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

Techniques d'enquête

Date de diffusion : 19 décembre 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

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

Comment accéder à ce produit

Le produit no 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

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 « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

This publication is also available in English.

Note de reconnaissance

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

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence ($p<0,05$)

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*.

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte 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 faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (rte@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca/Techniquesdenquete).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 40, numéro 2, décembre 2014

Table des matières

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, p. 149
Statistique Canada, n° 12-001-X au catalogue

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg.

Veuillez consulter la section avis à la fin de la revue pour des informations sur le processus de nomination et de sélection du prix Waksberg 2016.

Ce numéro de *Techniques d'enquête* commence par le quatorzième article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé de Steve Heeringa (président), Cynthia Clark, Louis-Paul Rivest et J.N.K. Rao, d'avoir choisi Constance Citro comme auteure de l'article du prix Waksberg de cette année.

Communication sollicitée pour le prix Waksberg 2014

Auteure : Constance F. Citro

Constance F. Citro est directrice du *Committee on National Statistics* (CNSTAT), poste qu'elle occupe depuis mai 2004. Auparavant, elle a été chef de cabinet par intérim (de décembre 2003 à avril 2004) et directrice principale d'études (de 1986 à 2003). Elle a commencé sa carrière au sein du CNSTAT en 1984 à titre de directrice d'études du groupe d'experts qui a rédigé *The Bicentennial Census: New Directions for Methodology* en 1990. Dr. Citro est titulaire d'un baccalauréat en sciences politiques de la *University of Rochester* et d'une maîtrise et d'un doctorat en sciences politiques de la *Yale University*. Avant d'arriver au CNSTAT, elle a occupé des postes de vice-présidente chez *Mathematica Policy Research Inc.* et chez *Data Use and Access Laboratories Inc*. Elle a été chargée de recherches de la *American Statistical Association* (ASA)/*National Science Foundation* (NSF)/*Census* en 1985-1986, en plus d'être membre de l'ASA et membre élue du *International Statistical Institute*. Pour le compte du CNSTAT, elle a dirigé les évaluations du recensement de 2000, du *Survey of Income and Program Participation*, des modèles de microsimulation pour des programmes de sécurité sociale, et du système de données sur la main-d'œuvre scientifique et technique de la NSF, en plus de diriger des études sur les conseils de révision institutionnelle et les recherches en sciences sociales, sur les estimations de la pauvreté pour de petites régions géographiques, et sur les données et méthodes utilisées pour la modélisation du revenu de retraite et sur une nouvelle approche de mesure de la pauvreté. Elle a coédité les éditions deux à cinq des *Principles and Practices for a Federal Statistical Agency*, et a contribué à des études sur la mesure de la discrimination raciale, sur l'élargissement de l'accès aux données de recherche, sur la convivialité des estimations du *American Community Survey*, sur le plan de recherche de la *National Children's Study*, sur le programme d'expériences et d'évaluations du recensement de 2010 du *Census Bureau*.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 151-181
Statistique Canada, n° 12-001-X au catalogue

Des modes multiples pour les enquêtes à des sources de données multiples pour les estimations

Constance F. Citro[1]

Résumé

Les utilisateurs et les fournisseurs de statistiques officielles, ainsi que ceux qui en assurent le financement, veulent des estimations « plus vastes, plus approfondies, plus rapides, de meilleure qualité et moins coûteuses » (selon Tim Holt, ancien chef de l'*Office for National Statistics* du Royaume-Uni), attributs auxquels j'ajouterais « plus pertinentes » et « moins fastidieuses ». Depuis la Deuxième Guerre mondiale, nous dépendons dans une large mesure des enquêtes sur échantillon probabiliste - celles-ci étant très bonnes dans les faits - pour atteindre ces objectifs pour les estimations dans de nombreux domaines, y compris le revenu des ménages et le chômage, l'état de santé autodéclaré, l'emploi du temps, les victimes d'actes criminels, l'activité des entreprises, les flux de produits, les dépenses des consommateurs et des entreprises, etc. Par suite des taux de plus en plus faibles de réponse totale et partielle et des preuves d'erreur de déclaration, nous avons réagi de nombreuses façons, y compris en utilisant des modes d'enquête multiples, des méthodes de pondération et d'imputation plus raffinées, l'échantillonnage adaptable, des essais cognitifs des questions d'enquête et d'autres méthodes pour maintenir la qualité des données. Dans le cas des statistiques sur le secteur des entreprises, afin de réduire le fardeau et les coûts, nous avons cessé depuis longtemps de recourir uniquement à des enquêtes pour produire les estimations nécessaires, mais jusqu'à présent, nous ne l'avons pas fait pour les enquêtes auprès des ménages, du moins pas aux États-Unis. Je soutiens que nous pouvons et que nous devons passer du paradigme de production des meilleures estimations possible à partir d'une enquête à la production des meilleures estimations possible pour répondre aux besoins des utilisateurs, à partir de sources de données multiples. Ces sources comprennent les dossiers administratifs et, de plus en plus, des données sur les transactions et des données en ligne. Je me sers de deux exemples - ceux du revenu des ménages et des installations de plomberie - pour illustrer ma thèse. Je propose des moyens d'inculquer une culture de la statistique officielle dont l'objectif est d'aboutir à des statistiques pertinentes, à jour, exactes et peu coûteuses, et qui traite les enquêtes, de même que les autres sources de données, comme des moyens d'atteindre cet objectif.

Mots clés : Enquêtes; dossier administratif; erreur totale; mégadonnées; revenu; logement.

1 Introduction

Tim Holt, ancien chef de l'*Office for National Statistics* du Royaume-Uni et ancien président de la *Royal Statistical Society*, a cerné cinq grands défis pour les statistiques officielles - à savoir qu'elles soient « plus vastes, plus approfondies, plus rapides, de meilleure qualité et moins coûteuses » (Holt 2007) - liste à laquelle j'ajouterais « moins fastidieuses » et « plus pertinentes ». Selon moi, pour relever comme il convient un ou plusieurs de ces défis, sans parler de tous les sept, les organismes statistiques officiels doivent passer du paradigme des enquêtes par échantillon probabiliste en vigueur depuis les 75 dernières années à un paradigme de sources de données multiples. Certains organismes ont procédé à ce changement pour la plupart de leurs programmes statistiques (voir, par exemple, Nelson et West 2014, au sujet de l'usage très étendu de statistiques fondées sur des données de registre au Danemark), et presque tous ont effectué ce changement pour certains de leurs programmes, mais il existe néanmoins des

1. Constance F. Citro, directrice, Committee on National Statistics, U.S. National Academy of Sciences/National Research Council. Courriel : ccitro@nas.edu.

programmes qui ne sont pas encore rendus très loin sur cette voie. Dans le cas des programmes de statistiques sur les ménages des États-Unis, il reste beaucoup à faire.

Cette transition ne devrait pas simplement élever une autre source de données au rang de panacée de la statistique officielle remplaçant l'enquête par échantillon probabiliste. Le recensement de la République allemande de 2011 - le premier réalisé dans ce pays depuis 1983 - nous rappelle justement les dangers d'une telle approche. Les résultats du recensement ont indiqué que les dossiers administratifs sur lesquels l'Allemagne avait fondé les chiffres de population officiels pendant plusieurs décennies surestimaient la population parce que les émigrants nés à l'étranger n'étaient pas enregistrés correctement (voir http://www.nytimes.com/2013/06/01/world/europe/census-shows-new-drop-in-germanys-population.html?_r=0 [November 2014]).

Ma thèse est que les programmes de statistiques officielles doivent prendre pour point de départ l'information dont ont besoin les utilisateurs pour l'élaboration des politiques, l'évaluation des programmes et la compréhension des tendances sociétales, et travailler à rebours des concepts vers les sources de données appropriées. Ces sources peuvent fort probablement inclure des enquêtes à échantillonnage probabiliste, mais aussi un ou plusieurs autres types de données. Ma thèse peut dans un certain sens passer pour un truisme, mais les personnes qui consacrent leur vie à perfectionner un outil particulier pour la collecte des données pourraient trop souvent considérer que cet outil est nécessaire en toute situation, plutôt que d'envisager le moyen le plus rentable d'obtenir les statistiques que souhaitent les décideurs, les chercheurs et d'autres utilisateurs des données.

Je ne doute pas un seul instant que Joe Waksberg, que j'ai eu l'honneur de connaître alors qu'il participait à un panel d'experts sur la méthodologie du recensement décennal du *Committee on National Statistics* (CNSTAT) au milieu des années 1980, approuverait mon sujet. Non seulement Joe était un être doué d'une bienveillance et d'un charme hors du commun, mais il possédait aussi une aptitude de premier plan à résoudre les problèmes et à innover. Joe insistait sur le fait qu'il importe d'examiner non seulement ce que l'on vous demande, mais aussi ce à quoi, selon vous, pense l'analyste (Morganstein et Marker 2000). Joe s'écartait invariablement des sentiers battus afin de cerner des sources de données et des modèles répondant aux besoins sous-jacents d'information au lieu de partir d'un concept a priori quant aux outils qu'il convenait d'utiliser.

Dans l'exposé qui suit, à la section 2, je passe brièvement en revue l'ascension et les avantages de l'échantillonnage probabiliste pour la statistique officielle aux États-Unis et, à la section 3, les menaces croissantes qui pèsent sur la pertinence, l'exactitude, l'actualité et la rentabilité des estimations fondées sur des enquêtes, ainsi que leur acceptation par le public. À la section 4 et à la section 5, j'examine les points forts et les points faibles des dossiers administratifs et d'autres sources de données non probabilistes qui pourraient être utiles, seules ou regroupées, pour la production de statistiques officielles. À la section 6, je décris de belles occasions pour les États-Unis de transformer les programmes d'enquêtes-ménages courants afin d'utiliser de multiples sources de données pour fournir de l'information d'une plus grande valeur. Je conclus à la section 7 en énumérant les obstacles à l'évolution vers le paradigme des sources de données multiples et propose des moyens de les aplanir.

Je me concentre sur ce que je connais le mieux, à savoir la statistique officielle aux États-Unis et les programmes de statistiques sur les ménages en particulier. D'autres programmes statistiques et d'autres organismes découvriront des analogies avec leurs propres travaux. Je critique le paradigme des enquêtes dans une perspective d'amélioration des statistiques officielles, tout en continuant d'apprécier grandement

la valeur des enquêtes à échantillonnage probabiliste, seules ou combinées à d'autres sources de données, et d'éprouver une profonde admiration pour le travail important des organismes statistiques voués à servir le bien public (voir National Research Council 2013c).

2 L'ascension de l'échantillonnage probabiliste en statistique officielle aux États-Unis

Il n'est pas exagéré de dire que les enquêtes par échantillonnage probabiliste à grande échelle ont été la réponse du 20e siècle au besoin de statistiques officielles plus vastes, plus approfondies, plus rapides, de meilleure qualité, moins coûteuses, plus pertinentes et moins fastidieuses. Ces enquêtes fournissaient des renseignements d'une précision connue contrairement aux enquêtes non probabilistes; en outre, ils fournissaient des renseignements détaillés, plus rapidement et à moindre coût, que les recensements. Duncan et Shelton (1978) et Harris-Kojetin (2012) passent en revue l'ascension de l'échantillonnage probabiliste dans les statistiques officielles aux États-Unis.

Il n'était pas évident durant les années 1930, époque à laquelle ont été élaborées la théorie et la pratique de l'échantillonnage probabiliste moderne aux États-Unis, que les enquêtes probabilistes seraient acceptées de manière aussi générale. L'arrivée de Jerzy Neyman au milieu des années 1930 a donné un élan aux travaux de W. Edwards Deming, Calvin Dedrick, Morris Hansen et leurs collègues au *Census Bureau* qui cherchaient à élaborer la théorie nécessaire pour l'échantillonnage de populations finies. Des enquêtes-échantillons à petite échelle, réalisées au cours des années 1930 par des universités et des organismes fédéraux sur des sujets comme les achats des consommateurs, le chômage, le logement urbain et la santé, ont fourni des preuves de concept et des conseils pratiques.

Les statisticiens innovateurs de l'administration fédérale avaient encore à surmonter les obstacles bureaucratiques jusqu'à la Maison-Blanche avant de pouvoir faire entrer l'échantillonnage dans la statistique fédérale officielle. Donc, les « vétérans » du *Census Bureau* étaient sceptiques quant à la possibilité d'utiliser des méthodes d'enquête pour obtenir des renseignements sur le chômage, tandis que les politiciens avaient des avis partagés quant à l'idée d'accepter des estimations (Anderson 1988). En 1937, une percée importante a eu lieu quand un échantillonnage de 2 % des ménages inclus dans les tournées postales non commerciales conçu par Dedrick, Hansen et d'autres a donné une estimation nettement plus élevée - et plus crédible - du nombre de chômeurs qu'un recensement « complet » de toutes les adresses résidentielles mené sur une base volontaire. S'inspirant de cet effort, de 1940 à 1942, la *Works Progress Administration* a produit le *Monthly Report on the Labor Force* fondé sur un échantillon, qui était le précurseur de la *Current Population Survey* (CPS). La CPS demeure aujourd'hui la source des estimations mensuelles officielles du chômage aux États-Unis produites par le *Census Bureau* et publiées par le *Bureau of Labor Statistics* (BLS).

Une autre percée a eu lieu quand le *Census Bureau*, qui essayait depuis des décennies de répondre aux demandes de questions supplémentaires dans le recensement décennal sans transformer le questionnaire en un cauchemar pour les répondants et les intervieweurs, a posé six questions à un échantillon de 5 % de la population au recensement de 1940. En raison du succès de l'échantillonnage, il a été décidé d'administrer les deux cinquièmes des questions du recensement de 1950 à un échantillon, et la même décision a été prise pour les recensements suivants. Le tableau 2.1 énumère certaines enquêtes-ménages,

certaines enquêtes-entreprises et certaines enquêtes par panel en cours aux États-Unis, ainsi que la date de leur lancement. La variété des sujets abordés et la longévité de ces enquêtes attestent de la dominance et de la valeur accordée au paradigme des enquêtes en statistique officielle aux États-Unis.

Tableau 2.1
Quelques enquêtes probabilistes menées par les organismes statistiques aux États-Unis, selon l'année de lancement

Décennie et année/type d'enquête	Enquête auprès des ménages transversale répétée	Enquête auprès des établissements commerciaux transversale répétée	Enquête par panel de particuliers
1940	1940 - Current Population Survey (CPS) 1947 - CPS Annual Social and Economic Supplement (CPS/ASEC)	1946 - Monthly Wholesale Trade Survey	
1950	1950 - Consumer Expenditure Survey (CE) 1955 - National Survey of Fishing, Hunting, and Wildlife-Associated Recreation 1957 - National Health Interview Survey (NHIS)	1953 - Advance Monthly Retail Sales Survey 1953 - Business R&D and Innovation Survey (BRDIS) 1959 - Building Permits Survey	
1960	1960 - Decennial Census Long-Form Sample (devenu l'American Community Survey en 2005)	1965 - National Hospital Care Survey	1966-1990 - National Longitudinal Survey of Older Men
1970	1972 - National Crime Victimization Survey (NCVS) 1973 - American Housing Survey (AHS); 1973 - National Survey of College Graduates (NSCG) 1979 - Residential Energy Consumption Survey (RECS)	1975 - Farm Costs and Returns Survey and Cropping Practices and Chemical Use Surveys (combiné à l'Agricultural Resource Management Survey en 1996) 1979 - Commercial Buildings Energy Consumption Survey (CBECS)	1972-1986 - National Longitudinal Survey of High School Class of 72 Depuis 1973 - Survey of Doctorate Recipients (SDR) Depuis 1979 - National Longitudinal Survey of Youth (NLSY79)
1980	1983 - Survey of Consumer Finances (SCF)	1985 - Manufacturing Energy Consumption Survey (MECS)	Depuis 1984 - Survey of Income and Program Participation (SIPP)
1990	1991 - Medicare Current Beneficiary Survey (MCBS)	1996 - Agricultural Resource Management Survey (ARMS)	Depuis 1997 - National Longitudinal Survey of Youth (NLSY97)
2000	2005 - American Community Survey (ACS)		2001-2008 - Early Childhood Longitudinal Study (Birth Cohort)

Notes : Le nom courant de l'enquête est utilisé; la périodicité de l'interview pour les enquêtes transversales et par panel répétées varie; certaines enquêtes transversales répétées ont une composante de panel (groupes de renouvellement); la durée des enquêtes par panel (nombre d'années durant lesquelles les répondants sont dans l'échantillon) varie.

Source : Liste compilée par l'auteur.

3 Des défauts dans la cuirasse : menaces croissantes pesant sur le paradigme des enquêtes

Les enquêtes par échantillonnage probabiliste sont indispensables aux organismes statistiques officiels et autres pour de nombreux types de mesures : par exemple, pour suivre des phénomènes tels que l'approbation du public à l'égard du président des États-Unis ou les sentiments exprimés de bien-être. En outre, les enquêtes probabilistes visant principalement à produire des mesures de concept, comme le revenu du ménage, qui pourraient être obtenues à partir d'autres sources, offrent deux grands avantages : 1) elles permettent d'obtenir des données sur une grande variété de covariables pouvant être utilisées dans l'analyse de la ou des variables principales d'intérêt et 2) elles sont sous le contrôle de leur concepteur. Pourtant, les menaces qui pèsent sur le paradigme des enquêtes par échantillonnage probabiliste font boule de neige d'une façon qui ne présage rien de bon pour l'avenir. Manski (2014) va jusqu'à accuser les organismes statistiques d'enfouir sous le tapis les principaux problèmes liés à leurs données et de sous-estimer nettement l'incertitude présente dans leurs estimations. Il considère la non-réponse aux enquêtes comme un exemple d'« incertitude permanente ».

3.1 Caractérisation de la qualité des enquêtes

Une classification des erreurs et des autres problèmes qui peuvent compromettre la qualité des estimations d'enquête est essentielle à la compréhension et à l'amélioration des statistiques officielles. Brackstone (1999) a écrit un article majeur concernant le développement des cadres de qualité des données. Plus récemment, Biemer, Trewin, Bergdahl et Lilli (2014) ont passé en revue la littérature sur les cadres systématiques de la qualité, en soulignant, en particulier, les six dimensions proposées par Eurostat (2000), à savoir la pertinence, l'exactitude, l'actualité et la ponctualité, l'accessibilité et la clarté, la comparabilité (temporelle et géographique) et la cohérence (normes cohérentes). Iwig, Berning, Marck et Prell (2013) ont examiné les cadres de la qualité établis par Eurostat, l'*Australian Bureau of Statistics*, l'*Office for National Statistics* du Royaume-Uni, Statistique Canada et d'autres organismes, et élaboré des questions fondées sur six dimensions de la qualité de leur cru - la pertinence, l'accessibilité, la cohérence, l'intelligibilité, l'exactitude et l'environnement institutionnel - destinées à être utilisées par les organismes statistiques américains pour évaluer l'utilité des dossiers administratifs. Daas, Ossen, Tennekes et Nordholt (2012) ont construit un cadre d'évaluation de l'utilisation de dossiers administratifs pour produire des données de recensement pour les Pays-Bas.

Biemer et coll. (2014) sont allés plus loin et ont utilisé le cadre d'Eurostat (en combinant la comparabilité et la cohérence en une seule dimension) comme fondement pour concevoir, tester et mettre en œuvre un système de cotes numériques pour évaluer et améliorer continuellement la qualité des produits de données de Statistics Sweden. Pour que l'évaluation soit complète, elle devrait aussi porter sur les dimensions de la qualité en regard du coût et du fardeau de réponse. Utilement en ce qui concerne mes objectifs, Biemer et coll. ont décomposé la dimension d'« exactitude », conçue comme étant l'erreur totale d'enquête (ou l'erreur totale de produit pour les programmes statistiques non fondés sur des enquêtes, comme les comptes nationaux), en une erreur d'échantillonnage et sept types d'erreurs non dues à l'échantillonnage, à savoir 1) l'erreur de base de sondage, y compris le sous-dénombrement et le surdénombrement, ainsi que les variables auxiliaires manquantes ou erronées dans la base de sondage; 2) l'erreur due à la non-réponse (totale et partielle); 3) l'erreur de mesure (surdéclaration, sous-

déclaration, autre); 4) l'erreur de traitement des données; 5) l'erreur de modélisation/estimation, telle que celle découlant de l'ajustement de modèles pour l'imputation ou de l'ajustement des valeurs des données afin qu'elles concordent avec les valeurs de référence; 6) l'erreur de révision (la différence entre les estimations publiées provisoires et définitives); et 7) l'erreur de spécification (la différence entre la variable réelle non observable et la variable indicatrice observée). Pour les enquêtes permanentes, j'ajouterais l'*erreur de concept dépassé*, qui est apparentée à l'erreur de spécification mais différente de celle-ci. Par exemple, le concept de revenu monétaire ordinaire du *Census Bureau* pour le calcul des estimations officielles du revenu des ménages et de la pauvreté d'après *l'Annual Social et Economic Supplement* (ASEC) de la CPS est devenu progressivement dépassé en raison de l'évolution des programmes d'imposition et de transferts des États-Unis (voir, par exemple, Czajka et Denmead 2012; National Research Council 1995).

3.2 Quatre sources d'erreur dans les statistiques américaines sur les ménages

3.2.1 Déficiences des bases de sondage

Obtenir une base de sondage complète et exacte pour les enquêtes peut être aussi difficile qu'obtenir des réponses auprès des unités sélectionnées dans l'échantillon à partir de la base de sondage et, dans de nombreux cas, la difficulté a persisté, voire même augmenté, au fil du temps. Joe Waksberg serait d'accord sur le problème des déficiences des bases de sondage : non seulement il a élaboré, en collaboration avec Warren Mitofsky, la méthode de composition aléatoire (CA) pour créer des bases de sondage et des échantillons pour réaliser des enquêtes téléphoniques résidentielles de haute qualité durant les années 1970 (voir Waksberg 1978; Tourangeau 2004), mais il a aussi assisté aux premiers signes de déclin de la popularité de la méthode en raison de phénomènes tels que l'existence de ménages ne possédant qu'un téléphone mobile.

L'une des bases de sondage utilisées fréquemment pour réaliser les enquêtes-ménages aux États-Unis est le Fichier maître des adresses (FMA) du *Census Bureau* élaboré pour le recensement décennal. Lors des quelques derniers recensements, la couverture nette des adresses résidentielles dans le FMA n'a cessé de s'améliorer, particulièrement pour les logements occupés (Mule et Konicki 2012). Le problème que continuent de poser les enquêtes-ménages est celui du sous-dénombrement des membres individuels dans les logements échantillonnés. Les ratios de couverture (c.-à-d. les estimations avant ajustement des ratios sur les chiffres de population de contrôle) dans le cas de la CPS de mars 2013, par exemple, ne sont que de 85 % pour l'ensemble de la population, et il existe des écarts prononcés entre les hommes et les femmes, les jeunes et les personnes âgées, ainsi que les blancs et les groupes minoritaires, les ratios de couverture étant aussi faibles que 61 % pour les hommes et les femmes de race noire âgés de 20 à 24 ans (voir http://www.census.gov/prod/techdoc/cps/cpsmar13.pdf [November 2014]). Aucune étude systématique de la série chronologique de ratios de couverture pour les enquêtes-ménages américaines n'a été réalisée, mais il existe des preuves que les ratios se sont dégradés.

Il est certes utile de corriger les erreurs de couverture pour tenir compte de l'âge, du sexe, de la race et du groupe ethnique, mais les ajustements des ratios effectués à l'heure actuelle pour les enquêtes-ménages ne fournissent indubitablement pas de correction pour d'autres écarts de couverture conséquents. (Les chiffres de contrôle pour l'ajustement des ratios, dans le cadre de l'un des usages les moins controversés et

les plus anciens des dossiers administratifs dans les enquêtes-ménages des États-Unis, sont tirés des estimations démographiques produites d'après les données du recensement précédent et mises à jour au moyen de dossiers administratifs et de données d'enquête.) Donc, tout ce que l'on sait au sujet du sous-dénombrement au recensement décennal des États-Unis indique que, si l'on maintient constantes la race et l'origine ethnique, les populations désavantagées sur le plan socioéconomique sont moins bien dénombrées que les autres (voir, par exemple, National Research Council 2004, annexe D). Il est peu probable que de meilleurs résultats soient obtenus dans le cas des enquêtes-ménages - par exemple, Czajka, Jacobson et Cody (2004) constatent que la *Survey of Income and Program Participation* (SIPP) sous-représente considérablement les familles à revenu élevé comparativement à la *Survey of Consumer Finances* (SCF), qui comprend un échantillon de ménages à revenu élevé tiré d'une liste basée sur les dossiers fiscaux. En tenant compte des différences de couverture socioéconomique, Shapiro et Kostanich (1988) estiment au moyen de simulations que les estimations de la pauvreté présentent un important biais à la baisse pour les hommes noirs dans la CPS/ASEC. Par ailleurs, comparativement à l'échantillon ayant reçu le questionnaire complet du Recensement de 2000, Heckman et LaFontaine (2010) constatent que le sous-dénombrement au supplément d'octobre sur les études de la CPS de 2000 contribue peu à la sous-estimation des taux d'achèvement des études secondaires; d'autres facteurs sont plus importants.

3.2.2 Tendance à la baisse de la réponse totale

Un groupe d'étude du National Research Council des États-Unis (2013b) vient d'achever un examen complet des causes et conséquences de la non-réponse totale aux enquêtes-ménages, qui confirme le phénomène bien connu voulant que le public soit de moins en moins disponible et disposé à répondre aux enquêtes, même celles menées par les organismes statistiques officiels jugés fiables. Aux États-Unis, déjà durant les années 1980, il existait des preuves que les taux de réponse ont été à la baisse depuis pratiquement le début de l'usage répandu des enquêtes par échantillonnage probabiliste (voir, par exemple, Steeh 1981; Bradburn 1992). De Leeuw et De Heer (2002) ont estimé un taux séculaire de diminution de la participation aux enquêtes de 3 points de pourcentage par année en examinant les enquêtes permanentes menées dans 16 pays occidentaux du milieu des années 1980 à la fin des années 1990. Le taux de participation mesure la réponse des cas échantillonnés admissibles effectivement contactés; les taux de réponse (il existe plusieurs variantes acceptées) possèdent des dénominateurs plus généraux, comprenant les cas admissibles qui n'ont pas été rejoints (National Research Council 2013c, p. 9-12). Le National Research Council (2013b, tableaux 1 et 2, p. 104) fournit les taux de réponse initiaux ou à la présélection pour une gamme d'enquêtes américaines officielles pour 1990-1991 (alors que les taux de réponse avaient déjà diminué considérablement pour de nombreuses enquêtes) et pour 2007-2009 montrant clairement que le problème ne disparaît pas.

On a longtemps supposé que des taux de réponse plus faibles, même avec repondération pour tenir compte de la non-réponse, entraînent inévitablement un biais dans les estimations d'enquête. Selon des travaux de recherche récents (voir, par exemple, Groves et Peytcheva 2008), la relation entre la non-réponse et le biais est complexe. Lorsqu'on prend des mesures extraordinaires pour accroître le taux de réponse, il est possible qu'on augmente aussi le biais, par inadvertance, si l'on obtient une réponse plus importante auprès de certains groupes seulement et non d'autres (voir, par exemple, Fricker et Tourangeau 2010). Toutefois, il serait imprudent de la part des organismes officiels de statistique de supposer que l'accroissement de la non-réponse n'a que peu d'effet, voire aucun, sur l'exactitude des estimations,

particulièrement si la non-réponse totale est couplée à la non-réponse partielle. Par exemple, on estime que les non-répondants aux enquêtes sur la santé sont en moins bonne santé, en moyenne, que les répondants, et que les non-répondants aux enquêtes sur le bénévolat sont moins susceptibles de faire du bénévolat que les répondants (National Research Council 2013b, p. 44-45). De surcroît, les études des effets de la non-réponse sur les associations bivariées ou multivariées ou sur la variance sont peu nombreuses, sauf en ce qui concerne le fait évident - et non sans importance - que la non-réponse totale réduit la taille effective de l'échantillon.

3.2.3 Réponse partielle souvent faible et à la baisse

Ni les enquêtes ni les recensements ne peuvent s'attendre à obtenir que les répondants fournissent une réponse à chacune des questions. Dans le cas du recensement des États-Unis, la vérification de certaines questions pour s'assurer de la cohérence est une pratique de longue date, mais jusqu'au milieu du 20e siècle, aucun ajustement n'était effectué pour la non-réponse partielle - les tableaux contenaient des lignes intitulées « pas de réponse » ou un énoncé similaire. Le premier recours à l'imputation a eu lieu en 1940 quand Deming a élaboré une méthode « cold deck » pour imputer l'âge en sélectionnant aléatoirement une valeur d'âge dans un ensemble approprié de cartes sélectionnées en fonction des autres renseignements connus au sujet de la personne dont l'âge manquait. À partir de 1960, grâce à l'émergence des ordinateurs à haute vitesse, des méthodes d'imputation « hot deck » ont été utilisées pour imputer les valeurs manquantes pour de nombreuses questions du recensement (Citro 2012). La méthode hot deck consiste à utiliser la valeur la plus récente enregistrée dans une matrice pour la personne ou le ménage traité précédemment et, par conséquent, ne requiert pas l'hypothèse que les données manquent entièrement au hasard (MCAR pour *missing completely at random*), bien qu'il soit nécessaire de supposer que les données manquent au hasard (MAR pour *missing at random*) dans les catégories définies par les variables dans la matrice hot deck. Des méthodes d'imputation fondées sur un modèle ne nécessitant pas d'aussi fortes hypothèses que celles de type MAR ou MCAR ont été élaborées (voir National Research Council 2010b), mais leur usage n'est pas très répandu dans les enquêtes-ménages aux États-Unis. Font exception la *Survey of Consumer Finance*s (SCF) (Kennickell 2011) et la *Consumer Expenditure (CE) Interview Survey* (Passero 2009).

Quelle que soit la méthode, l'imputation a l'avantage de créer un enregistrement de données complet pour chaque répondant, ce qui facilite l'analyse multivariée et réduit la probabilité que les chercheurs utilisent différentes méthodes de traitement des données manquantes donnant des résultats différents. Cependant, l'imputation peut introduire un biais dans les estimations, et la mesure dans laquelle des données manquent accentuera vraisemblablement l'importance de tout biais. Par conséquent, il est troublant de constater que la non-réponse a augmenté pour des questions importantes des enquêtes-ménages, comme celles sur le revenu, les actifs, les impôts et les dépenses de consommation, qui obligent les répondants à fournir des montants en dollars - par exemple, Czajka (2009, tableau A-8) compare les taux d'imputation d'une question pour le revenu total et pour plusieurs sources de revenus dans le cas de la CPS/ASEC et la SIPP pour 1993, 1997 et 2002 - un bon tiers des données sur le revenu sont imputées à l'heure actuelle dans le cas de la CPS/ASEC, en hausse par rapport à environ le quart en 1993 - et la situation n'est pas meilleure pour la SIPP. Clairement, étant donné des taux d'imputation aussi élevés, il est impératif de procéder à une évaluation minutieuse des effets des méthodes d'imputation. Hoyakem, Bollinger et Ziliak (2014), par exemple, estiment que la méthode d'imputation hot deck pour les revenus

dans la CPS/ASEC a systématiquement entraîné une sous-estimation de un point de pourcentage, en moyenne, de la pauvreté en se basant sur l'évaluation des revenus manquants dans les enregistrements des revenus de la CPS/ASEC et de la sécurité sociale.

3.2.4 L'erreur de mesure pose problème et n'est pas bien étudiée

Même en cas de déclaration complète, ou, plus fréquemment, d'ajustements pour tenir compte de la non-réponse totale et partielle, les estimations d'après les données d'enquête contiendront encore une erreur découlant des déclarations inexactes faites par les répondants qui devinent la réponse, évitent délibérément de donner une réponse correcte ou ne comprennent pas l'intention de la question. Même si les organismes statistiques reconnaissent l'existence de l'erreur de mesure, la portée de celle-ci est habituellement moins bien étudiée que celle de l'erreur d'échantillonnage ou des données manquantes. De nombreuses études de l'erreur de mesure comparent les estimations agrégées provenant d'une enquête à des estimations similaires provenant d'une autre enquête ou à un ensemble approprié de dossiers administratifs, ajustés autant qu'il est possible pour qu'ils soient comparables. Il est impossible de dégager de ces études le rôle joué par l'erreur de mesure comparativement à d'autres facteurs, mais les résultats indiquent l'ordre de grandeur des problèmes. Les auteurs de certaines études arrivent à apparier des enregistrements individuels et par conséquent à examiner les composantes de l'erreur de mesure.

Il est connu qu'une erreur de mesure importante affecte les estimations socioéconomiques clés produites d'après les enquêtes-ménages américaines. Donc, une foule d'études ont donné des preuves, enquête après enquête, d'une sous-estimation nette du revenu des ménages américains et, constatation encore plus troublante, d'une diminution de la complétude des déclarations, même après imputation et pondération. Ainsi, Fixler et Johnson (2012, tableau 2) ont estimé qu'entre 1999 et 2010, les estimations moyennes et médianes calculées d'après la CPS/ASEC sont devenues progressivement inférieures aux estimations des *National Income and Product Accounts* (NIPA) en raison de facteurs tels que 1) la sous-représentation des ménages à revenu très élevé dans l'échantillon de la CPS/ASEC, 2) la non-déclaration ou la sous-déclaration par les ménages à revenu élevé qui sont inclus dans l'échantillon et 3) la non-déclaration ou la sous-déclaration par les ménages à revenu moyen ou faible. Les études portant sur les sources individuelles de revenu révèlent une erreur encore pire. Par exemple, Meyer et Goerge (2011) constatent, en appariant les enregistrements du *Supplemental Nutrition Assistance Program* (SNAP) obtenus dans deux États, que près de 35 % et 50 %, respectivement, de véritables bénéficiaires ne déclarent pas avoir reçu des prestations dans le cadre de l'*American Community Survey* (ACS) ou de la CPS/ASEC. De même, Meyer, Mok et Sullivan (2009) fournissent des preuves d'écarts importants et souvent croissants entre les estimations d'enquête et les estimations fondées sur les dossiers administratifs correctement ajustés des bénéficiaires du revenu et des montants totaux pour de nombreuses sources.

La richesse est, comme on le sait, difficile à mesurer dans les enquêtes-ménages, et de nombreux organismes n'essaient pas de le faire. Czajka (2009, p. 143-145) résume les travaux de recherche sur la qualité des estimations de la richesse d'après la SIPP en les comparant aux estimations d'après la SCF et la *Panel Study of Income Dynamics* (PSID). En simplifiant considérablement les résultats, historiquement, la SIPP s'est avérée assez efficace pour mesurer les éléments de passif, comme la dette hypothécaire, et la valeur d'éléments d'actif possédés tels que les logements, les véhicules et les obligations d'épargne. Par contre, la SIPP n'a pas fourni de bonnes mesures de la valeur des actifs détenus principalement par les ménages à revenu élevé, comme les actions, les fonds communs de placement, ainsi que les comptes IRA

et KEOGH, tandis que la PSID a donné d'un peu meilleurs résultats. Sur une base nette, la SIPP sous-estime considérablement la valeur nette.

Une étude menée par le National Research Council (2013a) sur la *CE Interview and Diary Surveys* du BLS sur les dépenses de consommation comportant une interview et la tenue d'un journal a révélé des différences de qualité de la déclaration de divers types de dépenses comparativement aux estimations des dépenses de consommation personnelles (PCE pour *personal consumption expenditure*) ajustées de manière appropriée provenant des NIPA. Bee, Meyer et Sullivan (2012, tableau 2) ont également constaté une diminution de la déclaration de certaines dépenses - par exemple, la déclaration des dépenses en essence dans l'estimation des dépenses de consommation des ménages est passée de plus de 100 % de l'estimation des PCE comparables en 1986 à un peu moins de 80 % en 2010, tandis que la déclaration des dépenses en meubles et accessoires d'ameublement est passée de 77 % à 44 % au cours d'une période comparable.

4 Que peut-on faire?

Les spécialistes de la recherche sur les enquêtes ne sont pas restés inactifs face aux menaces multiples et croissantes qui pèsent sur le paradigme des enquêtes. Durant au moins les 15 dernières années, ils ont cherché activement des moyens de réduire ou de compenser l'erreur de couverture, la non-réponse totale et partielle, l'erreur de mesure et, plus récemment, le fardeau de réponse. Les stratégies adoptées comprenaient 1) consacrer plus d'argent à l'achèvement des cas (mais les contraintes budgétaires limitent la viabilité de cette stratégie), 2) utiliser les paradonnées et l'information auxiliaire afin de déterminer et de corriger plus efficacement le biais dû à la non-réponse totale, 3) employer des ajustements plus perfectionnés pour tenir compte des données manquantes qui ne reposent pas sur l'hypothèse qu'elles sont de type MAR, 4) utiliser des méthodes reposant sur un plan de collecte adaptatif afin d'optimiser le coût et la qualité des réponses, 5) utiliser de multiples bases de sondage pour réduire l'erreur de couverture (p. ex. listes de numéros de téléphone mobile et de numéros de téléphone fixe pour les enquêtes téléphoniques), 6) utiliser de multiples modes de collecte pour que la réponse soit plus efficace par rapport au coût, comme dans l'ACS, qui a ajouté récemment une option de réponse en ligne aux options d'envoi par la poste, d'ITAO et d'IPAO, 7) réduire le fardeau de réponse en optimisant les nombres d'appels et de visites de suivi, et 8) décrire les besoins de données d'enquête. Aux États-Unis, on fait souvent appel aux utilisateurs des données pour plaider la cause devant le Congrès et d'autres parties prenantes. Par exemple, l'*Association of Public Data Users*, le *Council of Professional Associations on Federal Statistics* et la *Population Association of America* mobilisent fréquemment les utilisateurs des données au nom des programmes des organismes statistiques.

Selon moi, quoique louables et nécessaires, ces étapes ne suffisent pas à restaurer le paradigme fondé sur l'enquête par échantillonnage probabiliste pour la production de statistiques officielles sur les ménages et d'autres types de répondants. Je propose plutôt que les organismes statistiques commencent systématiquement par cerner les besoins des décideurs et du public et qu'ils travaillent à rebours afin de déterminer quelles sont les sources de données appropriées pour répondre aux besoins de la façon la plus rentable et la moins lourde possible. Ce paradigme des sources de données multiples devrait s'appliquer à

tous les programmes statistiques qui sont habituellement fondés sur des enquêtes, des dossiers administratifs ou d'autres sources.

Certains programmes statistiques importants, comme les NIPA et l'Indice des prix à la consommation (voir Horrigan 2013) aux États-Unis et dans d'autres pays, utilisent des sources de données multiples depuis des décennies. L'une des raisons est que ces programmes s'appuient sur un cadre conceptuel généralement accepté qui détermine les éléments requis pour constituer un ensemble satisfaisant d'estimations. Il n'est pas acceptable d'omettre une ou plusieurs composantes du revenu des NIPA simplement parce que les données ne peuvent pas être obtenues à partir d'une source unique. En outre, comme les estimations clés des NIPA sont révisées périodiquement afin d'ajouter des données, d'améliorer la méthodologie et de peaufiner les concepts, il existe un biais positif intégré en faveur de la recherche de sources de données nouvelles et améliorées pour combler les lacunes et accroître l'exactitude. Les recensements économiques menés aux États-Unis s'appuient aussi sur des sources multiples, plus précisément les données de l'impôt sur le revenu pour les entreprises individuelles et les très petits employeurs, ainsi que des données d'enquête pour les plus grandes entreprises. En revanche, les programmes de statistiques sur les ménages des États-Unis ont adhéré plus étroitement au paradigme des enquêtes par échantillonnage probabiliste. De surcroît, comme les intervalles sont habituellement longs entre les révisions des concepts et du plan des enquêtes-ménages, les enquêtes perdent trop souvent du terrain en ce qui concerne leur capacité à servir les décideurs et le public, alors que l'utilisation de sources de données additionnelles permettrait d'importantes améliorations.

5 Quelles sources de données utiliser pour soutenir les enquêtes?

Pendant des décennies après l'introduction de l'échantillonnage probabiliste en statistique officielle, la seule autre source de données était les dossiers administratifs - provenant de divers paliers de gouvernement, selon la structure gouvernementale du pays (fédéral, État et local aux États-Unis), et de diverses entités non gouvernementales (p. ex. dossiers de paye des employeurs ou dossiers d'admission des hôpitaux). Un certain nombre d'organismes statistiques nationaux dans le monde ont commencé à intégrer des dossiers administratifs dans leurs programmes - cette intégration allant de leur utilisation accessoire au transfert, sans distinction aucune, des enquêtes et des recensements à un paradigme axé sur les dossiers administratifs.

