,260	0,249	0,301	0,52	0,98	0,98
	H	0,186	0,048	0,056	0,197	0,078	0,096	0,263	0,253	0,316	0,18	0,99	0,98
7	L	0,041	0,053	0,058	0,076	0,088	0,098	0,261	0,281	0,316	0,97	0,95	0,96
	M	0,096	0,049	0,045	0,116	0,087	0,095	0,259	0,289	0,340	0,75	0,94	0,99
	H	0,145	0,046	0,050	0,158	0,086	0,102	0,254	0,293	0,356	0,36	0,96	0,98
8	L	0,044	0,039	0,045	0,076	0,075	0,085	0,249	0,261	0,295	0,99	0,99	1,0
	M	0,118	0,039	0,041	0,135	0,074	0,086	0,261	0,258	0,309	0,62	0,99	0,99
	H	0,184	0,049	0,058	0,195	0,081	0,101	0,262	0,262	0,329	0,24	0,98	0,97
9	L	0,047	0,043	0,046	0,077	0,078	0,088	0,253	0,268	0,301	0,98	0,98	0,99
	M	0,116	0,041	0,042	0,133	0,076	0,089	0,260	0,265	0,322	0,64	1,0	1,0
	H	0,184	0,050	0,059	0,194	0,084	0,103	0,258	0,274	0,338	0,21	0,98	0,97
10	L	0,062	0,046	0,059	0,088	0,078	0,095	0,246	0,257	0,296	0,94	0,98	0,97
	M	0,135	0,041	0,050	0,149	0,073	0,090	0,260	0,245	0,300	0,48	0,99	0,99
	H	0,191	0,055	0,058	0,202	0,082	0,097	0,264	0,247	0,310	0,17	0,99	1,0
11	L	0,053	0,070	0,074	0,085	0,103	0,112	0,265	0,292	0,327	0,96	0,92	0,95
	M	0,070	0,065	0,053	0,096	0,100	0,101	0,258	0,303	0,348	0,88	0,95	0,99
	H	0,131	0,058	0,058	0,145	0,098	0,109	0,245	0,313	0,368	0,46	0,94	0,97
12	L	0,047	0,043	0,043	0,077	0,077	0,086	0,250	0,264	0,306	0,97	0,99	0,99
	M	0,122	0,047	0,048	0,139	0,080	0,091	0,260	0,263	0,317	0,56	0,97	0,99
	H	0,173	0,044	0,050	0,184	0,078	0,095	0,261	0,266	0,333	0,27	0,99	0,99

Tableau 4.10
Comparaisons des estimations fondées sur le biais absolu (BA), la racine de l'erreur quadratique moyenne (REQM) a posteriori, la couverture (C) et la largeur (W pour *width*) des intervalles de HDP à 95 % pour les cas de corrélation faible (L pour *low*), moyenne (M) et élevée (H pour *high*) selon l'hypothèse d'hétérogénéité

domaine	ρ	BA			REQM			W			C		
		SI	SniHo	SniHe	SI	SniHo	SniHe	SI	SniHo	SniHe	SI	SniHo	SniHe
1	L	0,050	0,052	0,047	0,078	0,085	0,084	0,215	0,235	0,251	0,91	0,92	0,94
	M	0,106	0,101	0,041	0,125	0,123	0,084	0,229	0,248	0,268	0,55	0,67	0,99
	H	0,163	0,145	0,056	0,176	0,162	0,109	0,242	0,261	0,343	0,16	0,39	0,99
2	L	0,045	0,047	0,050	0,079	0,085	0,094	0,231	0,252	0,287	0,95	0,97	0,99
	M	0,087	0,092	0,049	0,109	0,116	0,105	0,231	0,250	0,345	0,75	0,78	1,0
	H	0,161	0,174	0,090	0,173	0,187	0,145	0,221	0,235	0,382	0,26	0,23	0,89
3	L	0,056	0,058	0,068	0,088	0,094	0,108	0,235	0,259	0,287	0,91	0,92	0,91
	M	0,085	0,080	0,062	0,111	0,111	0,114	0,239	0,264	0,345	0,77	0,86	0,97
	H	0,151	0,133	0,060	0,165	0,152	0,129	0,233	0,261	0,405	0,32	0,55	0,97
4	L	0,056	0,059	0,056	0,088	0,095	0,099	0,240	0,263	0,290	0,92	0,93	0,96
	M	0,117	0,123	0,053	0,135	0,144	0,108	0,242	0,265	0,341	0,57	0,59	0,99
	H	0,175	0,194	0,081	0,188	0,207	0,140	0,245	0,267	0,404	0,23	0,20	0,96
5	L	0,047	0,049	0,047	0,078	0,084	0,089	0,224	0,245	0,273	0,98	0,97	0,99
	M	0,105	0,098	0,052	0,126	0,124	0,102	0,239	0,263	0,313	0,60	0,71	0,98
	H	0,164	0,144	0,071	0,178	0,163	0,130	0,246	0,271	0,399	0,28	0,46	1,0
6	L	0,045	0,051	0,059	0,076	0,084	0,097	0,218	0,238	0,273	0,95	0,95	0,95
	M	0,060	0,063	0,058	0,086	0,091	0,109	0,211	0,227	0,331	0,87	0,90	0,97
	H	0,103	0,113	0,055	0,118	0,128	0,112	0,195	0,203	0,353	0,51	0,53	0,99
7	L	0,043	0,046	0,056	0,079	0,085	0,098	0,234	0,257	0,284	0,97	0,98	0,97
	M	0,093	0,086	0,051	0,117	0,116	0,107	0,241	0,265	0,337	0,73	0,81	0,98
	H	0,152	0,134	0,065	0,168	0,155	0,129	0,240	0,266	0,400	0,31	0,53	0,98
8	L	0,045	0,048	0,060	0,077	0,083	0,098	0,223	0,244	0,277	0,96	0,99	0,96
	M	0,066	0,069	0,053	0,092	0,097	0,109	0,217	0,235	0,338	0,87	0,90	0,98
	H	0,103	0,115	0,064	0,121	0,132	0,123	0,207	0,217	0,375	0,56	0,48	0,99
9	L	0,048	0,050	0,045	0,080	0,086	0,088	0,227	0,249	0,275	0,98	0,97	0,99
	M	0,103	0,096	0,052	0,125	0,123	0,101	0,241	0,264	0,312	0,63	0,70	0,99
	H	0,166	0,146	0,057	0,180	0,166	0,120	0,245	0,272	0,382	0,24	0,42	0,99
10	L	0,046	0,047	0,057	0,076	0,080	0,095	0,214	0,235	0,272	0,92	0,96	0,95
	M	0,051	0,053	0,062	0,078	0,082	0,110	0,207	0,222	0,327	0,93	0,94	0,97
	H	0,093	0,102	0,054	0,108	0,117	0,110	0,191	0,198	0,348	0,61	0,60	1,0
11	L	0,057	0,058	0,071	0,088	0,094	0,110	0,237	0,260	0,293	0,91	0,95	0,92
	M	0,075	0,068	0,059	0,102	0,102	0,112	0,238	0,260	0,344	0,79	0,86	0,99
	H	0,136	0,120	0,067	0,152	0,142	0,128	0,229	0,257	0,398	0,43	0,58	0,99
12	L	0,044	0,046	0,051	0,076	0,082	0,092	0,222	0,243	0,277	0,98	0,96	1,0
	M	0,072	0,076	0,050	0,097	0,103	0,105	0,219	0,236	0,337	0,79	0,82	1,0
	H	0,115	0,127	0,064	0,131	0,143	0,123	0,211	0,222	0,369	0,51	0,45	0,98

Tableau 4.11
Comparaison de trois modèles au moyen du biais absolu (BA), la racine de l'erreur quadratique moyenne (REQM) a posteriori, la couverture (C) et la largeur (W) des intervalles de HDP à 95 % selon l'hypothèse d'homogénéité

ρ	ℓ	Modèle	BA	REQM	W	C
Faible	12	SI	$0{,}0503_{0{,}0075}$	$0{,}0812_{0{,}0056}$	$0{,}2539_{0{,}0086}$	$0{,}9600_{0{,}0292}$
		SniHo	$0{,}0514_{0{,}0145}$	$0{,}0853_{0{,}0131}$	$0{,}2709_{0{,}0155}$	$0{,}9583_{0{,}0381}$
		SniHe	$0{,}0551_{0{,}0133}$	$0{,}0943_{0{,}0117}$	$0{,}3062_{0{,}0161}$	$0{,}9733_{0{,}0303}$
	24	SI	$0{,}0486_{0{,}0040}$	$0{,}0809_{0{,}0029}$	$0{,}2594_{0{,}0123}$	$0{,}9721_{0{,}0177}$
		SniHo	$0{,}0509_{0{,}0101}$	$0{,}0861_{0{,}0103}$	$0{,}2778_{0{,}0215}$	$0{,}9713_{0{,}0201}$
		SniHe	$0{,}0568_{0{,}0105}$	$0{,}0975_{0{,}0101}$	$0{,}3166_{0{,}0210}$	$0{,}9792_{0{,}0177}$

NOTE : La notation a_b signifie que a est la moyenne des domaines et que b est l'erreur-type.

Tableau 4.11(suite)
Comparaison de trois modèles au moyen du biais absolu (BA), la racine de l'erreur quadratique moyenne (REQM) a posteriori, la couverture (C) et la largeur (W) des intervalles de HDP à 95 % selon l'hypothèse d'homogénéité

ρ	ℓ	Modèle	BA	REQM	W	C
Moyenne	12	SI	$0{,}1060_{0{,}0261}$	$0{,}1254_{0{,}0212}$	$0{,}2594_{0{,}0011}$	$0{,}6742_{0{,}1556}$
		SniHo	$0{,}0491_{0{,}0127}$	$0{,}0835_{0{,}0126}$	$0{,}2714_{0{,}0237}$	$0{,}9683_{0{,}0221}$
		SniHe	$0{,}0498_{0{,}0086}$	$0{,}0939_{0{,}0078}$	$0{,}3227_{0{,}0203}$	$0{,}9883_{0{,}0094}$
	24	SI	$0{,}1010_{0{,}0232}$	$0{,}1212_{0{,}0198}$	$0{,}2592_{0{,}0104}$	$0{,}7025_{0{,}1267}$
		SniHo	$0{,}0502_{0{,}0092}$	$0{,}0861_{0{,}0110}$	$0{,}2812_{0{,}0295}$	$0{,}9775_{0{,}0159}$
		SniHe	$0{,}0521_{0{,}0080}$	$0{,}0964_{0{,}0093}$	$0{,}3279_{0{,}0281}$	$0{,}9904_{0{,}0127}$
Élevée	12	SI	$0{,}1653_{0{,}0281}$	$0{,}1771_{0{,}0263}$	$0{,}2568_{0{,}0078}$	$0{,}2875_{0{,}1399}$
		SniHo	$0{,}0510_{0{,}0064}$	$0{,}0857_{0{,}0088}$	$0{,}2776_{0{,}0270}$	$0{,}9783_{0{,}0159}$
		SniHe	$0{,}0562_{0{,}0062}$	$0{,}1014_{0{,}0075}$	$0{,}3389_{0{,}0237}$	$0{,}9825_{0{,}0114}$
	24	SI	$0{,}1582_{0{,}0316}$	$0{,}1704_{0{,}0300}$	$0{,}2540_{0{,}0170}$	$0{,}3242_{0{,}1618}$
		SniHo	$0{,}0522_{0{,}0046}$	$0{,}0886_{0{,}0090}$	$0{,}2882_{0{,}0335}$	$0{,}9788_{0{,}0099}$
		SniHe	$0{,}0570_{0{,}0039}$	$0{,}1025_{0{,}0069}$	$0{,}3412_{0{,}0302}$	$0{,}9875_{0{,}0126}$

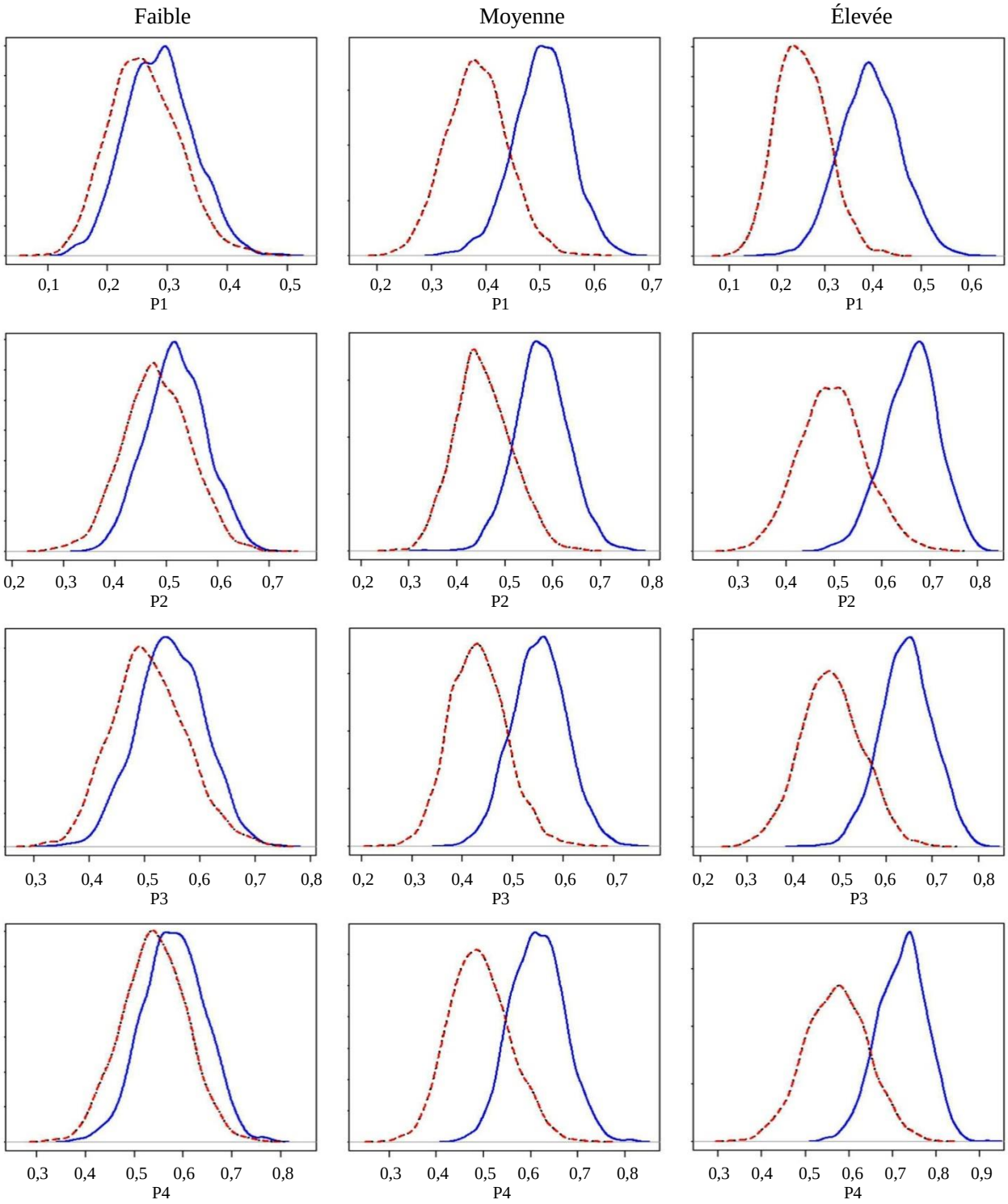
NOTE : La notation a_b signifie que a est la moyenne des domaines et que b est l'erreur-type.

Tableau 4.12
Comparaison de trois modèles au moyen du biais absolu (BA), la racine de l'erreur quadratique moyenne (REQM) a posteriori, la couverture (C) et la largeur (W) des intervalles de HDP à 95 % selon l'hypothèse d'hétérogénéité

ρ	ℓ	Modèle	BA	REQM	W	C
Faible	12	SI	$0{,}0501_{0{,}0050}$	$0{,}0842_{0{,}0051}$	$0{,}2703_{0{,}0109}$	$0{,}9733_{0{,}0156}$
		SniHo	$0{,}0509_{0{,}0049}$	$0{,}0862_{0{,}0049}$	$0{,}2800_{0{,}0112}$	$0{,}9758_{0{,}0168}$
		SniHe	$0{,}0556_{0{,}0083}$	$0{,}0959_{0{,}0075}$	$0{,}3132_{0{,}0127}$	$0{,}9767_{0{,}0227}$
	24	SI	$0{,}0457_{0{,}0054}$	$0{,}0768_{0{,}0054}$	$0{,}2478_{0{,}0135}$	$0{,}9675_{0{,}0205}$
		SniHo	$0{,}0470_{0{,}0055}$	$0{,}0818_{0{,}0053}$	$0{,}2703_{0{,}0146}$	$0{,}9779_{0{,}0169}$
		SniHe	$0{,}0537_{0{,}0093}$	$0{,}0922_{0{,}0082}$	$0{,}3018_{0{,}0167}$	$0{,}9792_{0{,}0193}$
Moyenne	12	SI	$0{,}0850_{0{,}0206}$	$0{,}1085_{0{,}0179}$	$0{,}2586_{0{,}0147}$	$0{,}8033_{0{,}1026}$
		SniHo	$0{,}0837_{0{,}0196}$	$0{,}1109_{0{,}0170}$	$0{,}2816_{0{,}0187}$	$0{,}8517_{0{,}0720}$
		SniHe	$0{,}0534_{0{,}0060}$	$0{,}1054_{0{,}0077}$	$0{,}3672_{0{,}0248}$	$0{,}9917_{0{,}0111}$
	24	SI	$0{,}0824_{0{,}0218}$	$0{,}1056_{0{,}0190}$	$0{,}2531_{0{,}0179}$	$0{,}7992_{0{,}1040}$
		SniHo	$0{,}0810_{0{,}0205}$	$0{,}1074_{0{,}0179}$	$0{,}2736_{0{,}0218}$	$0{,}8546_{0{,}0825}$
		SniHe	$0{,}0533_{0{,}0077}$	$0{,}1046_{0{,}0090}$	$0{,}3637_{0{,}0289}$	$0{,}9954_{0{,}0093}$
Élevée	12	SI	$0{,}1401_{0{,}0289}$	$0{,}1547_{0{,}0278}$	$0{,}2537_{0{,}0232}$	$0{,}4417_{0{,}1333}$
		SniHo	$0{,}1372_{0{,}0263}$	$0{,}1543_{0{,}0253}$	$0{,}2741_{0{,}0319}$	$0{,}5375_{0{,}1231}$
		SniHe	$0{,}0653_{0{,}0110}$	$0{,}1249_{0{,}0113}$	$0{,}4215_{0{,}0257}$	$0{,}9833_{0{,}0287}$
	24	SI	$0{,}1336_{0{,}0302}$	$0{,}1481_{0{,}0289}$	$0{,}2487_{0{,}0263}$	$0{,}4746_{0{,}1316}$
		SniHo	$0{,}1302_{0{,}0270}$	$0{,}1470_{0{,}0261}$	$0{,}2663_{0{,}0348}$	$0{,}5567_{0{,}1396}$
		SniHe	$0{,}0601_{0{,}0090}$	$0{,}1203_{0{,}0108}$	$0{,}4141_{0{,}0360}$	$0{,}9900_{0{,}0153}$

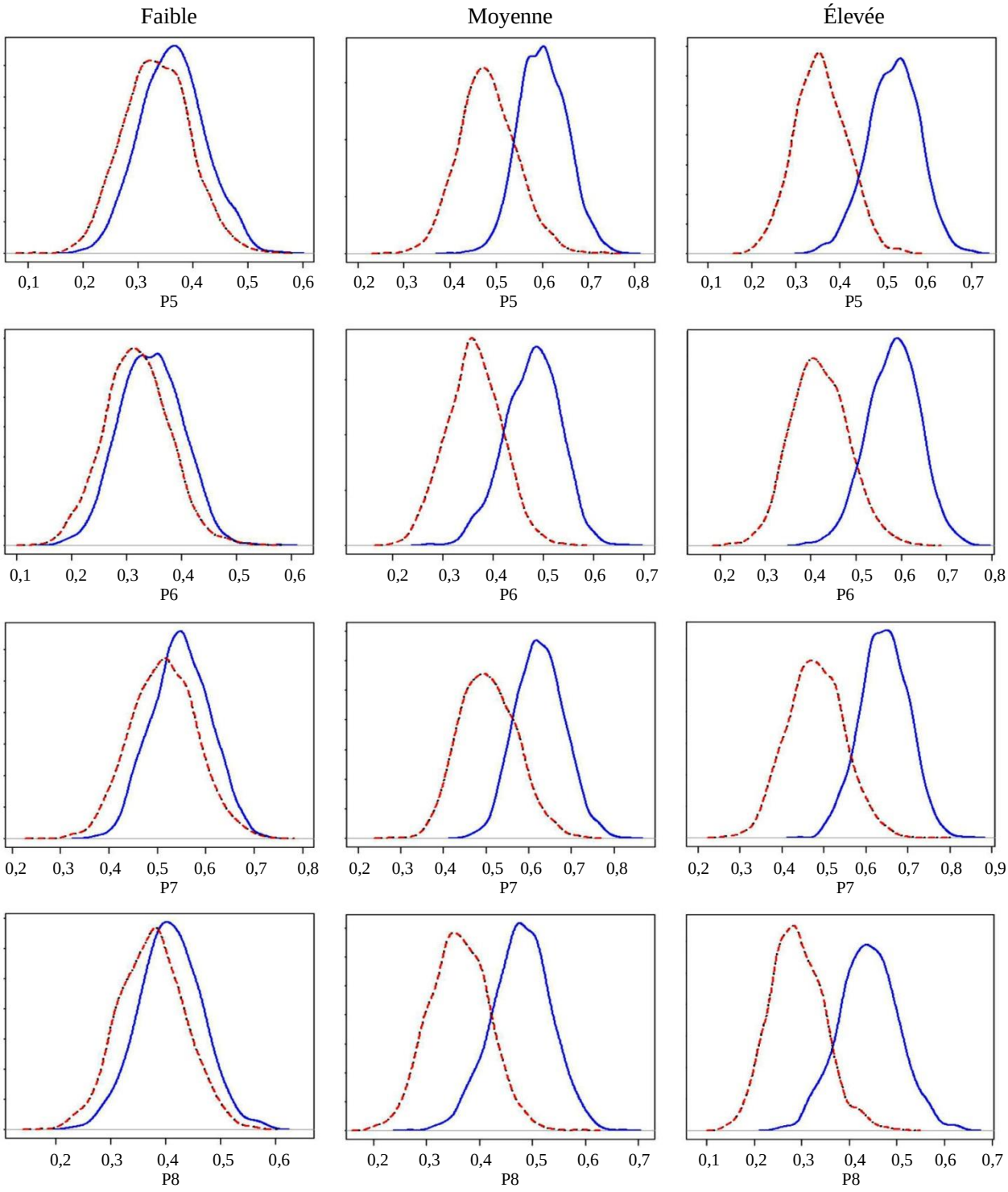
NOTE : La notation a_b signifie que a est la moyenne des domaines et que b est l'erreur-type.

Figure 4.3 Densités a posteriori des proportions de population finie pour les cas de corrélation faible, moyenne et élevée au moyen de données simulées selon l'hypothèse d'homogénéité.



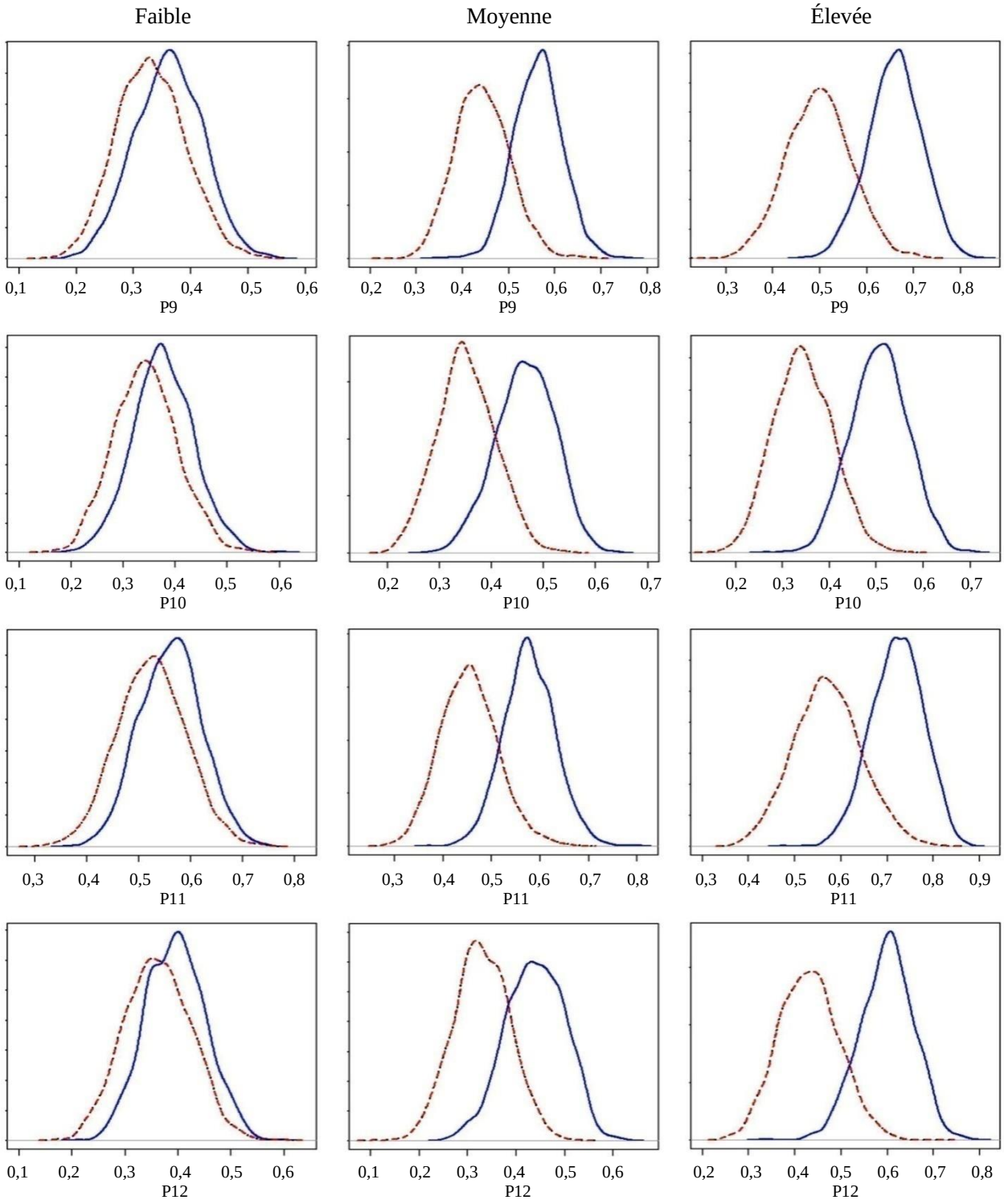
NOTE : Ici, la ligne pleine représente les densités de P selon le modèle de SI, la ligne pointillée représente les densités selon le modèle de SniHo, et la ligne en tirets représente celles selon le modèle de SniHe. (La ligne pointillée et la ligne en tirets coïncident.)

Figure 4.4 Densités a posteriori des proportions de population finie pour les cas de corrélation faible, moyenne et élevée au moyen de données simulées selon l'hypothèse d'homogénéité (suite).



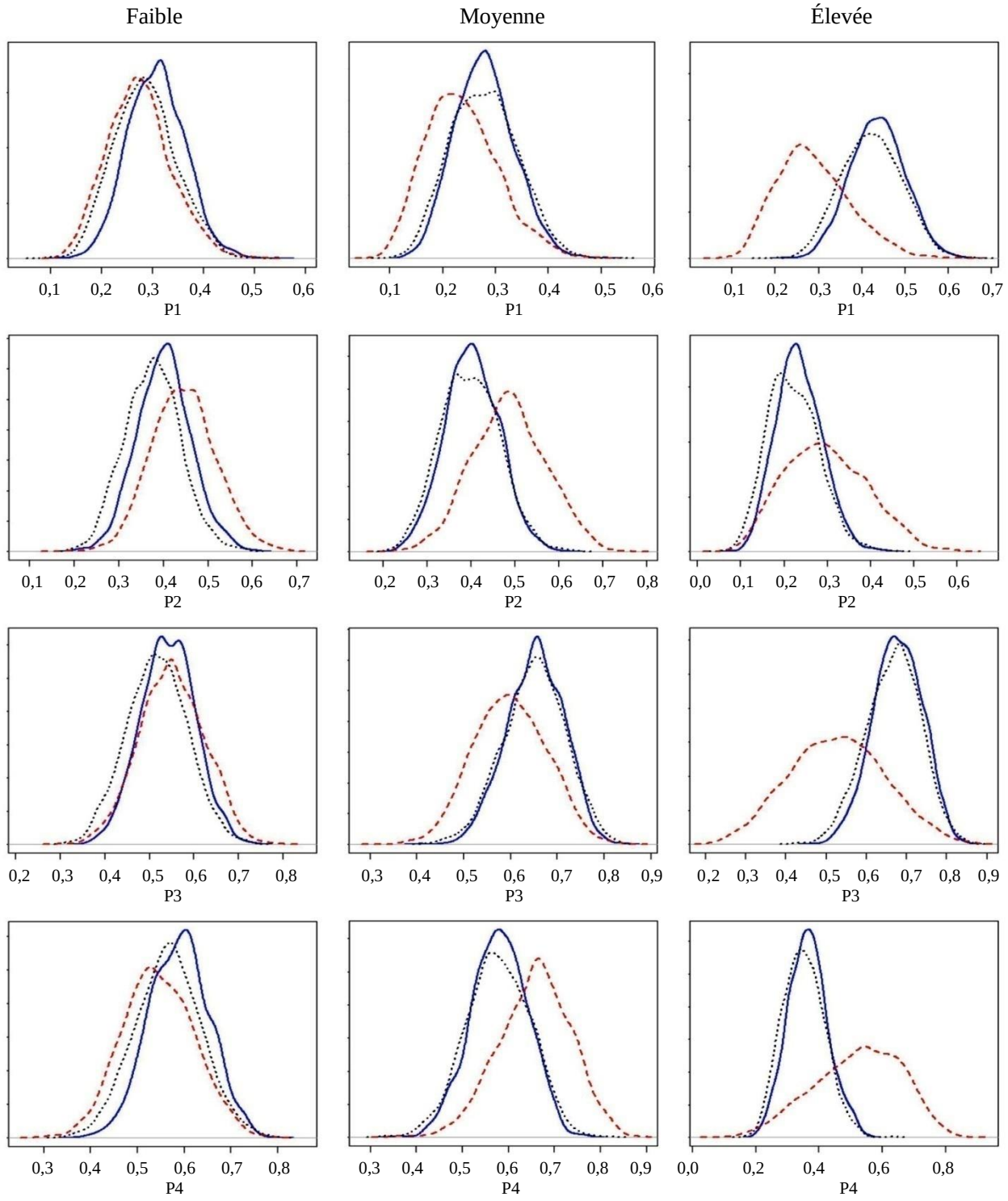
NOTE : Ici, la ligne pleine représente les densités de P selon le modèle de SI, la ligne pointillée représente les densités selon le modèle de SniHo, et la ligne en tirets représente celles selon le modèle de SniHe. (La ligne pointillée et la ligne en tirets coïncident.)

Figure 4.5 Densités a posteriori des proportions de population finie pour les cas de corrélation faible, moyenne et élevée au moyen de données simulées selon l'hypothèse d'homogénéité (suite).



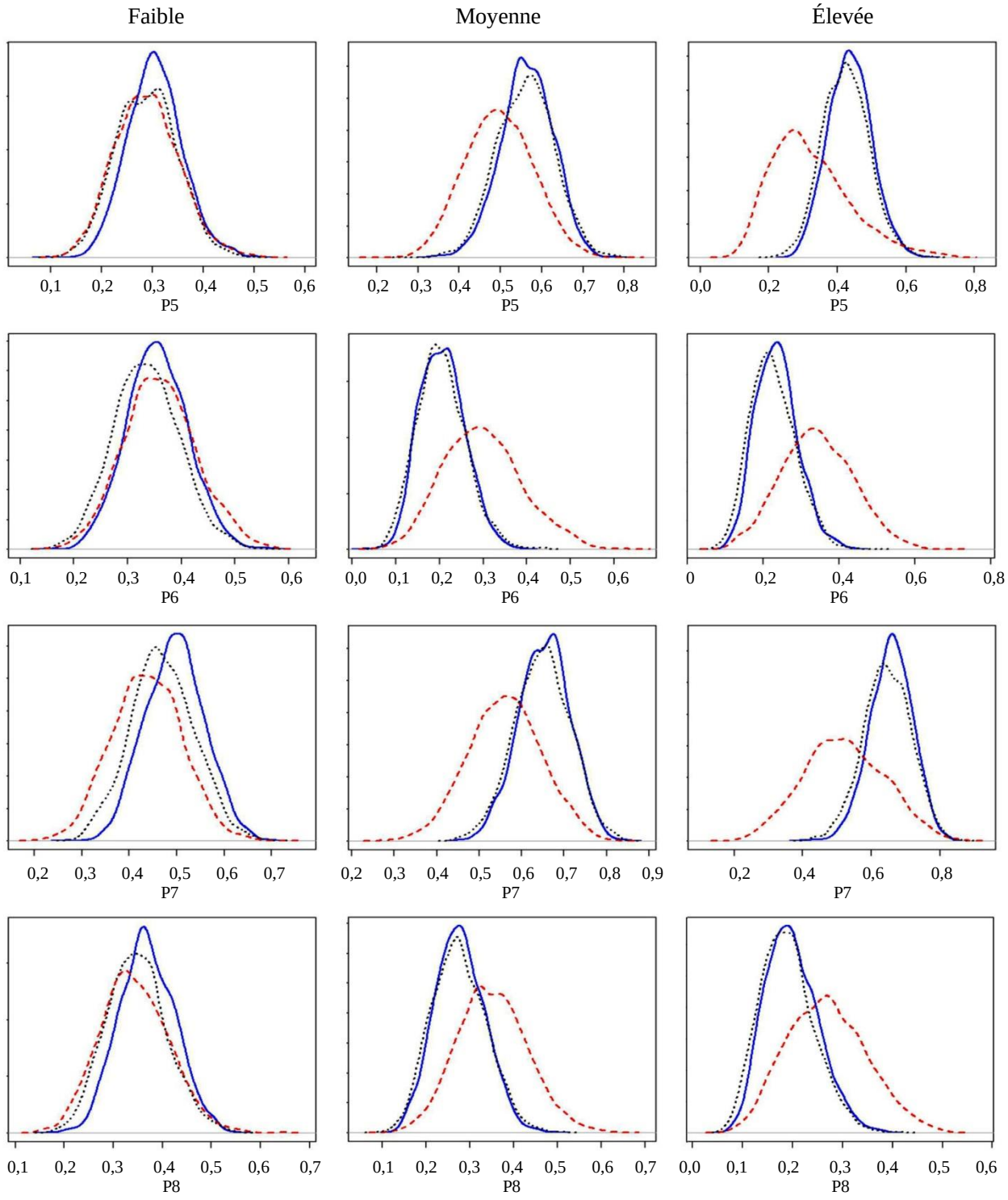
NOTE : Ici, la ligne pleine représente les densités de P selon le modèle de SI, la ligne pointillée représente les densités selon le modèle de SniHo, et la ligne en tirets représente celles selon le modèle de SniHe. (La ligne pointillée et la ligne en tirets coïncident.)

Figure 4.6 Densités a posteriori des proportions de population finie pour les cas de corrélation faible, moyenne et élevée au moyen de données simulées selon l'hypothèse d'hétérogénéité.



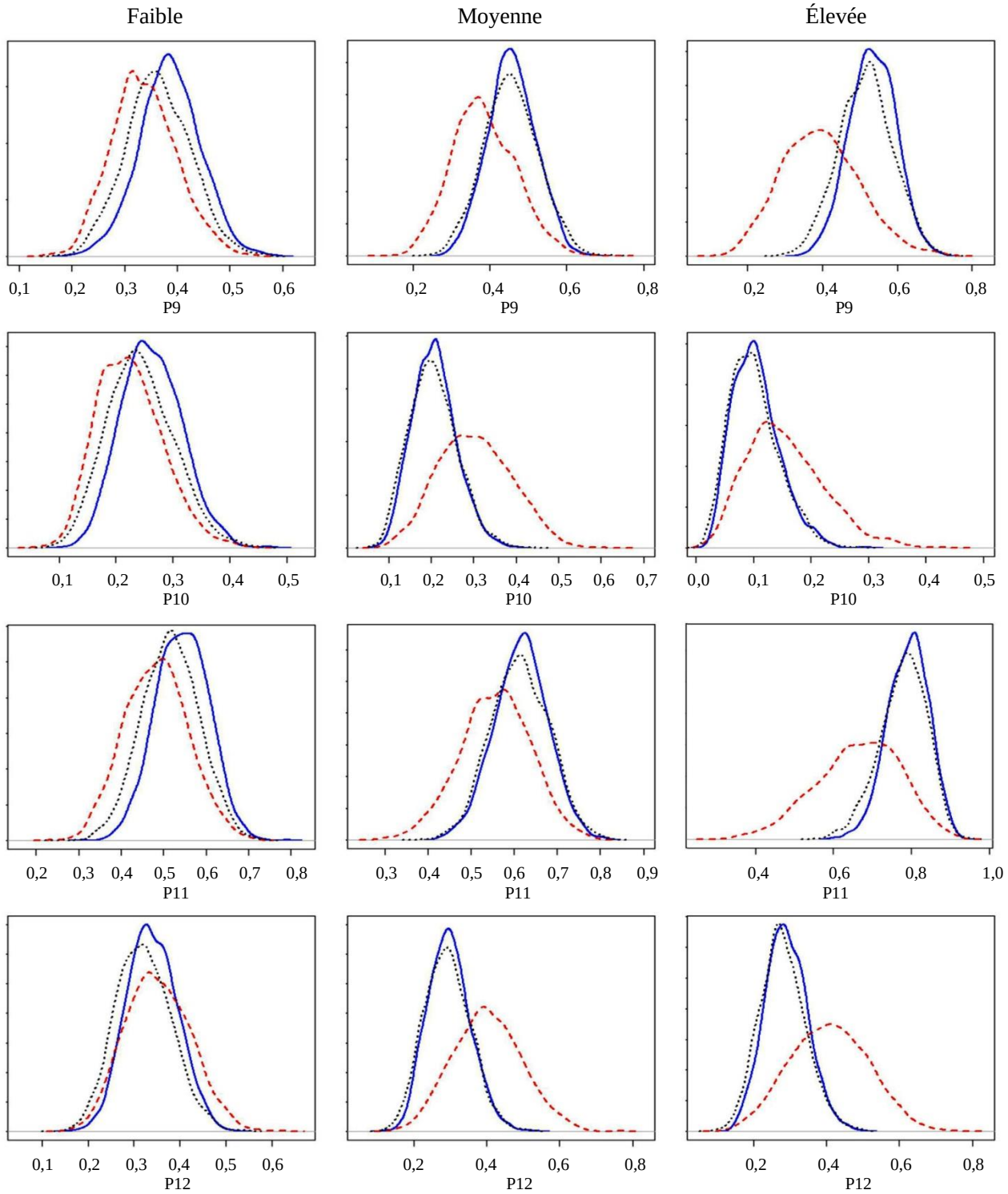
NOTE : Ici, la ligne pleine représente les densités de P selon le modèle de SI, la ligne pointillée représente les densités selon le modèle de SniHo, et la ligne en tirets représente celles selon le modèle de SniHe.

Figure 4.7 Densités a posteriori des proportions de population finie pour les cas de corrélation faible, moyenne et élevée au moyen de données simulées selon l'hypothèse d'hétérogénéité (suite).



NOTE : Ici, la ligne pleine représente les densités de P selon le modèle de SI, la ligne pointillée représente les densités selon le modèle de SniHo, et la ligne en tirets représente celles selon le modèle de SniHe.

Figure 4.8 Densités a posteriori des proportions de population finie pour les cas de corrélation faible, moyenne et élevée au moyen de données simulées selon l'hypothèse d'hétérogénéité (suite).



NOTE : Ici, la ligne pleine représente les densités de P selon le modèle de SI, la ligne pointillée représente les densités selon le modèle de SniHo, et la ligne en tirets représente celles selon le modèle de SniHe.

5. Conclusion

Nous avons étendu le modèle de sélection non ignorable homogène de CNK pour tenir compte de probabilités de sélection ayant des distributions différentes dans différents domaines. Nous l'avons ainsi amélioré afin de traiter la relation entre les variables binaires et les probabilités de sélection. En raison des nombreux paramètres supplémentaires, le calcul était beaucoup plus difficile, c'est pourquoi nous avons utilisé l'échantillonneur de Metropolis-Hastings pour surmonter cette difficulté.

Nous avons utilisé l'exemple de la limitation d'activité sévère dans la *National Health Interview Survey* (Enquête nationale sur la santé réalisée par interviews aux États-Unis) dans laquelle des petits domaines sont formées par le croisement du niveau d'études (pré-collégial, collégial, post-collégial), du sexe (masculin, féminin) et de la race (blanc, non blanc). Il semblerait que le modèle de sélection non ignorable hétérogène donne de meilleurs résultats que le modèle de sélection homogène et le modèle de sélection ignorable de référence.

Nous avons employé une étude par simulations pour évaluer les performances de notre modèle de sélection non ignorable hétérogène. Nous avons tiré des données du modèle de sélection non ignorable homogène et ajusté le modèle de sélection ignorable et le modèle de sélection non ignorable hétérogène. Nous avons constaté peu de différence entre les deux modèles de sélection non ignorable, mais une différence importante par rapport au modèle de sélection ignorable. Cependant, quand nous avons tiré les données à partir du modèle de sélection non ignorable hétérogène et ajusté le modèle de sélection ignorable et le modèle de sélection non ignorable homogène, nous avons constaté que le modèle de sélection non ignorable hétérogène était plus performant, particulièrement quand le biais de sélection est moyen ou fort, au moyen du biais, de l'erreur quadratique moyenne et de la couverture. Cette constatation apparaît manifestement dans les tableaux 4.9 à 4.12 et les figures 4.3 à 4.8.

Dans le cadre de notre modèle de sélection non ignorable hétérogène, nous pouvons penser à plusieurs autres problèmes. Premièrement, nous pouvons tenir compte des données polychotomiques, qui sont nombreuses dans les problèmes d'enquête. Deuxièmement, bien que notre approche soit un peu différente, nous pouvons tenir compte des covariables (par exemple âge, race, sexe dans notre application sur la limitation d'activité). Troisièmement, comme de nombreuses enquêtes comportent des non-répondants, nous pouvons essayer de tenir compte de la non-réponse non ignorable et de la sélection simultanément (par exemple Nandram et Choi, 2010). Cela peut s'effectuer dans notre cadre. Quatrièmement, comme dans un modèle double, nous pouvons tenir compte de la mise en grappes ou de la stratification dans chaque domaine (par exemple Nandram, 2016; Lee, Nandram et Kim, 2017). Notre travail peut également servir à résoudre des problèmes dans les échantillons non probabilistes (par exemple Elliot et Valliant, 2017).

Remerciements

Cette recherche a été financée par une subvention de la *Simons Foundation* (nº 353953, Balgobin Nandram).

Bibliographie

Chambers, R., Dorfman, A. et Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society, Series B*, 60, 397-411.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2010). Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 36, 1, 25-37. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2010001/article/11250-fra.pdf.

Choi, S., Nandram, B. et Kim, D. (2017). A hierarchical Bayesian model for binary data incorporating selection bias. *Communications in Statistics - Simulation and Computation*, 46, 6, 4767-4782.

Cox, N.R. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics*, 30, 171-178.

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

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Lee, D., Nandram, B. et Kim, D. (2017). Inférence bayésienne prédictive sur une proportion sous un modèle double pour petits domaines avec corrélations hétérogènes. *Techniques d'enquête*, 43, 1, 77-102. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2017001/article/14822-fra.pdf.

Ma, J., Sedransk, J., Nandram, B. et Chen, L. (2018). Bayesian predictive inference for finite population quantities under informative sampling. *Statistics and Application*, 16, 207-226.

Malec, D., Davis, W. et Cao, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine*, 18, 3189-3200.

Nandram, B. (2007). Bayesian predictive inference under informative sampling via surrogate samples. *Bayesian Statistics and its Applications*, (Éds. S.K. Upadhyay, U. Singh et D. Dey), Anamaya, New Delhi, Chapitre 25, 356-374.

Nandram, B. (2016). Bayesian predictive inference of a proportion under a two-fold small area model. *Journal of Official Statistics*, 32, 1, 187-208.

Nandram, B., et Choi, J.W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, 105, 120-135.

Nandram, B., Bai, Y. et Choi, J.W. (2011). Hierarchical Bayesian models for assessing possible changes in prevalence of activity limitation. *Advances and Applications in Statistical Sciences*, 6, 285-311.

Nandram, B., Bhatta, D., Bhadra, D. et Shen G. (2013). Bayesian predictive inference of a finite population proportion under selection bias. *Statistical Methodology*, 11, 1-21.

Nandram, B., Choi, J.W., Shen, G. et Burgos, C. (2006). Bayesian predictive inference under informative sampling and transformation. *Applied Stochastic Models in Business and Industry*, 22, 559-572.

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

Pfeffermann, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. *Journal of the American Statistical Association*, 83, 824-833.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

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

Pfeffermann, D., Krieger, A.M. et Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Si, Y., Pillai, N.S. et Gelman, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10, 3, 605-625.

Sverchkov, M., et Pfeffermann, D. (2004). Prévision des totaux de population finie basée sur la distribution échantillonnale. *Techniques d'enquête*, 30, 1, 87-101. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2004001/article/6996-fra.pdf.

Zangeneh, S.Z., et Little, R.J.A. (2015). Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology*, 3, 162-192.

Techniques d'enquête, juin 2021
Vol. 47, nº 1, pp. 133-162
Statistique Canada, nº 12-001-X au catalogue

Estimation sur petits domaines réconciliée sous le modèle de base au niveau de l'unité lorsque les taux d'échantillonnage sont non négligeables

Marius Stefan et Michael A. Hidiroglou[1]

Résumé

Nous examinons l'estimation d'une moyenne sur petits domaines sous le modèle de base au niveau de l'unité. La somme des estimateurs dépendant d'un modèle qui en résultent peut ne pas correspondre aux estimations obtenues au moyen d'un estimateur d'enquête direct qui est considéré comme précis pour l'ensemble de ces petits domaines. La réconciliation force la concordance des estimateurs fondés sur un modèle avec l'estimateur direct au niveau du domaine agrégé. L'estimateur par la régression généralisée est l'estimateur direct que nous utilisons pour réaliser la réconciliation. Dans le présent document, nous comparons des estimateurs sur petits domaines réconciliés d'après quatre procédures. La première procédure permet d'obtenir des estimateurs réconciliés au moyen d'un ajustement par le ratio. La deuxième procédure repose sur le meilleur estimateur linéaire sans biais empirique obtenu sous le modèle au niveau de l'unité augmenté à l'aide d'une variable adéquate qui assure la réconciliation. La troisième procédure utilise des estimateurs pseudo-empiriques construits au moyen de poids de sondage convenablement choisis de sorte que, une fois agrégés, ils concordent avec l'estimateur direct fiable pour le plus grand domaine. La quatrième procédure permet d'obtenir des estimateurs réconciliés qui résultent d'un problème de minimisation sous la contrainte donnée par la condition de réconciliation. Ces procédures de réconciliation sont appliquées aux estimateurs sur petits domaines lorsque les taux d'échantillonnage sont non négligeables. Les estimateurs réconciliés qui en résultent sont comparés quant au biais relatif et à l'erreur quadratique moyenne dans une étude par simulations fondée sur un plan de sondage ainsi qu'un exemple fondé sur des données d'enquête réelles.

Mots-clés : Petit domaine; réconciliation; estimateur empirique; estimateur pseudo-empirique; estimateur contraint.

1. Introduction

L'estimation sur petits domaines (EPD) a gagné en importance au cours des dernières années en raison de la demande pour des statistiques fiables sur les petits domaines. Des estimateurs directs sont utilisés pour estimer des paramètres d'intérêt lorsque la taille de l'échantillon est suffisamment grande. Cependant, ils présentent des erreurs-types et des coefficients de variation importants lorsque vient le temps de les appliquer aux petits domaines, car la taille d'échantillon réalisée sera assez limitée. Il est par conséquent nécessaire d'utiliser des modèles qui empruntent de l'information à d'autres domaines adjacents ou à des enquêtes précédentes pour obtenir des estimateurs stables pour ces petits domaines. Les estimations fondées sur un modèle présentent généralement une amélioration considérable par rapport aux estimations directes quant à l'erreur quadratique moyenne (EQM).

La théorie disponible pour l'estimation sur petits domaines repose sur des modèles au niveau du domaine ou au niveau de l'unité, selon le niveau des données auxiliaires disponibles. Les méthodes fondées sur des modèles au niveau de l'unité utilisent les données de chacune des unités comme données auxiliaires, tandis que les méthodes fondées sur des modèles au niveau du domaine utilisent des agrégats ou des moyennes des données sur les unités dans les petits domaines. Le modèle de Fay et Herriot (1979),

1. Marius Stefan, Faculty of Applied Sciences, Polytechnic University of Bucharest, Splaiul Independentei, nr. 313. Courriel : mastefan@gmail.com; Michael A. Hidiroglou, ancien employé de Statistique Canada. Courriel : hidirog@yahoo.ca.

désigné ci-après le modèle FH, est le modèle au niveau du domaine le plus utilisé dans l'estimation sur petits domaines. Le modèle de régression à erreurs emboîtées simple proposé dans Battese, Harter et Fuller (1988), également appelé le modèle de base au niveau de l'unité, est fréquemment utilisé lorsque des données sont disponibles au niveau de l'unité. Nous désignons ce modèle le modèle BHF. Il s'agit de deux cas particuliers d'un modèle linéaire mixte général dans l'EPD (voir Rao et Molina, 2015 pour consulter un excellent exposé sur l'estimation sur petits domaines).

Les moyennes ou les totaux de petits domaines sont les paramètres linéaires les plus fréquemment estimés dans l'EPD. En pareil cas, la méthode d'estimation sur petits domaines la plus populaire est l'utilisation de modèles linéaires mixtes afin de calculer les meilleurs prédicteurs linéaires sans biais (BLUP) pour la moyenne ou le total de petits domaines. Les estimateurs BLUP minimisent l'EQM dans la catégorie des estimateurs linéaires sans biais. Par ailleurs, on peut démontrer que l'estimateur BLUP peut être obtenu par la résolution des équations d'un modèle mixte comportant des inconnues données par les paramètres fixes et aléatoires du modèle. Les équations du modèle mixte résultent de la maximisation de la densité conjointe des données et du vecteur des effets aléatoires de petit domaine. Un estimateur BLUP dépend des variances (et des covariances) des effets aléatoires qu'on peut estimer à l'aide de la méthode d'ajustement de constantes (FC) d'Henderson, du maximum de vraisemblance (MV) ou du maximum de vraisemblance restreint (MVRE). L'utilisation de ces composantes estimées dans l'estimateur BLUP permet d'obtenir un estimateur à deux degrés appelé le meilleur prédicteur linéaire sans biais empirique (EBLUP).

Une difficulté potentielle avec les estimateurs EBLUP est que, lorsqu'ils sont agrégés pour l'ensemble des petits domaines, ils peuvent ne pas correspondre à l'estimation globale pour un plus grand domaine obtenue au moyen d'une estimation directe. Les organismes statistiques favorisent la correspondance globale entre la somme des estimations sur petits domaines fondées sur un modèle et l'estimation directe à un niveau supérieur qui correspond à l'agrégation des petits domaines. La réconciliation est une méthode permettant de modifier les estimations fondées sur un modèle pour qu'elles correspondent à l'estimateur direct pour le plus grand domaine.

Les méthodes de réconciliation actuelles sont fréquentistes ou bayésiennes. Dans le présent document, nous nous attardons à l'approche de réconciliation fréquentiste (pour connaître les procédures de réconciliation bayésiennes, voir You, Rao et Dick, 2004; Datta, Ghosh, Steorts et Maples, 2011 et Nandram et Sayit, 2011). On peut appliquer les méthodes fréquentistes afin d'obtenir des estimations sur petits domaines réconciliées pour les modèles tant au niveau du domaine et qu'au niveau de l'unité.

Nous passons brièvement en revue la documentation actuelle sur les deux types de modèle. Nous décrivons en premier lieu les procédures élaborées pour réconcilier les estimations au niveau du domaine. Pfeffermann et Barnard (1991) ont obtenu un estimateur réconcilié contraint en maximisant la densité conjointe des données et le vecteur des effets aléatoires de petit domaine compte tenu de la restriction de réconciliation. Leur estimateur réconcilié a été construit à l'aide d'estimations ajustées des effets fixes et de petit domaine qui solutionnent le problème de maximisation sous contrainte. Wang, Fuller et Qu (2008)

ont élaboré un estimateur EBLUP réconcilié pour le modèle FH au niveau du domaine, en minimisant une fonction de perte sous la contrainte donnée par la condition de réconciliation. Ils ont obtenu un second estimateur réconcilié en ajoutant une variable auxiliaire adéquate au modèle FH sans imposer de contrainte. Ils ont démontré que l'estimateur EBLUP fondé sur le modèle FH augmenté est autoréconcilié : l'estimateur a satisfait la condition de réconciliation sans autres rajustements. Bell, Datta et Ghosh (2013) ont généralisé les résultats présentés dans Wang et coll. (2008) au cas de multiples contraintes de réconciliation en considérant une fonction de perte plus générale. You, Rao et Hidiroglou (2013) ont obtenu un autre estimateur autoréconcilié sous le modèle FH en remplaçant le vecteur de régression utilisé dans l'estimateur EBLUP par un autre estimateur qui dépend des poids de réconciliation.