Grâce aux innovations technologiques des années 1970 et des années 1980, certaines sources de données supplémentaires, comme les enregistrements des dépenses aux caisses (rendus possibles par le développement des codes à barres et des scanneurs), et les images aériennes et par satellite pour catégoriser l'utilisation des terres, sont devenues disponibles, du moins potentiellement, pour la production de statistiques officielles. Cependant, l'univers des sources de données demeurait relativement limité. À partir des années 1990, l'avènement d'Internet et de la technologie de l'informatique à haute-vitesse a donné le jour à un extraordinaire éventail de nouvelles sources de données, dont les données envoyées par les caméras de circulation, la localisation des téléphones mobiles, les termes de recherche utilisés sur le Web et les affichages sur les sites des médias sociaux. Le défi pour les organismes statistiques consiste à classer et à évaluer toutes ces sources de données d'une manière qui les aide à en déterminer l'utilité.

5.1 Le concept des « mégadonnées » est-il utile?

Bon nombre de nouvelles catégories de données devenues disponibles au cours des quelque 15 dernières années sont souvent de très grande taille, ce qui a donné naissance au terme de « mégadonnées ». Je soutiens que ce terme à la mode n'aide que fort peu, voire nullement, les organismes statistiques à déterminer quelles sont les combinaisons de données convenant pour leurs programmes. En sciences informatiques, « les mégadonnées sont des fonds d'information à grand volume, grande vélocité et/ou grande variété qui nécessitent de nouvelles formes de traitement pour permettre la prise de meilleures décisions, la découverte d'idées et l'optimisation des processus » [*Traduction*] (Laney 2001). Ces propriétés ne sont pas inhérentes à un type particulier de données ou à une plateforme particulière, telle qu'Internet. Ce qui peut être considéré comme des « mégadonnées » est plutôt une cible en évolution à mesure que l'informatique à haute vitesse et les techniques d'analyse des données progressent. Dans l'environnement informatique actuel, les données de recensement, d'enquête et de dossiers administratifs peuvent rarement être qualifiées de « mégadonnées », même si elles auraient pu l'être à une époque antérieure. Aujourd'hui, les gens ont tendance à considérer comme étant des « mégadonnées » les flux de données provenant de caméras, de détecteurs et d'interactions en grande partie libres avec Internet, comme les messages sur les médias sociaux. À l'avenir, bon nombre de ces types de données pourraient ne plus rentrer dans cette catégorie. De plus, en ce qui concerne Internet, celui-ci génère non seulement une grande quantité de « mégadonnées » contemporaines, mais il facilite aussi l'accès à des données de volume plus habituel - par exemple, accès aux sondages d'opinion ou aux registres fonciers locaux.

À mon avis, les organismes statistiques souhaiteront le plus souvent, et devraient, être des « adeptes suivant de près les leaders » plutôt que des leaders de l'utilisation des mégadonnées. Il me paraît plus approprié que le milieu universitaire et le secteur privé soient les premiers à s'attaquer à l'utilisation de données aussi volumineuses et d'une telle vélocité et variété qu'elles nécessitent de grands pas en avant dans l'élaboration de nouvelles formes de traitement et d'analyse. Les organismes statistiques devraient se tenir au courant des avancées dans le domaine des mégadonnées qui pourraient être prometteuses pour leurs programmes et ils seraient bien avisés d'appuyer la recherche dans ces domaines pour s'assurer que les applications pertinentes pour leurs programmes voient le jour. Toutefois, je pense que les ressources des organismes statistiques devraient être consacrées principalement à l'utilisation de sources de données qui offrent des avantages dont l'utilité est plus immédiate.

Groves (2011) a tenté de passer à une classification plus pertinente pour les organismes statistiques que celle comprenant les « mégadonnées », d'une part, et toutes les autres données, d'autre part, en faisant la distinction entre ce qu'il appelle les « données conçues » qui sont « produites pour découvrir ce qui n'est pas mesuré » et les « données organiques » qui sont « produites secondairement aux processus, pour enregistrer le processus ». Keller, Koonin et Shipp (2012) énumèrent des exemples de sources de données sous les deux en-têtes de Groves. Leur liste de données conçues comprend les données administratives (p. ex. dossiers fiscaux), les enquêtes fédérales, les recensements de la population et les « autres données recueillies pour répondre à des questions stratégiques particulières ». Leur liste de données organiques comprend les données de localisation (« données externes » de téléphones mobiles, de transpondeurs pour postes de péage, de caméras de surveillance), les préférences politiques (dossiers d'enregistrement des électeurs, votes aux élections primaires, contributions aux partis politiques), les renseignements commerciaux (transactions sur carte de crédit, ventes de propriété, recherches en ligne, identification de radiofréquences), les renseignements sur la santé (dossiers médicaux électroniques, admissions à l'hôpital,

appareils pour surveiller les signes vitaux, ventes des pharmacies), et autres données organiques (imagerie optique, infrarouge et spectrale, mesures météorologiques, mesures sismiques et acoustiques, rayonnements ionisants biologiques et chimiques). Sans omettre, sous chaque catégorie, des données telles que les messages affichés sur Facebook ou Twitter, bien qu'ils puissent se retrouver sous la rubrique plus générale des « recherches en ligne ».

La question est de savoir si la classification en deux catégories de Keller et coll. (2012) est plus utile que celle de « mégadonnées » pour les besoins des organismes statistiques. Par exemple, classer les dossiers d'inscription des électeurs ou les dossiers de santé électroniques comme des données organiques plutôt que comme des données administratives conçues semble ne pas tenir compte des façons dont elles diffèrent de sources telles que les recherches en ligne et des façons dont elles sont similaires aux dossiers administratifs de l'administration fédérale et des États. En outre, même les données organiques sont « conçues », si ce n'est que de manière minimale, en ce sens que le fournisseur a spécifié certains paramètres, tels que les 140 caractères pour un message sur Twitter ou un angle de vision particulier pour une caméra de circulation. Néanmoins, la distinction entre données conçues et données organiques met en relief une dimension utile, qui est le degré auquel les organismes statistiques ont déjà accès à une source de données, contrôlent les changements apportés à une source de données et sont capables de comprendre facilement les propriétés d'une source de données.

5.2 Dimensions des sources de données : illustrations pour quatre grandes catégories

Établir une nomenclature et des critères d'évaluation satisfaisants qui peuvent aider les organismes statistiques à évaluer l'utilité éventuelle de diverses sources de données pour leurs programmes, dans le but de comprendre aussi bien les propriétés d'erreur des sources de données de rechange qu'ils ne comprennent l'erreur totale dans le cas des enquêtes, demandera un effort considérable de la part des organismes statistiques du monde entier (Iwig et coll. 2013 et Daas et coll. 2012, sont des exemples de tels efforts). Je ne prétends pas pouvoir m'approcher de ce but dans le présent article. Mon objectif est plus modeste - à savoir donner certaines illustrations afin que ceux et celles qui sont des inconditionnels du paradigme des enquêtes par échantillonnage probabiliste (ou du paradigme des dossiers administratifs) puissent voir que la tâche de comprendre d'autres sources de données est à la fois faisable et souhaitable. Je fournis des illustrations pour quatre sources de données variant du classique à l'avant-garde :

(1) Enquêtes et recensements, ou un ensemble de données tirées des réponses de particuliers qui sont interrogés sur un ou plusieurs sujets selon le plan établi par l'enquêteur (organisme statistique, autre organisme gouvernemental ou organisme universitaire ou privé d'enquête) conformément aux principes de la recherche par enquête dans le but de produire des données généralisables pour une population définie.

(2) Dossiers administratifs ou un ensemble de données obtenues au moyen de formulaires conçus par un organisme administratif conformément à une loi, un règlement ou une politique pour exploiter un programme, comme le versement de prestations à des bénéficiaires admissibles ou pour le versement de salaires. Les dossiers administratifs sont habituellement permanents et peuvent être gérés par des organismes gouvernementaux ou des organisations non gouvernementales.

(3) Dossiers de transactions commerciales, ou un ensemble de données obtenues par saisie électronique d'achats (p. ex. épicerie, biens immobiliers) effectués par un acheteur, mais sous une forme déterminée par un vendeur (p. ex. renseignements sur les produits et prix sous forme de codes à barres enregistrés par les scanneurs des caisses, enregistrements de renseignements sur les produits et les prix provenant des ventes en ligne, comme par l'intermédiaire d'Amazon).

(4) Interactions des particuliers avec le Web en utilisant des outils fournis commercialement, comme un navigateur Web ou un site de média social. Cette catégorie englobe un éventail vaste et en constante évolution de sources de données possibles pour lesquelles il n'existe aucune classification simple. L'une des caractéristiques déterminantes est que les personnes qui fournissent l'information, comme un message sur Twitter, agissent de manière autonome : elles ne doivent pas répondre à un questionnaire ou fournir des renseignements administratifs, mais choisissent plutôt de lancer une interaction.

Je commence par classer chaque source en fonction de deux dimensions, qui sont liées au cadre décrit dans Biemer et coll. (2014). J'attribue le classement en supposant qu'un organisme statistique n'a pas encore pris de mesure proactive afin de l'améliorer (p. ex. en intégrant du personnel dans un organisme administratif afin qu'il se familiarise en profondeur avec les dossiers de cet organisme). Les deux dimensions sont les suivantes :

(1) Degré d'accessibilité de l'organisme statistique national à la source et de contrôle qu'il exerce sur la source : élevé (l'organisme statistique conçoit la source de données et contrôle les changements qui y sont apportés); moyen (l'organisme statistique est autorisé à utiliser la source de données et influe sur les changements qui y sont apportés); faible (l'organisme statistique doit s'arranger pour obtenir la source de données conformément aux conditions établies par le fournisseur et n'a que peu d'influence, voire aucune, sur les changements qui y sont apportés). Une gradation peut être ajoutée à chacune de ces catégories selon, par exemple, la force de l'autorité dont dispose l'organisme pour acquérir un ensemble de dossiers administratifs.

(2) Degré possible de détermination et de mesure des composantes de l'erreur : élevé, comme dans le cas des enquêtes et des recensements conçus par l'organisme; moyen, comme dans le cas des dossiers administratifs des secteurs public et privé; et faible, comme dans le cas des flux de données provenant de choix autonomes de particuliers.

Je détermine ensuite des aspects de la qualité des données pour chaque source, à l'instar de Biemer et coll. (2014). J'indique aussi les variations pour la plupart des dimensions selon le fournisseur, comme un organisme statistique national, une autre unité gouvernementale nationale, un autre palier de gouvernement, une institution universitaire ou une entité commerciale. Toute cette information est regroupée dans le tableau 5.1 au mieux de mes connaissances.

Une source idéale pour un organisme statistique, toutes choses étant égales par ailleurs, est une source qui est fournie, conçue et contrôlée par l'organisme, et pour laquelle les erreurs peuvent être identifiées et mesurées et sont généralement maîtrisées, comme dans le cas d'une enquête à échantillonnage probabiliste de haute qualité, mise sur pied par l'organisme. À l'autre extrême se trouve une source de données qui est contrôlée par une ou plusieurs entreprises privées (p. ex. données de scanneur) ou, peut-être, des centaines

ou des milliers d'administrations publiques locales (p. ex. caméras de circulation), pour laquelle les données résultent de choix autonomes ou de mouvements non contrôlés, et pour laquelle il est difficile de conceptualiser, sans parler de mesurer, les erreurs dans la source de données. Pourtant, étant donné qu'un organisme statistique est chargé de fournir aux décideurs et aux membres du public des statistiques pertinentes, à jour et exactes dont le coût et le fardeau de réponse sont réduits au minimum, il pourrait fort bien exister des sources de données autres que les enquêtes qui justifient l'effort de les rendre utilisables à des fins statistiques. Je soutiens que les menaces qui pèsent sur le paradigme des enquêtes passées en revue plus haut rendent impérative la prise en considération d'autres sources de données, car il n'est plus possible de démontrer que les enquêtes représentent en tout temps et en toutes circonstances un meilleur choix que d'autres sources - elles n'obtiennent pas systématiquement une cote « élevée » sur les dimensions prises en compte dans le tableau 5.1.

Je soutiens aussi que les dossiers administratifs gouvernementaux, qui, comme l'indique le tableau 5.1, possèdent plus souvent les propriétés souhaitables pour la production de statistiques officielles que d'autres sources de données non issues d'enquêtes, devraient être considérés par les organismes statistiques comme une option toute désignée pour une intégration aussi étendue que possible dans leurs programmes d'enquêtes s'ils ne l'ont pas déjà fait. Les dossiers administratifs sont créés conformément à des règles concernant la population admissible, les personnes qui doivent fournir quel type d'information, les mesures qui doivent être prises par l'organisme administratif pertinent en se basant sur l'information (p. ex. remboursement d'impôt, versement de prestations), et ainsi de suite. Cela devrait permettre à un organisme statistique, moyennant l'effort requis, de se familiariser avec les structures d'erreur des dossiers administratifs comme ils le sont avec l'erreur totale d'enquête. Couper (2013) offre une discussion utile quelque peu semblable à la mienne. Il découvre des failles dans la capacité des sources de données organiques à être aussi utiles qu'on l'affirme, sans parler des affirmations quant à leur capacité de remplacer les enquêtes par échantillonnage probabiliste, mais il avertit les chercheurs d'enquête que s'ils ignorent les sources de données organiques, ils le font à leurs risques et périls. Ironiquement, sa conclusion qu'il faut utiliser certaines sources organiques est renforcée par l'erreur qu'il commet en classant les dossiers administratifs comme étant des données organiques. Leur classification correcte est celle de données conçues, même si elles ne le sont pas par un organisme statistique.

Tableau 5.1
Classement (ÉLEVÉ, MOYEN, FAIBLE, TRÈS FAIBLE ou VARIABLE) de quatre sources de données sur les dimensions d'utilisation dans les statistiques officielles

Dimension/ Source de données	Recensement/enquête probabiliste (p. ex. CPS/ASEC, ACS, NHIS - voir tableau 2.1)	Dossiers administratifs (p. ex. impôt sur le revenu, sécurité sociale, chômage, paye)	Enregistrements de transactions commerciales (p. ex. données de scanneurs et de cartes de crédit)	Interactions des particuliers avec Internet (p. ex. Twitter; volumes de termes de recherche dans Google)
Degré de contrôle/ d'accessibilité de la source par l'organisme statistique	ÉLEVÉ (enquête menée pour l'organisme statistique); MOYEN à FAIBLE (enquête menée pour un organisme privé).	ÉLEVÉ à MOYEN (dossiers d'un organisme national); MOYEN à FAIBLE (dossiers d'État ou dossiers locaux); MOYEN à FAIBLE (dossiers commerciaux).	MOYEN à FAIBLE	TRÈS FAIBLE

Dimension/ Source de données	**Recensement/enquête probabiliste (p. ex. CPS/ASEC, ACS, NHIS - voir tableau 2.1)**	**Dossiers administratifs (p. ex. impôt sur le revenu, sécurité sociale, chômage, paye)**	**Enregistrements de transactions commerciales (p. ex. données de scanneurs et de cartes de crédit)**	**Interactions des particuliers avec Internet (p. ex. Twitter; volumes de termes de recherche dans Google)**
Degré de capacité de l'organisme statistique à déterminer/ évaluer les propriétés/ erreurs	ÉLEVÉ (enquête menée pour l'organisme statistique); VARIABLE (enquête menée pour un organisme privé, dépend de la documentation et de la transparence).	ÉLEVÉ à MOYEN (dossiers d'un organisme national); MOYEN à FAIBLE (dossiers d'État ou dossiers locaux); MOYEN à FAIBLE (dossiers commerciaux).	MOYEN (dans la mesure où les enregistrements sont conformes aux normes reconnues (p. ex. pour les codes à barres et les renseignements sur les prix).	TRÈS FAIBLE
Attributs de la qualité des données (Biemer et coll. 2014)				
Pertinence pour les décideurs et les membres du public - Concepts et mesures	ÉLEVÉE pour une enquête menée pour l'organisme statistique, en supposant qu'elle est bien conçue et que les concepts et les mesures sont à jour; VARIABLE pour des enquêtes menées pour des organismes privés.	VARIABLE d'un système de dossiers à l'autre et à l'intérieur des systèmes de dossiers (p. ex. les dossiers de versement de prestations peuvent être très pertinents, tandis que les renseignements sur la composition de la famille peuvent s'appuyer sur un concept différent).	VARIABLE	VARIABLE, mais TRÈS FAIBLE dans l'état actuel des moyens d'acquérir, évaluer et analyser ces types de données.
Pertinence - Covariables utiles	ÉLEVÉE pour la plupart des enquêtes.	VARIABLE, mais rarement aussi élevée que pour la plupart des enquêtes.	VARIABLE, mais rarement aussi élevée que pour la plupart des enquêtes.	VARIABLE, mais habituellement FAIBLE.
Fréquence de collecte des données	D'hebdomadaire à toutes les deux ou trois années (toutes les décennies pour le recensement de la population des États-Unis); quelques enquêtes privées, comme les sondages électoraux, peuvent être exécutées à chaque jour.	En général, les dossiers sont mis à jour fréquemment (p. ex. quotidiennement) et continuellement.	En général, les enregistrements sont mis à jour fréquemment (p. ex. au moment de la transaction ou quotidiennement) et continuellement.	Les interactions sont saisies instantanément.
Actualité des données diffusées	VARIABLE, dépend de l'effort de l'organisme statistique ou de l'organisme privé, mais un certain décalage par rapport à la période de référence de la réponse est inévitable.	VARIABLE, mais un certain décalage par rapport à la date de référence à laquelle les dossiers ont été acquis par l'organisme statistique est probable.	VARIABLE, mais vraisemblablement de longs délais pour l'acquisition de données exclusives par l'organisme statistique.	VARIABLE, mais vraisemblablement de longs délais (quoique le *Billion Prices Project* du MIT ait établi des modalités d'accès très rapide aux prix sur Internet; voir bpp.mit.edu).

Dimension/ Source de données	Recensement/enquête probabiliste (p. ex. CPS/ASEC, ACS, NHIS - voir tableau 2.1)	Dossiers administratifs (p. ex. impôt sur le revenu, sécurité sociale, chômage, paye)	Enregistrements de transactions commerciales (p. ex. données de scanneurs et de cartes de crédit)	Interactions des particuliers avec Internet (p. ex. Twitter; volumes de termes de recherche dans Google)
Comparabilité et cohérence	ÉLEVÉES dans le temps et dans l'espace (géographie) au sein d'une enquête (sauf en cas de changement délibéré ou de changement sociétal affectant les mesures qui n'est pas pris en compte); VARIABLES, selon les enquêtes.	ÉLEVÉES au sein du système de dossiers (changements apportés aux dossiers gouvernementaux généralement annoncés par un changement juridique/ réglementaire/ de politique; changements apportés aux dossiers commerciaux vraisemblablement opaques). VARIABLES, selon les systèmes de dossiers.	ÉLEVÉES au sein du système d'enregistrements (changements généralement opaques pour l'organisme statistique); VARIABLES entre les systèmes d'enregistrements.	TRÈS FAIBLES, en ce sens que les fournisseurs (p. ex. Twitter) peuvent ajouter/soustraire des caractéristiques ou abandonner complètement un produit; changements généralement opaques pour l'organisme statistique; les auteurs des interactions peuvent avoir des cadres de référence très différents.
Exactitude (composantes de l'erreur)*				
Erreur de base de sondage	VARIABLE, possibilité d'un sous-dénombrement ou d'un surdénombrement important.	La base de sondage est habituellement bien définie par une loi, un règlement ou une politique; le problème en cas d'utilisation par un organisme statistique est que la base de sondage pourrait ne pas être exhaustive.	La base de sondage est mal définie pour les besoins d'un organisme statistique, en ce sens qu'elle représente quiconque a eu un achat scanné par un vendeur spécifié ou a utilisé une carte de crédit particulière pour un achat durant une période spécifiée; pose un grand défi à l'organisme statistique en ce qui concerne la détermination de l'usage approprié.	La base de sondage est mal définie pour les besoins d'un organisme statistique, en ce sens qu'elle représente quiconque a décidé, par exemple, de créer un compte Twitter ou d'effectuer une recherche dans Google durant une période spécifiée; pose un grand défi à l'organisme statistique en ce qui concerne la détermination de l'usage approprié.
Non-réponse (totale et partielle)	VARIABLE; peut être importante.	VARIABLE (p. ex. les dossiers de la sécurité sociale couvrent vraisemblablement presque toutes les personnes admissibles, mais les dossiers fiscaux reflètent vraisemblablement la fraude fiscale sous forme d'omission de produire une déclaration de revenus ou de non-déclaration de certains revenus).	SANS OBJET, en ce sens que les « répondants » sont autosélectionnés; le défi pour l'organisme statistique consiste à déterminer l'utilisation appropriée qui ne requiert pas l'hypothèse d'un mécanisme probabiliste.	SANS OBJET, en ce sens que les « répondants » sont autosélectionnés; le défi pour l'organisme statistique consiste à déterminer l'utilisation appropriée qui ne requiert pas l'hypothèse d'un mécanisme probabiliste.

Exactitude (composantes de l'erreur)* (SUITE)				
Dimension/ Source de données	**Recensement/enquête probabiliste (p. ex. CPS/ASEC, ACS, NHIS - voir tableau 2.1)**	**Dossiers administratifs (p. ex. impôt sur le revenu, sécurité sociale, chômage, paye)**	**Enregistrements de transactions commerciales (p. ex. données de scanneurs et de cartes de crédit)**	**Interactions des particuliers avec Internet (p. ex. Twitter; volumes de termes de recherche dans Google)**
Erreur de mesure	VARIABLE, au sein des enquêtes, par question, et entre les enquêtes pour des questions comparables; souvent mal évaluée, même pour les enquêtes réalisées par les organismes statistiques.	VARIABLE entre les systèmes de dossiers et au sein des systèmes de dossiers, par question, selon la mesure dans laquelle la question joue un rôle central dans le fonctionnement du programme (p. ex. une question sur le versement de prestations est vraisemblablement plus exacte que des éléments de données obtenus auprès des bénéficiaires, comme la situation d'emploi).	SANS OBJET pour la source de données en tant que telle, quoique toute caractéristique ajoutée par le vendeur en provenance d'une autre source peut ou non être valide; le défi pour l'organisme statistique consiste à ne pas introduire d'erreur de mesure en utilisant les données de manière inappropriée.	SANS OBJET pour la source de données en tant que telle, quoique toute caractéristique ajoutée par le vendeur en provenance d'une autre source peut ou non être valide; le défi pour l'organisme statistique consiste à ne pas introduire d'erreur de mesure en utilisant les données de manière inappropriée.
Erreur de traitement des données	VARIABLE (p. ex. possibilité d'erreur de saisie des données ou de recodage), mais fait habituellement l'objet d'un bon contrôle statistique, bien que cela soit plus difficile à évaluer pour les enquêtes réalisées par des organismes privés.	VARIABLE (p. ex. possibilité d'erreurs de saisie-clavier ou de codage), vraisemblablement mieux contrôlée pour les variables clés (p. ex. versements de prestations) que pour d'autres variables, mais difficile pour l'organisme statistique de l'évaluer.	VARIABLE (p. ex. possibilité d'erreurs lors de l'attribution des codes à barres ou des prix), vraisemblablement bien contrôlée, mais difficile pour l'organisme statistique de l'évaluer.	SANS OBJET, en ce sens que l'erreur n'est pas définie, quoiqu'il puisse y avoir à l'occasion des problèmes tels que, disons, l'écrasement et la perte d'une journée complète de messages Twitter.
Erreur de modélisation/ estimation	Biais découlant de processus tels que la pondération et l'imputation VARIABLE; souvent, l'organisme statistique déploie d'intenses efforts afin de bien concevoir l'enquête au départ, mais ne procède pas à un réexamen pour s'assurer que les procédures continuent d'être valides.	SANS OBJET (habituellement), en ce sens que les dossiers sont des données « brutes », sauf peut-être dans le cas de certaines variables recodées, mais un biais peut être introduit par l'organisme statistique durant le retraitement.	SANS OBJET (habituellement), en ce sens que les enregistrements sont des données « brutes », sauf peut-être dans le cas de certaines variables recodées ou résumées, mais un biais peut être introduit par l'organisme statistique durant le retraitement.	SANS OBJET (habituellement), en ce sens que les enregistrements sont des données « brutes », mais le retraitement par l'organisme statistique peut introduire un biais important (p. ex. en considérant que le terme « licencié » est toujours indicateur de chômage dans l'analyse des messages Twitter).

Exactitude (composantes de l'erreur)* (SUITE)				
Dimension/ Source de données	**Recensement/enquête probabiliste (p. ex. CPS/ASEC, ACS, NHIS - voir tableau 2.1)**	**Dossiers administratifs (p. ex. impôt sur le revenu, sécurité sociale, chômage, paye)**	**Enregistrements de transactions commerciales (p. ex. données de scanneurs et de cartes de crédit)**	**Interactions des particuliers avec Internet (p. ex. Twitter; volumes de termes de recherche dans Google)**
Erreur de spécification	VARIABLE (p. ex. l'état de santé autodéclaré peut indiquer validement la perception du répondant, mais pas nécessairement l'état de santé physique ou mental diagnostiqué); peut évoluer au cours du temps (p. ex. à mesure que l'usage des mots évolue parmi le public).	VARIABLE; peut être importante si les concepts dans les dossiers administratifs diffèrent de ceux dont l'organisme statistique a besoin (p. ex. les règles concernant la déclaration des revenus sur les formulaires de déclaration peuvent ne pas tenir compte de composantes telles que des avantages de cafétéria).	VARIABLE; peut être faible ou élevée en fonction de la mesure dans laquelle les données correspondent aux besoins de l'organisme statistique.	VARIABLE, mais vraisemblablement importante dans l'état actuel des moyens d'acquérir, d'évaluer et d'analyser ces types de données émanant de choix relativement libres effectués par des individus autonomes.
Fardeau*	VARIABLE, peut être élevé.	PAS DE FARDEAU SUPPLÉMENTAIRE imposé par l'organisme statistique à la population pertinente (p. ex. bénéficiaires), mais fardeau imposé à l'organisme administratif.	PAS DE FARDEAU SUPPLÉMENTAIRE imposé par l'organisme statistique à la population pertinente (p. ex. acheteurs), mais fardeau imposé au fournisseur.	PAS DE FARDEAU SUPPLÉMENTAIRE imposé par l'organisme statistique à la population pertinente (p. ex. afficheurs de messages sur Twitter), mais fardeau imposé au fournisseur.
Coût*	VARIABLE, peut être élevé; l'organisme statistique assume la totalité des coûts de conception, de collecte, de traitement et d'estimation.	VARIABLE, mais peut être plus faible que pour une enquête comparable, parce que l'organisme administratif assume les coûts de collecte des données, mais l'organisme statistique assume vraisemblablement des coûts de manipulation/ traitement spécial.	VARIABLE comme pour les dossiers administratifs, mais le fournisseur souhaite vraisemblablement un paiement; l'organisme statistique assume vraisemblablement des coûts de traitement spécial/ manipulation/ analyse.	VARIABLE comme pour les dossiers administratifs, mais le fournisseur souhaite vraisemblablement un paiement; les coûts supplémentaires assumés par l'organisme statistique pour le traitement/ analyse de données non structurées peuvent être élevés.

*La direction de l'échelle change; autrement dit « élevée » est indésirable et « faible » est désirable.
Note : N'inclut pas l'erreur de révision comprise dans la classification de Biemer et coll. (2014).
Source : Évaluation grossière de l'auteure.

5.3 Utilisations des dossiers administratifs dans les programmes fondés sur des enquêtes-ménages

Les participants aux enquêtes-ménages ont prouvé maintes fois que leurs réponses à de nombreuses questions importantes sur le revenu, la richesse, les dépenses et d'autres sujets ne sont pas très exactes. Dans de nombreux cas, l'utilisation de dossiers administratifs offre la possibilité de remédier à cette

situation. Une autre stratégie adoptée par de nombreux programmes d'enquêtes-ménages aux États-Unis consiste à inviter les répondants eux-mêmes à consulter leurs propres dossiers, comme les déclarations de revenus, lorsqu'ils répondent aux questions sur le revenu ou des sujets similaires. Sans aucun doute, les réponses sont vraisemblablement plus exactes lorsque les dossiers sont consultés, comme Johnson et Moore (pas de date) le constatent dans une comparaison de dossiers fiscaux aux réponses à la SCF pour l'exercice 2000. Cependant, la stratégie proprement dite semble être en grande partie un exercice futile. Selon la même étude de la SCF réalisée par Johnson et Moore, seulement 10 % des ménages dont le revenu brut ajusté est inférieur à 50 000 $ consultent leurs dossiers et seulement 22 % des ménages à revenu élevé le font. Voir National Research Council (2013a, p. 89-91) ainsi que Moore, Marquis et Bogen (1996) pour des constatations similaires au sujet des difficultés à obtenir que les répondants consultent leurs dossiers.

En me penchant maintenant sur les stratégies que les organismes statistiques peuvent adopter pour travailler directement avec des données administratives, je cerne huit façons selon lesquelles les dossiers administratifs peuvent contribuer à la qualité des données des enquêtes-ménages, à savoir 1) aider à évaluer la qualité des données d'enquête, en les comparant à des estimations agrégées, ajustées comme il convient pour tenir compte des différences d'univers et de concepts entre les populations, et par appariement exact des enregistrements de l'enquête et des enregistrements administratifs; 2) fournir des totaux de contrôle pour l'ajustement des poids de sondage afin de tenir compte des erreurs de couverture; 3) fournir des bases de sondage supplémentaires pouvant être utilisées dans un plan à bases de sondage multiples; 4) fournir des renseignements supplémentaires à annexer aux enregistrements d'enquête appariés pour améliorer la pertinence et l'utilité des données; 5) fournir des covariables pour les estimations fondées sur un modèle pour des régions géographiques plus petites que celles pour lesquelles des estimations peuvent être produites directement d'après l'enquête; 6) améliorer les modèles pour l'imputation des données manquantes dans les enregistrements de l'enquête; 7) remplacer « non » pour les participants à l'enquête qui auraient dû répondre à une question, remplacer « oui » pour les participants à l'enquête qui n'auraient pas dû répondre à une question, et remplacer les valeurs déclarées pour les participants à l'enquête qui ont fourni une réponse erronée à une question; et 8) remplacer les questions de l'enquête et utiliser les valeurs des dossiers administratifs directement. Dans une version non publiée plus longue du présent article, je donne des exemples actuels et possibles de chaque type d'utilisation et énumère les avantages, les problèmes de confidentialité et de perception du public, ainsi que les limites et les problèmes de faisabilité pour chaque utilisation, de manière générique et en particulier pour les enquêtes-ménages américaines portant sur des sujets tels que le revenu, les actifs et les dépenses. Ce qui importe, en ce qui me concerne, est que les avantages doivent surpasser les inconvénients, étant donné un programme soutenu, pour intégrer des systèmes de dossiers administratifs à des programmes statistiques.

5.4 Utilisations possibles de sources de données non habituelles

Ayant indiqué antérieurement que les données provenant d'autres sources que les enquêtes et les dossiers administratifs posent un certain nombre de problèmes pour la production de statistiques officielles, il serait négligent de ma part de ne pas discuter brièvement des raisons pour lesquelles ces données semblent si intéressantes. Les entreprises privées ont des fonctions de perte très différentes de celles des organismes statistiques - elles cherchent à avoir un avantage sur leurs concurrents. Des données qui sont plus à jour et qui indiquent des moyens d'accroître les ventes et les profits sont

vraisemblablement utiles à l'entreprise privée, même si elles ne couvrent pas entièrement une population ou qu'elles ont d'autres inconvénients pour les statistiques officielles. Dans cette perspective, les types d'expériences que réalise une entreprise telle que Google, en utilisant ses propres « mégadonnées », afin de trouver des moyens d'augmenter les publicités visionnées sont de bons investissements (voir, par exemple, McGuire, Manyika et Chui 2012). De même, les organismes chargés des programmes, à tous les paliers de gouvernement, souvent en collaboration avec des centres universitaires, regroupent et analysent leurs propres données et d'autres de façons novatrices afin de déceler des tendances, « points chauds », etc., non seulement pour améliorer leurs programmes et planifier de nouveaux services, mais aussi pour classer les ressources par ordre de priorité et améliorer la réponse en temps réel (voir, par exemple, le *Center for Urban Science and Progress* à l'Université de New York (http://cusp.nyu.edu/), ainsi que le *Urban Center for Computation and Data* à l'Université de Chicago (https://urbanccd.org)).

Les organismes statistiques ont besoin, avant tout et par-dessus tout, de sources de données qui couvrent une population connue et présentent des propriétés d'erreur qui sont raisonnablement bien comprises et qui ne sont pas susceptibles de changer sans qu'on s'y attende, c'est-à-dire exemptes de caractéristiques qui sont inhérentes à des sources comme les interactions autonomes avec des sites Web sur Internet. Les programmes fondés sur les enquêtes-ménages des organismes statistiques disposent toutefois d'au moins deux moyens qui pourraient leur permettre de tirer un « avantage » de sources de données non habituelles : l'un consiste à améliorer l'actualité des estimations provisoires des statistiques clés, et l'autre consiste à fournir des indicateurs avancés de l'évolution sociale (p. ex. l'émergence de nouveaux domaines de formation et professions) qui avertissent les organismes statistiques qu'il est nécessaire de modifier leurs concepts et leurs mesures.

6 Des besoins de données aux sources de données : deux exemples aux États-Unis

Afin d'illustrer concrètement mes propos, voici deux exemples relevés aux États-Unis - revenu des ménages et caractéristiques des logements - pour lesquels, selon moi, les organismes statistiques peuvent et doivent transformer leurs programmes d'enquête en programmes à sources de données multiples afin de mieux répondre aux besoins des utilisateurs. Le *U.S. Office of Management and Budget* (2014) a fait un pas dans cette direction dans un mémoire récent où il affirme que les utilisations statistiques des dossiers administratifs des organismes fédéraux représentent un bien positif et énonce les étapes à suivre pour institutionnaliser leur usage.

6.1 Revenu des ménages

Les statistiques officielles sur la répartition des revenus des ménages comptent parmi les indicateurs les plus importants du bien-être économique produits régulièrement par les bureaux nationaux de la statistique, et elles sont encore plus importantes à la lumière des débats actuels concernant les inégalités croissantes et d'autres sujets apparentés. Pourtant, il existe une abondance de preuves que la qualité des mesures du revenu des ménages obtenues d'après les réponses aux enquêtes menées aux États-Unis est altérée considérablement par l'erreur de couverture, la non-réponse totale, la non-réponse partielle et les

erreurs de déclaration. En outre, le concept de revenu monétaire ordinaire appliqué dans les enquêtes américaines est périmé étant donné les moyens complexes et continuellement en évolution par lesquels les ménages obtiennent des ressources pour soutenir leur consommation quotidienne et leur épargne. Il semble impératif que le système statistique des États-Unis améliore ses estimations vedettes du revenu calculées d'après les données de la CPS/ASEC, de la SIPP et, dans la mesure du possible, de l'ACS, en passant d'une approche fondée en grande partie sur les réponses aux enquêtes à une approche visant à intégrer les données de dossiers administratifs à celles de ces enquêtes. Le *Census Bureau* met en œuvre de nouvelles questions et des questions modifiées afin de mieux mesurer le revenu de pension et d'autres sources dans la CPS/ASEC, à la suite d'un examen majeur de la mesure du revenu dans cette enquête réalisée par Czajka et Denmead (2012) et d'un rapport sur les essais cognitifs des modifications apportées au questionnaire de l'ASEC (Hicks et Kerwin 2011). Récemment, le *Census Bureau* a également procédé à une refonte importante de la SIPP, en faisant appel à des méthodes axées sur l'utilisation de calendriers biographiques et d'interviews annuelles à la place d'interviews tous les quatre mois afin de réduire le fardeau et les coûts, et dont les effets sur la qualité doivent être évalués (voir https://www.census.gov/programs-surveys/sipp/about/re-engineered-sipp.html [November 2014]). Un processus a été mis en place pour examiner les questions de l'ACS, mais jusqu'à présent, les questions sur le revenu n'ont pas été abordées. En plus de poursuivre la recherche classique sur les questionnaires afin de trouver des moyens de réduire le fardeau de réponse, de clarifier la signification des questions et de faciliter la réponse aux questions sur le revenu dans la mesure du possible, les enquêtes vedettes seraient améliorées considérablement si :

(1) Le *U.S. Census Bureau* et le *Bureau of Economic Analysis* (BEA) se mettaient d'accord - et réexaminaient périodiquement la situation et la mettaient à jour au besoin - sur un concept contemporain du revenu ordinaire du ménage sur lequel fonder les estimations d'après les données de la CPS/ASEC, de la SIPP et de l'ACS, et la série d'estimations du revenu personnel dans les NIPA, qui sont établies en grande partie d'après des dossiers administratifs. Il existe à l'heure actuelle des différences conceptuelles entre les enquêtes et les NIPA, tel que le traitement des prestations de retraite, qui devraient être conciliées. L'utilisation d'un concept intégré du revenu du ménage rendrait les comptes des revenus personnels et les enquêtes-ménages plus utiles pour analyser les tendances sous les angles macro et microéconomiques.

(2) Le *Census Bureau* effectuait une étude sur les avantages probables de la mise en œuvre d'ajustements des pondérations des enquêtes socioéconomiques en plus des ajustements des pondérations démographiques. En supposant qu'il existe un avantage, le *Census Bureau* déterminerait ensuite quelles sont les sources appropriées, qui pourraient être les dossiers fiscaux ou la SCF, pour ajuster les pondérations dans la CPS, la SIPP et l'ACS afin de tenir compte des différences de couverture par grande catégorie socioéconomique.

(3) Le *Census Bureau* progressait stratégiquement, source par source, afin d'améliorer les imputations des montants de revenu dans la CPS/ASEC et la SIPP en utilisant les valeurs des dossiers administratifs. Le *Census Bureau* a déjà accès à de nombreux dossiers et entreprend des démarches pour en obtenir d'autres (p. ex. dossiers du SNAP provenant des États) dans le cadre de la planification du recensement de 2020.

(4) Le *Census Bureau* évoluait - prudemment, étant donné les obstacles supplémentaires à l'utilisation des dossiers administratifs aux États-Unis - vers le modèle de Statistique Canada, qui permet aux répondants de sauter des blocs entiers de questions sur le revenu en autorisant l'accès à leurs dossiers administratifs (voir http://www.statcan.gc.ca/eng/survey/household/5200 [November 2014]).

Je ne sous-estime aucunement les difficultés que présentent les étapes susmentionnées pour la production des statistiques sur le revenu aux États-Unis. Ces difficultés, sans ordre particulier, comprennent 1) les obstacles juridiques et bureaucratiques à l'obtention d'un accès facile aux dossiers administratifs, lesquels sont considérablement plus importants pour les dossiers détenus par les organismes d'État en raison de différences entre les lois, les politiques et les normes et systèmes de données des États; 2) les considérations concernant l'obtention du consentement des répondants, particulièrement si les valeurs tirées des dossiers sont substituées à des questions; 3) les perceptions de type « Big Brother » et de menace pour la vie privée, qui peuvent limiter l'accessibilité des microdonnées pour la recherche et l'analyse des politiques; 4) le manque de ressources permettant aux organismes statistiques d'entreprendre des activités telles que le remaniement des systèmes d'imputation; 5) les effets indésirables sur l'actualité des données dans la mesure où les dossiers sont mis à la disposition des organismes statistiques avec un certain décalage temporel, problème qui pourrait être résolu en diffusant des estimations provisoires suivies par des estimations définitives quand suffisamment de données administratives deviennent disponibles; 6) la connaissance insuffisante des structures d'erreur des dossiers, qui pourrait donner lieu à des surprises désagréables; 7) les différences conceptuelles entre les mesures appliquées dans les dossiers et dans les enquêtes, dont il n'est pas tenu compte facilement (p. ex. les revenus déclarés à l'IRS ne sont pas les revenus bruts, mais les revenus imposables); 8) le fardeau supplémentaire imposé aux employés déjà fortement sollicités des bureaux centraux des organismes statistiques; 9) la nécessité de réécrire les systèmes de traitement afin de relier de multiples flux de données et d'exécuter toutes les tâches nécessaires d'appariement, de réconciliation et d'estimation dans les délais prévus; 10) la méfiance de nombreux utilisateurs de microdonnées aux États-Unis, qui semblent préférer un ensemble de données provenant d'une source unique, comme une enquête, indépendamment des inexactitudes dans les données, à un ensemble de données provenant de sources multiples, qui pourraient contenir des valeurs fondées sur un modèle pour certaines variables; et 11) l'hésitation des employés de l'organisme statistique, qui semblent souvent croire qu'il n'est pas approprié d'utiliser, disons, des dossiers administratifs pour imputer un revenu à un répondant qui n'a pas fait état d'un revenu ou d'utiliser des dossiers administratifs pour remplacer certaines questions ou pour améliorer certaines imputations, à moins que cela ne puisse être fait pour toutes les questions. Dans sa planification du recensement de 2020, le *Census Bureau* envisage d'utiliser de manière limitée des dossiers administratifs pour le suivi des cas de non-réponse, et cela pourrait servir de modèle pour l'usage sélectif de dossiers administratifs dans les enquêtes-ménages. Même si elles sont formidables, ces difficultés ne sont nullement insurmontables. Un plan stratégique par étapes, bien formulé, en vue d'adopter une approche à sources de données multiples pourrait donner aux organismes statistiques la possibilité de travailler à l'amélioration des estimations du revenu et de réussir simultanément à réduire le fardeau de réponse et, peut-être, les coûts des programmes d'enquêtes clés.

6.2 Caractéristiques des logements, y compris les installations de plomberie

En réponse aux préoccupations concernant la mauvaise qualité des logements au pays, exprimées dans le *New Deal*, le recensement décennal des États-Unis de 1940 comprenait quelques questions sur les caractéristiques des logements. Cette préoccupation était bien fondée - le recensement de 1940 a révélé, par exemple, que 45 % des logements n'étaient pas équipés d'une installation de plomberie complète (eau courante chaude et froide, toilette avec chasse d'eau, douche ou baignoire). Voir https://www.census.gov/hhes/www/housing/census/historic/plumbing.html [November 2014]. Les questions sur le logement se sont multipliées et ont été incluses dans les recensements jusqu'en 2000. Quand l'*American Community Survey* est entrée en vigueur, elle comprenait les questions sur le logement qui figuraient antérieurement dans le questionnaire complet du recensement. L'*American Housing Survey* (AHS), qui est une enquête bisannuelle de beaucoup plus petite portée, recueille une gamme encore plus vaste de renseignements sur les logements et les quartiers.

La raison principale de chercher des moyens de transférer les questions sur le logement de l'ACS d'un programme fondé sur une enquête à un programme fondé sur une enquête plus d'autres sources de données est le fardeau de réponse, tant réel que perçu, qui, dans le climat politique actuel aux États-Unis, menace la viabilité de l'ACS. Comme le travail sur le terrain de l'ACS se déroule auprès d'un grand échantillon d'environ 280 000 ménages chaque mois, au lieu de tous les dix ans comme pour l'échantillon du questionnaire complet du recensement qu'elle a remplacé, l'enquête génère un flot modéré, mais continu, de plaintes pour les membres du Congrès, qui ont donné lieu à la tenue d'audiences. Le *Census Bureau* a déterminé les quatre questions de l'ACS qui suscitent le plus de plaintes, à savoir le revenu, l'incapacité, l'heure de départ pour le travail et les installations de plomberie (voir http://www.census.gov/acs/www/Downloads/operations_admin/2014_content_review/ACSContentReviewSummit.pdf[November 2014]). Les questions sur les installations de plomberie figurant dans le questionnaire complet du recensement faisaient aussi régulièrement l'objet de plaisanteries et de plaintes. En fait, les gens répondent de manière assez exhaustive à ces questions (voir http://www.census.gov/acs/www/methodology/item_allocation_rates_data/[October 2014]), mais elles continuent de susciter du ressentiment et sont parfois mal comprises (voir Woodward, Wilson et Chestnut 2007). De surcroît, un examen du questionnaire complet de l'ACS donne à penser que l'ensemble complet d'environ 30 questions sur le logement impose un fardeau considérable à de nombreux ménages, particulièrement les propriétaires ayant un prêt hypothécaire.

En réponse à ces préoccupations au sujet du fardeau de l'ACS, le *Census Bureau* a réduit le nombre d'appels et de visites de suivi (voir Zelenak et Davis 2013), a établi un « champion des répondants » et a fourni aux membres du public des renseignements justifiant les questions. Néanmoins, le 30 mai 2014, la Chambre des représentants des États-Unis a voté une loi de crédits qui, si elle est adoptée, transformera l'ACS en une enquête à participation volontaire plutôt qu'obligatoire. Même si des données de bonne qualité pouvaient vraisemblablement être recueillies avec un suivi suffisant, le coût de l'ACS augmenterait considérablement (voir Griffin 2011). Récemment, le *Census Bureau* a demandé aux organismes fédéraux de fournir une justification législative ou réglementaire pour chaque question, et il est fort possible que certaines questions soient abandonnées (voir http://www.census.gov/acs/www/about_the_survey/acs_content_review/ [November 2014]). Les questions sur les installations de plomberie semblent être de bonnes candidates à l'élimination dans l'ACS, étant donné qu'en 2012 aux États-Unis, seulement 0,4 % des logements n'étaient pas équipés d'une installation de plomberie complète (voir

http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_1YR_DP04&prodType=table [November 2014]). Cependant, ce faible pourcentage est concentré dans des régions particulières, comme les réserves habitées par les Autochtones et les régions rurales. En outre, l'élimination de toute question de l'ACS semble être une mesure radicale sans d'abord examiner si d'autres sources pourraient fournir les données.

En fait, les réponses à certaines questions sur le logement du questionnaire de l'ACS pourraient fort vraisemblablement être tirées d'une gamme d'autres sources, annexées au Fichier maître des adresses (FMA) du *Census Bureau*, et être disponibles en vue de leur inclusion dans l'ACS et d'autres enquêtes qui utilisent le FMA comme base de sondage. D'autres sources comprennent les dossiers administratifs des administrations locales sur les impôts fonciers établis, l'année de la construction et d'autres caractéristiques des propriétés, renseignements qui sont de plus en plus fréquemment compilés par des fournisseurs commerciaux, ce qui réduit la nécessité d'interagir individuellement avec les milliers d'administrations publiques aux États-Unis. Elles comprennent aussi des sources comme *Google Street View* pour les caractéristiques extérieures des propriétés, les sites Web des agents immobiliers pour la valeur des logements et les caractéristiques intérieures (p. ex. nombre de pièces), les compteurs intelligents pour les coûts des services publics (installés dans certaines régions et dont l'usage se répandra vraisemblablement), et les bases de données sur les prêts hypothécaires détenues par les organismes fédéraux et les vendeurs commerciaux. Les caractéristiques des logements qui changent rarement peuvent aussi être extraites des réponses fournies par les échantillons antérieurs ayant reçu le questionnaire complet du recensement. Les installations de plomberie en sont un parfait exemple - une fois qu'un logement est doté d'une installation de plomberie, celle-ci n'est presque jamais démantelée (même s'il peut arriver qu'elle ne soit pas fonctionnelle).

Ces sources de données de rechange varient en ce qui concerne la facilité d'acquisition et d'évaluation des données, la menace réelle ou perçue qu'elles font peser sur la vie privée et la confidentialité, et la mesure dans laquelle elles couvrent l'entièreté ou la majorité du pays. L'élaboration d'un Fichier maître des adresses et des logements (FMAL) augmenté, qui peut servir l'ACS et d'autres programmes statistiques du *Census Bureau*, prendra du temps et, pour certains sujets (p. ex. installations de plomberie), il pourrait être nécessaire d'utiliser une version distincte (plus longue) du questionnaire dans des régions géographiques sélectionnées auquel sont ajoutées les questions pertinentes. Tout cela sera complexe et délicat, mais les avantages potentiels à long terme sont considérables. Pour passer à un FMAL augmenté, le *Census Bureau* peut tirer parti des travaux de l'*Office of Policy Development & Research* du *U.S. Department of Housing and Urban Development* pour simplifier le long questionnaire de l'AHS en utilisant d'autres sources de données pour de nombreuses caractéristiques du logement et du quartier en vue de remplacer les questions de l'enquête; voir http://www.huduser.org/portal/datasets/ahs.html#planning [November 2014].

7 Défis et stratégies pour procéder à un changement de paradigme

J'ai présenté des arguments en faveur d'un nouveau paradigme en vertu duquel les organismes statistiques conçoivent et mettent à jour leurs programmes vedettes en déterminant quelle est la meilleure combinaison de sources de données et de méthodes pour répondre aux besoins des utilisateurs dans un

domaine dont l'importance est croissante. J'utilise les enquêtes-ménages réalisées aux États-Unis comme exemple d'une situation où il existe des preuves solides que s'appuyer uniquement sur les réponses aux enquêtes ne suffira pas pour répondre aux besoins critiques d'information de haute qualité sur le revenu, les dépenses et des sujets apparentés. Je pense qu'il est également vrai que l'utilisation de dossiers administratifs seulement, comme dans certains pays dotés de registres de population détaillés, pourrait ne pas fournir de renseignements suffisamment complets et de haute qualité en l'absence d'efforts réguliers en vue d'examiner la qualité des données des registres et d'augmenter et de corriger ces derniers au moyen d'information provenant d'autres sources, telles que les enquêtes. Par exemple, Axelson, Homberg, Jansson, Werner et Westling (2012) décrivent l'utilité des enquêtes pour évaluer la qualité des données sur les logements et les ménages provenant d'un nouveau registre des logements créé pour le recensement de 2011 en Suède.

Je conclus par une liste de facteurs qui rendent le changement de paradigme difficile, en énonçant aussi des moyens de procéder au changement que je recommande et de l'intégrer dans la culture des organismes statistiques. Les systèmes statistiques des États-Unis et d'autres pays ont fait preuve d'innovation dans de nombreux aspects de leurs programmes, mais changer les paradigmes est toujours un exercice difficile, comme en témoigne la bataille en vue d'introduire l'échantillonnage probabiliste dans la statistique officielle aux États-Unis durant les années 1930. Il est particulièrement difficile de repenser des programmes statistiques permanents, établis depuis longtemps, avec lesquels l'organisme producteur et les utilisateurs se sentent à l'aise.

Les facteurs qui peuvent entraver le changement comprennent 1) l'inertie, particulièrement quand un programme était au départ novateur et très bien conçu, de sorte qu'il peut se reposer sur ses lauriers; 2) le décalage par rapport à l'évolution des besoins des parties prenantes, lequel peut être exacerbé quand un organisme se voit comme la seule source des données nécessaires et sans concurrence; 3) la crainte d'amoindrir les programmes existants conjuguée à la crainte du « non inventé ici »; 4) l'évaluation continue inadéquate de toutes les dimensions de la qualité des données; et 5) la compression des ressources humaines et budgétaires, conjuguée à une hésitation compréhensible du personnel de l'organisme ou de leur base établie d'utilisateurs à réduire l'une ou l'autre des séries statistiques établies de longue date afin de réaliser d'importants progrès dans d'autres séries.

Pourtant, il existe de nombreux exemples remarquables d'innovations importantes mises en place par les organismes statistiques aux États-Unis et dans d'autres pays, de sorte qu'il existe manifestement des moyens de surmonter les obstacles énumérés plus haut pour procéder au changement de paradigme. Selon moi, l'ingrédient essentiel à un changement de paradigme est le ralliement des cadres et le soutien permanent de la haute direction d'un organisme statistique, déployé proactivement de manière à rallier le personnel à tous les niveaux de l'organisme. À titre d'exemple exceptionnel d'un tel leadership, voir dans National Research Council (2010a) la discussion du rôle de Morris Hansen et de ses collègues dans le remaniement de ce qui avait été un recensement effectué par des agents recenseurs en un recensement avec envoi et retour du questionnaire par la poste. Les travaux de remaniement ont été lancés et poursuivis après que l'on ait dégagé des preuves de l'existence d'un biais et d'une variance d'intervieweur considérables pour des éléments de données importants. On craignait également qu'il devienne plus difficile de recruter des agents recenseurs à mesure que les femmes entraient sur le marché du travail.

Des mesures particulières en vue d'obtenir l'appui des cadres de l'organisme dans le but précis d'inculquer l'utilisation de multiples sources de données dans les programmes permanents de statistiques officielles comprennent (voir Prell et coll. 2009, qui ont effectué des études de cas d'utilisation statistique fructueuse de dossiers administratifs aux États-Unis, pour des conclusions similaires) : 1) l'établissement d'attentes et d'objectifs précis pour les employés, par exemple l'attente que les programmes statistiques combineront d'office des sources telles que les enquêtes et les dossiers administratifs afin de produire des données pertinentes, exactes et à jour de manière rentable et en imposant un fardeau de réponse minimal; 2) l'attribution d'un rôle important aux spécialistes du domaine - interactions avec les utilisateurs externes et les producteurs internes des données; 3) la dotation des programmes opérationnels en personnel compétent en ce qui concerne toutes les sources de données pertinentes, ce qui inclut mettre sur un pied d'égalité les spécialistes de la conception des enquêtes et les spécialistes des dossiers administratifs ou d'autres sources de données; 4) la rotation des affectations, y compris des rotations internes, des rotations entre organismes statistiques, des rotations avec les organismes utilisateurs des données et des rotations avec les entités fournissant les sources de données de rechange; 5) la mise en place de ressources pour l'évaluation continue et 6) le traitement des organisations possédant des sources de données de rechange qui jouent un rôle important dans les programmes statistiques comme des partenaires. Sur ce dernier point, voir, par exemple, Hendriks (2012, p. 1473), qui, en décrivant les expériences de Statistics Norway au sujet de son premier recensement fondé sur des registres en 2011, insiste sur le fait que « Les trois C des statistiques fondées sur des registres (afin de produire des données de qualité) sont la coopération, la communication et la coordination » [*Traduction*].

Les organismes statistiques ont montré qu'ils étaient capables d'effectuer des changements de grande portée en réponse aux menaces qui pèsent sur les moyens établis de fonctionnement. La deuxième moitié du 20e siècle a donné le paradigme des enquêtes à échantillonnage probabiliste en réponse à la croissance des coûts et du fardeau associée à la réalisation de dénombrements complets et aux défauts des plans d'échantillonnage non probabilistes. Le 21e siècle peut sans aucun doute nous donner le paradigme de l'utilisation des meilleures sources de données, incluant les enquêtes, les dossiers administratifs et d'autres sources, pour répondre au besoin de statistiques officielles pertinentes, exactes, à jour et rentables des décideurs et des membres du public.

Remerciements

Le présent article est fondé sur les années d'expérience acquise par l'auteure auprès du *Committee on National Statistics* (CNSTAT), mais les opinions exprimées sont les siennes et ne doivent pas être interprétées comme représentant celles du CNSTAT ni celles de la *National Academy of Sciences*. L'auteure remercie John Czajka, David Johnson et Rochelle Martinez de leurs commentaires constructifs au sujet d'une version antérieure. Une version plus longue du présent article peut être obtenue sur demande auprès de l'auteure.

Bibliographie

Anderson, M.J. (1988). *The American Census: A Social History*. New Haven. CT: Yale University Press.

Axelson, M., Homberg, A, Jansson, I., Werner, P. et Westling, S. (2012). Doing a register-based census for the first time: The Swedish experience. *Paper presents at the Joint Statistical Meetings*, San Diego, CA (août). Statistics Sweden, Stockholm.

Bee, A., Meyer, B.D. et Sullivan, J.X. (2012). The validity of consumption data: Are the consumer expenditure interview and diary surveys informative? *NBER Working Paper No. 18308*. Cambridge, MA: National Bureau of Economic Research.

Biemer, P., Trewin, D., Bergdahl, H. et Lilli, J. (2014). A system for managing the quality of official statistics, with discussion. *Journal of Official Statistics*, 30(3, septembre), 381-442.

Brackstone, G. (1999). La gestion de la qualité des données dans un bureau de statistique. *Techniques d'enquête*, 25(2), 159-171.

Bradburn, N.H. (1992). A response to the nonresponse problem. 1992 AAPOR Presidential Address. *Public Opinion Quarterly*, 56(3), 391-397.

Citro, C.F. (2012). *Editing, Imputation and Weighting*. Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey, Second Edition, M. J. Anderson, C.F. Citro et J.J. Salvo, eds, 201-204. Washington, DC: CQ Press.

Couper, M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. Keynote presentation at the *5th European Survey Research Association Conference*. Ljubliana, Slovenia. http://www.europeansurveyresearch.org/sites/default/files/files/Couper%20keynote.pdf [July 2014].

Czajka, J.L. (2009). SIPP data quality. Appendix A in *Reengineering the Survey of Income and Program Participation*. National Research Council. Washington, DC: The National Academies Press.

Czajka, J.L. et Denmead, G. (2012). Income measurement for the 21st century: Updating the current population survey. Washington, DC: *Mathematica Policy Research*. Disponible au http://www.mathematica-mpr.com/~/media/publications/PDFs/family_support/income_measurement__21_century.pdf [July 2014].

Czajka, J.L., Jacobson, J.E. et Cody, S. (2004). Survey estimates of wealth: A comparative analysis and review of the Survey of Income and Program Participation. *Social Security Bulletin*, 65(1). Disponible au http://www.ssa.gov/policy/docs/ssb/v65n1/v65n1p63.html [July 2014].

Daas, P.J.H., Ossen, S.J.L., Tennekes, M. et Nordholt, E.S. (2012). Evaluation of the quality of administrative data used in the Dutch virtual census. *Paper presents at the Joint Statistical Meetings*, San Diego, CA (août). Methodology Sector and Division of Social and Spatial Statistics, Statistics Netherland, The Hague.

De Leeuw, E.D. et De Heer, W. (2002). *Trends in Household Survey Nonresponse: A Longitudinal and International Comparison*. R.M. Groves, D.A. Dillman, J. L. Eltinge et R.J.A. Little, eds. Survey Nonresponse, 41-54. New York: Wiley.

Duncan, J. W. et Shelton, W. C. (1978). *Revolution in United States Government Statistics 1926–1976*. Office of Federal Statistical Policy and Standards, U.S. Department of Commerce. Washington, DC: Government Printing Office.

Eurostat. (2000). Assessment of the quality in statistics. *Doc. Eurostat/A4/Quality/00/General/Standard report*. Luxembourg (4-5 avril). Disponible au http://www.unece.org/fileadmin/DAM/stats/documents/2000/11/ metis/crp.3.e.pdf [July 2014].

Fixler, D. et D.S. Johnson (2012). Accounting for the distribution of income in the U.S. National Accounts. *Paper prepared for the NBER Conference on Research in Income and Wealth*, 30 septembre. Disponible au http://www.bea.gov/about/pdf/Fixler_Johnson.pdf.

Fricker, S. et R. Tourangeau (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, 74(5), 935-955.

Griffin, D. (2011). Cost and workload implications of a voluntary American community survey. *U.S. Census Bureau,* Washington, DC (June 23).

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75(9), 861-871. Special 75th Anniversary Issue.

Groves, R.M. et Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72(2), 167-189.

Harris-Kojetin, B. (2012). *Federal Household Surveys*. Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey, Second Edition, M. J. Anderson, C.F. Citro et J.J. Salvo, eds, 226-234. Washington, DC: CQ Press.

Heckman, J. J. et LaFontaine, P.A. (2010). The American high school graduation rate: trends and levels. *NBER Working Paper 13670*. Cambridge, MA, National Bureau of Economic Research. Disponible au http://www.nber.org/papers/w13670 [July 2014].

Hendriks, C. (2012). Input data quality in register based statistics-The Norwegian experience. Proceedings of the *International Association of Survey Statisticians-JSM 2012*, 1473-1480. Article présenté au Joint Statistical Meetings, San Diego, CA (août). Statistics Norway, Kongsvinger, Norway.

Hicks, W. et Kerwin, J. (2011). Cognitive testing of potential changes to the Annual Social and Economic Supplement of the Current Population Survey. *Report to the U.S. Census Bureau*, Westat, Rockville, MD (25 juillet).

Holt, D.T. (2007).The official statistics Olympics challenge: Wider, deeper, quicker, better, cheaper. *The American Statistician*, 61(1, février), 1-8. Avec les commentaries de G. Brackstone et J.L. Norwood.

Horrigan, M.W. (2013). Big data: A BLS perspective. *Amstat News*, 427(janvier), 25-27.

Hoyakem, C., Bollinger, C. et Ziliak, J. (2014). The role of CPS nonresponse on the level and trend in poverty. *UKCPR Discussion Paper Series*, DP 2014-05. Lexington, KY: University of Kentucky Center for Poverty Research.

Iwig, W., Berning, M., Marck, P. et Prell, M. (2013). Data quality assessment tool for administrative data. Prepared for a subcommittee of the *Federal Committee on Statistical Methodology*, Washington, DC (février).

Johnson, B et Moore, K. [pas de date]. Consider the source: Differences in estimates of income and wealth from survey and tax data. Disponible au http://www.irs.gov/pub/irs-soi/johnsmoore.pdf [July 2014].

Keller, S.A., Koonin, S.E. et Shipp, S. (2012). Big data and city living - what can it do for us? *Statistical Significance*, 9(4), 4-7, août.

Kennickell, A. (2011). Look again: Editing and imputation of SCF panel data. *Paper prepared for the Joint Statistical Meetings*, Miami, FL (août).

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. *META Group [now Gartner] Research Note*, 6 février. Voir http://goo.gl/Bo3GS [July 2014].

Manski, C.F. (2014). Communicating uncertainty in official economic statistics. *NBER Working Paper No. 20098*. Cambridge, MA: National Bureau of Economic Research.

McGuire, T., Manyika, J. et Chui, M. (2012). Why big data is the new competitive advantage. *Ivey Business Journal* (juillet-août).

Meyer, B. D. et Goerge, R.M. (2011). Errors in survey reporting and imputation and their effects on estimates of Food Stamp Program participation. Working Paper. *Chicago Harris School of Public Policy*, University of Chicago.

Meyer, B.D., Mok, W. K.C. et Sullivan, J.X. (2009). The under-reporting of transfers in household surveys: Its nature and consequences. *NBER Working Paper No. 15181*. Cambridge, MA: National Bureau of Economic Research.

Moore, J.C., Marquis, K.H. et Bogen, K. (1996). The SIPP cognitive research evaluation experiment: Basic results and documentation. *SIPP Working Paper No. 212*. U.S. Census Bureau, Washington, DC (janvier). Disponible au http://www.census.gov/sipp/workpapr/wp9601.pdf [July 2014].

Morganstein, D. et Marker, D. (2000). A conversation with Joseph Waksberg. *Statistical Science*, 15(3), 299-312.

Mule, T. et Konicki, S. (2012). *2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Housing Units in the United States*. U.S. Census Bureau, Washington, DC.

National Research Council (1995). *Measuring Poverty: A New Approach*. Washington, DC: The National Academies Press.

National Research Council (2004). *The 2000 Census: Counting Under Adversity*. Washington, DC: The National Academies Press.

National Research Council (2010a). *Envisioning the 2010 Census*. Washington, DC: The National Academies Press.

National Research Council (2010b). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press.

National Research Council (2013a). *Measuring What We Spend: Toward a New Consumer Expenditure Survey*. Washington, DC: The National Academies Press.

National Research Council (2013b). *Nonresponse in Social Science Surveys: A Research Agenda*. Washington, DC: The National Academies Press.

National Research Council (2013c). *Principles and Practices for a Federal Statistical Agency*. Washington, DC: The National Academies Press.

Nelson, N. et West, K. (2014). Interview with Lars Thygesen. *Statistical Journal of the IAOS*, 30, 67-73.

Passero, B. (2009). The impact of income imputation in the Consumer Expenditure Survey. *Monthly Labor Review* (août), 25-42.

Prell, M., Bradsher-Fredrick, H., Comisarow, C., Cornman, S., Cox, C., Denbaly, M., Martinez, R.W., Sabol, W. et Vile, M. (2009). Profiles in success of statistical uses of administrative records. Report of a subcommittee of the *Federal Committee on Statistical Methodology*, U.S. Office of Management and Budget, Washington, DC.

Shapiro, G.M. et Kostanich, D. (1988). High response error and poor coverage are severely hurting the value of household survey data. *Proceedings of the Section on Survey Research Methods*, 443-448, American Statistical Association, Alexandria, VA. Disponible au http://www.amstat.org/sections/srms/Proceedings/papers/1988_081.pdf [July 2014].

Steeh, C.G. (1981). Trends in nonresponse rates, 1952-1979. *Public Opinion Quarterly*, 45, 40-57.

Tourangeau, R. (2004). Survey research and societal change. *Annual Review of Psychology*, 55, 775-801.

U.S. Office of Management and Budget. (2014). *Guidance for Providing and Using Administrative Data for Statistical Purposes*. Memorandum M-14-06. Washington, DC.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Woodward, J., Wilson, E. et Chesnut, J. (2007). *Evaluation Report Covering Facilities - Final Report. 2006 American Community Survey Content Test Report H.3.U.S.* Census Bureau. Washington, DC: U.S. Department of Commerce. Janvier.

Zelenak, M.F. et M.C. David (2013). *Impact of Multiple Contacts by Computer-Assisted Telephone Interview and Computer-Assisted Personal Interview on Final Interview Outcome in the American Community Survey*. U.S. Census Bureau, Washington, DC.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 183-210
Statistique Canada, n° 12-001-X au catalogue

Approches fréquentiste et bayésienne pour comparer les composantes de l'écart intervieweurs dans deux groupes d'intervieweurs d'enquête

Brady T. West et Michael R. Elliott[1]

Résumé

Les méthodologistes d'enquête étudient depuis longtemps les effets des intervieweurs sur la variance des estimations d'enquête. Les modèles statistiques tenant compte des effets aléatoires des intervieweurs sont souvent intégrés à ce genre d'études, et l'intérêt de la recherche repose sur l'ampleur de la composante de la variance de l'intervieweur. Une question peut se poser au cours d'une étude méthodologique : différents groupes d'intervieweurs (p. ex. ceux ayant de l'expérience relative à une enquête donnée par rapport aux nouvelles recrues, ou les intervieweurs IPAO par rapport aux intervieweurs ITAO) ont-ils des composantes de variance considérablement différentes dans ces modèles? Des écarts importants peuvent indiquer un besoin de formation supplémentaire pour certains sous-groupes, les propriétés moins optimales de différents modes ou styles d'interview pour certaines questions d'enquête (en ce qui concerne l'erreur quadratique moyenne globale des estimations d'enquête). Les chercheurs d'enquête désirant des réponses à ces types de questions disposent de différents outils statistiques. Le présent article cherche à fournir un aperçu des approches fréquentiste et bayésienne de rechange de la comparaison des composantes de la variance dans différents groupes d'intervieweurs d'enquête, au moyen d'un cadre de modélisation linéaire généralisée hiérarchique qui tient compte de différents types de variables d'enquête. Nous considérons d'abord les avantages et les limites de chaque approche, en comparant les méthodes utilisées pour l'estimation et l'inférence. Nous présentons ensuite une étude de simulation, en évaluant de façon empirique la capacité de chaque approche d'estimer efficacement les différences entre les composantes de la variance. Nous appliquons alors les deux approches à une analyse des données d'enquête réelles recueillies dans le cadre de la *National Survey of Family Growth* (NSFG) aux États-Unis. Nous concluons que les deux approches ont tendance à donner des inférences très semblables et nous présentons des suggestions à mettre en pratique, compte tenu des différences subtiles observées.

Mots clés : Écart intervieweurs; analyse bayésienne; modèles linéaires généralisés hiérarchiques; test du rapport des vraisemblances.

1 Introduction

La variance entre intervieweurs des méthodes d'enquête (p. ex. West, Kreuter et Jaenichen 2013; West et Olson 2010; Gabler et Lahiri 2009; O'Muircheartaigh et Campanelli 1998; Biemer et Trewin 1997; Kish 1962) se produit lorsque les réponses à une enquête obtenues par un intervieweur donné se ressemblent davantage que les réponses recueillies par différents intervieweurs. La variance entre intervieweurs peut augmenter la variance des estimations d'enquête des moyennes, et peut se produire à cause d'écarts de réponse corrélés introduits par un intervieweur (p. ex. Biemer et Trewin 1997), compte tenu de la complexité des questions de l'enquête (p. ex. Collins et Butcher 1982) ou des interactions entre l'intervieweur et le répondant (p. ex. Mangione, Fowler et Louis 1992), ou de la variance de l'erreur de non-réponse entre les intervieweurs (West et coll. 2013; Lynn, Kaminska et Goldstein 2011; West et Olson 2010).

Les organismes de recherche par sondages forment les intervieweurs de manière à éliminer cette composante de la variance des estimations d'enquête, qui est parfois plus grande que la composante de la

1. Brady T. West, Survey Methodology Program, Institute for Social Research, 426, rue Thompson, Ann Arbor, MI, 48106, courriel : bwest@umich.edu; Michael R. Elliott, Survey Methodology Program, Institute for Social Research, 426, rue Thompson, Ann Arbor, MI, 48106, courriel : mrelliot@umich.edu.

variance due à l'échantillonnage en grappes (Schnell et Kreuter 2005). En fait, une composante de l'écart intervieweurs ne peut jamais être égale à 0 (ce qui supposerait que les moyennes de la variable d'intérêt sont identiques pour tous les intervieweurs), mais les gestionnaires d'enquête cherchent à réduire cette composante au moyen de formation spécialisée des intervieweurs. Par exemple, les intervieweurs peuvent s'exercer à poser certaines questions sous la supervision directe du personnel de formation, avant de recevoir des commentaires sur toute variance de l'administration observée par le personnel (afin de normaliser l'administration; voir Fowler et Mangione 1990). Dans certains plans non imbriqués, où les intervieweurs sont généralement affectés exclusivement à une seule zone d'échantillonnage primaire (p. ex. la National Survey of Family Growth aux États-Unis; voir Lepkowski, Mosher, Davis, Groves et Van Hoewyk 2010), les effets de l'intervieweur et les effets de la zone sont confondus, ce qui empêche l'estimation de la variance des estimations d'enquête qui est attribuable exclusivement aux intervieweurs. Les plans d'échantillonnage imbriqués astucieux (Mahalanobis 1946) permettent aux intervieweurs de travailler dans plusieurs zones d'échantillonnage, et dans ces cas, des modèles multiniveaux recoupés peuvent être utilisés pour estimer les composantes de la variance due aux intervieweurs et aux zones (p. ex. Durrant, Groves, Staetsky et Steele 2010; Gabler et Lahiri 2009; Schnell et Kreuter 2005; O'Muircheartaigh et Campanelli 1999; O'Muircheartaigh et Campanelli 1998).

En général, l'estimation de l'ampleur globale de l'écart intervieweurs en ce qui concerne les mesures d'une variable d'enquête donnée ou le résultat du processus de collecte des données est un exercice utile pour les praticiens d'enquête. Si des sous-échantillons aléatoires d'unités d'échantillonnage sont attribués aux intervieweurs après une conception imbriquée, on peut estimer la composante de la variance due aux intervieweurs et par la suite, les effets particuliers des intervieweurs sur la variance d'une moyenne d'enquête estimée (p. ex. Groves 2004, p. 364). Les grosses estimations peuvent indiquer des difficultés de mesure potentielles éprouvées par certains intervieweurs, ou une réussite différentielle possible du recrutement de types particuliers d'unités échantillonnées. Compte tenu d'une estimation relativement importante d'une composante de l'écart intervieweurs et d'un test statistique approprié indiquant que la composante est considérablement plus grande que zéro (ou « non négligeable », étant donné que les composantes de la variance ne peuvent pas en principe être exactement égales à zéro; voir Zhang et Lin 2010), les gestionnaires d'enquête peuvent utiliser diverses méthodes pour calculer les prévisions des effets aléatoires associés aux intervieweurs individuels, et pour cerner les intervieweurs qui éprouvent peut-être des difficultés avec certains aspects du processus de collecte des données.

Bien que l'estimation des composantes de l'écart intervieweurs et les rajustements subséquents de la formation des intervieweurs et des protocoles de collecte des données font partie de la documentation des méthodes d'enquête depuis longtemps (voir Schaeffer, Dykema et Maynard 2010, pour un examen récent), aucune étude des méthodes d'enquête à ce jour n'a examiné les approches de rechange qui sont offertes aux chercheurs d'enquête pour *comparer* les composantes de la variance dans deux groupes indépendants d'intervieweurs d'enquête. En général, les approches statistiques de rechange sont disponibles pour estimer les composantes de l'écart intervieweurs, et les estimations (et les inférences correspondantes au sujet des composantes de la variance) peuvent être sensibles à la méthodologie d'estimation qu'utilise un chercheur d'enquête. Il en va de même pour les chercheurs d'enquête qui peuvent souhaiter comparer les composantes de la variance associées à différents groupes d'intervieweurs, pour différentes raisons (p. ex. cerner les groupes qui ont besoin de plus de formation ou trouver des modes plus efficaces pour certains types de questions) : il existe différentes approches statistiques pour

effectuer ces types de comparaisons, et les inférences au sujet des différences peuvent varier en fonction de l'approche utilisée. Dans le présent document, nous cherchons à évaluer des approches fréquentistes et bayésiennes de rechange afin de faire des inférences à propos des fluctuations des composantes de la variance entre deux groupes indépendants d'intervieweurs d'enquête, et à fournir des conseils pratiques aux chercheurs d'enquête qui s'intéressent à ce type d'analyse.

Le document est structuré comme suit. À la section 2, nous présentons le cadre de modélisation général qui permet de faire ces comparaisons des composantes de l'écart intervieweurs pour les variables d'enquête normales et non normales (p. ex. binaire, compte), et nous passons en revue la documentation existante en comparant les approches fréquentistes et bayésiennes de l'estimation et de l'inférence, en décrivant les avantages et inconvénients de ces approches. Nous présentons ensuite une étude de simulation à la section 3, en évaluant la capacité des deux approches d'estimer efficacement les différences entre les composantes de la variance des deux groupes hypothétiques d'intervieweurs. La section 4 applique les deux approches aux données d'enquête réelles recueillies dans le cadre de la National Survey of Family Growth (NSFG) des États-Unis (Lepkowski et coll. 2010; Groves, Mosher, Lepkowski et Kirgis 2009). Enfin, la section 5 présente le mot de la fin, des suggestions aux praticiens et des orientations pour les recherches à venir. Nous incluons à l'annexe le code SAS, R et WinBUGS, que les lecteurs peuvent utiliser pour mettre en œuvre les deux approches.

2 Approches de rechange pour comparer les composantes de la variance dans les modèles linéaires généralisés hiérarchiques

Nous examinons d'abord une catégorie générale de modèles que les chercheurs d'enquête peuvent utiliser pour comparer les composantes de la variance dans différents groupes d'intervieweurs. Les modèles linéaires généralisés hiérarchiques (MLGH) sont des outils analytiques flexibles qui peuvent être utilisés pour modéliser des observations pour les variables d'enquête d'intérêt normales et non normales (p. ex. binaire, compte), où les observations recueillies par un même intervieweur ne peuvent pas être considérées comme indépendantes (Raudenbush et Bryk 2002; Goldstein 1995). Nous examinons des approches de rechange pour faire des inférences au sujet des composantes de l'écart intervieweurs dans une catégorie particulière de MLGH, où les composantes de l'écart intervieweurs pour les deux groupes indépendants d'intervieweurs définis par une caractéristique connue de l'intervieweur n'ont pas à être égales. Ce type de MLGH peut être écrit comme suit

$$\begin{aligned} &g\left(E\left[y_{ij} \mid u_i\right]\right) = \beta_0 + \beta_1 I\left(Groupe = 1\right)_i + u_{i(1)} I\left(Groupe = 1\right)_i + u_{i(2)} I\left(Groupe = 2\right)_i \\ &u_{i(1)} \sim N\left(0, \tau_1^2\right),\ u_{i(2)} \sim N\left(0, \tau_2^2\right), \end{aligned} \tag{2.1}$$

où $g(x)$ est la fonction de lien appariant une transformation de la valeur prévue de la variable dépendante, y_{ij}, à la combinaison linéaire des effets fixes et aléatoires (p. ex. $g(x) = log\left[x/(1-x)\right]$ pour une répartition Bernoulli présumée [résultat binaire], $g(x) = log(x)$ pour une répartition de Poisson présumée [résultat du compte]), i est un indice pour l'intervieweur, j est un indice pour le répondant associé à un seul intervieweur, et $I(\bullet)$ représente une variable indicatrice, équivalant à 1 si la

condition à l'intérieur des parenthèses est vraie et 0 autrement. Les effets aléatoires des intervieweurs du groupe 1, $u_{i(1)}$, sont présumés suivre une répartition normale, avec la moyenne 0 et la variance τ_1^2, tandis que les effets aléatoires des intervieweurs du groupe 2, $u_{i(2)}$, sont présumés suivre une répartition normale avec la moyenne 0 et la variance τ_2^2. D'autres répartitions peuvent être postulées pour les effets aléatoires, et le modèle général dans (2.1) peut tenir compte de la surdispersion de la variable dépendante observée par rapport à la répartition postulée pour cette variable. L'aspect clé de la spécification dans (2.1) est que les effets aléatoires pour différents groupes d'intervieweurs ont des variances *différentes*. Le paramètre des effets fixes β_1 dans (2.1) représente un effet fixe du groupe 1 sur le résultat par rapport au groupe 2 dans le MLGH, et des effets fixes d'autres covariables peuvent être inclus facilement. De même, des sous-groupes supplémentaires d'intervieweurs peuvent être considérés au moyen d'effets aléatoires supplémentaires $u_{i(\mathrm{k})}$, pour $k > 2$. L'intérêt analytique repose sur l'ampleur de la différence entre les composantes de la variance.

Les modèles de la forme dans (2.1) peuvent être appliqués lorsque les études méthodologiques sont conçues de manière à comparer deux groupes différents d'intervieweurs en ce qui concerne leurs composantes de la variance. Par exemple, il y a un débat dans la documentation sur les méthodes d'enquête en vue de déterminer si les intervieweurs devraient utiliser des interviews normalisées ou interactives. Les partisans de l'interview normalisée soutiennent que tous les intervieweurs devraient administrer les enquêtes exactement de la même façon, afin de permettre aux répondants de décoder les questions comme ils l'entendent (p. ex. Fowler et Mangione 1990). D'autres recherches ont démontré qu'une méthode d'interview plus flexible axée sur un style interactif pourrait accroître la compréhension des répondants des questions d'enquête et réduire l'erreur de mesure (p. ex. Schober et Conrad 1997). Pour tester l'hypothèse voulant qu'un style d'interview entraîne une plus faible variance entre intervieweurs, un chercheur peut randomiser les intervieweurs en deux groupes ayant reçu une formation dans les deux styles différents, recueillir des données d'enquête sur diverses variables, puis intégrer le modèle (2.1), y compris les variables indicatrices pour les deux groupes d'intervieweurs. Cette même approche pourrait être utilisée pour comparer les composantes de l'écart intervieweurs en deux groupes d'intervieweurs affectés aléatoirement à différents modes de collecte des données (p. ex. IPAO ou ITAO). À ce jour, aucune étude publiée n'a tenté d'effectuer ce genre de comparaisons, qui sont toutefois importantes pour comprendre les effets généraux de ces décisions relatives au plan sur l'erreur quadratique moyenne (EQM) des estimations d'enquête.

Les approches fréquentistes de l'estimation des paramètres dans les MLGH reposent sur diverses approches numériques ou théoriques de l'approximation de fonctions complexes de probabilité, en particulier pour les modèles comme (2.1), qui supposent des structures complexes des effets aléatoires (p. ex. Faraway 2006, chapitre 10; Molenberghs et Verbeke 2005). En général, les inférences sont basées sur ces approches approximatives axées sur la probabilité, qui comprennent la pseudo-vraisemblance résiduelle (qui diffère de l'approche de l'estimation du pseudo-maximum de vraisemblance élaborée par Binder (1983) pour les analyses fondées sur le plan des données d'enquêtes-échantillons complexes), la quasi-vraisemblance pénalisée et le maximum de vraisemblance basé sur une approximation de Laplace. Des travaux antérieurs ont donné des résultats de simulation favorables pour l'approche de pseudo-vraisemblance résiduelle, qui indiquent une estimation presque sans biais des composantes de la variance dans un MLGH comparativement au maximum de vraisemblance au moyen de l'approximation de

Laplace ou de la quadrature adaptative (Pinheiro et Chao 2006). Ces résultats ressemblent au cas de l'estimation du maximum de vraisemblance restreint (REML, pour *restricted maximum likelyhood*) dans un modèle pour une variable de résultat normalement réparti. Pour les variables de résultat binaire, les techniques de quasi-vraisemblance marginale ou pénalisée peuvent donner lieu à un biais vers le bas dans les estimations des paramètres et à des problèmes de convergence, et les approches exclusivement bayésiennes peuvent avoir des propriétés favorables dans ce cas (Browne et Draper 2006; Rodriguez et Goldman 2001). Nous considérons donc l'approche de pseudo-vraisemblance résiduelle dans les simulations et les applications présentées dans cette étude, et comparons cette approche à une approche exclusivement bayésienne.