Nous décrivons maintenant les procédures qui permettent de réconcilier les estimations fondées sur un modèle au niveau de l'unité. L'objectif est d'obtenir des estimateurs sur petits domaines qui sont réconciliés à un estimateur direct à un niveau d'agrégation donné des petits domaines. Les estimateurs directs principalement utilisés par les organismes statistiques sont l'estimateur par la régression généralisée (GREG) décrit dans Särndal, Swensson et Wretman (1989) ou plus généralement l'estimateur par calage fondé sur les procédures décrites dans Deville et Särndal (1992). You et Rao (2002) ont élaboré un prédicteur pseudo-EBLUP (prédicteur YR) qui intègre des poids de sondage. Une propriété de cet estimateur est qu'il est autoréconcilié, c'est-à-dire que la somme des estimations sur petits domaines correspond à un estimateur ayant la même forme que l'estimateur GREG. Cependant, il ne s'agit pas d'un estimateur direct, car le vecteur de régression estimé faisant partie de cet estimateur témoigne de la structure des erreurs du modèle à erreurs emboîtées. En supposant que les taux d'échantillonnage sont négligeables, Stefan et Hidiroglou (2020) ont proposé plusieurs procédures pour assurer que les estimateurs EBLUP et pseudo-EBLUP soient réconciliés à l'estimateur GREG, étant donné que le modèle et l'estimateur GREG utilisaient tous deux le même vecteur de variables auxiliaires. Ugarte, Militino et Goicoa (2009) ont élaboré un estimateur EBLUP restreint pour un total de petits domaines satisfaisant la propriété de réconciliation à un estimateur synthétique.

Le présent document a pour objectif de comparer plusieurs estimateurs réconciliés d'une moyenne de petits domaines pour le modèle de base au niveau de l'unité lorsque les taux d'échantillonnage sont non négligeables. Nous comparons six estimateurs réconciliés : deux estimateurs réconciliés fondés sur les procédures proposées par Stefan et Hidiroglou (2020), deux estimateurs restreints fondés sur la procédure proposée par Ugarte et coll. (2009) et deux estimateurs par le ratio obtenus en multipliant les estimateurs EBLUP et YR de chaque petit domaine par un facteur d'ajustement commun. Le document est structuré de la manière suivante. La section 2 présente un résumé des estimateurs EBLUP et pseudo-EBLUP sous le modèle de base au niveau de l'unité. La section 3 décrit les six estimateurs réconciliés. Les deux premiers estimateurs sont fondés sur des ajustements par le ratio simples. Ensuite, nous démontrons comment on peut adapter les deux procédures de réconciliation proposées par Stefan et Hidiroglou (2020) dans le cas de taux d'échantillonnage négligeables pour produire des estimateurs de la moyenne de petits domaines réconciliés lorsque les taux d'échantillonnage sont non négligeables. Enfin, nous décrivons l'estimateur

EBLUP restreint d'Ugarte et coll. (2009) et nous proposons un pseudo-estimateur restreint qui est une variante de l'estimateur EBLUP restreint intégrant des poids de sondage. Nous proposons également une méthode utilisant le maximum de vraisemblance restreint reparamétré (MVREre) pour estimer les composantes de la variance. Cette méthode d'estimation est utile pour calculer des estimateurs EBLUP restreints de la moyenne de petits domaines, car elle permet d'obtenir des estimations des composantes de la variance strictement positives. La section 4 présente les résultats d'une simulation de Monte-Carlo fondée sur des ensembles de données générés, tandis que la section 5 présente les résultats d'une étude par simulations fondée sur un ensemble de données réel. Enfin, la section 6 présente quelques observations finales.

2. Estimation EBLUP et pseudo-EBLUP

Prenons le modèle de régression à erreurs emboîtées simple

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, N_i, \tag{2.1}$$

où y_{ij} est la variable d'intérêt pour la j^{e} unité de population dans le i^{e} petit domaine, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ est un vecteur de variables auxiliaires avec $x_{ij1} = 1$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ est un vecteur de paramètres de régression $p \times 1$ et N_i est le nombre d'unités de population dans le i^{e} petit domaine, U_i. Les effets aléatoires de petit domaine v_i sont présumés indépendants et identiquement distribués (i.i.d.) $N(0, \sigma_v^2)$, et indépendants des erreurs au niveau de l'unité e_{ij}, qui sont présumées i.i.d. $N(0, \sigma_e^2)$. Nous prélevons des échantillons s_i de taille n_i de manière indépendante dans chaque petit domaine i, selon un plan d'échantillonnage spécifié, les probabilités d'inclusion de premier ordre étant désignées π_{ij}, pour $j = 1, \ldots, N_i$. La taille totale de l'échantillon est n, où $n = \sum_{i=1}^m n_i$. Les poids de sondage de base qui en résultent sont donnés par $d_{ij} = 1/\pi_{ij}$. Nous supposons que le plan d'échantillonnage est ignorable et qu'il n'y a pas de biais de sélection. Cela suppose que le modèle (2.1) vaut également pour les données de l'échantillon :

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, n_i, \tag{2.2}$$

Le modèle (2.2) est un cas particulier du modèle linéaire mixte général. En définissant $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, $\mathbf{X}_i = (\mathbf{x}_{i1}^T, \ldots, \mathbf{x}_{in_i}^T)^T$, $\mathbf{v} = (v_1, \ldots, v_m)^T$ et $\mathbf{e}_i = (e_{i1}, \ldots, e_{in_i})^T$, il s'ensuit que le modèle (2.2) peut s'exprimer sous une forme matricielle par empilement des observations. L'équation qui en résulte est

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{2.3}$$

où $\mathbf{y} = \text{col}_{1 \le i \le m}(\mathbf{y}_i)$, $\mathbf{X} = \text{col}_{1 \le i \le m}(\mathbf{X}_i)$, $\mathbf{Z} = \text{diag}_{1 \le i \le m}\{\mathbf{1}_{n_i}\}$ et $\mathbf{e} = \text{col}_{1 \le i \le m}(\mathbf{e}_i)$, avec $\mathbf{1}_{n_i}$ un vecteur de dimension n_i composé de uns. Nous désignons $\mathbf{G}$ et $\mathbf{R}$ les matrices de variance des vecteurs

aléatoires $\mathbf{v}$ et $\mathbf{e}$ respectivement. Alors $\mathbf{G} = \sigma_v^2 \mathbf{I}_m$ et $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. Il s'ensuit que la matrice de variance du vecteur $\mathbf{y}$, désignée $\mathbf{V}$, est donnée par $\mathbf{V} = \mathbf{R} + \mathbf{ZGZ}^T$.

Les paramètres d'intérêt sont les moyennes de petits domaines $\bar{Y}_i$, où $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, $i = 1, \ldots, m$. Si N_i est grand, la fraction d'échantillonnage $f_i = N_i^{-1} n_i$ du i^{e} petit domaine est négligeable. Cette configuration correspond au cas d'une *population infinie* ou à des taux d'échantillonnage négligeables. Il s'ensuit que les moyennes de petits domaines $\bar{Y}_i$ peuvent être approximées par μ_i (voir Rao et Molina, 2015, page 174), où $\mu_i = \bar{\mathbf{X}}_i^T \boldsymbol{\beta} + v_i$ et $\bar{\mathbf{X}}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij} / N_i$ est le vecteur des moyennes de population des $\mathbf{x}_{ij}$ pour le i^{e} domaine. Un estimateur de μ_i est donné par $\hat{\mu}_i = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}} + \hat{v}_i$ (Rao et Molina, 2015, page 175), où $\hat{\boldsymbol{\beta}}$ et $\hat{v}_i$ sont des estimateurs de $\boldsymbol{\beta}$ et v_i respectivement. Si N_i n'est pas suffisamment grand ou si les taux d'échantillonnage f_i ne sont pas négligeables, les paramètres $\bar{Y}_i$ ne peuvent être approximés par des combinaisons linéaires de $\boldsymbol{\beta}$ et v_i. Cela correspond au cas d'une *population finie*. Soit r_i l'ensemble des valeurs y non observées $N_i - n_i$ dans le petit domaine i. Si nous supposons que nous connaissons les $\mathbf{x}_{ij}$ pour chaque personne dans la population, un estimateur $\hat{\bar{Y}}_i$ de $\bar{Y}_i$ est fondé sur les valeurs observées de y_{ij}, $j \in s_i$ et les valeurs prédites de $\hat{y}_{ij} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{v}_i$ pour $j \in r_i$. Autrement dit, l'estimateur $\hat{\bar{Y}}_i$ est donné par

$$\hat{\bar{Y}}_i = \frac{1}{N_i} \left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij} \right). \tag{2.4}$$

Une grande partie de la théorie sur l'EPD concerne le cas d'une population infinie, alors que la documentation sur le cas d'une population finie est plus limitée. Dans le présent document, nous nous attardons au cas d'une population finie (ou de taux d'échantillonnage non négligeables) et nous construisons par conséquent des estimateurs fondés sur l'équation (2.4).

2.1 Estimation EBLUP

Nous désignons $\tilde{\boldsymbol{\beta}}$ et $\tilde{\mathbf{v}} = (\tilde{v}_1, \ldots, \tilde{v}_m)^T$ les prédicteurs EBLUP de $\boldsymbol{\beta}$ et de $\mathbf{v}$ respectivement. Ces estimateurs sont donnés par $\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}$ et $\tilde{\mathbf{v}} = \mathbf{GZ}^T \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)$. Dans l'hypothèse de la normalité de $\mathbf{e}$ et de $\mathbf{v}$, on peut démontrer que $\tilde{\boldsymbol{\beta}}$ et $\tilde{\mathbf{v}}$ peuvent être obtenus en maximisant la densité conjointe de $\mathbf{y}$ et de $\mathbf{v}$ en ce qui concerne $\boldsymbol{\beta}$ et $\mathbf{v}$. Ceci équivaut à minimiser la fonction

$$\phi = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Zv})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Zv}) + \mathbf{v}^T \mathbf{G}^{-1} \mathbf{v}. \tag{2.5}$$

Il en résulte les équations d'un modèle mixte suivantes

$$\mathbf{A} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{v}} \end{pmatrix} = \mathbf{b}, \tag{2.6}$$

où

$$\mathbf{A} = \begin{pmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \quad \text{et} \quad \mathbf{b} = \begin{pmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{y} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{y} \end{pmatrix}. \tag{2.7}$$

(voir Rao et Molina, 2015, page 99 pour plus de détails). Les composantes de la variance (σ_v^2, σ_e^2) dans les équations (2.6) et (2.7) sont généralement inconnues. Trois méthodes d'estimation, la méthode FC, le MV et le MVRE, sont couramment utilisées dans l'EPD pour estimer les composantes de la variance (σ_v^2, σ_e^2). Une difficulté bien connue avec ces méthodes est que l'estimation de σ_v^2 peut prendre des valeurs négatives. En pareil cas, cette estimation est tronquée à zéro, c'est-à-dire que $\hat{\sigma}_v^2$ reçoit la valeur 0. Des versions empiriques de $\mathbf{A}$ et de $\mathbf{b}$, désignées $\hat{\mathbf{A}}$ et $\hat{\mathbf{b}}$, sont obtenues si les composantes inconnues de la variance (σ_v^2, σ_e^2) sont remplacées par des estimateurs $(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$. Il découle de l'équation (2.6) que des estimateurs EBLUP des paramètres du modèle $(\boldsymbol{\beta}, \mathbf{v})$, désignés $\hat{\boldsymbol{\beta}}$ et $\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_m)^T$, sont donnés par

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{v}} \end{pmatrix} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{b}}. \tag{2.8}$$

Au moyen de l'équation (2.8), il peut être démontré que $\hat{\boldsymbol{\beta}}$ et $\hat{\mathbf{v}}$ sont

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{v}} \end{pmatrix} = \begin{pmatrix} \left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y} \\ \hat{\mathbf{G}}\mathbf{Z}^T\hat{\mathbf{V}}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \end{pmatrix}, \tag{2.9}$$

où $\hat{\mathbf{G}} = \hat{\sigma}_v^2\mathbf{I}_m$ et $\hat{\mathbf{V}} = \hat{\sigma}_e^2\mathbf{I}_n + \hat{\sigma}_v^2\mathbf{Z}\mathbf{Z}^T$.

Remarque 1. Il est plus facile d'inverser les matrices $\hat{\mathbf{G}} = \hat{\sigma}_v^2\mathbf{I}_m$ et $\hat{\mathbf{R}} = \hat{\sigma}_e^2\mathbf{I}_n$ que $\hat{\mathbf{V}}$. En conséquence, il est plus simple d'utiliser les équations du modèle mixte (2.8) que les équations (2.9) pour calculer $\hat{\boldsymbol{\beta}}$ et $\hat{\mathbf{v}}$. Cependant, lorsque $\hat{\sigma}_v^2$ est égal à zéro, les équations (2.8) ne peuvent être utilisées, car le terme $\hat{\mathbf{G}}^{-1}$ dans la matrice $\hat{\mathbf{A}}$ n'existe pas. En pareil cas, $\hat{\boldsymbol{\beta}}$ et $\hat{\mathbf{v}}$ ne peuvent être calculés qu'avec les équations (2.9).

Dans le modèle (2.2), on peut démontrer que $\hat{\boldsymbol{\beta}}$ et $\hat{v}_i$ dans $\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_m)^T$ satisfont l'équation

$$\sum_{i=1}^{m}\sum_{j\in s_i}\mathbf{x}_{ij}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} - \hat{v}_i\right) = 0. \tag{2.10}$$

Des estimateurs $\hat{\boldsymbol{\beta}}$ et $\hat{v}_i$ sont utilisés pour calculer des prédicteurs EBLUP $\hat{y}_{ij}^{\text{EBLUP}}$ pour les unités non observées $N_i - n_i$ dans le petit domaine i : $\hat{y}_{ij}^{\text{EBLUP}} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{v}_i$ pour $j \in r_i$. Un estimateur EBLUP de $\bar{Y}_i$, désigné $\hat{\bar{Y}}_i^{\text{EBLUP}}$, est obtenu en remplaçant $\hat{y}_{ij}$ par $\hat{y}_{ij}^{\text{EBLUP}}$ dans l'équation (2.4). Il s'ensuit que $\hat{\bar{Y}}_i^{\text{EBLUP}}$ est

$$\hat{\bar{Y}}_i^{\text{EBLUP}} = \frac{1}{N_i}\left[\sum_{j\in s_i} y_{ij} + \mathbf{x}_{ir}^T\hat{\boldsymbol{\beta}} + (N_i - n_i)\,\hat{v}_i\right], \tag{2.11}$$

où $\mathbf{x}_{ir} = \sum_{j\in r_i}\mathbf{x}_{ij}$ représente la somme des valeurs non échantillonnées $\mathbf{x}_{ij}$.

2.2 Estimation de You-Rao

You et Rao (2002) ont proposé un estimateur de la moyenne de petits domaines pseudo-EBLUP (estimateur YR) qui intègre les poids de sondage d_{ij} à la formule de l'estimateur EBLUP. Une propriété

de l'estimateur pseudo-EBLUP est que la cohérence du plan est préservée lorsque la taille d'échantillon dans le domaine augmente. De plus, le prédicteur YR protège contre l'échec du modèle ou un plan d'échantillonnage informatif (voir entre autres Hidiroglou et Estevao, 2016 et Verret, Rao et Hidiroglou, 2015 pour plus de détails). On peut construire des estimateurs pseudo-EBLUP en utilisant la procédure décrite dans You et Rao (2002) avec des poids de sondage w_{ij} qui peuvent être calés sur un vecteur de variables auxiliaires. Soient $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ et $\hat{\mathbf{v}}^{\mathrm{YR}} = \left(\hat{v}_1^{\mathrm{YR}}, \ldots, \hat{v}_m^{\mathrm{YR}}\right)^T$ les estimateurs YR de $\boldsymbol{\beta}$ et $\mathbf{v}$ respectivement fondés sur des poids w_{ij} (voir You et Rao, 2002 pour plus de détails). Les estimateurs $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ et $\hat{v}_i^{\mathrm{YR}}$ satisfont les équations d'estimation au niveau de l'unité

$$\sum_{i=1}^{m} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \left(y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}^{\mathrm{YR}} - \hat{v}_i^{\mathrm{YR}} \right) = 0. \tag{2.12}$$

Les équations (2.12) représentent la version pondérée par les poids de sondage des équations (2.10). Les prédicteurs de You-Rao $\hat{y}_{ij}^{\mathrm{YR}}$ de y_{ij} sont calculés comme $\hat{y}_{ij}^{\mathrm{YR}} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}^{\mathrm{YR}} + \hat{v}_i^{\mathrm{YR}}$ pour $j \in r_i$. Le remplacement de $\hat{y}_{ij}$ par $\hat{y}_{ij}^{\mathrm{YR}}$ dans l'équation (2.4) permet d'obtenir l'estimateur YR de $\bar{Y}_i$ dans le cas de taux d'échantillonnage non négligeables :

$$\hat{\bar{Y}}_i^{\mathrm{YR}} = \frac{1}{N_i} \left[\sum_{j \in s_i} y_{ij} + \mathbf{x}_{ir}^T \hat{\boldsymbol{\beta}}^{\mathrm{YR}} + (N_i - n_i)\, \hat{v}_i^{\mathrm{YR}} \right]. \tag{2.13}$$

On peut obtenir autrement les estimateurs $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ et $\hat{\mathbf{v}}^{\mathrm{YR}}$ comme solutions à des équations d'un modèle mixte pondéré semblables à l'équation (2.6) (voir Huang et Hidiroglou, 2003 pour plus de détails). À cette fin, nous définissons les matrices $\mathbf{W}_i = \mathrm{diag}_{1 \le j \le n_i} \{w_{ij}\}$, $\mathbf{W} = \mathrm{diag}_{1 \le i \le m} \{\mathbf{W}_i\}$ et $\boldsymbol{\Omega} = \mathrm{diag}_{1 \le i \le m} \{\omega_i\}$, où $\omega_i = \sum_{j \in s_i} w_{ij}^2 \Big/ \sum_{j \in s_i} w_{ij}$ pour $i = 1, \ldots, m$. Soit ϕ_w la version pondérée de l'échantillon ϕ, où

$$\phi_w = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v})^T\, \mathbf{W}^{1/2} \mathbf{R}^{-1} \mathbf{W}^{1/2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}) + \mathbf{v}^T \boldsymbol{\Omega}^{1/2} \mathbf{G}^{-1} \boldsymbol{\Omega}^{1/2} \mathbf{v}, \tag{2.14}$$

$\mathbf{W}^{1/2}$ et $\boldsymbol{\Omega}^{1/2}$ représentant la racine carrée des matrices $\mathbf{W}$ et $\boldsymbol{\Omega}$ respectivement. Dans le premier terme de ϕ_w, l'erreur du modèle associée à l'observation y_{ij} est pondérée par le poids de sondage correspondant w_{ij}, tandis que dans le second terme de ϕ_w, le facteur ω_i dans la matrice diagonale $\boldsymbol{\Omega}$ représente le poids lié à l'effet de petit domaine v_i. On peut démontrer que la minimisation de ϕ_w en ce qui concerne $\boldsymbol{\beta}$ et $\mathbf{v}$ permet d'obtenir $\left(\hat{\boldsymbol{\beta}}^{\mathrm{YR}}, \hat{\mathbf{v}}^{\mathrm{YR}}\right)$. Il s'ensuit que $\left(\hat{\boldsymbol{\beta}}^{\mathrm{YR}}, \hat{\mathbf{v}}^{\mathrm{YR}}\right)$ sont donnés par

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{YR}} \\ \hat{\mathbf{v}}^{\mathrm{YR}} \end{pmatrix} = \hat{\mathbf{A}}_w^{-1} \hat{\mathbf{b}}_w, \tag{2.15}$$

où les valeurs connues de $\mathbf{A}_w$ et $\mathbf{b}_w$ sont données par

$$\mathbf{A}_w = \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{1/2} \mathbf{R}^{-1} \mathbf{W}^{1/2} \mathbf{X} & \mathbf{X}^T \mathbf{W}^{1/2} \mathbf{R}^{-1} \mathbf{W}^{1/2} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W}^{1/2} \mathbf{R}^{-1} \mathbf{W}^{1/2} \mathbf{X} & \mathbf{Z}^T \mathbf{W}^{1/2} \mathbf{R}^{-1} \mathbf{W}^{1/2} \mathbf{Z} + \boldsymbol{\Omega}^{1/2} \mathbf{G}^{-1} \boldsymbol{\Omega}^{1/2} \end{pmatrix} \text{ et } \mathbf{b}_w = \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{1/2} \mathbf{R}^{-1} \mathbf{W}^{1/2} \mathbf{y} \\ \mathbf{Z}^T \mathbf{W}^{1/2} \mathbf{R}^{-1} \mathbf{W}^{1/2} \mathbf{y} \end{pmatrix}, \tag{2.16}$$

et $\hat{\mathbf{A}}_w$ et $\hat{\mathbf{b}}_w$ sont des versions empiriques de $\mathbf{A}_w$ et $\mathbf{b}_w$ obtenues en estimant $\mathbf{G}$ et $\mathbf{R}$ par $\hat{\mathbf{G}} = \hat{\sigma}_v^2 \mathbf{I}_m$ et $\hat{\mathbf{R}} = \hat{\sigma}_e^2 \mathbf{I}_n$ respectivement. L'équation (2.15) peut également s'écrire sous la forme

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{YR}} \\ \hat{\mathbf{v}}^{\mathrm{YR}} \end{pmatrix} = \begin{pmatrix} \left(\mathbf{X}^T \hat{\mathbf{V}}_w^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \hat{\mathbf{V}}_w^{-1} \mathbf{y} \\ \hat{\mathbf{G}}_\omega \mathbf{Z}^T \hat{\mathbf{V}}_w^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{YR}}\right) \end{pmatrix}, \tag{2.17}$$

où $\hat{\mathbf{G}}_\omega = \boldsymbol{\Omega}^{-1/2}\hat{\mathbf{G}}\boldsymbol{\Omega}^{-1/2}$ et $\hat{\mathbf{V}}_w = \mathbf{W}^{-1/2}\hat{\mathbf{R}}\mathbf{W}^{-1/2} + \mathbf{Z}\hat{\mathbf{G}}_\omega\mathbf{Z}^T$.

3. Estimateurs réconciliés

Nous élaborons maintenant des estimateurs réconciliés des moyennes de petits domaines $\bar{Y}_i$ à l'aide du modèle au niveau de l'unité (2.2) ou de versions augmentées de ce dernier. Nous supposons qu'un estimateur direct fiable $\hat{Y}_w = \sum_{i=1}^m \sum_{j \in s_i} w_{ij} y_{ij}$ du total de population Y est disponible, où $Y = \sum_{i=1}^m Y_i$ et $Y_i = N_i \bar{Y}_i$ est le total du petit domaine i. Soit $\hat{\bar{Y}}_i$ l'estimateur sur petits domaines fondé sur un modèle de $\bar{Y}_i$. Il est souhaitable de veiller à ce que les valeurs agrégées de $\hat{\bar{Y}}_i$ correspondent à l'estimateurs fiable $\hat{Y}_w$. Les estimateurs de moyennes de petits domaines $\hat{\bar{Y}}_i$, $i = 1, \ldots, m$ sont considérés réconciliés à $\hat{Y}_w$ si

$$\sum_{i=1}^m N_i \hat{\bar{Y}}_i = \hat{Y}_w. \tag{3.1}$$

Soit $\hat{Y}_w$ un estimateur GREG avec des poids calés au niveau de la population sur un vecteur de variables auxiliaires $\mathbf{x}_{ij}^*$. Cet estimateur est analogue à l'estimateur par la régression combiné si l'on considère les petits domaines comme des strates. Le vecteur de variables auxiliaires $\mathbf{x}_{ij}^*$ peut ou non être le même que $\mathbf{x}_{ij}$. Nous distinguons deux cas dans ce contexte : $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$ et $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$. Le premier cas, $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$, suppose que toutes les composantes de $\mathbf{x}_{ij}$ appartiennent également à $\mathbf{x}_{ij}^*$ et que $\mathbf{x}_{ij}^*$ peut ou non comporter d'autres composantes qui sont différentes de celles contenues dans $\mathbf{x}_{ij}$. Le second cas, $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$, suppose que certaines des composantes de $\mathbf{x}_{ij}$ n'apparaissent pas dans $\mathbf{x}_{ij}^*$. Nous supposons que la première composante des deux vecteurs $\mathbf{x}_{ij}$ et $\mathbf{x}_{ij}^*$ est égale à un, car ces derniers représentent un terme d'ordonnée à l'origine.

Pour un échantillon donné s, les données auxiliaires $\mathbf{x}_{ij}^*$ et les poids de sondage de base $d_{ij} = 1/\pi_{ij}$, l'estimateur GREG du total de population Y est donné par

$$\hat{Y}^{\mathrm{GREG}} = \sum_{i=1}^m \sum_{j \in s_i} w_{ij}^{\mathrm{GREG}} y_{ij},$$

où les poids GREG w_{ij}^{GREG} sont donnés par

$$w_{ij}^{\mathrm{GREG}} = d_{ij}\left(1 + \left(\mathbf{X}^* - \hat{\mathbf{X}}^{*\mathrm{HT}}\right)^T \left(\sum_{i=1}^m \sum_{j \in s_i} d_{ij}\mathbf{x}_{ij}^*\mathbf{x}_{ij}^{*T}\right)^{-1} \mathbf{x}_{ij}^*\right). \tag{3.2}$$

Dans l'équation (3.2), $\mathbf{X}^* = \sum_{i=1}^m \mathbf{X}_i^*$, où $\mathbf{X}_i^* = \sum_{j=1}^{N_i} \mathbf{x}_{ij}^*$ représente le total de petits domaines connu, tandis que $\hat{\mathbf{X}}^{*\mathrm{HT}} = \sum_{i=1}^m \hat{\mathbf{X}}_i^{*\mathrm{HT}}$ et $\hat{\mathbf{X}}_i^{*\mathrm{HT}} = \sum_{j \in s_i} d_{ij}\mathbf{x}_{ij}^*$ représentent respectivement les estimateurs directs de Horvitz-Thompson fondés sur un plan de sondage de $\mathbf{X}^*$ et $\mathbf{X}_i^*$. Il convient de souligner que

$$\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}^{\mathrm{GREG}}\mathbf{x}_{ij}^{*} = \mathbf{X}^{*}. \tag{3.3}$$

À l'aide des poids GREG w_{ij}^{GREG}, les estimateurs de N_i et $\mathbf{X}_i$ sont donnés par

$$\hat{N}_i^{\mathrm{GREG}} = \sum_{j\in s_i} w_{ij}^{\mathrm{GREG}} \quad \text{et} \quad \hat{\mathbf{X}}_i^{\mathrm{GREG}} = \sum_{j\in s_i} w_{ij}^{\mathrm{GREG}}\mathbf{x}_{ij}. \tag{3.4}$$

Les estimations sur petits domaines $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ et $\hat{\bar{Y}}_i^{\mathrm{YR}}$ données respectivement par les équations (2.11) et (2.13) ne satisfont pas l'équation de réconciliation (3.1) pour $\hat{Y}_w = \hat{Y}^{\mathrm{GREG}}$, c'est-à-dire que les estimations totales $\hat{Y}^{\mathrm{EBLUP}} = \sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ et $\hat{Y}^{\mathrm{YR}} = \sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{YR}}$ ne correspondent pas à l'estimateur GREG $\hat{Y}^{\mathrm{GREG}}$. Nous devons ajuster $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ et $\hat{\bar{Y}}_i^{\mathrm{YR}}$ de sorte que la somme de ces estimateurs sur petits domaines modifiés corresponde à $\hat{Y}^{\mathrm{GREG}}$ lorsque la somme de tous les petits domaines m est faite.

Une modification très simple des $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ et des $\hat{\bar{Y}}_i^{\mathrm{YR}}$ est appelée la réconciliation par le ratio. Elle consiste à multiplier chaque $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ et $\hat{\bar{Y}}_i^{\mathrm{YR}}$ par les facteurs d'ajustement communs $\hat{Y}^{\mathrm{GREG}}\Big/\sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ et $\hat{Y}^{\mathrm{GREG}}\Big/\sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{YR}}$ respectivement, permettant d'obtenir les estimateurs réconciliés par le ratio

$$\hat{\bar{Y}}_{ib}^{\mathrm{EBRat}} = \hat{\bar{Y}}_i^{\mathrm{EBLUP}}\frac{\hat{Y}^{\mathrm{GREG}}}{\sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{EBLUP}}} \quad \text{et} \quad \hat{\bar{Y}}_{ib}^{\mathrm{YRat}} = \hat{\bar{Y}}_i^{\mathrm{YR}}\frac{\hat{Y}^{\mathrm{GREG}}}{\sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{YR}}}. \tag{3.5}$$

Il s'ensuit facilement que $\hat{\bar{Y}}_{ib}^{\mathrm{EBRat}}$ et $\hat{\bar{Y}}_{ib}^{\mathrm{YRat}}$ satisfont tous deux l'équation (3.1) avec $\hat{Y}_w = \hat{Y}^{\mathrm{GREG}}$. Dans l'équation (3.5) et ci-après, l'indice b indique que les estimateurs sont réconciliés à $\hat{Y}^{\mathrm{GREG}}$.

Il convient de souligner que les $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ et les $\hat{\bar{Y}}_i^{\mathrm{YR}}$ dans l'équation (3.5) sont multipliés par le même facteur peu importe leur précision et en ignorant les caractéristiques des petits domaines en particulier, comme la variabilité des unités dans un petit domaine ou la taille d'échantillon des petits domaines. En conséquence, les estimateurs réconciliés qui en résultent, $\hat{\bar{Y}}_{ib}^{\mathrm{EBRat}}$ et $\hat{\bar{Y}}_{ib}^{\mathrm{YRat}}$, d'après cette procédure simple, ne sont que des modifications proportionnelles des estimateurs $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ et $\hat{\bar{Y}}_i^{\mathrm{YR}}$ respectivement, pour obtenir la concordance voulue. On peut éviter cette limite en utilisant le modèle pour petits domaines (2.2) afin de construire les estimateurs réconciliés.

Nous démontrons maintenant comment le modèle (2.2) peut être utilisé pour obtenir des estimateurs réconciliés à $\hat{Y}^{\mathrm{GREG}}$. Aux sections 3.1 et 3.2, nous adaptons les procédures décrites dans Stefan et Hidiroglou (2020) pour obtenir des estimateurs réconciliés au cas des taux d'échantillonnage non négligeables. Aux sections 3.3 et 3.4, nous présentons deux estimateurs réconciliés restreints d'après la procédure proposée par Ugarte et coll. (2009). Les estimateurs réconciliés des sections 3.1 et 3.2 reposent sur l'hypothèse selon laquelle $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^{*}$, tandis que les estimateurs des sections 3.3 et 3.4 peuvent être calculés pour n'importe quel vecteur $\mathbf{x}_{ij}$ et $\mathbf{x}_{ij}^{*}$.

3.1 Estimateurs réconciliés EBLUP augmentés

Les poids GREG w_{ij}^{GREG} doivent être utilisés dans l'estimation pour réaliser la réconciliation à $\hat{Y}^{\mathrm{GREG}}$. Une façon possible d'intégrer w_{ij}^{GREG} à l'estimation consiste à augmenter le modèle pour petits domaines

(2.2) au moyen d'une variable auxiliaire adéquate qui est une fonction de w_{ij}^{GREG}. Cette procédure repose sur l'approche de modèle augmenté adoptée par Wang et coll. (2008), dans laquelle les estimations obtenues au moyen du modèle FH au niveau du domaine ont pu être contraintes de correspondre à des totaux déterminés. Stefan et Hidiroglou (2020) ont adapté l'approche de Wang et coll. (2008) sous le modèle de base au niveau de l'unité et pour des taux d'échantillonnage négligeables. Ils ont démontré que la réconciliation à $\hat{Y}^{\mathrm{GREG}}$ pouvait être obtenue en augmentant le modèle (2.2) avec les poids GREG w_{ij}^{GREG}. Nous étendons l'approche de Stefan et Hidiroglou (2020) au cas des taux d'échantillonnage non négligeables. Dans ce cas, la réconciliation à $\hat{Y}^{\mathrm{GREG}}$ est réalisée en augmentant le modèle (2.2) avec $q_{ij} = w_{ij}^{\mathrm{GREG}} - 1$. Il en résulte le modèle augmenté donné par

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_{1a} + q_{ij}\beta_{2a} + v_{ia} + e_{ija},\ i = 1, \ldots, m;\ j \in s_i. \tag{3.6}$$

Les effets aléatoires v_{ia} sont présumés i.i.d. $N\left(0, \sigma_{va}^2\right)$ et indépendants des erreurs au niveau de l'unité e_{ija}, et les e_{ija} sont présumés i.i.d. $N\left(0, \sigma_{ea}^2\right)$. Les estimateurs EBLUP de $\boldsymbol{\beta}_a = \left(\boldsymbol{\beta}_{1a}^T, \beta_{2a}\right)^T$ et v_{ia} sous l'équation (3.6) sont respectivement désignés $\hat{\boldsymbol{\beta}}_a = \left(\hat{\boldsymbol{\beta}}_{1a}^T, \hat{\beta}_{2a}\right)^T$ et $\hat{v}_{ia}$. Nous pouvons maintenant démontrer le résultat 1 pour $\hat{\boldsymbol{\beta}}_a$ et $\hat{v}_{ia}$.

Résultat 1. Les estimateurs EBLUP $\hat{\boldsymbol{\beta}}_a$ et $\hat{v}_{ia}$ reposant sur le modèle (3.6) obéissent à l'équation suivante

$$\sum_{i=1}^{m}\sum_{j\in s_i} y_{ij} + \left(\sum_{i=1}^{m}\mathbf{x}_{ir}\right)^T \hat{\boldsymbol{\beta}}_{1a} + \sum_{i=1}^{m} q_{iw}\hat{\beta}_{2a} + \sum_{i=1}^{m}\left(\hat{N}_i^{\mathrm{GREG}} - n_i\right)\hat{v}_{ia} = \hat{Y}^{\mathrm{GREG}}, \tag{3.7}$$

où $q_{iw} = \sum_{j\in s_i} q_{ij}^2 = \sum_{j\in s_i} (w_{ij}^{\mathrm{GREG}} - 1)^2$.

Preuve : Voir l'annexe A.

Il découle de l'équation (3.7) que les estimateurs sur petits domaines réconciliés à $\hat{Y}^{\mathrm{GREG}}$ sont donnés par

$$\hat{\bar{Y}}_{iab}^{\mathrm{EBLUP}} = \frac{1}{N_i}\left[\sum_{j\in s_i} y_{ij} + \mathbf{x}_{ir}^T\hat{\boldsymbol{\beta}}_{1a} + q_{iw}\hat{\beta}_{2a} + \left(\hat{N}_i^{\mathrm{GREG}} - n_i\right)\hat{v}_{ia}\right]. \tag{3.8}$$

L'indice a indique que $\hat{\bar{Y}}_{iab}^{\mathrm{EBLUP}}$ repose sur le modèle pour petits domaines augmenté.

3.2 Estimateurs réconciliés de You-Rao

On peut utiliser la procédure proposée par You et Rao (2002) avec des poids de sondage w_{ij} quelconques. Cependant, il n'est pas garanti que l'estimateur YR qui en résultera sera réconcilié à $\hat{Y}^{\mathrm{GREG}}$. Pour des taux d'échantillonnage négligeables, Stefan et Hidiroglou (2020) ont obtenu des estimateurs réconciliés au moyen de la procédure de You et Rao (2002) d'après les poids $w_{ij} = w_{ij}^{\mathrm{GREG}}$ de l'estimateur GREG. Pour des taux d'échantillonnage non négligeables, nous démontrons maintenant que les poids $w_{ij} = w_{ij}^{\mathrm{GREG}} - 1$ permettent d'obtenir des estimateurs YR réconciliés.

Soient $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ et $\hat{\mathbf{v}}^{\mathrm{YR}} = \left(\hat{v}_1^{\mathrm{YR}}, \ldots, \hat{v}_m^{\mathrm{YR}}\right)^T$ des estimateurs YR de $\boldsymbol{\beta}$ et $\mathbf{v}$ respectivement, w_{ij} étant remplacé par $w_{ij}^{\mathrm{GREG}} - 1$. En utilisant $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$, $\hat{v}_i^{\mathrm{YR}}$ et les $N_i - n_i$ estimations $\hat{y}_{ij}^{\mathrm{YR}} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} + \hat{v}_i^{\mathrm{YR}}$ pour

$j \in r_i$, on peut calculer un estimateur YR, désigné $\hat{\bar{Y}}_i^{\mathrm{YR}}$, à l'aide de l'équation (2.13). Cependant, $\hat{\bar{Y}}_i^{\mathrm{YR}}$ n'est pas réconcilié à $\hat{Y}^{\mathrm{GREG}}$ même s'il utilise les poids $w_{ij}^{\mathrm{GREG}} - 1$. La procédure YR originale permet d'obtenir un estimateur autoréconcilié dans un nombre limité de cas.

Pour réaliser la réconciliation à $\hat{Y}^{\mathrm{GREG}}$, un estimateur YR modifié, désigné $\hat{\bar{Y}}_{ib}^{\mathrm{YR}}$, est défini comme suit :

$$\hat{\bar{Y}}_{ib}^{\mathrm{YR}} = \frac{1}{N_i}\left[\sum_{j \in s_i} y_{ij} + \mathbf{x}_{ir}^T \hat{\boldsymbol{\beta}}^{\mathrm{YR}} + \left(\hat{N}_i^{\mathrm{GREG}} - n_i\right) \hat{v}_i^{\mathrm{YR}}\right]. \tag{3.9}$$

Ce qui suit démontre que $\hat{\bar{Y}}_{ib}^{\mathrm{YR}}$ défini au moyen de l'équation (3.9) est réconcilié à $\hat{Y}^{\mathrm{GREG}}$.

Résultat 2. Soient $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ et $\hat{\mathbf{v}}^{\mathrm{YR}} = \left(\hat{v}_1^{\mathrm{YR}}, \ldots, \hat{v}_m^{\mathrm{YR}}\right)^T$ les estimateurs YR de $\boldsymbol{\beta}$ et $\mathbf{v}$ respectivement, construits avec les poids $w_{ij}^{\mathrm{GREG}} - 1$. Alors, $\left(\hat{\boldsymbol{\beta}}^{\mathrm{YR}}, \hat{\mathbf{v}}^{\mathrm{YR}}\right)$ satisfait l'équation suivante :

$$\sum_{i=1}^{m} \sum_{j \in s_i} y_{ij} + \sum_{i=1}^{m} \mathbf{x}_{ir}^T \hat{\boldsymbol{\beta}}^{\mathrm{YR}} + \sum_{i=1}^{m} \left(\hat{N}_i^{\mathrm{GREG}} - n_i\right) \hat{v}_i^{\mathrm{YR}} = \hat{Y}^{\mathrm{GREG}}.$$

Preuve : Voir l'annexe A.

Étant donné $\mathbf{x}_{ij}^*$, les poids w_{ij}^{GREG} sont calés sur $\mathbf{x}_{ij}^*$ au niveau du petit domaine s'ils satisfont les équations suivantes

$$\sum_{j \in s_i} w_{ij}^{\mathrm{GREG}} \mathbf{x}_{ij}^* = \mathbf{X}_i^*, \quad \text{pour} \quad i = 1, \ldots, m. \tag{3.10}$$

Les équations (3.10) supposent l'équation (3.3), mais l'inverse n'est pas vrai. Si les poids w_{ij}^{GREG} satisfont les équations (3.10), et étant donné que $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$, il s'ensuit que les poids w_{ij}^{GREG} sont également calés sur $\mathbf{x}_{ij}$ au niveau du petit domaine. Du même coup, cela suppose que $\hat{N}_i^{\mathrm{GREG}} = N_i$, car nous présumons que le vecteur $\mathbf{x}_{ij}$ contient le régresseur constant égal à 1. Il s'ensuit que $\hat{\bar{Y}}_i^{\mathrm{YR}} = \hat{\bar{Y}}_{ib}^{\mathrm{YR}}$. Ainsi, l'estimateur YR $\hat{\bar{Y}}_i^{\mathrm{YR}}$ construit avec $w_{ij}^{\mathrm{GREG}} - 1$ est autoréconcilié à $\hat{Y}^{\mathrm{GREG}}$ dans le cas particulier où les poids GREG sont calés au niveau du petit domaine (voir You et Rao, 2002).

3.3 Estimateur réconcilié EBLUP restreint

À la section 2, nous avons démontré qu'on peut obtenir les estimateurs EBLUP de $(\boldsymbol{\beta}, \mathbf{v})$ si la fonction ϕ définie dans l'équation (2.5) est minimisée en ce qui concerne $(\boldsymbol{\beta}, \mathbf{v})$. Il s'ensuit par conséquent qu'on peut considérer un estimateur EBLUP comme la solution à un problème de minimisation non restreinte. L'idée des estimateurs EBLUP restreints est d'obtenir de nouveaux estimateurs de $(\boldsymbol{\beta}, \mathbf{v})$ en minimisant ϕ sous la restriction donnée par la condition de réconciliation. La procédure a été utilisée par Pfeffermann et Barnard (1991) sous le modèle FH au niveau du domaine. Plus récemment, Ugarte et coll. (2009) ont appliqué la procédure sous le modèle BHF au niveau de l'unité pour obtenir la réconciliation à un estimateur synthétique. Ugarte et coll. (2009) ont décrit l'estimateur restreint comme un estimateur par les moindres carrés généralisés sous une restriction en constatant que la minimisation peut être réalisée

comme dans la théorie économétrique de l'estimation par régression sous contraintes linéaires. Nous décrivons maintenant la procédure présentée dans Ugarte et coll. (2009).

Nous désignons $\hat{\boldsymbol{\beta}}^R$ et $\hat{\mathbf{v}}^R = \left(\hat{v}_1^R, \ldots, \hat{v}_m^R\right)^T$ les nouveaux estimateurs EBLUP restreints de $(\boldsymbol{\beta}, \mathbf{v})$. Ensuite, l'estimateur EBLUP restreint de $\bar{Y}_i$, désigné $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, est donné par l'équation (2.4), où $\hat{y}_{ij}$ sont remplacés par $\hat{y}_{ij}^R = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^R + \hat{v}_i^R$, pour $j \in r_i$. Nous imposons la condition que les estimateurs $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, $i = 1, \ldots, m$ soient réconciliés à $\hat{Y}^{\text{GREG}}$, c'est-à-dire qu'ils satisfassent l'équation (3.1) avec $\hat{Y}_w = \hat{Y}^{\text{GREG}}$. Après avoir fait quelques calculs algébriques, on peut démontrer que la réconciliation à $\hat{Y}^{\text{GREG}}$ des estimateurs $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, $i = 1, \ldots, m$ équivaut à l'équation de contrainte linéaire suivante

$$\mathbf{a}_1^T\hat{\boldsymbol{\beta}}^R + \mathbf{a}_2^T\hat{\mathbf{v}}^R = \hat{Y}_r^{\text{GREG}}, \tag{3.11}$$

où $\mathbf{a}_1 = \sum_{i=1}^m \mathbf{x}_{ir}$, $\mathbf{a}_2 = (N_1 - n_1, \ldots, N_m - n_m)^T$, $Y_r = Y - \sum_{i=1}^m \sum_{j \in s_i} y_{ij}$ est le total des valeurs de y_{ij} non observées avec $i = 1, \ldots, m$; $j \in r_i$, et $\hat{Y}_r^{\text{GREG}} = \hat{Y}^{\text{GREG}} - \sum_{i=1}^m \sum_{j \in s_i} y_{ij}$ est un estimateur de Y_r d'après $\hat{Y}^{\text{GREG}}$. Les estimateurs EBLUP restreints $\left(\hat{\boldsymbol{\beta}}^R, \hat{\mathbf{v}}^R\right)$ sont par conséquent obtenus comme la solution à la fonction de minimisation ϕ donnée par l'équation (2.5) sous la contrainte linéaire (3.11).

On peut utiliser la méthode des multiplicateurs de Lagrange pour résoudre la minimisation sous contrainte de ϕ. Après des calculs algébriques simples, on peut démontrer que des estimateurs $\left(\hat{\boldsymbol{\beta}}^R, \hat{\mathbf{v}}^R\right)$. sont donnés par

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^R \\ \hat{\mathbf{v}}^R \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{v}} \end{pmatrix} + \frac{1}{\mathbf{a}^T\hat{\mathbf{A}}\mathbf{a}}\,\hat{\mathbf{A}}^{-1}\mathbf{a}\left[\hat{Y}_r^{\text{GREG}} - \mathbf{a}^T\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{v}} \end{pmatrix}\right], \tag{3.12}$$

où $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$ sont les estimateurs EBLUP (non contraints) de $(\boldsymbol{\beta}, \mathbf{v})$, $\hat{\mathbf{A}}$ est la version empirique de la matrice $\mathbf{A}$ définie dans l'équation (2.7) et $\mathbf{a} = \left(\mathbf{a}_1^T\ \mathbf{a}_2^T\right)^T$. Ensuite, en utilisant $\hat{y}_{ij}^R$ dans l'équation (2.4), on peut réécrire l'estimateur $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ sous la forme

$$\hat{\bar{Y}}_{ib}^{\text{REBLUP}} = \frac{1}{N_i}\left[\sum_{j \in s_i} y_{ij} + \mathbf{x}_{ir}^T\hat{\boldsymbol{\beta}}^R + (N_i - n_i)\,\hat{v}_i^R\right]. \tag{3.13}$$

Remarque 2. La matrice $\hat{\mathbf{A}}$ n'existe pas pour les échantillons lorsque $\hat{\sigma}_v^2 = 0$. En pareil cas, nous avons constaté que l'équation (2.8) ne peut servir à calculer les estimateurs non contraints $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$. Cependant, on peut tout de même calculer $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$ lorsque $\hat{\sigma}_v^2 = 0$ parce que l'autre équation (2.9) peut être utilisée pour $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$. L'équation (3.12) démontre clairement que l'estimateur contraint $\left(\hat{\boldsymbol{\beta}}^R, \hat{\mathbf{v}}^R\right)$ ne peut être calculé pour des échantillons lorsque l'estimateur $\hat{\sigma}_v^2$ est tronqué à zéro, et aucune autre équation n'existe en pareil cas.

Il s'ensuit par conséquent que les méthodes d'estimation des composantes de la variance couramment utilisées dans l'EPD ne peuvent être utilisées pour calculer l'estimateur EBLUP restreint. À la section 3.4 et à l'annexe B, nous décrivons une autre méthode permettant de produire une estimation strictement positive de σ_v^2 pouvant être appliquée de pair avec $\left(\hat{\boldsymbol{\beta}}^R, \hat{\mathbf{v}}^R\right)$, de sorte qu'un estimateur réconcilié restreint de $\bar{Y}_i$ existe toujours.

3.4 Estimateur réconcilié de You-Rao restreint

Nous avons démontré à la section 2.2 qu'on peut obtenir des estimateurs YR de $\boldsymbol{\beta}$ et $\mathbf{v}$ comme une solution aux équations d'un modèle mixte obtenues en minimisant la fonction pondérée de l'échantillon ϕ_w donnée par l'équation (2.14). Autrement dit, nous avons démontré que, en définissant une fonction ϕ_w avec des poids $\{w_{ij}\}$, $i = 1, \ldots, m;\ j \in s_i$ et $\{\omega_i\}$, $i = 1, \ldots, m,$ puis en minimisant ϕ_w, nous obtenons les mêmes estimateurs que ceux donnés par la procédure de You et Rao (2002). Nous minimisons maintenant la fonction ϕ_w sous la contrainte de réconciliation donnée par l'équation (3.11). Il en résulte un estimateur YR restreint qui est réconcilié à $\hat{Y}^{\text{GREG}}$.

La minimisation de ϕ_w compte tenu de la restriction de réconciliation (3.11) permet d'obtenir des estimateurs de $\bar{Y}_i$, $i = 1, \ldots, m$ dont la réconciliation est garantie pour des poids quelconques qui définissent la fonction ϕ_w. Ainsi, on peut choisir un ensemble de poids w_{ij} quelconque dans ϕ_w. Dans une étude par simulations limitée fondée sur un plan de sondage, nous avons comparé trois estimateurs YR restreints en fonction de trois options pour w_{ij} : i. $w_{ij} = w_{ij}^{\text{GREG}} - 1$; ii. $w_{ij} = w_{ij}^{\text{GREG}}$; et iii. $w_{ij} = d_{ij}$. Nous n'avons constaté aucune différence significative entre ces trois estimateurs pour ce qui est de l'erreur quadratique moyenne du plan. Compte tenu de ce dernier point et étant donné que les estimateurs YR réconciliés non restreints décrits à la section 3.2 reposaient sur $w_{ij} = w_{ij}^{\text{GREG}} - 1,$ nous avons choisi de définir l'estimateur YR restreint en fonction de ces poids.

Soit ϕ_w défini en fonction de $w_{ij} = w_{ij}^{\text{GREG}} - 1$ et $\omega_i = \sum_{j \in s_i} w_{ij}^2 \Big/ \sum_{j \in s_i} w_{ij}$. La minimisation de ϕ_w en ce qui concerne $(\boldsymbol{\beta}, \mathbf{v})$ sous la contrainte de réconciliation (3.11) permet d'obtenir les estimateurs YR restreints de $(\boldsymbol{\beta}, \mathbf{v})$, désignés $(\hat{\boldsymbol{\beta}}^{\text{RYR}}, \hat{\mathbf{v}}^{\text{RYR}})$. Ces estimateurs sont donnés par :

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{\text{RYR}} \\ \hat{\mathbf{v}}^{\text{RYR}} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}}^{\text{YR}} \\ \hat{\mathbf{v}}^{\text{YR}} \end{pmatrix} + \frac{1}{\mathbf{a}^T \hat{\mathbf{A}}_w \mathbf{a}} \hat{\mathbf{A}}_w^{-1} \mathbf{a} \left[\hat{Y}_r^{\text{GREG}} - \mathbf{a}^T \begin{pmatrix} \hat{\boldsymbol{\beta}}^{\text{YR}} \\ \hat{\mathbf{v}}^{\text{YR}} \end{pmatrix} \right], \tag{3.14}$$

où les estimateurs $(\hat{\boldsymbol{\beta}}^{\text{YR}}, \hat{\mathbf{v}}^{\text{YR}})$ sont donnés par l'équation (2.15) et $\hat{\mathbf{A}}_w$ est la version empirique de $\mathbf{A}_w$ donnée par l'équation (2.16). En utilisant $\hat{\boldsymbol{\beta}}^{\text{RYR}}$ et $\hat{v}_i^{\text{RYR}}$ de $\hat{\mathbf{v}}^{\text{RYR}} = (\hat{v}_1^{\text{RYR}}, \ldots, \hat{v}_m^{\text{RYR}})^T$, les estimations YR restreintes $\hat{y}_{ij}^{\text{RYR}} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}^{\text{RYR}} + \hat{v}_i^{\text{RYR}}$ de y_{ij} non observé pour $j \in r_i$ sont ensuite utilisées pour calculer un estimateur YR restreint réconcilié :

$$\hat{\bar{Y}}_{ib}^{\text{RYR}} = \frac{1}{N_i} \left[\sum_{j \in s_i} y_{ij} + \mathbf{x}_{ir}^T \hat{\boldsymbol{\beta}}^{\text{RYR}} + (N_i - n_i)\, \hat{v}_i^{\text{RYR}} \right]. \tag{3.15}$$

Comme dans le cas de l'estimateur EBLUP restreint, les estimateurs $(\hat{\boldsymbol{\beta}}^{\text{RYR}}, \hat{\mathbf{v}}^{\text{RYR}})$ donnés par l'équation (3.14) n'existent pas si la méthode FC, le MV et le MVRE permettent d'obtenir une estimation tronquée pour σ_v^2. En conséquence, $\bar{Y}_i$ peut seulement être estimé par $\hat{\bar{Y}}_{ib}^{\text{RYR}}$ au moyen d'une méthode d'estimation des composantes de la variance qui permet toujours d'obtenir des estimations strictement positives pour σ_v^2.

Une estimation nulle de σ_v^2 ne pose pas de problème dans le calcul des estimateurs EBLUP et YR. Cependant, nous avons constaté que les estimateurs EBLUP et YR restreints ne peuvent être calculés si $\hat{\sigma}_v^2 = 0$. Pour contourner ce problème, nous utilisons une méthode proposée par Moghtased-Azar,

Tehranchi et Amiri-Simkooei (2014) qui garantit que l'estimateur de σ_v^2 sera strictement positif. Cette méthode repose sur le concept d'un maximum de vraisemblance restreint reparamétré (MVREre). L'idée est d'utiliser des fonctions dont l'intervalle est l'ensemble de tous les nombres réels positifs, à savoir des fonctions à valeur positive (FVP), pour des composantes inconnues de la variance dans le modèle stochastique plutôt que d'utiliser les composantes de la variance elles-mêmes. Leurs résultats numériques ont démontré l'estimation réussie sous contrainte de non-négativité des composantes de la variance (comme des valeurs positives) ainsi que des composantes de la covariance (comme des valeurs négatives ou positives).

Nous avons utilisé un algorithme de cotation de Fisher pour obtenir de manière itérative des estimations par le MVREre des composantes de la variance du modèle de base au niveau de l'unité donné par l'équation (2.2) (voir l'annexe B pour plus de détails). Nous avons également effectué une petite simulation et constaté que, pour des tailles d'échantillon de petits domaines égales ou supérieures à 3, l'algorithme de cotation de Fisher convergeait en moins de 15 itérations. Lorsque nous avons uniquement considéré les échantillons qui permettaient de produire une estimation nulle $\hat{\sigma}_v^2 = 0,$ nous avons observé que l'algorithme convergeait encore plus rapidement (voir la figure 4.1 à la section 4).

4. Étude par simulations

Nous présentons dans cette section les résultats d'une étude par simulations fondée sur un plan de sondage, car une telle étude rejoint les mesures qui sont calculées par les organismes nationaux de statistique. Une étude fondée sur un plan de sondage en est une dans laquelle une population finie fixe est d'abord générée au moyen d'un modèle hypothétique puis, pour chaque exécution de la simulation, un échantillon est prélevé à partir de la population finie fixe. Le but de l'étude par simulations est d'évaluer les propriétés des estimateurs réconciliés décrits à la section 3 quant au biais sous le plan et à l'erreur quadratique moyenne du plan. Nous avons examiné deux scénarios : le *Scénario 1* correspond au cas d'une modélisation exacte, tandis que le *Scénario 2* correspond au cas d'une modélisation inexacte. Des diagnostics de modélisation comme ceux présentés dans Rao et Molina (2015, pages 114-118) peuvent être utilisés pour vérifier si les modèles sont exacts ou non. De tels diagnostics de modélisation comprennent l'analyse de résidus, pour détecter les écarts par rapport au modèle hypothétique, et les diagnostics de sélection des variables auxiliaires pour le modèle et de suppression de cas, pour détecter les observations influentes.

4.1 Configuration de la simulation pour générer les populations finies

Pour chaque scénario, nous avons considéré cinq populations. Chaque population comptait $m = 30$ petits domaines, $N_i = 100$ unités de population se trouvant dans chaque petit domaine. Les populations correspondant au *Scénario 1* ont été créées au moyen du modèle suivant

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + v_i + e_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, N_i, \tag{4.1}$$

où $\beta_0 = 10$ et $\beta_1 = 5$. Pour générer les populations dans le *Scénario 2*, nous avons divisé 30 petits domaines en trois groupes égaux de petites domaines, désignés G_ℓ, pour $\ell = 1, 2, 3$. Le premier groupe G_1 contient les domaines $i = 1, \ldots, 10$, le deuxième groupe G_2 contient les domaines $i = 11, \ldots, 20$ et le troisième groupe G_3 contient les domaines $i = 21, \ldots, 30$. Le modèle à l'intérieur d'un groupe donné est donné par

$$y_{ij} = \beta_{0,\ell} + x_{ij}\beta_{1,\ell} + v_i + e_{ij},\ i \in G_\ell;\ j = 1, \ldots, N_i, \tag{4.2}$$

où $(\beta_{0,1} = 10,\ \beta_{1,1} = 1)$ pour les domaines $i \in G_1$, $(\beta_{0,2} = 20,\ \beta_{1,2} = 5)$ pour les domaines $i \in G_2$ et $(\beta_{0,3} = 30,\ \beta_{1,3} = 10)$ pour les domaines $i \in G_3$. Les équations (4.1) et (4.2) utilisent toutes deux la variable auxiliaire $\mathbf{x}_{ij} = (1,\ x_{ij})^T$ dont les valeurs x_{ij}, $j = 1, \ldots, N_i$ ont été générées à partir d'une loi exponentielle de moyenne égale à 5 et de variance égale à 25.

Les composantes aléatoires dans les équations (4.1) et (4.2) ont été générées à partir des lois normales $v_i \sim N(0,\ \sigma_v^2)$ et $e_{ij} \sim N(0,\ \sigma_e^2)$. Les cinq populations correspondant au *Scénario 1*, désignées A1, B1, C1, D1 et E1, ont été générées d'après l'équation (4.1) et les doublets de paramètres de variance suivants : i. $(\sigma_{Av}^2 = 0,2,\ \sigma_e^2 = 20)$ pour la population A1; ii. $(\sigma_{Bv}^2 = 1,\ \sigma_e^2 = 20)$ pour la population B1; iii. $(\sigma_{Cv}^2 = 2,\ \sigma_e^2 = 20)$ pour la population C1; iv. $(\sigma_{Dv}^2 = 4,\ \sigma_e^2 = 20)$ pour la population D1; et $(\sigma_{Ev}^2 = 20,\ \sigma_e^2 = 20)$ pour la population E1. Il convient de souligner que, pour les populations A1 à E1, la valeur de σ_e^2 demeure fixe, tandis que les valeurs de σ_v^2 varient. Les σ_v^2 sont choisis pour obtenir les ratios de variance $\delta = \sigma_v^2 / \sigma_e^2$ à 0,01; 0,05; 0,1; 0,2 et 1. Les cinq populations dans le *Scénario 2*, désignées A2, B2, C2, D2 et E2, ont été générées d'après l'équation (4.2) avec les mêmes doublets de paramètres de variance que pour le *Scénario 1*.

Un plan d'échantillonnage stratifié a été utilisé en prélevant des échantillons probabilistes indépendants proportionnels à des tailles d'échantillon (PPT) de taille n_i dans le i^{e} petit domaine. Les tailles d'échantillon des petits domaines ont été prises $n_i = 3$ pour $i = 1, \ldots, m$. Les probabilités de sélection ont été calculées comme $p_{ij} = b_{ij} \Big/ \sum_{j=1}^{N_i} b_{ij}$, où les mesures de taille sont $b_{ij} = x_{ij}$. Nous avons utilisé l'échantillonnage de Poisson conditionnel (EPC) pour sélectionner les échantillons PPT dans chaque petit domaine (voir Tillé (2006), chapitre 5). Les poids de sondage de base sont donnés par $d_{ij} = 1/(n_i p_{ij})$.

Dans le *Scénario 1*, nous avons ajusté le modèle de régression à erreurs emboîtées (4.1) et sa version augmentée en fonction des données de l'échantillonnage PPT sélectionnées à partir d'une des cinq populations générées au moyen du modèle (4.1). Ce scénario représente une modélisation exacte, car le modèle ajusté et le modèle utilisé pour générer la population finie coïncident. Dans le *Scénario 2*, nous avons ajusté le modèle de régression à erreurs emboîtées (4.1) et sa version augmentée en fonction des données de l'échantillonnage PPT sélectionnées à partir d'une des cinq populations générées au moyen du modèle (4.2). Ce scénario représente une modélisation inexacte, car le modèle ajusté et le modèle utilisé pour générer la population finie ne coïncident pas.

Nous avons sélectionné $G =$ 30 000 échantillons PPT stratifiés à partir de chacune des dix populations finies : les populations A1 à E1 correspondant au *Scénario 1* et les populations A2 à E2 correspondant au *Scénario 2*. Pour $g = 1, \ldots, G$, soient $\left(\hat{\sigma}_v^{2\mathrm{RE}(g)}, \hat{\sigma}_e^{2\mathrm{RE}(g)}\right)$ et $\left(\hat{\sigma}_v^{2\mathrm{reRE}(g)}, \hat{\sigma}_e^{2\mathrm{reRE}(g)}\right)$ désignant respectivement les estimations de $\left(\sigma_v^2, \sigma_e^2\right)$ obtenues au moyen de la méthode du MVRE tronqué et de sa version reparamétrée, qui correspond au g^{e} échantillon. Les valeurs de départ dans l'équation (B.2) étaient $\alpha_1^{(0)} = \log\left(0{,}1 + \hat{\sigma}_v^{2\mathrm{RE}(g)}\right)$ et $\alpha_2^{(0)} = \log\left(\hat{\sigma}_e^{2\mathrm{RE}(g)}\right)$. L'équation (B.2) a atteint la convergence en moins de 15 itérations pour l'ensemble des populations et dans les deux scénarios. D'après les échantillons simulés G sélectionnés dans chacune des cinq populations correspondant au *Scénario 1*, nous avons calculé la valeur Monte-Carlo de la probabilité d'obtenir une estimation du MVRE tronquée à zéro pour σ_v^2, étant donné que

$$P_{\mathrm{MC}}\left(\hat{\sigma}_v^{2\mathrm{RE}} = 0\right) = \frac{1}{G}\sum_{g=1}^{G} I\left(\hat{\sigma}_v^{2\mathrm{RE}(g)} = 0\right),$$

où $I(A)$ est une fonction indicatrice ayant la valeur 1 si la condition A est vérifiée, et la valeur 0 autrement.

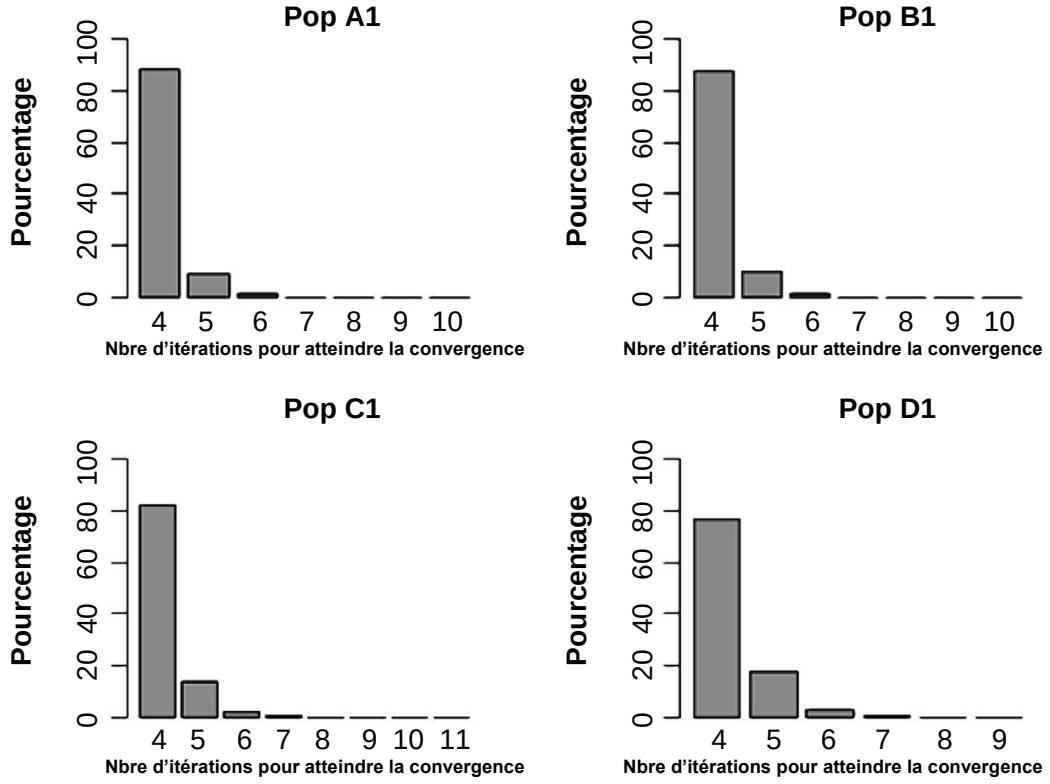
Le tableau 4.1 présente les valeurs Monte-Carlo de la probabilité d'obtenir une estimation nulle pour $\hat{\sigma}_v^{2\mathrm{RE}}$. On peut constater que la probabilité simulée $P_{\mathrm{MC}}\left(\hat{\sigma}_v^{2\mathrm{RE}} = 0\right)$ peut être aussi élevée que 0,47 pour $\delta = 0{,}01$. À mesure que δ augmente, cette probabilité empirique diminue. Le tableau 4.1 montre clairement que les estimations $\left(\hat{\sigma}_v^{2\mathrm{RE}}, \hat{\sigma}_e^{2\mathrm{RE}}\right)$ ne peuvent servir à calculer les estimateurs EBLUP et YR restreints pour des échantillons sélectionnés dans les populations A1, B1, C1 et D1.

Tableau 4.1
Valeurs de $P\left(\hat{\sigma}_v^{2\mathrm{RE}} = 0\right)$: *Scénario 1*

	Pop A1 $\delta = 0{,}01$	**Pop B1** $\delta = 0{,}05$	**Pop C1** $\delta = 0{,}1$	**Pop D1** $\delta = 0{,}2$	**Pop E1** $\delta = 1$
$P_{\mathrm{MC}}\left(\hat{\sigma}_v^{2\mathrm{RE}} = 0\right)$	0,47	0,40	0,21	0,06	0,00

La figure 4.1 présente le nombre d'itérations pour atteindre la convergence de l'algorithme de cotation de Fisher pour l'estimation $\hat{\sigma}_v^{2\mathrm{reRE}}$ de σ_v^2. L'algorithme s'arrête lorsque la valeur de $\left|\hat{\sigma}_v^{2\mathrm{reRE}(r+1)} - \hat{\sigma}_v^{2\mathrm{reRE}(r)}\right|$ est inférieure à 10^{-5}, où $\hat{\sigma}_v^{2\mathrm{reRE}(r)}$ représente la r^{e} itération calculée au moyen de l'équation (B.2) à l'annexe B. Les pourcentages présentés dans la figure 4.1 reposent uniquement sur les échantillons ayant une estimation du MVRE tronquée de σ_v^2, c'est-à-dire $\hat{\sigma}_v^{2\mathrm{RE}} = 0$. Nous avons uniquement considéré les populations A1, B1, C1 et D1, car ces quatre populations présentent des probabilités non négligeables que $\hat{\sigma}_v^{2\mathrm{RE}}$ soit nul. La figure 4.1 montre clairement que la convergence est atteinte en un maximum de 11 itérations.

Figure 4.1 Pourcentage d'itérations pour atteindre la convergence dans les échantillons avec $\hat{\sigma}_v^{2\mathrm{RE}} = 0$.



4.2 Comparaison des estimateurs réconciliés

Le but de l'étude par simulations est de comparer les estimateurs réconciliés décrits à la section 3 quant au biais sous le plan et à l'erreur quadratique moyenne du plan. Nous avons utilisé les deux scénarios, car nous voulions vérifier la protection qu'offre la réconciliation contre une modélisation inexacte. De plus, nous avons examiné la réconciliation à deux estimateurs GREG : $\hat{Y}_1^{\mathrm{GREG}}$ et $\hat{Y}_2^{\mathrm{GREG}}$. L'estimateur $\hat{Y}_1^{\mathrm{GREG}}$ a des poids donnés par l'équation (3.2) calés sur le vecteur auxiliaire $\mathbf{x}_{ij} = (1, x_{ij})$ associé au modèle pour petits domaines. Il s'ensuit que l'estimateur $\hat{Y}_1^{\mathrm{GREG}}$ correspond au cas de $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$. Le second estimateur GREG $\hat{Y}_2^{\mathrm{GREG}}$ a des poids donnés par l'équation (3.2) d'après le vecteur auxiliaire $\mathbf{x}_{ij}^* = (1, x_{ij}^*)$, où les valeurs x_{ij}^*, $j = 1, \ldots, N_i$ ont été générées à partir d'une loi exponentielle de moyenne égale à 5 et de variance égale à 25, et indépendamment des valeurs x_{ij}, $j = 1, \ldots, N_i$. Il s'ensuit que l'estimateur $\hat{Y}_2^{\mathrm{GREG}}$ correspond au cas de $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$, car la variable auxiliaire x_{ij} associée au modèle au niveau de l'unité (4.1) n'appartient pas au vecteur $\mathbf{x}_{ij}^*$ utilisé pour obtenir les poids associés à $\hat{Y}_2^{\mathrm{GREG}}$.

Pour une population finie fixe, soient $\bar{Y}_i$ la moyenne du petit domaine i et $\hat{\bar{Y}}_i$ un estimateur générique de $\bar{Y}_i$. Nous désignons $\hat{\bar{Y}}_i^{(g)}$ la valeur de $\hat{\bar{Y}}_i$ d'après le g^{e} échantillon simulé, pour $g = 1, \ldots, G$. Les estimateurs décrits à la section 3 respectent la propriété de réconciliation peu importe la méthode utilisée pour estimer les composantes de la variance. Étant donné que les estimateurs réconciliés restreints

reposent sur des estimations $\left(\hat{\sigma}_v^{2\mathrm{reRE}(g)}, \hat{\sigma}_e^{2\mathrm{reRE}(g)}\right)$, nous avons décidé d'utiliser le MVREre dans le calcul de $\hat{\bar{Y}}_i^{(g)}$ pour chaque estimateur $\hat{\bar{Y}}_i$ évalué dans cette étude par simulations.

Nous avons examiné les mesures de rendement suivantes :

Biais relatif absolu moyen

$$\overline{\mathrm{BRA}} = \frac{1}{m}\sum_{i=1}^{m}\mathrm{BRA}_i \quad \text{avec} \quad \mathrm{BRA}_i = \left|\frac{1}{G}\sum_{g=1}^{G}\frac{\hat{\bar{Y}}_i^{(g)}}{\bar{Y}_i} - 1\right|$$

Racine de l'erreur quadratique moyenne relative

$$\overline{\mathrm{REQMR}} = \frac{1}{m}\sum_{i=1}^{m}\mathrm{REQMR}_i \quad \text{avec} \quad \mathrm{REQMR}_i = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\hat{\bar{Y}}_i^{(g)}}{\bar{Y}_i} - 1\right)^2}.$$

Cette partie de la simulation est résumée dans quatre tableaux. Nous présentons les résultats séparément pour les *Scénarios 1* et *2*. Les résultats dans le cas où la réconciliation est pour $\hat{Y}_1^{\mathrm{GREG}}$ (le cas de $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$) sont résumés dans les tableaux 4.2 (*Scénario 1*) et 4.3 (*Scénario 2*). Les résultats dans le cas où la réconciliation est pour $\hat{Y}_2^{\mathrm{GREG}}$ (le cas de $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$) sont résumés dans les tableaux 4.4 (*Scénario 1*) et 4.5 (*Scénario 2*).

Réconciliation à $\hat{Y}_1^{\mathrm{GREG}}$ (le cas de $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$)

Nous avons calculé le $\overline{\mathrm{BRA}}$ et la $\overline{\mathrm{REQMR}}$ pour deux estimateurs non réconciliés, $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ et $\hat{\bar{Y}}_i^{\mathrm{YR}}$, ainsi que leurs estimateurs correspondants réconciliés à $\hat{Y}_1^{\mathrm{GREG}}$. Pour $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$, nous avons trois estimateurs réconciliés, $\hat{\bar{Y}}_{ib}^{\mathrm{EBRat}}$, $\hat{\bar{Y}}_{iab}^{\mathrm{EBLUP}}$ et $\hat{\bar{Y}}_{ib}^{\mathrm{REBLUP}}$, donnés respectivement par les équations (3.5), (3.8) et (3.13). Pour $\hat{\bar{Y}}_i^{\mathrm{YR}}$, les estimateurs réconciliés correspondants sont $\hat{\bar{Y}}_{ib}^{\mathrm{YRat}}$, $\hat{\bar{Y}}_{ib}^{\mathrm{YR}}$ et $\hat{\bar{Y}}_{ib}^{\mathrm{RYR}}$, donnés respectivement par les équations (3.5), (3.9) et (3.15).

Nous examinons d'abord leurs propriétés lorsque le modèle est exact (*Scénario 1*). En comparant les $\overline{\mathrm{BRA}}$ de l'ensemble des estimateurs au tableau 4.2, nous constatons qu'il n'y a pas de grande différence entre les estimateurs. Les estimateurs EBLUP présentent des $\overline{\mathrm{BRA}}$ un peu plus faibles que les estimateurs reposant sur la procédure YR. L'estimateur réconcilié $\hat{\bar{Y}}_{iab}^{\mathrm{EBLUP}}$ présente les $\overline{\mathrm{BRA}}$ les plus faibles, tandis que les $\overline{\mathrm{BRA}}$ des estimateurs réconciliés $\hat{\bar{Y}}_{ib}^{\mathrm{EBRat}}$ et $\hat{\bar{Y}}_{ib}^{\mathrm{REBLUP}}$ sont identiques à ceux de $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$. Les valeurs du $\overline{\mathrm{BRA}}$ associées aux estimateurs $\hat{\bar{Y}}_i^{\mathrm{YR}}$, $\hat{\bar{Y}}_{ib}^{\mathrm{YRat}}$ et $\hat{\bar{Y}}_{ib}^{\mathrm{RYR}}$ se rapprochent, tandis que l'estimateur $\hat{\bar{Y}}_{ib}^{\mathrm{YR}}$ présente un biais relatif un peu plus élevé, surtout pour des valeurs plus grandes de $\delta = \sigma_v^2 / \sigma_e^2$. Pour l'ensemble des estimateurs, les $\overline{\mathrm{BRA}}$ augmentent à mesure que δ augmente : de légères exceptions se produisent lorsque $\delta = 1$.

Ensuite, nous nous penchons sur les $\overline{\mathrm{REQMR}}$. Comme on pouvait s'y attendre, les $\overline{\mathrm{REQMR}}$ les plus faibles sont associées à $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$, tandis que l'estimateur $\hat{\bar{Y}}_i^{\mathrm{YR}}$ présente des valeurs de la $\overline{\mathrm{REQMR}}$ un peu plus grandes en raison de l'utilisation de poids de sondage sous une modélisation exacte. La réconciliation entraîne une augmentation de la $\overline{\mathrm{REQMR}}$. Il convient de souligner que les $\overline{\mathrm{REQMR}}$ des estimateurs réconciliés $\hat{\bar{Y}}_{iab}^{\mathrm{EBLUP}}$ et $\hat{\bar{Y}}_{ib}^{\mathrm{YR}}$, donnés aux sections 3.1 et 3.2 respectivement, sont plus élevées que celles

associées aux méthodes restreintes $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ et $\hat{\bar{Y}}_{ib}^{\text{RYR}}$, données aux sections 3.3 et 3.4 respectivement. Les procédures par le ratio simplistes $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ et $\hat{\bar{Y}}_{ib}^{\text{YRat}}$ présentent des $\overline{\text{REQMR}}$ qui sont assez comparables à celles réconciliées qui utilisent les méthodes restreintes. Les $\overline{\text{REQMR}}$ augmentent à mesure que δ augmente.

Nous concluons ce qui suit dans le cas de $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^{*}$ et lorsque le modèle pour petits domaines est correctement spécifié. Les estimateurs réconciliés restreints ou réconciliés par le ratio donnent de meilleurs résultats que ceux qui utilisent un modèle augmenté pour l'estimation EBLUP ou une méthode YR modifiée. Lorsque les techniques restreintes ou de réconciliation par le ratio sont utilisées, les estimateurs résultants ont des valeurs du biais semblables à celles associées à leurs versions non réconciliées, tandis que leurs valeurs de l'erreur quadratique moyenne sont légèrement plus élevées que celles des versions non réconciliées. Les estimateurs sur petits domaines et l'estimateur GREG $\hat{Y}_1^{\text{GREG}}$ reposent sur les mêmes variables auxiliaires, tandis que le modèle est exact. En conséquence, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ et $\hat{\bar{Y}}_i^{\text{YR}}$ n'ont pas à être modifiés considérablement pour réaliser la réconciliation à $\hat{Y}_1^{\text{GREG}}$.

Tableau 4.2
$\overline{\text{BRA}}$ (%) et $\overline{\text{REQMR}}$ (%) pour le *Scénario 1* : la réconciliation à $\hat{Y}_1^{\text{GREG}}\left(\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^{*}\right)$

Estimateur	Mesure	Pop A1 $\delta = 0{,}01$	Pop B1 $\delta = 0{,}05$	Pop C1 $\delta = 0{,}1$	Pop D1 $\delta = 0{,}2$	Pop E1 $\delta = 1$
$\hat{\bar{Y}}_i^{\text{EBLUP}}$	$\overline{\text{BRA}}$	1,1	1,9	2,3	2,7	2,6
	$\overline{\text{REQMR}}$	2,7	3,4	3,9	4,9	6,5
$\hat{\bar{Y}}_i^{\text{YR}}$	$\overline{\text{BRA}}$	1,2	2,0	2,4	2,9	3,1
	$\overline{\text{REQMR}}$	3,1	3,7	4,2	5,3	7,2
$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\overline{\text{BRA}}$	1,1	1,9	2,3	2,7	2,6
	$\overline{\text{REQMR}}$	3,2	3,8	4,3	5,2	6,9
$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\overline{\text{BRA}}$	1,2	2,0	2,4	2,9	3,1
	$\overline{\text{REQMR}}$	3,1	3,7	4,3	5,3	7,4
$\hat{\bar{Y}}_{iab}^{\text{EBLUP}}$	$\overline{\text{BRA}}$	1,0	1,6	2,1	2,4	2,3
	$\overline{\text{REQMR}}$	9,6	9,8	10,1	11,1	13,9
$\hat{\bar{Y}}_{ib}^{\text{YR}}$	$\overline{\text{BRA}}$	1,2	2,0	2,5	3,0	3,7
	$\overline{\text{REQMR}}$	3,5	4,8	5,4	11,7	14,5
$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\overline{\text{BRA}}$	1,1	1,9	2,3	2,7	2,6
	$\overline{\text{REQMR}}$	3,2	3,8	4,3	5,3	7,0
$\hat{\bar{Y}}_{ib}^{\text{RYR}}$	$\overline{\text{BRA}}$	1,2	2,0	2,4	2,9	3,2
	$\overline{\text{REQMR}}$	3,1	3,7	4,3	5,3	7,5

Les résultats lorsqu'un modèle exact n'est pas utilisé sont présentés au tableau 4.3. La valeur de δ n'a pas une grande incidence sur les $\overline{\text{BRA}}$ et les $\overline{\text{REQMR}}$ pour l'ensemble des estimateurs. Les $\overline{\text{BRA}}$ et les $\overline{\text{REQMR}}$ des estimateurs EBLUP, qu'ils soient réconciliés ou non, sont plus élevés que ceux associés aux estimateurs YR. Il s'ensuit que, pour une modélisation inexacte, l'utilisation des estimateurs YR est recommandée. Étant donné que $\hat{Y}_1^{\text{GREG}}$ et les estimateurs fondés sur la procédure YR utilisent le même

vecteur de données auxiliaires, il s'ensuit qu'il n'y a pas de grande différence quant au $\overline{\text{BRA}}$ et à la $\overline{\text{REQMR}}$ entre l'estimateur non réconcilié $\hat{\bar{Y}}_i^{\text{YR}}$ et ses versions réconciliées, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, $\hat{\bar{Y}}_{ib}^{\text{YR}}$ et $\hat{\bar{Y}}_{ib}^{\text{RYR}}$. Cependant, on peut constater que l'estimateur réconcilié $\hat{\bar{Y}}_{ib}^{\text{YR}}$ présente les valeurs du $\overline{\text{BRA}}$ les plus faibles, tandis que l'estimateur réconcilié restreint $\hat{\bar{Y}}_{ib}^{\text{RYR}}$ présente les $\overline{\text{REQMR}}$ les plus faibles.

Tableau 4.3
$\overline{\text{BRA}}$ (%) et $\overline{\text{REQMR}}$ (%) pour le *Scénario 2* : la réconciliation à $\hat{Y}_1^{\text{GREG}}$ $\left(\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*\right)$

Estimateur	Mesure	Pop A2 $\delta = 0{,}01$	Pop B2 $\delta = 0{,}05$	Pop C2 $\delta = 0{,}1$	Pop D2 $\delta = 0{,}2$	Pop E2 $\delta = 1$
$\hat{\bar{Y}}_i^{\text{EBLUP}}$	$\overline{\text{BRA}}$	42,3	42,7	43,2	43,0	41,5
	$\overline{\text{REQMR}}$	59,8	60,5	61,1	60,6	59,0
$\hat{\bar{Y}}_i^{\text{YR}}$	$\overline{\text{BRA}}$	13,5	13,8	13,8	13,6	13,5
	$\overline{\text{REQMR}}$	42,8	43,2	43,5	43,2	42,4
$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\overline{\text{BRA}}$	42,9	43,4	43,9	43,6	42,1
	$\overline{\text{REQMR}}$	61,2	61,9	62,7	62,1	60,3
$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\overline{\text{BRA}}$	13,8	14,1	14,1	13,9	13,8
	$\overline{\text{REQMR}}$	43,9	44,4	44,7	44,4	43,5
$\hat{\bar{Y}}_{iab}^{\text{EBLUP}}$	$\overline{\text{BRA}}$	19,8	20,2	20,2	20,2	19,6
	$\overline{\text{REQMR}}$	66,2	66,7	67,6	67,3	66,6
$\hat{\bar{Y}}_{ib}^{\text{YR}}$	$\overline{\text{BRA}}$	10,9	10,6	11,5	12,5	10,7
	$\overline{\text{REQMR}}$	47,3	47,6	48,1	47,9	47,8
$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\overline{\text{BRA}}$	41,2	41,8	41,8	41,7	40,6
	$\overline{\text{REQMR}}$	58,2	59,0	59,1	58,9	57,4
$\hat{\bar{Y}}_{ib}^{\text{RYR}}$	$\overline{\text{BRA}}$	12,5	12,7	12,6	12,5	12,5
	$\overline{\text{REQMR}}$	42,4	42,9	43,1	42,9	42,1

Réconciliation à $\hat{Y}_2^{\text{GREG}}$ (le cas de $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$)

Les résultats dans ce cas sont présentés aux tableaux 4.4 et 4.5 pour les *Scénarios 1* et *2*, respectivement. La pondération se rapporte à l'estimateur w_{ij}^{GREG} donné par l'équation (3.2). Nous avons examiné les quatre estimateurs suivants ($\hat{\bar{Y}}_{ib}^{\text{EBRat}}$, $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$ et $\hat{\bar{Y}}_{ib}^{\text{RYR}}$) qui sont réconciliés à $\hat{Y}_2^{\text{GREG}}$. Les deux premiers estimateurs, $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ et $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, sont donnés par les équations (3.5) et (3.13) respectivement, tandis que les deux derniers, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$ et $\hat{\bar{Y}}_{ib}^{\text{RYR}}$, sont donnés par les équations (3.5) et (3.15).

Au tableau 4.4, nous résumons les valeurs moyennes du BRA et de la REQMR lorsque le modèle est exact. Autrement dit, tant l'échantillon que les données sur la population respectent le modèle (4.1). Nous discutons en premier lieu de leurs propriétés quant aux $\overline{\text{BRA}}$. En comparant les $\overline{\text{BRA}}$ de l'ensemble des estimateurs au tableau 4.4, nous constatons encore une fois que, sous une modélisation exacte, l'estimateur EBLUP original, $\hat{\bar{Y}}_i^{\text{EBLUP}}$, présente les $\overline{\text{BRA}}$ les plus faibles. Les $\overline{\text{BRA}}$ augmentent lorsque la réconciliation est nécessaire, ce qui est différent de ce que nous avons constaté au tableau 4.2. Il n'y a pas de grande différence quant au $\overline{\text{BRA}}$ entre les estimateurs réconciliés obtenus au moyen des méthodes

fondées sur des ajustements par le ratio, $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ et $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, et ceux obtenus au moyen des méthodes restreintes, $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ et $\hat{\bar{Y}}_{ib}^{\text{RYR}}$. Les $\overline{\text{BRA}}$ augmentent à mesure que δ augmente : de légères exceptions se produisent lorsque $\delta = 1$.

Ensuite, nous nous penchons sur les $\overline{\text{REQMR}}$. Comme on pouvait s'y attendre, les $\overline{\text{REQMR}}$ les plus faibles sont associées à $\hat{\bar{Y}}_{i}^{\text{EBLUP}}$, ce qui est optimal sous une modélisation exacte. La réconciliation entraîne une augmentation de la $\overline{\text{REQMR}}$. Il convient de souligner que les $\overline{\text{REQMR}}$ associées à l'ensemble des quatre procédures de réconciliation au tableau 4.4 sont assez élevées comparativement aux $\overline{\text{REQMR}}$ associées aux estimateurs non réconciliés $\hat{\bar{Y}}_{i}^{\text{EBLUP}}$ et $\hat{\bar{Y}}_{i}^{\text{YR}}$. Les estimateurs $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ et $\hat{\bar{Y}}_{ib}^{\text{YRat}}$ ont une efficacité semblable, tandis que $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ et $\hat{\bar{Y}}_{ib}^{\text{RYR}}$ présentent des valeurs de la $\overline{\text{REQMR}}$ qui sont un peu plus élevées que celles de $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ et $\hat{\bar{Y}}_{ib}^{\text{YRat}}$. Les $\overline{\text{REQMR}}$ augmentent à mesure que δ augmente.

Lorsque $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^{*}$, il existe des différences plus grandes entre les estimateurs sur petits domaines fondés sur le modèle (2.2), qui utilise le vecteur $\mathbf{x}_{ij}$, et l'estimateur GREG, qui utilise $\mathbf{x}_{ij}^{*}$. Il convient de souligner que nous avons examiné une situation un peu extrême alors que $\mathbf{x}_{ij}$ et $\mathbf{x}_{ij}^{*}$ n'ont pas de variable en commun. Il s'ensuit que les modifications nécessaires pour obtenir des estimateurs réconciliés sont plus importantes dans ce cas comparativement au cas de $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^{*}$. Cela explique pourquoi, au tableau 4.4, les estimateurs réconciliés présentent des valeurs du $\overline{\text{BRA}}$ et de la $\overline{\text{REQMR}}$ significativement plus élevées que les estimateurs qui ne sont pas réconciliés à $\hat{Y}_{2}^{\text{GREG}}$.

Tableau 4.4
$\overline{\text{BRA}}$ (%) et $\overline{\text{REQMR}}$ (%) pour le *Scénario 1* : la réconciliation à $\hat{Y}_{2}^{\text{GREG}}\left(\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^{*}\right)$

Estimateur	Mesure	Pop A1 $\delta = 0{,}01$	Pop B1 $\delta = 0{,}05$	Pop C1 $\delta = 0{,}1$	Pop D1 $\delta = 0{,}2$	Pop E1 $\delta = 1$
$\hat{\bar{Y}}_{i}^{\text{EBLUP}}$	$\overline{\text{BRA}}$	1,1	1,9	2,3	2,7	2,6
	$\overline{\text{REQMR}}$	2,7	3,4	3,9	4,9	6,5
$\hat{\bar{Y}}_{i}^{\text{YR}}$	$\overline{\text{BRA}}$	1,2	2,0	2,4	2,9	3,1
	$\overline{\text{REQMR}}$	3,1	3,7	4,2	5,3	7,2
$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\overline{\text{BRA}}$	4,2	4,3	4,5	4,9	4,6
	$\overline{\text{REQMR}}$	13,0	13,2	13,5	14,0	14,6
$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\overline{\text{BRA}}$	4,2	4,3	4,5	5,0	4,8
	$\overline{\text{REQMR}}$	13,0	13,2	13,5	14,0	14,0
$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\overline{\text{BRA}}$	4,2	4,3	4,5	5,0	4,8
	$\overline{\text{REQMR}}$	13,1	13,3	13,5	14,1	15,0
$\hat{\bar{Y}}_{ib}^{\text{RYR}}$	$\overline{\text{BRA}}$	4,2	4,3	4,6	5,1	5,0
	$\overline{\text{REQMR}}$	13,5	13,7	13,8	14,5	16,2

L'incidence d'utiliser un modèle inexact est présentée au tableau 4.5. Nous constatons que l'estimateur $\hat{\bar{Y}}_{i}^{\text{EBLUP}}$ est le plus touché en ce qui a trait au $\overline{\text{BRA}}$ et à la $\overline{\text{REQMR}}$, car la procédure EBLUP suppose que le modèle est exact. Les versions réconciliées de l'estimateur EBLUP, $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ et $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, présentent

aussi des $\overline{\text{BRA}}$ et des $\overline{\text{REQMR}}$ élevés. Bien que l'estimateur original de You et Rao (2002), $\hat{\bar{Y}}_i^{\text{YR}}$, présente un $\overline{\text{BRA}}$ beaucoup plus faible que l'estimateur EBLUP, sa $\overline{\text{REQMR}}$ est assez élevée. Le $\overline{\text{BRA}}$ et la $\overline{\text{REQMR}}$ associés à la version réconciliée par le ratio de $\hat{\bar{Y}}_i^{\text{YR}}$, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, sont un peu plus élevés que ceux associés à $\hat{\bar{Y}}_i^{\text{YR}}$. L'estimateur YR réconcilié, $\hat{\bar{Y}}_{ib}^{\text{RYR}}$, qui repose sur la procédure restreinte donnée à la section 3.4, présente le $\overline{\text{BRA}}$ le plus faible parmi les estimateurs présentés au tableau 4.5. En raison de la réconciliation, sa $\overline{\text{REQMR}}$ est légèrement plus élevée que celle associée à $\hat{\bar{Y}}_i^{\text{YR}}$.

Tableau 4.5
$\overline{\text{BRA}}$ (%) et $\overline{\text{REQMR}}$ (%) pour le *Scénario 2* : la réconciliation à $\hat{Y}_2^{\text{GREG}}\left(\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*\right)$

Estimateur	Mesure	Pop A2 $\delta = 0{,}01$	Pop B2 $\delta = 0{,}05$	Pop C2 $\delta = 0{,}1$	Pop D2 $\delta = 0{,}2$	Pop E2 $\delta = 1$
$\hat{\bar{Y}}_i^{\text{EBLUP}}$	$\overline{\text{BRA}}$	42,3	42,6	43,2	43,0	41,6
	$\overline{\text{REQMR}}$	59,8	60,4	61,1	60,7	59,1
$\hat{\bar{Y}}_i^{\text{YR}}$	$\overline{\text{BRA}}$	13,6	13,6	13,9	13,7	13,5
	$\overline{\text{REQMR}}$	42,8	43,1	43,5	43,3	42,4
$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\overline{\text{BRA}}$	43,8	44,4	44,9	44,6	43,3
	$\overline{\text{REQMR}}$	65,4	66,1	67,0	66,4	64,5
$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\overline{\text{BRA}}$	15,0	15,2	15,6	15,2	14,9
	$\overline{\text{REQMR}}$	47,9	48,2	48,7	48,3	47,3
$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\overline{\text{BRA}}$	37,3	38,0	38,1	37,8	37,1
	$\overline{\text{REQMR}}$	57,4	58,2	58,5	58,2	56,7
$\hat{\bar{Y}}_{ib}^{\text{RYR}}$	$\overline{\text{BRA}}$	9,9	10,1	10,4	10,0	10,1
	$\overline{\text{REQMR}}$	43,4	43,8	44,2	43,9	43,1

5. Exemple fondé sur des données réelles

Dans cette section, nous comparons les estimateurs réconciliés dans une analyse fondée sur des données réelles. L'ensemble de données étudié est celui présenté par Battese et coll. (1988), dans le cadre d'une étude visant à estimer le nombre moyen d'hectares consacrés à la culture du maïs et du soja par segment dans douze comtés du Centre-Nord de l'Iowa. La variable de réponse y_{ij} est le nombre d'hectares consacrés à la culture du maïs dans le j^{e} segment du i^{e} comté. Les variables auxiliaires, x_{1ij} et x_{2ij}, représentent le nombre de pixels classés comme étant du maïs ou du soja, respectivement, dans le j^{e} segment du i^{e} comté. Nous présentons uniquement les résultats pour $\bar{Y}_i$, le nombre moyen d'hectares consacrés à la culture du maïs par segment pour le comté i.

Suivant la méthode de Battese et coll. (1988), nous avons supprimé les données d'échantillon du deuxième segment échantillonné dans le comté de Hardin, car la superficie consacrée à la culture du maïs pour ce segment semblait erronée. Parmi les douze comtés, trois comtés ne comportaient qu'un seul segment échantillonné. Suivant la méthode de Prasad et Rao (1990), nous avons regroupé ces trois comtés en un seul, donnant lieu à un ensemble de données comprenant 10 comtés dont la taille d'échantillon n_i

variait de 2 à 5 dans chaque comté. Le nombre total de segments N_i (taille de la population) dans chaque comté variait de 402 à 1 505. Suivant la méthode de You et Rao (2002), nous avons supposé un échantillonnage aléatoire simple dans chaque comté, et le poids de sondage de base a été calculé comme $d_{ij} = N_i / n_i$ pour l'unité j dans le i^{e} comté.

Nous fondons nos calculs sur le modèle d'échantillonnage au niveau de l'unité donné par

$$y_{ij} = \beta_0 + x_{1ij}\beta_1 + x_{2ij}\beta_2 + v_i + e_{ij},\ j = 1, \ldots, n_i;\ i = 1, \ldots, 10, \tag{5.1}$$

où v_i et e_{ij} sont des erreurs normalement distribuées à variances communes σ_v^2 et σ_e^2. Nous avons ajusté le modèle (5.1) en fonction des données d'échantillon pour obtenir des estimations par l'estimateur MPE/EBE de $\boldsymbol{\beta}$ et v_i, désignées $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)^T$ et $\hat{v}_i$, et reparamétré les estimations par le MVRE des composantes de la variance, désignées $\left(\hat{\sigma}_v^{2\text{reRE}}, \hat{\sigma}_e^{2\text{reRE}}\right)$. Les estimations EBLUP des effets fixes du modèle sont de $\hat{\beta}_0 = 58{,}5$; $\hat{\beta}_1 = 0{,}316$ et $\hat{\beta}_2 = -0{,}150$, tandis que les estimations par le MVREre des composantes de la variance sont de $\hat{\sigma}_v^{2\text{reRE}} = 135{,}6$ et $\hat{\sigma}_e^{2\text{reRE}} = 155{,}9$. Le δ estimé est de 0,869, ce qui se rapproche de 1. Pour chaque unité dans l'échantillon, nous avons répété le vecteur $\mathbf{x}_{ij} = \left(1, x_{1ij}, x_{2ij}\right)^T$ plusieurs fois égales à $\left[d_{ij}\right]$, le nombre entier le plus proche du poids d'échantillonnage $d_{ij} = N_i / n_i$. Ainsi, nous avons obtenu une pseudo-population de valeurs x, désignée $\mathbf{x}_{ij}^{\text{ps}} = \left(1, x_{1ij}^{\text{ps}}, x_{2ij}^{\text{ps}}\right)^T$, la taille de la population du comté étant égale à $N_i^{\text{ps}} = n_i\left[N_i / n_i\right]$. Les valeurs y de notre pseudo-population, désignée y_{ij}^{ps}, sont définies comme suit : $y_{ij}^{\text{ps}} = y_{ij}$ pour $j \in s_i$, et $y_{ij}^{\text{ps}} = \hat{\beta}_0 + x_{1ij}^{\text{ps}}\hat{\beta}_1 + x_{2ij}^{\text{ps}}\hat{\beta}_2 + \hat{v}_i + e_{ij}^{\text{ps}}$ pour $j \in r_i^{\text{ps}}$, où $e_{ij}^{\text{ps}} \sim N\left(0, \hat{\sigma}_e^{2\text{reRE}}\right)$ et r_i^{ps} se compose des unités non observées $N_i^{\text{ps}} - n_i$ dans le i^{e} petit domaine. Prasad et Rao (1990) ont utilisé une procédure semblable pour générer une pseudo-population comptant un plus grand nombre de comtés que l'ensemble de données présenté par Battese et coll. (1988). Leur pseudo-population composée de vingt comtés a été obtenue en deux étapes : en premier lieu, les valeurs des variables auxiliaires associées à l'ensemble de données original ont été reproduites; ensuite, les valeurs de la variable de réponse ont été calculées à partir du modèle, en utilisant les valeurs x reproduites et les estimations des paramètres du modèle.

Soient $\bar{Y}_i = \left(N_i^{\text{ps}}\right)^{-1} \sum_{j=1}^{N_i^{\text{ps}}} y_{ij}^{\text{ps}}$ et $Y = \sum_{i=1}^{10} N_i^{\text{ps}} \bar{Y}_i$ respectivement la moyenne du i^{e} petit domaine et le total de la pseudo-population. Au niveau de la population, nous estimons Y au moyen de l'estimateur GREG $\hat{Y}^{\text{GREG}}$ d'après les poids donnés par l'équation (3.2), où le vecteur $\mathbf{x}_{ij}^{\text{ps}*}$ est le vecteur bidimensionnel $\mathbf{x}_{ij}^{\text{ps}*} = \left(1, x_{1ij}^{\text{ps}}\right)^T$. Il s'ensuit que $\mathbf{x}_{ij}^{\text{ps}} \not\subset \mathbf{x}_{ij}^{\text{ps}*}$, étant donné que $\mathbf{x}_{ij}^{\text{ps}} = \left(1, x_{1ij}^{\text{ps}}, x_{2ij}^{\text{ps}}\right)^T$ et $\mathbf{x}_{ij}^{\text{ps}*} = \left(1, x_{1ij}^{\text{ps}}\right)^T$.

À partir de la pseudo-population $\left(y_{ij}^{\text{ps}}, \mathbf{x}_{ij}^{\text{ps}}\right)$, $j = 1, \ldots, N_i^{\text{ps}}$; $i = 1, \ldots, 10$, nous avons prélevé $G =$ 30 000 échantillons aléatoires simples stratifiés sans remise de taille n_i et traité chaque comté comme une strate. Ces tailles d'échantillon étaient égales à celles de l'ensemble de données original. Nous avons utilisé le biais relatif (BR) sous le plan et l'erreur quadratique moyenne (REQMR) pour évaluer le rendement de six estimateurs : deux estimateurs non réconciliés, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ et $\hat{\bar{Y}}_i^{\text{YR}}$, et quatre estimateurs

réconciliés, $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ et $\hat{\bar{Y}}_{ib}^{\text{RYR}}$, qu'on peut calculer dans le cas de $\mathbf{x}_{ij}^{\text{ps}} \not\subset \mathbf{x}_{ij}^{\text{ps*}}$. Soient $\hat{\bar{Y}}_i$ un estimateur générique de la moyenne du i^{e} petit domaine $\bar{Y}_i$ et $\hat{\bar{Y}}_i^{(g)}$ sa valeur associée au g^{e} échantillon, pour $g = 1, \ldots, G$. Les valeurs de son BR et de sa REQMR sont données par

$$\text{BR}_i = \frac{1}{G}\sum_{g=1}^{G}\frac{\hat{\bar{Y}}_i^{(g)}}{\bar{Y}_i} - 1 \quad \text{et} \quad \text{REQMR}_i = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\hat{\bar{Y}}_i^{(g)}}{\bar{Y}_i} - 1\right)^2}.$$

Le tableau 5.1 présente le BR sous le plan et la REQMR des six estimateurs de $\bar{Y}_i$ pour les dix comtés de la pseudo-population. Dans cet exemple, nous constatons que les BR et les REQMR sont assez semblables pour l'ensemble des estimateurs et des tailles d'échantillon, parce que le modèle ayant permis de générer les données sur la population est exact, tandis que le modèle pour petits domaines et l'estimateur GREG ont tous deux en commun la variable auxiliaire égale au nombre de pixels classés comme étant du maïs.

Tableau 5.1
BR (%) et REQMR (%) : la réconciliation à $\hat{Y}^{\text{GREG}}\left(\mathbf{x}_{ij}^{\text{ps}} \not\subset \mathbf{x}_{ij}^{\text{ps*}}\right)$

Comté	n_i	Mesure	$\hat{\bar{Y}}_i^{\text{EBLUP}}$	$\hat{\bar{Y}}_i^{\text{YR}}$	$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\hat{\bar{Y}}_{ib}^{\text{RYR}}$
Cerro Hamilton Worth	3	BR	1,6	1,4	1,3	1,3	1,0	1,2
		REQMR	5,2	5,4	5,3	5,4	5,6	5,4
Humboldt	2	BR	2,0	1,9	1,7	1,8	1,8	1,8
		REQMR	4,5	4,5	4,5	4,5	4,4	4,5
Franklin	3	BR	-3,3	-3,4	-3,5	-3,5	-3,5	-3,5
		REQMR	5,2	5,4	5,5	5,5	5,4	5,4
Pocahontas	3	BR	-3,1	-3,4	-3,4	-3,5	-3,3	-3,5
		REQMR	6,2	6,5	6,4	6,6	6,4	6,6
Winnebago	3	BR	2,6	2,3	2,3	2,2	2,3	2,2
		REQMR	5,4	5,3	5,3	5,3	5,3	5,2
Wright	3	BR	-0,4	-0,6	-0,7	-0,7	-0,6	-0,6
		REQMR	3,7	3,8	3,9	3,9	3,8	3,9
Webster	4	BR	-2,6	-2,9	-2,9	-3,0	-2,8	-2,9
		REQMR	5,2	5,4	5,5	5,5	5,4	5,5
Hancock	5	BR	0,9	0,7	0,6	0,6	0,8	0,7
		REQMR	4,2	4,1	4,2	4,2	4,2	4,2
Kossuth	5	BR	3,5	3,3	3,2	3,2	3,2	3,2
		REQMR	5,9	5,8	5,8	5,8	5,8	5,8
Hardin	5	BR	-1,5	-1,7	-1,8	-1,8	-1,7	-1,8
		REQMR	4,2	4,3	4,4	4,5	4,3	4,4

6. Conclusion

En général, la somme des estimations sur petits domaines fondées sur un modèle n'est pas égale à une estimation directe obtenue pour l'ensemble de ces petits domaines. Le poids qui est associé à l'estimateur direct peut être le poids d'échantillonnage ou un poids obtenu en utilisant l'estimateur GREG. Les

données auxiliaires qui sont utilisées pour obtenir les estimations GREG et les estimations sur petits domaines au niveau de l'unité peuvent ne pas nécessairement coïncider. Dans le présent document, nous avons proposé plusieurs procédures de réconciliation pour deux estimateurs sur petits domaines bien connus (EBLUP et YR) qui reposent sur le modèle au niveau de l'unité. Nous avons examiné le cas où les taux d'échantillonnage ne sont pas négligeables et le plan d'échantillonnage est ignorable. Dans l'éventualité où l'on considèrerait que le plan d'échantillonnage n'est pas ignorable pour certaines des questions d'enquête, le vecteur de données auxiliaires $\mathbf{x}_{ij}$ dans le modèle (2.2) pourrait être augmenté en intégrant une autre variable spécifiée g_{ij} fonction des poids de sondage pour compenser le biais potentiel des estimateurs EBLUP ou YR. Verret et coll. (2015) ont proposé un certain nombre d'options pour g_{ij} qui comprenaient le poids de sondage w_{ij}. Dans le cas de l'estimateur EBLUP, la réconciliation est réalisée en ajoutant la variable $q_{ij} = w_{ij}^{\text{GREG}} - 1$. Étant donné que q_{ij} doit présenter une forte corrélation avec w_{ij}, la procédure proposée pour réconcilier l'estimateur EBLUP doit offrir une bonne protection contre un possible plan d'échantillonnage non ignorable. Les simulations présentées dans Verret et coll. (2015) ont montré que la procédure YR, en elle-même, offre également une bonne protection contre un possible plan d'échantillonnage non ignorable. Leurs simulations ont également montré qu'on peut obtenir une protection supplémentaire en définissant g_{ij} comme étant égal à $n_i w_{ij}$.

Nous avons étendu les procédures de réconciliation décrites dans Stefan et Hidiroglou (2020) au cas de taux d'échantillonnage non négligeables dans chaque petit domaine. Ces procédures reposent sur des estimateurs qui ont à l'origine été élaborés par Battese et coll. (1988) (estimateur EBLUP) et You et Rao (2002) (estimateur YR) lorsque les taux d'échantillonnage dans chaque petit domaine sont négligeables. Ugarte et coll. (2009) ont proposé un estimateur réconcilié différent qui est un estimateur EBLUP restreint. Nous avons étendu la procédure décrite dans Ugarte et coll. (2009) pour obtenir un estimateur réconcilié qui intègre les poids de sondage et qui est essentiellement un estimateur YR restreint. Nous avons également examiné deux estimateurs réconciliés fondés sur des ajustements par le ratio simples appliqués aux estimateurs EBLUP et YR respectivement. Nous avons réalisé une étude par simulations pour comparer les propriétés de ces six estimateurs réconciliés.

Si les données auxiliaires utilisées pour estimer les moyennes de petits domaines sont les mêmes que celles utilisées dans l'estimateur GREG, et si le modèle est exact, la procédure restreinte décrite dans Ugarte et coll. (2009) et l'estimateur EBLUP ajusté par le ratio présenteront les plus faibles $\overline{\text{BRA}}$ et $\overline{\text{REQMR}}$. En revanche, si le modèle est inexact et si les données auxiliaires sont les mêmes, l'estimateur YR fondé sur la procédure décrite dans Stefan et Hidiroglou (2020), adaptée aux taux d'échantillonnage non négligeables, présente les plus faibles $\overline{\text{BRA}}$, tandis que l'estimateur YR restreint présente les plus faibles $\overline{\text{REQMR}}$. Par contre, si les données auxiliaires utilisées pour estimer les moyennes de petits domaines ne sont pas le mêmes que celles utilisées dans l'estimateur GREG, nous en arrivons aux conclusions suivantes. L'estimateur EBLUP restreint et l'estimateur EBLUP ajusté par le ratio sont les

estimateurs réconciliés qui présentent les plus faibles $\overline{\text{BRA}}$ et $\overline{\text{REQMR}}$ si le modèle est exact. Si le modèle n'est pas exact, l'estimateur YR restreint est l'option privilégiée quant au $\overline{\text{BRA}}$ et à la $\overline{\text{REQMR}}$.

La réconciliation doit être fondée sur la procédure EBLUP si le modèle linéaire à effets mixtes est adéquat. Si le modèle linéaire et la référence (l'estimateur GREG) ont en commun une grande quantité de données auxiliaires, les estimateurs réconciliés sont semblables à leurs versions non réconciliées, autrement la perte d'efficacité en raison de la réconciliation peut être importante. Si le modèle n'est pas exact, la procédure YR doit être utilisée pour réaliser la réconciliation. En pareil cas, la réconciliation peut entraîner des gains importants quant au $\overline{\text{BRA}}$ et à la $\overline{\text{REQMR}}$, surtout si le modèle pour petits domaines et l'estimateur GREG partagent un petit nombre de variables auxiliaires. Les populations finies associées à une modélisation inexacte ont été générées d'après le modèle (4.2), la fonction moyenne ayant été incorrectement spécifiée. Cependant, un modèle peut être mal spécifié de différentes façons, et les conclusions associées à ces cas peuvent être différentes.

Remerciements

Nous aimerions remercier deux évaluateurs anonymes et le rédacteur en chef adjoint pour leurs suggestions constructives.