Il existe deux approches pour faire une inférence à propos des différences entre les composantes de la variance dans le contexte fréquentiste. La première approche consiste à tester l'hypothèse nulle voulant que $\tau_1^2 = \tau_2^2$, par rapport à l'hypothèse de rechange voulant que $\tau_1^2 \neq \tau_2^2$. En principe, il s'agit d'un simple test d'hypothèse à effectuer au moyen de la méthodologie fréquentiste, puisque l'hypothèse nulle définit une contrainte d'égalité au lieu d'établir un paramètre pour une valeur à la limite d'un espace des paramètres. Le modèle sous l'hypothèse nulle est imbriqué dans le modèle sous l'hypothèse de rechange, où $\tau_2^2 = \tau_1^2 + k$. L'hypothèse nulle peut donc être réécrite $k = 0$, au lieu de $k \neq 0$. Un essai statistique est calculé par l'intégration d'une version limitée du modèle dans (2.1), les composantes de la variance de l'effet aléatoire dans les deux groupes étant réputées égales, puis par l'intégration du modèle à la forme plus générale dans (2.1). La différence positive entre les valeurs approximatives du logarithme du rapport de vraisemblance -2 de ces deux modèles est alors calculée, et appelée répartition du khi carré avec un degré de liberté.

La deuxième méthode nécessite le calcul de la différence entre les estimations pseudo-MV, $\hat{\tau}_1 - \hat{\tau}_2$, et un intervalle de confiance connexe de type Wald de 95 % pour la différence, donnée par $\hat{\tau}_1 - \hat{\tau}_2 \pm 1{,}96\sqrt{\hat{\text{var}}(\hat{\tau}_1) + \hat{\text{var}}(\hat{\tau}_2) - 2\hat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_2)}$. Cet intervalle nécessite des estimations asymptotiques des variances et des covariances des deux composantes estimées de la variance, qui sont calculées en fonction de la matrice d'Hessian (dérivée seconde) de la fonction objective utilisée pour la procédure d'estimation du maximum de vraisemblance. Si l'intervalle de Wald obtenu comprend zéro, on pourrait conclure que les données ne permettent pas de rejeter l'hypothèse nulle. Les intervalles de confiance pour les différences entre les composantes de la variance peuvent également être calculés au moyen des inversions des tests de vraisemblance des profils (p. ex. Viechtbauer 2007), mais le logiciel standard ne comprend pas d'options pour mettre en œuvre cette procédure (à notre connaissance).

Ces deux approches fréquentistes pour faire une inférence au sujet des différences entre les composantes de l'écart intervieweurs ont des limites. Lorsque le nombre d'intervieweurs dans chaque groupe est petit (disons moins de 30; voir Hox 1998, pour une discussion), les résultats asymptotiques pour le test du rapport des vraisemblances (Zhang et Lin 2010) risquent de ne plus tenir. Les méthodes fréquentistes (maximum de vraisemblance) ont également tendance à surévaluer la précision des estimations, puisqu'elles ne tiennent pas compte de l'incertitude des estimations des composantes de la variance (Carlin et Louis 2009, p. 335-336), qui est particulièrement problématique pour les petits échantillons (Goldstein 1995, p. 23). Les approches bayésiennes permettent aux analystes de disposer les répartitions a priori des composantes de la variance pour tenir compte de cette incertitude, contrairement aux approches fréquentistes. En outre, Molenberghs et Verbeke (2005, p. 277) soutiennent que les tests du rapport des vraisemblances ne devraient pas être utilisés pour tester des hypothèses lorsque les modèles

sont intégrés au moyen de méthodes de pseudo-vraisemblance. Les méthodes d'estimation approximative du maximum de vraisemblance peuvent également donner lieu à des estimations invalides (c.-à-d. négatives) des composantes de la variance dans ces modèles lorsque les composantes de la variance sont très petites. Les logiciels qui n'utilisent pas de procédures d'estimation limitant ces composantes de la variance à des valeurs supérieures à zéro répondent généralement à ce problème en établissant des estimations négatives des composantes de la variance égales à zéro (sans l'erreur type qui les accompagne), ce qui empêche le calcul de l'intervalle de confiance de type Wald susmentionné.

Une approche bayésienne pour intégrer les MLGH décrits dans (2.1) utilise l'échantillonneur de Gibbs basé sur les MCMC et la méthodologie d'échantillonnage-rejet d'adaptation (Gilks et Wild 1992) pour simuler des tirages de la répartition a posteriori pour les paramètres du modèle défini dans (2.1). En général, les répartitions a posteriori pour les paramètres dans un MLGH n'ont pas de formes de répartition connues et doivent être simulées (Gelman, Carlin, Stern et Rubin 2004, section 16.4). Les valeurs a priori diffuses et non informatives pour les effets fixes et les composantes de la variance dans (2.1) peuvent être précisées pour les simulations, afin de laisser les données fournir le plus d'information au sujet des répartitions a posteriori des paramètres (Gelman et Hill 2007; Gelman 2006, section 7). Cette approche donne lieu à des inférences basées sur les tirages simulés des répartitions a posteriori marginales des deux paramètres des effets fixes, les deux paramètres de la variance, les effets aléatoires des intervieweurs et toute fonction de ces paramètres. Cette étude s'intéresse à la répartition a posteriori marginale de la différence entre les variances des effets aléatoires de deux groupes d'intervieweurs définis par une caractéristique connue au niveau de l'intervieweur, calculée au moyen des tirages simulés des deux composantes de la variance.

Étant donné que les vérifications d'hypothèses traditionnelles ne sont pas utiles dans le contexte bayésien, l'inférence bayésienne sera axée sur la différence entre les composantes de l'écart intervieweurs. L'inférence pour la différence est basée sur plusieurs milliers de tirages pour les deux composantes de la variance de la répartition a posteriori combinée estimée au moyen de la formule d'échantillonnage de Gibbs. Pour chaque tirage d des deux composantes de la variance, la différence entre les composantes de la variance, définie comme $\tau_1^{2(d)} - \tau_2^{2(d)}$, peut être calculée. Les inférences seront donc basées sur la répartition marginale de ces différences, sans tenir compte des effets aléatoires des intervieweurs et des autres paramètres dérangeants. La médiane et les quantiles 0,025 et 0,975 (pour un ensemble crédible de 95 %) des différences simulées entre les deux composantes de la variance seront calculés en fonction du nombre réel de tirages de simulation des deux composantes de la variance à partir de la répartition a posteriori combinée estimée. Dans une analyse donnée, plusieurs milliers de tirages de la répartition a posteriori peuvent être générés au moyen de la formule d'échantillonnage de Gibbs, un grand nombre de tirages initiaux étant éliminés en tant que tirages de rodage, et le nombre réel de tirages de simulation sera calculé en fonction du nombre de tirages de rodage (Gelman et Hill 2007, chapitre 16). Si l'ensemble crédible de 95 % obtenu comprend 0, les données supporteront l'hypothèse des deux groupes ayant des composantes égales de la variance. Si l'ensemble crédible de 95 % ne comprend pas 0, les données supporteront l'hypothèse des deux groupes ayant des variances différentes, une médiane positive suggérant que le groupe 1 a la composante de variance plus élevée. L'inférence pour les deux effets fixes peut suivre une approche semblable.

Le fait de se concentrer sur les tirages des deux composantes de la variance à partir de la répartition a posteriori combinée complète (et leurs différences) et de faire fi des tirages des effets aléatoires des

intervieweurs et des effets fixes a pour conséquence d'intégrer ces autres paramètres de la répartition a posteriori combinée. Cette approche bayésienne offre donc une méthode pratique pour simuler les tirages à partir de la répartition marginale d'un paramètre complexe (la différence entre les deux composantes de la variance) et pour calculer un ensemble crédible de 95 % pour ce paramètre. Ces estimations peuvent également être obtenues dans l'approche fréquentiste, comme susmentionné, mais l'approche bayésienne n'exige pas de suppositions asymptotiques et comprend la variabilité des composantes de la variance estimée dans le calcul des ensembles crédibles de 95 % au moyen des tirages simulés.

Plusieurs (habituellement trois) chaînes de Markov peuvent être exécutées en parallèle selon l'algorithme d'échantillonnage itératif de Gibbs pour simuler des trajets aléatoires dans l'espace de la répartition a posteriori combinée. La statistique $\hat{R}$ de Gelman-Rubin, qui représente (à peu près) la racine carrée de la variance de l'amalgame des chaînes divisée par la variance moyenne dans la chaîne (Gelman et Rubin 1992), peut être utilisée pour évaluer la convergence (ou la combinaison) des chaînes pour chaque paramètre. Les valeurs inférieures à 1,1 dans cette statistique peuvent être considérées comme un indice de convergence des chaînes pour un paramètre donné. Les tirages a posteriori des paramètres peuvent être dérivés des trois chaînes combinées pour générer la taille finale de l'échantillon réel de tirages utilisés pour les inférences.

L'approche bayésienne susmentionnée a par ailleurs ses limites. La sélection des répartitions a priori utilisées pour calculer la répartition a posteriori pour les paramètres dans (2.1) est essentiellement arbitraire et dépend des choix d'un analyste donné et de la quantité d'information a priori disponible. De plus, le choix de la répartition a priori peut devenir essentiel lorsqu'il y a un petit nombre d'intervieweurs (disons moins de 20), où différentes distributions a priori peuvent entraîner des inférences très différentes au sujet des composantes de la variance (Lambert, Sutton, Burton, Abrams et Jones 2005); l'utilisation de l'information a priori au sujet des composantes de la variance peut accroître l'efficacité relative à l'utilisation de distributions a priori non informatives dans ces cas. L'erreur de spécification du modèle est également une nette possibilité, selon la variable de l'enquête modélisée, qui s'avère aussi une faiblesse de l'approche fréquentiste. La demande relative aux calculs peut également être problématique dans l'approche bayésienne (méthode d'échantillonnage de Gibbs) (Browne et Draper 2006), en particulier lorsque l'on souhaite faire des comparaisons des composantes de l'écart intervieweurs pour un grand nombre de variables d'enquête (avec des répartitions potentiellement différentes) et qu'il y a un nombre relativement élevé d'intervieweurs; ce n'est pas nécessairement aussi problématique compte tenu des progrès réalisés récemment en matière de rapidité du matériel et de l'efficacité des algorithmes. Enfin, les analystes ne sont peut-être pas à l'aise avec les logiciels disponibles pour les approches bayésiennes, ce qui fait qu'il pourrait y avoir une courbe d'apprentissage associée à la mise en œuvre de cette approche.

Plusieurs articles antérieurs ont comparé ces approches fréquentistes et bayésiennes de rechange au moyen d'études de simulation. Chaloner (1987) a envisagé des modèles d'analyse de la variance à un facteur avec effets aléatoires pour les données déséquilibrées (à l'instar du cas dans cette étude, où les intervieweurs ont des charges de travail variables), et observé des valeurs empiriques inférieures de l'EQM pour les modes a posteriori des composantes de la variance selon l'approche bayésienne et au moyen des distributions a priori non informatives par rapport à l'approche fréquentiste (maximum de vraisemblance). Van Tassell et Van Vleck (1996) ont déclaré que la formule d'échantillonnage de Gibbs (au moyen des répartitions préalables informatives ou non informatives) et le REML produisent tous deux des estimations empiriques non biaisées des composantes de la variance qui ont tendance à être très

semblables. Browne et Draper (2006) ont également constaté que les deux approches peuvent donner lieu à des estimations non biaisées, la nature plus « automatique » des approches fréquentistes étant un atout intéressant. Dans le contexte de la prédiction des moyennes pour les petites régions au moyen de modèles avec effets aléatoires des régions, Singh, Stukel et Pfeffermann (1998) ont déclaré que les approximations bayésiennes de l'EQM pour les prédictions ont de bonnes propriétés fréquentistes, mais que la méthode bayésienne a tendance à produire des biais fréquentistes et des EQM de prédiction plus marquées que les méthodes fréquentistes. Farrell (2000) a constaté que l'approche bayésienne entraînait des prédictions légèrement plus précises des proportions des petites régions, comportant peu de variations des taux de couverture entre les deux approches. Ugarte, Goicoa et Militino (2009) ont également conclu que les deux approches donnaient des résultats très similaires dans une application nécessitant la détection des zones à haut risque de maladie. Ces auteurs soulignent que la simplicité relative du calcul de l'approche fréquentiste est intéressante compte tenu de ces résultats. En général, d'après la documentation dans ce domaine, nous prévoyons un rendement similaire des deux méthodes dans le cas d'une comparaison des composantes de l'écart intervieweurs, et nous évaluons cette attente au moyen d'une étude de simulation (section 3).

Bien qu'il y ait de nombreuses procédures logicielles pour l'intégration de modèles multiniveaux et de composantes de la variance d'estimation au moyen des méthodes fréquentistes et bayésiennes (voir West et Galecki 2011 pour un examen), l'approche fréquentiste de la comparaison particulière des composantes de la variance dont il est question dans le présent document n'est facilement mise en œuvre que dans la procédure GLIMMIX de SAS/STAT (SAS 2010), au moyen de l'énoncé COVTEST avec l'option HOMOGENEITY (qui présume qu'une variable GROUP a été précisée dans l'énoncé RANDOM, indiquant différents groupes de grappes ayant des effets aléatoires découlant des différentes répartitions). Au moment de la rédaction, nous ne connaissons pas d'autres procédures permettant la mise en œuvre facile de l'approche comparative fréquentiste. Un exemple de code pouvant être utilisé pour l'intégration de ces modèles au moyen de la procédure GLIMMIX est disponible à l'annexe. L'approche bayésienne pour comparer les composantes de la variance peut être mise en œuvre dans le logiciel BUGS (inférence bayésienne au moyen de la méthode d'échantillonnage de Gibbs) (voir la bibliographie pour obtenir plus de détails). Nous incluons également dans l'annexe un exemple de code qui met en œuvre cette approche en ouvrant WinBUGS à partir de R dans l'annexe.

3 Étude de simulation

Nous avons réalisé une petite étude de simulation en vue d'examiner les propriétés empiriques de ces deux approches de rechange. Les données des deux variables d'enquête hypothétiques d'intérêt (répartition normale et répartition de Bernoulli) ont été simulées selon les deux modèles de super-population suivants :

$$\begin{aligned} y_{ij} &= 45 + 5 \times I\left(Groupe = 2\right)_i + u_{i(1)} I\left(Groupe = 1\right)_i + u_{i(2)} I\left(Groupe = 2\right)_i + \varepsilon_{ij} \\ u_{i(1)} &\sim N\left(0,1\right),\ u_{i(2)} \sim N\left(0,2\right),\ \varepsilon_{ij} \sim N\left(0,64\right) \end{aligned} \tag{3.1}$$

$$P\left(y_{ij}=1\right)=\frac{\exp\left[-1+u_{i(1)}I\left(Groupe=1\right)_i+u_{i(2)}I\left(Groupe=2\right)_i\right]}{1+\exp\left[-1+u_{i(1)}I\left(Groupe=1\right)_i+u_{i(2)}I\left(Groupe=2\right)_i\right]} \tag{3.2}$$
$$u_{i(1)}\sim N\left(0;0{,}03\right),\ u_{i(2)}\sim N\left(0;0{,}13\right).$$

La notation utilisée ici est conforme à celle qui a été utilisée à (2.1). Les valeurs de la deuxième variable de Bernoulli ont été générées pour les cas hypothétiques selon le modèle de régression logistique indiqué à (3.2). Pour obtenir la variable de Bernoulli observée, un tirage aléatoire a été obtenu à partir d'une répartition UNIFORME(0,1), et la variable a été établie à 1 si le tirage aléatoire était inférieur ou égal à la probabilité prédite, et à 0 autrement. Pour un groupe hypothétique d'intervieweurs à la fois, les effets aléatoires des intervieweurs ont été tirés, et les valeurs pour les cas de chaque intervieweur ont été générées en fonction du modèle précisé.

Nous avons généré 200 échantillons de cas hypothétiques et de données simulées pour chaque variable, 50 intervieweurs hypothétiques dans un groupe recueillant des données à partir de 50 cas hypothétiques chacun ($n=2\ 500$ pour chaque groupe d'intervieweurs). Nous avons ensuite généré 200 autres échantillons dans un scénario de petit échantillon, 20 intervieweurs dans chaque groupe recueillant des données à partir de 10 cas hypothétiques chacun ($n=200$ pour chaque groupe d'intervieweurs). Les choix des composantes de la variance à (3.1) correspondent aux corrélations intra-intervieweur de 0,015 et 0,030 pour les deux groupes hypothétiques d'intervieweurs, tandis que les choix des composantes de la variance à (3.2) correspondent aux corrélations intra-intervieweur de 0,009 et 0,038. Toutes ces valeurs seraient considérées comme plausibles dans un contexte d'enquête en personne ou par téléphone (West et Olson 2010). Les variations connues des composantes de la variance entre les groupes sont donc de 1 pour la variable normale, et de 0,1 pour la variable de Bernoulli.

Compte tenu de ces valeurs connues pour les composantes de l'écart intervieweurs dans la population hypothétique, nous avons appliqué chaque méthode décrite dans la section 2 [au moyen de valeurs uniformes diffuses et non informatives pour les composantes de la variance, selon les recommandations de Gelman (2006, section 7)] pour chaque échantillon hypothétique. Nous avons calculé les mesures empiriques suivantes à des fins de comparaison : 1) le biais empirique et relatif de l'estimateur; 2) l'EQM empirique de l'estimateur; 3) la couverture « fréquentiste » des intervalles de type Wald à 95 % (lorsque l'on utilise l'approche fréquentiste) et les ensembles crédibles à 95 % (lorsque l'on utilise l'approche bayésienne); et 4) les largeurs moyennes des intervalles de type Wald à 95 % et les ensembles crédibles. Le nombre d'intervalles de type Wald ne pouvant pas être calculés en raison des composantes estimées de la variance de 0 (sans erreur type connexe) a également été enregistré dans chaque cas. Toutes les simulations ont été effectuées au moyen de SAS, R et BUGS, et le code de simulation est disponible sur demande.

Le tableau 3.1 présente les résultats de l'étude de simulation. Les résultats suggèrent que pour les échantillons modérés à grands d'intervieweurs et de répondants, les deux approches donnent des estimateurs de la différence entre les composantes de la variance qui ont un biais relativement petit, comme prévu. On a découvert que l'approche fréquentiste donne des estimateurs ayant des valeurs empiriques plus petites de l'EQM. Ce n'est pas vraiment étonnant, étant donné la variabilité supplémentaire des estimations bayésiennes introduites par la prise en compte de l'incertitude des répartitions a priori des paramètres dans les distributions a priori non informatives. L'utilisation des distributions a priori informatives pourrait améliorer l'efficacité des estimations bayésiennes. Dans le

contexte d'un gros échantillon, les intervalles de confiance de 95 % et les ensembles crédibles calculés pour la différence entre les composantes de la variance semblent avoir des propriétés de couverture acceptables, l'approche bayésienne ayant un léger sous-dénombrement.

Tableau 3.1
Résultats de l'étude de simulation comparant les propriétés empiriques des approches fréquentistes et bayésiennes pour faire une inférence au sujet des différences entre les composantes de l'écart intervieweurs.

Tailles d'échantillon		Approche fréquentiste	Approche bayésienne
	Y Normale		
	Biais empirique	-0,0498	-0,0189
	Biais relatif	-4,98%	-1,89%
	EQM empirique	0,6546	0,8134
	IC/EC de 95%	0,960	0,920
	Largeur moyenne IC/EC 95%	3,1689	3,6283
50 intervieweurs / groupe	% des IC de Wald invalides	0,0%	--
50 cas / intervieweur	**Y Bernoulli**		
($n = 2\ 500$ / groupe)	Biais empirique	-0,0020	-0,0046
	Biais relatif	-2,0%	-4,6%
	EQM empirique	0,0029	0,0033
	IC/EC de 95%	0,938	0,940
	Largeur moyenne IC/EC 95%	0,2142	0,2372
	% des IC de Wald invalides	11,5%	--
	Y Normale		
	Biais empirique	-0,2341	-0,3508
	Biais relatif	-23,41%	-35,08%
	EQM empirique	6,9873	6,2869
	IC/EC de 95%	1,000	0,995
	Largeur moyenne IC/EC 95%	16,6313	18,3574
20 intervieweurs / groupe	% des IC de Wald invalides	54,0%	--
10 cas / intervieweur	**Y Bernoulli**		
($n = 200$ / groupe)	Biais empirique	-0,0348	-0,0196
	Biais relatif	-34,8%	-19,6%
	EQM empirique	0,0345	0,0861
	IC/EC de 95%	1,000	0,980
	Largeur moyenne IC/EC 95%	1,2604	1,7970
	% des IC de Wald invalides	65,5%	--

Fait intéressant, 11,5 % des intervalles de confiance de 95 % de type Wald n'ont pas pu être calculés pendant l'analyse du résultat binaire des gros échantillons, parce qu'une des composantes de la variance estimée était égale à zéro (sans erreur type). Ce taux de « rejet à la vérification » pour les intervalles de Wald s'aggravait beaucoup pour les deux variables des petits échantillons, où les deux méthodes

produisaient également des estimations ayant un biais négatif. L'approche fréquentiste peut donc fournir une estimation de la différence et des intervalles de confiance connexes qui fonctionnent bien dans les gros échantillons avec des variables réparties normalement, mais dans les petits échantillons ou même les échantillons de taille modérée à grande ayant des variables non normales, les intervalles simples de type Wald pouvant être calculés au moyen des logiciels standard peuvent échouer une bonne partie du temps. Cet échec est attribuable au fait que la matrice de Hessian n'est pas inversible lorsqu'une composante de variance estimée est établie à zéro (c.-à-d. que la vraisemblance ne peut pas être estimée par un quadratique). Dans l'ensemble, ces résultats de simulation suggèrent donc que : 1) les deux approches auront un rendement semblable lorsqu'elles seront appliquées à des données d'enquête réelles avec des échantillons modérés à gros d'intervieweurs et de répondants; 2) l'approche bayésienne pourrait être la meilleure option si les intervalles (ou ensembles crédibles) pour la différence sont souhaités; et 3) il faut faire preuve de prudence lorsque l'on applique l'une ou l'autre des méthodes aux échantillons relativement petits d'intervieweurs et de répondants.

4 Application : La National Survey of Family Growth (NSFG) des États-Unis

Nous appliquons maintenant les approches fréquentistes et bayésiennes aux données d'enquête réelles recueillies au septième cycle de la NSFG (de juin 2006 à juin 2010). Le plan original de ce cycle de la NSFG (Groves et coll. 2009) a nécessité 16 trimestres de collecte des données d'un échantillon continu qui était représentatif à l'échelle nationale lorsqu'il a été terminé en juin 2010. Les données analysées dans ce document ont été recueillies à partir d'un échantillon national de 11 609 femmes de 15 à 44 ans, par 87 intervieweuses (aux tailles d'échantillons variables pour chaque intervieweuse). Pour obtenir plus de détails sur le plan et le fonctionnement du septième cycle de la NSFG, voir Lepkowski et coll. (2010) ou Groves et coll. (2009).

Pour chacune des 87 intervieweuses, l'information est disponible concernant l'âge (47,1 % ont 55 ans ou plus), les années d'expérience (43,7 % ont au moins cinq ans d'expérience), le nombre d'enfants (33,3 % ont deux enfants ou plus), l'état matrimonial (19,5 % n'ont jamais été mariées), les autres emplois (46,0 % avaient un autre emploi), les études postsecondaires (57,5 % ont obtenu un diplôme d'un programme collégial de quatre ans), l'expérience antérieure dans le cadre de la NSFG (82,8 % avaient travaillé aux cycles précédents) et l'appartenance ethnique (81,6 % sont blanches). Ces caractéristiques observables au niveau de l'intervieweuse serviront à diviser les intervieweuses en deux groupes (en l'absence d'une expérience randomisée idéale, comme décrit à la section 2).

Pour chacune des 11 609 répondantes, on mesure la parité (ou nombre de naissances vivantes) et un indicateur de l'activité sexuelle actuelle (indiquée par au moins un partenaire masculin actuel ou au moins un partenaire masculin au cours des 12 derniers mois) en vue d'une analyse. Bien que ces mesures semblent simples, les concepts mesurés peuvent être communiqués différemment par différentes intervieweuses (ce qui donne lieu à l'écart intervieweuses). La principale question analytique vise à

déterminer si ces différents groupes d'intervieweuses ont des composantes de la variance considérablement différentes pour ces variables d'enquête en particulier.

Nous examinons d'abord un MLGH pour la variable de la parité. Supposons que Y soit une variable aléatoire de Poisson avec un paramètre λ. Nous autorisons la surdispersion (ou dispersion extra-Poisson) dans Y, ce qui est très fréquent dans les variables de dénombrement (par exemple, la parité moyenne pour l'échantillon de 11 609 femmes est de 1,19, et la variance des valeurs mesurées de la parité est de 1,99). En nous appuyant sur Hilbe (2007) et Durham, Pardoe et Vega (2004), nous supposons que $\lambda = r\mu$, où r est une variable aléatoire $\text{GAMMA}\left(\alpha^{-1}, \alpha^{-1}\right)$. On constate alors que Y a une répartition binomiale négative avec une moyenne μ et un paramètre d'échelle α :

$$E(Y) = E(\lambda) = E(r\mu) = \mu E(r) = \mu$$
$$\text{var}(Y) = E(\lambda) + \text{var}(\lambda) = E(r\mu) + \text{var}(r\mu) = \mu E(r) + \mu^2 \text{var}(r) = \mu(1 + \alpha\mu)$$

Nous précisons un MLGH pour la valeur observée de la parité pour la répondante j interviewée par l'intervieweuse i, y_{ij}, comme suit :

$$\begin{aligned}
&y_{ij} \sim Poisson(\lambda_i),\ \lambda_i = r_i \mu_i \\
&r_i \sim Gamma\left(\alpha^{-1}, \alpha^{-1}\right) \\
&\log(\mu_i) = \beta_0 + \beta_1 I(Groupe = 1)_i + u_{i(1)} I(Groupe = 1)_i + u_{i(2)} I(Groupe = 2)_i \\
&u_{i(1)} \sim N\left(0, \tau_1^2\right), \qquad u_{i(2)} \sim N\left(0, \tau_2^2\right).
\end{aligned} \tag{4.1}$$

Dans ce modèle multiniveau de régression binomiale négative, $\exp(\beta_0)$ représente la parité prévue pour le groupe 2, $\exp(\beta_1)$ représente le changement multiplicatif prévu pour le groupe 1 par rapport au groupe 2, $u_{i(1)}$ est un effet aléatoire associé à l'intervieweuse i dans le groupe 1, et $u_{i(2)}$ est un effet aléatoire associé à l'intervieweuse i du groupe 2.

Ensuite, nous examinons un MLGH pour l'indicateur binaire de l'activité sexuelle actuelle. Supposons que $z_{ij} = 1$ si une répondante j indique l'activité sexuelle actuelle à l'intervieweuse i, et 0 autrement. Nous précisons le modèle suivant pour cet indicateur binaire :

$$\begin{aligned}
&z_{ij} \sim Bernoulli(p_i) \\
&\ln\left[p_i/(1 - p_i)\right] = \beta_0 + \beta_1 I(Groupe = 1)_i + u_{i(1)} I(Groupe = 1)_i + u_{i(2)} I(Groupe = 2)_i \\
&u_{i(1)} \sim N\left(0, \tau_1^2\right), \qquad u_{i(2)} \sim N\left(0, \tau_2^2\right).
\end{aligned} \tag{4.2}$$

Dans ce modèle, $\exp(\beta_0)$ représente les probabilités prévues d'activité sexuelle actuelle pour le groupe 2, $\exp(\beta_1)$ représente la variation multiplicative prévue des probabilités d'activité sexuelle actuelle pour le groupe 1 par rapport au groupe 2, $u_{i(1)}$ est un effet aléatoire associé à l'intervieweuse i dans le groupe 1, et $u_{i(2)}$ est un effet aléatoire associé à l'intervieweuse i dans le groupe 2.

Nous intégrons les modèles (4.1) et (4.2) au moyen des deux approches décrites à la section 2. Pour l'approche fréquentiste, d'après les recommandations de la documentation mentionnée à la section 2, nous avons estimé les paramètres dans ces modèles au moyen de l'estimation de la pseudo-vraisemblance résiduelle (PVR), telle que mise en œuvre dans la procédure GLIMMIX du logiciel SAS/STAT. Toutes les analyses fréquentistes présentées dans cette section ont été répétées au moyen d'une quadrature adaptative pour estimer les fonctions de vraisemblance, et les principaux résultats n'ont pas changé; de plus, l'utilisation de la quadrature adaptative a donné des délais d'estimation plus longs.

Pour l'approche bayésienne, les répartitions a priori non informatives suivantes pour ces paramètres ont été utilisées. Ces répartitions a priori ont été sélectionnées en fonction d'une combinaison d'estimations à partir de l'intégration initiale du modèle naïf, ainsi que des recommandations de Gelman et Hill (2007) et de Gelman (2006, section 7) pour les répartitions a priori correctes mais non informatives pour les paramètres de la variance dans les modèles hiérarchiques ayant un nombre raisonnablement élevé de groupes (c.-à-d. plus de cinq) ou d'intervieweuses, dans le contexte actuel :

$$\begin{aligned} &\beta_0 \sim N(0,100) && \beta_1 \sim N(0,100) \\ &\tau_1^2 \sim Uniforme(0,10) && \tau_2^2 \sim Uniforme(0,10) \\ &\ln(\alpha) \sim N(0,100). \end{aligned}$$

Les valeurs a priori non informatives pour les effets fixes et le paramètre d'échelle (après transformation par le logarithme naturel) pour la variable du dénombrement binomial négatif (parité) indiquent qu'on s'attend à ce que ces paramètres soient quelque part dans la plage (-10, 10), tandis que les valeurs a priori non informatives pour les composantes de la variance sont des répartitions uniformes dans la plage (0, 10). Étant donné les estimations naïves initiales des effets fixes de -1 à 1 et les estimations initiales du paramètre d'échelle (non transformé) et des composantes de la variance de 0 à 5, ces valeurs a priori sont relativement diffuses, exprimant peu de connaissances antérieures au sujet de ces paramètres et laissant les données disponibles de la NSFG fournir la majeure partie de l'information. Les études antérieures comparant les composantes de l'écart intervieweurs pour des variables semblables sur le dénombrement pourraient également être utilisées dans les applications générales de cette technique afin de préciser des répartitions a priori plus informatives. Il est également important de souligner que le logiciel BUGS utilise des variances inversées pour la répartition normale, ce qui veut dire que 0,01 et les inverses des composantes de la variance seront indiqués dans les fonctions de la répartition normale (par exemple, le code WinBUGS utilisé pour les analyses est disponible à l'annexe).

Le tableau 4.1 présente des statistiques descriptives pour les intervieweuses dans chacun des groupes définis par les huit caractéristiques au niveau de l'intervieweuse. Ces statistiques descriptives comprennent le nombre d'intervieweuses dans chaque groupe (sur 87 au total), ainsi que la moyenne, l'écart-type (Éc.T.) et la plage pour le nombre de cas (tailles d'échantillons) affectés à chaque intervieweuse.

Tableau 4.1
Statistiques descriptives pour les intervieweuses de la NSFG dans chaque groupe défini par les huit caractéristiques au niveau de l'intervieweuse

	Nombre d'intervieweuses	Taille d'échantilllon totale	Taille d'échantilllon moyenne	Éc.T. des tailles d'échantillon	Étendue des tailles d'échantillon
Âge (Années)					
< 54	46	5 888	128,00	113,29	(18, 554)
55 et +	41	5 721	139,54	132,67	(12, 532)
Expérience					
< 5 ans	49	6 062	123,71	126,65	(12, 554)
5 ans et +	38	5 547	145,97	116,71	(18, 507)
Nbr. d'enfants					
< 2	58	7 756	133,72	113,28	(18, 532)
2 et +	29	3 853	132,86	140,53	(12, 554)
Déjà mariée					
Oui	70	9 923	141,76	129,00	(17, 554)
Non	17	1 686	99,18	83,49	(12, 377)
Autre emploi					
Non	47	5 406	115,02	95,49	(12, 532)
Oui	40	6 203	155,08	145,92	(17, 554)
Diplôme collégial					
Non	37	4 528	122,38	87,97	(18, 409)
Oui	50	7 081	141,62	142,71	(12, 554)
NSFG avant					
Non	15	1 155	77,00	39,17	(20, 166)
Oui	72	10 454	145,19	130,29	(12, 554)
Ethnicité					
Autre	16	1 781	111,31	75,53	(20, 297)
Blanche	71	9 828	138,42	130,35	(12, 554)

Les statistiques descriptives dans le tableau 4.1 indiquent une variance considérable des tailles des échantillons attribuées aux intervieweuses. Une approche de modélisation traitant les effets des intervieweurs comme des valeurs fixes ne conviendrait probablement pas à ces données, compte tenu des petites tailles pour certaines des intervieweuses (ce qui pourrait donner lieu à des estimations instables pour certaines intervieweuses). Au lieu de cela, une approche de modélisation empruntant de l'information à plusieurs intervieweuses (traitant les effets des intervieweuses comme des valeurs aléatoires) donnerait des estimations plus stables des moyennes pour chaque intervieweuse. Nous constatons également que pour trois des caractéristiques observables des intervieweuses (déjà mariée, expérience relative à la NSFG et ethnicité), un des deux groupes a moins de 20 intervieweuses, ce qui n'est pas idéal pour une estimation fiable des composantes de la variance (Hox 1998). Compte tenu des résultats de la simulation pour les

petits échantillons (section 3), nous tenons compte des effets des groupes sociodémographiques dans nos analyses.

De simples examens des répartitions des moyennes des mesures de la parité observée pour les intervieweuses dans chaque groupe sont présentés à la figure 4.1 ci-après, afin d'obtenir une première impression de l'ampleur de l'écart intervieweurs dans chaque groupe. La figure 4.1 présente des boîtes à pattes côte à côte des moyennes des intervieweuses pour la variable de la parité pour chaque groupe, la moyenne étant pondérée par l'attribution des tailles des échantillons, ainsi que la répartition globale des 11 609 mesures de la parité dans l'ensemble de données complet.

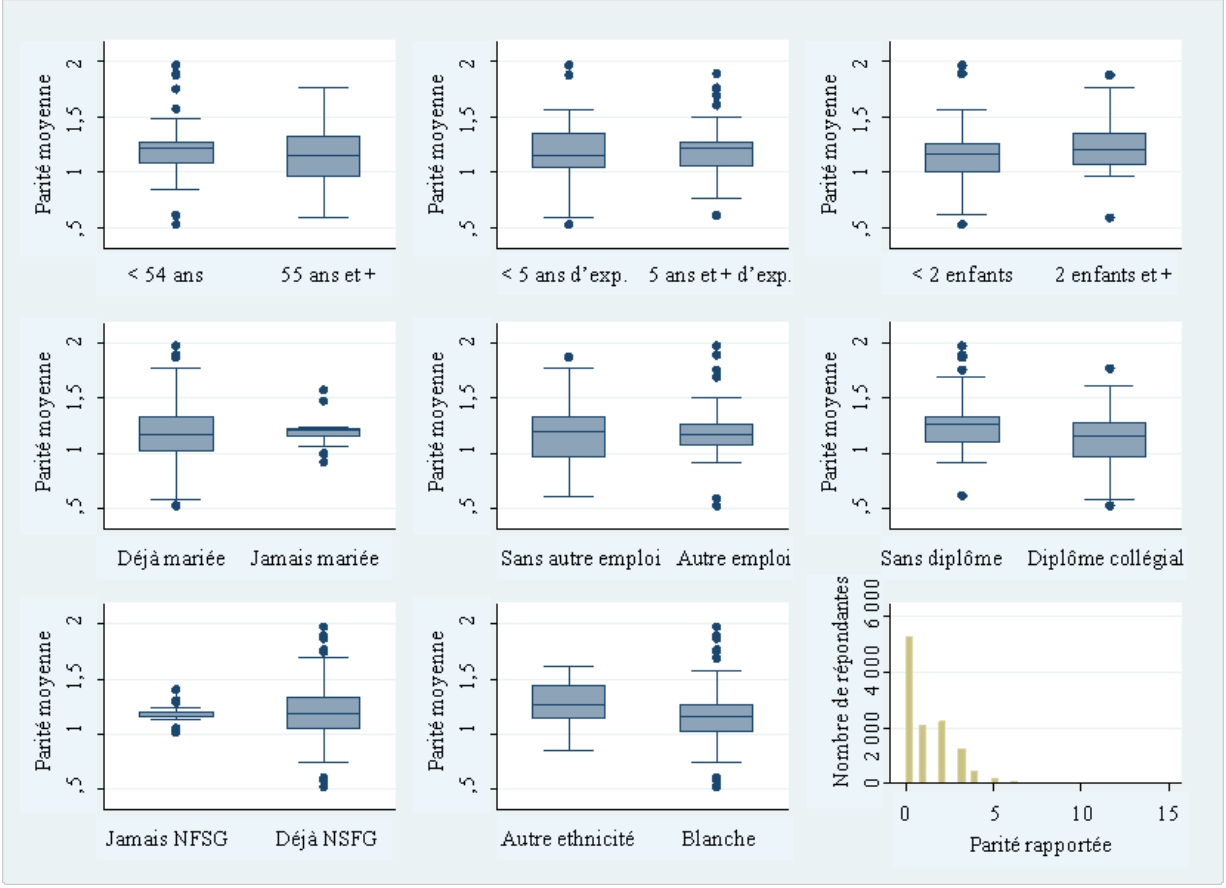

Figure 4.1 Répartitions des moyennes observées pour les intervieweuses de chaque groupe, les moyennes des intervieweuses étant pondérées par l'attribution de la taille de l'échantillon, ainsi que la répartition globale des mesures déclarées de la parité.

Les répartitions des moyennes des valeurs mesurées de la parité pour les intervieweuses dans la figure 4.1 donnent une première impression des groupes qui ont tendance à diverger pour ce qui est des composantes de l'écart intervieweurs. Le groupe d'intervieweuses qui n'ont jamais été mariées semble avoir une variance réduite, à l'instar du groupe d'intervieweuses sans expérience préalable relative à la NSFG. Les boîtes à pattes suggèrent également que les groupes ne varient pas beaucoup en ce qui concerne les moyennes de la parité, ce qui est rassurant (c.-à-d. que des groupes différents d'intervieweuses ne produisent pas de moyennes marginales différentes pour l'estimation d'intérêt). Enfin,

la répartition des valeurs observées de la parité pour la totalité des 11 609 répondantes a l'apparence prévue pour une variable mesurant certains événements relativement rares (naissances vivantes), avec une moyenne de 1,19 et une variance de 1,99.

Nous considérons ensuite les répartitions des proportions de femmes indiquant l'activité sexuelle actuelle parmi les intervieweuses dans chaque groupe (figure 4.2).