Annexe A

Preuve du résultat 1. Les estimateurs EBLUP $\hat{\boldsymbol{\beta}}_a = \left(\left(\hat{\boldsymbol{\beta}}_{1a}\right)^T, \hat{\beta}_{2a}\right)^T$ et $\hat{v}_{ia}$, qui sont fondés sur le modèle (3.6), satisfont l'équation

$$\sum_{i=1}^{m}\sum_{j\in s_i}\begin{pmatrix}\mathbf{x}_{ij}\\ q_{ij}\end{pmatrix}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{1a} - q_{ij}\hat{\beta}_{2a} - \hat{v}_{ia}\right) = 0. \tag{A.1}$$

L'équation (A.1) prend la forme de l'équation (2.10) et correspond au modèle augmenté (3.6). En étendant la seconde équation sous (A.1), nous obtenons que

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{1a} + \sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}^2\hat{\beta}_{2a} + \sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}\hat{v}_{ia} = \sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}y_{ij}. \tag{A.2}$$

La variable q_{ij} est définie comme $q_{ij} = w_{ij}^{\text{GREG}} - 1$. Le côté droit de l'équation (A.2) est

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}y_{ij} = \sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\text{GREG}} - 1\right) y_{ij} = \hat{Y}^{\text{GREG}} - \sum_{i=1}^{m}\sum_{j\in s_i} y_{ij}. \tag{A.3}$$

Les sommes qui figurent du côté gauche de l'équation (A.2) sont données par

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}\mathbf{x}_{ij}^T = \sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\text{GREG}} - 1\right)\mathbf{x}_{ij}^T = \left(\sum_{i=1}^{m}\mathbf{X}_i - \sum_{j\in s_i}\mathbf{x}_{ij}\right)^T = \left(\sum_{i=1}^{m}\mathbf{x}_{ir}\right)^T, \tag{A.4}$$

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}^2 = \sum_{i=1}^{m} q_{iw}, \tag{A.5}$$

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij} = \sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right) = \sum_{i=1}^{m}\left(\hat{N}_i^{\mathrm{GREG}} - n_i\right). \tag{A.6}$$

En établissant la dernière égalité de l'équation (A.4), nous avons utilisé le fait que $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$ et que les poids w_{ij}^{GREG} satisfont l'équation (3.3). Le résultat 1 s'obtient en remplaçant (A.3), (A.4), (A.5) et (A.6) par (A.2).

Preuve du résultat 2. Les équations d'estimation pondérées par les poids de sondage qui définissent $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ et $\hat{\mathbf{v}}^{\mathrm{YR}}$ sont données par l'équation (2.12) construite avec les poids $w_{ij}^{\mathrm{GREG}} - 1$:

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right)\mathbf{x}_{ij}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} - \hat{v}_i^{\mathrm{YR}}\right) = 0.$$

Étant donné que le premier terme de $\mathbf{x}_{ij}$ est un (représentant une ordonnée à l'origine), il s'ensuit que

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right)\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} - \hat{v}_i^{\mathrm{YR}}\right) = 0. \tag{A.7}$$

Les termes dans l'équation (A.7) sont donnés par :

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right) y_{ij} = \hat{Y}^{\mathrm{GREG}} - \sum_{i=1}^{m}\sum_{j\in s_i} y_{ij}, \tag{A.8}$$

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right)\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} = \left(\sum_{i=1}^{m}\mathbf{X}_i - \sum_{j\in s_i}\mathbf{x}_{ij}\right)^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} = \left(\sum_{i=1}^{m}\mathbf{x}_{ir}\right)^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}}, \tag{A.9}$$

et

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right)\hat{v}_i^{\mathrm{YR}} = \sum_{i=1}^{m}\left(\hat{N}_i^{\mathrm{GREG}} - n_i\right)\hat{v}_i^{\mathrm{YR}}. \tag{A.10}$$

L'ajout de (A.8), (A.9) et (A.10) à (A.7) permet d'obtenir

$$\sum_{i=1}^{m}\sum_{j\in s_i} y_{ij} + \left(\sum_{i=1}^{m}\mathbf{x}_{ri}\right)^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} + \sum_{i=1}^{m}\left(\hat{N}_i^{\mathrm{GREG}} - n_i\right)\hat{v}_i^{\mathrm{YR}} = \hat{Y}^{\mathrm{GREG}}. \tag{A.11}$$

L'équation (A.11) démontre le résultat 2.

Annexe B

Estimation des composantes de la variance par le MVRE reparamétré

Soit $\boldsymbol{\delta} = (\delta_1, \delta_2)$ le vecteur des composantes de la variance, où $\delta_1 = \sigma_v^2$ et $\delta_2 = \sigma_e^2$. Nous définissons le vecteur $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ de sorte que $\sigma_v^2 = e^{\alpha_1}$ et $\sigma_e^2 = e^{\alpha_2}$. La fonction du maximum de vraisemblance logarithmique restreint, désignée $l(\boldsymbol{\alpha})$ est

$$l(\boldsymbol{\alpha}) = l(\alpha_1, \alpha_2) = c - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{y}, \tag{B.1}$$

où c est une constante générique, $\mathbf{V} = e^{\alpha_1}\mathbf{Z}\mathbf{Z}^T + e^{\alpha_2}\mathbf{I}_n$ et $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}$. Il convient de souligner que $\mathbf{P}\mathbf{X} = \mathbf{0}$. La solution à la maximisation de $l(\boldsymbol{\alpha})$ est obtenue de manière itérative au moyen de l'algorithme de cotation de Fisher en mettant à jour l'équation suivante

$$\boldsymbol{\alpha}^{(r+1)} = \boldsymbol{\alpha}^{(r)} + \mathbf{I}\left(\boldsymbol{\alpha}^{(r)}\right)^{-1}\mathbf{s}\left(\boldsymbol{\alpha}^{(r)}\right). \tag{B.2}$$

Dans ce cas-ci, $\mathbf{s}\left(\boldsymbol{\alpha}^{(r)}\right) = \left(\partial l\left(\boldsymbol{\alpha}^{(r)}\right)/\partial\alpha_1, \partial l\left(\boldsymbol{\alpha}^{(r)}\right)/\partial\alpha_2\right)^T$ est le vecteur de dérivées partielles de premier ordre de $l(\boldsymbol{\alpha})$ en ce qui concerne $\boldsymbol{\alpha}$, et $\mathbf{I}\left(\boldsymbol{\alpha}^{(r)}\right) = \left(I_{jk}\left(\boldsymbol{\alpha}^{(r)}\right)\right)_{j,k=1,2}$ est la matrice des dérivées de deuxième ordre attendues de $-l(\boldsymbol{\alpha})$ en ce qui concerne $\boldsymbol{\alpha}$, où $I_{jk}\left(\boldsymbol{\alpha}^{(r)}\right) = E\left(-\partial^2 l\left(\boldsymbol{\alpha}^{(r)}\right)/\partial\alpha_j\partial\alpha_k\right)$.

Sous le modèle BHF, les dérivées partielles de premier ordre de $l(\boldsymbol{\alpha})$ sont données par

$$\frac{\partial l}{\partial\alpha_j}(\boldsymbol{\alpha}) = \left[-\frac{1}{2}\operatorname{tr}\left(\mathbf{P}\mathbf{V}_{(j)}\right) + \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{V}_{(j)}\mathbf{P}\mathbf{y}\right]e^{\alpha_j},\ j = 1, 2, \tag{B.3}$$

où $\mathbf{V}_{(1)} = \mathbf{Z}\mathbf{Z}^T$ et $\mathbf{V}_{(2)} = \mathbf{I}_n$. Les valeurs attendues des dérivées partielles de deuxième ordre de $l(\boldsymbol{\alpha})$ sont

$$E\left(\frac{\partial^2 l}{\partial\alpha_j\partial\alpha_k}(\boldsymbol{\alpha})\right) = -\frac{1}{2}\operatorname{tr}\left(\mathbf{P}\mathbf{V}_{(j)}\mathbf{P}\mathbf{V}_{(k)}\right)e^{\alpha_j+\alpha_k},\ j, k = 1, 2. \tag{B.4}$$

L'estimateur par le MVRE reparamétré de $\boldsymbol{\delta}$ est obtenu au moyen de

$$\hat{\boldsymbol{\delta}}^{\text{reRE}} = \left(\hat{\sigma}_v^{2\text{reRE}}, \hat{\sigma}_e^{2\text{reRE}}\right) = \left(e^{\hat{\alpha}_1}, e^{\hat{\alpha}_2}\right). \tag{B.5}$$

Bibliographie

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

Bell, W.R., Datta, G.S. et Ghosh, M. (2013). Benchmarking small area estimators. *Biometrika*, 100, 189-202.

Datta, G.S., Ghosh, M., Steorts, R. et Maples, J. (2011). Bayesian benchmarking with applications to small area estimation. *Test*, 20, 574-588.

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

Hidiroglou, M.A., et Estevao, V.M. (2016). A comparison of small area and traditional estimators via simulation. *Statistics in Transition*, 17, 133-154.

Huang, R., et Hidiroglou, M.A. (2003). Design consistent estimators for a mixed linear model on survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1897-1904.

Moghtased-Azar, K., Tehranchi, R. et Amiri-Simkooei, A.R. (2014). An alternative method for non-negative estimation of variance components. *Journal of Geodesy*, 88, 427-439.

Nandram, B., et Sayit, H. (2011). Une analyse bayésienne des probabilités de réponse dans les petits domaines sous une contrainte. *Techniques d'enquête*, 37, 2, 147-162. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011002/article/11603-fra.pdf.

Pfeffermann, D., et Barnard, C. (1991). Some new estimators for small area means with applications to the assessment of farmland values. *Journal of Business and Economic Statistics*, 9, 73-83.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

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

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

Stefan, M., et Hidiroglou, M.A. (2020). Benchmarked estimators for a small area mean under a one-fold nested regression model. *Revue Internationale de Statistique*, (À paraître).

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.

Ugarte, M.D., Militino, A.F. et Goicoa, T. (2009). Benchmarked estimates in small areas using linear mixed models with restrictions. *Test*, 18, 342-364.

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

Wang, J., Fuller, W.A. et Qu, Y. (2008). Estimation pour petits domaines sous une contrainte. *Techniques d'enquête*, 34, 1, 33-40. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008001/article/10619-fra.pdf.

You, Y., et Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K. et Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

You, Y., Rao, J.N.K. et Hidiroglou, M.A. (2013). De la performance des estimateurs sur petits domaines autocalés sous le modèle au niveau du domaine de Fay-Herriot. *Techniques d'enquête*, 39, 1, 243-255. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013001/article/11830-fra.pdf.

Techniques d'enquête, juin 2021
Vol. 47, nº 1, pp. 163-204
Statistique Canada, nº 12-001-X au catalogue

Estimation des discontinuités de domaine au moyen de modèles de Fay-Herriot hiérarchiques bayésiens

Jan A. van den Brakel et Harm-Jan Boonstra[1]

Résumé

Les changements dans le plan d'une enquête répétée entraînent généralement des effets systématiques dans les estimations de l'échantillon, qu'on appellera discontinuités dans la suite du texte. Pour ne pas confondre les changements réels d'une période à l'autre avec les effets d'un remaniement, on quantifie souvent les discontinuités en mettant en œuvre parallèlement le plan précédent et le nouveau pendant un certain temps. Les tailles d'échantillon de ces exécutions parallèles sont généralement trop petites pour que des estimateurs directs puissent être appliqués aux discontinuités des domaines. On propose un modèle de Fay-Herriot (FH) hiérarchique bayésien bivarié pour prédire plus précisément les discontinuités de domaine et on l'applique à un remaniement de l'enquête néerlandaise sur la victimisation criminelle (*Dutch Crime Victimzation Survey*). Cette méthode est comparée à un modèle de FH univarié où les estimations directes selon l'approche ordinaire sont utilisées comme covariables dans un modèle de FH pour l'autre approche appliquée sur une taille d'échantillon réduite et un modèle de FH univarié où les estimations directes des discontinuités sont modélisées directement. On propose une procédure de sélection ascendante corrigée qui minimise le critère d'information de Watanabe-Akaike (*Watanabe-Akaike Information Criterion* ou WAIC) jusqu'à ce que la réduction du WAIC soit inférieure à l'erreur-type de ce critère. Au moyen de cette approche, on choisit des modèles plus parcimonieux, ce qui empêche de sélectionner des modèles complexes qui tendent à surajuster les données.

Mots-clés : Modèles au niveau du domaine; modèle de Fay-Herriot bivarié; estimation sur petits domaines; remaniement d'enquête; biais de mesure; méthode de Monte Carlo par chaîne de Markov (MCMC); échantillonneur de Gibbs.

1. Introduction

Les statistiques officielles produites par les instituts nationaux de statistique sont généralement fondées sur des enquêtes-échantillons répétées. Une grande partie de leur valeur réside dans leur continuité, qui permet de suivre l'évolution de la société et de l'économie, de façon à décider des mesures à prendre. Les échantillons d'enquête contiennent, outre des erreurs d'échantillonnage, différentes sources d'erreurs non dues à l'échantillonnage qui ont un effet systématique sur les résultats d'une enquête. Tant que le processus d'enquête est constant, cette composante de biais n'est pas visible. On avance souvent cet argument pour conserver les processus d'enquête des enquêtes répétées aussi longtemps que possible. De temps à autre, il faut toutefois les modifier pour améliorer l'efficacité des enquêtes, réduire les coûts liés à ces dernières ou répondre à de nouvelles exigences, ce qui est fréquemment observé dans l'utilisation d'enquêtes à mode mixte comprenant des questionnaires en ligne dans les statistiques officielles. Le remaniement d'un processus d'enquête a généralement des effets systématiques sur les estimations de l'enquête, puisque les biais induits par les erreurs non dues à l'échantillonnage susmentionnées sont modifiés, ce qui perturbe la comparabilité avec les chiffres publiés auparavant.

On appelle discontinuités les différences systématiques dans les résultats d'une enquête répétée découlant du remaniement du processus d'enquête. Pour éviter la mise en œuvre d'un nouveau processus

1. Jan A. van den Brakel, Statistics Netherlands, Department of Statistical Methods et Maastricht University, Department of Quantitative Economics. Courriel : ja.vandenbrakel@cbs.nl; Harm-Jan Boonstra, Statistics Netherlands, Department of Statistical Methods.

d'enquête perturbant la comparabilité des estimations au fil du temps, il est important de quantifier ces discontinuités. Cela évite de confondre les changements réels des paramètres d'intérêt avec la modification du biais de mesure causée par la modification du processus d'enquête.

Plusieurs méthodes de quantification des discontinuités sont proposées dans la littérature (van den Brakel, Smith et Compton, 2008). Une approche fiable et simple consiste à exécuter les deux méthodes, la précédente et la nouvelle, simultanément pendant un certain temps, procédé que nous désignerons ci-après par l'expression *exécution parallèle*. Idéalement, il s'agit d'une expérience randomisée qui peut être incluse dans l'échantillon probabiliste de l'enquête (van den Brakel, 2008). Dans le présent article, nous proposons des méthodes d'estimation sur petits domaines pour estimer les discontinuités de domaine. Nous examinons une situation où l'enquête ordinaire, utilisée aux fins de production de chiffres officiels, est réalisée sur toute la taille de l'échantillon et parallèlement à une autre approche. En raison des limites budgétaires, l'échantillon attribué à l'approche de rechange n'est souvent pas suffisamment grand pour permettre d'observer les différences détectables minimales aux niveaux préalablement spécifiés de signification et de puissance à l'aide d'estimateurs directs standards, en particulier pour les sous-populations ou les domaines.

Pour expliquer le problème traité dans l'article, il nous faut introduire certaines notations. Soit θ_i la valeur de population réelle d'une variable d'intérêt pour le domaine i. De plus, $\hat{y}_i^r$ et $\hat{y}_i^a$ désignent des estimations directes de θ_i basées respectivement sur la méthode ordinaire d'enquête et l'autre méthode. Étant donné que l'enquête ordinaire est menée à la taille habituelle de l'échantillon, $\hat{y}_i^r$ est une estimation directe fiable de θ_i, au moins pour les domaines planifiés. En raison de la taille réduite de l'échantillon de la nouvelle enquête exécutée parallèlement, $\hat{y}_i^a$ n'apportera toutefois pas une précision suffisante. On peut obtenir des estimations de domaine plus précises sur le petit échantillon disponible dans la nouvelle approche selon le modèle de Fay-Herriot (FH) (Fay et Herriot, 1979), qui est défini comme étant $\hat{y}_i^a = \mathbf{x}_i^t\beta + \nu_i + e_i^a$, $\mathbf{x}_i$ étant un vecteur avec des covariables au niveau du domaine, β les coefficients de régression, ν_i les effets aléatoires du domaine et e_i^a l'erreur d'échantillonnage. Pour obtenir des estimations de domaine plus précises dans l'approche de rechange, van den Brakel, Buelens et Boonstra (2016) ont proposé un modèle de FH univarié hiérarchique bayésien (HB), où les estimations d'échantillon de l'enquête ordinaire sont considérées comme des variables auxiliaires potentielles dans une procédure de sélection de modèle. Ceci implique que $\hat{y}_i^r$ sert de covariable dans $\mathbf{x}_i$, en plus des covariables habituelles issues de registres ou de recensements. Il en résulte un modèle au niveau du domaine, avec erreur de mesure (Ybarra et Lohr, 2008). L'utilisation d'estimations directes fiables observées dans l'enquête ordinaire a augmenté significativement la précision des estimations de domaine dans l'approche de rechange qui est exécutée sur une taille d'échantillon réduite (van den Brakel et coll., 2016).

Soit $\tilde{y}_i^a$ la prédiction sur petits domaines de θ_i basée sur le modèle de FH susmentionné appliqué au petit échantillon attribué à la méthode d'enquête de rechange. Suivant l'approche suivie par van den Brakel et coll. (2016), les estimations ponctuelles des discontinuités de domaine sont obtenues

comme étant la différence entre l'estimation directe obtenue par l'enquête ordinaire et la prédiction de domaine fondée sur le modèle obtenue par l'approche de rechange, c'est-à-dire $\tilde{\Delta}_i = \hat{y}_i^r - \tilde{y}_i^a$. L'utilisation de l'estimation directe de l'enquête ordinaire comme variable auxiliaire dans les prédictions pour petits domaines de l'enquête de rechange donne lieu à des corrélations positives fortes entre les deux estimateurs, qui ne peuvent être ignorées lors du calcul des erreurs-types pour les discontinuités. Plus précisément, $\mathrm{Var}(\tilde{\Delta}_i) = \mathrm{Var}(\hat{y}_i^r) + \mathrm{EQM}(\tilde{y}_i^a) - 2\mathrm{Cov}(\hat{y}_i^r, \tilde{y}_i^a)$. Étant donné que la variable $\hat{y}_i^r$ est également utilisée comme covariable dans $\mathbf{x}_i$ selon le modèle de FH pour $\tilde{y}_i^a$, $\mathrm{Cov}(\hat{y}_i^r, \tilde{y}_i^a)$ est non nulle. À cette fin, nous proposons deux approximations analytiques des erreurs-types des discontinuités. La première méthode combine l'estimation de la variance fondée sur le plan de l'estimateur direct de l'enquête ordinaire $(\mathrm{Var}(\hat{y}_i^r))$ avec la variance a posteriori des prédictions de domaine HB de l'enquête de rechange $(\mathrm{EQM}(\tilde{y}_i^a))$ et un estimateur fondé sur le plan pour la covariance entre les deux estimations ponctuelles $(\mathrm{Cov}(\hat{y}_i^r, \tilde{y}_i^a))$. Cette méthode est instable, en ce sens que même des estimations de la variance négatives se produisent en cas d'estimations de covariance positives fortes. Elle pose aussi un autre problème : les approximations de la variance fondées sur le plan et sur le modèle sont combinées dans une seule mesure de l'incertitude pour les discontinuités. C'est pourquoi nous proposons une seconde approximation analytique, dans laquelle un estimateur fondé sur le plan pour la variance des prédictions de domaine HB $(\mathrm{EQM}(\tilde{y}_i^a))$ est calculé et combiné à la variance fondée sur le plan pour l'estimateur direct de l'enquête ordinaire et la covariance fondée sur le plan entre les deux estimations ponctuelles.

La littérature comporte plusieurs références à l'estimation de l'erreur quadratique moyenne fondée sur le plan dans l'estimation sur petits domaines. Gonzalez et Waksberg (1973) ont introduit le concept d'erreur quadratique moyenne fondée sur le plan d'un ensemble d'estimateurs synthétiques et ils ont proposé un estimateur qui peut cependant être instable et prendre des valeurs négatives. Marker (1995) a proposé un estimateur plus stable mais biaisé pour l'erreur quadratique moyenne fondée sur le plan pour les estimations sur petits domaines, qui peut également prendre des valeurs négatives. Lahiri et Pramanik (2019) ont proposé un estimateur fondé sur le plan de sondage qui ne peut pas prendre de valeurs négatives, en suivant les concepts d'erreur quadratique moyenne fondée sur le plan de sondage, introduite à l'origine par Gonzalez et Waksberg (1973). Rivest et Belmonte (2000) ont proposé un estimateur pour l'erreur quadratique moyenne qui mesure l'incertitude par rapport au plan d'échantillonnage conditionnellement aux effets aléatoires du modèle et en supposant la normalité du modèle d'échantillonnage. Rao, Rubin-Bleuer et Estevao (2018) et Pfeffermann et Ben-Hur (2018) ont également proposé un modèle pour l'erreur quadratique moyenne fondée sur le plan dans les estimateurs sur petits domaines. Rao et coll. (2018) estiment les paramètres du modèle par le maximum de vraisemblance restreint, tandis que Pfeffermann et Ben-Hur (2018) appliquent une méthode bootstrap.

On peut aussi contourner les complications de l'estimation de la variance des discontinuités de domaine dans un modèle de FH univarié en établissant un cadre totalement bayésien aux fins d'analyse des discontinuités de domaine. Deux approches sont proposées dans le présent article. La première approche est un modèle de FH bivarié servant à modéliser les estimations directes simultanément selon la

méthode ordinaire et la méthode de rechange, c'est-à-dire un modèle au niveau du domaine bivarié pour le vecteur $\left(\hat{y}_i^r, \hat{y}_i^a\right)^t$. La composante aléatoire de ce modèle prend en compte la corrélation entre les paramètres de domaine selon la méthode ordinaire et la méthode de rechange. La précision des discontinuités estimées est améliorée par l'augmentation de la taille effective de l'échantillon à l'intérieur des domaines au moyen de corrélations transversales. De plus, une corrélation positive entre les effets de domaine aléatoires diminue encore davantage l'erreur-type des discontinuités estimées. La deuxième approche fait appel à un modèle de FH univarié pour les estimations directes des discontinuités, c'est-à-dire un modèle de FH univarié pour $\hat{\Delta}_i = \hat{y}_i^r - \hat{y}_i^a$. Cette méthode est considérée comme une solution de rechange moins complexe pour le modèle de FH bivarié. On prévoit toutefois qu'elle rend plus difficile la construction de bons modèles de prédiction, puisque les covariables issues des registres pourraient être de bons prédicteurs pour les variables cibles de l'enquête-échantillon, mais probablement pas pour les différences systématiques entre les différences de deux estimations de la même variable obtenues au moyen de processus d'enquête différents.

On a appliqué le modèle de FH univarié proposé par van den Brakel et coll. (2016) pour estimer les discontinuités de domaine dans cinq variables cibles clés de l'Enquête néerlandaise sur la victimisation criminelle (EVC) à partir de données obtenues par une exécution parallèle, dans laquelle l'enquête ordinaire est effectuée à la taille d'échantillon normale et l'enquête de rechange à une taille d'échantillon représentant environ le quart de la taille d'échantillon normale. Dans l'article, le modèle de FH bivarié et le modèle de FH univarié pour les discontinuités de domaine sont également appliqués au même remaniement de l'EVC. Les résultats sont comparés au modèle de FH univarié proposé dans van den Brakel et coll. (2016).

La sélection du modèle dans le présent article se fonde sur une procédure de sélection ascendante qui minimise les critères du WAIC (Watanabe, 2010, 2013). Pour éviter de sélectionner des modèles trop paramétrés, on propose d'ajouter des covariables dans une procédure de sélection ascendante seulement si elles réduisent le critère d'information de Watanabe-Akaike (*Watanabe-Akaike Information Criterion* ou WAIC) de plus que l'erreur-type de ce critère. Cela empêche la sélection de plusieurs covariables qui n'améliorent que légèrement le WAIC, ce qui donne des modèles tendant à surajuster les données.

Le modèle de FH (Fay et Herriot, 1979) est fréquemment appliqué dans le contexte de l'estimation sur petits domaines (Rao et Molina, 2015). Les modèles de FH sont particulièrement appropriés si l'on dispose d'informations auxiliaires au niveau du domaine. Datta, Ghosh, Nangia et Natarjan (1996) ont employé un modèle de FH multivarié ajusté dans un cadre HB pour estimer le revenu médian. Les modèles de FH multivariés ajustés dans un cadre fréquentiste sont examinés dans Gonzales-Manteiga, Lombardia, Molina, Morales et Santamaria (2008) et Benavent et Morales (2016). Plusieurs auteurs ont fourni des modèles de FH de séries chronologiques pour utiliser l'information des échantillons des éditions précédentes d'une enquête comme forme d'estimation sur petits domaines (Rao et Yu, 1994; Datta, Lahiri, Maiti et Lu, 1999; You et Rao, 2000; Estaban, Morales, Perez et Santamaria, 2012; Marhuenda, Molina et Morales, 2013). Pfeffermann et Burck (1990), Pfeffermann et Tiller (2006),

van den Brakel et Krieg (2016), Bollineni-Balabay, van den Brakel, Palm et Boonstra (2017), notamment, donnent des exemples de modèles de FH de séries chronologiques dans un cadre d'espaces d'états. Boonstra et van den Brakel (2019) discutent de la façon dont les modèles de FH de séries chronologiques peuvent être exprimés soit dans un cadre d'espaces d'état et ajustés par filtre de Kalman, soit sous forme de modèles multiniveaux de séries chronologiques dans un cadre hiérarchique bayésien, et estimés au moyen d'un échantillonneur de Gibbs.

L'article est structuré selon le plan suivant. La section 2 décrit l'enquête sur la victimisation criminelle, son remaniement et la configuration de l'exécution parallèle. La section 3 explique le modèle de FH bivarié, y compris le cadre HB et l'approche de sélection et d'évaluation du modèle. Les résultats sont présentés dans la section 4. Enfin, à la section 5, nous concluons l'article par une discussion.

2. Enquête sur la victimisation criminelle

L'enquête néerlandaise sur la victimisation criminelle (EVC) est une enquête de longue date réalisée chaque année par Statistique Pays-Bas (*Statistics Netherlands*) dans le but de publier des chiffres fiables sur les taux de criminalité, le sentiment de sécurité et la satisfaction à l'égard du travail de la police aux Pays-Bas. L'EVC est conçue pour fournir des chiffres fiables au niveau national et au niveau des districts de police, qui correspondent à la subdivision des Pays-Bas en 25 régions. L'EVC est fondée sur un plan d'échantillonnage aléatoire simple stratifié pour les personnes âgées de 15 ans et plus résidant aux Pays-Bas. Les strates sont formées par les régions des districts de police afin que la précision de ces estimations de domaines planifiés soit contrôlée. La base de sondage se fonde sur le registre de tous les résidents de l'État néerlandais, la *Municipal Basis Administration* (MBA). L'échantillon annuel de l'EVC ordinaire est conçu de façon à ce qu'environ 19 000 répondants soient interrogés. L'échantillon est divisé également sur les strates, de sorte qu'environ 760 observations sont obtenues dans chaque strate. L'estimateur par la régression généralisée (ERG) (Särndal, Swensson et Wretman, 1992) sert à estimer les paramètres de population à l'échelle nationale et à celle des districts de police.

L'EVC a été remaniée en 2008. La collecte de données est passée d'un plan à mode mixte reposant sur des interviews sur place assistées par ordinateur (IPAO) et des interviews téléphoniques assistées par ordinateur (ITAO) à un plan séquentiel à mode mixte qui commence par des interviews sur le Web (IW) suivies d'interviews des non-répondants par IPAO et ITAO. De plus, des modifications ont été apportées au questionnaire pour en améliorer la formulation et l'ordre des questions. Pour quantifier les discontinuités découlant du remaniement, l'enquête ordinaire utilisée aux fins de publication officielle a été menée parallèlement à l'approche d'enquête de rechange avec un échantillon d'environ 6 000 répondants. Dans cette application, la méthode ordinaire était fondée sur le nouveau plan d'enquête utilisant une collecte par IW, ITAO et IPAO, tandis que la méthode de rechange se basait sur le plan précédent qui repose sur une collecte de données par IPAO et ITAO seulement. Le plan d'échantillonnage de l'exécution parallèle est basé sur un échantillonnage aléatoire simple stratifié où les districts de police

sont les strates, au moyen d'une allocation proportionnelle. On obtient un plan d'échantillonnage optimal pour estimer les chiffres à l'échelle nationale, mais sous-optimal pour l'estimation de domaine.

L'enquête rend compte de nombreuses variables de résultats différentes. Dans la présente étude, cinq variables clés sont examinées (voir le tableau 2.1). Les estimations de ces variables à l'échelle nationale selon l'enquête ordinaire et l'enquête de rechange sont indiquées dans le tableau 2.2. La taille de l'échantillon allouée à l'approche de rechange est suffisamment grande pour estimer les discontinuités au niveau national au moyen de l'estimateur ERG, mais insuffisante pour estimer les discontinuités au niveau du domaine des 25 districts de police. Les estimations directes des discontinuités à l'échelle nationale sont significativement très différentes de zéro, contrairement aux moyennes non pondérées des estimations de domaine directes et de leurs erreurs-types. Aux fins d'obtention de prédictions plus précises des discontinuités de domaine, une méthode d'estimation sur petits domaines fondée sur un modèle basé sur des modèles au niveau du domaine (Fay et Herriot, 1979) est proposée dans la section suivante.

Tableau 2.1
Les cinq variables clés de l'EVC considérées dans la présente étude

Variable	Description
`nuisance`	Nuisance perçue dans le quartier sur une échelle de dix points : cette mesure comprend la nuisance causée par des personnes ivres, des voisins ou des groupes de jeunes, du harcèlement et des problèmes liés à la consommation de drogue.
`unsafe`	Pourcentage de personnes se sentant parfois en danger.
`propvict`	Pourcentage de personnes ayant déclaré avoir été victimes d'un crime contre des biens au cours des 12 derniers mois.
`offtot`	Nombre total d'infractions pour 100 personnes.
`satispol`	Pourcentage de personnes satisfaites de leur dernier contact avec la police (si le contact a eu lieu au cours des 12 derniers mois).

Tableau 2.2
Les estimations ERG pour l'approche d'enquête ordinaire et de rechange ont été calculées en moyenne par districts et à l'échelle nationale. Erreurs-types entre crochets

Variable	Moyenne sur les 25 districts de police						Échelle nationale					
	ordinaire		de rechange		$\hat{\Delta}$		ordinaire		de rechange		$\hat{\Delta}$	
`offtot`	42,29	(4,73)	33,28	(5,73)	9,01	(7,69)	43,79	(1,07)	34,09	(1,04)	9,70	(1,49)
`unsafe`	24,38	(2,03)	19,86	(2,87)	4,52	(3,57)	25,07	(0,44)	20,48	(0,52)	4,59	(0,68)
`nuisance`	1,61	(0,11)	1,28	(0,13)	0,33	(0,17)	1,67	(0,02)	1,34	(0,02)	0,33	(0,03)
`satispol`	60,61	(4,23)	55,58	(6,88)	5,04	(8,21)	59,88	(0,92)	55,10	(1,25)	4,78	(1,55)
`propvict`	12,55	(1,60)	9,78	(2,19)	2,78	(2,77)	13,02	(0,36)	10,32	(0,39)	2,70	(0,53)

3. Méthodes

3.1 Estimation sur petits domaines des discontinuités de domaine

Le test d'hypothèses sur les différences entre les estimations d'un paramètre de population finie observé selon les différents processus d'enquête suppose l'existence d'erreurs de mesure. Par conséquent,

il faut un modèle d'erreur de mesure pour expliquer les différences systématiques entre les estimations d'enquête pour le même paramètre de population observé selon deux méthodes d'enquête différentes. Soit θ_i le paramètre de population du domaine $i = 1, \ldots, m$. Supposons que y_i^r et y_i^a désignent la valeur observée pour θ_i dans le cas d'un dénombrement complet respectivement selon l'approche régulière et l'approche de rechange. Les estimations directes pour y_i^r et y_i^a sont obtenues au moyen de l'estimateur ERG à partir des échantillons attribués à l'enquête ordinaire et à l'enquête de rechange et sont désignées respectivement par $\hat{y}_i^r$ et $\hat{y}_i^a$.

La relation entre les valeurs observées dans un dénombrement complet et le paramètre de population réel est la suivante : $y_i^q = \theta_i + \lambda_i^q$, $i = 1, \ldots, m$, $q = r, a$, λ_i^q étant le biais de mesure réel si θ_i est mesuré selon l'approche d'enquête q. Sans information externe, il est impossible d'estimer λ_i^q. À partir des données de l'échantillon, seul le biais relatif par exemple $\Delta_i = y_i^r - y_i^a = \lambda_i^r - \lambda_i^a$ est identifiable. Les estimations directes de ces discontinuités sont obtenues à partir des données d'enquête comme le contraste entre les estimations ERG, c'est-à-dire $\hat{\Delta}_i = \hat{y}_i^r - \hat{y}_i^a$.

Dans le cas de l'EVC des Pays-Bas, la taille de l'échantillon de l'enquête ordinaire est suffisamment grande pour obtenir des estimations directes suffisamment précises pour les domaines planifiés, puisque l'échantillon est conçu aux fins de la publication de statistiques officielles pour ces domaines. L'échantillon attribué à l'enquête de rechange pour l'exécution parallèle n'a qu'un tiers de la taille d'échantillon ordinaire, ce qui est insuffisant pour obtenir des estimations directes précises pour les domaines planifiés. Dans un article antérieur (van den Brakel et coll., 2016), nous avons élaboré des modèles de FH univariés afin d'obtenir des prédictions plus précises pour les paramètres de domaine observés avec la petite taille d'échantillon attribuée à l'approche d'enquête de rechange qui utilise des variables auxiliaires issues de trois sources différentes. La première source comporte des variables démographiques issues de la *Municipal Basis Administration*, qui est l'administration recensant tous les résidents du pays. La deuxième source contient des variables connexes disponibles dans le Registre des infractions déclarées par la police (RIDP). La troisième source, qui est unique dans le cas d'une exécution parallèle, contient des estimations directes pour les mêmes variables observées selon l'enquête ordinaire, qui sont suffisamment précises au moins pour les domaines planifiés comme les districts de police. Les estimations directes de l'enquête ordinaire sont souvent sélectionnées comme variables auxiliaires pour ces modèles de FH univariés. Cela n'est pas surprenant puisque ce sont des estimations d'enquête pour les mêmes paramètres de population. Bien qu'elles soient mesurées au moyen d'un processus d'enquête différent, on peut s'attendre à des corrélations positives fortes. De fortes améliorations de la précision de la prédiction sur petits domaines sont effectivement constatées si l'ensemble de variables auxiliaires potentielles, c'est-à-dire de la MBA et du RIDP, est étendu avec les estimations directes de l'EVC ordinaire.

Dans cette application, l'erreur d'échantillonnage dans les variables auxiliaires provenant de l'EVC ordinaire peut être ignorée dans le modèle de FH, puisque la taille de l'échantillon et, par conséquent, l'erreur d'échantillonnage pour ces domaines sont plus ou moins égales pour les domaines (Ybarra et

Lohr, 2008). Cela signifie que la composante de variance des effets de domaine aléatoires est augmentée avec l'erreur d'échantillonnage des variables auxiliaires, ce qui convient tant que l'erreur d'échantillonnage ne diffère pas entre domaines. Dans la plupart des applications, ce n'est pas le cas et il faudrait employer les méthodes proposées par Ybarra et Lohr (2008) pour tenir compte de l'erreur d'échantillonnage dans les variables auxiliaires.

Les modèles multiniveaux de FH peuvent être ajustés selon une approche fréquentiste utilisant le meilleur prédicteur linéaire sans biais empirique (EBLUP) ou selon une méthode HB (Rao et Molina, 2015). Dans van den Brakel et coll. (2016) l'approche HB est privilégiée à l'EBLUP, puisque l'information auxiliaire forte dans la partie à effet fixe du modèle donne souvent des estimations nulles pour la composante de variance des effets de domaine aléatoires; ce qui attribue trop de poids à la partie de la régression synthétique et trop peu de poids aux estimations directes dans l'EBLUP (Bell, 1999; Rao et Molina, 2015). Ce problème peut également être résolu par une estimation par la méthode du maximum de vraisemblance corrigée, voir par exemple Li et Lahiri (2010) et Hirose et Lahiri (2018).

Soit $\tilde{y}_i^a$ la prédiction HB pour le domaine i dans l'approche de rechange. Désormais, les discontinuités de domaine sont obtenues par $\tilde{\Delta}_i = \hat{y}_i^r - \tilde{y}_i^a$. L'utilisation d'estimations directes de l'enquête ordinaire comme variables auxiliaires dans la partie fixe du modèle de FH pour l'enquête de rechange augmente considérablement la complexité de l'estimation de la variance des discontinuités. La variance de $\tilde{\Delta}_i$ peut être exprimée comme étant $\mathrm{Var}(\tilde{\Delta}_i) = \mathrm{Var}(\hat{y}_i^r) + \mathrm{EQM}(\tilde{y}_i^a) - 2\mathrm{cov}(\hat{y}_i^r, \tilde{y}_i^a)$. L'utilisation de $\hat{y}_i^r$ ou d'estimations d'échantillons connexes comme variables auxiliaires pour prédire $\tilde{y}_i^a$ donne des valeurs non nulles pour $\mathrm{cov}(\hat{y}_i^r, \tilde{y}_i^a)$ qui ne peuvent pas être ignorées. Pour calculer par approximation $\mathrm{Var}(\tilde{\Delta}_i)$, van den Brakel et coll. (2016) ont proposé un estimateur approximativement sans biais par rapport au plan de sondage pour $\mathrm{cov}(\hat{y}_i^r, \tilde{y}_i^a)$ et $\mathrm{Var}(\hat{y}_i^r)$, tandis que l'$\mathrm{EQM}(\tilde{y}_i^a)$ est obtenue par approximation avec la variance a posteriori des prédictions de domaine HB. L'un des principaux inconvénients de cette approche réside dans l'interrelation entre les mesures d'incertitude fondées sur le modèle et celles fondées sur le plan. D'une part, les EQM pour $\tilde{y}_i^a$ sont obtenues par approximation avec leurs variances a posteriori. D'autre part, les covariances entre $\hat{y}_i^r$ et $\tilde{y}_i^a$ sont obtenues par approximation dans une perspective fondée sur le plan. Par conséquent, l'application naïve de cette approche peut donner des estimations de la variance négatives pour les discontinuités. Cet inconvénient a été résolu au moyen d'une approximation fondée sur le plan pour l'$\mathrm{EQM}(\tilde{y}_i^a)$, ce qui donne une approximation entièrement fondée sur le plan pour l'incertitude des discontinuités de domaine estimées.

Dans le présent article, nous proposons un cadre totalement HB pour l'estimation des discontinuités de domaine comme solution de rechange en élaborant un modèle de FH bivarié pour les paramètres de domaine observés selon l'approche ordinaire et de l'approche de rechange. L'avantage de cette approche est qu'elle améliore la précision des estimations directes des estimations de domaine selon la méthode ordinaire et de rechange en empruntant de l'information d'autres domaines et des deux enquêtes. Les estimations de la variance négatives pour les discontinuités de domaine estimées sont exclues par

définition dans ce cadre HB multivarié. Cette méthode est comparée à une solution de rechange simple, à savoir un modèle de FH univarié pour les estimations directes des discontinuités. Comme nous l'avons indiqué dans l'introduction, il se peut qu'il soit difficile de trouver dans les registres disponibles des covariables qui expliquent les discontinuités. Les deux modèles présentent notamment l'avantage d'éviter les complications liées à la prise en compte de l'erreur d'échantillonnage dans les variables auxiliaires, ce qui est nécessaire si les estimations d'enquête de l'enquête ordinaire sont utilisées comme covariables dans les modèles de FH univariés et que l'erreur d'échantillonnage diffère entre domaines.

3.2 Modèle de Fay-Herriot bivarié

Une version bivariée du modèle de FH (Fay et Herriot, 1979) commence par un modèle de mesure pour les deux estimations par régression généralisée (ERG) observées dans chaque domaine :

$$\hat{\mathbf{y}}_i = \mathbf{y}_i + \mathbf{e}_i,\ i = 1, \ldots, m, \tag{3.1}$$

avec $\mathbf{y}_i = \left(y_i^r, y_i^a\right)^t$, $\hat{\mathbf{y}}_i$ étant un vecteur contenant les estimations ERG de $\mathbf{y}_i$ et $\mathbf{e}_i = \left(e_i^r, e_i^a\right)^t$ un vecteur avec les erreurs d'échantillonnage de $\hat{\mathbf{y}}_i$ pour lesquelles on suppose que

$$\mathbf{e}_i \overset{\text{ind}}{\sim} \mathcal{N}\left(\mathbf{0}_2, \mathbf{\Psi}_i\right),\ i = 1, \ldots, m. \tag{3.2}$$

Ici $\mathbf{0}_2$ est un vecteur-colonne en 2 dimensions avec chaque élément égal à zéro. Comme l'échantillon de l'enquête ordinaire et celui de l'enquête de rechange sont tirés de façon indépendante, on suppose que $\mathbf{\Psi}_i = \text{Diag}\left(\psi_i^r, \psi_i^a\right)$ où ψ_i^q est la variance par rapport au plan de $\hat{y}_i^q$. On suppose également que ces variances par rapport au plan sont connues, bien qu'elles soient remplacées par leurs estimations en pratique. Les paramètres de domaine vrais sont modélisés au moyen d'un modèle multiniveau. Pour les effets fixes, on suppose que la méthode ordinaire et la méthode de rechange partagent les mêmes covariables. Dans le cas le plus général, les coefficients de régression de la partie fixe sont différents pour les deux variables, c'est-à-dire $y_i^q = \mathbf{x}_i^t \boldsymbol{\beta}^q + \nu_i^q$, avec $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)^t$ un vecteur p avec des covariables de domaine i, $\boldsymbol{\beta}^q$ un vecteur p des coefficients de régression, qui sont égaux dans tous les domaines, mais qui peuvent différer selon l'approche d'enquête. On suppose que $x_{i1} = 1$ correspond à l'ordonnée à l'origine. De plus, ν_i^q sont des effets de domaine aléatoires. Cela donne lieu au modèle multiniveau bivarié suivant pour les deux paramètres de domaine :

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i,\ i = 1, \ldots, m, \tag{3.3}$$

où $\mathbf{X}_i = \mathbf{I}_2 \otimes \mathbf{x}_i^t$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{r^t}, \boldsymbol{\beta}^{a^t}\right)^t$, $\mathbf{I}_2$ a une matrice d'identité bidimensionnelle, et $\mathbf{v}_i = \left(\nu_i^r, \nu_i^a\right)^t$. Pour les effets de domaine aléatoires, on suppose que

$$\mathbf{v}_i \overset{\text{IID}}{\sim} \mathcal{N}\left(\mathbf{0}_2, \boldsymbol{\Sigma}\right),\ i = 1, \ldots, m, \tag{3.4}$$

avec $\boldsymbol{\Sigma}$ une matrice de covariance 2×2 générale pour les effets de domaine aléatoires. L'insertion de (3.3) dans (3.1) donne :

$$\hat{\mathbf{y}}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i + \mathbf{e}_i,\ i = 1, \ldots, m, \tag{3.5}$$

avec les hypothèses de modèle (3.2) et (3.4).

Comme le nombre de domaines dans cette application est petit, il est important de sélectionner des modèles parcimonieux. Une façon de réduire la complexité du modèle consiste à supposer que les coefficients de régression sont égaux dans les deux approches d'enquête. Dans ce cas, un indicateur fictif, disons δ_i, est introduit, il est égal à zéro pour l'enquête ordinaire et égal à un pour l'enquête de rechange. Dans ce cas, $x_{i1} = 1$ correspond à l'ordonnée à l'origine globale et $x_{i2} = \delta_i$ est l'indicateur dont le coefficient mesure les différences entre les ordonnées à l'origine des variables observées selon les deux enquêtes. Alors $y_i^q = \mathbf{x}_i^t\boldsymbol{\beta} + \nu_i^q$ et dans (3.3) $\mathbf{X}_i = \mathbf{x}_i^t$, et $\boldsymbol{\beta}$ un vecteur avec les coefficients de régression correspondants. Par conséquent, deux versions des effets fixes sont examinées :

- `FE_uq` : modèle à effet fixe où l'approche régulière et l'approche de rechange partagent les mêmes covariables, mais ont des coefficients de régression différents. Dans ce cas, les discontinuités de domaine sont données par

$$\Delta_i = \sum_{j=1}^{p} x_{i,j}\left(\beta_j^r - \beta_j^a\right) + \left(\nu_i^r - \nu_i^a\right). \tag{3.6}$$

- `FE_eq` : une version plus parcimonieuse pour la composante à effet fixe, qui suppose que les coefficients de régression sont égaux dans l'approche régulière et l'approche de rechange. Dans ce cas, les discontinuités de domaine sont données par

$$\Delta_i = -\beta_2 + \left(\nu_i^r - \nu_i^a\right), \tag{3.7}$$

  avec β_2 le coefficient de régression pour $x_{i2} = \delta_i$.

Les structures de covariance suivantes pour les effets de domaine aléatoires sont examinées :

- `RE_f` : une matrice de covariance complète $\boldsymbol{\Sigma}$ pour les effets de domaine aléatoires. La corrélation positive entre les effets de domaine aléatoires augmentera davantage la précision des estimations pour les discontinuités de domaine puisque les estimations de domaine empruntent de l'information non seulement de domaines différents, mais aussi entre les deux enquêtes.
- `RE_d` : une matrice de covariance diagonale avec des variances séparées pour l'approche ordinaire et l'approche de rechange, c'est-à-dire : $\boldsymbol{\Sigma} = \mathrm{Diag}\left(\sigma_r^2, \sigma_a^2\right)$. Cette structure de covariance combinée au modèle `FE_uq` revient à appliquer un modèle de FH univarié aux deux enquêtes séparément. Dans ce cas, les modèles utilisent seulement l'information des échantillons provenant d'autres domaines de la même enquête, mais pas de l'autre enquête, afin d'améliorer la précision des estimations pour les discontinuités de domaine.
- `RE_s` : une matrice de covariance diagonale avec des variances égales pour l'approche ordinaire et l'approche de rechange, c'est-à-dire : $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_2$.

3.3 Modèle de Fay-Herriot univarié pour les discontinuités de domaine

Le modèle de FH univarié pour les estimations directes des discontinuités commence par la définition d'un modèle d'erreur de mesure pour les estimations par la régression généralisée des discontinuités :

$$\hat{\Delta}_i = \Delta_i + z_i \tag{3.8}$$

avec $\Delta_i = y_i^r - y_i^a$ la vraie discontinuité du domaine i selon un dénombrement complet de la population dans les deux approches, $\hat{\Delta}_i = \hat{y}_i^r - \hat{y}_i^a$ l'estimation par la régression généralisée pour Δ_i fondée sur l'exécution parallèle et $z_i = e_i^r - e_i^a$ l'erreur d'échantillonnage pour $\hat{\Delta}_i$. On suppose que $z_i \simeq \mathcal{N}(0, \psi_i^r + \psi_i^a)$ et que les variances par rapport au plan des erreurs d'échantillonnage sont connues. Pour Δ_i, on suppose le modèle linéaire suivant :

$$\Delta_i = \mathbf{x}_i^t \boldsymbol{\beta} + v_i, \tag{3.9}$$

avec $v_i \simeq \mathcal{N}(0, \sigma_v^2)$. L'insertion de (3.9) dans (3.10) donne :

$$\hat{\Delta}_i = \mathbf{x}_i^t \boldsymbol{\beta} + v_i + z_i. \tag{3.10}$$

3.4 Estimation du modèle de Fay-Herriot bivarié

Les modèles élaborés dans les sous-sections 3.2 et 3.3 sont ajustés au moyen d'une méthode HB utilisant un échantillonnage de Monte Carlo par chaînes de Markov (MCMC). On utilise en particulier l'échantillonneur de Gibbs. Les lois a priori suivantes sont utilisées pour les paramètres du modèle et les hyperparamètres. Pour les coefficients de régression, on suppose des lois a priori impropres uniformes, c'est-à-dire $\boldsymbol{\beta} \sim 1$. Pour les effets de domaine aléatoires, on utilise une loi a priori normale multivariée : $\mathbf{v} \mid \boldsymbol{\Sigma} \simeq \mathcal{N}(\mathbf{0}_{2m}, \mathbf{I}_m \otimes \boldsymbol{\Sigma})$.

Dans le cas d'une matrice de covariance complète pour les effets de domaine aléatoires, la loi a priori pour $\boldsymbol{\Sigma}$ est considérée comme une loi de Wishart inversée et mise à l'échelle (O'Malley et Zaslavsky, 2008). On obtient cette distribution en écrivant $\boldsymbol{\Sigma} = \mathrm{Diag}(\boldsymbol{\xi})\tilde{\boldsymbol{\Sigma}}\mathrm{Diag}(\boldsymbol{\xi})$, avec $\boldsymbol{\xi} = (\xi^r, \xi^a)^t$ et en supposant une distribution normale standard pour ξ^r et ξ^a, c'est-à-dire $\xi^x \simeq \mathcal{N}(0, 1)$, $(x \in (a, r))$ et une loi de Wishart inverse pour $\tilde{\boldsymbol{\Sigma}}$, c'est-à-dire $\tilde{\boldsymbol{\Sigma}} \simeq \mathrm{Inv-Wish}(v_v, \boldsymbol{\Phi}_v)$, avec $v_v = d + 1$ degrés de liberté, avec d la dimension de $\tilde{\boldsymbol{\Sigma}}$ qui est égale à 2 dans cette application, et le paramètre d'échelle $\boldsymbol{\Phi}_v = \mathbf{I}_2$. Dans le cas d'une matrice de covariance diagonale pour les effets de domaine aléatoires, les lois a priori pour σ_q (en cas de variances inégales) ou σ (en cas de variances égales) sont des lois demi-Cauchy. Ces distributions a priori sont plus robustes que la loi du khi carré inverse plus couramment utilisée (Gelman, 2006). La loi du khi carré inverse peut être informative, même en cas de paramètres dont l'échelle et la forme sont petites. De plus, des problèmes de convergence peuvent se produire avec l'échantillonneur de Gibbs. Les deux problèmes sont en grande partie évités par une paramétrisation multiplicative et redondante des effets aléatoires (Gelman, 2006; Gelman, Van Dyk, Huang et Boscardin, 2008; Polson et Scott, 2012). Voir van den Brakel et Boonstra (2018) pour en savoir plus sur les lois a priori de ce modèle.

Soit $\hat{\mathbf{y}}$ le $2m$ vecteur-colonne obtenu en empilant les m vecteurs-colonnes $\hat{\mathbf{y}}_i$, et $\mathbf{X}$ la matrice obtenue par l'empilement des matrices $\mathbf{X}_i$. En cas de coefficients de régression inégaux, $\mathbf{X}$ est une matrice $2m \times 2p$. En cas de coefficients de régression égaux, $\mathbf{X}$ est une matrice $2m \times p$. La fonction de vraisemblance peut s'écrire comme suit :

$$p(\hat{\mathbf{y}} \,|\, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{v}, \boldsymbol{\Psi}), \tag{3.11}$$

avec $\boldsymbol{\Psi} = \oplus_{i=1}^{m} \boldsymbol{\Psi}_i$ une matrice diagonale $2m \times 2m$ avec les variances par rapport au plan des estimations directes $\hat{\mathbf{y}}$ et $\boldsymbol{\theta}$ un vecteur contenant tous les paramètres du modèle. La distribution a priori conjointe $p(\boldsymbol{\theta})$ est égale au produit des lois a priori mentionnées ci-dessus. La distribution a posteriori de $\boldsymbol{\theta}$ est proportionnelle à la densité conjointe, c'est-à-dire $p(\boldsymbol{\theta} \,|\, \hat{\mathbf{y}}) \propto p(\boldsymbol{\theta})\, p(\hat{\mathbf{y}} \,|\, \boldsymbol{\theta})$. Le modèle est ajusté au moyen de l'échantillonneur de Gibbs; Geman et Geman (1984); Gelfand et Smith (1990). Les distributions conditionnelles complètes utilisées dans l'échantillonneur de Gibbs sont spécifiées dans van den Brakel et Boonstra (2018).

Pour chaque modèle considéré, l'échantillonneur de Gibbs est exécuté en trois chaînes indépendantes avec des valeurs de départ générées aléatoirement. La longueur de chaque chaîne après la période de rodage pour chaque analyse est de 10 000 itérations. Cela donne 30 000 tirages pour calculer les estimations et les erreurs-types. On évalue la convergence de la simulation par MCMC au moyen de tracés et de graphiques d'autocorrélation ainsi que du facteur de réduction d'échelle potentiel de Gelman-Rubin (Gelman et Rubin, 1992), qui diagnostique le mélange de chaînes. Les diagnostics laissent penser que toutes les chaînes convergent bien dans les 500 tirages. De plus, les erreurs de simulation de Monte Carlo estimées sont faibles comparativement aux erreurs-types a posteriori pour tous les paramètres, si bien que le nombre de tirages suffit largement à nos besoins.

Les paramètres à estimer d'intérêt sont exprimés sous forme de fonctions des paramètres, et l'application de ces fonctions aux données de sortie de la méthode MCMC pour les paramètres donne des résultats tirés des lois a posteriori pour ces paramètres à estimer. Les prédictions de domaine pour les variables cibles selon le modèle de FH bivarié sont obtenues comme étant les moyennes a posteriori calculées approximativement au moyen de la sortie de l'échantillonneur de Gibbs et elles sont désignées par $\tilde{y}_i^{q,b\mathrm{FH}}$. Les prédictions de domaine pour les discontinuités sont obtenues comme les moyennes a posteriori de (3.6) ou (3.7) calculées approximativement au moyen de la sortie de l'échantillonneur de Gibbs et sont désignées par $\tilde{\Delta}_i^{b\mathrm{FH}}$. Les erreurs quadratiques moyennes pour $\tilde{y}_i^{q,b\mathrm{FH}}$ et $\tilde{\Delta}_i^{b\mathrm{FH}}$ sont obtenues sous forme de variances a posteriori calculées approximativement au moyen de la sortie de l'échantillonneur de Gibbs.

Les méthodes sont mises en œuvre dans R au moyen du module R `mcmcsae` (Boonstra, 2020).

3.5 Variances par rapport au plan de regroupement

Les estimations des variances par rapport au plan ψ_i^r et ψ_i^a sont disponibles à partir de l'estimateur ERG et sont utilisées comme si les vraies variances par rapport au plan étaient connues. Il s'agit d'une

hypothèse standard dans l'estimation sur petits domaines. C'est pourquoi il est important de fournir des estimations fiables de ces variances par rapport au plan. Dans l'enquête ordinaire, les estimations de la variance des estimations ERG sont jugées suffisamment fiables pour être utilisées dans le modèle de FH. Pour l'enquête de rechange, les estimations des variances par rapport au plan ne sont pas fiables et sont donc lissées pour que soit améliorée la stabilité des estimations de ψ_i^a. En supposant que les variances de population des résidus de l'estimation par la régression généralisée selon l'approche de rechange sont égales dans tous les domaines, on utilise le type d'analyse de variance de l'estimateur de la variance regroupé :

$$\psi_i^a = \frac{1 - f_i^a}{n_i^a} \frac{1}{n^a - m} \sum_{i=1}^{m} \left(n_i^a - 1\right) S_{i;\mathrm{GREG}}^{a^2}$$

avec f_i^a la fraction d'échantillon dans le domaine i de l'enquête de rechange, n_i^a le nombre de répondants dans le domaine i selon l'enquête de rechange, $n^a = \sum_{i=1}^{m} n_i^a$, et $S_{i;\mathrm{GREG}}^{a^2}$ la variance de la population estimée des résidus de l'estimation par la régression généralisée.

Il est aussi possible de lisser les estimations de la variance des estimations directes en modélisant les estimations de la variance ainsi que les estimations ERG elles-mêmes (You et Chapman, 2006) et Sugasawa, Tamae et Kubokawa (2017). Cette méthode remonte à Arora et Lahiri (1997). Une autre possibilité consiste à lisser les estimations de la variance en appliquant les fonctions de variance généralisée (Wolter (2007), chapitre 7, et Hawala et Lahiri (2018)).

3.6 Sélection et évaluation du modèle

Les critères de sélection de modèle souvent appliqués dans les scénarios HB sont le critère d'information largement applicable aussi dit critère d'information Watanabe-Akaike (*Watanae-Akaike Information Criterion* ou WAIC) (Watanabe, 2010, 2013) et le critère d'information de déviance (DIC) (Spiegelhalter, Best, Carlin et van der Linde, 2002). Leur emploi est courant, car ils se calculent facilement à partir des résultats de la simulation par MCMC et parce qu'ils permettent un compromis raisonnable entre l'ajustement du modèle et la complexité du modèle. Le WAIC est considéré comme une amélioration par rapport au DIC, puisque ce dernier peut produire des estimations négatives pour le nombre effectif de paramètres et qu'il n'est pas défini pour les modèles singuliers (Vehtari, Gelman et Gabry, 2017). La pénalité utilisée aux fins de complexité du modèle dans le DIC et le WAIC est étroitement liée au nombre effectif de paramètres proposé par Hodges et Sargent (2001) pour les modèles linéaires multiniveaux, pour lesquels chaque effet fixe contribue à un degré de liberté et les effets aléatoires contribuent à une valeur dans la fourchette entre zéro et m, selon la taille de la composante de variance. Comme l'indique la définition du WAIC, les modèles aux valeurs de WAIC plus petites sont préférables. Les estimations du WAIC sont incertaines, et une approximation de son erreur-type est fournie par l'équation (23) de Vehtari et coll. (2017) et peut se calculer au moyen du module R `loo` (Vehtari, Gelman et Gabry, 2015).

Les covariables sont sélectionnées à partir de l'ensemble de variables auxiliaires répertoriées dans van den Brakel et Boonstra (2018) au moyen d'une procédure de sélection ascendante. Plusieurs modèles sont comparés au moyen des estimations du WAIC mentionnées ci-dessus. À partir de l'ensemble de covariables potentielles, la covariable ayant la valeur de WAIC la plus petite est sélectionnée dans le modèle. Ce processus de sélection est répété itérativement tant que l'ajout d'une nouvelle covariable diminue davantage la valeur du WAIC. Dans cette application, la procédure de sélection ascendante, abrégée en `étape-WAIC`, donne souvent des modèles comportant un grand nombre de covariables. Puisque les valeurs du WAIC sont des estimations contenant des erreurs, il semble qu'il ne serait pas souhaitable de minimiser le WAIC en ajoutant des covariables au modèle tant que cela réduit les estimations ponctuelles du WAIC. À titre de solution de rechange, nous avons appliqué une procédure de sélection ascendante dans laquelle des covariables sont ajoutées au modèle tant qu'une nouvelle covariable réduit le WAIC d'une valeur supérieure à l'erreur-type estimée du WAIC. Nous appellerons cette méthode `étape-WAIC-et`.

La procédure de sélection ascendante est appliquée à chacune des six combinaisons différentes des deux versions à effet fixe (`FE_uq` et `FE_eq`) et des trois structures de covariance de la composante aléatoire (`RE_f`, `RE_d` et `RE_s`). Des six modèles obtenus, on sélectionne celui ayant la plus petite valeur de WAIC. L'adéquation de modèle des six modèles sélectionnés est évaluée au moyen de vérifications prédictives a posteriori. Cela signifie que l'on compare les ensembles de données répliquées – simulés à partir de la distribution prédictive a posteriori – aux données observées à l'origine pour étudier les écarts systématiques et évaluer la mesure dans laquelle le modèle sélectionné ajuste les données observées (Gelman, Carlin, Stern, Dunson, Vehtari et Rubin, 2004). Les valeurs p prédictives a posteriori sont calculées pour six tests différents qui évaluent des aspects particuliers de la distribution prédictive a posteriori. Les valeurs p prédictives a posteriori pour les discontinuités de domaine sont définies comme étant $p = P\left(T\left(\hat{\boldsymbol{\Delta}}^{\text{sim}}, \boldsymbol{\Delta}\right) \geq T\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}\right) \middle| \hat{\mathbf{y}}\right)$, où $\hat{\boldsymbol{\Delta}}^{\text{sim}} = \left(\hat{\Delta}_1^{\text{sim}}, \ldots, \hat{\Delta}_m^{\text{sim}}\right)^t$ sont des répliques des discontinuités observées pour les m domaines selon la distribution prédictive a posteriori, $\hat{\boldsymbol{\Delta}} = \left(\hat{\Delta}_1, \ldots, \hat{\Delta}_m\right)^t$ les estimations directes observées pour les m discontinuités de domaine et $T(\hat{\boldsymbol{\Delta}}^{\text{sim}}, \boldsymbol{\Delta})$ (ou $T(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta})$) une statistique de test qui dépend de $\hat{\boldsymbol{\Delta}}^{\text{sim}}$ (ou $\hat{\boldsymbol{\Delta}}$) et des valeurs vraies inconnues pour les m discontinuités de domaine $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_m)^t$. Les valeurs p prédictives a posteriori sont estimées à partir de la sortie de l'échantillonneur de Gibbs comme la moyenne sur les échantillons de Monte Carlo S

$$\hat{p} = \frac{1}{S} \sum_{s=1}^{S} I\left(T\left(\hat{\boldsymbol{\Delta}}^s, \boldsymbol{\Delta}^s\right) \geq T\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}^s\right)\right), \tag{3.12}$$

avec $I(A)$ la fonction d'indicateur qui a une valeur de un si la condition A est remplie et nulle sinon, avec $\hat{\boldsymbol{\Delta}}^s = \left(\hat{\Delta}_1^s, \ldots, \hat{\Delta}_m^s\right)^t$ les m discontinuités de domaine observées dans la s^{e} réplique de la simulation par MCMC et $\boldsymbol{\Delta}^s = \left(\Delta_1^s, \ldots, \Delta_m^s\right)^t$ les vraies valeurs des m discontinuités de domaine dans la réplique s^{e} de la simulation par MCMC. Si un modèle ajuste les données observées adéquatement, on s'attend à ce que $T\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}^s\right)$ se trouve dans la majeure partie de l'histogramme des répliques $T\left(\hat{\boldsymbol{\Delta}}^s, \boldsymbol{\Delta}^s\right)$. Par

conséquent, les valeurs p proches de zéro ou un indiquent un mauvais ajustement pour ce qui est de cette statistique de test. Dans les expressions ci-dessous, il est entendu que T_x, pour $x = 1, \ldots, 6$, est une fonction de $(\hat{\boldsymbol{\Delta}}^s, \boldsymbol{\Delta}^s)$ ou $(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}^s)$, selon le composant évalué dans (3.12). On définit les tests prédictifs a posteriori suivants (You, 2008) :

1. Une statistique générale de test de la qualité de l'ajustement $T_1 = \sum_{i=1}^{m} (\hat{\Delta}_i - \Delta_i)^2 \big/ \mathrm{Var}(\hat{\Delta}_i \,|\, \Delta_i)$. Ici $\mathrm{Var}(\hat{\Delta}_i \,|\, \Delta_i) = \psi_i^r + \psi_i^a$;
2. $T_2 = \max(\hat{\Delta}_i)$ et $T_3 = \min(\hat{\Delta}_i)$ qui sont sensibles aux écarts dans la queue de la distribution;
3. $T_4 = \frac{1}{m} \sum_{i=1}^{m} \hat{\Delta}_i \equiv \bar{\hat{\Delta}}$, c'est-à-dire la moyenne qui est sensible au biais dans les prédictions de domaine;
4. $T_5 = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\Delta}_i - \bar{\hat{\Delta}})^2$, c'est-à-dire la variance des estimations de domaine, qui est sensible, par exemple, au rétrécissement excessif;
5. $T_6 = |\max(\hat{\Delta}_i) - \bar{\Delta}| - |\min(\hat{\Delta}_i) - \bar{\Delta}|$, avec $\bar{\Delta} = \frac{1}{m} \sum_{i=1}^{m} \Delta_i$ qui est sensible à l'asymétrie dans la distribution.

4. Résultats

4.1 Sélection du modèle

Dans la sous-section 3.2, deux versions différentes des effets fixes (`FE_uq` et `FE_eq`) et trois structures de covariance différentes des effets aléatoires (`RE_f`, `RE_d` et `RE_s`) sont examinées pour le modèle de FH bivarié. On applique la procédure de sélection ascendante de la sous-section 3.6 à chacune de ces six combinaisons séparément pour sélectionner les covariables. Nous avons vu dans la sous-section 3.1 que, pour le modèle de FH bivarié et le modèle de FH univarié aux fins des discontinuités, les covariables potentielles peuvent se trouver dans la *Municipal Base Administration* et le Registre des infractions déclarées par la police. Les noms de ces covariables commencent respectivement par `MBA_` et `PR_`. Pour le modèle de FH univarié selon l'enquête de rechange, les estimations directes de l'EVC ordinaire sont également considérées comme des covariables (van den Brakel et coll., 2016). Les noms de ces covariables commencent par `CVSR_`. Voir l'aperçu des covariables dans l'annexe.

Le tableau 4.1 résume les modèles finalement sélectionnés pour le modèle de FH bivarié. Les modèles présentés dans le tableau 4.1 sont sélectionnés selon la procédure `étape-WAIC-et`. La procédure `étape-WAIC` permet de sélectionner des modèles comportant une quantité nettement plus grande de covariables qui n'améliorent que légèrement le WAIC. Pour `offtot` et `unsafe`, l'`étape-WAIC` donne un modèle avec 4 covariables ayant des matrices de covariance diagonale pour les effets aléatoires (`RE_d` et `RE_s`) et des coefficients de régression inégaux. Pour `satispol` et `propvict`, l'`étape-WAIC` donne un modèle ayant respectivement 3 et 2 covariables avec des coefficients de régression inégaux, et aussi avec des matrices de covariance diagonale ayant des variances égales pour les effets

aléatoires (RE_s). Avec seulement 25 domaines, le risque est grand que ces modèles surajustent les données. On a l'exception nuisance, où les deux procédures de sélection parviennent au même modèle.

Au moyen de la procédure étape-WAIC-et, on obtient des modèles plus parcimonieux indiqués ci-dessous dans le tableau 4.1. Pour les infractions totales, offtot et nuisance, un modèle avec une seule covariable aux coefficients de régression égaux pour les deux enquêtes (FE_eq) est obtenu en combinaison avec une matrice de covariance complète (RE_f), des effets de domaine aléatoires importants et une forte corrélation positive de 0,98 pour offtot et de 0,81 pour nuisance. On obtient aussi pour unsafe un modèle plus parcimonieux, avec une covariable et des coefficients de régression égaux (FE_eq) au moyen de la procédure étape-WAIC-et. Dans ce cas, une matrice de covariance diagonale avec des variances égales (RE_s) est sélectionnée. Pour propvict et satispol, la procédure étape-WAIC-et évite que soient sélectionnées de grandes quantités de covariables, ce qui se produit avec la méthode étape-WAIC. Le modèle sélectionné pour propvict possède une matrice de covariance complète avec une corrélation positive faible de 0,1 (RE_f) et une covariable avec des coefficients de régression inégaux (FE_uq). Le modèle pour satispol possède une matrice de covariance diagonale avec des variances égales (RE_s) et une covariable avec des coefficients de régression inégaux (FE_uq). Puisque les modèles parcimonieux sont privilégiés dans cette application, les modèles obtenus au moyen de la méthode étape-WAIC-et sont finalement sélectionnés. Voir van den Brakel et Boonstra (2018) pour lire une analyse plus détaillée de la sélection de modèle qui a donné les modèles finalement sélectionnés.

Le tableau 4.2 résume les modèles sélectionnés au moyen du modèle de FH univarié pour les estimations directes des discontinuités, élaborées à la sous-section 3.3. Les modèles sont sélectionnés selon la procédure étape-WAIC-et. Pour unsafe, l'étape-WAIC donne un modèle ayant quatre covariables. Pour les autres variables, les mêmes modèles sont sélectionnés comme au moyen de la procédure étape-WAIC-et. Les modèles de FH univariés élaborés dans van den Brakel et coll. (2016) pour l'approche d'enquête de rechange sont résumés dans le tableau 4.3.

Les diagnostics du modèle standard testent les hypothèses sous-jacentes selon lesquelles les effets aléatoires de domaine et les résidus sont normalement et indépendamment distribués. Comme le nombre de domaines dans cette application est petit, la puissance des tests de normalité est faible et n'indique pas d'écarts par rapport à la normalité. Par conséquent, les tests prédictifs a posteriori résumés dans la sous-section 3.5 sont utilisés aux fins d'évaluation de l'adéquation du modèle. De plus, on compare les prédictions de domaine agrégées à l'échelle nationale aux estimations directes à l'échelle nationale afin d'évaluer le biais introduit par les procédures d'estimation sur petits domaines à la sous-section 4.2. Les valeurs p prédictives a posteriori pour les estimations de domaine des variables cibles et des discontinuités sont résumées dans le tableau 4.4 selon le modèle de FH bivarié et le tableau 4.5 selon le modèle de FH univarié pour les discontinuités. La mesure générale de l'adéquation de l'ajustement (T_1)

indique que l'ajustement des discontinuités de `offtot` est de qualité réduite (les autres modèles examinés avaient des valeurs élevées semblables). Les valeurs pour le modèle de FH bivarié sont légèrement supérieures à celles du modèle de FH univarié pour les discontinuités. Les valeurs p prédictives a posteriori pour les valeurs maximales (T_2) et minimales (T_3) n'indiquent pas de problèmes dans les queues des distributions. Pour ces valeurs p prédictives a posteriori, il n'y a pas de différences systématiques entre les modèles de FH bivarié et univarié. Les valeurs de T_4, T_5 et T_6 pour les discontinuités du modèle bivarié sont comparables aux valeurs du modèle univarié. Les valeurs prédictives a posteriori pour la moyenne (T_4) et l'asymétrie de la distribution (T_6) indiquent que les distributions sont concentrées symétriquement autour de leur moyenne. Les valeurs p prédictives a posteriori pour la variance (T_5) indiquent un rétrécissement quelque peu insuffisant pour les discontinuités de `nuisance`, `propvict` et `offtot` selon les modèles de FH bivarié et univarié.

Tableau 4.1
Modèles finaux du modèle de FH bivarié sélectionnés au moyen de l'`étape-WAIC-et`. Tous les modèles contiennent une ordonnée à l'origine. ρ : corrélation entre les effets aléatoires

Variable	Modèle	Effets aléatoires de la structure de covariance			
		Type	σ_r^2	σ_a^2	ρ
offtot	FE_eq: δ_i + PR_weapon	RE_f	8,77	5,32	0,98
unsafe	FE_eq: δ_i + PR_propcrim	RE_s	1,17	1,17	-
nuisance	FE_eq: δ_i + MBA_immigrnw	RE_f	0,20	0,14	0,81
satispol	FE_uq: MBA_immigr	RE_s	0,78	0,78	-
propvict	FE_uq: PR_propcrim	RE_f	0,79	0,39	0,2

Tableau 4.2
Modèles finaux du modèle de FH univarié pour les estimations directes des discontinuités. Tous les modèles contiennent une ordonnée à l'origine

Variable	Modèle	Effets aléatoires de la variance (σ_v^2)
offtot	PR_propcrim	0,80
unsafe	MBA_benefit	1,03
nuisance	PR_threat	0,038
satispol	MBA_benefit	0,928
propvict	PR_assault	0,485

Tableau 4.3
Modèles finaux du modèle de FH univarié selon l'EVC de rechange, tirés de van den Brakel et coll. (2016). Tous les modèles contiennent une ordonnée à l'origine

Variable	Modèle	Effets aléatoires de la variance (σ_v^2)
offtot	CVSR_victim	0,003
unsafe	CVSR_nuisance + MBA_benefit + PR_propcrim + PR_drugs	2,997
nuisance	CVSR_nuisance + MBA_old	0,805
satispol	CVSR_funcpol	4,995
propvict	PR_propcrim + MBA_old	7,725

Tableau 4.4
Valeurs *p* prédictives a posteriori pour les modèles de FH multivariés finaux du tableau 4.1

Variable	T_1	T_2	T_3	T_4	T_5	T_6
	Discontinuités					
offtot	0,980	0,797	0,069	0,337	0,968	0,416
unsafe	0,343	0,841	0,833	0,454	0,437	0,912
nuisance	0,927	0,940	0,034	0,345	0,988	0,465
satispol	0,772	0,595	0,392	0,610	0,762	0,484
propvict	0,925	0,261	0,029	0,258	0,970	0,070
	Variables cibles					
offtot	0,859	0,249	0,024	0,317	0,524	0,089
unsafe	0,308	0,779	0,492	0,420	0,474	0,708
nuisance	0,766	0,317	0,108	0,433	0,504	0,156
satispol	0,742	0,929	0,584	0,457	0,875	0,797
propvict	0,695	0,339	0,168	0,379	0,655	0,194

Tableau 4.5
Valeurs *p* prédictives a posteriori pour les modèles de FH univariés finaux pour les estimations directes des discontinuités du tableau 4.2

Variable	T_1	T_2	T_3	T_4	T_5	T_6
	Discontinuités					
offtot	0,985	0,828	0,071	0,390	0,972	0,438
unsafe	0,382	0,885	0,816	0,464	0,523	0,920
nuisance	0,970	0,941	0,072	0,434	0,978	0,554
satispol	0,814	0,607	0,378	0,607	0,783	0,488
propvict	0,946	0,272	0,052	0,383	0,963	0,092

4.2 Résultats de l'estimation

Dans cette sous-section, nous traitons des résultats de l'estimation selon chacune des trois approches de modélisation. Dans la sous-section 4.2.1, les prédictions HB pour les variables cibles de l'enquête ordinaire et de l'enquête de rechange obtenues avec le modèle de FH bivarié sont comparées aux estimations directes et aux prédictions de domaine obtenues avec le modèle de FH univarié, où les estimations directes de l'approche ordinaire sont des variables auxiliaires potentielles dans la sélection du modèle. Les résultats subséquents des discontinuités de domaine sont abordés à la sous-section 4.2.2. Ici, les résultats obtenus avec le modèle de FH univarié pour les discontinuités sont également analysés.

Avec l'estimation sur petits domaines fondée sur un modèle, la variance par rapport au plan des estimateurs directs est réduite au coût de l'acceptation d'une certaine quantité de biais par rapport au plan. Pour évaluer les différences entre les estimations ponctuelles directes et les prédictions sur petits domaines, les deux mesures suivantes sont définies. La première est la différence relative moyenne (DRM), qui résume les différences entre les estimations directes et les prédicions de domaine :

$$\text{DRM} = \frac{100\ \%}{m} \sum_{i=1}^{m} \frac{\hat{y}_i^q - \tilde{y}_i^q}{\hat{y}_i^q},\ q = r,\ a, \tag{4.1}$$

et $\tilde{y}_i^q$ est la prédiction de domaine fondée sur le modèle de FH bivarié ou le modèle de FH univarié. La deuxième mesure est la différence relative moyenne absolue (DRMA) entre l'estimation directe et la prédiction de domaine, qui est définie comme suit :

$$\text{DRMA} = \frac{100\ \%}{m} \sum_{i=1}^{m} \left| \frac{\hat{y}_i^q - \tilde{y}_i^q}{\hat{y}_i^q} \right|, \ q = r, a, \tag{4.2}$$

la précision accrue des prédictions sur petits domaines est mesurée au moyen de la différence relative moyenne des erreurs-types (DRMET) entre les estimations directes et les prédictions de domaine et est définie comme suit :

$$\text{DRMET} = \frac{100\ \%}{m} \sum_{i=1}^{m} \frac{\text{ET}\left(\hat{y}_i^q\right) - \text{ET}\left(\tilde{y}_i^q\right)}{\text{ET}\left(\hat{y}_i^q\right)}, \ q = r, a, \tag{4.3}$$

ces mesures sont définies de manière similaire pour les estimations et les prédictions des discontinuités de domaine $\hat{\Delta}_i$ et $\tilde{\Delta}_i$.

4.2.1 Résultats pour les variables selon l'enquête ordinaire et l'enquête de rechange

Dans le tableau 4.6, les prédictions de domaine et leurs erreurs-types calculées en moyenne sur les domaines ainsi que la DRM, la DRMA et la DRMET sont données pour l'enquête de rechange selon le modèle de FH univarié avec les modèles présentés dans le tableau 4.3. Les résultats selon le modèle de FH bivarié, fondés sur les modèles finaux du tableau 4.1, sont présentés dans le tableau 4.7 pour les variables de l'enquête de rechange et dans le tableau 4.8 pour les variables de l'enquête ordinaire. La comparaison des erreurs-types (ET) et de la DRMET dans les tableaux 4.6 et 4.7 montre que le modèle de FH bivarié entraîne des réductions plus grandes des erreurs-types pour toutes les variables, à l'exception de `nuisance`. Cela se produit au coût d'un biais accru. La comparaison de la DRM et de la DRMA dans les deux tableaux montre que les écarts entre les estimations directes et les prédictions sur petits domaines sont plus importants dans le modèle de FH bivarié. La comparaison de l'ET et de la DRMET dans les tableaux 4.7 et 4.8 montre que l'amélioration de la précision au moyen du modèle de FH bivarié selon l'enquête ordinaire est plus petite, comme cela était prévisible en raison de la plus grande taille d'échantillon de cette dernière. Le biais dans les prédictions du modèle de FH bivarié pour l'enquête ordinaire est également plus faible, ce qui découle d'une comparaison de la DRM et de la DRMA dans les tableaux 4.7 et 4.8.

Les prédictions de domaine selon le modèle de FH univarié et bivarié sont représentées graphiquement par rapport aux estimations ERG dans les figures 4.1 à 4.5. Les graphiques contiennent également l'estimation ERG à l'échelle nationale par rapport aux prédictions de domaine agrégées à l'échelle nationale selon (4.4). Les figures 4.1 et 4.3 montrent qu'il n'y a qu'un faible rétrécissement pour

offtot et nuisance. La figure 4.2 montre pour unsafe que le modèle de FH bivarié rétrécit les prédictions de domaine selon l'enquête de rechange, tandis que le degré de rétrécissement est plus petit pour le modèle de FH univarié selon l'EVC de rechange et le modèle de FH bivarié selon l'enquête ordinaire. Pour propvict (voir la figure 4.4), il y a une petite différence entre le degré de rétrécissement de l'EVC de rechange selon le modèle bivarié et le modèle univarié. D'après la figure 4.5, il s'ensuit que le modèle de FH bivarié pour satispol ne peut pas modéliser adéquatement les observations dans l'enquête de rechange avec l'information auxiliaire des deux registres (MBA et RIDP). Dans ce cas, les prédictions de domaine de satispol selon l'approche de rechange comportent un rétrécissement extrêmement excessif. En effet, le modèle de FH univarié sélectionne la même variable auxiliaire de l'enquête ordinaire seulement (voir le tableau 4.3) et donne des prédictions de domaine plus réalistes.

Pour les variables liées aux opinions et points de vue telles que unsafe et satispol, la réduction des erreurs-types s'accompagne d'une augmentation relativement forte du biais. Cela est particulièrement le cas pour la prédiction sur petits domaines du modèle de FH bivarié selon l'enquête de rechange. Pour ces variables, il n'y a pas de covariables fortement corrélées dans la MBA et le RIDP. Dans ces cas, le modèle de FH univarié donne de meilleurs résultats puisque les covariables connexes de l'enquête ordinaire sont sélectionnées (voir le tableau 4.3), tandis que le modèle bivarié ne détecte pas la corrélation entre les effets aléatoires (voir le tableau 4.1).

Tableau 4.6
Moyenne des prédictions de domaine selon l'enquête de rechange avec le modèle de FH univarié, tirée de van den Brakel et coll. (2016)

Variable	Est. HB	ET	DRM (%)	DRMA (%)	DRMET (%)
offtot	33,21	2,90	-0,44	7,03	47,74
unsafe	19,83	1,64	-0,96	7,58	41,16
nuisance	1,29	0,08	-0,74	5,02	37,96
satispol	55,09	2,54	-0,11	6,43	61,98
propvict	9,85	0,84	-3,17	11,86	60,69

Tableau 4.7
Moyenne des prédictions de domaine selon l'enquête de rechange avec le modèle de FH bivarié

Variable	Est. HB	ET	DRM (%)	DRMA (%)	DRMET (%)
offtot	33,26	2,82	-0,99	6,93	49,36
unsafe	19,82	1,21	-2,54	11,97	56,47
nuisance	1,28	0,08	-0,98	4,28	35,72
satispol	55,08	1,97	-0,49	8,97	70,06
propvict	9,91	0,73	-4,81	14,70	65,35

Tableau 4.8
Moyenne des prédictions de domaine selon l'enquête ordinaire avec le modèle de FH bivarié

Variable	Est. HB	ET	DRM (%)	DRMA (%)	DRMET (%)
`offtot`	41,34	3,76	0,96	4,56	17,95
`unsafe`	24,22	1,07	-0,01	6,02	46,78
`nuisance`	1,60	0,09	0,38	2,62	15,93
`satispol`	60,82	1,47	-0,77	5,38	64,83
`propvict`	12,18	0,88	1,58	7,84	43,70

Les estimations directes à l'échelle nationale sont des estimations exactes puisqu'elles sont fondées sur des échantillons dont la taille est suffisamment grande. Par conséquent, on évalue souvent le biais dans les prédictions de domaine fondées sur un modèle en comparant les estimations directes à l'échelle nationale aux prédictions de domaine agrégées à l'échelle nationale. Les variables cibles de cette application sont toutes définies comme étant des moyennes de population. Par conséquent, les prédictions de domaine agrégées sont obtenues comme étant la moyenne sur les domaines pondérée avec les tailles de domaine relatives.

$$\tilde{y}^q = \sum_{i=1}^{m} \frac{N_i}{N} \tilde{y}_i^q \tag{4.4}$$

avec N_i la taille de la population du domaine i et N la taille de la population totale.

Le tableau 4.9 compare la moyenne pondérée des prédictions de domaine selon (4.4) avec les estimations par la régression généralisée nationales. Pour le modèle de FH univarié selon l'EVC de rechange, les prédictions de domaine agrégées sont presque exactement égales aux estimations par la régression généralisée à l'échelle nationale. Pour le modèle de FH bivarié, les différences sont légèrement plus grandes, mais les prédictions de domaine agrégées sont encore très proches des estimations par la régression généralisée à l'échelle nationale. La plus grande différence relative s'élève à 3 % et est observée pour `offtot` selon l'enquête ordinaire.

Tableau 4.9
Estimations par la régression généralisée et prédictions HB nationales et agrégées selon l'approche d'enquête ordinaire et de rechange (4.4)

Variable	Ordinaire		De rechange			Discontinuité			
	ERG	biv. FH	ERG	biv. FH	uni. FH	ERG	biv. FH	uni. FH	Δ FH *)
`offtot`	43,79	42,47	34,09	34,02	34,09	9,7	8,45	9,7	9,04
`unsafe`	25,07	24,89	20,48	20,49	20,48	4,59	4,40	4,59	4,69
`nuisance`	1,67	1,66	1,34	1,34	1,34	0,33	0,32	0,34	0,33
`satispol`	59,88	60,36	55,10	55,06	55,12	4,78	5,29	5,04	5,07
`propvict`	13,02	12,76	10,32	10,33	10,32	2,70	2,43	2,70	2,63

*) : Δ FH sont les prédictions HB avec le modèle de FH univarié pour les estimations directes des discontinuités, pondérées de la même façon que (4.4).

Figure 4.1 Estimations de domaine ERG comparativement aux prédictions HB du total des infractions. Graphique supérieur : enquête régulière au moyen du modèle de FH bivarié; graphique central : enquête de rechange au moyen du modèle de FH bivarié; graphique inférieur : enquête de rechange au moyen du modèle de FH univarié. Les prédictions de domaine sont agrégées à l'échelle nationale selon (4.4).

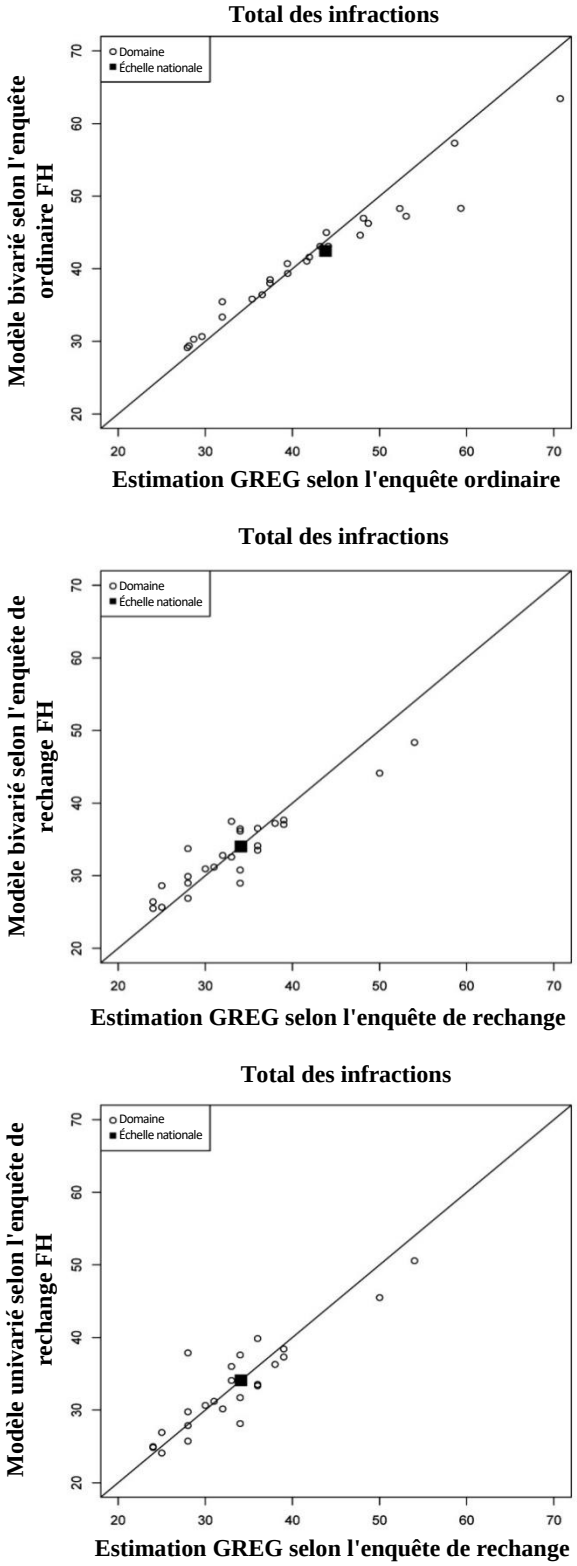


Figure 4.2 Estimations de domaine ERG comparativement aux prédictions HB du sentiment d'insécurité. Graphique supérieur : enquête régulière au moyen du modèle de FH bivarié; graphique central : enquête de rechange au moyen du modèle de FH bivarié; graphique inférieur : enquête de rechange au moyen du modèle de FH univarié. Les prédictions de domaine sont agrégées à l'échelle nationale selon (4.4).

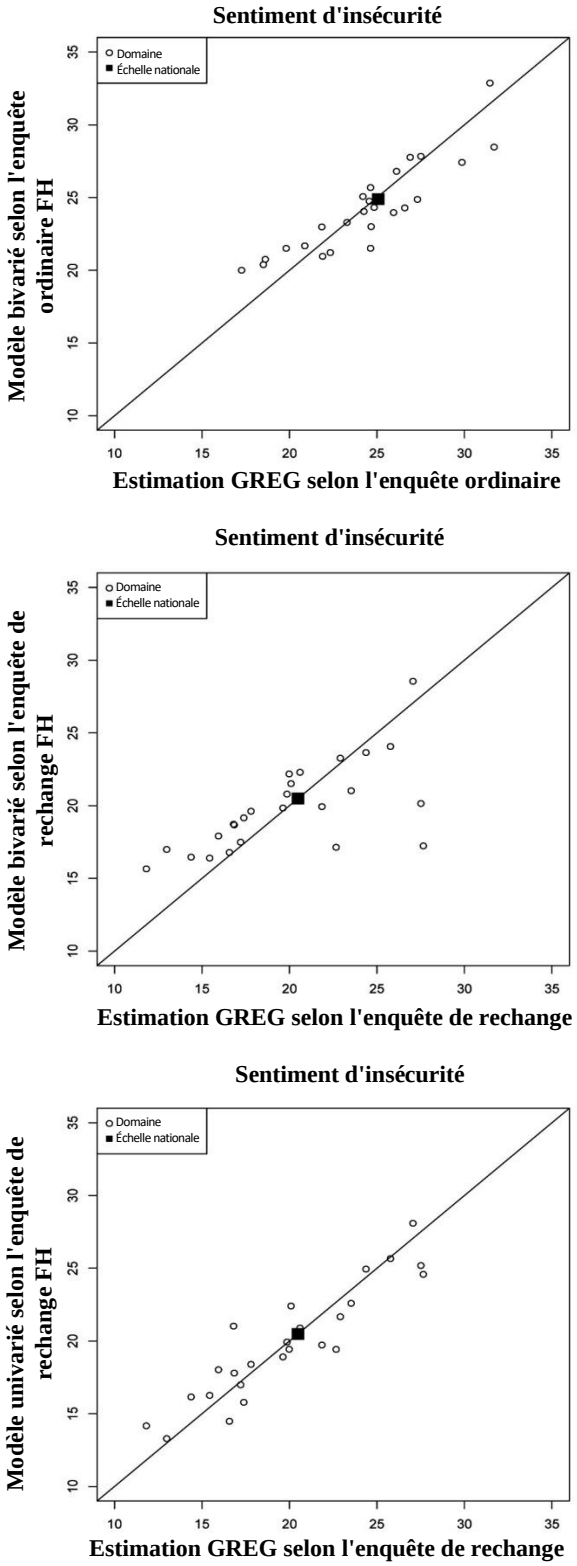


Figure 4.3 Estimations de domaine ERG comparativement aux prédictions HB des nuisances. Graphique supérieur : enquête régulière au moyen du modèle de FH bivarié; graphique central : enquête de rechange au moyen du modèle de FH bivarié; graphique inférieur : enquête de rechange au moyen du modèle de FH univarié. Les prédictions de domaine sont agrégées à l'échelle nationale selon (4.4).

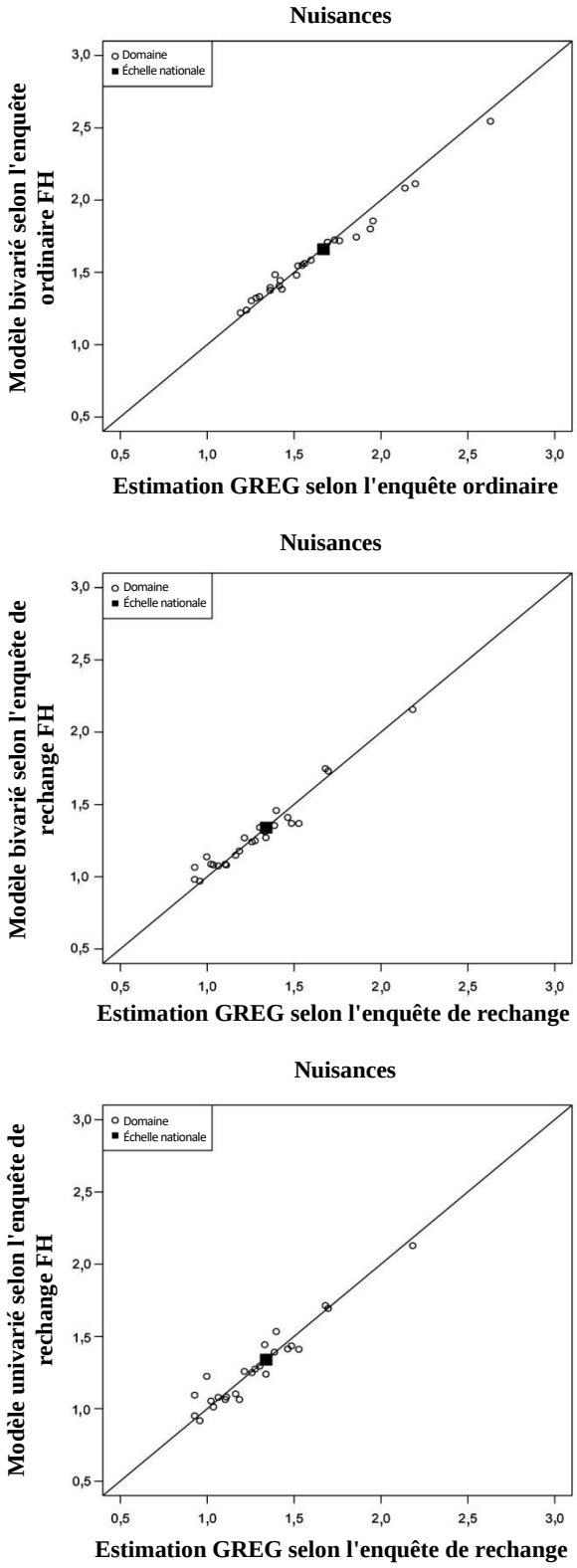


Figure 4.4 Estimations de domaine ERG comparativement aux prédictions HB des crimes contre des biens. Graphique supérieur : enquête régulière au moyen du modèle de FH bivarié; graphique central : enquête de rechange au moyen du modèle de FH bivarié; graphique inférieur : enquête de rechange au moyen du modèle de FH univarié. Les prédictions de domaine sont agrégées à l'échelle nationale selon (4.4).

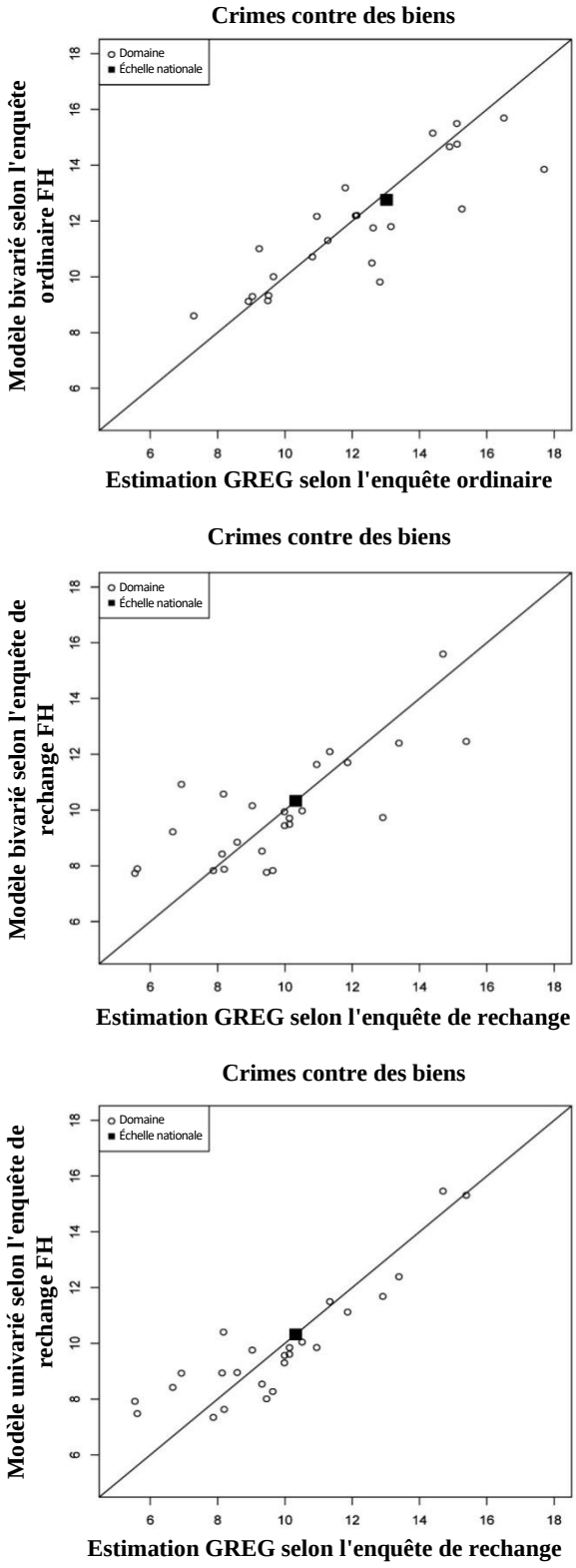
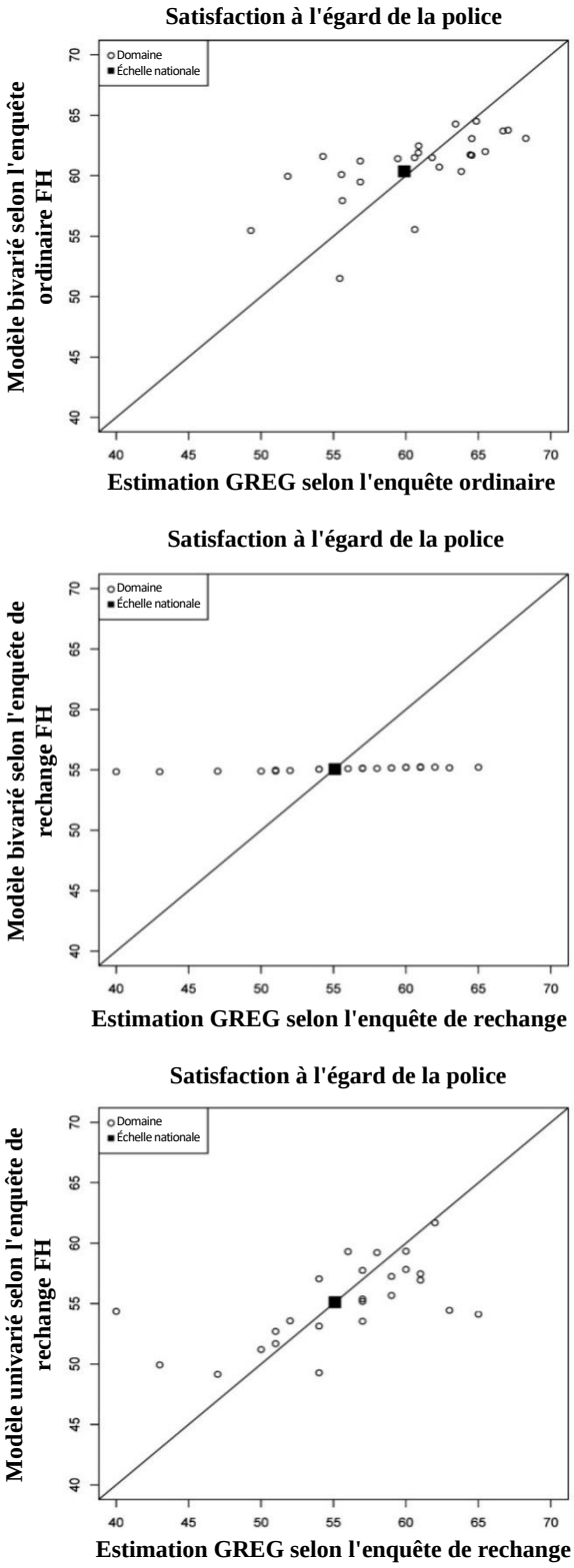


Figure 4.5 Estimations de domaine ERG comparativement aux prédictions HB de la satisfaction à l'égard de la police. Graphique supérieur : enquête régulière au moyen du modèle de FH bivarié; graphique central : enquête de rechange au moyen du modèle de FH bivarié; graphique inférieur : enquête de rechange au moyen du modèle de FH univarié. Les prédictions de domaine sont agrégées à l'échelle nationale selon (4.4).



4.2.2 Résultats des estimations de discontinuité

Dans les quatre dernières colonnes du tableau 4.9, les estimations par la régression généralisée des discontinuités à l'échelle nationale sont comparées aux prédictions de domaine obtenues par le modèle de FH univarié pour l'EVC de rechange, le modèle de FH bivarié et le modèle de FH univarié pour les discontinuités, agrégées à l'échelle nationale au moyen de (4.4). Les plus grandes différences entre les estimations ERG pour les discontinuités à l'échelle nationale et les prédictions de domaine agrégées sont obtenues pour le modèle bivarié et les plus petites pour le modèle de FH univarié selon l'EVC de rechange. On pouvait s'y attendre puisque le modèle de FH bivarié rétrécit à la fois les estimations de domaine pour l'enquête ordinaire et l'enquête de rechange. Au moyen du modèle de FH univarié pour l'EVC de rechange, seules les estimations pour l'enquête de rechange sont remplacées par des prédictions de domaine, tandis que les estimations pour l'enquête ordinaire ne sont pas corrigées. De plus, les prédictions de domaine pour l'enquête de rechange ont des DRM et des DRMA plus grandes selon le modèle de FH bivarié comparativement au modèle de FH univarié (comparer les tableaux 4.6 et 4.7). Les différences selon le modèle de FH univarié pour les discontinuités sont plus petites comparativement au modèle de FH bivarié, mais plus grandes comparativement au modèle de FH univarié pour l'EVC de rechange.

Dans les tableaux 4.10, 4.11 et 4.12, les prédictions de domaine et leurs erreurs-types pour les discontinuités calculées en moyenne sur les domaines ainsi que la DRM et la DRMET sont résumées respectivement selon le modèle de FH univarié pour l'EVC de rechange, le modèle de FH bivarié et le modèle de FH univarié pour les discontinuités. Les DRM sont grandes parce que les estimations ERG pour les discontinuités dans le dénominateur de (4.11) prennent souvent des valeurs proches de zéro, ce qui rend ces indicateurs instables. C'est pourquoi la DRMA est remplacée par la médiane des différences relatives absolues, $\left|\left(\hat{y}_i^q - \tilde{y}_i^q\right)/\hat{y}_i^q\right|$, et est abrégée en MDRA. Ces dernières sont en effet des indicateurs plus stables du biais. La plus petite MDRA est obtenue selon le modèle de FH univarié pour l'EVC de rechange, car cette méthode corrige seulement les prédictions de domaine de l'EVC de rechange. Quant aux valeurs de la MDRA pour le modèle de FH bivarié, elles sont plus petites que celles du modèle de FH univarié pour les discontinuités.

À l'exception des erreurs-types `nuisance` pour les prédictions de domaine selon le modèle de FH bivarié, elles sont plus petites que celles du modèle de FH univarié pour l'EVC de rechange. Dans le cas de `propvict` et `offtot`, il s'agit du résultat de prédictions de domaine légèrement plus précises pour l'enquête de rechange par rapport au modèle de FH univarié (comparer les tableaux 4.6 et 4.7), ce qui représente une nette amélioration de la précision des prédictions de domaine de l'enquête ordinaire comparativement aux estimateurs ERG (tableaux 4.8 et 2.2) et la corrélation positive entre les effets aléatoires. Dans le cas de `satispol` et `unsafe`, cela est principalement le résultat d'une nette amélioration de la précision des prédictions de domaine selon le modèle de FH bivarié pour les estimateurs ordinaires comparativement aux estimateurs ERG (tableaux 4.8 et 2.2) et aussi une nette amélioration de la précision des prédictions de domaine selon le modèle de FH bivarié pour l'enquête de rechange comparativement au modèle univarié (comparer le tableau 4.6 au tableau 4.7).

Pour les cinq variables, les plus petites erreurs-types sont obtenues selon le modèle de FH univarié pour les discontinuités. Cela se fait au coût d'un biais plus important, comme le montrent les valeurs de la MDRA. On a l'exception `satispol`, pour laquelle le biais en termes de MDRA selon le modèle de FH bivarié est manifestement plus important que selon le modèle de FH univarié pour les discontinuités. Pour cette variable, on obtient le biais le plus faible au moyen du modèle de FH univarié pour l'EVC de rechange, mais la réduction des erreurs-types est également plus petite.

Les dernières colonnes des tableaux 4.11 et 4.12 contiennent les facteurs de rétrécissement des discontinuités de domaine calculées en moyenne sur les domaines. Selon le modèle de FH univarié pour les discontinuités, les facteurs de rétrécissement pour les prédictions des discontinuités de domaine, c'est-à-dire les poids donnés à l'estimateur direct pour les discontinuités, sont définis comme étant $\gamma_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + (\psi_i^r + \psi_i^a))$. Selon le modèle bivarié, les facteurs de rétrécissement pour les prévisions des discontinuités de domaine sont définis comme étant $\gamma_i = \boldsymbol{\iota}\hat{\boldsymbol{\Sigma}}\boldsymbol{\iota}^t / (\boldsymbol{\iota}\hat{\boldsymbol{\Sigma}}\boldsymbol{\iota}^t + (\psi_i^r + \psi_i^a))$, avec $\boldsymbol{\iota} = (1, -1)^t$. Le facteur de rétrécissement moyen est défini comme étant $\bar{\gamma} = 1 / M \sum_{i=1}^{M} \gamma_i$. Notons que cette statistique n'est pas disponible pour les discontinuités obtenues au moyen du modèle de FH univarié pour l'EVC de rechange, puisque dans cette approche, les discontinuités de domaine sont obtenues comme étant le contraste entre l'estimation ERG pour l'enquête ordinaire et la prédiction de domaine pour l'approche de rechange. À l'exception de `satispol`, les facteurs de rétrécissement dans le modèle de FH univarié pour les discontinuités sont plus petits d'un facteur 10 que ceux du modèle de FH bivarié. La question se pose de savoir si les facteurs de rétrécissement extrêmement faibles du modèle de FH univarié pour les discontinuités rétrécissent excessivement les estimations directes des discontinuités.

Tableau 4.10
Prédictions de domaine des discontinuités selon le modèle de FH univarié pour l'EVC de rechange

Variable	Est. HB	ET	DRM (%)	MDRA (%)	DRMET (%)
`offtot`	9,08	3,92	-5,67	22,14	48,47
`unsafe`	4,55	2,46	42,98	23,44	29,45
`nuisance`	0,33	0,07	-5,80	11,49	57,49
`satispol`	5,52	4,72	99,26	47,72	40,86
`propvict`	2,70	1,83	-142,60	30,49	32,75

Tableau 4.11
Prédictions de domaine des discontinuités selon le modèle de FH bivarié

Variable	Est. HB	ET	DRM (%)	MDRA (%)	DRMET (%)	$\bar{\gamma}$
`offtot`	8,07	2,68	1,94	23,34	63,60	0,208
`unsafe`	4,40	1,56	55,49	41,24	55,09	0,317
`nuisance`	0,31	0,09	-4,01	18,43	44,47	0,327
`satispol`	5,74	2,46	228,50	73,43	68,75	0,019
`propvict`	2,27	1,10	-113,00	27,11	59,02	0,376

Tableau 4.12
Prédictions de domaine des discontinuités selon le modèle de FH univarié pour les estimations directes des discontinuités

Variable	Est. HB	ET	DRM (%)	MDRA (%)	DRMET (%)	$\bar{\gamma}$
offtot	8,35	2,25	-33,39	26,63	69,35	0,012
unsafe	4,43	1,48	45,45	42,14	59,99	0,082
nuisance	0,32	0,06	-13,85	20,97	63,04	0,049
satispol	5,67	2,39	186,07	65,33	69,61	0,014
propvict	2,54	0,91	-55,59	45,73	65,84	0,032

Les graphiques des discontinuités estimées au moyen de l'estimateur ERG, le modèle de FH univarié pour l'EVC de rechange, le modèle de FH bivarié et le modèle de FH univarié pour les discontinuités sont présentés dans les figures 4.6 à 4.10. Les prédictions pour les discontinuités de domaine obtenues par les trois modèles sont plus stables comparativement aux estimations par la régression généralisée. Cela apparaît manifestement, par exemple, avec unsafe (figure 4.7), où les estimations par la régression généralisée pour la discontinuité sont parfois positives et parfois négatives. Les prédictions des discontinuités de domaine selon le modèle de FH bivarié et le modèle de FH univarié pour les discontinuités sont constamment positives, ce qui semble plus plausible puisqu'il est peu probable que les discontinuités de domaine aient des signes opposés. Les prédictions des discontinuités de domaine selon le modèle de FH univarié pour l'EVC de rechange sont plus proches des estimations par la régression généralisée et, par conséquent, moins stables. On observe une tendance semblable dans les autres variables.