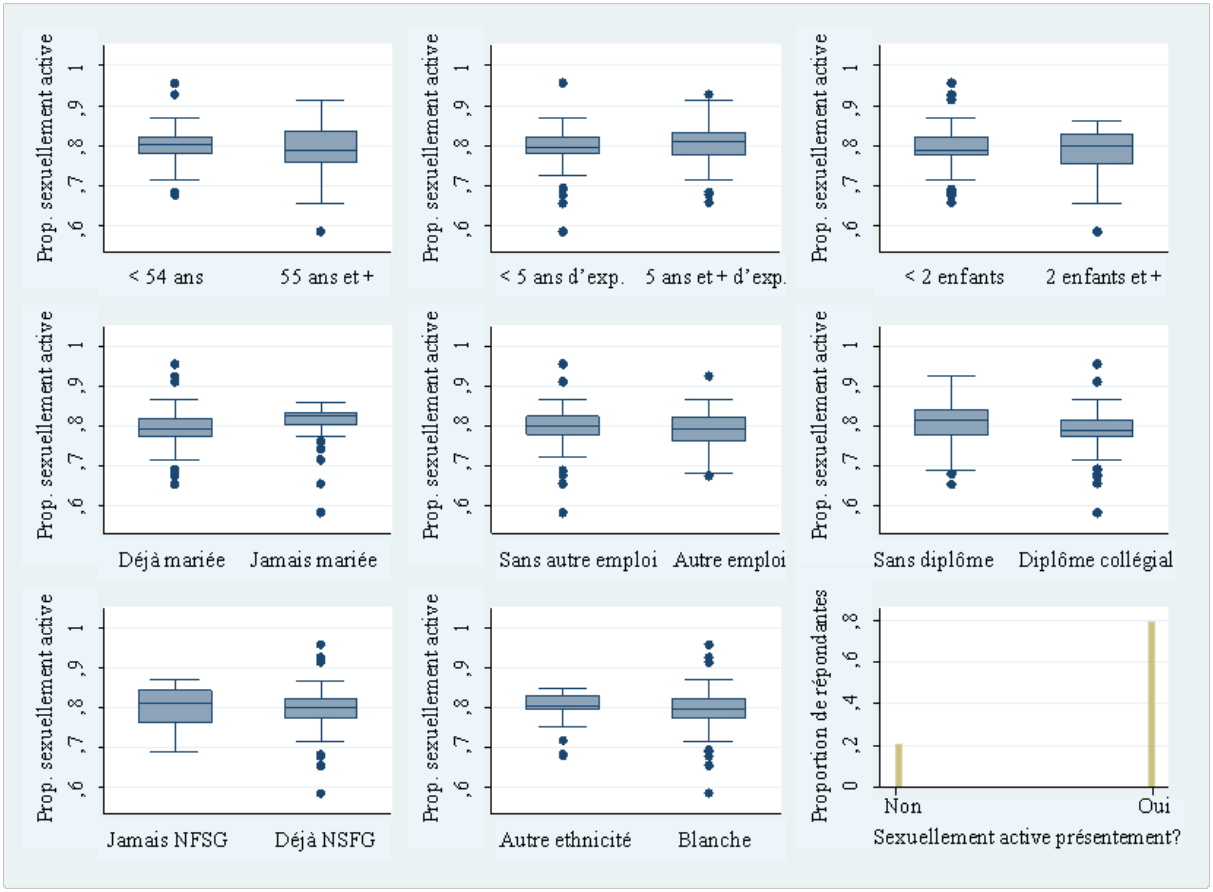

Figure 4.2 Répartitions des proportions observées de répondantes indiquant l'activité sexuelle actuelle pour les intervieweuses dans chaque groupe, les moyennes des intervieweuses étant pondérées par l'attribution de la taille de l'échantillon, ainsi que la répartition globale de l'indicateur d'activité sexuelle.

Nous voyons moins d'indices de variations de l'écart intervieweurs entre les groupes en général pour cette proportion, par rapport à la parité moyenne. Environ 80 % des répondantes ont indiqué qu'elles étaient actuellement actives sexuellement.

Le tableau 4.2 présente des estimations des paramètres dans chaque modèle binomial négatif pour la variable de la parité mesurée en fonction des deux approches analytiques. Ce tableau présente également les résultats des tests du rapport des vraisemblances comparant les deux composantes de l'écart intervieweurs (pour chaque paire de groupes) dans le contexte de l'approche fréquentiste, et ensembles crédibles de 95 % pour la différence entre les deux composantes de la variance dans le contexte de l'approche bayésienne.

Tableau 4.2
Estimations des paramètres des modèles de régression binomiale négative pour la parité et les comparaisons des composantes de l'écart intervieweurs dans le contexte des approches analytiques de rechange fréquentistes et bayésiennes.

	Approche fréquentiste (SAS PROC GLIMMIX)				Approche bayésienne (WinBUGS)			
Groupe de variables de l'intervieweuse	$\hat{\beta}_0$ **(ET)/** $\hat{\beta}_1$ **(ET)**	$\hat{\alpha}$ **(ET)**	$\hat{\tau}_1^2$ **(ET)/** $\hat{\tau}_2^2$ **(ET)**	**Test du rapport de vrais. :** $\tau_1^2 = \tau_2^2$	$\hat{\beta}_0$ **(Éc.T.)/** $\hat{\beta}_1$ **(Éc.T.)**	$\hat{\alpha}$ **(Éc.T.)**	$\hat{\tau}_1^2$ **(Éc.T.)/** $\hat{\tau}_2^2$ **(Éc.T.)**	**95% EC :** $\tau_1^2 - \tau_2^2$
Âge (1 = < 54 ans, 2 = 55 ans et +)	0,185(0,031)/ -0,007(0,043)	0,538 (0,018)	0,026(0,009)/ 0,024(0,008)	χ_1^2 =0,03, p= 0,873	0,183(0,033)/ -0,003(0,046)	0,685 (0,024)	0,025(0,010)/ 0,024(0,009)	(-0,026; 0,028)
Expérience (1 = < 5 ans, 2 = 5 ans et +)	0,201(0,033)/ -0,036(0,044)	0,537 (0,018)	0,024(0,008)/ 0,027(0,010)	χ_1^2 =0,04, p= 0,835	0,197(0,034)/ -0,031(0,045)	0,694 (0,027)	0,024(0,009)/ 0,027(0,011)	(-0,032; 0,024)
Nbr. d'enfants (1 = < 2, 2 = 2 et +)	0,254(0,036)/ -0,109(0,044)	0,537 (0,018)	0,023(0,007)/ 0,022(0,009)	χ_1^2 =0,01, p= 0,926	0,253(0,038)/ -0,109(0,045)	0,692 (0,025)	0,023(0,007)/ 0,023(0,012)	(-0,032; 0,024)
Déjà mariée (1 = Oui, 2 = Non)	0,184(0,029)/ -0,001(0,039)	0,537 (0,018)	0,030(0,008)/ 0,000(S.O.)*	χ_1^2 =5,41, p= 0,020	0,181(0,037)/ 0,004(0,045)	0,694 (0,025)	0,030(0,008)/ 0,003 (0,007)	(0,002; 0,048)
Autre emploi (1 = Oui, 2 = Non)	0,186(0,031)/ -0,009(0,043)	0,538 (0,018)	0,022(0,009)/ 0,027(0,008)	χ_1^2 =0,15, p= 0,699	0,188(0,032)/ -0,010(0,044)	0,688 (0,025)	0,020(0,010)/ 0,028(0,010)	(-0,036; 0,021)
Diplôme collégial (1 = Oui, 2 = Non)	0,242(0,031)/ -0,108(0,042)	0,538 (0,018)	0,023(0,008)/ 0,022(0,008)	χ_1^2 < 0,01, p= 0,963	0,240(0,032)/ -0,106(0,044)	0,693 (0,024)	0,024(0,009)/ 0,021(0,010)	(-0,025; 0,030)
NSFG avant (1 = Oui, 2 = Non)	0,174(0,035)/ 0,010(0,043)	0,537 (0,018)	0,031(0,008)/ 0,000(S.O.)*	χ_1^2 = 8,26, p= 0,004	0,169(0,036)/ 0,013(0,045)	0,692 (0,026)	0,030(0,008)/ 0,001(0,005)	(0,006; 0,050)
Ethnicité (1 = Blanche, 2 = Autre)	0,217(0,046)/ -0,044(0,052)	0,537 (0,018)	0,027(0,007)/ 0,018(0,011)	χ_1^2 =0,38, p= 0,536	0,220(0,051)/ -0,050(0,058)	0,690 (0,025)	0,026(0,008)/ 0,020(0,017)	(-0,045; 0,027)

* PROC GLIMMIX indiquait que la matrice estimée de variance-covariance des effets aléatoires n'était pas définie positive, et que l'estimation était établie à zéro à cause que l'estimation de la PVR de la composante de variance était négative. Le même résultat s'est produit lorsque l'on a utilisé la quadrature adaptative au lieu de la PVR.

Notes : Les estimations suivant l'approche bayésienne sont des médianes des tirages de répartitions a posteriori. ET = erreur-type asymptotique. Éc.T. = écart type des tirages de la répartition a posteriori. EC = ensemble crédible.

Conformément à notre étude de simulation à la section 3, les résultats dans le tableau 4.2 indiquent qu'il n'est pas rare que l'approche fréquentiste donne des estimations négatives des composantes de l'écart intervieweurs (qui amène SAS PROC GLIMMIX à établir les estimations égales à zéro, et à ne pas déclarer les erreurs types estimées pour les estimations), en particulier pour les groupes ayant de petits

échantillons d'intervieweurs. Dans deux cas, il s'ensuit une importante statistique sur le test du rapport des vraisemblances, qui pourrait suggérer que les deux composantes de la variance sont différentes. En revanche, l'approche bayésienne produit de très petites estimations des composantes de la variance, et un ensemble crédible de 95 % pour la différence entre les composantes de la variance. Par exemple, dans les cas de l'état matrimonial et de l'expérience préalable relative à la NSFG, nous voyons des estimations qui sont conformes à la figure 4.1, ce qui porte à croire qu'il y a une variance considérablement plus faible dans les mesures de la parité dans le groupe d'intervieweuses jamais mariées et le groupe d'intervieweuses sans expérience. Les ensembles crédibles pour les différences dans ces deux cas cadrent avec les tests fréquentistes, mais les limites inférieures de ces ensembles sont très proches de zéro, ce qui suggère que les différences, bien qu'importantes, ne sont peut-être pas très marquées. Nous considérons qu'il s'agit d'un avantage de l'approche bayésienne.

L'approche bayésienne donne des erreurs-types un peu plus importantes (ou des écarts-types a posteriori) pour les estimations des paramètres dans presque tous les cas, ce qui indique l'incertitude des estimations des paramètres qui est prise en compte par les répartitions a priori. L'utilisation des distributions a priori non informatives dans ce cas, qui pourrait entraîner une répartition a posteriori dominée par la fonction de vraisemblance, est la raison probable de la similarité dans ces mesures de l'incertitude, et des distributions a priori plus informatives pourraient augmenter l'efficacité des estimations bayésiennes. Les estimations des paramètres individuels et les inférences correspondantes connexes sont généralement très semblables lorsque l'on suit les deux approches, comme le suggère la documentation à la section 2, et les effets fixes estimés suggèrent que les groupes différents d'intervieweuses n'ont pas tendance à recueillir des mesures différentes pour la variable de la parité. Fait intéressant, les deux approches conviennent que les intervieweuses ayant moins d'enfants et/ou un diplôme collégial de quatre ans ont tendance à recueillir des mesures plus faibles pour la variable de la parité, mais ces différences pourraient bien être attribuables à d'autres covariables non prises en compte dans ces analyses. Enfin, nous voyons des estimations légèrement différentes du paramètre d'échelle binomiale négative selon les deux approches. Il fallait s'y attendre, puisque l'approche bayésienne utilise les médianes des répartitions a posteriori, tandis que l'approche fréquentiste utilise les modes des fonctions de vraisemblance. De plus, les répartitions a posteriori ne sont pas exactement égales aux fonctions de vraisemblance lorsque les distributions a priori appropriées sont utilisées. Les estimations fréquentistes du paramètre d'échelle étaient beaucoup plus proches des estimations bayésiennes lorsque l'on utilise la quadrature adaptative avec cinq points de quadrature pour estimer les vraisemblances binomiales négatives (résultats non montrés); les inférences fréquentistes pour les autres paramètres n'ont pas changé lorsqu'on utilisait cette méthode d'estimation de rechange.

Nous avons répété ces analyses pour l'indicateur d'activité sexuelle actuelle. Le tableau 4.3 présente les paramètres estimés des modèles de régression logistique multiniveaux suivant chacune des deux approches. Conformément à la figure 4.2, ces analyses ne révèlent aucun indice de différences entre les divers groupes d'intervieweuses pour les composantes de la variance ou les valeurs prévues de ce résultat. Les inférences étaient encore une fois très semblables lorsque l'on suivait les deux approches, et les variances des composantes de la variance estimée étaient encore une fois légèrement plus marquées selon l'approche bayésienne.

Tableau 4.3
Estimations des paramètres dans les modèles de régression logistique pour l'activité sexuelle actuelle et comparaisons des composantes de l'écart intervieweurs d'après les approches analytiques fréquentistes et bayésiennes de rechange.

	Approche fréquentiste (SAS PROC GLIMMIX)			Approche bayésienne (WinBUGS)		
Groupe de variables de l'intervieweuse	$\hat{\beta}_0$ **(ET)/** $\hat{\beta}_1$ **(ET)**	$\hat{\tau}_1^2$ **(ET)/** $\hat{\tau}_2^2$ **(ET)**	**Test du rapport de vrais. :** $\tau_1^2 = \tau_2^2$	$\hat{\beta}_0$ **(Éc.T.)/** $\hat{\beta}_1$ **(Éc.T.)**	$\hat{\tau}_1^2$ **(ET)/** $\hat{\tau}_2^2$ **(ET)**	**95% EC:** $\tau_1^2 - \tau_2^2$
Âge (1 = < 54 ans, 2 = 55 ans et +)	1,333 (0,066) / 0,032 (0,076)	0,008 (0,013) / 0,045 (0,024)	$\chi_1^2 = 2{,}05$, $p = 0{,}153$	1,344 (0,055) / 0,024 (0,066)	0,009 (0,013) / 0,046 (0,028)	(-0,107; 0,016)
Expérience (1 = < 5 ans, 2 = 5 ans et +)	1,378 (0,064) / -0,050 (0,073)	0,004 (0,017) / 0,037 (0,020)	$\chi_1^2 = 1{,}52$, $p = 0{,}217$	1,384 (0,051) / -0,061 (0,064)	0,005 (0,017) / 0,039 (0,023)	(-0,087; 0,024)
Nbr. d'enfants (1 = < 2, 2 = 2 et +)	1,362 (0,080) / -0,015 (0,088)	0,022 (0,015) / 0,033 (0,024)	$\chi_1^2 = 0{,}16$, $p = 0{,}689$	1,363 (0,059) / -0,012 (0,070)	0,024 (0,016) / 0,037 (0,030)	(-0,094; 0,037)
Déjà mariée (1 = Oui, 2 = Non)	1,387 (0,130) / -0,045 (0,134)	0,020 (0,012) / 0,048 (0,041)	$\chi_1^2 = 0{,}58$, $p = 0{,}447$	1,398 (0,090) / -0,053 (0,097)	0,021 (0,013) / 0,051 (0,055)	(-0,180; 0,035)
Autre emploi (1 = Oui, 2 = Non)	1,374 (0,043) / -0,046 (0,072)	0,026 (0,016) / 0,024 (0,020)	$\chi_1^2 = 0{,}01$, $p = 0{,}927$	1,381 (0,045) / -0,051 (0,065)	0,029 (0,019) / 0,022 (0,022)	(-0,055; 0,063)
Diplôme collégial (1 = Oui, 2 = Non)	1,388 (0,051) / -0,063 (0,071)	0,016 (0,014) / 0,035 (0,022)	$\chi_1^2 = 0{,}60$, $p = 0{,}439$	1,394 (0,052) / -0,072 (0,064)	0,014 (0,016) / 0,038 (0,024)	(-0,079; 0,033)
NSFG avant (1 = Oui, 2 = Non)	1,363 (0,103) / -0,012 (0,111)	0,020 (0,012) / 0,069 (0,055)	$\chi_1^2 = 1{,}20$, $p = 0{,}273$	1,381 (0,113) / -0,024 (0,118)	0,021 (0,013) / 0,083 (0,084)	(-0,301; 0,019)
Ethnicité (1 = Blanche, 2 = Autre)	1,354 (0,077) / -0,004 (0,088)	0,024 (0,014) / 0,032 (0,031)	$\chi_1^2 = 0{,}05$, $p = 0{,}816$	1,365 (0,080) / -0,013 (0,088)	0,025 (0,015) / 0,032 (0,044)	(-0,131; 0,044)

Notes : Les estimations suivant l'approche bayésienne sont des médianes de tirages de répartitions a posteriori. ET = erreur-type asymptotique. Éc.T. = écart-type des tirages de la répartition a posteriori. EC = ensemble crédible.

5 Mot de la fin

Le présent document a examiné les méthodes fréquentistes et bayésiennes pour comparer les composantes de l'écart intervieweurs pour les questions d'enquête non réparties normalement entre deux groupes indépendants d'intervieweurs d'enquête. Les méthodes sont basées sur une catégorie flexible de modèles linéaires généralisés hiérarchiques (les MLGH), qui permettent aux composantes de la variance pour deux groupes absolument exclusifs d'intervieweurs de varier, et des approches inférentielles de rechange basées sur ces modèles. Les résultats d'une étude de simulation suggèrent que les deux approches ont peu de biais empirique, des valeurs empiriques comparables de l'EQM et une bonne couverture pour des échantillons moyens ou gros d'intervieweurs et de répondants. Les analyses des données réelles de la National Survey of Family Growth (NSFG) des États-Unis suggèrent que les inférences basées sur les deux approches ont tendance à se ressembler. Nous constatons que le rendement similaire de ces deux approches est une bonne nouvelle pour les chercheurs d'enquête, en ce que les

fréquentistes et les bayésiens disposent des outils nécessaires pour analyser ce problème et tireront des conclusions similaires.

Quelques distinctions subtiles entre les deux approches sont ressorties des analyses, en particulier en ce qui concerne la taille des échantillons et les estimations des composantes de la variance qui sont extrêmement petites ou égales à zéro. Ces problèmes méritent une discussion approfondie, compte tenu de leurs conséquences pour la réalisation d'enquêtes. L'approche bayésienne illustrée ici est en mesure de tenir compte de l'incertitude de l'estimation des composantes de la variance pour former des ensembles crédibles et ne s'appuie pas sur la théorie asymptotique, mais nous avons constaté que les inférences au sujet des variations des composantes de la variance entre plusieurs sous-groupes différents d'intervieweuses de la NSFG (chacun d'une taille modérée) ne variaient pas par rapport à celles qui seraient faites au moyen des approches fréquentistes. Il faudrait approfondir l'analyse pour déterminer si nous obtiendrions ou non les mêmes résultats pour des groupes encore plus petits d'intervieweurs; l'étude de simulation présentée à la section 3 suggère qu'aucune des deux méthodes ne fonctionne bien dans un contexte où deux groupes de 20 intervieweurs recueillent des données auprès de 10 répondants chacun. Une application initiale de ces deux méthodes aux données à partir du premier trimestre de collecte des données pendant ce cycle de la NSFG (une vingtaine d'intervieweurs dans chacun des deux groupes interviewent une vingtaine de répondants chacun en moyenne) a donné des résultats semblables à ceux décrits ici pour les gros échantillons, certains indices portant à croire que l'approche bayésienne est plus conservatrice (West 2011).

En général, l'approche bayésienne offre une forme plus naturelle d'inférence pour ce problème, ce qui indique un éventail de valeurs pour la différence, où environ 95 % des différences se trouveront. Ces résultats pourraient intéresser certains consommateurs de produits d'une enquête donnée, au lieu de la simple valeur p pour un test du rapport des vraisemblances, qui ne donne pas aux utilisateurs une idée de l'ampleur des différences possibles. Dans le contexte fréquentiste, le test du rapport des vraisemblances pourrait être la seule méthode d'inférence disponible si l'estimation ponctuelle du maximum de la pseudo-vraisemblance pour une ou plusieurs des composantes de la variance est de zéro, sans erreur-type correspondante (empêchant le calcul d'intervalles de type Wald). Cette situation a été observée aussi bien dans les simulations que dans les analyses de la NSFG, en particulier pour les groupes ayant des petits échantillons d'intervieweurs; compte tenu de l'utilisation des tests du rapport des vraisemblances pour la théorie asymptotique, l'approche bayésienne pourrait être un meilleur choix pour les petits échantillons. Le rendement de l'approche bayésienne n'est toutefois pas idéal pour les très petits échantillons, comme le démontre l'étude de simulation à la section 3.

Nous avons observé deux importantes différences entre les sous-groupes d'intervieweurs pour les données de la NSFG, et dans chaque cas, le groupe ayant la plus petite variance avait une composante de variance estimée de zéro (sans erreur-type calculée) dans le contexte de l'approche fréquentiste. Les inférences obtenues en fonction de ces estimations (où les valeurs de vraisemblances ont été calculées au moyen des estimations de zéro pour les sous-groupes en question pendant les tests du rapport des vraisemblances) correspondaient à l'approche bayésienne. Nous rappelons aux lecteurs qui utilisent les méthodes fréquentistes que les petits échantillons d'intervieweurs ou les très petites quantités de variance parmi les intervieweurs pour des variables en particulier peuvent entraîner des estimations du maximum de vraisemblance négatives des composantes de la variance, ce qui peut être problématique pour l'interprétation de l'écart intervieweurs pour les groupes individuels. Certaines procédures logicielles

capables d'intégrer des modèles multiniveaux (p. ex. la procédure gllamm dans Stata, ou la fonction lmer() dans R) limitent les composantes de la variance aux valeurs supérieures à zéro pendant l'estimation pour empêcher ce problème, ce qui peut prolonger les délais d'estimation. D'autres procédures logicielles (comme GLIMMIX dans SAS) fixeront tout simplement ces estimations négatives à zéro, et échoueront à calculer une erreur-type. Bien que ces composantes de la variance ne puissent pas en principe être égales à zéro, nous suggérons d'interpréter ces résultats pour prouver qu'il y a une variance négligeable entre les intervieweurs d'un groupe donné. Bates (2009) s'oppose à l'utilisation d'erreurs-types pour faire des inférences au sujet des composantes de la variance dans le contexte fréquentiste, en particulier lorsque les composantes de la variance sont proches de zéro, au lieu de suggérer que la fonction de la déviance profilée devrait être utilisée pour visualiser la précision des estimations. Cette approche et l'approche de Wald pour calculer les intervalles de confiance seront toujours limitées aux petits échantillons.

Nous ne voyons pas de problème empirique à utiliser ces estimations zéro pour effectuer les tests du rapport des vraisemblances démontrés ici pour comparer des groupes d'intervieweurs, étant donné que les tirages bayésiens des composantes de la variance dans ces groupes seraient également très petits. Cependant, dans le cas de la variance de l'estimation de l'écart intervieweurs pour les groupes individuels, un examen de la sensibilité des inférences bayésiennes aux choix de différentes répartitions a priori pour les composantes de la variance devrait être effectué lorsque des composantes de la variance proches de zéro sont prévues, ou que le nombre d'intervieweurs est relativement faible (Browne et Draper 2006; Lambert et coll. 2005). De plus, si les chercheurs d'enquête souhaitent *prédire* les effets aléatoires des intervieweurs dans le cas où les composantes de l'écart intervieweurs devraient être proches de zéro, les méthodes fréquentistes et bayésiennes donnent toutes deux des résultats très médiocres, et la prédiction n'est pas recommandée dans ce cas (Singh et coll. 1998, p. 390). Voir Savalei et Kolenikov (2008) pour obtenir plus d'information sur la question de la variance zéro.

Cette étude n'était certainement pas sans limites. Nous reconnaissons que la conception de la NSFG, où les intervieweurs sont habituellement affectés à une seule zone d'échantillonnage principale, ne permettait pas une attribution imbriquée de cas échantillonnés aux intervieweurs. Par conséquent, il est difficile de démêler les effets des intervieweurs et les effets des zones d'échantillonnage principales. Les méthodes décrites dans le présent document peuvent facilement intégrer des covariables supplémentaires au niveau de l'intervieweur ou de la région, afin d'expliquer la variance parmi les intervieweurs ou les régions en raison des covariables observables. Il faudrait continuer la recherche en général pour déterminer comment estimer l'écart intervieweurs en présence d'un plan d'échantillonnage strictement non imbriqué, et nous n'avons pas abordé cette question ouverte dans le présent document. Comme mentionné à la section 1, des plans d'échantillonnage ont été utilisés dans les études récentes pour démêler les effets des intervieweurs des effets des régions. Les études ultérieures devraient étudier la capacité des deux approches examinées dans le présent document de détecter les différences entre les composantes de l'écart intervieweurs lorsque l'on utilise des modèles multiniveaux recoupés qui comprennent aussi les effets des régions dans un plan d'échantillonnage imbriqué.

De même, nous n'avons pas tenu compte des caractéristiques d'échantillonnage complexe de la NSFG (c.-à-d. pondération ou échantillonnage en grappes stratifié) dans les analyses. La théorie à la base de l'estimation des paramètres dans les modèles multiniveaux en présence de poids d'enquête nécessite des poids pour les répondants et les grappes aux niveaux supérieurs, en l'occurrence les intervieweurs (Rabe-Hesketh et Skrondal 2006; Pfefferman, Skinner, Holmes, Goldstein et Rasbash 1998). Les analyses

présentées ici présument effectivement que nous avons un échantillon d'intervieweurs pour une population plus grande qui a été sélectionnée avec une probabilité égale, et que tous les répondants pour chaque intervieweur avaient un poids égal. Les méthodes décrites par Gabler et Lahiri (2009) pourraient être utiles pour pallier cette faiblesse, et les analystes pourraient également inclure les effets fixes des poids d'enquête ou des codes de stratification dans les modèles proposés ici. Nous réservons ces approfondissements à des recherches futures.

Enfin, ce document n'a pas tenu compte d'un autre aspect riche de l'approche bayésienne, en ce que les tirages a posteriori des 87 effets aléatoires des intervieweurs dans les modèles ont également été générés par l'algorithme d'échantillonnage de Gibbs BUGS. Ces tirages permettraient aux gestionnaires d'enquête de faire des inférences au sujet des effets qu'ont certains intervieweurs sur certaines mesures d'enquête. Une mise à jour constante et régulière de ces répartitions a posteriori au fil du déroulement du cycle de collecte des données permettrait aux gestionnaires d'enquête d'intervenir lorsque les répartitions a posteriori pour des intervieweurs en particulier suggèrent que ces intervieweurs ont des effets non nuls sur les mesures d'enquête.

Remerciements

Les auteurs sont reconnaissants du soutien d'un contrat avec le National Center for Health Statistics, qui a permis la réalisation du septième cycle de la National Survey of Family Growth (contrat 200-2000-07001).

Annexe

A.1 Exemple de code

Voici un exemple de code pour intégrer les types de modèles abordés dans le document au moyen de SAS PROC GLIMMIX. Dans ce code, PARITY et SEXMAIN sont les variables de dénombrement et binaire respectivement, mesurées pour les répondants de la NSFG, FINAL_INT_ID est un code d'utilisateur final de l'intervieweur, et INT_NVMARRIED est une variable qui indique qu'un intervieweur n'a jamais été marié. L'option ASYCOV indique les estimations asymptotiques des variances et des covariances des composantes de la variance estimée.

```
/* état matrimonial */

proc glimmix data = bayes.final_analysis asycov;
   class final_int_id int_nvmarried;
   model parity = int_nvmarried / dist = negbin link = log solution cl;
   random int / subject = final_int_id group = int_nvmarried;
   covtest homogeneity / cl (type = plr);
   nloptions tech=nrridg;
run;
```

```
proc glimmix data = bayes.final_analysis asycov;
   class final_int_id int_nvmarried;
   model sexmain (event = "1") = int_nvmarried / dist = binary link = logit
solution cl;
   random int / subject = final_int_id group = int_nvmarried;
   covtest homogeneity / cl (type = plr);
   nloptions tech=nrridg;
run;
```

Nous fournissons également un exemple de code WinBUGS pour intégrer les modèles au moyen des approches bayésiennes dont il est question ci-après. Nous appelons le code WinBUGS du logiciel R. Dans ce code, LOWAGE.G est un indicateur au niveau de l'intervieweur (avec 87 valeurs) d'appartenance à la tranche d'âge plus jeune des intervieweurs, et HIGHAGE.G est un indicateur d'appartenance au groupe plus âgé. Le code complet, y compris le code créant les variables utilisées ci-après, peut être obtenu des auteurs sur demande.

```
# Charger les progiciels nécessaires pour utiliser BUGS à partir de R

library(arm)
library(R2WinBUGS)

############# Analyses de parité

# Fichiers BUGS pour les groupes d'âge et la parité (age_nb.bug)

model {
   for (i in 1:n){
      parity[i] ~ dpois(lambda[i])
      lambda[i] <- rho[i]*mu[i]
      log(mu[i]) <- b0[intid[i]]
      rho[i]~dgamma(alpha,alpha)
   }

   for (j in 1:J){
      b0[j] ~ dnorm(b0.hat[j], tau.b0[highage.g[j]+1])
      b0.hat[j] <- beta0 + beta1*lowage.g[j]
   }

   beta0 ~ dnorm(0,0.01)
   beta1 ~ dnorm(0,0.01)
   alpha <- exp(logalpha)
   logalpha ~ dnorm(0,0.01)

   for (k in 1:2){
      tau.b0[k] <- pow(sigma.b0[k], -2)
      sigma.b0[k] ~ dunif(0,10)
   }
}

# Simulations pour le modèle parité/groupe d'âge dans BUGS

n <- length(parity)
J <- 87
age.data <- list("n", "J", "parity", "intid", "highage.g", "lowage.g")
```

```
age.inits <- function(){
   list (b0=rnorm(J), beta0=rnorm(1), beta1=rnorm(1), sigma.b0=runif(2),
logalpha=rnorm(1))}
age.parameters <- c("b0", "beta0", "beta1", "sigma.b0", "alpha")
age.1 <- bugs(age.data, age.inits, age.parameters, "age_nb.bug", n.chains = 3,
n.iter=5000, debug=TRUE,
bugs.directory="C:/Users/bwest/Desktop/winbugs14/WinBUGS14")

attach.bugs(age.1)

# Pour les tableaux de resultats et d'inférences

resultsmat <- cbind(numeric(6),numeric(6),numeric(6),numeric(6))

resultsmat[1,1] <- quantile(beta0,0.5)
resultsmat[1,2] <- sd(beta0)
resultsmat[1,3] <- quantile(beta0,0.025)
resultsmat[1,4] <- quantile(beta0,0.975)

resultsmat[2,1] <- quantile(beta1,0.5)
resultsmat[2,2] <- sd(beta1)
resultsmat[2,3] <- quantile(beta1,0.025)
resultsmat[2,4] <- quantile(beta1,0.975)

resultsmat[3,1] <- quantile(sigma.b0[,1]^2,0.5)
resultsmat[3,2] <- sd(sigma.b0[,1]^2)
resultsmat[3,3] <- quantile(sigma.b0[,1]^2,0.025)
resultsmat[3,4] <- quantile(sigma.b0[,1]^2,0.975)

resultsmat[4,1] <- quantile(sigma.b0[,2]^2,0.5)
resultsmat[4,2] <- sd(sigma.b0[,2]^2)
resultsmat[4,3] <- quantile(sigma.b0[,2]^2,0.025)
resultsmat[4,4] <- quantile(sigma.b0[,2]^2,0.975)

resultsmat[5,1] <- quantile(1/alpha,0.5)
resultsmat[5,2] <- sd(1/alpha)
resultsmat[5,3] <- quantile(1/alpha,0.025)
resultsmat[5,4] <- quantile(1/alpha,0.975)

vardiff <- sigma.b0[,1]^2 - sigma.b0[,2]^2
resultsmat[6,1] <- quantile(vardiff,0.5)
resultsmat[6,2] <- sd(vardiff)
resultsmat[6,3] <- quantile(vardiff,0.025)
resultsmat[6,4] <- quantile(vardiff,0.975)

resultsmat

############# Analyses de l'activité sexuelle actuelle

# Fichiers BUGS pour les groupes d'âge et l'activité sexuelle (age_bin.bug)

model {
   for (i in 1:n){
      sexmain[i] ~ dbern(p[i])
      logit(p[i]) <- b0[intid[i]]
   }
```

```
   for (j in 1:J){
      b0[j] ~ dnorm(b0.hat[j], tau.b0[highage.g[j]+1])
      b0.hat[j] <- beta0 + beta1*lowage.g[j]
   }
   beta0 ~ dnorm(0,0.01)
   beta1 ~ dnorm(0,0.01)

   for (k in 1:2){
      tau.b0[k] <- pow(sigma.b0[k], -2)
      sigma.b0[k] ~ dunif(0,10)
   }
}

# Simulations pour le modèle parité/groupe d'âge dans BUGS

n <- length(sexmain)
J <- 87
age.data <- list("n", "J", "sexmain", "intid", "highage.g", "lowage.g")
age.inits <- function(){
   list (b0=rnorm(J), beta0=rnorm(1), beta1=rnorm(1), sigma.b0=runif(2))}
age.parameters <- c("b0", "beta0", "beta1", "sigma.b0")
age.1 <- bugs(age.data, age.inits, age.parameters, "age_bin.bug", n.chains =
3, n.iter=5000, debug=TRUE,
bugs.directory="C:/Users/bwest/Desktop/winbugs14/WinBUGS14")

   attach.bugs(age.1)
```

Bibliographie

Bates, D. (2009). Assessing the precision of estimates of variance components. *Presentation to the Max Planck Institute for Ornithology*, Seewiesen, 21 juillet 2009. La présentation peut être téléchargée à partir de http://lme4.r-forge.r-project.org/slides/2009-07-21-Seewiesen/4PrecisionD.pdf.

Biemer, P.P. et Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. Chapitre 27 de *Survey Measurement and Process Quality*, Editeurs Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz et Trewin. Wiley-Interscience, 603-632.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Browne, W.J. et Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3), 473-514.

BUGS, http://www.mrc-bsu.cam.ac.uk/bugs/welcome.html.

Carlin, B.P. et Louis, T.A. (2009). *Bayesian Methods for Data Analysis*. Chapman and Hall / CRC Press.

Chaloner, K. (1987). A Bayesian approach to the estimation of variance components for the unbalanced one-way random model. *Technometrics*, 29(3), 323-337.

Collins, M. et Butcher, B. (1982). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.

Durham, C.A., Pardoe, I. et Vega, E. (2004). A methodology for evaluating how product characteristics impact choice in retail settings with many zero observations: An application to restaurant wine purchase. *Journal of Agricultural and Resource Economics*, 29(1), 112-131.

Durrant, G.B., Groves, R.M., Staetsky, L. et Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly*, 74, 1-36.

Faraway, J.J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall / CRC Press: Boca Raton, FL.

Farrell, P.J. (2000). Bayesian inference for small area proportions. Sankhya: The *Indian Journal of Statistics, Series B* (1960-2002), 62(3), 402-416.

Fowler, F.J. et Mangione, T.W. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Newbury Park, CA: Sage.

Gabler, S. et Lahiri, P. (2009). De la définition et de l'interprétation de la variabilité d'intervieweur pour un plan d'échantillonnage complexe. *Techniques d'enquête*, 35(1), 91-106.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533.

Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (2004). *Bayesian Data Anaylsis*. Chapman and Hall / CRC Press.

Gelman, A. et Hill, J. (2007). *Data Analysis using Regression and Multilevel / Hierarchical Models*. Cambridge University Press.

Gelman, A. et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-511.

Gilks, W.R. et Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41, 337-348.

Goldstein, H. (1995). *Multilevel Statistical Models, Second Edition*. Kendall's Library of Statistics 3, Edward Arnold.

Groves, R.M. (2004). *Survey Errors and Survey Costs (2nd Edition)*. Dans le chapitre 8 : The Interviewer as a Source of Survey Measurement Error. Wiley-Interscience.

Groves, R.M., Mosher, W.D., Lepkowski, J.M. et Kirgis, N.G. (2009). Planning and development of the continuous National Survey of Family Growth. National Center for Health Care Statistics. *Vital Health Statistics*, 1(48).

Hilbe, J.M. (2007). *Negative Binomial Regression*. Cambridge University Press.

Hox, J. (1998). *Multilevel Modeling: When and Why*. Dans I. Balderjahn, R. Mathar et M. Schader (Eds.). Classification, data anaylsis, and data highways. New York: Springer-Verlag, 147-154.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

Lambert, P.C., Sutton, A.J., Burton, P.R., Abrams, K.R. et Jones, D.R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15), 2401-2428.

Lepkowski, J.M., Mosher, W.D., Davis, K.E., Groves, R.M. et Van Hoewyk, J. (2010). The 2006-2010 National Survey of Family Growth: sample design and analysis of a continuous survey. National Center for Health Statistics, *Vital and Health Statistics*, 2(150), juin 2010.

Lynn, P., Kaminska, O. et Goldstein, H. (2011). Panel attrition: how important is it to keep the same interviewer? *ISER Working Paper Series*, Article 2011-02.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Mangione, T.W., Fowler, F.J. et Louis, T.A. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics*, 8(3), 293-307.

Molenberghs, G. et Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag, Berlin.

O'Muircheartaigh, C. et Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society, Series A*, 161(1), 63-77.

O'Muircheartaigh, C. et Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A*, 162(3), 437-446.

Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60(1), 23-40.

Pinheiro, J.C. et Chao, E.C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15, 58-81.

Rabe-Hesketh, S. et Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series A*, 169, 805-827.

Raudenbush, S.W. et Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA.

Rodriguez, G. et Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society, Series A*, 164(2), 339-355.

SAS Institute, Inc. (2010). Documentation en ligne pour la procédure GLIMMIX.

Savalei, V. et Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, 13(2), 150-170.

Schaeffer, N.C., Dykema, J. et Maynard, D.W. (2010). *Handbook of Survey Research, Second Edition*. Dans Interviewers and Interviewing. J.D. Wright et P.V. Marsden (eds). Bingley, U.K.: Emerald Group Publishing Limited.

Schnell, R. et Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21(3), 389-410.

Schober, M. et Conrad, F. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576-602.

Singh, A.C., Stukel, D.M. et Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society, Series B*, 60(2), 377-396.

Ugarte, M.D., Goicoa, T. et Militino, A.F. (2009). Empirical bayes and fully bayes procedures to detect high-risk areas in disease mapping. *Computational Statistics and Data Analysis*, 53, 2938-2949.

Van Tassell, C.P. et Van Vleck, L.D. (1996). Multiple-trait Gibbs sampler for animal models: Flexible programs for Bayesian and likelihood-based (co)variance component inference. *Journal of Animal Science*, 74, 2586-2597.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26, 37-52.

West, B.T. (2011). Bayesian analysis of between-group differences in variance components in hierarchical generalized linear models. Dans JSM Proceedings, Survey Research Methods Section. Alexandria, VA: *American Statistical Association*, 1828-1842.

West, B.T. et Galecki, A.T. (2011). An overview of current software procedures for fitting linear mixed models. *The American Statistician*, 65(4), 274-282.

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(2), 277-297.

West, B.T. et Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74(5), 1004-1026.

Zhang, D. et Lin, X. (2010). Variance component testing in generalized linear mixed models for longitudinal / clustered data and other related topics. Random Effect and Latent Variable Model Selection. *Springer Lecture Notes in Statistics*, Volume 192.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 211-233
Statistique Canada, n° 12-001-X au catalogue

L'agrégation bootstrap des estimateurs non différenciables dans les enquêtes complexes

Jianqiang C. Wang, Jean D. Opsomer et Haonan Wang[1]

Résumé

L'agrégation bootstrap est une puissante méthode de calcul utilisée pour améliorer la performance des estimateurs inefficaces. Le présent article est le premier à explorer l'utilisation de l'agrégation bootstrap dans l'estimation par sondage. Nous y examinons les effets de l'agrégation bootstrap sur les estimateurs d'enquête non différenciables, y compris les fonctions de répartition de l'échantillon et les quantiles. Les propriétés théoriques des estimateurs d'enquête agrégés par bootstrap sont examinées sous le régime fondé sur le plan de sondage et le régime fondé sur le modèle. En particulier, nous montrons la convergence par rapport au plan des estimateurs agrégés par bootstrap et obtenons la normalité asymptotique des estimateurs dans un contexte fondé sur le modèle. L'article explique comment la mise en œuvre de l'agrégation bootstrap des estimateurs d'enquête peut tirer parti des répliques produites pour l'estimation par sondage de la variance, facilitant l'application de l'agrégation bootstrap dans les enquêtes existantes. Un autre défi important dans la mise en œuvre de l'agrégation bootstrap en contexte d'enquête est l'estimation de la variance pour les estimateurs agrégés par bootstrap eux-mêmes, et nous examinons deux façons possibles d'estimer la variance. Les expériences par simulation révèlent une amélioration de l'estimateur par agrégation bootstrap proposé par rapport à l'estimateur original et comparent les deux approches d'estimation de la variance.

Mots-clés : Bootstrap; fonction de distribution; estimation des quantiles.

1 Introduction

L'agrégation bootstrap, appelée « bagging » en anglais, est une méthode de rééchantillonnage initialement introduite pour améliorer les algorithmes d'apprentissage « faibles ». L'agrégation bootstrap a été proposée par Breiman (1996), qui a démontré de façon heuristique qu'elle améliorait la performance des prédicteurs à structure arborescente. L'agrégation bootstrap a depuis été appliquée dans un large éventail de contextes et analysée par de nombreux auteurs. Bühlmann et Yu (2002) ont démontré les effets de lissage de l'agrégation bootstrap et de ses variations sur les algorithmes de classification des décisions difficiles et formalisé la notion de « prédicteurs instables ». Chen et Hall (2003) ont dérivé des résultats théoriques de l'agrégation bootstrap d'estimateurs définis par des équations d'estimation. Buja et Stuetzle (2006) ont envisagé l'agrégation bootstrap des statistiques U et soutenu que l'agrégation bootstrap [*Traduction*] « réduit souvent, mais pas toujours, la variance, mais augmente toujours le biais ». Friedman et Hall (2007) ont examiné l'incidence de l'agrégation bootstrap sur les estimateurs non linéaires. Plus récemment, Hall et Robinson (2009) ont discuté des effets de l'agrégation bootstrap sur le choix par validation croisée des paramètres de lissage et présenté des résultats intrigants concernant l'amélioration, par agrégation bootstrap, de l'ordre du choix par validation croisée de la bande passante du noyau.