Ces graphiques montrent que pour propvict, offtot et unsafe le modèle de FH bivarié entraîne une nette amélioration des prédictions pour les discontinuités de domaine comparativement au modèle de FH univarié pour l'EVC de rechange. Pour nuisance, les erreurs-types des discontinuités augmentent dans le modèle de FH bivarié comparativement au modèle de FH univarié pour l'EVC de rechange. Le modèle de FH bivarié pour satispol ne peut pas modéliser adéquatement les observations dans l'enquête de rechange avec l'information auxiliaire des deux registres (MBA et RIDP). Dans ce cas, les prédictions de domaine de satispol selon l'approche de rechange comportent un rétrécissement excessif. En effet, le modèle de FH univarié sélectionne une variable auxiliaire à partir de l'enquête ordinaire (voir le tableau 4.3) et donne manifestement de meilleures résultats.

On s'attendait à ce qu'il soit difficile de produire des prédictions raisonnables pour les discontinuités de domaine au moyen du modèle de FH univarié pour les estimations directes des discontinuités, étant donné qu'il est difficile d'imaginer que les variables auxiliaires disponibles dans des registres comme la MBA et le RIDP contiennent de bons prédicteurs des différences systématiques dans les erreurs d'enquête. Néanmoins, cette méthode plus pragmatique donne des résultats acceptables. Une des interprétations possibles serait que les discontinuités sont, dans une certaine mesure, proportionnelles aux valeurs de la variable cible et qu'elles montrent par conséquent une tendance systématique qui s'explique en partie par les covariables sélectionnées. Les très petits facteurs de rétrécissement dans ce modèle représentent un

sujet de préoccupation, ce qui pourrait indiquer que le modèle attribue trop de poids à l'estimateur synthétique.

Figure 4.6 Discontinuités du total des infractions fondées sur l'estimateur ERG (graphique supérieur), le modèle de FH univarié (deuxième graphique), le modèle de FH bivarié (troisième graphique) et le modèle de FH univarié pour les discontinuités des estimations directes (graphique inférieur) avec un intervalle de confiance de 95 %.

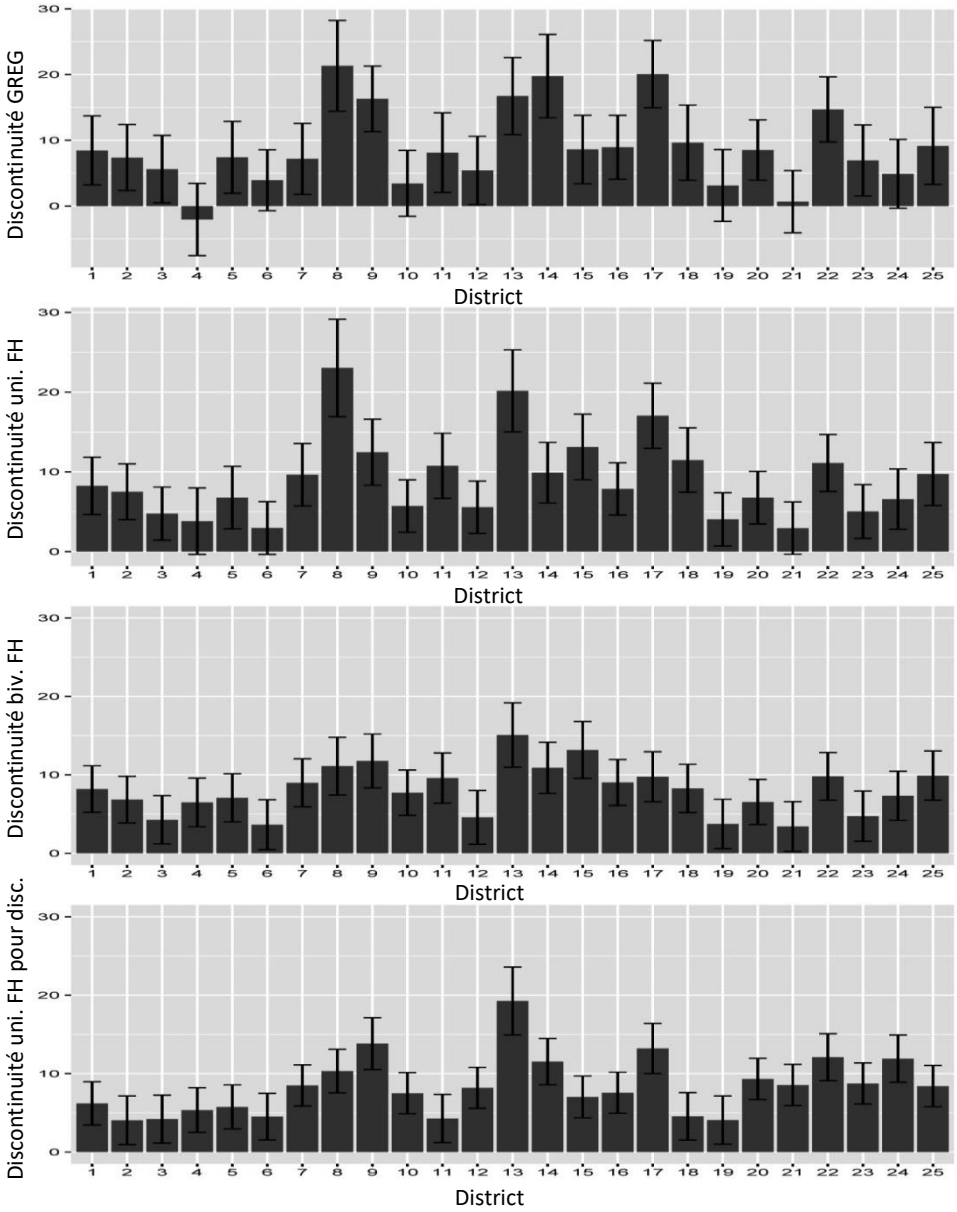


Figure 4.7 Discontinuités du sentiment d'insécurité fondées sur l'estimateur ERG (graphique supérieur), le modèle de FH univarié (deuxième graphique), le modèle de FH bivarié (troisième graphique) et le modèle de FH univarié pour les discontinuités des estimations directes (graphique inférieur) avec un intervalle de confiance de 95 %.

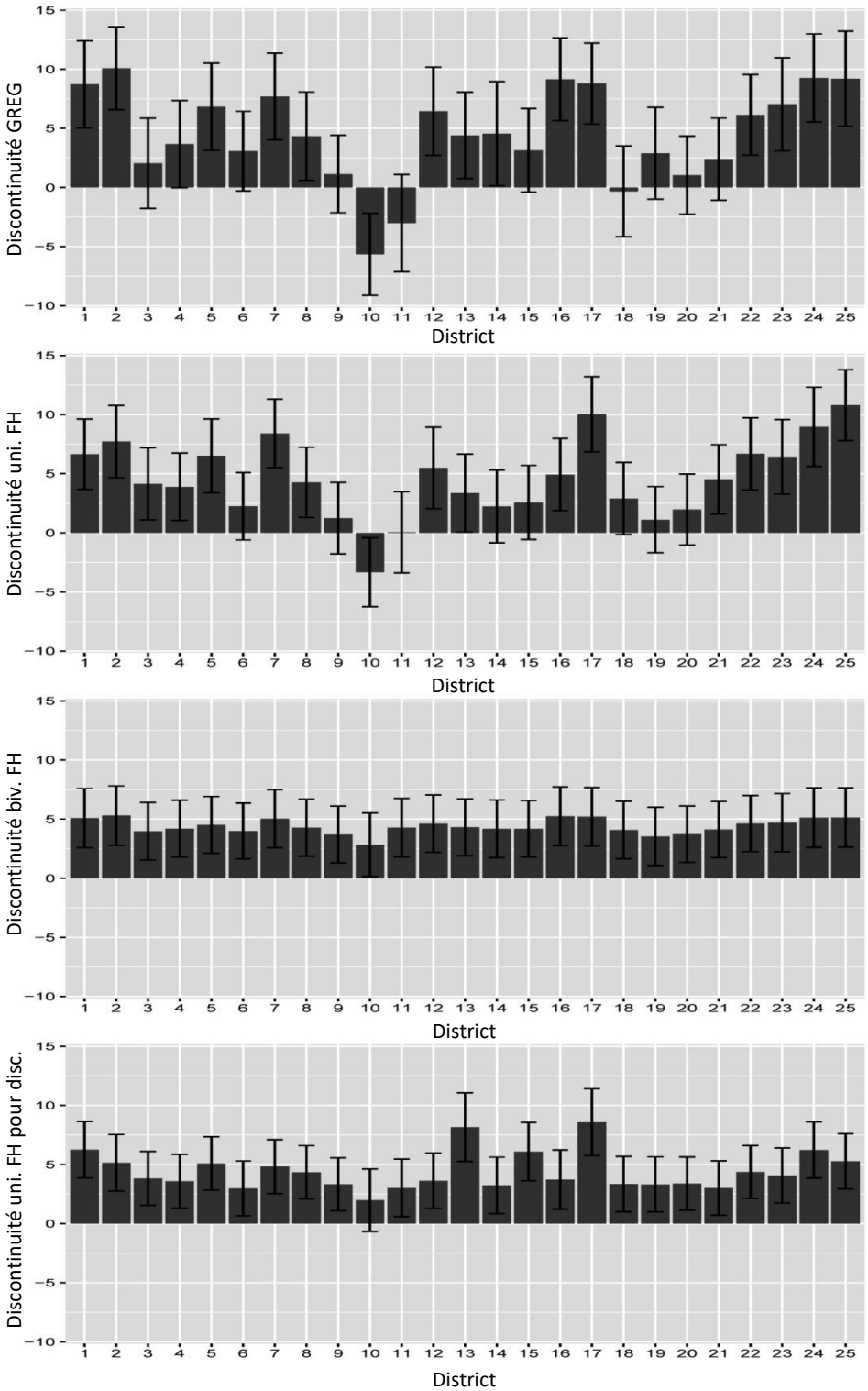


Figure 4.8 Discontinuités des nuisances fondées sur l'estimateur ERG (graphique supérieur), le modèle de FH univarié (deuxième graphique), le modèle de FH bivarié (troisième graphique) et le modèle de FH univarié pour les discontinuités des estimations directes (graphique inférieur) avec un intervalle de confiance de 95 %.

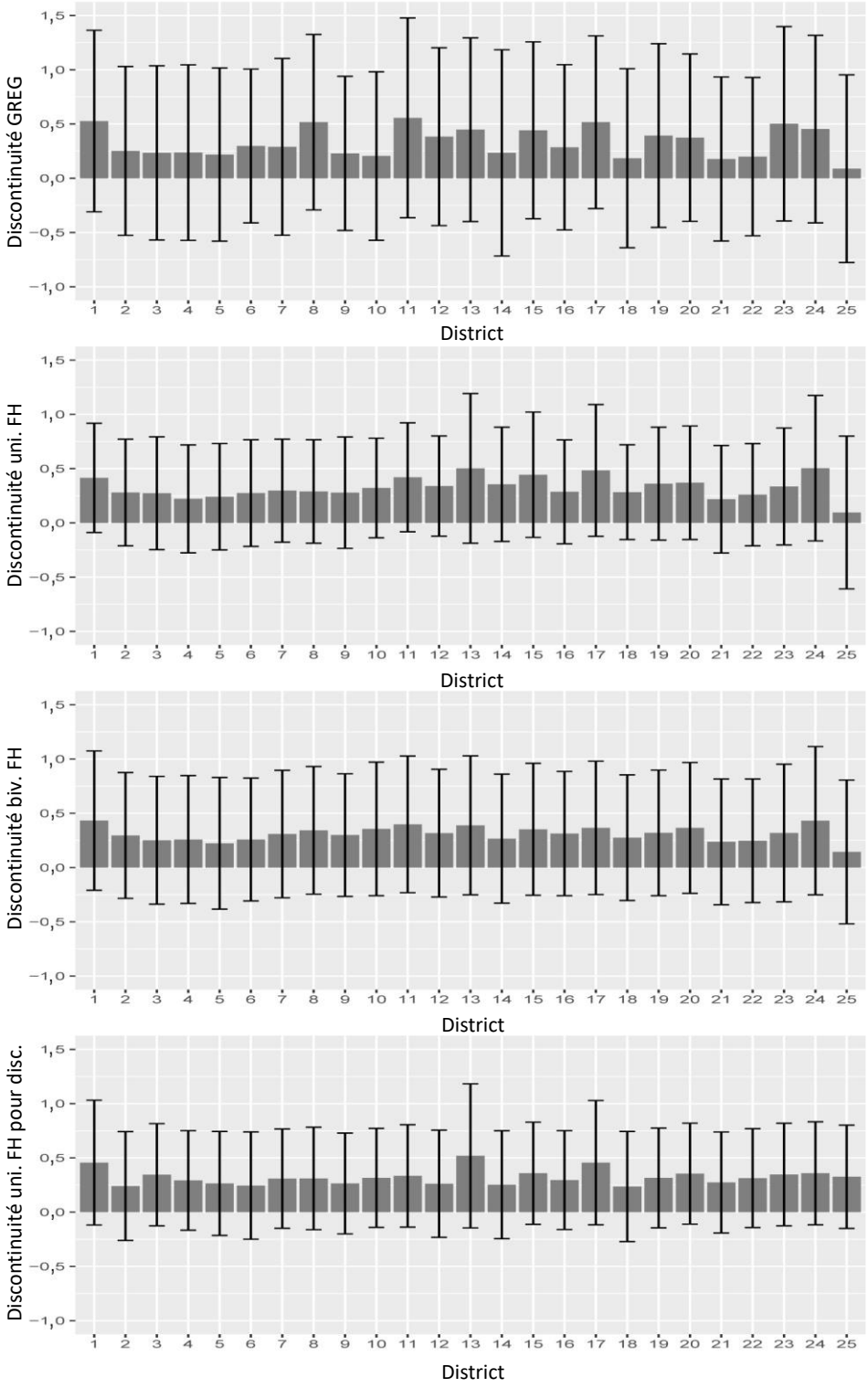


Figure 4.9 Discontinuités des crimes contre des biens fondées sur l'estimateur ERG (graphique supérieur), le modèle de FH univarié (deuxième graphique), le modèle de FH bivarié (troisième graphique) et le modèle de FH univarié pour les discontinuités des estimations directes (graphique inférieur) avec un intervalle de confiance de 95 %.

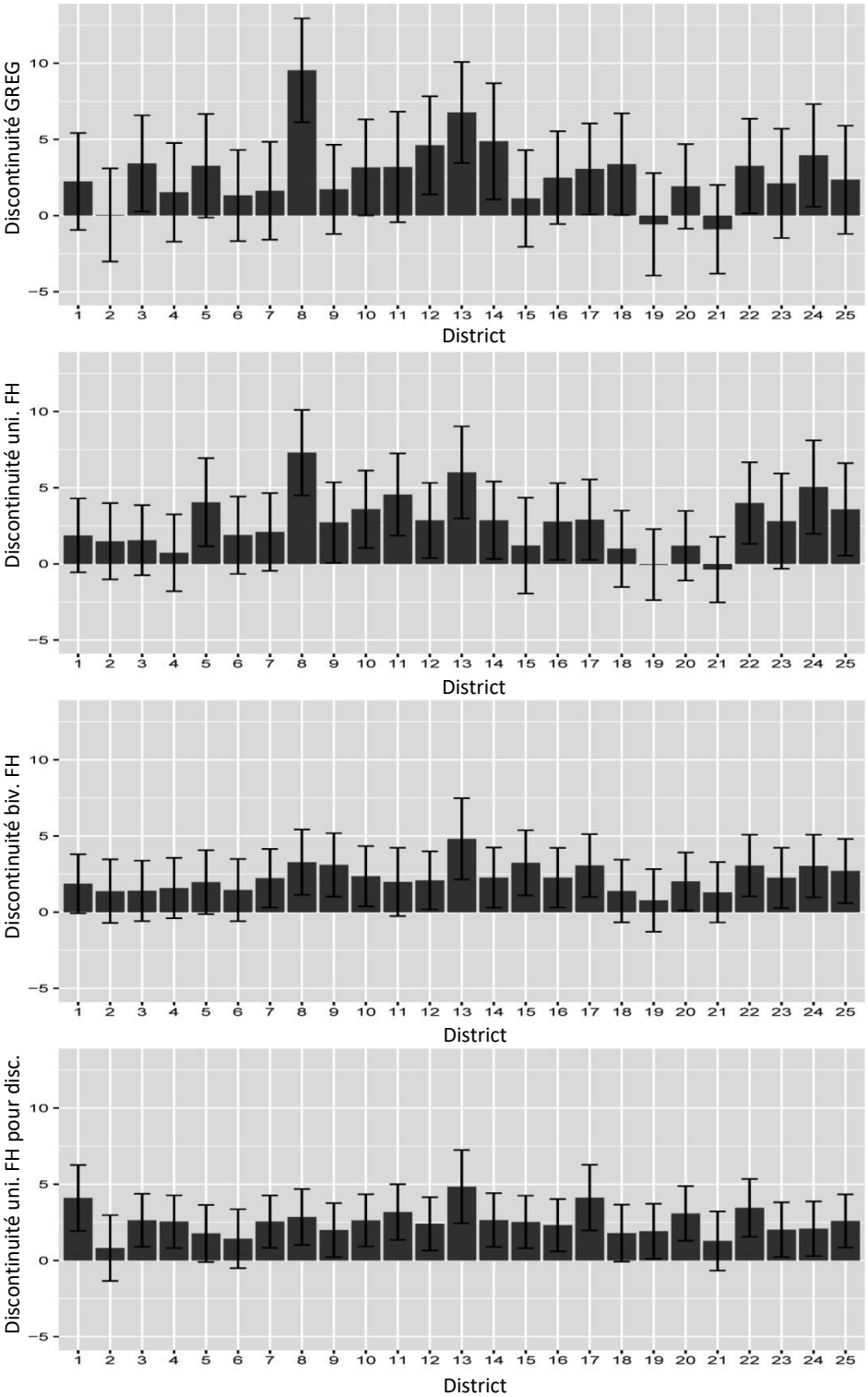
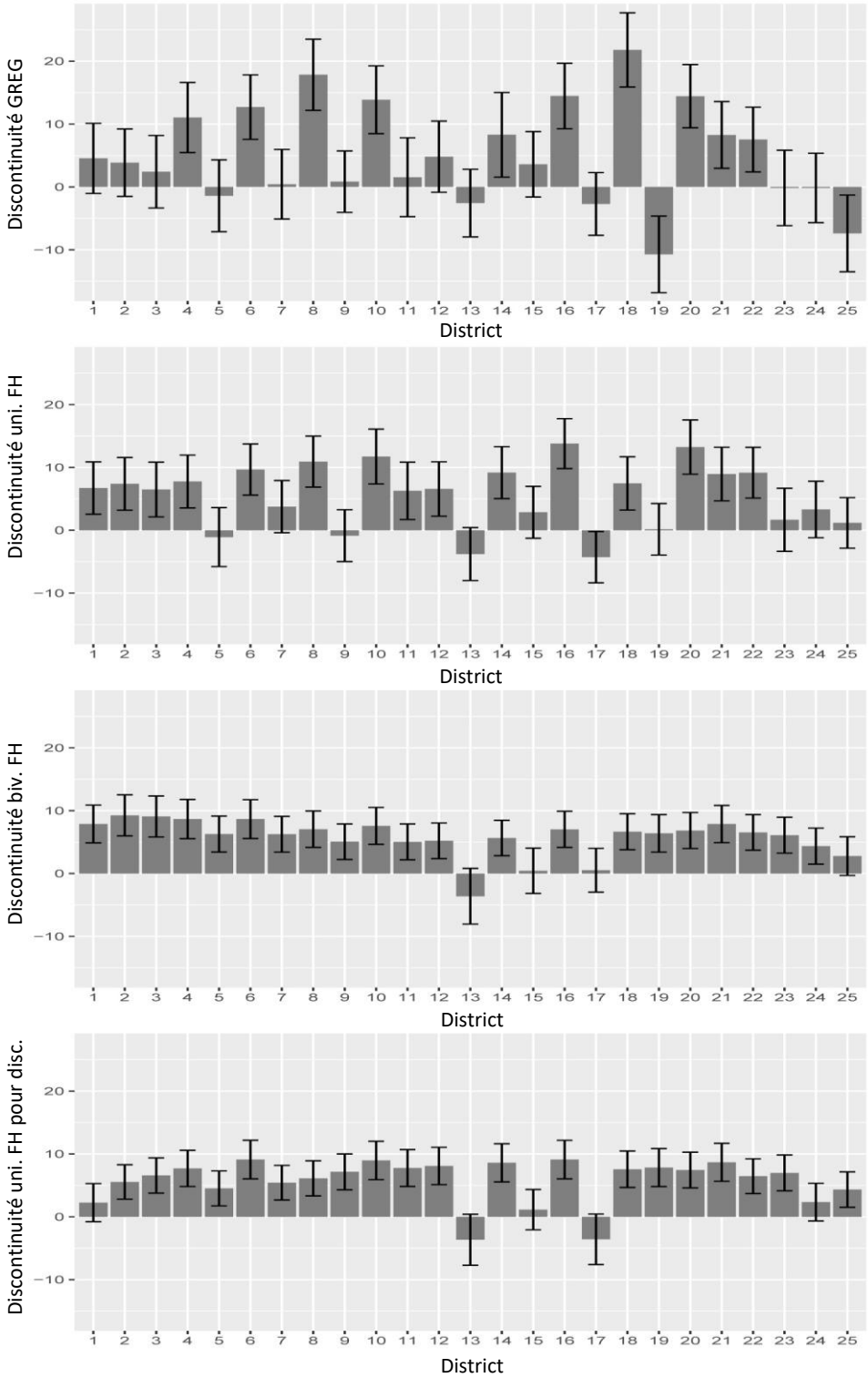


Figure 4.10 Discontinuités de la satisfaction à l'égard de la police fondées sur l'estimateur ERG (graphique supérieur), le modèle de FH univarié (deuxième graphique), le modèle de FH bivarié (troisième graphique) et le modèle de FH univarié pour les discontinuités des estimations directes (graphique inférieur) avec un intervalle de confiance de 95 %.



5. Discussion

Les remaniements de processus d'enquête entraînent souvent des discontinuités qui perturbent la comparabilité dans le temps des résultats obtenus au moyen d'une enquête répétée. Pour éviter de confondre les changements réels d'une période à l'autre avec les différences de biais de mesure, il est important que ces discontinuités soient quantifiées pendant la mise en œuvre d'un nouveau processus d'enquête. Une approche simple consiste à recueillir des données parallèlement selon le modèle antérieur et le nouveau modèle pendant une certaine période. Souvent, les budgets attribués à la collecte de données parallèle ne permettent pas de disposer des tailles d'échantillon minimales requises, qui proviennent des calculs de puissance aux fins de détection des différences préalablement spécifiées minimales à certains niveaux de signification et de puissance. Ces tailles pourraient suffire à la quantification des discontinuités à l'échelle nationale, mais pas à un niveau de domaine, y compris pour les domaines planifiés de l'enquête ordinaire. Pour obtenir des prédictions plus précises des discontinuités de domaine, nous proposons des méthodes d'estimation sur petits domaines fondées sur des modèles bayésiens de Fay-Herriot (FH).

Dans un article précédent (van den Brakel et coll., 2016), un modèle de FH univarié est proposé, où des estimations de domaine directes fiables de l'enquête ordinaire sont considérées comme des variables auxiliaires potentielles dans la procédure de sélection de modèle ascendante afin de construire des modèles adéquats pour la prédiction sur petits domaines du petit échantillon attribué à l'enquête de rechange. Dans le présent article, nous proposons un modèle de FH bivarié pour les estimations directes obtenues dans l'enquête ordinaire et l'enquête de rechange comme autre solution d'obtention de prédictions adéquates pour les discontinuités de domaine. De plus, un modèle de FH univarié appliqué aux estimations directes des discontinuités est considéré comme une solution de rechange simple. Les méthodes sont appliquées à une exécution parallèle à petite échelle visant à quantifier les discontinuités d'un remaniement du processus d'enquête de l'Enquête néerlandaise sur la victimisation criminelle (EVC).

L'utilisation d'estimations directes de l'enquête ordinaire comme variables auxiliaires dans les modèles pour petits domaines selon l'approche de rechange améliore considérablement la précision, comparativement aux modèles univariés qui utilisent seulement des variables auxiliaires des registres disponibles. Cela était prévisible puisque les deux enquêtes cherchent à mesurer les mêmes variables au moyen d'une approche d'enquête différente. L'approche univariée présente toutefois un inconvénient : la procédure d'estimation de la variance pour les discontinuités est complexe, car une covariance non négligeable apparaît entre les estimations directes du plan de l'enquête ordinaire et les prédictions fondées sur le modèle pour le plan de sondage de rechange. La méthode est complexe, car une EQM fondée sur un modèle est combinée à une variance fondée sur le plan d'un estimateur direct. Elle peut même donner des estimations de la variance négatives pour les discontinuités. Ces complications sont partiellement

contournées par l'élaboration d'un estimateur fondé sur le plan pour l'EQM des prédictions sur petits domaines et de la composante de covariance (van den Brakel et coll., 2016).

Selon un modèle de FH bivarié dans un cadre totalement bayésien, on évite les estimations de variance négatives puisque les variances des discontinuités sont calculées à partir des matrices de covariance définie positive du modèle bivarié. Le modèle de FH bivarié améliore les prédictions des discontinuités de domaine puisque le modèle accroît la précision des estimations dans les deux approches, ordinaire et de rechange, et que la forte corrélation positive entre les effets de domaine aléatoires réduit davantage encore la variance des contrastes. Pour quatre des cinq variables de l'EVC des Pays-Bas, le modèle de FH bivarié produit en effet des prédictions plus précises pour les discontinuités de domaine comparativement au modèle de FH univarié. Un autre avantage du modèle bivarié est qu'il améliore les prédictions de domaine du modèle ordinaire et du modèle de rechange, tandis que le modèle univarié suppose que la taille de l'échantillon de l'enquête ordinaire est suffisamment grande pour produire des estimations de domaine directes fiables et précises. Par conséquent, le modèle bivarié convient également dans les exécutions parallèles où, par exemple, on réduit la taille de l'échantillon de l'enquête ordinaire afin d'augmenter la taille de l'échantillon pour l'enquête de rechange. Enfin, le modèle de FH bivarié évite les complications qu'implique le fait de tenir compte de l'erreur d'échantillonnage dans les covariables, ce qui est souvent nécessaire si les estimations directes de l'enquête ordinaire sont utilisées comme covariables dans un modèle de FH univarié.

Pour une variable (satisfaction à l'égard du rendement de la police), aucun modèle adéquat n'a pu être construit seulement au moyen des variables auxiliaires disponibles des registres. Pour cette variable, le modèle multivarié semble entraîner un rétrécissement excessif des prédictions pour les discontinuités de domaine. Les résultats du modèle univarié sont nettement meilleurs dans ce cas, car les estimations directes de l'enquête ordinaire sont les seules variables auxiliaires qui donnent un modèle adéquat pour les prédictions sur petits domaines.

Le modèle de FH univarié pour les estimations directes des discontinuités de domaine apparaît comme une solution de rechange raisonnable. Il évite les complications du modèle de FH univarié pour l'EVC de rechange et la méthode est largement plus simple que le modèle de FH bivarié. Les facteurs de rétrécissement extrêmement faibles représentent une source de préoccupation, car ils indiquent que le modèle attribue trop de poids à la partie synthétique des prédictions de domaine. Le biais de ces prédictions de domaine est en effet plus important que celui du modèle de FH bivarié.

Le problème général dans cette application avec une procédure de sélection ascendante du modèle, où les covariables sont incluses dans le modèle tant que la valeur du WAIC est réduite, est qu'elle donne des modèles ayant des ensembles de covariables relativement grands. Le nombre limité de domaines dans cette application entraîne un risque réel de surajustement des données. Pour certaines variables, les covariables semblent être de solides prédicteurs des variables de domaine, ce qui donne de petits effets

aléatoires. L'ajustement d'un modèle sans ces covariables donne des modèles ayant des effets aléatoires importants et de fortes corrélations positives entre les estimations de l'enquête ordinaire et de l'enquête de rechange. Pour les autres variables, un modèle à la structure de covariance complète produit automatiquement des modèles parcimonieux pour la partie à effet fixe, probablement parce que l'ensemble de covariables disponibles représente des prédicteurs moins solides pour ces variables cibles.

Le problème indiqué plus haut de sélection de modèles comportant un trop grand nombre de covariables est contourné au moyen d'une autre méthode de sélection ascendante. Étant donné que les valeurs du WAIC sont estimées à partir de la sortie de l'échantillonneur de Gibbs, ces valeurs sont observées avec un certain degré d'incertitude. Cela pèse en faveur de la non-inclusion des covariables si elles n'entraînent qu'une petite réduction du WAIC. Dans une autre méthode de sélection ascendante, les covariables sont sélectionnées seulement si la diminution de la valeur du WAIC est supérieure à l'erreur-type estimée du WAIC. Au moyen de cette approche, on choisit des modèles parcimonieux puisqu'ils évitent la sélection d'une ou de plusieurs covariables qui n'améliorent le WAIC que légèrement. Pour ce qui est des variables pour lesquelles de grands ensembles de covariables ont été sélectionnés au départ, cette approche donne un compromis raisonnable entre l'ajustement du modèle et la complexité du modèle. Une autre solution consiste à envisager des modèles avec des coefficients de régression égaux. Ces modèles conviennent toutefois moins à la prédiction des discontinuités de domaine si les effets aléatoires sont faibles. Dans de telles situations, l'indicateur fictif est la seule composante du modèle qui distingue l'approche régulière de l'approche de rechange. Cela donne des prédictions synthétiques pour les discontinuités de domaine qui sont presque égales sur les domaines, et approximativement égales à l'estimateur direct pour la discontinuité à l'échelle nationale. Selon le type de changements apportés au processus d'enquête, on pourrait supposer à bon escient que les discontinuités de domaine sont égales. Dans ce cas, la meilleure estimation est obtenue au moyen de l'estimateur direct à l'échelle nationale.

Pour mieux comprendre les propriétés et le comportement des trois modèles différents d'estimation des discontinuités de domaine, y compris l'approche de sélection de modèle proposée, il faut une simulation exhaustive. Cela permet de mieux comprendre quelle approche de modélisation parmi les trois il faut préférer selon les conditions. Cette étude devra faire l'objet de recherches futures.

Remerciements

Les opinions exprimées dans l'article sont celles des auteurs et ne correspondent pas nécessairement à la politique de l'institut de la statistique des Pays-Bas, *Statistics Netherlands*. Nous tenons à remercier le rédacteur associé et les deux examinateurs pour leur lecture attentive de notre manuscrit et leurs précieux commentaires, qui se sont révélés très utiles pour améliorer la qualité de notre article.

Annexe

Tableau A.1
Aperçu des données auxiliaires

Variable	Description	Source
MBA_benefit	Pourcentage de prestataires de prestations sociales	MBA
MBA_immigr	Pourcentage d'immigrants dans la population	MBA
MBA_immigrnw	Pourcentage d'immigrants non occidentaux dans la population	MBA
MBA_old	Pourcentage de personnes âgées (65 ans et plus)	MBA
MBA_benefit	Pourcentage de prestataires de prestations sociales	MBA
PR_assault	Agressions physiques déclarées	RIDP
PR_propcrim	Crimes contre des biens	RIDP
PR_threat	Menaces déclarées	RIDP
PR_weapon	Infractions relatives aux armes	RIDP
PR_drugs	Infractions relatives aux drogues	RIDP
CVSR_nuisance	Nuisances perçues dans le quartier	enquête ordinaire
CVSR_victim	Pourcentage de personnes ayant déclaré avoir été victimes d'un crime	enquête ordinaire
CVSR_funcpol	Opinion sur le fonctionnement des services de police sur une échelle de 10 points	enquête ordinaire

Bibliographie

Arora, V., et Lahiri, P. (1997). On the superiority of the bayesian method over the blup in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Bell, W.R. (1999). Accounting for uncertainty about variances in small area estimation. *Bulletin of the International Statistical Institute*, 52nd Session.

Benavent, R., et Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis*, 94, 372-390.

Bollineni-Balabay, O., van den Brakel, J.A., Palm, F. et Boonstra, H.J. (2017). Multilevel hierarchical Bayesian vs. state space approach in time series small area estimation: The Dutch Travel Survey. *Journal of the Royal Statisitcal Society*, à venir.

Boonstra, H.-J. (2020). mcmcsae: *Markov Chain Monte Carlo Small Area Estimation*. R package version 0.5.0.

Boonstra, H.J., et van den Brakel, J.A. (2019). Estimation du niveau et de la variation du chômage au moyen de modèles de séries chronologiques structurels. *Techniques d'enquête*, 45, 3, 421-454. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019003/article/00005-fra.pdf.

Datta, G., Ghosh, M., Nangia, N. et Natarjan, K. (1996). Estimation of median income in four-peron families: A Bayesian approach. Dans *Bayesian analysis in Statistics and Econometrics*, (Éds., D.A. Berry, K.M. Chaloner et J.M. Geweke), New York: John Wiley & Sons, Inc., 129-140.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94(448), 1074-1082.

Estaban, M.D., Morales, D., Perez, A. et Santamaria, L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, 56, 2840-2855.

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(366), 269-277.

Gelfand, A.E., et Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533.

Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472.

Gelman, A., Van Dyk, D.A., Huang, Z. et Boscardin, W.J. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17(1), 95-122.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. et Rubin, D.B. (2004). *Bayesian Data Analysis*. New York: Chapman and Hall.

Geman, S., et Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, 6, 721-741.

Gonzales-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. et Santamaria, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics and Data Analysis*, 52, 5242-5252.

Gonzalez, M.E., et Waksberg, J. (1973). Estimation of the error of synthetic estimates. Rapport technique, document présenté à la première réunion de l'International Association of Survey Statisticians, Vienne, 18-25 août 1973.

Hawala, S., et Lahiri, P. (2018). Variance modelling for domains. *Statisitcs and Applications*, 16, 399-409.

Hirose, M., et Lahiri, P. (2018). Estimating variance of random effects to solve multiple problems simultaneously. *Annals of Statistics*, 46, 1721-1741.

Hodges, J.S., et Sargent, D.J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 88(2), 367-379.

Lahiri, P., et Pramanik, S. (2019). Evaluation of synthetic small-area estimators using design-based methods. *Austrian Journal of Statistics*, 48, 43-57.

Li, H., et Lahiri, P. (2010). Adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.

Marhuenda, Y., Molina, I. et Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 57, 308-325.

Marker, D. (1995). *Small Area Estimation: A Bayesian Perspective*. Thèse de doctorat, University of Michigan.

O'Malley, A.J., et Zaslavsky, A.M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103(484), 1405-1418.

Pfeffermann, D., et Ben-Hur, D. (2018). Estimation of randomisation mean square error in small area estimation. *Revue Internationale de Statistique*, 87, 31-49.

Pfeffermann, D., et Burck, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 2, 229-249. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1990002/article/14534-fra.pdf.

Pfeffermann, D., et Tiller, R. (2006). Small area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

Polson, N.G., et Scott, J.G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4), 887-902.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*. Wiley-Interscience.

Rao, J.N.K., et Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

Rao, J.N.K., Rubin-Bleuer, S. et Estevao, V.M. (2018). Measuring uncertainty associated with model-based small area estimators. Document de recherche non publié.

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.

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

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. et van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64(4), 583-639.

Sugasawa, S., Tamae, H. et Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44(1), 150-167.

van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statisitcal Society*, 171, 581-613.

van den Brakel, J.A., et Boonstra, H.J. (2018). Hierarchical Bayesian bivariate Fay-Herriot model for estimating domain discontinuities. Document de travail, Statistics Netherlands.

van den Brakel, J.A., et Krieg, S. (2016). Small area estimation with state-space common factor models for rotating panels. *Journal of the Royal Statisitcal Society*, 179, 763-791.

van den Brakel, J.A., Buelens, B. et Boonstra, H.J. (2016). Small area estimation to quantify discontinuities in repeated sample surveys. *Journal of the Royal Statisitcal Society*, 179, 229-250.

van den Brakel, J.A., Smith, P.A. et Compton, S. (2008). Quality procedures for survey transitions-experiments, time series and discontinuities. *Survey Research Methods*, 2, 123-141.

Vehtari, A., Gelman, A. et Gabry, J. (2015). `loo`: *Efficient Leave-one-out Cross-validation and WAIC for Bayesian Models*. R package version 0.1.3.

Vehtari, A., Gelman, A. et Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistical Computation*, 27, 1413-1432.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571-3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867-897.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer.

Ybarra, L.M.R., et Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4), 919-931.

You, Y. (2008). Small area estimation using area level models with model checking and applications. *Proceedings of the Survey Methods Section of the Statistical Society of 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., 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, juin 2021
Vol. 47, nº 1, pp. 205-230
Statistique Canada, nº 12-001-X au catalogue

Le regroupement bayésien aux fins d'analyse des données catégoriques sur petits domaines

Aejeong Jo, Balgobin Nandram et Dal Ho Kim[1]

Résumé

Les stratégies bayésiennes de regroupement servent à résoudre les problèmes de précision liés aux analyses statistiques des données sur petits domaines. Dans ces cas, les échantillons des sous-populations sont généralement petits, même si la population ne l'est pas nécessairement. Une autre solution consiste à regrouper des données semblables en vue de réduire le nombre de paramètres dans le modèle. De nombreuses enquêtes recueillent des données catégoriques par domaine, lesquelles sont ensuite réunies dans un tableau de contingence. Nous examinons les modèles de regroupement bayésiens hiérarchisés avec une loi a priori de processus de Dirichlet pour analyser les données catégoriques sur de petits domaines. Toutefois, la loi a priori utilisée aux fins de regroupement de ces données entraîne souvent un problème de rétrécissement excessif. Pour corriger le problème, nous séparons les paramètres en effets globaux et locaux. Cette étude porte sur le regroupement de données au moyen d'un processus de Dirichlet. Nous comparons les modèles de regroupement utilisant des données sur la densité minérale osseuse (DMO) tirées de la *Third National Health and Nutrition Examination Survey*, portant sur la période de 1988 à 1994 aux États-Unis. Nos analyses des données sur la DMO sont effectuées au moyen d'un échantillonneur de Gibbs et d'un échantillonnage par tranche pour effectuer les calculs a posteriori.

Mots-clés : Données catégoriques; processus de Dirichlet; regroupement bayésien hiérarchique non paramétrique; échantillonnage par tranche; petits domaines.

1. Introduction

De nombreuses enquêtes recueillent des données catégoriques par domaine, lesquelles sont ensuite conservées dans un tableau de contingence. Par exemple, dans une étude comparative typique sur le taux d'obésité, le chercheur pourrait classer les données de l'échantillon mesurées par degré d'obésité. Ensuite, le taux d'obésité régional est estimé au moyen du nombre d'échantillons attribués à chaque catégorie. Dans de tels cas, nous devons tenir compte de la façon dont la taille de l'échantillon et le nombre de paramètres dans le modèle influent sur la précision, en particulier quand il s'agit d'estimations pour petits domaines (Rao et Molina, 2015). En général, la précision d'un modèle diminue à mesure que le nombre de paramètres augmente, si l'on suppose que les données de l'échantillon sont identiques. Pour empêcher cette réduction du niveau de précision, le modèle construit doit être aussi simple que possible. Autrement dit, il faut réduire le nombre de paramètres du modèle. Ce dernier perd alors de sa capacité à refléter en détail les effets dans chaque domaine. Une autre façon de résoudre le problème de précision consiste à augmenter la taille d'échantillon autorisée par paramètre. Cela signifie que nous pouvons utiliser des stratégies de regroupement dans l'analyse de données catégoriques sur de petits domaines.

Les méthodes de regroupement suscitent un intérêt croissant chez les chercheurs. Malec et Sedransk (1992) ont élaboré une procédure bayésienne pour estimer la moyenne d'une expérience dans un ensemble d'expériences apparemment semblables. Ils ont construit la distribution a priori pour un paramètre d'emplacement afin de traduire leurs hypothèses. Ils ont déterminé des sous-ensembles de paramètres,

1. Aejeong Jo, chercheur, Division for Healthcare Technology Assessment Research, National Evidence-based Healthcare Collaborating Agency, 173, Toegye-ro, Jung-gu, Séoul, 100-705, Corée. Courriel : joaejung@neca.re.kr; Balgobin Nandram, professeur, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, États-Unis. Courriel : balnan@wpi.edu; Dal Ho Kim, professeur, Department of Statistics, Kyungpook National University, 80 Daehakro, Bukgu, Daegu, 702-701, Corée. Courriel : dalkim@knu.ac.kr.

avec des indices indiquant la similarité entre les sous-ensembles, dans lesquels il existe une incertitude au sujet de la composition des sous-ensembles. Ils ont spécifié la distribution a priori d'un paramètre au moyen d'un conditionnement sur le même indice dans des expériences similaires. Grâce à leur distribution a priori souple, les données de l'échantillon agissent sur l'intensité et la nature du regroupement. Ultérieurement, Evans et Sedransk (1999) ont proposé un modèle bayésien plus souple utilisant des covariables. En outre, Evans et Sedransk (2003) ont fourni une justification entièrement bayésienne des résultats de Malec et Sedransk (1992). Depuis, ces trois travaux ont été étendus à partir du même concept principal de spécification d'un modèle dans lequel des indices servent à indiquer des expériences similaires (Consonni et Veronese, 1995 et DuMouchel et Harris, 1983). Par ailleurs, Dunson (2009) a proposé une généralisation du processus de Dirichlet (PD) avancé par Ferguson (1973), qui permet un regroupement local dépendant et l'emprunt d'information. L'objectif est d'emprunter de l'information afin d'estimer plus efficacement chaque fonction. Le processus proposé de regroupement local offre une méthode simple et souple de spécification du processus de sélection local. Ils suggèrent d'utiliser l'échantillonnage par tranche, proposé par Walker (2007), pour réaliser les calculs a posteriori. Il s'agit d'une méthode simple et efficace qui permet des calculs a posteriori pour un processus à dimensions infinies semblables à ceux d'un processus à dimensions finies. Ici, nous construisons un modèle de regroupement au moyen de ces concepts de base pour des données fondées sur de petits domaines. Récemment, Nandram, Zhou et Kim (2019) ont proposé un test de Bayes regroupé d'indépendance pour des tableaux de contingence éparses. Ils ont construit le modèle basé sur un modèle bayésien hiérarchique multinomial selon le processus de Dirichlet; voir aussi Nandram (1998) qui a construit une loi a priori utilisant une distribution de Dirichlet afin de regrouper les données dans les modèles. Bien sûr, un PD est présumé pour les paramètres d'intérêt, qui sont les paramètres de probabilité de la cellule dans notre tableau de contingence.

Dans la présente étude, nous utilisons les données sur la densité minérale osseuse (DMO), tirées de la *Third National Health and Nutrition Examination Survey* (NHANES III) concernant une période de six ans allant d'octobre 1988 à septembre 1994. La DMO est la quantité de minéraux dans les tissus osseux, mesurée comme étant la densité optique par cm^2 de surface osseuse par imagerie médicale. La DMO est utilisée dans les milieux cliniques comme indicateur indirect de l'ostéopénie, de l'ostéoporose, de fractures, etc. La DMO est statistiquement corrélée avec la probabilité de fractures, qui constituent un problème de santé publique important, surtout chez les femmes âgées. C'est pourquoi les données sur la DMO sont des indicateurs clés dans l'identification des patients ostéoporotiques dont la solidité osseuse est susceptible de s'améliorer en cas de prise en charge rapide.

La NHANES III contient des données cliniques concernant 33 994 participants à l'enquête et elle est échantillonnée par région. Chaque personne est classée selon trois niveaux de DMO : 1) normale, 2) ostéopénie et 3) ostéoporose. Notre étude a utilisé l'inférence bayésienne sur des tableaux catégoriques. Voir Agresti et Hitchcock (2005) et Leonard (1977) pour l'inférence sur des tableaux multinomiaux. Les données originales ont été obtenues de centres d'examen mobiles aux États-Unis. La NHANES III, un important programme du *National Center for Health Statistics* (NCHS), examine l'état de santé et la nutrition aux États-Unis. Lancé dans les années 1960, le programme a réalisé des enquêtes sur plusieurs sujets liés à la santé et à la nutrition. La NHANES fournit ainsi des enquêtes reposant sur de grands

échantillons de la population des États-Unis. Toutefois, le NCHS s'intéresse également aux estimations pour de petites régions géographiques et de petits domaines d'étude. Quand la taille de l'échantillon d'une sous-population est petite, nous devons envisager un autre estimateur fondé sur une stratégie de regroupement afin d'analyser les données.

Par conséquent, nous nous concentrons sur la prédiction de la proportion de la population finie de chaque région. Nous estimons la proportion de la population finie en entrant les données de l'échantillon dans le modèle pour prédire la partie hors de l'échantillon non observée de la population finie, puis en obtenant la somme pondérée des données de l'échantillon observées et la fraction non observée prédite obtenue à la proportion de l'échantillon. Premièrement, nous estimons les paramètres de probabilité de cellule à partir des données de l'échantillon. Au cours du processus, le nombre observé par catégorie dans la NHANES III est utilisé comme partie de l'échantillon de la population finie. Deuxièmement, ces paramètres servent à prédire des valeurs pour la partie hors de l'échantillon. Enfin, nous obtenons les proportions de population finie en combinant les données de l'échantillon et la valeur de prédiction de la partie hors de l'échantillon.

L'article est structuré comme suit. Dans la section 2, nous présentons les stratégies de regroupement bayésiennes hiérarchiques utilisées dans l'analyse des données catégoriques sur petits domaines. Dans la section 3, nous présentons et examinons les résultats de notre analyse des données de l'ensemble de données de l'enquête sur la DMO. La section 4 présente les conclusions de l'article. Les annexes A et B comprennent le processus de calcul du modèle de regroupement bayésien hiérarchique.

2. Modèles bayesiens hiérarchiques

2.1 Modèles paramétriques

Pour un modèle bayésien hiérarchique de référence, nous considérons un tableau de contingence $I \times K$, où n_{ik} indique la k^{e} réponse dans le i^{e} domaine, pour $i = 1, \ldots, I$, $k = 1, \ldots, K$. Soit π_{ik} la proportion correspondante pour chaque cellule. Ensuite, supposons que

$$\mathbf{n}_i \,|\, \boldsymbol{\pi}_i \overset{\text{iid}}{\sim} \text{multinominal}(n_{i.}, \boldsymbol{\pi}_i), \tag{2.1}$$

où $\mathbf{n}_i = (n_{i1}, \ldots, n_{iK})$ pour $i = 1, \ldots, I$, est un vecteur de réponses, $n_{i.} = \sum_{k=1}^{K} n_{ik}$ est la somme des réponses dans le domaine i, et $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, $0 \le \pi_{ik} \le 1$, $\sum_{k=1}^{K} \pi_{ik} = 1$, est le vecteur de proportion pour chaque domaine. Le modèle ne permet pas de regroupement et est désigné comme modèle de référence aux fins de comparaison de nos modèles. Les paramètres π_i sont indépendants et ne possèdent pas d'effet commun. Cela signifie que les domaines ne sont pas liés.

Il existe cinq catégories de modèles de regroupement paramétriques, classées selon les probabilités a priori des vecteurs de proportion dans une distribution multinomiale. Premièrement, les quatre distributions a priori pour les inférences bayésiennes paramétriques sont données comme suit :

1) Pas de regroupement, $\boldsymbol{\pi}_i \overset{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{1})$;
2) Regroupement complet, $\boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{1})$ avec $\boldsymbol{\pi}_1 = \ldots = \boldsymbol{\pi}_I = \boldsymbol{\pi}$;

3) Regroupement adaptatif, $\boldsymbol{\pi}_i \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}\tau)$;

4) Regroupement restreint, $\boldsymbol{\pi}_i \overset{\text{ind}}{\sim} \phi \ \text{Dirichlet}(\boldsymbol{\mu}\tau) + (1-\phi)\,\text{Dirichlet}(\mathbf{1})$;

où $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$, $0 \le \mu_k \le 1$, $\sum_{k=1}^{K} \mu_k = 1$ et $\tau > 0$ sont les hyperparamètres de la distribution de Dirichlet. Nous supposons en outre que $\pi(\boldsymbol{\mu}, \tau) = (K-1)!/(1+\tau)^2$, une loi a priori de rétrécissement. Notons que Yin et Nandram (2020a, b) placent un processus de Dirichlet sur le processus d'échantillonnage pour tenir compte des lacunes, des valeurs aberrantes et des égalités dans les données d'enquête; voir aussi Nandram et Yin (2016a, b) pour plus de discussions sur le processus de Dirichlet. La valeur $\phi \sim \text{uniforme}(1/2, 1)$, $\phi > 1/2$ signifie qu'un poids plus grand est donné au regroupement adaptatif.

Le modèle 1 est un modèle sans regroupement qui estime le paramètre sans partage de données d'autres domaines. En revanche, le modèle 2 est un modèle de regroupement complet qui estime le paramètre en traitant les différents domaines comme s'ils étaient un seul domaine. Quand on estime des paramètres sur un petit domaine avec un petit nombre de données au moyen du modèle 1, l'estimation peut se heurter au problème de la petite taille du domaine, car le paramètre est estimé à partir de données insuffisantes. Bien que le modèle de regroupement complet corrige ce problème, il présente d'autres inconvénients : il est impossible de discerner un rétrécissement excessif et les domaines individuellement. En partant de ce constat, l'article présente plusieurs méthodes de regroupement afin de trouver un modèle produisant de meilleures estimations. Dans le modèle 3, le modèle de regroupement adaptatif introduit par Nandram, Zhou et Kim (2019), tous les domaines ont en commun les mêmes hyperparamètres; par conséquent, ils partagent aussi leurs données de domaine. Il s'agit d'une méthode indirecte de regroupement complet qui préserve une certaine variation dans les domaines, mais qui suppose généralement que tous ont des caractéristiques identiques; voir aussi Nandram (1998). Cela produit des variations dans l'estimation des hyperparamètres. Nous proposons alors un modèle de regroupement restreint, qui modifie le modèle 3 en supprimant l'information sur le partage de données entre domaines distincts. Dans ce nouveau modèle, les domaines distincts utilisent leurs données locales pour estimer les paramètres, et les domaines ayant des caractéristiques semblables partagent de l'information par un regroupement fondé sur le même hyperparamètre, ce qui améliore l'estimation. Toutefois, ce modèle attribue le même hyperparamètre aux domaines présentant des caractéristiques similaires, ce qui peut entraîner un problème de rétrécissement excessif quand le lissage se produit dans la catégorie. Pour pallier ce problème, nous proposons le modèle 5, un modèle de regroupement global-local, voir Dunson (2009). Le modèle de regroupement global-local regroupe l'information de données provenant de domaines ayant des caractéristiques semblables, mais il préserve également chaque variation dans la catégorie au moyen du modèle d'effet local, ce qui réduit le lissage dans l'effet catégorique. De ce fait, le modèle 5 est à la fois souple et robuste.

La cinquième distribution a priori utilisée pour les inférences bayésiennes paramétriques est appelée regroupement global-local. Dans ce cas, nous utilisons une notation différente pour le vecteur de proportion de chaque domaine. Soit $\mathbf{p}_i$, pour $i = 1, \ldots, I$, qui désigne le vecteur de proportion de cellule correspondant dans le i^{e} domaine. Nous supposons que

$$\mathbf{n}_i \,|\, \mathbf{p}_i \overset{\text{ind}}{\sim} \text{multinomial}(n_{i.}, \mathbf{p}_i),\ i = 1, \ldots, I, \tag{2.2}$$

où $\mathbf{p}_i = \left(e^{\theta+\eta_{i1}} \Big/ \left(1 + \sum_{k=1}^{K-1} e^{\theta+\eta_{ik}}\right), \ldots, e^{\theta+\eta_{i(K-1)}} \Big/ \left(1 + \sum_{k=1}^{K-1} e^{\theta+\eta_{ik}}\right), 1 \Big/ \left(1 + \sum_{k=1}^{K-1} e^{\theta+\eta_{ik}}\right)\right)'$. Ici, $\mathbf{p}_i$ est composé de deux composantes, à savoir θ et η_{ik}, et θ reflète la probabilité de base qui a rassemblé tous les domaines. L'effet global-local est représenté dans la composante $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{i(K-1)})'$. Plus particulièrement,

$$\boldsymbol{\eta}_i \overset{\text{iid}}{\sim} I_{(z_i=0)} \prod_{k=1}^{K-1} \mathrm{N}(0, \sigma^2) + I_{(z_i=1)} \prod_{k=1}^{K-1} \mathrm{N}(0, \sigma_k^2), \tag{2.3}$$

où $\mathrm{N}(0, \sigma^2)$ est la distribution normale du paramètre global σ^2, $\mathrm{N}(0, \sigma_k^2)$ est la distribution normale des paramètres locaux σ_k^2 pour chaque catégorie, où chaque domaine est désigné par un indice différent. Ensuite, $z_i, i = 1, \ldots, I,$ suit une distribution de Bernoulli avec un hyperparamètre ϕ, qui corrige la proportion entre les effets globaux et locaux. Ainsi, si nous devons nous concentrer sur l'effet global, la distribution a priori pour ϕ est établie au moyen de la distribution uniforme sur $(1/2, 1)$. Plus précisément, nous supposons que

$$z_i \,|\, \phi \overset{\text{iid}}{\sim} \text{Bernoulli}(\phi),\ i = 1, \ldots, I,$$

$$\phi \sim \text{uniforme}\left(\frac{1}{2}, 1\right),$$

$$\pi(\theta) = \frac{1}{\pi(1 + \theta^2)}, \text{ une loi a priori de Cauchy,}$$

$$\pi(\sigma^2, \sigma_1^2, \ldots, \sigma_{K-1}^2) = \frac{1}{(1 + \sigma^2)^2} \prod_{k=1}^{K-1} \frac{1}{(1 + \sigma_k^2)^2},$$

où $-\infty < \eta_{ik} < \infty, k = 1, \ldots, K-1, \sigma^2 > 0,$ et $\sigma_k^2 > 0, k = 1, \ldots, K-1.$ Nous avons utilisé la loi a priori de Cauchy pour θ et les lois a priori de rétrécissement pour les composantes de la variance. Les lois a priori de rétrécissement ressemblent à la loi a priori demi-Cauchy et elles sont mathématiquement plus pratiques quand on effectue des transformations sur $(0, 1)$.

Les modèles proposés dans l'article sont fondés sur le modèle de regroupement adaptatif (Nandram, Zhou et Kim, 2019), qui applique le principe d'attribution d'indices de paramètres identiques dans la distribution a priori à des expériences similaires (Malec et Sedransk, 1992) aux données catégoriques. Ainsi, le modèle sans regroupement et le modèle de regroupement complet représentent deux cas extrêmes de regroupement adaptatif, avec des paramètres $\boldsymbol{\mu}$ et τ. Par ailleurs, le modèle de regroupement restreint comporte un principe de regroupement comme celui du modèle de regroupement adaptatif, mais le modèle représente également l'incertitude par le paramètre de pondération ϕ. Toutefois, le même paramètre dans la loi a priori utilisée aux fins de regroupement de ces données entraîne souvent un problème de rétrécissement excessif. Pour corriger ce problème, nous proposons le modèle de regroupement

global-local. Dans ce modèle, les effets des paramètres sont séparés en effets globaux et effets locaux. Nous proposons par conséquent deux nouveaux modèles, le modèle de regroupement restreint et le modèle de regroupement global-local, et nous les comparons aux modèles existants.

2.2 Modèles non paramétriques

Nous examinons la loi a priori de Dirichlet pour $\boldsymbol{\pi}_i$ de (2.1) à la section 2.1. La loi a priori est structurée comme suit :

$$\pi_i \,|\, G \overset{\text{iid}}{\sim} G$$

$$G \sim \text{PD}(\alpha, G_0),$$

où G_0 est la distribution de base et le nombre réel positif, α, est le paramètre de concentration dans la loi a priori de Dirichlet. Le modèle est spécifié par la structure de la distribution de base. Nous remarquons ici que Yin et Nandram (2020a, b) ont utilisé un processus de Dirichlet dans le processus d'échantillonnage pour tenir compte des lacunes, des valeurs aberrantes et des égalités dans les données d'enquête. Premièrement, nous définissons le modèle au moyen de deux distributions a priori précédentes, comme suit :

6) Regroupement adaptatif non paramétrique $G_0 \equiv \text{Dirichlet}(\boldsymbol{\mu}\tau)$;

7) Regroupement restrictif non paramétrique $G_0 \equiv \phi\ \text{Dirichlet}(\boldsymbol{\mu}\tau) + (1-\phi)\,\text{Dirichlet}(\mathbf{1}_{1\times K})$.

Supposons que $\pi(\boldsymbol{\mu}, \tau) = (K-1)!/(1+\tau)^2$, $\pi(\alpha) \propto 1/(1+\alpha)^2$, et $\phi \sim \text{uniforme}(1/2, 1)$. Dans les modèles 6 et 7, nous utilisons un processus de rupture de bâton pour la loi a priori de processus de Dirichlet (Sethuraman, 1994).

Le dernier modèle est une version non paramétrique de (2.2) dans la section 2.1, utilisée aux fins de regroupement global-local. Ici, les variables $\boldsymbol{\eta}_i$ servent à construire le paramètre bayésien non paramétrique, comme suit :

$$\boldsymbol{\eta}_i \overset{\text{iid}}{\sim} I_{(z_i=0)} G_0 + I_{(z_i=1)} \prod_{k=1}^{K-1} G_k$$

$$G_0 \sim \text{PD}(\alpha_0, \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I})),\ G_k \sim \text{PD}(\alpha_1, N(0, \sigma_k^2)),$$

où $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{i(K-1)})'$, $z_i \sim \text{Bernouilli}(\phi)$, $\pi(\theta) \propto 1$, $\phi \sim \text{uniforme}(1/2, 1)$, $\pi(\sigma^2, \sigma_1^2, \ldots, \sigma_{K-1}^2) = 1\Big/\left\{(1+\sigma^2)^2 \prod_{k=1}^{K-1}(1+\sigma_k^2)^2\right\}$, pour $i = 1, \ldots, I$. Il s'agit du modèle (A.1), de regroupement global-local non paramétrique.

La distribution de η_i implique un mélange de domaines de regroupement global et local. Alors que le regroupement global est mené selon le même principe que les modèles non paramétriques mentionnés plus haut, la loi a priori de Dirichlet, où la distribution normale est la distribution de base pour chaque cellule, est définie indépendamment, ce qui permet de résoudre le problème de rétrécissement excessif susceptible

de découler du regroupement global. Ici, z_i, qui représente le poids du modèle, suit la distribution de Bernouilli avec ϕ plus grand que $1/2$ comme paramètre, et ϕ suit la distribution uniforme. Par conséquent, le domaine de regroupement local est pondéré de façon à être défini comme la forme la plus à même d'atténuer le degré de rétrécissement. De plus, pour assurer la simplicité du modèle, la queue lourde et la caractéristique non informative de la distribution θ suivent la loi a priori impropre qui est identique au modèle paramétrique précédent, mais plus simple, et la distribution a posteriori est présentée de façon propre.

3. Analyse des données

Dans cette section, nous présentons les résultats empiriques de la comparaison des performances des cinq modèles de regroupement paramétrique décrits à la section 2 et des trois versions de regroupement non paramétrique décrites à la section 3. Nous utilisons les données sur la DMO pour la période de 1988 à 1994; celles-ci proviennent de la NHANES III et ont été recueillies auprès des centres d'examen mobiles aux États-Unis.

Notre analyse est effectuée au moyen de tableaux de contingence avec un nombre de cellules pour trois catégories de DMO dans 31 comtés aux États-Unis. Ici, la DMO est classée à l'un des trois niveaux. La catégorie normale correspond aux valeurs de DMO inférieures à un écart-type (ET) sous la moyenne adulte des personnes blanches non hispaniques (NHW pour *non-Hispanic white*) $(0{,}82\,\mathrm{mg/cm^2} < \mathrm{DMO})$. La catégorie de l'ostéopénie correspond aux valeurs de DMO entre 1 et 2,5 ET inférieures à la moyenne des jeunes adultes blancs non hispaniques $(0{,}64\,\mathrm{mg/cm^2} < \mathrm{DMO} \leq 0{,}82\,\mathrm{mg/cm^2})$. Enfin, la catégorie de l'ostéoporose correspond aux valeurs de DMO supérieures à 2,5 ET sous la moyenne des jeunes adultes blancs non hispaniques $(\mathrm{DMO} \leq 0{,}64\,\mathrm{mg/cm^2})$.

Nous prédisons la proportion de population finie pour la distribution de la DMO dans chaque domaine au moyen du modèle bayésien de regroupement. L'enquête porte sur environ 0,02 % de la population, et la prédiction nécessaire pour les 99,98 % restants représente un travail considérable. Le tableau 3.1 présente les données de l'échantillon, qui comprennent un nombre de cellules pour chaque niveau catégorisé dans chaque région. Nous estimons la proportion de la population finie en prédisant la partie hors de l'échantillon de la population finie à partir d'une distribution multinomiale avec un paramètre $\boldsymbol{\pi}_i$, $i = 1, \ldots, I\,(= 31)$ à chaque itération MCMC. Plus précisément, supposons que N_{ik} pour $k = 1, 2, 3$ soit le niveau total de DMO dans le domaine i, où la valeur est inconnue. Nous avons la valeur (n_{ik}) pour la partie échantillonnée de la population finie. Ensuite, nous calculons la proportion de population finie (P_{ik}) pour $i = 1, \ldots, I\,(I = 31)$ comme suit :

$$P_{ik} = \frac{1}{N_i}\{n_{ik} + (N_{ik} - n_{ik})\},\ k = 1, 2, 3, \tag{3.1}$$

où $N_i = \sum_{k=1}^{3} N_{ik}$, $N_{ik} - n_{ik}$ est la partie hors de l'échantillon pour chaque niveau de DMO $k\,(k = 1, 2, 3)$ dans le domaine i, pris de la distribution multinomiale avec un paramètre $\hat{\boldsymbol{\pi}}_i$, estimé à l'aide d'une méthode MCMC dans chaque modèle. Ensuite, la moyenne a posteriori et l'écart-type de P_{ik} sont obtenus au moyen de la distribution empirique estimée de P_{ik}.

Tableau 3.1
Données catégoriques de DMO pour 31 comtés tirées de la NHANES III

Domaines	DMO		
	Normale	Ostéopénie	Ostéoporose
1	33	24	9
2	46	39	5
3	40	25	8
4	48	25	6
5	40	15	10
6	74	30	12
7	47	19	7
8	38	15	6
9	49	16	11
10	99	40	14
11	39	18	2
12	63	27	4
13	48	18	5
14	42	16	4
15	40	15	4
16	110	44	7
17	37	14	3
18	55	18	5
19	47	12	6
20	296	95	17
21	59	18	4
22	78	21	7
23	196	55	15
24	149	44	9
25	69	19	5
26	49	10	6
27	73	19	3
28	76	14	3
29	77	13	4
30	96	13	6
31	88	12	4

Nous utilisons 1 000 itérations pour mettre « en rodage » les échantillons MCMC, et prenons une valeur toutes les 10 valeurs pour obtenir les 1 000 itérations. De plus, nous utilisons des graphiques

d'autocorrélation du modèle pour corriger le nombre de répétitions et d'intervalles d'élimination sélective. Par exemple, dans un modèle non paramétrique avec un nombre relativement élevé de paramètres, nous prenons une valeur estimée sur 20 de 1 001 à 20 000. Nous établissons la valeur initiale de la proportion $\boldsymbol{\pi}_i$ pour $i = 1, \ldots, 31$ en fonction de la proportion de la colonne des valeurs de l'échantillon dans chaque domaine.

Les groupes sont classés selon les valeurs de quartile de la proportion de la première colonne. Le paramètre de réglage ξ_j, $j = 1, \ldots, J$ est initialement réglé à $\xi = 0{,}5$, puis il est révisé en fonction des performances de chaque modèle.

Dans le tableau 3.2, nous présentons les moyennes a posteriori (MP) et les écarts-types a posteriori (ETP) des proportions de population finie pour les huit modèles et certains domaines. Les cas des modèles 1 et 2 sont les structures de regroupement les plus extrêmes. La MP du modèle 1 donne les résultats 0,511; 0,496; 0,901; 0,826 et 0,820 pour les domaines correspondant respectivement à 1; 2; 28; 29 et 31, dans la catégorie de DMO normale, ce qui signifie que les fluctuations des domaines sont supérieures à celles de la MP du modèle 2 (0,652; 0,654; 0,714; 0,716; 0,719). Pour le modèle 3, les fluctuations (0,644; 0,612; 0,793; 0,816; 0,798) montrent une tendance semblable à celle des MP du modèle 1 pour chaque domaine, mais elles sont plus lisses que celles du modèle 1. Cela pourrait être interprété comme l'effet de regroupement indirect par l'hyperparamètre plutôt que l'effet de regroupement direct du modèle 2. Pour le regroupement restreint, dans les domaines qui présentent des caractéristiques similaires, les valeurs estimées sont calculées par regroupement indirect et les hyperparamètres comme dans le modèle 3, et les valeurs estimées sont plus lisses que celles du modèle 1 (MP des domaines 28; 29; 31 : 0,827; 0,779 et 0,864, respectivement). Cependant, dans les domaines où des caractéristiques similaires ne sont pas présentes, les valeurs estimées sont proches de celles du modèle 1, car le paramètre est estimé uniquement à partir de l'information des domaines visés (MP des domaines 1; 2 : 0,523 et 0,510, respectivement.) Le modèle 5, que nous proposonspour pallier le problème de rétrécissement excessif susceptible de découler du regroupement de l'information, partage l'information de domaines à proximité et corrige le rétrécissement excessif en reflétant l'effet local de chaque cellule, ce qui rend les valeurs estimées situées entre celles du modèle 1 et du modèle 2 (MP du modèle 5 : 0,581; 0,511; 0,711; 0,836; 0,808). Notons que le modèle bayésien non paramétrique attribue les domaines avec des indices comme étant le même groupe en fonction des caractéristiques de l'information par l'hyperparamètre d_i, $i = 1, \ldots, I\,(= 31)$, et que le même groupe partage directement le paramètre. Pour les données de DMO, le nombre du groupe varie de un à trois, montrant la fréquence la plus élevée, et la MP est estimée comme dans le modèle 2. Les caractéristiques des valeurs estimées de MP pour chaque modèle sont également indiquées de façon identique dans la DMO de l'ostéopénie et la DMO de l'ostéoporose.

La moyenne a posteriori, les écarts-types (ET) et les coefficients de variation (c.v.) a posteriori des proportions de population finie pour les huit modèles sont illustrés dans les figures 3.1 à 3.3.

Dans la figure 3.1, la variation de la moyenne a posteriori dans les huit modèles est plus grande dans la catégorie de la DMO normale, qui représente la plus grande proportion. En particulier, nous constatons que les MP du modèle non paramétrique ressemblent à celles du modèle de regroupement complet. Pendant l'analyse des données, nous avons trouvé un à trois groupes d'indices. À partir de là, nous avons pu découvrir que la distribution de la DMO est assez semblable dans tous les domaines de la NHANES III. Par conséquent, tous les domaines partagent leur information avec d'autres pour estimer le même hyperparamètre. Les modèles de regroupement non paramétrique et complet donnent alors des estimations similaires. Par conséquent, le regroupement peut résoudre les problèmes associés aux estimations pour petits domaines quand les domaines ont des caractéristiques en commun. De plus, la figure 3.2 montre que les modèles non paramétriques ont de bonnes performances, car l'écart-type des modèles non paramétriques comportant de nombreux paramètres est similaire ou plus petit que celui des modèles paramétriques.

Le coefficient de variation des modèles par statut de DMO se trouve à la figure 3.3. Dans le cas du c.v., la DMO de l'ostéoporose est celle pour laquelle on constate la plus grande différence entre modèles. De plus, le c.v. des versions non paramétriques est relativement faible comparativement à celui de la version paramétrique, qui n'est pas différent pour chaque statut de DMO. Ajoutons qu'il est très significatif que la version non paramétrique ait un c.v. plus petit que les modèles de la version paramétrique, bien qu'elle ait un espace de paramètre infini.

Pour estimer les paramètres, nous utilisons un échantillonneur de Gibbs. Alors que les paramètres aux espaces de paramètres restreints sont échantillonnés au moyen de la méthode de grille, les autres paramètres sont échantillonnés à l'aide de l'algorithme de Metropolis-Hastings. Nous cherchons à obtenir un taux d'acceptation de 30 à 70 %. Dans l'analyse réelle, le taux d'acceptation de l'algorithme est de 34 à 49 %. Nous comparons les deux mesures pour ce qui est des performances de chaque modèle. Premièrement, nous calculons le critère d'information de déviance (DIC), un critère typique de choix de modèle bayésien, pour comparer les modèles bayésiens hiérarchiques. Dans le critère d'information de déviance, qui a été proposé par Spiegelhalter et coll. (2002), une valeur de DIC plus faible indique de meilleures performances. Deuxièmement, nous évaluons les performances des huit modèles en calculant l'ordonnée prédictive conditionnelle logarithmique (LCPO pour *logarithmic conditional predictive ordinate*), qui est une méthode de comparaison utilisant la validation croisée. La LCPO moyenne, proposée par Gneiting et Raftery (2007), est calculée comme suit :

$$\overline{\text{LCPO}} = -\frac{1}{I}\sum_{i=1}^{I}\log\left(\text{C}\hat{\text{P}}\text{O}_i\right), \tag{3.2}$$

où $\text{C}\hat{\text{P}}\text{O}_i = \sum_{h=1}^{H} w_h P\left(Y = y \mid \boldsymbol{\Omega}^{(h)}\right)$, $w_h = \sum_{h=1}^{H} f\left(Y = y \mid \boldsymbol{\Omega}^{(h)}\right) / f\left(Y = y \mid \boldsymbol{\Omega}^{(h)}\right)$, pour $i = 1, \ldots, I$, et $P\left(Y = y \mid \boldsymbol{\Omega}^{(h)}\right)$ est la probabilité d'une observation unique d'un paramètre donné $\boldsymbol{\Omega}^{(h)}$, et $h = 1, \ldots, H$ désigne les itérations du résultat de la méthode MCMC sous le modèle de regroupement bayésien hiérarchique.

Tableau 3.2
Valeurs sommaires a posteriori des proportions de population finies des données sur la DMO selon les huit modèles par région

Domaines	Modèle	DMO normale			DMO d'ostéopénie			DMO d'ostéoporose		
		MP	ETP	IC à 95 %	MP	ETP	IC à 95 %	MP	ETP	IC à 95 %
1	1	0,511	0,024	(0,467; 0,561)	0,380	0,024	(0,333; 0,427)	0,109	0,015	(0,082; 0,139)
	2	0,652	0,023	(0,603; 0,692)	0,264	0,021	(0,227; 0,306)	0,084	0,013	(0,061; 0,109)
	3	0,644	0,024	(0,594; 0,691)	0,288	0,023	(0,242; 0,338)	0,068	0,011	(0,048; 0,091)
	4	0,523	0,024	(0,476; 0,570)	0,385	0,024	(0,336; 0,432)	0,092	0,013	(0,070; 0,118)
	5	0,581	0,022	(0,536; 0,627)	0,327	0,022	(0,285; 0,370)	0,092	0,013	(0,067; 0,118)
	6	0,651	0,022	(0,606; 0,696)	0,258	0,020	(0,218; 0,297)	0,090	0,014	(0,067; 0,117)
	7	0,656	0,023	(0,612; 0,700)	0,260	0,021	(0,218; 0,300)	0,084	0,013	(0,061; 0,109)
	8	0,658	0,022	(0,615; 0,703)	0,267	0,021	(0,230; 0,311)	0,075	0,012	(0,052; 0,100)
2	1	0,496	0,021	(0,458; 0,534)	0,442	0,021	(0,402; 0,484)	0,061	0,010	(0,042; 0,080)
	2	0,654	0,021	(0,617; 0,693)	0,277	0,019	(0,242; 0,316)	0,068	0,010	(0,047; 0,087)
	3	0,612	0,020	(0,570; 0,652)	0,313	0,019	(0,273; 0,353)	0,075	0,012	(0,053; 0,098)
	4	0,510	0,022	(0,470; 0,551)	0,438	0,022	(0,396; 0,478)	0,052	0,009	(0,036; 0,071)
	5	0,511	0,023	(0,466; 0,554)	0,425	0,023	(0,381; 0,472)	0,064	0,011	(0,045; 0,085)
	6	0,653	0,020	(0,618; 0,693)	0,271	0,017	(0,237; 0,304)	0,075	0,012	(0,053; 0,096)
	7	0,659	0,020	(0,618; 0,702)	0,273	0,018	(0,239; 0,311)	0,068	0,011	(0,047; 0,089)
	8	0,651	0,021	(0,609; 0,689)	0,292	0,019	(0,254; 0,329)	0,058	0,011	(0,040; 0,079)
28	1	0,901	0,011	(0,881; 0,920)	0,081	0,010	(0,062; 0,099)	0,018	0,005	(0,010; 0,028)
	2	0,714	0,020	(0,677; 0,751)	0,223	0,017	(0,189; 0,256)	0,064	0,011	(0,043; 0,088)
	3	0,793	0,017	(0,757; 0,826)	0,154	0,015	(0,126; 0,184)	0,053	0,010	(0,034; 0,073)
	4	0,827	0,016	(0,796; 0,856)	0,123	0,013	(0,099; 0,148)	0,050	0,010	(0,032; 0,069)
	5	0,711	0,021	(0,672; 0,751)	0,244	0,020	(0,209; 0,283)	0,046	0,010	(0,028; 0,064)
	6	0,715	0,018	(0,680; 0,748)	0,216	0,016	(0,184; 0,249)	0,070	0,011	(0,049; 0,092)
	7	0,721	0,019	(0,683; 0,757)	0,217	0,018	(0,185; 0,251)	0,062	0,011	(0,043; 0,084)
	8	0,729	0,022	(0,686; 0,770)	0,220	0,020	(0,181; 0,261)	0,052	0,011	(0,034; 0,073)
29	1	0,826	0,015	(0,796; 0,855)	0,153	0,015	(0,124; 0,184)	0,020	0,005	(0,011; 0,030)
	2	0,716	0,019	(0,679; 0,755)	0,219	0,017	(0,187; 0,253)	0,065	0,010	(0,047; 0,085)
	3	0,816	0,016	(0,785; 0,847)	0,139	0,014	(0,113; 0,168)	0,045	0,008	(0,030; 0,060)
	4	0,779	0,017	(0,746; 0,815)	0,134	0,014	(0,105; 0,160)	0,087	0,012	(0,065; 0,111)
	5	0,836	0,016	(0,807; 0,866)	0,117	0,013	(0,095; 0,145)	0,047	0,009	(0,031; 0,067)
	6	0,715	0,020	(0,677; 0,755)	0,213	0,018	(0,179; 0,248)	0,072	0,012	(0,051; 0,096)
	7	0,721	0,019	(0,685; 0,756)	0,214	0,017	(0,183; 0,251)	0,066	0,011	(0,045; 0,085)
	8	0,729	0,021	(0,690; 0,768)	0,217	0,020	(0,179; 0,257)	0,053	0,011	(0,034; 0,075)
31	1	0,820	0,015	(0,792; 0,849)	0,136	0,014	(0,110; 0,161)	0,044	0,008	(0,030; 0,061)
	2	0,719	0,019	(0,680; 0,758)	0,216	0,017	(0,181; 0,253)	0,065	0,010	(0,046; 0,087)
	3	0,796	0,016	(0,765; 0,827)	0,159	0,015	(0,133; 0,188)	0,046	0,008	(0,031; 0,063)
	4	0,864	0,013	(0,838; 0,889)	0,095	0,011	(0,075; 0,119)	0,041	0,008	(0,027; 0,056)
	5	0,808	0,018	(0,773; 0,841)	0,153	0,017	(0,122; 0,188)	0,038	0,008	(0,024; 0,054)
	6	0,721	0,018	(0,685; 0,756)	0,207	0,017	(0,177; 0,239)	0,072	0,011	(0,052; 0,092)
	7	0,724	0,018	(0,688; 0,758)	0,211	0,017	(0,180; 0,243)	0,065	0,010	(0,045; 0,085)
	8	0,739	0,020	(0,698; 0,774)	0,208	0,019	(0,171; 0,247)	0,053	0,011	(0,033; 0,073)

Figure 3.1 Graphique des moyennes a posteriori de la proportion de population finie.

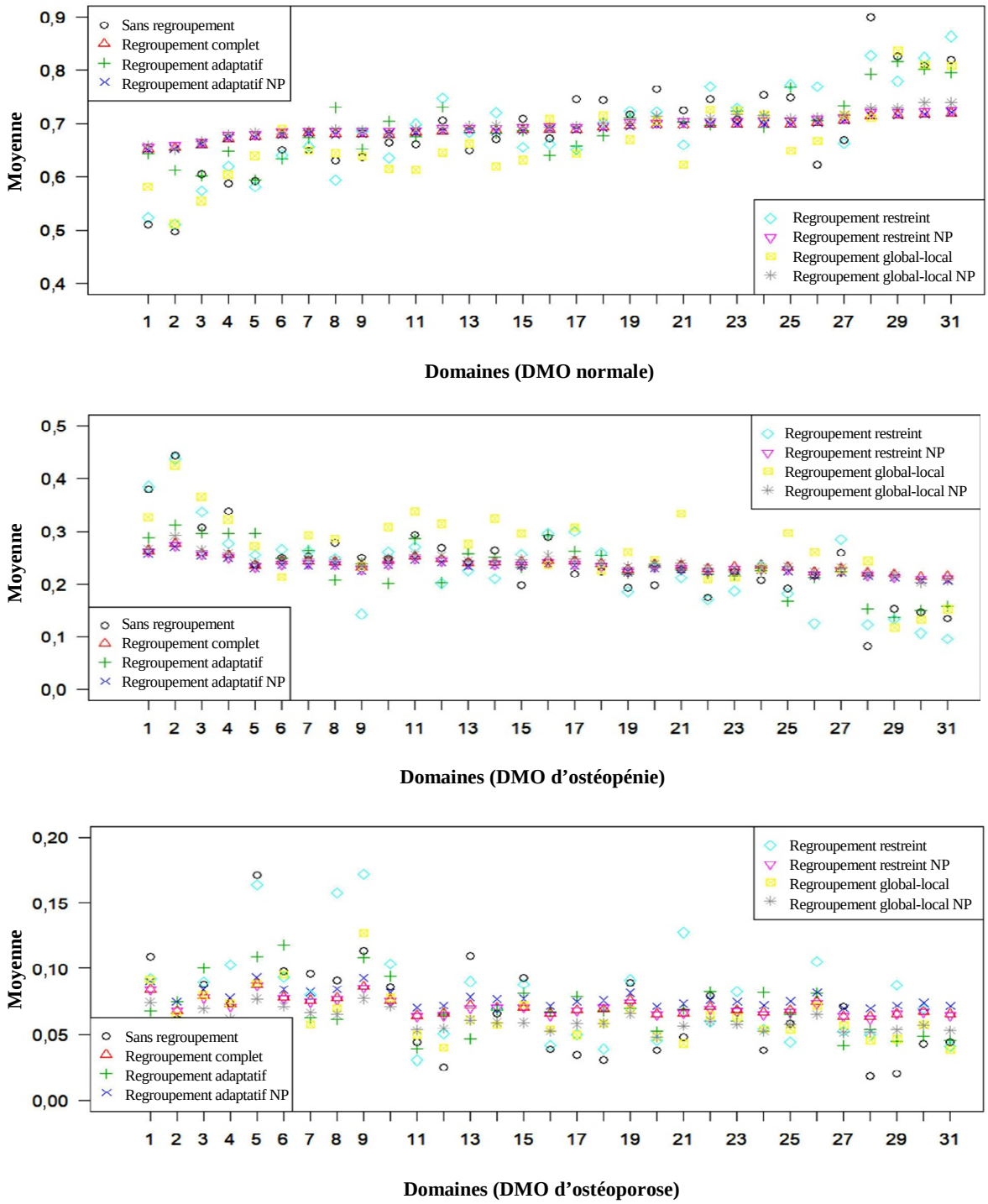


Figure 3.2 Graphique des écarts-types a posteriori de la proportion de la population finie.

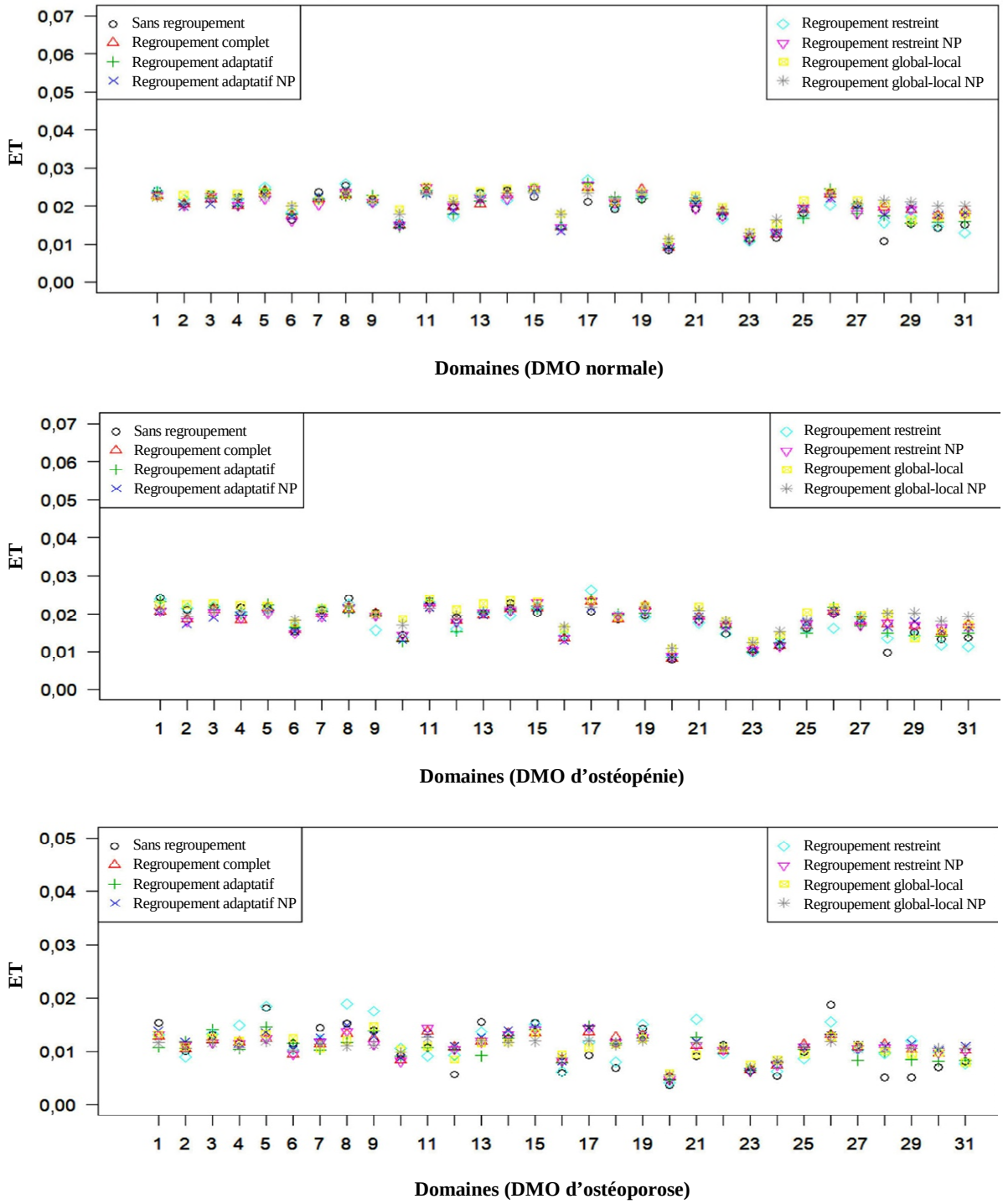
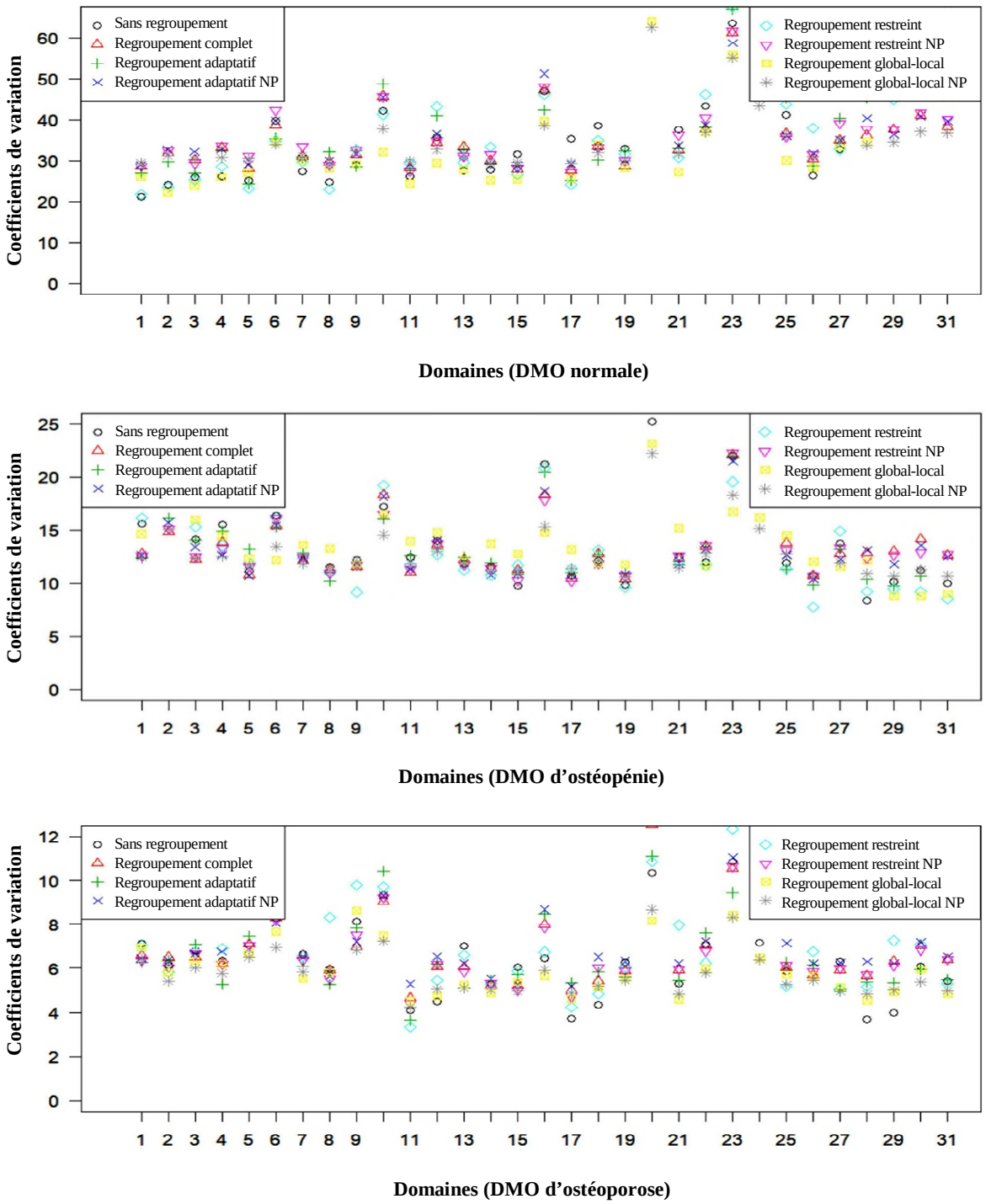


Figure 3.3 Graphique des coefficients de variation de la proportion finie de la population.



Le tableau 3.3 montre les résultats des deux mesures pour chaque modèle de regroupement bayésien. Un modèle est considéré comme plus performant quand ses mesures estimées sont plus petites. La LCPO et le DIC dans les modèles paramétriques sont comparés au moyen du modèle global-local, et on obtient pour la LCPO et le DIC respectivement 12,707 et 4 984,56, ce qui indique les meilleures performances. Le modèle de regroupement restreint du modèle 4 a une valeur de LCPO de 12,886, ce qui indique la deuxième meilleure performance après le modèle de regroupement global-local, mais le DIC du modèle de regroupement adaptatif a une valeur de 4 998,90, ce qui est inférieur à celui du modèle de regroupement restreint.

Tableau 3.3
Comparaisons de $\overline{\text{LCPO}}$ et du DIC (IC à 95 %) selon les modèles 1 à 8

Modèle		$\overline{\text{LCPO}}$	DIC
Modèles paramétriques	Sans regroupement	13,167	4 999,19
	Regroupement complet	13,151	5 011,45
	Regroupement adaptatif	13,411	4 998,90
	Regroupement restreint	12,886	5 000,21
	Regroupement global-local	12,707	4 984,56
Modèles non paramétriques	Regroupement adaptatif	13,105	5 001,17
	Regroupement restreint	12,837	4 983,88
	Regroupement global-local	12,694	4 768,47

Il faut aussi garder à l'esprit que la version non paramétrique utilisant la même méthode de regroupement donne des valeurs similaires à celles de la méthode paramétrique. En particulier, bien que le modèle de regroupement global-local non paramétrique ait le plus grand nombre de paramètres à estimer, sa LCPO et son DIC ont respectivement des valeurs de 12,694 et 4 768,47, ce qui indique que ses performances sont supérieures à celles de tous les modèles. De plus, dans le modèle de regroupement restreint, les valeurs de la LCPO et du DIC du modèle non paramétrique sont respectivement de 12,837 et de 4 983,88, ce qui représente des performances meilleures que celles du modèle paramétrique. Ces résultats sont identiques dans l'échelle de LCPO du modèle de regroupement adaptatif (c'est-à-dire de base) (LCPO (paramétrique c. non paramétrique) = (13,411 c. 13,105)). Cela signifie que les performances de la version non paramétrique sont très bonnes pour nos données, bien que l'espace des paramètres ait des dimensions infinies.

Le tableau 3.4 illustre les valeurs statistiques calculées aux fins d'estimation du rétrécissement du modèle. Pour estimer le rétrécissement, nous avons calculé la moyenne et l'écart-type de la différence absolue par rapport au modèle sans rétrécissement pour chaque catégorie de DMO. Soit MP_{ik}, $i = 1, \ldots, I$, $k = 1, 2, 3$ désigne la moyenne a posteriori de la proportion de population finie pour chaque cellule dans le domaine i. La moyenne (MER) et l'écart-type (ETER) de l'estimateur de réduction sont

$$\text{MER}_k = \frac{1}{I}\sum_{i=1}^{I}\frac{|\text{MP}_{ik} - \text{MP}_{0ik}|}{\text{MP}_{0ik}},\ I = 31, \tag{3.3}$$

$$\mathrm{ETER}_k = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}\left(\frac{|\mathrm{MP}_{ik} - \mathrm{MP}_{0ik}|}{\mathrm{MP}_{0ik}} - \mathrm{MER}_k\right)^2}, \tag{3.4}$$

où MP_{0ik} est la moyenne a posteriori pour la proportion de population finie du modèle 1, un modèle sans regroupement.

En nous fondant sur ce calcul, nous montrons que la moyenne de la MER et l'ETER dans la cellule normale sont les plus petits selon le modèle de regroupement global-local. D'après l'analyse des données, avec les données de DMO appliquées dans la présente étude, le nombre de groupes d'échantillonnage par tranche est de 1 à 3, et nous pouvons confirmer que les caractéristiques des données sont identiques pour la plupart des régions. Cependant, dans le modèle global-local, on peut montrer que le problème de rétrécissement excessif causé par le regroupement se résout si l'on observe le plus petit degré de rétrécissement. En outre, les cellules de l'ostéopénie et de l'ostéoporose pourraient confirmer la tendance à un rétrécissement faible par rapport à d'autres modèles, et l'ETER dans le modèle de regroupement global-local peut se révéler faible. Entre-temps, dans le cas des modèles non paramétriques, l'indice de groupe affichait le nombre le plus élevé de 1 en raison de l'échantillonnage par tranche. On peut par conséquent supposer que la dépendance aux données était excessive dans le modèle sans regroupement.

De plus, le test de Geweke, le graphique d'autocorrélation et la taille effective de l'échantillon (TEE) ont été appliqués aux fins de diagnostic du modèle, et ils ont montré un mélange de chaînes important.

Tableau 3.4
Comparaison du rétrécissement selon les huit modèles, voir les équations (3.3) et (3.4)

Modèle	Normale		Ostéopénie		Ostéoporose	
	Moyenne	Type	Moyenne	Type	Moyenne	Type
Sans regroupement (pas de modèle de rétrécissement)						
Regroupement complet	0,087	0,072	0,217	0,320	0,522	0,622
Regroupement adaptatif	0,066	0,064	0,147	0,166	0,474	0,501
Regroupement restreint	0,088	0,071	0,210	0,302	0,587	0,728
Regroupement global-local	0,054	0,043	0,157	0,137	0,456	0,682
Regroupement adaptatif NP	0,088	0,074	0,209	0,306	0,524	0,609
Regroupement restreint NP	0,065	0,049	0,249	0,354	0,335	0,362
Regroupement global-local NP	0,085	0,072	0,206	0,311	0,427	0,430

NP : non paramétrique

4. Conclusion

Dans la présente étude, nous construisons des modèles de regroupement bayésien paramétrique hiérarchique et leurs versions non paramétriques au moyen d'une loi a priori de processus de Dirichlet de Ferguson (1973) pour regrouper les données. Les méthodes de regroupement élaborées ici sont utiles dans l'analyse des données d'enquête. Par une méthode de grille, nous avons tiré les paramètres avec des densités a posteriori non standard et un support qui se trouve dans un intervalle fini. En revanche, nous avons utilisé l'algorithme de Metropolis-Hastings pour tirer les paramètres avec support dans un intervalle infini.

Le processus de Dirichlet est supposé pour le paramètre d'intérêt $\boldsymbol{\pi}_i$, pour $i = 1, \ldots, I$, dans nos modèles. Nous appliquons l'algorithme d'échantillonnage par tranche aux fins de spécification de la loi a priori de Dirichlet, qui est une extension de la loi a priori de rupture de bâton très couramment utilisée qui a été proposée par Ishwaran et James (2001). Cinq modèles paramétriques sont modélisés dans un espace de paramètres aux dimensions finies, et trois versions non paramétriques ont un espace de paramètres aux dimensions infinies. Les huit modèles bayésiens hiérarchiques se distinguent également selon le type d'effets dans les paramètres du modèle. Pour le modèle de base (2.1), nous pouvons construire des modèles plus efficaces et efficients permettant un effet d'emprunt des domaines voisins dans les estimations pour petits domaines. Toutefois, les distributions a priori échangeables dans un modèle bayésien hiérarchique peuvent causer un problème de rétrécissement excessif. Pour pallier ce problème, l'effet d'un paramètre est divisé en deux éléments, comme le montre le modèle de base (2.2), dit modèle de regroupement global-local bayésien hiérarchique. Le modèle permet de regrouper des expériences similaires (domaine) et d'emprunter de l'information dans chaque domaine.

Pour comparer les huit modèles à partir de données réelles, nous utilisons les données sur la DMO fournies par la NHANES III. La DMO est statistiquement corrélée avec la probabilité de fractures, qui constituent un problème de santé publique important, surtout chez les femmes âgées. C'est pourquoi la DMO est un indicateur important du diagnostic de l'ostéoporose, qui pourrait permettre de prendre en charge les patients plus tôt pour améliorer leur solidité osseuse. Dans chaque échantillon, nous attribuons un indicateur basé sur trois catégories (normale, ostéopénie, ostéoporose) avant d'analyser les données. Les modèles hiérarchiques obtenus ayant une distribution a priori de regroupement pour les données de DMO donnent de meilleures performances que les autres modèles. Pour comparer les performances des modèles, nous avons calculé le critère d'information de déviance (DIC) et l'ordonnée prédictive conditionnelle logarithmique (LCPO). Nous avons alors constaté que le modèle de regroupement global-local était le plus performant. Bien que les versions non paramétriques des modèles aient un espace de paramètres aux dimensions infinies, elles présentaient pour les deux mesures de comparaison des valeurs semblables à celles du modèle de regroupement paramétrique avec un espace de paramètres aux dimensions finies. C'est pourquoi les résultats doivent être interprétés prudemment.

Remerciements

Ces travaux ont été soutenus par une subvention de la *National Research Foundation of Korea* (NRF) financée par le gouvernement coréen (MSIT) (2016R1D1A1B03932261). Cette recherche a aussi été financée par une subvention de la *Simons Foundation* (nº 353953, Balgobin Nandram).

Annexe

A. Calculs des modèles paramétriques

Soit $\mathbf{n} = (\mathbf{n}_1, \ldots, \mathbf{n}_I)'$ la matrice de réponses et $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_I)'$ la matrice des paramètres de proportion pour (2.1). Dans un modèle de regroupement adaptatif, soit $\boldsymbol{\Omega}_i = (\boldsymbol{\pi}_i, \boldsymbol{\mu}, \tau)$, $i = 1, \ldots, I$. Ici,

le modèle sans regroupement et le modèle de regroupement complet sont des cas particuliers de regroupement adaptatif, avec des paramètres $\boldsymbol{\mu}$ et τ. La densité a posteriori conditionnelle complète des paramètres pour les données est obtenue de la façon habituelle par combinaison de la vraisemblance et des distributions a priori, comme suit :

$$\pi(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_I \mid \mathbf{n}) \propto \left\{\prod_{i=1}^{I} f(\mathbf{n} \mid \boldsymbol{\pi}_i)\, \pi(\boldsymbol{\pi}_i)\right\} \pi(\boldsymbol{\mu}, \tau)$$
$$\propto \prod_{i=1}^{I} \left\{\frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{ik}^{n_{ik}+\mu_k\tau-1}\right\} \frac{(K-1)!}{(\tau+1)^2},$$

où $D(\boldsymbol{\mu}\tau) = \prod_{k=1}^{K} \Gamma(\mu_k\tau) \Big/ \Gamma\left(\sum_{k=1}^{K} \mu_k\tau\right)$.

Pour exécuter un échantillonneur de Gibbs, nous tirons les valeurs comme suit :

(a) Conditionnelle complète pour $\boldsymbol{\pi}_i$, $i = 1, \ldots, I$: Tirer $\boldsymbol{\pi}_i \mid \mathbf{n}_i, \boldsymbol{\mu}, \tau \sim \text{Dirichlet}((\mathbf{n}_i + \boldsymbol{\mu})\tau)$.

(b) Conditionnelle complète pour $\boldsymbol{\mu}$: Tirer

$$\pi(\boldsymbol{\mu} \mid \mathbf{n}, \boldsymbol{\pi}, \tau) \propto \prod_{i=1}^{I} \left\{\frac{\Gamma\left(\sum_{k=1}^{K} \mu_k\tau\right)}{\prod_{k=1}^{K} \Gamma(\mu_k\tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k\tau-1}\right\}. \tag{A.1}$$

Soit $\boldsymbol{\mu}_{(k)}$, $k = 1, \ldots, K$ le vecteur de paramètres autre que la k^{e} composante μ_k. Nous obtenons ensuite la densité a posteriori conditionnelle de μ_k, étant donné $\boldsymbol{\mu}_{(k)}$, à chaque étape. Ici, nous devons estimer les $K-1$ composantes du paramètre $\boldsymbol{\mu}$ de façon séquentielle. Nous pouvons ensuite calculer la valeur de la K^{e} composante de $\boldsymbol{\mu}$ au moyen de $\mu_K = 1 - \sum_{k=1}^{K-1} \mu_k$. À partir de la densité a posteriori conditionnelle (3.1), nous pouvons tirer μ_k, $k = 1, \ldots, K-1$ au moyen de la méthode de grille, avec le support $1 - \sum_{k'=1,\, k' \neq k}^{K-1} \mu_{k'}$.

(c) Conditionnelle complète pour τ : Tirer

$$\pi(\tau \mid \mathbf{n}, \boldsymbol{\pi}, \boldsymbol{\mu}) \propto \prod_{i=1}^{I} \left\{\frac{\Gamma\left(\sum_{k=1}^{K} \mu_k\tau\right)}{\prod_{k=1}^{K} \Gamma(\mu_k\tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k\tau-1}\right\} \frac{1}{(\tau+1)^2}.$$

Dans ce cas aussi, nous pouvons utiliser la méthode de grille pour τ. Étant donné que la méthode de grille peut être utilisée pour un support fermé, nous transformons τ en $\rho = 1/(1+\tau)$, $0 < \rho < 1$. La valeur absolue du jacobien est $1/\rho^2$. Ensuite, la densité a posteriori conditionnelle de ρ peut être exprimée comme suit :

$$\pi(\rho \mid \mathbf{n}, \boldsymbol{\pi}, \boldsymbol{\mu}) \propto \prod_{i=1}^{I} \left\{\frac{\Gamma\left(\sum_{k=1}^{K} \mu_k \frac{1-\rho}{\rho}\right)}{\prod_{k=1}^{K} \Gamma\left(\mu_k \frac{1-\rho}{\rho}\right)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \frac{1-\rho}{\rho} - 1}\right\}.$$

La densité a posteriori conditionnelle complète dans le cas d'un regroupement restreint est obtenue à partir de la vraisemblance et des distributions a priori, construites à partir de $\pi(\boldsymbol{\mu}, \tau)$, et d'une distribution a priori supplémentaire pour ϕ, $i = 1, \ldots, I$:

$$\pi(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_I \mid \mathbf{n}) = \left\{\prod_{i=1}^{I} f(\mathbf{n} \mid \boldsymbol{\pi}_i)\, \pi(\boldsymbol{\pi}_i)\right\} \pi(\boldsymbol{\mu}, \tau)\, \pi(\phi),$$

où $\mathbf{\Omega}_i = (\boldsymbol{\pi}_i, \boldsymbol{\mu}, \tau, \phi)$, $i = 1, \ldots, I$.

Pour l'échantillonneur de Gibbs, nous considérons les variables latentes z_i, $i = 1, \ldots, I$ d'une loi de Bernoulli avec le paramètre ϕ. Ici, la densité a posteriori conjointe s'obtient comme suit :

$$\begin{aligned}\pi(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_I \mid \mathbf{n}) &\propto \left\{\prod_{i=1}^{I} f(\mathbf{n} \mid \boldsymbol{\pi}_i)\, \pi(\boldsymbol{\pi}_i \mid z_i)\, \pi(z_i)\right\} \pi(\boldsymbol{\mu}, \tau)\, \pi(\phi) \\ &\propto \left\{\prod_{i=1}^{I}\left[\phi \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau + n_{ik} - 1}\right]^{z_i} \left[(1-\phi)\prod_{k=1}^{K} \pi_{ik}^{n_{ik}}\right]^{1-z_i}\right\} \\ &\quad \times \frac{(K-1)!}{(\tau+1)^2} I_{\left(\phi \in \left(\frac{1}{2}, 1\right)\right)}.\end{aligned}$$

Puis, notre échantillonneur de Gibbs s'écrit comme suit :

(a) Conditionnelle complète pour ϕ : Tirer

$$\pi(\phi \mid \mathbf{n}, \boldsymbol{\mu}, \tau, \mathbf{z}) \propto \phi^{\sum_{i=1}^{I} z_i} (1-\phi)^{I - \sum_{i=1}^{I} z_i}, \ \frac{1}{2} \le \phi \le 1.$$

En d'autres termes, étant donné $\boldsymbol{\mu}, \tau, \mathbf{z}$, et les données, ϕ suit une loi bêta tronquée avec des paramètres $\sum_{i=1}^{I} z_i$ et $I - \sum_{i=1}^{I} z_i$, et la borne inférieure de ϕ est $1/2$.

(b) Conditionnelle complète pour z_i, $i = 1, \ldots, I$: Tirer $z_i \mid \mathbf{n}, \boldsymbol{\pi}_i, \boldsymbol{\mu}, \tau \sim \text{Bernoulli}(p_i)$, où

$$p_i = \left(\phi \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1} \Big/ D(\boldsymbol{\mu}\tau)\right) \Big/ \left(\phi \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1} \Big/ D(\boldsymbol{\mu}\tau) + (1-\phi) I_{\left(\sum_{k=1}^{K} \pi_{ik} = 1\right)} \Big/ K\right).$$

(c) Conditionnelle complète pour $\boldsymbol{\pi}_i$, $i = 1, \ldots, I$: Tirer

$$\pi(\boldsymbol{\pi}_i \mid \mathbf{n}, z_i, \boldsymbol{\mu}, \tau) \propto \left[\prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau + n_{ik} - 1}\right]^{z_i} \left[\prod_{k=1}^{K} \pi_{ik}^{1 + n_{ik} - 1}\right]^{1-z_i}.$$

Dans le cas de $z_i = 1$, nous pouvons générer la valeur de $\boldsymbol{\pi}_i$ à partir d'une loi de Dirichlet avec des paramètres $\boldsymbol{\mu}$ et τ. Dans d'autres cas, nous pouvons l'interpréter comme étant l'incertitude de la modélisation. Autrement dit, pour $\boldsymbol{\pi}_i$ étant donné $z_1 = 0$, $\mathbf{n}$ tire sa valeur d'une loi de Dirichlet uniforme avec un vecteur de paramètre $K \times 1$, où chaque composante a la valeur un.

(d) Conditionnelle complète pour $\boldsymbol{\mu}$: Tirer

$$\pi(\boldsymbol{\mu} \mid \mathbf{n}, \boldsymbol{\pi}, \tau) \propto \prod_{z_i = 1} \left\{\frac{\Gamma\left(\sum_{k=1}^{K} \mu_k \tau\right)}{\prod_{k=1}^{K} \Gamma(\mu_k \tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1}\right\}.$$

(e) Conditionnelle complète pour τ : Tirer

$$\pi(\tau \mid \mathbf{n}, \boldsymbol{\pi}, \boldsymbol{\mu}) \propto \prod_{z_i = 1} \left\{\frac{\Gamma\left(\sum_{k=1}^{K} \mu_k \tau\right)}{\prod_{k=1}^{K} \Gamma(\mu_k \tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1}\right\} \frac{1}{(\tau+1)^2}.$$

Bien entendu, le processus de génération des paramètres $\boldsymbol{\mu}$ et τ est similaire à celui du paramètre du regroupement adaptatif. En outre, les données utilisées pour $\boldsymbol{\mu}$ et τ sont $z_i = 1$, $i = 1, \ldots, I$.

Sinon, le vecteur de paramètre correspondant au domaine i selon un regroupement global-local se compose de θ, $\boldsymbol{\eta}$, $\mathbf{z}$, ϕ, σ^2, $\sigma_1^2, \ldots,$ et σ_{K-1}^2. Ensuite, la densité a posteriori conditionnelle complète pour les données dans le modèle s'écrit comme suit :

$$\begin{aligned}
\pi(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_I \,|\, \mathbf{n}) \propto & \left\{\prod_{i=1}^{I} f(\mathbf{n}_i \,|\, \boldsymbol{\eta}_i, \theta)\right\} \left\{\prod_{i=1}^{I} \pi(\boldsymbol{\eta}_i \,|\, z_i, \sigma^2, \sigma_1^2, \ldots, \sigma_{K-1}^2)\right\} \pi(\sigma^2) \\
& \times \pi(\sigma_1^2, \ldots, \sigma_{K-1}^2)\, \pi(\theta)\, \pi(\mathbf{z} \,|\, \phi)\, \pi(\phi) \\
= & \left\{\prod_{i=1}^{I} \frac{n_{i\cdot}!}{\prod_{k=1}^{K} n_{ik}!} \left\{\prod_{k=1}^{K-1} \left(\frac{e^{\theta+\eta_{ik}}}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{il}}}\right)^{n_{ik}}\right\} \left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{ik}}}\right)^{n_{iK}}\right\} \\
& \times \left\{\prod_{i=1}^{I} \left[\prod_{k=1}^{K-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\eta_{ik}^2\right)\right]^{z_i} \left[\prod_{k=1}^{K-1} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}\eta_{ik}^2\right)\right]^{1-z_i}\right\} \\
& \times \frac{1}{(1+\sigma^2)^2} \left\{\prod_{k=1}^{K-1} \frac{1}{(1+\sigma_k^2)^2}\right\} \frac{1}{\pi(1+\theta^2)} \left\{\prod_{i=1}^{I} \phi^{z_i} (1-\phi)^{1-z_i}\right\} I_{(1/2 \le \phi \le 1)},
\end{aligned}$$

où $\boldsymbol{\Omega}_i = (\theta, \boldsymbol{\eta}_i, z_i, \phi, \sigma^2, \sigma_1^2, \ldots, \sigma_{K-1}^2)$, $i = 1, \ldots, I$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_I)'$ et $\mathbf{z} = (z_1, \ldots, z_I)'$.