La littérature susmentionnée étudiait les effets de l'agrégation bootstrap sur différents estimateurs, particulièrement les estimateurs non différenciables non linéaires, sous l'hypothèse d'échantillonnage de données *iid* (indépendantes et identiquement distribuées). Pour les données dépendantes, Lee et Yang

1. Jianqiang C. Wang, Hewlett-Packard Labs, Palo Alto, CA 94304. Jean D. Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523. Courriel : jopsomer@stat.colostate.edu; Haonan Wang, Department of Statistics, Colorado State University, Fort Collins, CO 80523.

(2006) ainsi que Inoue et Kilian (2008) ont étudié les effets de l'agrégation bootstrap sur les séries chronologiques économiques. Les premiers ont étudié les effets de l'agrégation bootstrap sur les prédicteurs non différenciables comme les fonctions de signe et les quantiles, tandis que les seconds ont mis l'accent sur l'agrégation bootstrap des prédicteurs de prétest applicables à la prévision de l'inflation des prix à la consommation aux États-Unis.

Comme le montre ce bref examen de la littérature, l'agrégation bootstrap est une méthode prometteuse utilisée pour améliorer l'efficacité des estimateurs. L'agrégation bootstrap pour les estimateurs d'enquête n'a toutefois pas été envisagée jusqu'ici. Le présent article est le premier à examiner l'utilisation de l'agrégation bootstrap en contexte d'enquête; il comprend une évaluation du gain d'efficacité potentiel, avance un certain nombre de résultats théoriques, et explore les questions de mise en œuvre et d'estimation de la variance. Conformément aux pratiques générales d'enquête, nous examinons seulement les estimateurs qui peuvent s'exprimer sous forme de fonctions de Horvitz-Thompson (HT). Plus précisément, nous étudions trois types d'estimateurs. Premièrement, de nombreux estimateurs d'usage courant peuvent s'exprimer en tant que fonctions différenciables des estimateurs de HT. Par exemple, l'estimateur de Hajek, l'estimateur par le ratio et l'estimateur par la régression généralisée peuvent tous être considérés comme des fonctions différenciables des estimateurs de HT. Deuxièmement, il existe d'autres estimateurs d'enquête non différenciables, dont ceux de Dunstan et Chambers (Dunstan et Chambers 1986) et de Rao-Kovar-Mantel (Rao, Kovar et Mantel 1990), l'estimateur de post-stratification endogène (Breidt et Opsomer 2008) et les estimateurs de proportion de personnes à faible revenu (Berger et Skinner 2003). Troisièmement, d'autres estimateurs sont définis seulement comme solutions d'équations d'estimation pondérées. Pour plus de renseignements sur les équations d'estimation en contexte d'enquête, voir Godambe et Thompson (2009), Fuller (2009) et leurs références.

Bien que l'agrégation bootstrap puisse être considérée comme un type de méthode de répliques, elle est très différente de la méthode bootstrap et d'autres méthodes de répliques conçues pour estimer la variance. Contrairement à ces méthodes, l'agrégation bootstrap a pour but d'améliorer l'estimateur même. L'agrégation bootstrap peut être naturellement intégrée aux enquêtes complexes à grande échelle, car nous pouvons tirer parti des poids de réplication facilement disponibles dans de nombreuses enquêtes pratiques. Dans le présent article, nous montrons comment les répliques créées pour l'estimation bootstrap de la variance peuvent être modifiées et utilisées dans l'agrégation bootstrap de l'estimateur original. Malheureusement, une difficulté inhérente à l'application de l'agrégation bootstrap dans les enquêtes est l'absence d'estimateur de variance fondé sur le plan de sondage. Nous examinons un certain nombre de méthodes proposées pour estimer la variance des estimateurs agrégés par bootstrap, mais il reste du travail à faire dans ce domaine.

Le reste de cet article est organisé comme suit. Dans la section 2, nous définissons nos estimateurs cibles et présentons la version agrégée par bootstrap de chaque estimateur. Dans la section 3, nous présentons les propriétés théoriques des estimateurs agrégés par bootstrap. Dans la section 4, nous montrons comment utiliser les répliques pour appliquer les versions agrégées par bootstrap des estimateurs, et nous examinons l'estimation de la variance pour les estimateurs agrégés par bootstrap résultants. Dans la section 5, nous exposons les résultats des simulations et, dans la section 6, nous présentons quelques conclusions et remarques finales.

2 Agrégation bootstrap des estimateurs

2.1 Approche générale

Dans cette section, nous discutons de la mise en œuvre de l'agrégation bootstrap en contexte d'estimation par sondage. Nous commençons par présenter la notation nécessaire. Soit U une population finie de taille N, où chaque élément $i \in U$ est associé à un vecteur de mesures $\mathbf{y}_i$, dans l'espace euclidéen $\mathbb{R}^q$ à q dimensions. Nous utilisons le plan d'échantillonnage $p(\ )$ pour tirer un échantillon aléatoire $A \subseteq U$ de taille n. Soit $\mathcal{Y} = \{\mathbf{y}_i \mid i \in A\}$ l'ensemble des observations de l'échantillon. Ici, le plan d'échantillonnage pourrait être un échantillonnage aléatoire simple sans remise (EASSR), un échantillonnage de Poisson ou un plan de sondage complexe qui prévoit une stratification et/ou un échantillonnage à plusieurs degrés. Dans chaque plan, la probabilité qu'un élément i soit inclus dans l'échantillon est π_i.

La moyenne de population du vecteur de mesure $\mathbf{y}$ est $\boldsymbol{\mu}$. Elle peut être estimée au moyen de l'estimateur de Horvitz-Thompson (HT) défini comme étant

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i \in A}\frac{\mathbf{y}_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{y}_i}{\pi_i} I_i, \tag{2.1}$$

où I_i est l'**indicateur d'appartenance à l'échantillon** pour le i-ième élément. De façon plus générale, soit θ une quantité d'intérêt de la population et $\hat{\theta}(\mathcal{Y})$ l'estimateur de θ selon les observations de l'échantillon $\mathcal{Y}$. L'estimateur $\hat{\theta}(\mathcal{Y})$ est abrégé en $\hat{\theta}$ s'il n'y a aucune confusion. Comme il est noté dans la section qui précède, nous supposons que $\hat{\theta}$ peut s'exprimer en tant que fonction d'estimateurs simples de la forme (2.1).

Sous sa forme la plus générale, l'algorithme d'agrégation bootstrap pour l'estimation par sondage se présente comme suit :

1. Pour $b = 1, 2, \ldots, B$:
   a. Tirer le nouvel échantillon A_b de l'échantillon aléatoire A, et désigner les observations du nouvel échantillon par $\mathcal{Y}_b^* = \{\mathbf{y}_i \mid i \in A_b\}$.
   b. Calculer l'estimation du paramètre en se fondant sur le nouvel échantillon A_b, désigné par $\hat{\theta}(\mathcal{Y}_b^*)$.

2. Faire la moyenne sur les estimations répétées $\hat{\theta}(\mathcal{Y}_1^*)$, $\hat{\theta}(\mathcal{Y}_2^*)$, …, $\hat{\theta}(\mathcal{Y}_B^*)$ afin d'obtenir l'estimateur agrégé par bootstrap,

$$\hat{\theta}_{bag} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}(\mathcal{Y}_b^*). \tag{2.2}$$

Dans la littérature sur l'agrégation bootstrap, les nouveaux échantillons A_b sont souvent appelés **échantillons bootstrap** (Breiman 1996) et c'est ce que nous faisons ici, même si nous ne les utilisons pas pour estimer la variance.

Dans l'algorithme, les échantillons bootstrap pouvaient être tirés selon le plan d'échantillonnage plutôt que par répartition empirique des observations de l'échantillon, méthode qui est plus couramment utilisée

dans la littérature sur l'agrégation bootstrap ordinaire (Breiman 1996) et qui équivaut à un échantillonnage aléatoire simple (avec ou sans remise). Par exemple, si l'échantillon A est tiré par échantillonnage stratifié ou en grappes, ce plan de sondage pourrait être pris en considération lors de la sélection des nouveaux échantillons. Plus généralement en contexte d'enquête, l'étape 1 de l'algorithme d'agrégation bootstrap proposé peut être traitée dans le cadre d'un échantillonnage à deux phases, la première phase correspondant à l'échantillon original A et la deuxième, au nouvel échantillon A_b. Ainsi, l'estimateur par extension classique pour plans de sondage à deux phases de Särndal, Swensson et Wretman (1997) est mis en œuvre pour calculer l'estimateur répété $\hat{\theta}\left(\mathcal{Y}_b^*\right)$. Dans le nouvel échantillon A_b, la pseudo probabilité d'inclusion pour le i-ième élément est $\pi_i^* = \pi_i \pi_{i|A}$ où $\pi_{i|A} = \Pr\left(i \in A_b \mid i \in A\right)$ est la probabilité d'inclusion du i-ième élément du nouvel échantillon A_b étant donné son inclusion dans l'échantillon A. Par conséquent, l'estimateur agrégé par bootstrap est une approximation de l'espérance de l'estimateur à deux phases en ce qui concerne la deuxième phase de l'échantillonnage, qui est aussi appelée **espérance bootstrap** dans les méthodes d'agrégation bootstrap ordinaire (Bühlmann et Yu 2002). Bien qu'un plan d'échantillonnage bootstrap général soit possible, nous nous limitons à l'EASSR dans les parties théoriques de cet article. Pour élargir la portée de notre discussion, dans la section sur l'estimation de la variance et dans la section numérique, nous présentons le cas où les échantillons bootstrap sont tirés par EASSR stratifié avec les mêmes strates que l'échantillon original A, ce qui est une extension utile et réaliste.

Prenons, par exemple, l'estimateur de HT défini en (2.1). Le rééchantillonnage bootstrap de l'échantillon réalisé A est tiré sous EASSR de taille k. Selon ce plan de rééchantillonnage, l'estimateur par échantillon répété est défini comme suit :

$$\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right) = \frac{1}{N} \sum_{i \in A_b} \frac{\mathbf{y}_i}{\pi_i^*}, \tag{2.3}$$

où la pseudo probabilité d'inclusion est $\pi_i^* = \pi_i \pi_{i|A} = k\pi_i / n$. La formule (2.2) peut alors être utilisée pour calculer la version agrégée par bootstrap de l'estimateur π^* classique. Un calcul simple montre que l'estimateur agrégé par bootstrap est identique à l'estimateur de HT original si tous les échantillons EASSR de taille k sont dénombrés dans le calcul de (2.2). Il en va de même pour tous les autres estimateurs linéaires. En général, le calcul de l'estimateur agrégé par bootstrap $\hat{\theta}_{bag}$ n'est pas aussi facile. Dans le reste de cette section, nous nous attardons aux calculs de ce genre pour les trois types d'estimateurs non linéaires abordés dans la section 1.

2.2 Agrégation bootstrap des estimateurs différenciables

Pour les estimateurs d'enquête qui sont des fonctions différenciables des estimateurs de HT, la quantité d'intérêt de la population peut aussi s'exprimer sous forme de fonction différenciable des moyennes de population, c'est-à-dire $\theta_d = m\left(\boldsymbol{\mu}\right)$, où $m(\cdot)$ est une fonction différenciable connue. L'indice « d » veut dire **différenciable** par opposition à **non différenciable** $\left(\theta_{nd}\right)$ et à **équation d'estimation** $\left(\theta_{ee}\right)$ à venir plus tard. Un estimateur direct de type « plug-in » de θ_d, fondé sur les observations de l'échantillon $\mathcal{Y}$, peut s'écrire

$$\hat{\theta}_d = m(\hat{\boldsymbol{\mu}}), \tag{2.4}$$

où $\hat{\boldsymbol{\mu}}$ est défini en (2.1). Ainsi, la version d'échantillon répété de $\hat{\theta}_d$ peut s'écrire

$$\hat{\theta}_d\left(\mathcal{Y}_b^*\right) = m\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)\right),$$

où $\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)$ est défini en (2.3). L'estimateur agrégé par bootstrap de θ_d, dénoté par $\hat{\theta}_{d,bag}$, est alors défini au moyen de la formule en (2.2).

2.3 Agrégation bootstrap des estimateurs non différenciables explicitement définis

Comme exemple de ce type d'estimateurs, prenons la proportion de ménages dont le revenu est inférieur au seuil de pauvreté pour une population donnée. Cette proportion peut s'écrire $(1/N)\sum_{i=1}^{N} I(y_i \le \lambda_N)$, où y_i est la valeur du revenu pour le i-ième ménage de la population, et λ_N est le seuil de pauvreté de la population. On peut voir que cette quantité d'intérêt est la moyenne des fonctions noyau de l'indicateur, et que la fonction noyau n'est pas différenciable en λ_N. Ici, nous considérons une classe plus générale où le noyau est une fonction arbitraire non différenciable, mais bornée. Ce type de quantité de population peut s'écrire

$$\theta_{nd} = \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{y}_i - \boldsymbol{\lambda}_N),$$

où $\boldsymbol{\lambda}_N$ est un paramètre de population inconnu, par exemple la moyenne, un quantile ou une autre quantité de population, et $h(\mathbf{y} - \boldsymbol{\lambda}) : \mathbb{R}^p \to \mathbb{R}$ est une fonction non différenciable de $\boldsymbol{\lambda}$. La quantité de population θ_{nd} généralise la notion de proportion inférieure à un niveau estimé et ressemble à la forme générale d'une statistique U.

Wang et Opsomer (2011) ont étudié une classe d'estimateurs ressemblant à des statistiques U, à savoir les estimateurs d'enquête non différenciables,

$$\hat{\theta}_{nd} = \frac{1}{N}\sum_{i \in A} \frac{1}{\pi_i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right), \tag{2.5}$$

où $\hat{\boldsymbol{\lambda}}$ est un estimateur fondé sur le plan de $\boldsymbol{\lambda}_N$. Hors contexte d'une enquête, les estimateurs de ce type sont considérés comme des [*Traduction*] « fonctions non différenciables de la répartition empirique » (Bickel, Götze et van Zwet 1997). Les procédures bootstrap appropriées pour ces estimateurs ont notamment été étudiées par Beran et Srivastava (1985) et par Dümbgen (1993). Nous définissons la version répétée de $\hat{\theta}_{nd}$ fondée sur le nouvel échantillon A_b comme suit

$$\hat{\theta}_{nd}\left(\mathcal{Y}_b^*\right) = \frac{1}{N}\sum_{i \in A_b} \frac{1}{\pi_i^*} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right),$$

où $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ dépend uniquement du nouvel échantillon bootstrap A_b, et l'estimateur agrégé par bootstrap est alors défini comme étant la moyenne des estimateurs répétés. Supposons que le processus de

rééchantillonnage est l'EASSR de taille k, et que chaque sous-échantillon est choisi pour calculer l'estimateur d'agrégation bootstrap; l'estimateur bootstrap prend alors la forme suivante après manipulation :

$$\hat{\theta}_{nd,bag} = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i\binom{n-1}{k-1}}\sum_{A_b\ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right), \tag{2.6}$$

qui remplace $h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right)$ en (2.5) par une quantité « lisse » $\sum_{A_b\ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\Big/\binom{n-1}{k-1}$, en calculant la moyenne des « sauts » de l'estimateur. Dans bien des cas, on peut réduire la variance en effectuant ce remplacement. Le terme $\sum_{A_b\ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\Big/\binom{n-1}{k-1}$ est l'**espérance bootstrap** de $h\left(\mathbf{y}_i - \cdot\right)$ et peut être approximé en utilisant la convolution de $h\left(\mathbf{y}_i - \cdot\right)$ avec la distribution d'échantillonnage de $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$. Les aspects théoriques de $\hat{\theta}_{nd,bag}$ sont examinés dans la section 3.

2.4 Agrégation bootstrap des estimateurs définis par des équations d'estimation non différenciables

Enfin, nous expliquons comment faire l'agrégation bootstrap des estimateurs définis par des équations d'estimation non différenciables. Pour faciliter la présentation, nous considérons un paramètre d'intérêt unidimensionnel. Le paramètre de population θ_{ee} d'intérêt est défini comme suit

$$\theta_{ee} = \inf\left\{\gamma : S(\gamma) \geq 0\right\},$$

où

$$S(\gamma) = \frac{1}{N}\sum_{i=1}^{N}\psi\left(y_i - \gamma\right),$$

et $\psi(\cdot)$ est une fonction non différenciable réelle. Nous pouvons estimer le paramètre de population θ_{ee} au moyen de $\hat{\theta}_{ee}$, où

$$\hat{\theta}_{ee} = \inf\left\{\gamma : \hat{S}(\gamma) \geq 0\right\}$$

avec

$$\hat{S}(\gamma) = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\psi\left(y_i - \gamma\right).$$

Un estimateur de ce type qu'on rencontre souvent est le quantile d'échantillon défini par inversion de la fonction de distribution cumulative de l'échantillon (Francisco et Fuller 1991), où $\psi\left(y_i - \gamma\right) = I_{\left(y_i \leq \gamma\right)} - \alpha$ pour le quantile α.

Sur le plan conceptuel, il existe deux versions de l'agrégation bootstrap $\hat{\theta}_{ee}$: l'une résout l'« équation d'estimation agrégée par bootstrap » définie par agrégation bootstrap de la fonction de score, tandis que

l'autre calcule la moyenne sur les estimations rééchantillonnées de $\hat{\theta}_{ee}$. Comme il est expliqué dans la section 2.1, la première version produit un estimateur équivalant à l'estimateur original, parce que l'espérance bootstrap des échantillons bootstrap de $\hat{S}(\gamma)$ est égale à $\hat{S}(\gamma)$ pour un γ fixe. Nous considérons donc seulement la deuxième version. Pour définir l'estimateur d'équation agrégé par bootstrap, nous commençons par définir la fonction de score répétée $\hat{S}_b(\gamma)$ fondée sur le nouvel échantillon A_b comme suit

$$\hat{S}_b(\gamma) = \frac{1}{N}\sum_{i \in A_b} \frac{1}{\pi_i^*}\psi(y_i - \gamma).$$

L'estimateur répété fondé sur A_b est alors défini comme étant $\hat{\theta}_{ee}\left(\mathcal{Y}_b^*\right) = \inf\left\{\gamma : \hat{S}_b(\gamma) \geq 0\right\}$ et l'estimateur agrégé par bootstrap devient

$$\hat{\theta}_{ee,bag} = \frac{1}{\binom{n}{k}}\sum \hat{\theta}_{ee}\left(\mathcal{Y}_b^*\right), \tag{2.7}$$

où la moyenne est calculée sur tous les échantillons sans remise possibles de taille k choisis à partir de A. Chen et Hall (2003) ont examiné les estimateurs agrégés par bootstrap définis par des équations d'estimation non linéaires sous les conditions iid et ils ont conclu que l'agrégation bootstrap n'améliore pas toujours la précision des estimateurs étudiés.

3 Résultats théoriques

Nous commençons par décrire brièvement l'analyse asymptotique des estimateurs agrégés par bootstrap sous échantillonnage général d'une population finie, c.-à-d. dans un contexte fondé sur le plan de sondage. Nous procédons ainsi dans le cadre habituel d'une population croissante, où nous considérons une séquence croissante de populations emboîtées, disons U_N, $N = 1, 2, \ldots,$ avec une moyenne de population finie $\boldsymbol{\mu}_N$. Associée à la séquence de populations est une séquence de plans d'échantillonnage utilisés pour tirer un échantillon aléatoire $A_N \subseteq U_N$ de taille n_N, avec probabilités d'inclusion connexes π_{iN}. Comme cela se fait couramment dans la littérature sur les enquêtes, nous supprimons l'indice N dans l'échantillon A, la taille de l'échantillon n et les probabilités d'inclusion π_i. Par souci de concision, nous fournissons seulement les résultats asymptotiques fondés sur le plan pour l'agrégation bootstrap de l'estimateur différenciable $\hat{\theta}_d$ et non différenciable $\hat{\theta}_{nd}$. Les hypothèses formelles qui sous-tendent les résultats et les théorèmes associés aux estimateurs différenciables et non différenciables figurent à l'annexe A.1. La principale conclusion que nous pouvons tirer dans ce contexte fondé sur le plan est que, si nous partons d'un estimateur conforme au plan et que nous laissons le nombre d'échantillons bootstrap k augmenter avec n, les versions agrégées par bootstrap des estimateurs seront elles aussi conformes au plan. Il s'agit clairement d'une propriété clé de ces estimateurs, puisqu'il n'y aurait aucune raison d'en tenir compte s'ils n'étaient pas conformes au plan.

Malheureusement, les résultats fondés sur le plan ci-dessus sont très limités et, surtout, ils ne fournissent pas une distribution asymptotique qui permettrait de faire de l'inférence, autre propriété

hautement souhaitable des estimateurs d'enquête. Nous considérons donc également un contexte fondé sur un modèle, dans lequel nous pouvons obtenir une approximation asymptotique de la variance. En présentant des résultats fondés sur le modèle, nous supposons que le plan d'échantillonnage choisissant l'échantillon original A est un plan d'échantillonnage avec probabilités égales, et les caractéristiques de la population peuvent être considérées comme un échantillon iid d'une répartition de superpopulation. Dans ce contexte, l'estimateur agrégé par bootstrap peut être traité comme une statistique U. Nous pouvons alors appliquer la théorie des statistiques U pour obtenir une expansion asymptotique des estimateurs agrégés par bootstrap. L'analyse est comparable à celles de Bühlmann et Yu (2002) et de Buja et Stuetzle (2006). Aux fins du présent article, nous nous limitons aux échantillons bootstrap de taille k où k est bornée et fixe. Selon cette hypothèse, les estimateurs agrégés par bootstrap peuvent être considérés comme des statistiques U de degré fixe pour lesquelles une théorie asymptotique a été bien élaborée. Un cas plus intéressant survient lorsque la taille k du rééchantillonnage croît avec la taille de l'échantillon n, et que cela aboutit à des statistiques U de degré infini. Ces statistiques ont des applications dans l'étude de l'estimateur de Kaplan-Meier et des estimateurs bootstrap m-sur-n, et les lecteurs sont invités à consulter Frees (1989), Heilig (1997), Heilig et Nolan (2001) et leurs références sur les propriétés statistiques de ces estimateurs. Schick et Wefelmeyer (2004) ont étudié les propriétés des statistiques U de degré infini produites à partir des moyennes mobiles des innovations dans des séries chronologiques. L'étude des estimateurs agrégés par bootstrap considérés comme des statistiques U de degré infini dépasse la portée du présent article; nous nous limitons donc aux échantillons bootstrap de tailles fixes et bornées dans le cas fondé sur un modèle.

Considérons d'abord l'estimateur agrégé par bootstrap en (2.5). Sous l'EASSR, l'estimateur (2.5) peut être simplifié comme suit

$$\hat{\theta}_{nd} = \frac{1}{n}\sum_{i\in A} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right)$$

et la version agrégée par bootstrap de $\hat{\theta}_{nd}$ est définie comme étant

$$\hat{\theta}_{nd,bag} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right) \tag{3.1}$$

où $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ dépend seulement du nouvel échantillon A_b. Pour faciliter la présentation, nous prenons $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ comme moyenne de l'échantillon. Dans le cas qui nous intéresse, un simple calcul algébrique révèle que

$$\hat{\theta}_{nd,bag} = \frac{1}{\binom{n}{k}}\sum_{A_b\in\mathcal{A}}\left\{\frac{1}{k}\sum_{i\in A_b} h\left(\frac{k-1}{k}\mathbf{y}_i - \frac{1}{k}\sum_{j\neq i}\mathbf{y}_j\right)\right\},$$

où $\mathcal{A}$ représente les sous-ensembles de taille k de l'ensemble $\{1,2,\ldots,n\}$. L'estimateur $\hat{\theta}_{nd,bag}$ est une statistique U de degré k avec noyau

$$g\left(y_1,\ldots,y_k\right) = \frac{1}{k}\sum_{i=1}^{k} h\left(\frac{k-1}{k}\mathbf{y}_i - \frac{1}{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\mathbf{y}_j\right)$$

à condition que k reste finie.

On peut voir que l'estimateur agrégé par bootstrap $\hat{\theta}_{nd,bag}$ est une statistique symétrique de $\mathbf{y}_i$, et la théorie standard des statistiques symétriques (Lee 1990) s'applique. Les résultats sont énoncés dans le théorème 1, et les hypothèses et preuves figurent à l'annexe A.2.

Théorème 1 *Sous les hypothèses M.1 à M.4 concernant la répartition de superpopulation et les plans d'échantillonnage et de rééchantillonnage*

$$\mathrm{AV}\left(\hat{\theta}_{nd,bag}\right)^{-1/2}\left(\hat{\theta}_{nd,bag}-\theta_{nd,\infty}\right)\xrightarrow{p}N(0,1), \tag{3.2}$$

où la valeur limite $\theta_{nd,\infty}=\lim_{n\to\infty}\mathrm{E}\left[h\left(\mathbf{y}_i-\hat{\boldsymbol{\lambda}}\right)\right]$, *la variance asymptotique*

$$\mathrm{AV}\left(\hat{\theta}_{nd,bag}\right)=\frac{1}{n}\mathrm{Var}\left[u\left(\mathbf{y}_i\right)\right]+\frac{(k-1)^2}{n}\mathrm{Var}\left[v\left(\mathbf{y}_i\right)\right]+\frac{2(k-1)}{n}\mathrm{Cov}\left[u\left(\mathbf{y}_i\right),v\left(\mathbf{y}_i\right)\right], \tag{3.3}$$

et

$$u(\mathbf{y})=\mathrm{E}\left[h\left(\mathbf{y}-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y}\right)\right)\right],$$
$$v(\mathbf{y})=\mathrm{E}\left[h\left(\mathbf{y}_1-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y}\right)\right)\right].$$

Comme il est indiqué en (3.3), la variance asymptotique de l'estimateur agrégé par bootstrap dépend des fonctions inconnues $u(\mathbf{y})$ et $v(\mathbf{y})$, qui sont des espérances de $h(\cdot)$ en ce qui concerne la répartition de superpopulation. En $u(\mathbf{y})$ et $v(\mathbf{y})$, $\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y}\right)$ est calculé à partir de $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1}$ avec un vecteur arbitraire $\mathbf{y}$. L'espérance porte sur la distribution de vecteurs aléatoires *iid* $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1}$. Cette espérance de haute dimension est difficile à calculer et pourrait ne pas avoir une expression explicite en général. On ne peut pas obtenir la forme exacte de $u(\cdot)$ et $v(\cdot)$ mais on peut l'approximer en utilisant une approche fondée sur le rééchantillonnage. Les fonctions inconnues $u(\cdot)$ et $v(\cdot)$ sont définies comme étant des espérances de quantités respectives liées à la répartition de superpopulation, qui peuvent être approximées par l'espérance concernant la distribution empirique.

La variance asymptotique fondée sur un modèle peut être estimée dans le cadre du processus d'agrégation bootstrap. Nous pouvons calculer les intégrandes $h\left(\mathbf{y}-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*,\mathbf{y}\right)\right)$ et $h\left(\mathbf{y}_1-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*,\mathbf{y}\right)\right)$ en fonction de chaque échantillon bootstrap, $\mathbf{y}$ étant où nous voulons évaluer $u(\cdot)$ et $v(\cdot)$, et $\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*$ étant les valeurs rééchantillonnées. Nous pouvons alors calculer la moyenne de chaque quantité pour approximer l'espérance. Enfin, nous pouvons estimer la variance en calculant la variance d'échantillon des espérances évaluées à chacun des points d'échantillonnage. Pour les estimateurs non lisses comme ceux qui nous intéressent, il est souvent recommandé d'utiliser la méthode bootstrap lisse pour approximer la variance (Efron 1979; Davison et Hinkley 1997). Nous appliquons la méthode bootstrap lisse et ajoutons une petite quantité de bruit à chaque valeur rééchantillonnée afin de lisser la fonction sous-jacente. L'algorithme détaillé est expliqué au moyen d'un exemple dans la section 5.

Nous passons maintenant au résultat fondé sur un modèle des estimateurs agrégés par bootstrap définis par les équations d'estimation en (2.7). Un cas spécial dans ce contexte est l'agrégation bootstrap des

quantiles d'échantillon, qui a été étudiée par Knight et Bassett (2002). Knight et Bassett (2002) ont examiné le rééchantillonnage par bootstrap et par EASSR et étudié les effets de l'agrégation bootstrap sur le terme restant de la représentation des quantiles de Bahadur (Bahadur 1966). Nous abordons la question sous un angle légèrement différent et traitons l'estimateur agrégé par bootstrap comme une statistique U. Les hypothèses et les preuves figurent à l'annexe A.2. Il est à noter que, sous l'hypothèse M.5, la fonction d'estimation non différenciable doit avoir une limite lisse. Dans le théorème suivant, nous linéarisons l'estimateur de l'équation d'estimation agrégé par bootstrap et donnons une expression pour la variance asymptotique.

Théorème 2 *Sous les hypothèses M.1 à M.3 et M.5, le résultat asymptotique suivant tient pour l'estimateur de l'équation d'estimation agrégé par bootstrap (2.7),*

$$\mathrm{AV}\left(\hat{\theta}_{ee,bag}\right)^{-1/2}\left(\hat{\theta}_{ee,bag}-\theta_{ee,\infty}\right)\xrightarrow{p}N\left(0,1\right), \tag{3.4}$$

où $\theta_{ee,\infty}$ *est la limite asymptotique de la quantité de population* θ_{ee}, *la variance asymptotique de* $\hat{\theta}_{ee,bag}$ *est*

$$\mathrm{AV}\left(\hat{\theta}_{ee,bag}\right)=\frac{k^2}{n}\mathrm{Var}\left[u\left(y_i\right)\right], \tag{3.5}$$

et

$$u\left(y\right)=\mathrm{E}\ \inf\left\{\gamma:\frac{1}{k}\sum_{i=1}^{k-1}\psi\left(y_i-\gamma\right)+\frac{1}{k}\psi\left(y-\gamma\right)\geq 0\right\}. \tag{3.6}$$

Comme nous l'avons vu pour l'estimateur agrégé par bootstrap (3.1), les résultats asymptotiques du théorème 2 font intervenir une fonction inconnue. Là encore, cette fonction peut être calculée par rééchantillonnage des échantillons répétés disponibles.

4 Estimation de la variance

Bien que l'approche fondée sur un modèle permette d'obtenir des distributions asymptotiques et donc une inférence asymptotiquement correcte, nous nous intéressons surtout aux applications fondées sur le plan de l'agrégation bootstrap. En contexte fondé sur le plan de sondage, nous pouvons combiner naturellement la construction de l'estimateur agrégé par bootstrap avec l'estimation de la variance de la statistique originale, en tirant parti des échantillons répétés diffusés par les organismes statistiques. Dans le présent article, nous prenons l'exemple précis d'un échantillonnage aléatoire simple stratifié avec plan d'échantillonnage bootstrap d'un échantillon EASSR stratifié.

Nous commençons par appliquer une version de la procédure bootstrap de Rao et Wu (1988) afin d'estimer la variance des estimateurs d'enquête avant l'agrégation bootstrap. Soient N_h, n_h et k_h, la taille de la population, la taille de l'échantillon et la taille du sous-échantillon dans la strate h, $h=1,2,\ldots,H$. Ici, B échantillons bootstrap sont tirés par échantillonnage aléatoire simple stratifié sans remise de taille

k_h pour calculer la variance bootstrap de la statistique originale et l'estimateur agrégé par bootstrap. Pour chaque échantillon bootstrap, nous attribuons un poids de

$$\frac{N_h}{N}\left(1-k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\right)\frac{1}{n_h}+\frac{N_h}{N}k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\frac{1}{k_h}$$

à chaque élément échantillonné de la strate h, et

$$\frac{N_h}{N}\left(1-k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\right)\frac{1}{n_h}$$

aux éléments non échantillonnés. Nous utilisons ensuite la variance ordinaire des estimateurs d'échantillons répétés comme estimateur de variance. Le schéma de pondération susmentionné est algébriquement identique à l'équation 4.1 de Rao et Wu (1988), où la correction pour population finie est intégrée aux poids de rééchantillonnage. L'estimateur de variance par rééchantillonnage dérivé de la méthode de pondération se réduit à un estimateur de variance ordinaire sous l'EASSR stratifié et garantit l'absence de biais sous le plan. Afin de combiner l'agrégation bootstrap et l'estimateur de variance bootstrap, nous utilisons les mêmes échantillons bootstrap afin de construire les estimateurs agrégés par bootstrap pour les quantités de population qui nous intéressent.

Sous le cadre fondé sur le plan de sondage, aucun estimateur de variance analytique n'est disponible pour l'estimateur agrégé par bootstrap en général. Pour le moment, nous suggérons d'appliquer les deux approches suivantes d'estimation de la variance :

(Var. 1) Utiliser la variance estimée de l'estimateur original même si l'estimateur agrégé par bootstrap a une variance plus faible. Cette méthode produit des intervalles de confiance de même largeur, mais son taux de couverture est supérieur à celui de l'intervalle de confiance original.

(Var. 2) Multiplier la variance estimée de l'estimateur original par un facteur de correction tenant compte de l'amélioration probable de l'efficacité. Ce facteur pourrait être le gain d'efficacité si l'on présume que l'échantillon est un échantillon *iid* d'une superpopulation infinie. On peut déterminer le facteur en utilisant les résultats des théorèmes 1 et 2, ou par expérience bootstrap non paramétrique. Une procédure bootstrap possible est le bootstrap double, qu'on met en œuvre par rééchantillonnage bootstrap ordinaire afin d'estimer la variance de l'estimateur original, et un autre niveau de rééchantillonnage EASSR afin de déterminer la variance de l'estimateur agrégé par bootstrap. On peut estimer le ratio de la variance entre l'estimateur agrégé par bootstrap et l'estimateur original en utilisant ces échantillons bootstrap emboîtés, et multiplier la variance sous le plan de l'estimateur original par ce ratio.

Nous examinons les deux approches dans les simulations de la section 5, mais il s'agit clairement d'un domaine qui devrait faire l'objet de recherches plus approfondies.

5 Simulations

Pour évaluer le comportement pratique de l'agrégation bootstrap en contexte d'enquête, nous générons une population finie de taille $N = 2\ 000$ à trois strates. La taille de chaque strate est N_h où $h = 1,2,3,$ et les proportions des strates sont fixées à $(N_1; N_2; N_3)/N = (0,5;\ 0,3;\ 0,2)$. La distribution de la variable cible y_i dans chaque strate est $y_{1i} \sim |N(-1,1)|$, $y_{2i} \sim \Gamma(1,1)$ et $y_{3i} \sim |N(3,2)|$. Une variable auxiliaire x_i est générée par $x_i = A_0 + A_1 y_i + A_2(G_i - \alpha/\beta)$ où $A_0 = A_1 = 2,\ A_2 = 1,\ \alpha = 2,\ \beta = 1$ et $G_i \overset{iid}{\sim} \Gamma(2,1)$. Nous tirons de façon répétée des échantillons de taille n par échantillonnage aléatoire simple stratifié de la population d'intérêt et la répartition de la taille de l'échantillon est $(n_1; n_2; n_3)/n = (0,3; 0,3; 0,4)$. Dans ce contexte, le plan de sondage est clairement informatif, parce que les observations ne sont pas iid dans la population globale et sont corrélées avec les probabilités d'inclusion.

Nous nous intéressons à trois quantités de population : un quantile α de population, une proportion de la population inférieure à une fraction donnée d'un quantile de population (voir Berger et Skinner 2003, par exemple) et l'estimateur de Rao-Kovar-Mantel (RKM) de la fonction de répartition (Rao et coll. 1990). Le premier est un exemple d'un estimateur fondé sur une équation d'estimation non différenciable, tandis que les deux derniers sont des estimateurs non différenciables explicitement définis. L'estimateur sur échantillon du quantile est obtenu par inversion de la fonction de répartition cumulative estimée. L'estimateur sur échantillon de la proportion inférieure à une fraction donnée d'un quantile de population est l'estimateur HT de la proportion des observations inférieures à la médiane d'échantillon d'une variable d'intérêt multipliée par une constante c,

$$\hat{\theta}_{pr} = \frac{1}{N} \sum_{i \in A} \frac{1}{\pi_i} \mathrm{I}_{\left(y_i \le c\hat{\theta}_{\mathrm{med}}\right)},$$

où $\hat{\theta}_{\mathrm{med}}$ est la médiane d'échantillon de y_i. L'estimateur par la différence de RKM fondé sur le plan de sondage et sur un modèle de ratio est

$$\hat{\theta}_{\mathrm{RKM}} = \frac{1}{N} \left\{ \sum_{i \in A} \frac{1}{\pi_i} \mathrm{I}_{(y_i \le t)} + \sum_{i=1}^{N} \mathrm{I}_{\left(\hat{R}x_i \le t\right)} - \sum_{i \in A} \frac{1}{\pi_i} \mathrm{I}_{\left(\hat{R}x_i \le t\right)} \right\}, \tag{5.1}$$

où $\hat{R}$ est le ratio estimé entre y et x.

La variance sous le plan de ces estimateurs non différenciables est un peu fastidieuse à estimer. Pour le calcul des variances et des intervalles des quantiles d'échantillon, les lecteurs sont invités à consulter Francisco et Fuller (1991), Sitter et Wu (2001) et leurs références. Pour les proportions inférieures à un niveau estimé, voir Shao et Rao (1993) et Berger et Skinner (2003).

Les variances sous le plan des estimateurs originaux $\hat{\theta}_{qr}$, $\hat{\theta}_{pr}$ et $\hat{\theta}_{\mathrm{RKM}}$, sont estimées au moyen de la procédure bootstrap sans remise décrite dans la section qui précède. Nous employons une taille d'échantillon bootstrap de $k_h = n_h/2$. Les estimateurs agrégés par bootstrap ainsi construits sont souvent

qualifiés de « subagging estimators » (estimateurs sous-agrégés par bootstrap) (Bühlmann et Yu 2002). Il a été établi que les échantillons sans remise de taille $n/2$ produisent des résultats semblables à ceux des échantillons avec remise de taille n en agrégation bootstrap (Buja et Stuetzle 2006; Friedman et Hall 2007). Nous appliquons les deux approches de la variance pour les estimateurs agrégés par bootstrap proposées dans la section qui précède, c.-à-d. une approche identique à celle de l'estimateur non agrégé par bootstrap (Var. 1), et une autre qui multiplie l'estimation de la variance originale par un facteur de correction fondé sur un modèle (Var. 2). Le facteur est déterminé par bootstrap double sur un échantillon particulier. En principe, il faudrait répéter l'exercice pour chaque échantillon, mais ce scénario est exclu par le lourd fardeau de calcul. Les intervalles de confiance des trois estimateurs sont construits par approximation normale. Les intervalles de confiance pour la proportion et l'estimateur de RKM sont construits par approximation normale sur une échelle transformée *logit*, $\log\left[\hat{\theta}\big/\left(1-\hat{\theta}\right)\right]$ ou $\log\left[\hat{\theta}_{bag}\big/\left(1-\hat{\theta}_{bag}\right)\right]$, puis par rétrotransformation (Agresti 2002; Korn et Graubard 1998).

Le tableau 5.1 résume le biais, l'écart-type et le ratio de l'EQM des quantiles d'échantillon originaux et agrégés par bootstrap, tandis que le tableau 5.2 examine les estimateurs de la variance et les intervalles de confiance. Les tailles d'échantillon choisies sont $n = 100$ et 200. Dans le tableau 5.1, nous pouvons voir que l'estimateur de quantile agrégé par bootstrap est plus efficace que l'estimateur original puisque le ratio de l'EQM est inférieur à un dans cette expérience par simulation. En général, plus la taille de l'échantillon diminue, plus les effets de lissage de l'agrégation bootstrap deviennent évidents. Dans le tableau 5.2, nous comparons les deux intervalles de confiance avec estimateur de point d'agrégation bootstrap aux intervalles de confiance originaux. Comme prévu, l'intervalle de confiance construit selon la méthode 1 a la même longueur et une couverture plus étendue que l'original. Dans cet exemple, les intervalles de confiance construits selon la méthode 2 sont plus étroits, mais leur niveau de couverture reste proche du niveau nominal.