Ensuite, l'échantillonneur de Gibbs se présente ainsi :

(a) Conditionnelle complète pour ϕ : Tirer $\phi \,|\, \text{autres} \sim$ bêta tronquée$\left(\sum_{i=1}^{I} z_i, I - \sum_{i=1}^{I} z_i, \frac{1}{2}, 1\right)$.

(b) Conditionnelle complète pour z_i, $i = 1, \ldots, I$: Tirer $z_i \,|\, \text{autres} \sim \text{Bernoulli}(p_i)$, avec $p_i = \phi \prod_{k=1}^{K-1} \left(\exp(-\eta_{ik}^2/2\sigma^2)/\sqrt{2\pi\sigma^2}\right) \Big/ \left\{\phi \prod_{k=1}^{K-1} \exp(-\eta_{ik}^2/2\sigma^2)/\sqrt{2\pi\sigma^2} + (1-\phi) \prod_{k=1}^{K-1} \left(\exp(-\eta_{ik}^2/2\sigma_k^2)/\sqrt{2\pi\sigma_k^2}\right)\right\}$.

(c) Conditionnelle complète pour θ : Tirer

$$\pi(\theta \,|\, \text{autres}) \propto \left[\prod_{i=1}^{I} \prod_{k=1}^{K-1} \left(\frac{e^{\theta+\eta_{ik}}}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{il}}}\right)^{n_{ik}} \left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{il}}}\right)^{n_{iK}}\right] \frac{1}{1+\theta^2}.$$

(d) Conditionnelle complète pour η_{ik}, $i = 1, \ldots, I$, $k = 1, \ldots, (K-1)$: Tirer

$$\begin{aligned}
\pi(\theta \,|\, \text{autres}) \propto & \; I_{(z_i=1)} \left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{il}}}\right)^{\sum_{k=1}^{K} n_{ik}} \exp\left\{\sum_{k=1}^{K-1} \eta_{ik} n_{ik} - \frac{1}{2} \sum_{k=1}^{K-1} \frac{1}{\sigma^2} \eta_{ik}^2\right\} \\
& + I_{(z_i=0)} \left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{il}}}\right)^{\sum_{k=1}^{K} n_{ik}} \exp\left\{\sum_{k=1}^{K-1} \eta_{ik} n_{ik} - \frac{1}{2} \sum_{k=1}^{K-1} \frac{1}{\sigma_k^2} \eta_{ik}^2\right\}.
\end{aligned}$$

(e) Conditionnelle complète pour σ^2 : Tirer

$$\pi(\sigma^2 \,|\, \text{autres}) \propto \frac{1}{\sigma^{(K-1)\sum_{i=1}^{I} I(z_i=1)} (1+\sigma^2)^2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{z_i=1} \sum_{k=1}^{K-1} \eta_{ik}^2\right\}.$$

(f) Conditionnelle complète pour σ_k^2 $k = 1, \ldots, (K-1)$: Tirer

$$\pi(\sigma^2 \mid \text{autres}) \propto \frac{1}{\sigma_k^{\sum_{i=1}^{I} I_{(z_i=0)}} \left(1+\sigma_k^2\right)^2} \exp\left\{-\frac{1}{2\sigma_k^2} \sum_{z_i=0} \eta_{ik}^2\right\}.$$

Dans ce modèle, nous proposons d'utiliser l'algorithme de Metropolis-Hastings, qui est l'algorithme de méthode de Monte Carlo par chaîne de Markov (MCMC) le plus couramment employé pour estimer la valeur du paramètre d'emplacement, $(\theta, \boldsymbol{\eta})$. Bien entendu, $\left(\sigma^2, \sigma_1^2, \ldots, \sigma_{(K-1)}^2\right)$ est tiré des distributions conditionnelles complètes ci-dessus, et l'échantillonneur de Gibbs est exécuté au moyen de la méthode de grille.

B. Calculs des modèles non paramétriques

Afin de regrouper les paramètres dans des modèles bayésiens non paramétriques, nous appliquons la méthode d'échantillonnage par tranche présentée par Kalli, Griffin et Walker (2011). Ils ont proposé une version efficace de l'échantillonneur par tranche pour des modélisations par mélange selon le processus de Dirichlet construites par Walker (2007). Supposons que les observations y_i, $i = 1, \ldots, I$ sont générées dans le modèle par mélange selon le processus de Dirichlet avec le paramètre θ. Autrement dit,

$$\begin{aligned} y_i \mid \mathrm{G} &\overset{\text{iid}}{\sim} \mathrm{G} \\ \mathrm{G} &\sim \mathrm{PD}(\alpha, \mathrm{G}_0) \end{aligned}$$

Ici, nous écrivons $\mathrm{G} \sim \mathrm{PD}(\alpha, \mathrm{G}_0)$ pour indiquer que G suit un processus Dirichlet avec le paramètre $\alpha > 0$. Alors, G a une représentation en rupture de bâton (Sethuraman, 1994) donnée par

$$P(y_i \mid G, \alpha) = \sum_{j=1}^{\infty} w_j f(y_i \mid \theta_j)$$

où $\theta_1, \theta_2, \theta_3, \ldots$ sont des valeurs indépendantes et identiquement distribuées (iid) à partir de P_0 et

$$w_1 = v_1, \qquad w_j = v_j \prod_{l<j} (1 - v_l)$$

avec v_j étant la valeur iid à partir de $\text{Bêta}(1, \alpha)$, voir aussi Antoniak (1974).

L'algorithme de l'échantillonneur par tranche proposé par Walker (2007) introduit une variable latente $u \in (0, 1)$, $d_i = 1, 2, \ldots$ pour effectuer l'échantillonnage sur la distribution conjointe. Premièrement, la variable latente u_i a une densité conjointe comme suit :

$$P(y_i, u_i \mid G, \alpha) = \sum_{j=1}^{\infty} I_{(u_i < w_j)} f(y_i \mid \theta_j).$$

Plus tard, ils ont introduit une variable latente d_i représentant l'attribution de groupe de l'observation i. À ce moment, la densité conjointe de (y_i, u_i, d_i) est la suivante :

$$P(y_i, u_i, d_i \mid G, \alpha) = \sum_{j=1}^{\infty} I_{(u_i < w_{d_i = j})} f(y_i \mid \theta_j).$$

Ensuite, nous devons échantillonner le paramètre θ, α, et v y compris les variables latentes u et d à chaque itération d'un échantillonneur de Gibbs. Kalli et coll. (2011) ont montré comment effectuer un échantillonnage par tranche pour les modélisations par mélanges selon le processus de Dirichlet en traitant u et v comme des blocs dans l'algorithme de base décrit par Walker (2007). L'algorithme est le suivant :

1. $\pi(\theta_j \mid \ldots) \propto G_0(\theta_j) \prod_{d_i = j} f(y_i \mid \theta_j)$,
2. $\pi(v_j) \propto \text{Bêta}(a_j, b_j)$, où $a_j = 1 + \sum_{i=1}^{I} I_{(d_i = j)}$ et $b_j = \alpha + \sum_{i=1}^{I} I_{(d_i > j)}$,
3. $\pi(u_i \mid \ldots) \propto I_{(0 < u_i < \xi_{d_i})}$, où $\xi_j = (1 - \kappa)\kappa^{j-1}$, et κ est une constante,
4. $P(d_i = k \mid \ldots) \propto I_{(k:\xi_k > u_i)} w_k / \xi_k f(y_i \mid \theta_k)$,
5. $\pi(\alpha \mid \ldots) \propto \alpha^J \prod_{j=1}^{J} (1 - v_j)^{\alpha - 1} \pi(\alpha)$.

Aux fins de l'article, nous devons échantillonner les variables suivantes à chaque itération d'un échantillonneur de Gibbs :

$$\{(\boldsymbol{\pi}_j, v_j),\ j = 1, 2, \ldots, ; (d_i, u_i),\ i = 1, \ldots, I\}.$$

En général, κ est égal à 0,5. Toutefois, nous utilisons κ comme paramètre de réglage pour le modèle bayésien hiérarchique. Ensuite, la densité a posteriori complète pour le regroupement adaptatif non paramétrique des données est obtenue comme suit :

$$\begin{aligned} P(\mathbf{n}, \mathbf{d}, \mathbf{u} \mid \mathbf{v}, \boldsymbol{\pi}) & \left\{ \prod_{i=1}^{I} \pi(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}, \tau) \right\} \pi(\boldsymbol{\mu}, \tau)\, \pi(\mathbf{v} \mid \alpha)\, \pi(\alpha) \\ & \propto \left\{ \prod_{i=1}^{I} I_{(u_i < \xi_{d_i})} \frac{w_{d_i}}{\xi_{d_i}} n_{i.}! \prod_{k=1}^{K} \frac{\pi_{d_i k}^{n_{ik}}}{n_{ik}!} \right\} \left\{ \prod_{i=1}^{I} \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1} \right\} \\ & \times \left\{ \frac{(K-1)!}{(1+\tau)^2} \right\} \left\{ \prod_{j=1}^{J} \frac{1}{B(1, \alpha)} (1 - v_j)^{\alpha - 1} \right\} \frac{1}{(1+\alpha)^2}, \end{aligned}$$

où $\mathbf{d} = (d_1, \ldots, d_I)$, $\mathbf{u} = (u_1, \ldots, u_I)$, $\mathbf{v} = (v_1, \ldots, v_j)$, et les hyperparamètres sont mutuellement indépendants. On peut alors effectuer l'échantillonneur de Gibbs en deux étapes.

La première consiste à regrouper les données :

(a) Tirer pour u_i à partir de uniforme$(0, \xi_{d_i})$.

(b) Tirer pour d_i à partir de $P(d_i = j \mid \text{autres}) = I_{(u_i < \xi_j)} w_j / \xi_j \prod_{k=1}^{K} \pi_{jk}^{n_{ik}} / n_{ik}!$. Puis, nous pouvons générer la valeur de chaque paramètre à partir de la densité conditionnelle suivante :

(c) Tirer $\boldsymbol{\pi}_j$, $j = 1, \ldots, J$, à partir de Dirichlet$\left(\boldsymbol{\mu}\tau + \sum_{d_i = j} \mathbf{n}_i - 1\right)$.

(d) Tirer v_j, $j = 1, \ldots, J$, à partir de Bêta$\left(1 + \sum_{i=1}^{I} I_{(d_i = j)}, \alpha + \sum_{i=1}^{I} I_{(d_i = j)}\right)$.

(e) Tirer $\boldsymbol{\mu}$ à partir de

$$\pi(\boldsymbol{\mu} \mid \text{autres}) \propto \prod_{j=1}^{J} \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{jk}^{\mu_k \tau - 1}.$$

(f) Tirer τ à partir de

$$\pi(\tau \mid \text{autres}) \propto \left\{\prod_{j=1}^{J} \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{jk}^{\mu_k \tau - 1}\right\} \frac{(K-1)!}{(\tau+1)^2}.$$

Pour notre échantillonneur de Gibbs, nous devons transformer τ en $\rho = 1/(1+\tau)$, $0 < \rho < 1$ parce que nous devons utiliser la méthode de grille pour τ, qui est tirée de la distribution a priori non informative avec un support de variable égal à $(0, \infty)$.

$$\pi(\rho \mid \text{autres}) \propto \left\{\prod_{j=1}^{J} \frac{1}{D\left(\boldsymbol{\mu}\frac{1-\rho}{\rho}\right)} \prod_{k=1}^{K} \pi_{jk}^{\mu_k \frac{1-\rho}{\rho} - 1}\right\} (K-1)!.$$

(g) Tirer α à partir de

$$\pi(\alpha \mid \text{autres}) \propto \left\{\prod_{j=1}^{J} (1-v_j)^{\alpha-1}\right\} \frac{\alpha^J}{(1+\alpha)^2}.$$

Le paramètre α est également tiré de la distribution a priori non informative avec un support de variable égal à $(0, \infty)$, comme dans le cas de τ. Par conséquent, nous devons transformer α en δ, avec le jacobien $1/\delta^2$. Ensuite, la densité conditionnelle pour δ est obtenue comme suit :

$$\pi(\delta \mid \text{autres}) \propto \left\{\prod_{j=1}^{J} (1-v_j)^{\frac{1-\delta}{\delta}-1}\right\} \left(\frac{1-\delta}{\delta}\right)^J.$$

La version non paramétrique du regroupement restreint a $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK})$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, $0 \le \mu_k \le 1$, $\sum_{k=1}^{K} \mu_k = 1$, $\mathbf{d} = (d_1, \ldots, d_I)$, et $d_i = j$ pour $j = 1, 2, \ldots, J$. Ensuite, nous composons la densité a posteriori complète pour ces données au moyen de l'équation suivante :

$$P(\mathbf{n}, \mathbf{d}, \mathbf{u} \mid \mathbf{v}, \boldsymbol{\pi}) \left\{\prod_{j=1}^{J} \pi(\boldsymbol{\pi}_j \mid \boldsymbol{\mu}, \tau, \phi)\right\} \pi(\boldsymbol{\mu}, \tau)\, \pi(\phi)\, \pi(\mathbf{v} \mid \alpha)\, \pi(\alpha).$$

Dans notre échantillonneur de Gibbs, nous considérons les variables latentes z_j, pour $j = 1, \ldots, J$, comme la version paramétrique du regroupement restreint, où les indices pour le paramètre sont égaux au paramètre $\boldsymbol{\pi}$. À ce moment, l'étape de regroupement des données est la même que ci-dessus, et on génère le paramètre comme suit :

(a) Tirer ϕ de Bêta$\left(\sum_{j=1}^{J} z_j,\ J - \sum_{j=1}^{J} z_j,\ \phi \in \left(\frac{1}{2}, 1\right)\right)$ tronquée.

(b) Tirer z_j de Bernoulli(p_j),

où $p_j = \left(\phi \prod_{k=1}^{K} \pi_{jk}^{\mu_k \tau - 1} \Big/ D(\boldsymbol{\mu}\tau)\right) \Big/ \left\{\phi \prod_{k=1}^{K} \pi_{jk}^{\mu_k \tau - 1} \Big/ D(\boldsymbol{\mu}\tau) + (1-\phi)\, I_{\left(\sum_{k=1}^{K} \pi_{jk} = 1\right)} \Big/ K - 1\right\}$.

(c) Tirer $\boldsymbol{\pi}_j$ de Dirichlet$\left(I_{(z_j=1)}\boldsymbol{\mu}\tau + I_{(z_j=0)}\, \mathbf{1} + \sum_{d_i = j} \mathbf{n}_i\right)$.

(d) Tirer v_j de Bêta$\left(1 + \sum_{i=1}^{I} I_{(d_i = j)},\ \alpha + \sum_{i=1}^{I} I_{(d_i > j)}\right)$.

(e) Tirer $\boldsymbol{\mu}$ à partir de

$$\pi(\boldsymbol{\mu} \mid \text{autres}) \propto \prod_{j=1}^{J} \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{jk}^{\mu_k \tau - 1}.$$

(f) Tirer τ à partir de

$$\pi(\tau \mid \text{autres}) \propto \left[\prod_{j=1}^{J} \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{jk}^{*\mu_k \tau - 1}\right] \frac{(K-1)!}{(\tau+1)^2}.$$

(g) Tirer α à partir de

$$\pi(\alpha \mid \text{autres}) \propto \left\{\prod_{j=1}^{m} (1-\nu_j)^{\alpha-1}\right\} \frac{\alpha^J}{(1+\alpha)^2}.$$

Enfin, la densité a posteriori complète est calculée au moyen de la densité a posteriori conjointe et des distributions a priori pour leurs paramètres dans le modèle de regroupement global-local bayésien non paramétrique. Ici, l'algorithme de regroupement des données est le même que l'inférence ci-dessus. Ensuite, nous estimons chaque paramètre de la manière suivante :

(a) Tirer ϕ de Bêta$\left(\sum_{j=1}^{j} z_j,\ I - \sum_{j=1}^{J} z_j,\ \frac{1}{2},\ 1\right)$ tronquée.

(b) Tirer z_j de Bernoulli(p_j), où $p_j = \phi \prod_{k=1}^{K-1}\left(\exp\left(-\eta_{jk}^2 / 2\sigma^2\right) / \sqrt{2\pi\sigma^2}\right) \Big/ \left\{\phi \prod_{k=1}^{K-1} \left(\exp(-\eta_{jk}^2 / 2\sigma^2) / \sqrt{2\pi\sigma^2}\right) + (1-\phi) \prod_{k=1}^{K-1}\left(\exp\left(-\eta_{jk}^2 / 2\sigma_k^2\right) / \sqrt{2\pi\sigma_k^2}\right)\right\}$.

(c) Tirer θ à partir de

$$\pi(\theta \mid \text{autres}) \propto \left[\prod_{i=1}^{I} \prod_{k=1}^{K-1} \left(\frac{e^{\theta+\eta_{d_i k}}}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{d_i l}}}\right)^{n_{ik}} \left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{d_i l}}}\right)^{n_{iK}}\right].$$

(d) Tirer pour η_{jk}, $k = 1, \ldots, (K-1)$:

$$\pi(\eta_{jk} \mid \text{autres}) \propto I_{(z_j=1)} \left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{jl}}}\right)^{\sum_{d_i=j}\sum_{k=1}^{K} n_{ik}} \exp\left\{\sum_{d_i=j}\sum_{k=1}^{K-1} \eta_{jk} n_{ik} - \frac{1}{2\sigma^2}\sum_{k=1}^{K-1} \eta_{jk}^2\right\}$$

$$+ I_{(z_i=0)} \left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{jl}}}\right)^{\sum_{d_i=j}\sum_{k=1}^{K} n_{ik}} \exp\left\{\sum_{d_i=j}\sum_{k=1}^{K-1} \eta_{jk} n_{ik} - \frac{1}{2}\sum_{k=1}^{K-1} \frac{1}{\sigma_k^2} \eta_{jk}^2\right\}.$$

(e) Tirer ν_j de Bêta$\left(1 + \sum_{i=1}^{I} I_{(d_i=j)},\ I_{(z_j=1)}\alpha + I_{(z_j=0)}\alpha_0 + \sum_{i=1}^{I} I_{(d_i>j)}\right)$.

(f) Tirer σ^2 à partir de

$$\pi(\sigma^2 \mid \text{autres}) \propto \frac{1}{\sigma^{K-1\sum_{j=1}^{J} I_{(z_j=1)}} (1+\sigma^2)^2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{z_j=1} \sum_{k=1}^{K-1} \eta_{jk}^2\right\}.$$

(g) Tirer σ_k^2 $k = 1, \ldots, (K-1)$ à partir de

$$\pi(\sigma_k^2 \mid \text{autres}) \propto \frac{1}{\sigma_k^{\sum_{j=1}^{J} I_{(z_j=0)}} (1+\sigma_k^2)^2} \exp\left\{-\frac{1}{2\sigma_k^2} \sum_{z_j=0} \eta_{jk}^2\right\}.$$

(h) Tirer α à partir de

$$\pi(\alpha \mid \text{autres}) \propto \frac{\alpha^{\sum_{j=1}^{J} I_{(z_j=1)}}}{(1+\alpha)^2} \left(\prod_{z_j=1} (1-\nu_j)\right)^{\alpha-1}.$$

(i) Tirer α_0 from

$$\pi(\alpha_0 \mid \text{autres}) \propto \frac{\alpha_0^{\sum_{j=1}^{J} I_{(z_j=0)}}}{(1+\alpha_0)^2} \left(\prod_{z_j=0} (1-\nu_j)\right)^{\alpha_0 - 1}.$$

Bibliographie

Agresti, A., et Hitchcock, D.B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14, 297-330.

Antoniak, C.E. (1974). Mixture of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2, 1152-1174.

Consonni, G., et Veronese, P. (1995). A Bayesian method for combining results from several binomial experiments. *Journal of the American Statistical Association*, 90, 935-944.

DuMouchel, W.H., et Harris, J.E. (1983). Bayes methods for combining the results of cancer studies in humans and other species. *Journal of the American Statistical Association*, 78, 293-308.

Dunson, D.B. (2009). Nonparametric Bayes local partition models for random effects. *Biometrika*, 96, 249-262.

Evans, R., et Sedransk, J. (1999). Methodology for pooling subpopulation regressions when sample sizes are small and there is uncertainty about which subpopulations are similar. *Statistica Sinica*, 9, 345-359.

Evans, R., et Sedransk, J. (2003). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain: II. *Journal of Statiatical Planning and Inference*, 111, 95-100.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 209-230.

Gneiting, T., et Reftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359-378.

Ishwaran, H., et James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96, 161-173.

Kalli, M., Griffin, J.E. et Walker, S.G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21, 93-105.

Leonard, T. (1977). Bayes simultaneous estimation for several multinomial distributions. *Communications in Statistics: Theory and Methods*, 6, 619-630.

Malec, D., et Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, 79, 593-601.

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-112.

Nandram, B., et Yin, J. (2016a). Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline. *Statistical Methodology*, 28, 1-17.

Nandram, B., et Yin, J. (2016b). A nonparametric Bayesian prediction interval for a finite population mean. *Journal of Statistical Computation and Simulation*, 86, 3141-3157.

Nandram, B., Zhou, J. et Kim, D.H. (2019). A pooled Bayes test of independence for sparse contingency tables from small area. *Journal of Statistical Computation and Simulation*, 89(5), 899-926.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*, Second Edition. New York: John Wiley & Sons, Inc.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, 4, 639-650.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. et Linde, A.V.D. (2002). Bayesian measures of model complexity and fit. *Journal of royal statistical society, Series B*, 64(4), 538-639.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation*, 36, 45-54.

Yin, J., et Nandram, B. (2020a). A nonparametric Bayesian analysis of response data with gaps, outlier and ties. *Statistics and Applications*, 18(2), 121-141.

Yin, J., et Nandram, B. (2020b). A Bayesian small area model with Dirichlet processes on the responses. *Statistics in Transition*, 21(1), 1-17.

Techniques d'enquête, juin 2021
Vol. 47, nº 1, pp. 231-239
Statistique Canada, nº 12-001-X au catalogue

Note sur l'imputation multirobuste par appariement selon la moyenne prédictive réalisée avec des données d'enquête complexes

Sixia Chen, David Haziza et Alexander Stubblefield[1]

Résumé

L'imputation par appariement selon la moyenne prédictive est couramment utilisée pour régler les cas de non-réponse dans le cadre d'enquêtes. L'approche habituelle consiste à spécifier un modèle de régression unique. Dans la présente note, nous proposons une nouvelle procédure par appariement selon la moyenne prédictive permettant à l'utilisateur de spécifier plusieurs modèles de régression. L'estimateur obtenu est multirobuste en ce sens qu'il demeure convergent si l'un des modèles de régression spécifié est défini correctement. Les résultats d'une étude de simulation indiquent que la méthode proposée fonctionne bien en termes de biais et d'efficacité.

Mots-clés : Multirobuste; imputation par la méthode du plus proche voisin; données d'enquête; estimation de la variance.

1. Introduction

L'appariement selon la moyenne prédictive (AMP), qui est étroitement lié à l'imputation par la méthode du plus proche voisin (Chen et Shao, 2000; Beaumont et Bocci, 2009; Yang et Kim, 2019), est une procédure d'imputation populaire (Little, 1988; Yang et Kim, 2020). Pour la méthode du plus proche voisin, une valeur manquante de variable y est remplacée par la valeur y du répondant le plus proche par rapport à un vecteur de variables $\mathbf{x}$ complètement observées. Toutefois, dans le cas de cette méthode, l'estimateur obtenu peut être affecté par un biais non négligeable lorsque $\mathbf{x}$ a une grande dimension (Yang et Kim, 2019), un problème souvent appelé le « fléau de la dimension ». L'AMP, quant à elle, commence par l'ajustement d'un modèle paramétrique (par exemple un modèle de régression linéaire) fondé sur les unités répondantes, y étant la variable réponse et $\mathbf{x}$ l'ensemble de variables explicatives, ce qui permet d'obtenir un ensemble de valeurs prédites, $\hat{m}$, pour toutes les unités de l'échantillon (répondants et non-répondants). Une valeur manquante de la variable d'enquête y est ensuite remplacée par la valeur y du répondant le plus proche par rapport à $\hat{m}$. Ce dernier peut être considéré comme un résumé scalaire de l'information contenue dans le vecteur $\mathbf{x}$. Par conséquent, contrairement à la méthode du plus proche voisin, l'AMP n'est pas vulnérable à la dimension de $\mathbf{x}$, ce qui est souhaitable.

Les deux procédures susmentionnées sont des procédures non paramétriques. Par conséquent, elles sont moins vulnérables aux erreurs de spécification du modèle, contrairement aux méthodes paramétriques (par exemple l'imputation par la régression linéaire). De plus, elles font toutes deux partie de la classe des procédures d'imputation par donneur, c'est-à-dire qu'elles produisent des valeurs imputées admissibles obtenues avec des valeurs observées réelles « empruntées » à des répondants.

1. Sixia Chen, Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, États-Unis; David Haziza, Département de mathématiques et de statistique, Université d'Ottawa, Ottawa, Canada. Courriel : dhaziza@uottawa.ca; Alexander Stubblefield, Department of Economics, Michigan State University, East Lansing, MI 48823, États-Unis.

Pendant la première étape de l'AMP, l'information du vecteur $\mathbf{x}$ est convertie en un score unique $\hat{m}$ grâce à un modèle paramétrique (par exemple un modèle de régression linéaire). Si le modèle spécifié décrit bien le lien entre y et $\mathbf{x}$, cette procédure devrait être efficace sur le plan du biais. Par contre, s'il y a des erreurs importantes de spécification, elle peut produire des estimateurs biaisés.

Les approches multirobustes s'appuyant sur des modèles de régression et de non-réponse multiples améliorent la robustesse et réduisent les erreurs de spécification du modèle; voir Han et Wang (2013), Han (2014) et Chen et Haziza (2019a), entre autres. Dans la présente note, nous présentons une nouvelle procédure d'AMP qui permet d'utiliser plusieurs modèles, chacun pouvant être fondé sur différentes fonctionelles ou sur un ensemble différent de variables explicatives. Spécifier de multiples modèles peut s'avérer utile dans un certain nombre de situations (Chen et Haziza, 2017; Chen et Haziza, 2019b). Ces modèles peuvent être paramétriques ou non paramétriques. La logique sous-tendant la méthode proposée est d'ajuster chacun de ces modèles en fonction des unités répondantes, ce qui fournit plusieurs ensembles de valeurs prédites pour toutes les unités d'échantillonnage. Après avoir introduit le cadre théorique dans la section 2, nous décrivons comment combiner ces résultats pour obtenir les valeurs imputées dans la section 3. La procédure proposée est multirobuste en ce sens que l'estimateur obtenu est convergent si l'un seul des modèles est correctement spécifié. Compte tenu du fait que le vrai modèle qui lie y et $\mathbf{x}$ est inconnu, l'approche proposée est attrayante parce qu'elle protège contre les erreurs de spécification du modèle. De plus, contrairement à la procédure d'imputation multirobuste étudiée par Chen et Haziza (2017), la méthode proposée appartient à la classe des procédures d'imputation par donneur. Dans la section 4, nous effectuons une étude par simulation afin d'évaluer la méthode proposée en termes de biais et d'efficacité.

2. Configuration de base

Considérons une population finie $\mathcal{F}_N = \{(\mathbf{x}_i, y_i),\ i = 1, 2, \ldots, N\}$, générée à partir du modèle de superpopulation suivant :

$$y_i = m(\mathbf{x}_i) + \varepsilon_i, \tag{2.1}$$

où $m(\cdot)$ est une fonction inconnue, $\mathbf{x}_i$ est un vecteur de variables complètement observées associées à l'unité i et les ε_i représentent des variables aléatoires mutuellement indépendantes, où $\mathrm{E}(\varepsilon_i \mid \mathbf{x}_i) = 0$ et $\mathrm{V}(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2$. Par souci de simplicité, nous supposons que la variance est homoscédastique, mais notre méthode peut facilement se généraliser au cas de variances non constantes.

L'objectif est d'estimer la moyenne de la population $\theta = \mathrm{E}(y)$. Étant donné la population finie, un échantillon probabiliste S de taille n est sélectionné selon un plan d'échantillonnage avec des

probabilités d'inclusion du premier ordre π_i et des probabilités d'inclusion du deuxième ordre π_{ij}. Le poids d'échantillonnage associé à l'unité i est $w_i = \pi_i^{-1}$.

L'indicateur de réponse lié à l'unité i est r_i et $r_i = 1$ si y_i est observée, et $r_i = 0$ si y_i est manquante. L'ensemble $S_r = \{i \in S : r_i = 1\}$ représente l'ensemble de répondants pour la variable d'enquête y. Nous supposons que les données manquantes sont réparties au hasard (MAR) :

$$\Pr(r_i = 1 | \mathbf{x}_i, y_i) = \Pr(r_i = 1 | \mathbf{x}_i). \tag{2.2}$$

La procédure d'AMP habituelle peut être décrite comme suit : Pour commencer, nous utilisons le modèle de régression paramétrique suivant $\mathcal{M} = \{m(\mathbf{x}_i; \boldsymbol{\beta})\}$, où $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus (Yang et Kim, 2020). Pour $i \in S$, nous calculons le score $\hat{m}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$, où $\hat{\boldsymbol{\beta}}$ est un estimateur approprié de $\boldsymbol{\beta}$ fondé sur les unités répondantes. Ensuite, la valeur obtenue par imputation pour le y_i manquant est $y_i^* = y_j$, où j est l'indice du plus proche voisin de l'unité i, ce qui répond au critère $\mathcal{D}(\hat{m}_j, \hat{m}_i) \leq \mathcal{D}(\hat{m}_{j'}, \hat{m}_i)$ pour tout $j' \in S_r$, où $\mathcal{D}(\cdot,\cdot)$ indique une fonction de distance, par exemple la distance euclidienne. Pour que la procédure protège des erreurs de spécification, le modèle paramétrique doit satisfaire à la condition de continuité de Lipschitz (Yang et Kim, 2020). Cette condition peut ne pas être satisfaite pour certains modèles et fonctions couramment employés, y compris les modèles quadratiques (Yang et Kim, 2020).

3. Méthode proposée

La méthode proposée permet à l'utilisateur de spécifier des modèles de régression multiples pour la variable d'enquête y, ce qui augmente la probabilité de choisir un modèle qui reproduit efficacement la relation entre la variable réponse et les variables explicatives. L'approche est ainsi multirobuste et il n'est pas nécessaire de satisfaire à la condition de continuité de Lipschitz. Tant que l'un des modèles est correctement spécifié, l'estimateur obtenu sera convergent.

Nous employons une classe de modèles de régression : $\mathcal{M} = \{m^{(k)}(\mathbf{x}_i; \boldsymbol{\beta}^{(k)}), k = 1, 2, \ldots, K\}$. Pour imputer les valeurs manquantes, nous procédons comme suit :

(Étape 1). Obtenir les estimateurs $\hat{\boldsymbol{\beta}}^{(k)}$ de $\boldsymbol{\beta}^{(k)}$, $k = 1, 2, \ldots, K$, en résolvant les équations d'estimation pondérées ci-dessous :

$$\hat{U}_m^{(k)}(\hat{\boldsymbol{\beta}}^{(k)}) = \sum_{i \in S} w_i r_i \{y_i - m^{(k)}(\mathbf{x}_i; \boldsymbol{\beta}^{(k)})\} \frac{\partial m^{(k)}(\mathbf{x}_i; \boldsymbol{\beta}^{(k)})}{\partial \boldsymbol{\beta}^{(k)}} = 0. \tag{3.1}$$

(Étape 2). Pour $i \in S$, obtenir le vecteur de taille K de valeurs prédites

$$\mathbf{V}_i = \left(m^{(1)}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(1)}), m^{(2)}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(2)}), \ldots, m^{(K)}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(K)})\right)^{\mathrm{T}}.$$

(Étape 3). Ajuster un modèle de régression linéaire pondéré sans ordonnée à l'origine avec y comme variable réponse et $\mathbf{V}$ comme vecteur de variables explicatives. La valeur prédite obtenue associée à l'unité i est $\hat{M}\left(\mathbf{x}_i;\, \hat{\boldsymbol{\beta}},\, \hat{\boldsymbol{\eta}}\right)$:

$$\hat{M}\left(\mathbf{x}_i;\, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}\right) \;=\; \mathbf{V}_i^{\mathrm{T}} \hat{\boldsymbol{\eta}},$$

où $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \ldots, \hat{\boldsymbol{\beta}}^{(K)}\right)$ et

$$\hat{\boldsymbol{\eta}} \;=\; \left\{\sum_{i \in S} w_i r_i \mathbf{V}_i \mathbf{V}_i^{\mathrm{T}}\right\}^{-1} \sum_{i \in S} w_i r_i \mathbf{V}_i y_i. \tag{3.2}$$

(Étape 4). La valeur imputée pour le y_i manquant est $y_i^* = y_j$, où j est l'indice du voisin le plus proche de l'unité i, ce qui respecte le critère $\mathcal{D}\left\{\hat{M}(\mathbf{x}_j; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}), \hat{M}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})\right\} \leq \mathcal{D}\left\{\hat{M}(\mathbf{x}_{j'}; \hat{\boldsymbol{\eta}}), \hat{M}(\mathbf{x}_i;\, \hat{\boldsymbol{\beta}},\, \hat{\boldsymbol{\eta}})\right\}$ pour tout $j' \in S_r$.

Après les étapes 1 à 4, nous définissons l'estimateur imputé de θ :

$$\hat{\theta}_{\mathrm{MR}} \;=\; \frac{1}{\hat{N}} \sum_{i \in S} w_i \left\{r_i y_i \;+\; (1 - r_i)\; y_i^*\right\}, \tag{3.3}$$

où $\hat{N} = \sum_{i \in S} w_i$. En suivant une approche semblable à celle utilisée par Yang et Kim (2020), il est possible de montrer que l'estimateur $\hat{\theta}_{\mathrm{MR}}$ est multirobuste, puisqu'il est convergent si l'un des modèles est correctement spécifié.

La variance de $\hat{\theta}_{\mathrm{MR}}$ peut être estimée en suivant des procédures d'estimation de la variance par répliques (Rust et Rao, 1996; Wolter, 2007). En ce qui concerne l'AMP pour les données d'enquête, Yang et Kim (2020) ont également étudié des procédures de réplication. Le nombre de répliques est indiqué par L et $w_i^{(g)}$ représente le poids de réplication lié à l'unité i de la réplique g^{e}. L'équation ci-dessous permet d'obtenir un estimateur de la variance par répliques pour $\hat{\theta}_{\mathrm{MR}}$:

$$\hat{V}_{\mathrm{rep}}\left(\hat{\theta}_{\mathrm{MR}}\right) \;=\; \sum_{g=1}^{L} c_g \left(\hat{\theta}_{\mathrm{MR}}^{(g)} \;-\; \hat{\theta}_{\mathrm{MR}}\right)^2, \tag{3.4}$$

où

$$\hat{\theta}_{\mathrm{MR}}^{(g)} \;=\; \sum_{i \in S} w_i^{(g)} \left\{r_i y_i \;+\; (1 - r_i)\; y_i^{*(g)}\right\}$$

désigne l'estimateur $\hat{\theta}_{\mathrm{MR}}$ de la g^{e} réplique et $y_i^{*(g)}$ étant la valeur imputée liée à l'unité i de la g^{e} réplique, calculée en suivant les étapes 1 à 4 ci-dessus, en se fondant sur le poids de réplication $w_i^{(g)}$ plutôt que le poids d'échantillonnage originaux w_i. Le facteur c_g de l'équation (3.4) dépend de la méthode de réplication. Par exemple, avec l'estimateur jackknife avec suppression d'une unité, $L = n$, $c_g = n/(n-1)$ et $w_i^{(g)} = n/(n-1)\, w_i$ si $i \neq g$ et $w_i^{(g)} = 0$ si $i = g$.

4. Étude de simulation

Afin d'évaluer la performance de la méthode proposée en termes de biais et d'efficacité, nous avons réalisé une étude de simulation limitée. Nous avons généré $B = 2\,000$ populations finies, chacune ayant une taille de $N = 20\,000$. Tout d'abord, les variables explicatives $x_1 - x_4$ ont été générées à partir d'une distribution normale standard multidimensionnelle. Puis, nous avons généré la variable d'enquête y en fonction des variables $x_1 - x_4$ selon les modèles de régression suivants :

(M1). $y = 1 + x_1 + x_2 + x_3 + x_4 + \varepsilon$, où $\varepsilon \sim N(0, 1)$.

(M2). $y = 1 + x_1^2 + x_2^2 + x_3 + x_4 + x_3 x_4 + \varepsilon$, où $\varepsilon \sim N(0, 1)$.

Les modèles (M1) et (M2) sont linéaires et fondés sur les variables explicatives $x_1 - x_4$, sauf que (M2) comprend des termes quadratiques et une interaction.

Pour chaque population finie, un échantillon probabiliste S a été sélectionné selon un d'échantillonnage systématique avec probabilité proportionnelle à la taille au moyen de la variable de taille $z_i = \log(0{,}1 \,|\, y_i + v_i \,|\, + 4)$, où $v_i \sim N(0, 1)$. Les probabilités d'inclusion du premier ordre sont calculées à l'aide de la formule $\pi_i = n z_i \Big/ \sum_{i=1}^{N} z_i$ avec $n = 200$, 500 et 1 000.

Pour chaque échantillon, les indicateurs de réponse r_i ont été générés selon une distribution de Bernoulli de probabilité p_i, celle-ci valant :

$$p_i = 0{,}1 + 0{,}9 \times \frac{\exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})}{1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})}. \tag{4.1}$$

Nous avons utilisé deux ensembles de valeurs pour $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$: $(0; 1; 1; 1; 1)$ et $(1{,}38; 1; 1; 1; 1)$. Ils ont permis d'obtenir des taux de réponse d'environ 70 % et 50 %, respectivement.

Nous avons calculé les estimateurs de θ suivants :

(Naïf). La moyenne pondérée des répondants $\hat{\theta}_{\text{naïf}} = \sum_{i \in S_r} w_i y_i \Big/ \sum_{i \in S_r} w_i$.

(Rég). L'estimateur imputé obtenu selon une imputation par régression linéaire déterministe à l'aide du modèle (M1).

(AMP1). L'estimateur obtenu en suivant la procédure d'AMP où $\hat{m}_i$, $i \in S$, a été déterminé avec le modèle (M1).

(Nouveau1). L'estimateur imputé fondé sur la procédure multirobuste d'AMP proposée en utilisant les deux modèles (M1) et (M2).

(Nouveau2). L'estimateur imputé fondé sur la procédure multirobuste d'AMP proposée en utilisant les deux modèles (M1) et (M2) et deux modèles supplémentaires, soit (M3) et (M4). Pour (M3), seule x_1 est employé comme variable explicative et pour (M4) seule x_1^2 l'est.

Nous avons calculé le biais relatif Monte-Carlo (BRMMC), ainsi que son erreur-type relative (ETRMMC) et sa racine de l'erreur quadratique moyenne relative (REQMRMMC), ceux-ci étant respectivement définis comme suit :

$$\text{BRMMC} = \frac{2\,000^{-1}\sum_{b=1}^{2\,000}\left(\hat{\theta}_b - \theta_b\right)}{\theta_{\text{MC}}},$$

$$\text{ETRMMC} = \frac{\sqrt{(B-1)^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \hat{\theta}_{\text{MC}}\right)^2}}{\theta_{\text{MC}}}$$

et

$$\text{REQMRMMC} = \frac{\sqrt{(B-1)^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \theta_{\text{MC}}\right)^2}}{\theta_{\text{MC}}},$$

où θ_b désigne la moyenne de la b^{e} population et $\hat{\theta}_b$ indique l'estimateur $\hat{\theta}$ dans le b^{e} échantillon pour $b = 1, \ldots, 2\,000$ et

$$\theta_{\text{MC}} = \frac{1}{2\,000}\sum_{b=1}^{2\,000}\theta_b, \quad \hat{\theta}_{\text{MC}} = \frac{1}{2\,000}\sum_{b=1}^{2\,000}\hat{\theta}_b.$$

Les résultats sont présentés dans les tableaux 4.1 et 4.2. Le biais de l'estimateur naïf était important dans tous les scénarios, conformément aux attentes. Le tableau 4.1 indique que lorsque le modèle vrai était (M1) l'imputation par régression linéaire a très bien fonctionné en ce qui concerne le biais, conformément aux attentes. À la fois pour l'AMP1 et la méthode proposée, le biais était négligeable pour $n = 1\,000$ et petit pour $n = 500$ et $n = 200$. Par exemple, pour $n = 200$ et un taux de réponse de 70 %, le biais relatif était égal à 2,4 % pour l'AMP, Nouveau1 et Nouveau2. En ce qui concerne l'efficacité, l'imputation par régression linéaire a légèrement été meilleure que l'AMP et les méthodes proposées, conformément aux attentes. Par exemple, pour $n = 1\,000$ et un taux de réponse de 70 %, la REQM était égale à 7,5 % pour l'imputation par régression linéaire et égale à 8,0 % pour l'AMP, Nouveau1 et Nouveau2. Il faut souligner que la performance de l'AMP et celle des méthodes proposées étaient presque identiques pour tous les scénarios du tableau 4.1. Par conséquent, intégrer deux modèles supplémentaires n'a pas semblé avoir d'incidence sur l'efficacité de l'estimateur obtenu (Nouveau2).

Le tableau 4.2 indique que lorsque le modèle vrai était (M2), l'imputation par régression linéaire et l'AMP ont entrainé des biais importants dans tous les scénarios, conformément aux attentes. Étant une procédure d'imputation paramétrique, l'imputation par régression linéaire est vulnérable aux erreurs de spécification du modèle. Par contre, le biais de l'AMP était plus petit que celui de l'imputation par régression linéaire, ce qui indiquait une certaine résistance aux erreurs de spécification du modèle. Par exemple, pour $n = 1\,000$ et un taux de réponse de 70 %, le biais relatif était égal à -9,2 % pour l'imputation par régression linéaire et à -3,7 % pour l'AMP. La performance des méthodes proposées était

meilleure que celle de l'imputation par régression linéaire et de l'AMP en ce qui concerne le biais, l'erreur type et l'erreur quadratique moyenne, pour tous les scénarios. Enfin, la performance de Nouveau1 et celle de Nouveau2 étaient presque identiques.

Tableau 4.1
Biais relatif Monte-Carlo, erreur-type relative Monte-Carlo et racine de l'erreur quadratique moyenne relative Monte-Carlo lorsque le modèle vrai est (M1)

			Méthode				
Taux de réponse	**Taille de l'échantillon**	**Mesure $(\times 10^2)$**	**Naïf**	**Rég.**	**AMP1**	**Nouveau1**	**Nouveau2**
70 %	1 000	BRMMC	64,7	-0,1	0,4	0,4	0,4
		ETRMMC	7,5	7,5	8,0	8,0	8,0
		REQMRMMC	65,1	7,5	8,0	8,0	8,0
70 %	500	BRMMC	65,3	0,5	1,4	1,4	1,4
		ETRMMC	10,7	10,4	11,2	11,2	11,2
		REQMRMMC	66,1	10,4	11,3	11,3	11,3
70 %	200	BRMMC	64,6	0,3	2,4	2,4	2,4
		ETRMMC	16,5	16,7	17,5	17,5	17,6
		REQMRMMC	66,7	16,7	17,7	17,7	17,7
50 %	1 000	BRMMC	99,3	0,0	0,7	0,7	0,6
		ETRMMC	8,8	8,1	9,0	9,0	9,0
		REQMRMMC	99,7	8,1	9,1	9,1	9,1
50 %	500	BRMMC	98,9	-0,1	1,3	1,3	1,3
		ETRMMC	12,1	11,2	12,5	12,5	12,5
		REQMRMMC	99,6	11,2	12,6	12,6	12,6
50 %	200	BRMMC	99,8	0,8	4,3	4,3	4,4
		ETRMMC	19,3	17,7	19,6	19,6	19,6
		REQMRMMC	101,6	17,7	20,1	20,1	20,0

Tableau 4.2
Biais relatif Monte-Carlo, erreur-type relative Monte-Carlo et racine de l'erreur quadratique moyenne relative Monte-Carlo lorsque le modèle vrai est (M2)

			Méthode				
Taux de réponse	**Taille de l'échantillon**	**Mesure $(\times 10^2)$**	**Naïf**	**Rég.**	**AMP1**	**Nouveau1**	**Nouveau2**
70 %	1 000	BRMMC	7,5	-9,2	-3,7	0,1	0,1
		ETRMMC	3,5	3,5	3,9	3,1	3,1
		REQMRMMC	8,2	9,9	5,4	3,1	3,1
70 %	500	BRMMC	7,5	-9,4	-4,0	0,2	0,2
		ETRMMC	5,0	5,1	5,6	4,5	4,5
		REQMRMMC	9,0	10,7	6,9	4,5	4,5
70 %	200	BRMMC	7,6	-9,2	-4,0	0,1	0,1
		ETRMMC	7,8	7,9	8,5	6,8	6,8
		REQMRMMC	10,9	12,1	9,4	6,8	6,8
50 %	1 000	BRMMC	16,6	-11,3	-3,1	0,3	0,3
		ETRMMC	4,0	4,5	5,0	3,3	3,3
		REQMRMMC	17,1	12,2	5,9	3,3	3,3
50 %	500	BRMMC	16,5	-11,5	-3,5	0,3	0,3
		ETRMMC	5,7	6,3	7,0	4,8	4,7
		REQMRMMC	17,5	13,2	7,8	4,8	4,8
50 %	200	BRMMC	16,5	-12,0	-3,9	-0,1	-0,1
		ETRMMC	9,1	9,9	11,0	7,4	7,4
		REQMRMMC	18,8	15,6	11,7	7,4	7,4

Remerciements

S. Chen a reçu l'appui du *National Institute on Minority Health and Health Disparities* des *National Institutes of Health* [1R21MD014658-01A1] et de l'*Oklahoma Shared Clinical and Translational Resources* (U54GM104938) ainsi qu'un Institutional Development Award décerné par le *National Institute of General Medical Sciences*. Le contenu du document est sous la seule responsabilité des auteurs et n'est pas nécessairement représentatif du point de vue officiel des *National Institutes of Health*. Des subventions du Conseil de recherches en sciences naturelles et en génie du Canada ont contribué aux travaux de D. Haziza.

Bibliographie

Beaumont, J.-F., et Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Chen, S., et Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 102, 439-453.

Chen, S., et Haziza, D. (2019a). Multiply robust nonparametric multiple imputation for the treatment of missing data. *Statistica Sinica*, 29, 2035-2053.

Chen, S., et Haziza, D. (2019b). Recent developments in dealing with item nonresponse in surveys: A critical review. *Revue Internationale de Statistique*, 87, S192-S218.

Chen, J., et Shao, J. (2000). Nearest-neighbour imputation for survey data. *Journal of Official Statistics*, 16, 583-599.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109, 1159-1173.

Han, P., et Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, 100, 417-430.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-296.

Rust, K.F., et Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Wolter, K. (2007). *Introduction to Variance Estimation, 2nd Edition*. Springer, Berlin.

Yang, S., et Kim, J.K. (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. *Advances in Econometrics - The Econometrics of Complex Survey Data: Theory and Applications*, 39, 209-234.

Yang, S., et Kim, J.K. (2020). Predictive mean matching imputation in survey sampling. À paraître dans le *Scandinavian Journal of Statistics*.

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 36, No. 4, December 2020

Letter to the Editors
Andreas V. Georgiou 729

Basic Statistics of Jevons and Carli Indices under the GBM Price Model
Jacek Białek 737

Developing Land and Structure Price Indices for Ottawa Condominium Apartments
Kate Burnett-Isaacs, Ning Huang and W. Erwin Diewert 763

An Improved Fellegi-Sunter Framework for Probabilistic Record Linkage Between Large Data Sets
Marco Fortini 803

Three-Form Split Questionnaire Design for Panel Surveys
Paul M. Imbriano and Trivellore E. Raghunathan 827

Double Barreled Questions: An Analysis of the Similarity of Elements and Effects on Measurement Quality
Natalja Menold 855

The Representativeness of Online Time Use Surveys. Effects of Individual Time Use Patterns and Survey Design on the Timing of Survey Dropout
Petrus te Braak, Joeri Minnen and Ignace Glorieux 887

Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context
James Wagner, Brady T. West, Michael R. Elliott and Stephanie Coffey 907

Book Review: Paul C. Beatty, Debbie Collins, Lyn Kaye, Jose-Luis Padilla, Gordon B. Willis, and Amanda Wilmot. Advances in Questionnaire Design, Development, Evaluation and Testing. 2019, Wiley, ISBN: 978-1-119-26362-3, 816 pages
Jennifer Edgar 933

Book Review: Yuling Pan, Mandy Sha, and Hyunjoo Park. The Sociolinguistics of Survey Translation. 2020, New York: Routledge, ISBN 978-1-138-55087-2, 166 pages
Patricia Goerman 937

Book Review: Paul J. Lavrakes, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw, and Brady West. Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment. 2019, Wiley, ISBN: 978-1-119-08374-0, 544 pages
Katherine Jenny Thompson 941

Editorial Collaborators 945
Index to Volume 36, 2020 953

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. 1, March 2021

Building a Sample Frame of SMEs Using Patent, Search Engine, and Website Data
Sanjay K. Arora, Sarah Kelley and Sarvothaman Madhavan 1

Optimal Reconciliation of Seasonally Adjusted Disaggregates Taking Into Account the Difference Between Direct and Indirect Adjustment of the Aggregate
Francisco Corona, Victor M. Guerrero and Jesús López-Peréz 31

Panel Conditioning in the U.S. Consumer Expenditure Survey
Stephanie Eckman and Ruben Bach 53

Weighted Dirichlet Process Mixture Models to Accommodate Complex Sample Designs for Linear and Quantile Regression
Michael R. Elliott and Xi Xia 71

Identifying Outliers in Response Quality Assessment by Using Multivariate Control Charts Based on Kernel Density Estimation
Jiayun Jin and Geert Loosveldt 97

Can Smart City Data be Used to Create New Official Statistics?
Rob Kitchin and Samuel Stehle 121

An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges
Danielle McCool, Peter Lugtig, Ole Mussmann and Barry Schouten 149

Measuring and Modeling Food Losses
Marco Mingione, Carola Fabi and Giovanna Jona Lasinio 171

Survey Mode Effects on Objective and Subjective Questions: Evidence from the Labour Force Survey
Joachim Schork, Cesare A.F. Riillo and Johann Neumayr 213

Generalised Regression Estimation Given Imperfectly Matched Auxiliary Data
Li-Chun Zhang 239

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 48, No. 3, September/septembre 2020

Issue Information 339

Original Articles

Using ranked set sampling with binary outcomes in cluster randomized designs
Xinlei Wang, Mumu Wang, Johan Lim and Soohyun Ahn 342

A backward procedure for change-point detection with applications to copy number variation detection
Seung Jun Shin, Yichao Wu and Ning Hao 366

Empirical likelihood for nonlinear regression models with nonignorable missing responses
Zhihuang Yang and Niansheng Tang.....386

Robust multivariate change point analysis based on data depth
Shojaeddin Chenouri, Ahmad Mozaffari and Gregory Rice.....417

Post model-fitting exploration via a “Next-Door” analysis
Leying Guan and Robert Tibshirani.....447

A semiparametric stochastic mixed effects model for bivariate cyclic longitudinal data
Kexin Ji and Joel A. Dubin 471

Estimation of the additive hazards model with interval-censored data and missing covariates
Huiqiong Li, Han Zhang, Liang Zhu, Ni Li and Jianguo Sun.....499

Nonparametric change point detection for periodic time series
Lingzhe Guo and Reza Modarres.....518

Partial deconvolution estimation in nonparametric regression
Jianhong Shi, Xiuqin Bai and Weixing Song 535

On the role of local blockchain network features in cryptocurrency price formation
Asim K. Dey, Cuneyt G. Akcora, Yulia R. Gel and Murat Kantarcioglu.....561

Nonparametric beta kernel estimator for long and short memory time series
Taoufik Bouezmarni, Sébastien Bellegem and Yassir Rabhi 582

Optimal balanced block designs for correlated observations
Razieh Khodsiani and Saeid Pooladsaz 596

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 48, No. 4, December/décembre 2020

Issue Information 605

Original Articles

Estimation in the Cox cure model with covariates missing not at random, with application to disease screening/prediction
Lisha Guo, Yi Xiong and X. Joan Hu 608

Correlated and misclassified binary observations in complex surveys
Hon Yiu So, Mary E. Thompson and Changbao Wu.....633

Inference for misclassified multinomial data with covariates
Shijia Wang, Liangliang Wang and Tim B. Swartz.....655

Homogeneity testing under finite location-scale mixtures
Jiahua Chen, Pengfei Li and Guanfu Liu.....670

Copula-based predictions in small area estimation
Kanika Grover, Elif F. Acar and Mahmoud Torabi.....685

A sequential split-and-conquer approach for the analysis of big dependent data in computer experiments
Chengrui Li, Ying Hung and Minge Xie 712

A Bayesian mixture of experts approach to covariate misclassification
Michelle Xia, P. Richard Hahn and Paul Gustafson 731

Continuous threshold models with two-way interactions in survival analysis
Shuo Shuo Liu and Bingshu E. Chen.....751

A Gaussian alternative to using improper confidence intervals
André Plante 773

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.