Tableau 5.1
Biais, écart-type et ratios de l'EQM des quantiles d'échantillon et des quantiles d'échantillon agrégé par bootstrap; taille de la population $N = 2\ 000$, nombre de bootstraps $B = 2\ 000$ et résultats de $2\ 000$ simulations

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
α	**0,2**	**0,3**	**0,5**	**0,7**	**0,8**	**0,2**	**0,3**	**0,5**	**0,7**	**0,8**
biais$\left(\hat{\theta}_{qt}\right)$	0,002	0,008	0,000	-0,005	-0,035	-0,008	0,005	0,006	0,007	-0,005
biais$\left(\hat{\theta}_{qt,bag}\right)$	0,018	0,019	-0,001	-0,007	-0,043	-0,006	0,009	0,005	0,006	-0,022
é-t$\left(\hat{\theta}_{qt}\right)$	0,093	0,124	0,149	0,181	0,212	0,070	0,076	0,103	0,136	0,148
é-t$\left(\hat{\theta}_{qt,bag}\right)$	0,089	0,112	0,138	0,167	0,197	0,065	0,073	0,099	0,127	0,139
$\dfrac{EQM_p\left(\hat{\theta}_{qt,bag}\right)}{EQM_p\left(\hat{\theta}_{qt}\right)}$	0,946	0,844	0,859	0,854	0,875	0,866	0,924	0,919	0,862	0,912

Tableau 5.2
Biais relatif, probabilité de couverture et largeur de l'intervalle de confiance des estimateurs de variance bootstrap pour les quantiles d'échantillon et des estimateurs de variance non corrigé $\left(\hat{V}_1\right)$ et corrigé $\left(\hat{V}_2\right)$ pour les quantiles d'échantillon agrégé par bootstrap; mêmes conditions de simulation qu'au tableau 5.1

	$n=100,\ k=50$					$n=200,\ k=100$				
α	**0,2**	**0,3**	**0,5**	**0,7**	**0,8**	**0,2**	**0,3**	**0,5**	**0,7**	**0,8**
$\frac{\mathrm{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{qt}\right)\right]}{V\left(\hat{\theta}_{qt}\right)}$	1,208	1,091	1,099	1,135	1,205	1,067	1,117	1,093	1,098	1,180
$\frac{\mathrm{E}\left[\hat{V}_1\left(\hat{\theta}_{qt,bag}\right)\right]}{V\left(\hat{\theta}_{qt,bag}\right)}$	1,327	1,325	1,279	1,331	1,402	1,224	1,217	1,188	1,273	1,326
$\frac{\mathrm{E}\left[\hat{V}_2\left(\hat{\theta}_{qt,bag}\right)\right]}{V\left(\hat{\theta}_{qt,bag}\right)}$	1,307	1,217	1,196	1,184	1,383	1,245	1,249	1,392	1,107	1,104
P.C.(I.C.)	0,944	0,934	0,924	0,928	0,922	0,938	0,951	0,942	0,935	0,950
$\mathrm{P.C.}\left(\mathrm{I.C.1}_{\cdot bag}\right)$	0,950	0,946	0,938	0,938	0,939	0,942	0,950	0,946	0,943	0,954
$\mathrm{P.C.}\left(\mathrm{I.C.2}_{\cdot bag}\right)$	0,949	0,934	0,932	0,929	0,938	0,944	0,952	0,958	0,927	0,936
Largeur(I.C.)										
$\mathrm{Largeur}\left(\mathrm{I.C.1}_{\cdot bag}\right)$	0,386	0,492	0,597	0,729	0,880	0,277	0,309	0,414	0,544	0,612
$\mathrm{Largeur}\left(\mathrm{I.C.2}_{\cdot bag}\right)$	0,383	0,472	0,577	0,688	0,874	0,279	0,313	0,448	0,508	0,559

Les tableaux 5.3 et 5.4 résument les résultats fondés sur le plan pour l'estimateur de la proportion de personnes à faible revenu. Le ratio de l'EQM montre que l'estimateur agrégé par bootstrap est uniformément plus efficace que l'estimateur original et que l'EQM de cet estimateur agrégé est inférieur à 50% de l'EQM de l'estimateur original dans certains cas (voir $c=1,2$). Cela est probablement dû au fait que l'estimateur comporte deux « niveaux » de non-différenciabilité : la médiane d'échantillon est un estimateur non différenciable dont le gain d'efficacité est illustré au tableau 5.1, et la proportion de personnes à faible revenu est une fonction non différenciable de la médiane d'échantillon. Les « sauts » des estimateurs sont lissés par agrégation bootstrap, ce qui donne un estimateur plus stable. La comparaison des intervalles de confiance au tableau 5.4 donne des résultats semblables à ceux obtenus pour les quantiles.

Tableau 5.3
Biais, écart-type et ratio de l'EQM de la proportion estimée inférieure à une constante c multipliée par une médiane estimée et l'estimateur de proportion agrégé par bootstrap; taille de la population $N=2\ 000$, nombre de bootstraps $B=2\ 000$ et résultats de $2\ 000$ simulations

	$n=100,\ k=50$					$n=200,\ k=100$				
c	**0,2**	**0,4**	**0,6**	**1,2**	**1,5**	**0,2**	**0,4**	**0,6**	**1,2**	**1,5**
$\mathrm{biais}\left(\hat{\theta}_{pr}\right)$	-0,002	-0,002	-0,003	0,011	0,006	0,000	-0,002	-0,005	-0,004	-0,004
$\mathrm{biais}\left(\hat{\theta}_{pr,bag}\right)$	-0,004	-0,004	-0,007	0,017	0,009	-0,001	-0,005	-0,009	-0,001	-0,004
$\text{é-t}\left(\hat{\theta}_{pr}\right)$	0,034	0,039	0,038	0,034	0,046	0,023	0,027	0,026	0,026	0,036
$\text{é-t}\left(\hat{\theta}_{pr,bag}\right)$	0,031	0,035	0,031	0,020	0,034	0,022	0,025	0,022	0,017	0,029
$\frac{EQM_p\left(\hat{\theta}_{pr,bag}\right)}{EQM_p\left(\hat{\theta}_{pr}\right)}$	0,861	0,821	0,709	0,538	0,581	0,883	0,860	0,783	0,434	0,671

Tableau 5.4
Biais relatif, probabilité de couverture et largeur de l'intervalle de confiance des estimateurs de variance bootstrap pour les proportions d'échantillon et des estimateurs de variance non corrigé $\left(\hat{V}_1\right)$ et corrigé $\left(\hat{V}_2\right)$ pour les proportions d'échantillon agrégé par bootstrap; mêmes conditions de simulation qu'au tableau 5.3. Le sigle « I.C.T. » est utilisé pour désigner les intervalles de confiance obtenus par transformation logit.

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
c	**0,2**	**0,4**	**0,6**	**1,2**	**1,5**	**0,2**	**0,4**	**0,6**	**1,2**	**1,5**
$\frac{\mathrm{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{pr}\right)\right]}{V\left(\hat{\theta}_{pr}\right)}$	1,122	1,191	1,325	1,472	1,281	1,140	1,191	1,251	1,350	1,217
$\frac{\mathrm{E}\left[\hat{V}_1\left(\hat{\theta}_{pr,bag}\right)\right]}{V\left(\hat{\theta}_{pr,bag}\right)}$	1,323	1,471	1,959	4,095	2,307	1,293	1,428	1,766	3,064	1,821
$\frac{\mathrm{E}\left[\hat{V}_2\left(\hat{\theta}_{pr,bag}\right)\right]}{V\left(\hat{\theta}_{pr,bag}\right)}$	1,240	0,963	1,190	1,174	1,149	1,145	1,262	1,319	2,039	1,524
P.C.(I.C.T.)	0,969	0,970	0,984	0,991	0,980	0,964	0,974	0,977	0,983	0,946
P.C.$\left(\text{I.C.T.1}_{\cdot bag}\right)$	0,979	0,983	0,995	0,998	0,995	0,974	0,980	0,988	0,998	0,976
P.C.$\left(\text{I.C.T.2}_{\cdot bag}\right)$	0,976	0,944	0,973	0,922	0,942	0,962	0,969	0,968	0,993	0,957
Largeur(I.C.T)										
Largeur$\left(\text{I.C.T.1}_{\cdot bag}\right)$	0,144	0,166	0,168	0,157	0,197	0,098	0,115	0,114	0,113	0,149
Largeur$\left(\text{I.C.T.2}_{\cdot bag}\right)$	0,139	0,134	0,131	0,085	0,140	0,093	0,108	0,099	0,092	0,136

Les tableaux 5.5 et 5.6 résument les résultats fondés sur le plan de sondage pour l'estimateur de RKM. Là encore, nous observons le gain d'efficacité en appliquant la méthode de l'agrégation bootstrap, et le gain se situe entre 2 % et 12 %. Les deux estimateurs de variance de la quantité agrégée par bootstrap donnent d'assez bons résultats. Les deux versions des intervalles de confiance pour les estimateurs agrégés par bootstrap ont des taux de couverture réels proches de 95 %, et les intervalles de confiance calculés selon l'approche du facteur de correction (Var. 2) sont légèrement plus courts que ceux calculés selon la méthode 1.

Tableau 5.5
Biais, écart-type et ratios de l'EQM de l'estimateur de RKM et de l'estimateur de RKM agrégé par bootstrap (5.1); taille de la population $N = 2\ 000$, nombre de bootstraps $B = 2\ 000$ et résultats de $2\ 000$ simulations

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
t	**0,5**	**1,5**	**2,5**	**3,5**	**4,5**	**0,5**	**1,5**	**2,5**	**3,5**	**4,5**
biais$\left(\hat{\theta}_{\text{RKM}}\right)$	0,000	0,000	0,000	0,000	0,000	-0,001	0,001	0,000	0,000	0,001
biais$\left(\hat{\theta}_{\text{RKM},bag}\right)$	-0,001	0,000	-0,001	0,000	0,000	-0,001	0,001	0,000	0,001	0,001
é-t$\left(\hat{\theta}_{\text{RKM}}\right)$	0,043	0,044	0,030	0,015	0,012	0,030	0,030	0,020	0,011	0,009
é-t$\left(\hat{\theta}_{\text{RKM},bag}\right)$	0,042	0,042	0,028	0,014	0,012	0,030	0,029	0,019	0,011	0,009
$\frac{EQM_p\left(\hat{\theta}_{\text{RKM},bag}\right)}{EQM_p\left(\hat{\theta}_{\text{RKM}}\right)}$	0,965	0,911	0,877	0,914	0,917	0,976	0,928	0,917	0,918	0,981

Tableau 5.6
Biais relatif, probabilité de couverture et largeur de l'intervalle de confiance des estimateurs de variance bootstrap pour l'estimateur de RKM (5.1) et des estimateurs de variance non corrigé $\left(\hat{V}_1\right)$ et corrigé $\left(\hat{V}_2\right)$ pour l'agrégation bootstrap des estimateurs de RKM; mêmes conditions de simulation qu'au tableau 5.5

	$n=100,\ k=50$					$n=200,\ k=100$				
t	**0,5**	**1,5**	**2,5**	**3,5**	**4,5**	**0,5**	**1,5**	**2,5**	**3,5**	**4,5**
$\frac{\mathrm{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{\mathrm{RKM}}\right)\right]}{V\left(\hat{\theta}_{\mathrm{RKM}}\right)}$	1,081	1,192	1,078	1,082	1,078	1,016	1,045	1,138	1,121	1,016
$\frac{\mathrm{E}\left[\hat{V}_1\left(\hat{\theta}_{\mathrm{RKM},bag}\right)\right]}{V\left(\hat{\theta}_{\mathrm{RKM},bag}\right)}$	1,115	1,324	1,183	1,198	1,156	1,038	1,138	1,223	1,210	1,062
$\frac{\mathrm{E}\left[\hat{V}_2\left(\hat{\theta}_{\mathrm{RKM},bag}\right)\right]}{V\left(\hat{\theta}_{\mathrm{RKM},bag}\right)}$	1,087	1,117	0,962	1,042	1,019	1,009	1,083	1,106	1,118	1,002
P.C.(I.C.)	0,958	0,963	0,955	0,956	0,959	0,954	0,956	0,966	0,964	0,948
P.C.$\left(\mathrm{I.C.1}_{\cdot bag}\right)$	0,958	0,968	0,958	0,967	0,964	0,958	0,964	0,970	0,970	0,956
P.C.$\left(\mathrm{I.C.2}_{\cdot bag}\right)$	0,957	0,954	0,937	0,951	0,950	0,955	0,958	0,959	0,960	0,948
Largeur(I.C.)										
Largeur$\left(\mathrm{I.C.1}_{\cdot bag}\right)$	0,171	0,183	0,116	0,074	0,052	0,122	0,122	0,083	0,049	0,034
Largeur$\left(\mathrm{I.C.2}_{\cdot bag}\right)$	0,169	0,168	0,105	0,069	0,049	0,120	0,120	0,079	0,047	0,033

Dans le contexte des estimateurs non lisses comme ceux considérés ici, il est souvent recommandé d'utiliser un bootstrap lisse plutôt qu'un bootstrap simple pour estimer la variance. Nous avons envisagé de perturber chaque observation rééchantillonnée y_{hi}^* de la strate h pour obtenir

$$\tilde{y}_{hi}^* = \bar{y}_h + \left(1+\sigma_Z^2\right)^{-1/2}\left(y_{hi}^* - \bar{y}_h + s_h Z^*\right), \tag{5.2}$$

où $\bar{y}_h$ et s_h sont la moyenne de l'échantillon et l'écart-type de la strate de l'échantillon original, y_{hi}^* est la valeur rééchantillonnée à l'origine, et Z^* est le bruit aléatoire où $Z^* \overset{iid}{\sim} N\left(0,\sigma_Z^2\right)$. La variance de Z^* contrôle le degré de lissage. Nous avons appliqué cette méthode à l'estimation du quantile et la proportion inférieure à un niveau estimé, mais cela n'a pas semblé améliorer l'efficacité de la procédure d'estimation. Une explication possible est que la contamination par le bruit déstabilise les observations répétées découlant de l'échantillon avec remise et stabilise l'estimateur de variance subséquent dans une certaine mesure. Comme nous avons utilisé la méthode d'échantillonnage sans remise, nous avons évité ce problème en grande partie. Une étude plus approfondie est nécessaire pour comprendre les effets du lissage dans ce contexte.

6 Conclusions

Dans le présent article, nous avons examiné l'utilisation des procédures d'agrégation bootstrap pour les estimateurs d'enquête non linéaires et non différenciables. Nous avons présenté des résultats théoriques de

l'agrégation bootstrap des estimateurs fondés sur le plan de sondage et de ceux fondés sur le modèle. L'estimateur agrégé par bootstrap peut être traité comme l'espérance d'un estimateur à deux phases avec conditionnement sur la première phase, et cette espérance lisse les « sauts » de l'estimateur non différenciable. L'étude empirique révèle le potentiel de l'agrégation bootstrap des estimateurs d'enquête non différenciables. Bien que l'efficacité relative de l'agrégation bootstrap varie d'un scénario à l'autre, les résultats sont prometteurs.

Il reste à déterminer comment estimer la variance des estimateurs agrégés par bootstrap lorsque le plan d'échantillonnage est généralement complexe. Nous avons proposé deux méthodes d'estimation de la variance à des fins pratiques, mais une étude théorique plus approfondie de l'estimation de la variance dans un cadre fondé sur le plan de sondage serait certainement justifiée.

Annexe

A.1 Théorie fondée sur le plan de sondage

Les hypothèses D.1 à D.6 sont utilisées pour illustrer les résultats fondés sur le plan de sondage figurant ci-dessous (théorèmes 3 et 4). L'hypothèse D.1 spécifie les conditions de moments sur la variable étudiée $\mathbf{y}_i$, tandis que l'hypothèse D.2 spécifie les conditions sur la probabilité d'inclusion de second ordre du plan d'échantillonnage. L'hypothèse D.3 garantit que la taille de chaque rééchantillonnage converge à la limite à l'infini. L'hypothèse D.4 spécifie les conditions de lissage sur $m(\cdot)$ dans l'estimateur différenciable. Les hypothèses D.5-D.6 montrent la convergence par rapport au plan de l'agrégation bootstrap des estimateurs d'enquête non différenciables.

(D.1) La variable étudiée $\mathbf{y}_i$ a un moment de population finie $2+\delta$ pour une valeur arbitrairement petite $\delta > 0$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{y}_i^{2+\delta} \| < \infty,$$

où chaque élément de $\mathbf{y}_i^{2+\delta}$ est l'élément original élevé à la puissance $2+\delta$ et $\|\cdot\|$ est la norme euclidienne.

(D.2) Pour tous les N, $\min_{i\in U_N} \pi_i \geq \pi_N^* > 0$, où $N\pi_N^* \to \infty$, et

$$\lim_{N\to\infty} \sup n \cdot \max \left| \pi_{ij} - \pi_i \pi_j \right| < \infty,$$

où π_{ij} est la probabilité d'inclusion conjointe des éléments i, j.

(D.3) Le processus de rééchantillonnage générant A_b est l'EASSR de taille k, où $k = O\left(n^{\kappa}\right)$, $\kappa \in (0,1]$. De plus, chaque nouvel échantillon bootstrap de taille k est utilisé pour calculer l'estimateur agrégé par bootstrap.

(D.4) La fonction $m(\cdot)$ est différenciable et a une dérivée seconde continue non triviale dans un voisinage compact de $\boldsymbol{\mu}_N$.

(D.5) L'estimateur $\hat{\boldsymbol{\lambda}}$ converge vers la cible de population $\boldsymbol{\lambda}_N$ à une vitesse de $\sqrt{n}$, $\lim_{N\to\infty}\boldsymbol{\lambda}_N = \boldsymbol{\lambda}_\infty$ et l'estimateur $\hat{\boldsymbol{\lambda}}$ est une statistique symétrique.

(D.6) La fonction $h(\cdot)$ est bornée et la quantité de population est [*traduction*] « différenciable de manière compacte dans un sens faible » (Dümbgen 1993). Il existe une fonction $g(\cdot)$ telle que

$$\sup_{\mathbf{s}\in C_{\mathbf{s}}}\left|\frac{1}{N}\sum_{i=1}^{N}h\left(\mathbf{y}_i-\boldsymbol{\lambda}_\infty-N^{-\alpha}\mathbf{s}\right)-\frac{1}{N}\sum_{i=1}^{N}h\left(\mathbf{y}_i-\boldsymbol{\lambda}_\infty\right)-g\left(\boldsymbol{\lambda}_\infty\right)N^{-\alpha}\mathbf{s}\right|\to 0,$$

où $C_{\mathbf{s}}$ est un ensemble compact suffisamment important de $\mathbb{R}^p$, $0<\alpha\le 1/2$ et $g\left(\boldsymbol{\lambda}_\infty\right)$ est borné.

Le théorème suivant donne plusieurs approximations asymptotiques de l'estimateur agrégé par bootstrap, selon le taux de convergence de k par rapport à n. Dans les trois cas, l'estimateur agrégé par bootstrap est conforme au plan. Intuitivement, l'estimateur agrégé par bootstrap se comporte comme l'estimateur original lorsque la taille k du rééchantillonnage est importante (tend vers l'infini à une vitesse d'au moins $n^{1/2}$), mais converge à une vitesse différente lorsque le rééchantillonnage est de petite taille.

Théorème 3 *Sous les hypothèses D.1 à D.4, l'estimateur différenciable agrégé par bootstrap $\hat{\theta}_{d,bag}$ admet l'expansion de deuxième ordre suivante :*

$$\hat{\theta}_{d,bag}-\theta_d=\begin{cases}\left\{m'\left(\boldsymbol{\mu}_N\right)\right\}^T\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_N\right)+o_p\left(n^{-1/2}\right), & \textit{pour}\ \ \kappa>1/2\\ \left\{m'\left(\boldsymbol{\mu}_N\right)\right\}^T\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_N\right)+ & \\ \qquad\dfrac{1}{2\binom{n}{k}}\sum\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)^T m''\left(\boldsymbol{\mu}_N\right)\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)+o_p\left(n^{-1/2}\right), & \textit{pour}\ \kappa=1/2\\ \dfrac{1}{2\binom{n}{k}}\sum\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)^T m''\left(\boldsymbol{\mu}_N\right)\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)+o_p\left(k^{-1}\right), & \textit{pour}\ \ \kappa<1/2\end{cases}$$

où $\kappa>0$ est tel que la taille du rééchantillonnage $k=O\left(n^{\kappa}\right)$.

Preuve du théorème 3 :

La preuve découle facilement d'une expansion de Taylor de chaque estimateur fondé sur un rééchantillonnage $m\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)\right)$ autour de $\boldsymbol{\mu}_N$. Le terme d'expansion linéaire se réduit à

$\left\{m'\left(\boldsymbol{\mu}_N\right)\right\}^T\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_N\right)$ sur la base d'un argument antérieur. Sous D.1 et D.3, le terme quadratique a le même ordre que la variance de l'EASSR de $\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)$ et est donc $o_p\left(1/k\right)$.

Le théorème 4 donne ensuite la convergence par rapport au plan de l'estimateur agrégé par bootstrap non différenciable.

Théorème 4 *Sous les hypothèses D.1-D.3 et D.5-D.6, l'estimateur agrégé par bootstrap non différenciable* $\hat{\theta}_{nd,bag}$ *est conforme au plan pour sa cible de population* θ_{nd}, *i.e.,* $\hat{\theta}_{nd,bag}-\theta_{nd}=o_p\left(1\right)$.

Preuve du théorème 4 :

Nous pouvons établir que $\left(1/N\right)\sum_{i\in A}\left(1/\pi_i\right)h\left(\mathbf{y}_i-\boldsymbol{\lambda}_N\right)$ est conforme au plan pour θ_{nd} en conséquence de D.2 et que $h(\cdot)$ est borné (D.6). Il suffit alors de démontrer que $\hat{\theta}_{nd,bag}-\left(1/N\right)\sum_{i\in A}\left(1/\pi_i\right)h\left(\mathbf{y}_i-\boldsymbol{\lambda}_N\right)=o_p\left(1\right)$, ou

$$\frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\left\{\frac{1}{\binom{n-1}{k-1}}\sum_{A_b\ni i}h\left(\mathbf{y}_i-\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)-h\left(\mathbf{y}_i-\boldsymbol{\lambda}_N\right)\right\}=o_p\left(1\right)$$

suivant (2.6). Nous pouvons établir que l'ensemble d'estimateurs fondés sur le rééchantillonnage $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ est uniformément contenu dans un voisinage de $\boldsymbol{\lambda}_N$, ou, $\sup_{A_b}\left|\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\lambda}_N\right|=O\left(N^{-\alpha}\mathbf{s}\right)$ pour certains $\alpha>0$. Nous pouvons alors appliquer D.6 pour conclure à la convergence par rapport au plan de l'estimateur agrégé par bootstrap.

A.2 Théorie fondée sur le modèle

Les hypothèses M.1 à M.4 sont utilisées pour montrer les résultats fondés sur le modèle (théorèmes 1 et 2). L'hypothèse M.1 spécifie la répartition de superpopulation des caractéristiques de population $\mathbf{y}_i$. Les hypothèses M.2 et M.3 supposent un échantillonnage aléatoire simple sans remise pour le plan de sondage et le processus de rééchantillonnage. L'hypothèse M.5 est requise pour montrer les résultats asymptotiques fondés sur un modèle pour l'estimateur agrégé par bootstrap défini par les équations d'estimation.

(M.1) La séquence de caractéristiques de population $\mathbf{y}_i$ constitue un échantillon *iid* d'une distribution de probabilités de densité $f_Y\left(\mathbf{y}\right)$.

(M.2) Le plan d'échantillonnage est ignorable ou, ce qui revient au même, les observations échantillonnées et non échantillonnées sont distribuées de la même façon.

(M.3) Le processus de rééchantillonnage générant A_b est l'EASSR de taille k, où la taille de l'échantillon bootstrap k est bornée. De plus, chaque nouvel échantillon bootstrap de taille k est utilisé pour calculer l'estimateur agrégé par bootstrap.

(M.4) La fonction $h(\cdot)$ est bornée.

(M.5) Soit $S_\infty(\gamma) = \mathrm{E}\psi(y_i - \gamma)$ une fonction continue de γ, et $\theta_{ee,\infty}$ la plus petite racine de $S_\infty(\gamma) = 0$; pour un y arbitraire à l'appui de la variable aléatoire y_i, la quantité

$$\inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k-1}\psi(y_i - \gamma) + \frac{1}{k}\psi(y - \gamma) \geq 0\right\}$$

appartient à un ensemble compact avec probabilité 1.

Preuve du théorème 1 :

L'estimateur agrégé par bootstrap $\hat{\theta}_{nd,bag}$ est une statistique symétrique, à condition que $\hat{\boldsymbol{\lambda}}$ soit symétrique (Lee 1990). Nous pouvons le projeter sur une seule dimension, disons $\mathbf{y}_1$, mais les projections sur d'autres observations sont équivalentes en raison de la symétrie,

$$\begin{aligned}
&\mathrm{E}\left\{\hat{\theta}_{nd,bag}\middle|\mathbf{y}_1 = \mathbf{y}\right\} \\
&= \mathrm{E}\left\{\frac{1}{n}\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni 1} h\left(\mathbf{y}_1 - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\middle|\mathbf{y}_1 = \mathbf{y}\right\} + \mathrm{E}\left\{\frac{n-1}{n}\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni \{i,1\}, i \neq 1} h\left(\mathbf{y}_1 - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\middle|\mathbf{y}_1 = \mathbf{y}\right\} \\
&= \frac{1}{n}u(\mathbf{y}) + \frac{k-1}{n}v(\mathbf{y}).
\end{aligned}$$

Nous pouvons alors dériver la linéarisation suivante de l'estimateur agrégé par bootstrap en appliquant la théorie des statistiques symétriques,

$$\hat{\theta}_{nd,bag} - \theta_{nd,\infty} = \frac{1}{n}\sum_{i=1}^{n}\left\{u(\mathbf{y}_i) - \theta_{nd,\infty}\right\} + \frac{k-1}{n}\sum_{i=1}^{n}\left\{v(\mathbf{y}_i) - \theta_{nd,\infty}\right\} + o_p\left(n^{-1/2}\right),$$

où $u(\cdot)$, $v(\cdot)$ et $\theta_{nd,\infty}$ sont définies dans le théorème 1. Il est facile de calculer la variance asymptotique (3.3) étant donné l'hypothèse d'échantillonnage *iid* .

Preuve du théorème 2 :

L'estimateur agrégé par bootstrap défini en (2.7) peut être traité comme une statistique U d'ordre k d'un échantillon, avec fonction noyau

$$h(y_1, y_2, \ldots, y_k) = \inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k}\psi(y_i - \gamma) \geq 0\right\}.$$

Nous pouvons appliquer directement une formule bien connue pour linéariser la statistique U (Serfling 1980; van der Vaart 1998, p. 161) afin d'obtenir la linéarisation

$$\hat{\theta}_{ee,bag} - \theta_{ee,\infty} = \frac{k}{n}\sum_{i=1}^{n}\left\{u\left(y_i\right) - \theta_{ee,\infty}\right\} + o_p\left(n^{-1/2}\right),$$

où

$$\begin{aligned} u\left(y\right) &= \mathrm{E}\ h\left(y, y_1, y_2, \ldots, y_{k-1}\right) \\ &= \mathrm{E}\ \inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k-1}\psi\left(y_i - \gamma\right) + \frac{1}{k}\psi\left(y - \gamma\right) \geq 0\right\}. \end{aligned}$$

L'estimateur d'équation d'estimation agrégé par bootstrap (2.7) peut être linéarisé comme suit :

$$\hat{\theta}_{ee,bag} - \theta_{ee,\infty} = \frac{k}{n}\sum_{i=1}^{n}\left\{u\left(y_i\right) - \theta_{ee,\infty}\right\} + o_p\left(n^{-1/2}\right). \tag{A.1}$$

La variance asymptotique de $\hat{\theta}_{ee,bag}$ peut être obtenue directement de la linéarisation (A.1).

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition, New York: John Wiley and Sons.

Bahadur, R.R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37, 577-580.

Beran, R. et Srivastava, M. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, 13, 95-115.

Berger, Y.G. et Skinner, C.J. (2003). Variance estimation for a low income proportion. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 52 (4), 457-468.

Bickel, P., Götze, F. et van Zwet, W. (1997). Resampling fewer than n observations: gains, losses and remedies for losses. *Statistica Sinica*, 7, 1-31.

Breidt, F. et Opsomer, J. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Annals of Statistics*, 36, 403-427.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Bühlmann, P. et Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30 (4), 927-961.

Buja, A. et Stuetzle, W. (2006). Observations on bagging. *Statistica Sinica*, 16 (2), 323-351.

Chen, S.X. et Hall, P. (2003). Effects of bagging and bias correction on estimators defined by estimating equations. *Statistica Sinica*, 13 (1), 97-109.

Davison, A. et Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95, 125-140.

Dunstan, R. et Chambers, R.L. (1986). Model-based confidence intervals in multipurpose surveys. *Applied Statistics*, 35, 276-280.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1-26.

Francisco, C.A. et Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.

Frees, E.W. (1989). Infinite order U-statistics. *Scandinavian Journal of Statistics*, 16, 29-45.

Friedman, J.H. et Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137 (3), 669-683.

Fuller, W. (2009). *Sampling Statistics*. John Wiley and Sons.

Godambe, V. et Thompson, M. (2009). Estimating functions and survey sampling. Dans C. Rao et D. P. (éditeurs) (Eds.), *Handbook of Statistics, vol. 29: Sample Surveys: Inference and Analysis*, 669-687. Elsevier/North-Holland.

Hall, P. et Robinson, A. (2009). Reducing variability of crossvalidation for smoothing parameter choice. *Biometrika*, 96, 175-186.

Heilig, C.M. (1997). *An Empirical Process Approach to U-processes of Increasing Degree*. Thèse de doctorat, University of California, Berkeley.

Heilig, C.M. et Nolan, D. (2001). Limit theorems for the infinite-degree U-process. *Statistica Sinica*, 11, 289-302.

Inoue, A. et Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association*, 103 (482), 511-522.

Knight, K. et Bassett, J.G. (2002). Second order improvements of sample quantiles using subsamples. Unpublished manuscript.

Korn, E.L. et Graubard, B.I. (1998). Intervalles de confiance pour les proportions à petit nombre d'évènements positifs prévus estimées au moyen des données d'enquête. *Techniques d'enquête*, 24 (2), 209-218.

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

Lee, T.-H. et Yang, Y. (2006). Bagging binary and quantile predictors for time series. *Journal of Econometrics*, 135 (1-2), 465-497.

Rao, J., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Rao, J. et Wu, C. (1988). Resampling inference with complex surveys. *Journal of the American Statistical Association*, 83 (401), 231-241.

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1997). *Model Assisted Survey Sampling*. Springer-Verlag Inc (Berlin; New York).

Schick, A. et Wefelmeyer, W. (2004). Estimating invariant laws of linear processes by U-statistics. *The Annals of Statistics*, 32, 603-632.

Sering, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons.

Shao, J. et Rao, J. (1993). Standard errors for low income proportions estimated from stratified multistage samples. *Sankhya B,* 55, 393-414.

Sitter, R.R. et Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics & Probability Letters*, 52 (4), 353-358.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wang, J.C. et Opsomer, J.D. (2011). On the asymptotic normality and variance estimation of nondifferentiable survey estimators. *Biometrika*, 98, 91-106.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 235-256
Statistique Canada, n° 12-001-X au catalogue

Imputation fractionnaire hot deck pour une inférence robuste sous un modèle de non-réponse partielle en échantillonnage

Jae Kwang Kim et Shu Yang[1]

Résumé

L'imputation fractionnaire paramétrique (IFP) proposée par Kim (2011) est un outil d'estimation des paramètres à usage général en cas de données manquantes. Nous proposons une imputation fractionnaire hot deck (IFHD), qui est plus robuste que l'IFP ou l'imputation multiple. Selon la méthode proposée, les valeurs imputées sont choisies parmi l'ensemble des répondants, et des pondérations fractionnaires appropriées leur sont assignées. Les pondérations sont ensuite ajustées pour répondre à certaines conditions de calage, ce qui garantit l'efficacité de l'estimateur IFHD résultant. Deux études de simulation sont présentées afin de comparer la méthode proposée aux méthodes existantes.

Mots-clés : Algorithme EM; information de Kullback-Leibler; valeurs manquant au hasard; imputation multiple

1 Introduction

L'imputation est une méthode courante de compensation de la non-réponse partielle dans les enquêtes sur échantillon. Soit y la variable étudiée sujette à la non-réponse et $\mathbf{x}$ le vecteur des variables auxiliaires complètement observées. On utilise souvent un modèle de distribution conditionnelle $f(y|\mathbf{x})$ afin de générer des valeurs imputées pour la donnée y_i manquante. Cette méthode d'imputation fondée sur un modèle a fait l'objet de nombreuses études. L'imputation multiple de Rubin (1987) est une approche bayésienne d'imputation fondée sur un modèle. L'algorithme EM Monte Carlo de Wei et Tanner (1990) peut être traité comme une approche fréquentiste d'imputation fondée sur un modèle. Kim (2011) proposait une imputation fractionnaire paramétrique pour traiter les données manquantes multivariées.

Cependant, la méthode d'imputation fondée sur un modèle qui génère des valeurs imputées à partir de $f(y|\mathbf{x})$ n'est pas une imputation hot deck en ce sens que les valeurs artificielles sont construites après l'imputation. Une caractéristique souhaitable de l'imputation hot deck est que toutes les valeurs imputées sont des valeurs observées. Par exemple, les valeurs imputées pour des variables catégoriques seront elles aussi catégoriques et le nombre de catégories est le même que celui observé pour les répondants. Pour cette raison, l'imputation hot deck est la méthode d'imputation la plus populaire, particulièrement dans les enquêtes-ménages. L'imputation par la méthode du plus proche voisin est une autre imputation hot deck. Chen et Shao (2001), Beaumont et Bocci (2009), Kim, Fuller et Bell (2011) ont eux aussi examiné l'imputation par la méthode du plus proche voisin en contexte d'échantillonnage. Durrant (2009), Haziza (2009) et Andridge et Little (2010) ont donné des aperçus détaillés des méthodes d'imputation hot deck en échantillonnage.

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011. Courriel : jkim@iastate.edu; Shu Yang, Department of Statistics, Iowa State University, Ames, IA 50011.

Kalton et Kish (1984) ont proposé une imputation fractionnaire hot deck afin d'assurer l'efficacité de l'imputation hot deck. Kim et Fuller (2004) et Fuller et Kim (2005) ont soumis l'imputation fractionnaire hot deck à un examen rigoureux et examiné l'estimation de la variance. Cependant, leur approche s'applique seulement lorsque $\mathbf{x}$ est catégorique. Pour les covariables continues, l'appariement d'après la moyenne prédictive peut être traité comme une méthode d'imputation par le plus proche voisin fondée sur la valeur prédite obtenue à partir de $f(y \mid \mathbf{x})$, mais ses propriétés statistiques ne sont pas traitées de façon approfondie dans la littérature.

Dans le présent article, nous proposons une nouvelle méthode d'imputation fractionnaire hot deck (IFHD) fondée sur un modèle paramétrique de $f(y \mid \mathbf{x})$ qui permet des covariables continues. La méthode proposée présente plusieurs avantages par rapport aux méthodes existantes. Premièrement, cette imputation hot deck préserve la structure de corrélation entre les éléments. Deuxièmement, elle est robuste en ce sens que l'estimateur résultant est moins sensible à l'échec du modèle théorique $f(y \mid \mathbf{x})$. Troisièmement, elle fournit des estimateurs de variance convergents pour différents paramètres sans exiger la condition de compatibilité de Meng (1994). L'imputation multiple exige toutefois la condition de compatibilité pour valider l'estimation de la variance. Lorsque la condition de compatibilité n'est pas satisfaite, l'imputation multiple donne souvent lieu à une inférence prudente qui, à son tour, réduit la puissance des tests. Voir la section 5.2 pour plus de détails.

La présentation de l'article suit. Dans la section 2, nous décrivons la configuration de base. La méthode proposée est présentée dans la section 3. La robustesse de l'IFHD est traitée dans la section 4. Dans la section 5, nous présentons les résultats de deux études par simulation et, dans la section 6, nous formulons nos conclusions.

2 Configuration de base

Considérons une population finie de N éléments identifiés par un ensemble d'indices $U = \{1, 2, \ldots, N\}$, où N est connu. À chaque unité i de la population sont associées les variables étudiées $\mathbf{x}_i$ et y_i, où $\mathbf{x}_i$ est toujours observée et y_i est sujette à la non-réponse. Soit A l'ensemble d'indices pour les éléments d'un échantillon sélectionné par échantillonnage probabiliste. Nous voulons estimer η, définie comme étant une solution (unique) à l'équation d'estimation de population $\sum_{i=1}^{N} U(\eta; \mathbf{x}_i, y_i) = 0$. Par exemple, la moyenne de population peut être obtenue en posant que $U(\eta; \mathbf{x}_i, y_i) = \eta - y_i$. Sous réponse complète, un estimateur convergent de η est obtenu en résolvant

$$\sum_{i \in A} w_i U(\eta; \mathbf{x}_i, y_i) = 0, \tag{2.1}$$

où $w_i = \{Pr(i \in A)\}^{-1}$ est l'inverse de la probabilité d'inclusion d'ordre un de l'unité i. Binder et Patak (1994) et Rao, Yung et Hidiroglou (2002) ont examiné les propriétés asymptotiques de l'estimateur obtenu au moyen de l'équation (2.1). Lorsqu'il manque des données, nous définissons

$$\delta_i = \begin{cases} 1 & \text{si } y_i \text{ est observée;} \\ 0 & \text{sinon.} \end{cases}$$

Nous obtenons alors un estimateur convergent de η en prenant l'espérance conditionnelle et en résolvant

$$\sum_{i\in A} w_i \left[\delta_i U\left(\eta;\mathbf{x}_i,y_i\right) + \left(1-\delta_i\right) E\left\{ U\left(\eta;\mathbf{x}_i,Y\right) \mid \mathbf{x}_i,\delta_i=0 \right\} \right] = 0 \tag{2.2}$$

pour η. L'équation d'estimation (2.2) est parfois qualifiée d'équation d'estimation prévue (Wang et Pepe 2000).

Pour calculer l'espérance conditionnelle en (2.2), nous supposons que la population finie étudiée est réalisée à partir d'une population infinie appelée superpopulation. Dans le modèle de la superpopulation, nous postulons souvent une distribution conditionnelle paramétrique de y étant donnée $\mathbf{x}$, $f\left(y \middle| \mathbf{x};\theta\right)$, qui est connue jusqu'au paramètre θ dans l'espace des paramètres Ω. Sous le modèle spécifié, nous pouvons calculer un estimateur convergent $\hat{\theta}$ de θ puis utiliser une méthode Monte Carlo pour évaluer l'espérance conditionnelle en (2.2) étant donné l'estimation $\hat{\theta}$. Si le mécanisme de réponse est manquant au hasard ou est ignorable au sens de Rubin (1976), nous pouvons approximer ainsi l'équation d'estimation prévue en (2.2)

$$\sum_{i\in A} w_i \left\{ \delta_i U\left(\eta;\mathbf{x}_i,y_i\right) + \left(1-\delta_i\right) \frac{1}{m} \sum_{j=1}^{m} U\left(\eta;\mathbf{x}_i,y_i^{*(j)}\right) \right\} = 0, \tag{2.3}$$

où

$$y_i^{*(1)},\ldots,y_i^{*(m)} \overset{i.i.d.}{\sim} f\left(y_i \mid \mathbf{x}_i;\hat{\theta}\right).$$

Nous utilisons souvent l'estimateur du maximum de vraisemblance $\hat{\theta}$, ce qui résout

$$S\left(\theta\right) = \sum_{i\in A} w_i \delta_i S\left(\theta;\mathbf{x}_i,y_i\right) = 0, \tag{2.4}$$

où $S\left(\theta;\mathbf{x},y\right) = \partial \log f\left(y \mid \mathbf{x};\theta\right) / \partial\theta$. Il est à noter que nous utilisons les poids d'échantillonnage w_i dans l'équation de score (2.4). Nous supposons implicitement que le modèle d'imputation, qui génère les valeurs imputées, est le modèle concernant les valeurs de la population finie $f\left(y_i \mid \mathbf{x}_i\right)$, plutôt que les valeurs d'échantillon. Nous convenons ainsi que le mécanisme d'échantillonnage peut être informatif dans le sens de Pfeffermann (2011). Par contraste, l'imputation multiple utilise le modèle d'échantillon, $f_s\left(y_i \mid \mathbf{x}_i\right) \equiv f\left(y_i \mid \mathbf{x}_i, i\in A\right)$, pour générer les valeurs imputées et suppose souvent que le mécanisme d'échantillonnage n'est pas informatif. Ainsi, l'imputation multiple suppose que les données sont manquantes au hasard dans l'échantillon étudié tandis que l'imputation fractionnaire suppose que les données sont manquantes au hasard dans la population. Sous un plan d'échantillonnage informatif, la génération de valeurs imputées à partir du modèle d'échantillon $f_s\left(y_i \mid \mathbf{x}\right)$ ne mène pas nécessairement à une inférence valide même quand la condition de données manquant au hasard dans l'échantillon est remplie. Voir la section 8.4 de Kim et Shao (2013) pour un examen plus approfondi des données manquant au hasard sous échantillonnage informatif.

L'imputation fractionnaire paramétrique (IFP) de Kim (2011) peut être utilisée pour calculer l'espérance conditionnelle en (2.2) de manière efficace. En IFP, les valeurs imputées sont générées à partir d'une distribution proposée acceptable $h\left(y \mid \mathbf{x}_i\right)$ puis l'équation d'estimation imputée (2.3) est remplacée par

$$\sum_{i\in A} w_i \left\{ \delta_i U\left(\eta;\mathbf{x}_i, y_i\right) + \left(1-\delta_i\right)\sum_{j=1}^{m} w_{ij}^* U\left(\eta;\mathbf{x}_i, y_i^{*(j)}\right)\right\} = 0, \tag{2.5}$$

où

$$w_{ij}^* = \frac{f\left(y_i^{*(j)} \mid \mathbf{x}_i;\hat{\theta}\right)\Big/ h\left(y_i^{*(j)} \mid \mathbf{x}_i\right)}{\sum_{k=1}^{m}\left\{ f\left(y_i^{*(k)} \mid \mathbf{x}_i;\hat{\theta}\right)\Big/ h\left(y_i^{*(k)} \mid \mathbf{x}_i\right)\right\}}. \tag{2.6}$$

Le choix de distribution proposée $h(\cdot)$ est un peu arbitraire. Nous examinerons un choix particulier qui pourrait mener à une estimation robuste.

La convergence de l'estimateur $\hat{\eta}$ résultant de (2.3) ou (2.5) peut être établie en supposant que la distribution conditionnelle $f\left(y \mid \mathbf{x};\theta\right)$ est correctement spécifiée (selon un argument semblable à celui utilisé dans la preuve du corollaire II.2 d'Andersen et Gill (1982), et la preuve n'est pas faite ici). Dans le présent article, nous examinons un différent type d'imputation fractionnaire qui est plus robuste en cas d'échec de l'hypothèse du modèle d'imputation.

3 Méthode proposée

Nous examinons d'abord une méthode d'imputation fractionnaire hot deck appelée **imputation fractionnaire complète**, où les valeurs imputées sont tirées de l'ensemble de répondants désigné par $A_R = \left\{i \in A; \delta_i = 1\right\}$. C'est-à-dire que la j-ième valeur imputée de la donnée manquante y_i, désignée par $y_i^{*(j)}$, est égale à la j-ième valeur de y dans l'ensemble A_R. Nous proposons une méthode d'imputation fractionnaire hot deck qui utilise l'hypothèse du modèle paramétrique $f\left(y \mid \mathbf{x};\theta\right)$. Si tous les éléments de A_R sont choisis comme valeurs imputées de la donnée manquante y_i, nous pouvons traiter $\left\{y_j; j \in A_R\right\}$ comme une réalisation de $f\left(y_j \mid \delta_j = 1\right)$ et, si $h\left(y_j \mid \mathbf{x}_i\right) = f\left(y_j \mid \delta_j = 1\right)$ est choisi en (2.6), le poids fractionnaire assigné au donneur y_j pour la donnée manquante y_i devient

$$\begin{aligned} w_{ij}^* &\propto f\left(y_j \mid \mathbf{x}_i, \delta_i = 0;\hat{\theta}\right)\Big/ f\left(y_j \mid \delta_j = 1\right) \\ &\propto f\left(y_j \mid \mathbf{x}_{i;}\hat{\theta}\right)\Big/ f\left(y_j \mid \delta_j = 1\right), \end{aligned} \tag{3.1}$$

où $\sum_{j;\delta_j=1} w_{ij}^* = 1$ et $\hat{\theta}$ est l'estimateur du maximum de vraisemblance (EMV) obtenu de l'équation (2.4). La deuxième ligne découle de l'hypothèse des données manquant au hasard. Nous pouvons aussi écrire

$$\begin{aligned} f\left(y_j \mid \delta_j = 1\right) &= \int f\left(y_j \mid \mathbf{x}, \delta_j = 1\right) f\left(\mathbf{x} \mid \delta_j = 1\right) d\mathbf{x} \\ &= \int f\left(y_j \mid \mathbf{x}\right) f\left(\mathbf{x} \mid \delta_j = 1\right) d\mathbf{x} \\ &\cong \frac{1}{N_R}\sum_{k=1}^{N} \delta_k f\left(y_j \mid \mathbf{x}_k\right), \end{aligned} \tag{3.2}$$

où la deuxième égalité découle de l'hypothèse des valeurs manquant au hasard, et la dernière égalité (approximative) est obtenue en approximant l'intégrale par distribution empirique de la population. N_R est le nombre de répondants dans la population. En utilisant les poids d'enquête, nous pouvons approximer

$$f\left(y_j \mid \delta_j = 1\right) \cong \frac{\sum_{k \in A_R} w_k f\left(y_j \mid \mathbf{x}_k\right)}{\sum_{k \in A_R} w_k}$$

et les poids fractionnaires en (3.1) sont calculés comme suit :

$$w_{ij}^* \propto \frac{f\left(y_j \mid \mathbf{x}_i; \hat{\theta}\right)}{\sum_{k \in A_R} w_k f\left(y_j \mid \mathbf{x}_k; \hat{\theta}\right)} \tag{3.3}$$

où $\sum_{j \in A_R} w_{ij}^* = 1$. En (3.3), la masse ponctuelle w_{ij}^* assignée au donneur y_j pour l'unité manquante i est exprimée par le ratio de la densité $f\left(y \mid \mathbf{x}\right)$. Ainsi, pour chaque unité manquante i, $n_R = \left|A_R\right|$, nous utilisons les observations comme donneurs pour l'imputation hot deck et w_{ij}^* comme poids fractionnaires. Cette méthode d'imputation fractionnaire peut être qualifiée d'imputation fractionnaire complète (IFC) en l'absence de caractère aléatoire attribuable au mécanisme d'imputation. L'estimateur IFC de η, défini par $\sum_{i=1}^{N} U\left(\eta; \mathbf{x}_i, y_i\right) = 0$ est alors calculé en résolvant

$$\sum_{i \in A} w_i \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1 - \delta_i\right) \sum_{j \in A_R} w_{ij}^* U\left(\eta; \mathbf{x}_i, y_j\right) \right\} = 0, \tag{3.4}$$

où w_{ij}^* est défini en (3.3). Il est à noter que l'équation d'estimation imputée (3.4) est une bonne approximation de l'équation d'estimation prévue en (2.2).

En échantillonnage, un ensemble de données imputées où la quantité d'imputation est importante n'est pas toujours souhaitable. Au lieu d'utiliser toutes les observations en A_R comme donneurs pour chaque donnée manquante, nous pouvons sélectionner un sous-ensemble de A_R afin de réduire la taille de l'ensemble donneur de la donnée manquante y_i. Ainsi, la sélection des donneurs est considérée comme un problème d'échantillonnage et nous utilisons un plan d'échantillonnage et des techniques de pondération efficaces pour obtenir des estimateurs par imputation efficaces. Des plans d'échantillonnage efficaces, comme un échantillonnage stratifié ou un échantillonnage systématique avec probabilité proportionnelle à la taille (PPT), peuvent être utilisés pour sélectionner des donneurs de taille m. Un échantillonnage PPT systématique pour l'imputation fractionnaire hot deck peut être décrit comme suit :

1. Dans chaque i où $\delta_i = 0$, trier les donneurs de l'ensemble complet de répondants $\left\{y_j; \delta_j = 1\right\}$ par ordre croissant où $y_{(1)} \leq \cdots \leq y_{(r)}$ et utiliser $w_{i(j)}^*$ pour désigner le poids fractionnaire associé à $y_{(j)}$, c'est-à-dire $w_{i(j)}^* = w_{ik}^*$ pour $y_{(j)} = y_k$.

2. Partitionner $[0,1]$ par $\left\{I_j \equiv \left[\sum_{k=0}^{j} w_{i(j)}^*, \sum_{k=0}^{j+1} w_{i(j)}^*\right), j = 1, \ldots, r-1\right\}$, où $w_{i(0)}^* = 0$.

3. Générer $u \sim \text{uniforme}(0,1/m)$ et poser $u_k = u + k/m$, $k = 0,\ldots,m-1$. Pour $k = 0,\ldots,m-1$, si $u_k \in I_j$ pour certains $0 \le j \le r-1$, inclure j dans l'échantillon D_i.

Après avoir sélectionné D_i dans l'ensemble complet de répondants, nous assignons les poids fractionnaires initiaux $w_{ij0}^* = 1/m$ aux donneurs choisis en D_i. D'autres ajustements sont apportés aux poids fractionnaires afin de satisfaire

$$\sum_{i\in A} w_i \left\{ (1-\delta_i) \sum_{j\in D_i} w_{ij,c}^* \mathbf{q}(\mathbf{x}_i, y_j) \right\} = \sum_{i\in A} w_i \left\{ (1-\delta_i) \sum_{j\in A_R} w_{ij}^* \mathbf{q}(\mathbf{x}_i, y_j) \right\}, \tag{3.5}$$

pour certains $\mathbf{q}(\mathbf{x}_i, y_j)$, et $\sum_{j\in D_i} w_{ij,c}^* = 1$ pour tous les i où $\delta_i = 0$, w_{ij}^* étant les poids fractionnaires pour la méthode d'IFC définie en (3.3). En ce qui concerne le choix de la fonction de contrôle $\mathbf{q}(\mathbf{x}, y)$ en (3.5), nous pouvons utiliser $\mathbf{q}(\mathbf{x}, y) = (y, y^2)'$, ce qui rapproche le plus possible les distributions empiriques de y pour D_i et A_R en ce sens que les premier et second moments de y sont les mêmes. D'autres choix peuvent être envisagés. Voir Fuller et Kim (2005).

Le problème d'ajustement des poids initiaux afin de respecter certaines contraintes est souvent qualifié de calage et les poids fractionnaires résultants peuvent être qualifiés de poids fractionnaires calés. En utilisant la pondération par régression, nous pouvons calculer des poids fractionnaires finaux de calage qui satisfont à (3.5) et $\sum_j w_{ij,c}^* = 1$ comme suit :

$$w_{ij,c}^* = w_{ij0}^* + w_{ij0}^* \Delta \left(\mathbf{q}_{ij}^* - \overline{\mathbf{q}}_{i\cdot}^* \right), \tag{3.6}$$

où $\mathbf{q}_{ij}^* = \mathbf{q}(\mathbf{x}_i, y_j)$, $\overline{\mathbf{q}}_{i\cdot}^* = \sum_{j\in A_R} w_{ij0}^* \mathbf{q}_{ij}^*$,

$$\Delta = \left\{ C_q - \sum_{i\in A} w_i (1-\delta_i) \sum_{j\in A_R} w_{ij0}^* \mathbf{q}_{ij}^* \right\}^T \left\{ \sum_{i\in A} w_i (1-\delta_i) \sum_{j\in A_R} w_{ij0}^* \left(\mathbf{q}_{ij}^* - \overline{\mathbf{q}}_{i\cdot}^* \right)^{\otimes 2} \right\}^{-1}$$

et $C_q = \sum_{i\in A} w_i \left\{ (1-\delta_i) \sum_{j\in A_R} w_{ij}^* \mathbf{q}(\mathbf{x}_i, y_j) \right\}$. Ici, $B^{\otimes 2}$ désigne BB^T. Certains des poids fractionnaires calculés en (3.6) peuvent prendre des valeurs négatives. Le cas échéant, il faut utiliser des algorithmes remplaçant la pondération par régression. Par exemple, considérons la pondération par l'entropie, où les poids fractionnaires de la forme

$$w_{ij,c}^* = \frac{w_{ij}^* \exp\left(\Delta \mathbf{q}_{ij}^* \right)}{\sum_{k\in A_R} w_{ik}^* \exp\left(\Delta \mathbf{q}_{ik}^* \right)} \tag{3.7}$$

sont à peu près égaux aux poids fractionnaires par régression en (3.6) et sont toujours positifs. Après avoir obtenu les poids fractionnaires de calage, nous pouvons calculer l'estimateur IFHD de η en résolvant

$$\sum_{i\in A} w_i \left\{ \delta_i U(\eta; \mathbf{x}_i, y_i) + (1-\delta_i) \sum_{j\in D_i} w_{ij,c}^* U(\eta; \mathbf{x}_i, y_j) \right\} = 0. \tag{3.8}$$

Une méthode par rééchantillonnage peut être utilisée pour estimer la variance. L'annexe A.1 contient une brève discussion de l'estimateur de variance par rééchantillonnage pour la méthode proposée.

La méthode proposée peut aussi traiter la non-réponse non ignorable sous spécification correcte du modèle de réponse. Voir l'annexe A.3 pour l'extension à un cas de non-réponse non ignorable.

4 Robustesse

Nous examinons maintenant la robustesse de la méthode proposée pour tenir compte d'un léger écart par rapport au modèle paramétrique présumé. La robustesse de l'estimateur proposé protège contre les erreurs de spécification du modèle d'imputation, une légère inclinaison exponentielle du modèle vrai. Pour simplifier la présentation, supposons que le plan d'échantillonnage est un échantillonnage aléatoire simple et que l'échantillon réalisé est un échantillon aléatoire tiré du modèle de superpopulation.

Nous supposons que le modèle vrai $g(y \mid x)$ n'appartient pas à $\{f(y \mid x;\theta);\theta \in \Omega\}$. Nous pouvons quand même spécifier un modèle de travail $f(y \mid x;\theta)$ et calculer l'EMV de θ. Il est bien connu (White 1982) que l'EMV converge vers θ^*, le minimiseur de l'information de Kullback-Leibler

$$K(\theta) = E_g\left[\log\left\{\frac{g(Y \mid x)}{f(Y \mid x;\theta)}\right\}\right]$$

pour $\theta \in \Omega$. Sung et Geyer (2007) ont examiné les propriétés asymptotiques de l'EMV Monte Carlo de θ sous données manquantes.

Pour discuter formellement de la robustesse, supposons que la distribution véritable $g(y \mid x)$ appartient au voisinage

$$\mathcal{N}_\varepsilon = \left\{g; D(g,f) < \frac{1}{2}\varepsilon^2\right\} \tag{4.1}$$

pour un rayon $\varepsilon > 0$, où

$$D(g,f) = \int log\left(\frac{g}{f}\right) g\ dy, \tag{4.2}$$

est la mesure de la distance de Kullback-Leibler. Le voisinage (4.1) peut être décrit de la façon suivante. Posons que $z(x,y,\theta)$ est une fonction de x, y et θ, normalisée pour satisfaire $E_{Y|x}(z) = 0$ et $Var_{Y|x}(z) = 1$, et définissons

$$g(y \mid x) = f(y \mid x;\theta)\exp\{\varepsilon z(x,y,\theta) - \kappa(x,\theta)\}, \tag{4.3}$$

où

$$\kappa = \log\left(E_{Y|x}\left[\exp\{\varepsilon z(x,Y,\theta)\}\right]\right).$$

Pour un petit $\varepsilon > 0$, il peut être démontré que

$$\kappa \cong D(g,f) \cong \frac{1}{2}\varepsilon^2. \tag{4.4}$$

L'équation (4.3) représente un vaste ensemble de distributions proches de $f(y \mid x;\theta)$ créées en variant $z(x,y,\theta)$ sur différentes fonctions normalisées, où z et ε contiennent des interprétations géométriques qui représentent respectivement la direction et la grandeur des erreurs de spécification. Pour le paramètre θ, de dimension p, nous pouvons spécifier les directions des erreurs de spécification comme suit :

$$\left(z_1, z_2, \ldots, z_p\right)^T = I_\theta^{-1/2} s(x,y,\theta),$$

où $s(x,y,\theta) = \partial \log f(y \mid x;\theta) / \partial\theta$ et I_θ est la matrice d'information pour θ. Représentons $z(x,y,\theta)$ comme

$$z(x,y,\theta) = \lambda^T I_\theta^{-1/2} s(x,y,\theta),$$

où $\sum_{i=1}^{p} \lambda_i^2 = 1$. $z(x,y,\theta)$ satisfait alors aux critères de normalisation $E_{Y|x}(z) = 0$ et $Var_{Y|x}(z) = 1$. Voir Copas et Eguchi (2001) pour une discussion plus approfondie de cette expression.

Soit $w_{ij,g}^*$ le poids fractionnaire de la forme (3.3) selon la vraie densité g où $w_{ij,f}^*$ est le poids fractionnaire correspondant selon la « densité de travail » f. La construction spéciale du poids nous permet d'établir

$$w_{ij,g}^* \cong w_{ij,f}^* + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta}\left(w_{ij,f}^*\right). \tag{4.5}$$

La preuve de (4.5) est donnée à l'annexe A.2. Ainsi

$$\begin{aligned} \sum_i w_i \sum_j w_{ij,g}^* U\left(\eta; x_i, y_j\right) \cong & \sum_i w_i \sum_j w_{ij,f}^* U\left(\eta; x_i, y_j\right) \\ & + \varepsilon \lambda^T I_\theta^{-1/2} \sum_i w_i \sum_j \frac{\partial}{\partial\theta}\left(w_{ij,f}^*\right) U\left(\eta; x_i, y_j\right). \end{aligned} \tag{4.6}$$

Pour un petit ε, nous avons

$$\sum_i w_i \sum_j w_{ij,g}^* U\left(\eta; x_i, y_j\right) \cong \sum_i w_i \sum_j w_{ij,f}^* U\left(\eta; x_i, y_j\right),$$

et l'estimateur η résultant de $\sum_i w_i \sum_j w_{ij,f}^* U\left(\eta; x_i, y_j\right) = 0$ approchera donc la valeur réelle η_0.

5 Étude par simulation

Nous avons effectué deux études par simulation. Dans la section 5.1, nous avons comparé la performance de la méthode proposée avec celle d'autres méthodes d'imputation dans un modèle correctement spécifié et un modèle mal spécifié, respectivement, avec données manquantes ignorables. Dans la section 5.2, nous avons comparé la puissance statistique d'un test fondé sur l'IFHD plutôt que sur l'imputation multiple (IM).

5.1 Première simulation

La première étude de simulation testait la performance de la méthode proposée sous l'hypothèse de données manquantes ignorables. Nous avons utilisé deux ensembles de modèles pour générer les observations. Dans le modèle A, $y_i = 0{,}5x_i + e_i$, où $x_i \sim \exp(1)$, $e_i \sim N(0,1)$, x_i et e_i étant indépendants. Dans le modèle B, $y_i = 0{,}5x_i + e_i$, où $x_i \sim \exp(1)$, $e_i \sim \{\chi^2(2) - 2\}/2$, x_i et e_i étant indépendants. Les échantillons aléatoires de taille $n = 200$ ont été générés séparément à partir des deux modèles. Outre (x_i, y_i), nous avons généré δ_i à partir d'une Bernoulli(π_i), où $\pi_i = \{1 + \exp(-0{,}2 - x_i)\}^{-1}$. La variable x_i était toujours observée, mais la variable y_i l'était si et seulement si $\delta_i = 1$. Les taux de réponse globaux étaient d'environ 65% dans les deux cas. Nous avons utilisé $B = 2\ 000$ échantillons Monte Carlo lors de la simulation.

À partir de chacun des échantillons Monte Carlo, dont l'un avait été généré à l'aide du modèle A et l'autre à l'aide du modèle B, nous avons calculé les huit estimateurs suivants :

1. L'estimateur d'échantillon complet qui est calculé sur l'échantillon complet.

2. L'appariement d'après la moyenne prédictive (AMP) est une méthode d'imputation semi-paramétrique, qui attribue une valeur de manière aléatoire à partir des observations les plus proches de la valeur prédite tirée de $f(y \mid \mathbf{x})$. L'AMP a été mis en œuvre au moyen de la fonction « mice.impute.pmm » de R.

3. L'estimateur par imputation multiple (IM) où la taille du groupe d'imputation est $m = 10$, et où les valeurs imputées sont générées à partir du modèle de régression fondé sur la théorie normale, tel que l'envisagent Schenker et Welsh (1988).

4. L'estimateur par imputation fractionnaire paramétrique sans calage (IFP) où la taille du groupe d'imputation est $m = 10$.

5. L'estimateur par imputation fractionnaire paramétrique avec calage (IFP_cal) où la taille du groupe d'imputation est $m = 10$. Les poids fractionnaires sont calculés selon la méthode de calage en (3.6) où $\mathbf{q} = (y, y^2)$.

6. L'estimateur par imputation fractionnaire complète (IFC) utilisant l'ensemble complet de répondants comme valeurs d'imputation, c.-à-d. que la taille du groupe d'imputation est $m = n_R$, où n_R est la taille de A_R.

7. L'estimateur par imputation fractionnaire hot deck sans calage (IFHD) utilisant un petit sous-ensemble de répondants de taille $m = 10$ comme valeurs d'imputation.

8. L'estimateur par imputation fractionnaire hot deck avec calage (IFHD_cal) utilisant un petit sous-ensemble de répondants de taille $m = 10$ comme valeurs d'imputation. Les poids fractionnaires sont calculés selon la méthode de calage en (3.6) où $\mathbf{q} = (y, y^2)$.

L'imputation multiple est une façon de générer des valeurs imputées avec estimation simplifiée de la variance. Cette procédure envisage des méthodes bayésiennes de génération des valeurs imputées, où $m > 1$ valeurs imputées sont générées à partir d'une loi prédictive a posteriori. À partir des valeurs imputées $\mathbf{y}^{*(1)},\ldots,\mathbf{y}^{*(m)}$, l'estimateur par imputation multiple η, désigné par $\hat{\eta}_{IM}$ est

$$\hat{\eta}_{IM} = \frac{1}{m}\sum_{k=1}^{m}\hat{\eta}^{(k)}$$

où $\hat{\eta}^{(k)}$ est l'estimateur de réponse complète appliqué au k-ième ensemble de données imputées. La formule de Rubin peut être utilisée pour estimer la variance par IM,

$$\hat{V}_{IM}\left(\hat{\eta}_{IM}\right) = W_m + \left(1+\frac{1}{m}\right)B_m, \tag{5.1}$$

où $W_m = m^{-1}\sum_{k=1}^{m}\hat{V}^{(k)}$, $B_m = \left(m-1\right)^{-1}\sum_{k=1}^{m}\left(\hat{\eta}^{(k)} - \hat{\eta}_{IM}\right)^2$, et $\hat{V}^{(k)}$ est l'estimateur de la variance de $\hat{\eta}^{(k)}$ sous réponse complète appliqué au k-ième ensemble de données imputées.

Dans les deux modèles, nous avons utilisé la densité normale de moyenne $\beta_0 + \beta_1 x$ et de variance σ^2 comme modèle de travail pour l'imputation. Le modèle de travail est donc le vrai modèle dans le modèle A, mais pas dans le modèle B.

Nous avons considéré trois paramètres : $\theta_1 = E\left(Y\right)$, la moyenne de population de y, $\theta_2 = Pr\left(Y < 1\right)$, la proportion de Y inférieure à un, et θ_3, le quantile 0,5 de Y. Pour estimer θ_2 sous échantillon complet, nous avons utilisé $\hat{\theta}_{2,n} = n^{-1}\sum_{i=1}^{n} I\left(y_i < 1\right)$. Pour estimer θ_3 sous échantillon complet, nous avons utilisé $\hat{\theta}_{3,n} = \hat{F}^{-1}\left(p\right) = \inf\left\{y : \hat{F}\left(y\right) > p\right\}$, où $\hat{F}\left(y\right) = n^{-1}\sum_{i=1}^{n} I\left(y_i < y\right)$ et $p = 0{,}5$.

Les tableaux 5.1 et 5.2 montrent les moyennes Monte Carlo, la variance normalisée et les erreurs quadratiques moyennes normalisées des huit estimateurs sous les modèles A et B, respectivement. La variance normalisée (erreur quadratique moyenne) est le ratio de la variance (erreur quadratique moyenne) et la variance (erreur quadratique moyenne) de l'estimateur d'échantillon complet multiplié par 100, ce qui mesure la variance (erreur quadratique moyenne) accrue attribuable à l'imputation par rapport à l'estimateur d'échantillon complet. En ce qui concerne les moyennes Monte Carlo (4e colonne), les estimateurs par imputation sont tous sans biais pour l'estimation de θ_1, θ_2, et θ_3 sous le modèle A. Sous le modèle B, l'AMP, l'IM, l'IFP et l'IFP_cal pour l'estimation de θ_3 sont beaucoup plus biaisés en valeurs absolues que l'IFC, l'IFHD et l'IFHD_cal lorsque le modèle est mal spécifié dans cette simulation. Pour ce qui est de la variance normalisée et de l'erreur quadratique moyenne normalisée (5e et 6e colonnes), l'IFP est plus efficace que l'IFHD parce qu'en IFP, les valeurs imputées sont générées directement en fonction de la distribution conditionnelle $f\left(y \mid x\right)$ tandis qu'en IFHD, les valeurs imputées peuvent être tirées des répondants dont les poids fractionnaires sont dominants. La taille effective des données imputées est déterminée par les observations imputées qui ont des poids fractionnaires importants, ce qui contribue aussi à la perte d'efficacité. L'IFHD perd en efficacité, mais gagne en robustesse. Enfin, l'IFHD où $m = 10$ a une variance normalisée un peu plus grande que l'IFC pour θ_2, en raison de la variabilité additionnelle due à la procédure d'échantillonnage. En comparant l'IFP à l'IFP_cal et l'IFHD à l'IFHD_cal, l'étape du calage améliore légèrement l'efficacité. L'AMP affiche la plus grande variance quel que soit le scénario.

Tableau 5.1
Moyenne, variance normalisée (VN) et erreur quadratique moyenne normalisée (EQMN) Monte Carlo des estimateurs ponctuels dans le modèle A de la première simulation

Modèle	Paramètre	Méthode	Moyenne	VN	EQMN
A	μ_y	Complet	0,50	100	100
		AMP	0,50	175	175
		IM ($m = 10$)	0,50	135	135
		IFP ($m = 10$)	0,50	130	130
		IFP cal ($m = 10$)	0,50	130	130
		IFC ($m = n_R$)	0,50	130	130
		IFHD ($m = 10$)	0,50	156	156
		IFHD cal ($m = 10$)	0,50	130	130
	$Pr(Y < 1)$	Complet	0,68	100	100
		AMP	0,68	168	167
		IM ($m = 10$)	0,68	112	112
		IFP ($m = 10$)	0,68	110	110
		IFP cal ($m = 10$)	0,68	109	109
		IFC ($m = n_R$)	0,68	130	130
		IFHD ($m = 10$)	0,68	137	136
		IFHD cal ($m = 10$)	0,68	132	132
	Quantile	Complet	0,47	100	100
		AMP	0,47	184	184
		IM ($m = 10$)	0,47	111	111
		IFP ($m = 10$)	0,47	111	111
		IFP cal ($m = 10$)	0,47	111	111
		IFC ($m = n_R$)	0,47	135	135
		IFHD ($m = 10$)	0,47	142	142
		IFHD cal ($m = 10$)	0,47	141	141

Tableau 5.2
Moyenne, variance normalisée (VN) et erreur quadratique moyenne normalisée (EQMN) Monte Carlo des estimateurs ponctuels dans le modèle B de la première simulation

Modèle	Paramètre	Méthode	Moyenne	VN	EQMN
B	μ_y	Complet	0,50	100	100
		AMP	0,50	172	172
		IM ($m = 10$)	0,50	131	131
		IFP ($m = 10$)	0,50	131	131
		IFP cal ($m = 10$)	0,50	128	128
		IFC ($m = n_R$)	0,50	127	127
		IFHD ($m = 10$)	0,50	147	147
		IFHD cal ($m = 10$)	0,50	127	127
	$Pr(Y < 1)$	Complet	0,75	100	100
		AMP	0,75	166	166
		IM ($m = 10$)	0,73	140	170
		IFP ($m = 10$)	0,73	138	168
		IFP cal ($m = 10$)	0,73	137	169
		IFC ($m = n_R$)	0,75	137	137
		IFHD ($m = 10$)	0,75	145	145
		IFHD cal ($m = 10$)	0,75	140	141
	Quantile	Complet	0,26	100	100
		AMP	0,24	191	198
		IM ($m = 10$)	0,31	122	159
		IFP ($m = 10$)	0,31	123	160
		IFP cal ($m = 10$)	0,31	122	159
		IFC ($m = n_R$)	0,26	135	135
		IFHD ($m = 10$)	0,26	144	144
		IFHD cal ($m = 10$)	0,26	139	139

Nous avons examiné l'estimation de la variance par rééchantillonnage pour l'IFC et l'IFHD, particulièrement l'estimation de la variance jackknife avec suppression d'un groupe, qui est décrite à l'annexe A.1. Nous avons aussi envisagé l'estimation de la variance en IM, qui utilise la formule de Rubin (5.1).

Le tableau 5.3 montre les biais relatifs Monte Carlo des estimateurs de variance, qui sont calculés comme $\left[E_{MC}\left\{\hat{V}\right\}-V_{MC}\left\{\hat{\theta}\right\}\right]\Big/V_{MC}\left\{\hat{\theta}\right\}$, où $E_{MC}\left\{\hat{V}\right\}$ est la moyenne Monte Carlo des estimations de la variance $\hat{V}$, et $V_{MC}\left\{\hat{\theta}\right\}$ est la variance Monte Carlo des estimations ponctuelles $\hat{\theta}$. Le biais relatif de l'estimateur de la variance en IFC et IFHD est raisonnablement faible pour tous les paramètres envisagés dans les deux modèles, ce qui suggère que l'estimateur de la variance par rééchantillonnage est valide. Le biais relatif et la statistique t de l'estimateur de la variance en IM sont faibles pour θ_1 mais assez importants pour θ_2 même quand le modèle de travail est vrai (modèle A). La formule de Rubin repose sur la décomposition suivante :

$$V\left(\hat{\theta}_{IM}\right)=V\left(\hat{\theta}_n\right)+V\left(\hat{\theta}_{IM}-\hat{\theta}_n\right), \tag{5.2}$$

où $\hat{\theta}_n$ est l'estimateur d'échantillon complet de η. Essentiellement, le terme W_m en (5.1) estime $V\left(\hat{\theta}_n\right)$ et le terme $\left(1+m^{-1}\right)B_m$ en (5.1) estime $V\left(\hat{\theta}_{IM}-\hat{\theta}_n\right)$. La décomposition (5.2) est vérifiée lorsque $\hat{\theta}_n$ est l'EMV de θ, ce qui est la condition de compatibilité de $\hat{\theta}_n$ (Meng 1994). Dans les cas généraux, nous avons

$$V\left(\hat{\theta}_{IM}\right)=V\left(\hat{\theta}_n\right)+V\left(\hat{\theta}_{IM}-\hat{\theta}_n\right)+2Cov\left(\hat{\theta}_{IM}-\hat{\theta}_n,\hat{\theta}_n\right) \tag{5.3}$$

et l'estimateur de la variance de Rubin peut être biaisé si $Cov\left(\hat{\theta}_{IM}-\hat{\theta}_n,\hat{\theta}_n\right)\neq 0$. La condition de compatibilité est vérifiée pour l'estimation de la moyenne de population; elle ne l'est toutefois pas pour l'estimateur de $Pr\left(Y<1\right)$ par la méthode des moments. Il est à noter que l'estimateur imputé de $\theta_2=Pr\left(Y<1\right)$ peut s'exprimer comme suit :

$$\hat{\theta}_{2,I}=n^{-1}\sum_{i=1}^{n}\left[\delta_i I\left(y_i<1\right)+\left(1-\delta_i\right)E\left\{I\left(y_i<1\right)\,|\,x_i;\hat{\mu},\hat{\sigma}\right\}\right]. \tag{5.4}$$

Ainsi, les estimateurs imputés de θ_2 « empruntent de l'information » en utilisant des données additionnelles associées à $f\left(y\,|\,x\right)$, c'est-à-dire que la normalité de $f\left(y\,|\,x\right)$ est utilisée pour calculer l'espérance conditionnelle en (5.4), ce qui améliore l'efficacité de l'estimateur imputé pour θ_2. Le même phénomène s'applique à θ_3. Au tableau 5.1, l'augmentation de la variance due à l'imputation pour l'IM où $m=10$ est d'environ 35 % pour θ_1 mais de seulement 12 % et 11 % pour θ_2 et θ_3, respectivement, ce qui illustre le phénomène de l'« emprunt d'information » pour l'estimation de θ_2 et θ_3 grâce à l'utilisation de données additionnelles à l'étape de l'imputation. Ainsi, lorsque les conditions de compatibilité ne sont pas remplies, l'estimateur imputé améliore l'efficacité, mais l'estimateur de la variance de Rubin ne reconnaît pas cette amélioration.

Tableau 5.3
Biais relatif Monte Carlo de l'estimateur de la variance par rééchantillonnage dans la première simulation

Modèle	Paramètre	Méthode	B.R. (%)
*A	$V\left(\hat{\theta}_1\right)$	IM ($m=10$)	-2,33
		IFC ($m=n_R$)	-0,80
		IFHD_cal ($m=10$)	-0,80
	$V\left(\hat{\theta}_2\right)$	IM ($m=10$)	8,20
		IFC ($m=n_R$)	-5,01
		IFHD_cal ($m=10$)	-5,12
	$V\left(\hat{\theta}_3\right)$	IM ($m=10$)	19,84
		IFC ($m=n_R$)	4,50
		IFHD_cal ($m=10$)	3,78
*B	$V\left(\hat{\theta}_1\right)$	IM ($m=10$)	2,60
		IFC ($m=n_R$)	-0,56
		IFHD_cal ($m=10$)	-0,56
	$V\left(\hat{\theta}_2\right)$	IM ($m=10$)	-3,33
		IFC ($m=n_R$)	-1,89
		IFHD_cal ($m=10$)	-3,25
	$V\left(\hat{\theta}_3\right)$	IM ($m=10$)	-8,99
		IFC ($m=n_R$)	3,50
		IFHD_cal ($m=10$)	3,80

5.2 Deuxième simulation

La deuxième simulation testait la puissance de la méthode proposée dans un test d'hypothèse utilisant le modèle nul comme modèle d'imputation. Les échantillons de données bivariées $\left(x_i, y_i\right)$ de taille $n=100$ ont été générés à partir de

$$y_i = \beta_0 + \beta_1 x_i + \beta_2\left(x_i^2 - 1\right) + e_i \tag{5.5}$$

où $\left(\beta_0;\beta_1;\beta_2\right)=\left(0;0{,}9;0{,}06\right)$, $x_i \sim N\left(0;1\right)$, $e_i \sim N\left(0;0{,}16\right)$, x_i et e_i étant indépendants. La variable x_i est toujours observée, mais la probabilité que y_i réponde est de $0{,}5$. Les échantillons Monte Carlo ont été générés indépendamment $B=10\ 000$ fois. Nous voulons tester $H_0 : \beta_2 = 0$ pour les répondants. Nous avons comparé l'IFHD à l'IM en utilisant la même taille de groupe d'imputation, soit $m=30$. Le modèle d'imputation est le modèle nul,

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

C'est-à-dire que le modèle d'imputation utilise des informations supplémentaires de $\beta_2 = 0$. À partir des données imputées, nous avons ajusté le modèle (5.5) et calculé la puissance d'un test $H_0 : \beta_2 = 0$ au niveau significatif de $0{,}05$. Nous avons également envisagé la méthode des cas complets (MCC), qui utilise les répondants seulement pour la régression.

Le tableau 5.4 montre la moyenne et la variance Monte Carlo des estimateurs ponctuels, le biais relatif de l'estimateur de la variance et la puissance Monte Carlo des tests $H_0 : \beta_2 = 0$. Dans chacun des échantillons Monte Carlo, nous avons construit un intervalle de confiance à 95% de Wald de β_2 en utilisant la formule $\left(\hat{\beta}_2 - 1{,}96\hat{V}^{1/2}; \hat{\beta}_2 + 1{,}96\hat{V}^{1/2}\right)$ et nous rejetons l'hypothèse nulle si $\beta_2 = 0$ ne tombe pas dans l'intervalle de confiance de Wald. La puissance Monte Carlo est calculée comme étant la fréquence relative du rejet de l'hypothèse nulle dans les échantillons Monte Carlo. Dans la deuxième colonne, les estimateurs IFHD et IM sont biaisés pour β_2, comme prévu, car le modèle d'imputation est le modèle nul et il est légèrement différent du modèle vrai qui a généré l'échantillon. Le biais de l'IFHD est plus petit que celui de l'IM en raison de la robustesse de l'IFHD examinée à la section 4. En IM, 50% des données imputées viennent du modèle nul et les 50% qui restent sont tirées du modèle vrai, de sorte que la pente β_2 est ramenée à zéro par la moitié de la vraie pente. En IFHD, même si nous avons utilisé le modèle nul pour calculer les poids fractionnaires, les données imputées viennent du modèle vrai, ce qui réduit le biais. L'IM fournit des estimateurs ponctuels plus efficaces que la MCC, mais l'estimation de la variance est très prudente (surestimation d'environ 180%). Étant donné le biais positif marqué de l'estimateur de variance IM, la puissance statistique des tests fondés sur l'IM est plus faible que celle des tests fondés sur la MCC. Par ailleurs, l'IFHD fournit des estimateurs ponctuels plus efficaces que la MCC, l'estimation de la variance est essentiellement sans biais, et la puissance statistique des tests fondés sur l'IFHD est plus élevée que celle des tests fondés sur la MCC.

Tableau 5.4
Résultats fondés sur 10 000 échantillons Monte Carlo de la deuxième simulation

Méthode	$E\left(\hat{\beta}_2\right)$	$V\left(\hat{\beta}_2\right)$	B.R.$\left(\hat{V}\right)$	Puissance
IFHD	0,046	0,00146	0,02	0,314
IM	0,028	0,00056	1,81	0,044
MCC	0,060	0,00234	-0,01	0,285

6 Conclusion

Nous avons proposé une méthode d'imputation fractionnaire hot deck qui utilise un modèle paramétrique pour $f(y \mid \mathbf{x})$ quand $\mathbf{x}$ contient des composantes continues. La méthode proposée fournit une estimation robuste pour les paramètres en ce sens que le modèle d'imputation n'est pas nécessairement égal au modèle générateur de données. Le prix que nous payons dans l'IFHD est la perte d'efficacité dans l'estimation ponctuelle. Dans notre première simulation, l'estimateur IFHD pour $P(Y < 1)$ affiche la deuxième variance en importance, mais la plus petite erreur quadratique moyenne lorsque le modèle de travail n'est pas vrai, comparativement à d'autres estimateurs.

La perte d'efficacité tient principalement au fait que les poids fractionnaires sont plus variables que ceux obtenus selon la méthode de l'IFP parce que certains des $\mathbf{x}_j$ n'aident pas à imputer y_i, c'est-à-dire que la valeur de $f\left(y_i \mid \mathbf{x}_j; \hat{\theta}\right)$ peut être très faible. Lorsque le groupe d'imputation est très petit (p. ex.

$m = 10$), l'imputation fractionnaire hot deck ne fait pas augmenter la variance de façon significative, comme nous pouvons le voir au tableau 5.1 sous le modèle A.

En fait, la méthode d'imputation fractionnaire peut être utilisée pour élaborer une méthode d'imputation unique en appliquant l'IFHD où $m = 1$, ce qui sélectionne une valeur imputée ayant une probabilité proportionnelle au poids fractionnaire pour chaque unité manquante. En l'occurrence, l'IFHD peut être utilisée pour élaborer une méthode d'imputation unique qui reste robuste aux erreurs de spécification du modèle. Le calage de pondération est toutefois incompatible avec une imputation unique. Nous pouvons quand même respecter les contraintes de calage en employant la méthode d'imputation équilibrée examinée par Chauvet, Deville et Haziza (2011), ou l'échantillonnage réjectif de Poisson de Fuller (2009). Un examen plus approfondi suivant cette piste fera l'objet d'une prochaine étude.

Remerciements

Nous remercions deux examinateurs anonymes et le rédacteur associé de leurs commentaires très utiles. Les travaux de recherche ont été financés en partie par une subvention du Conseil de recherches en sciences naturelles et en génie (MMS-121339) et par l'entente de coopération conclue entre le *Natural Resources Conservation Service* de l'USDA et le *Center for Survey Statistics and Methodology* de l'*Iowa State University*.

Annexe

A.1 Estimation de la variance par rééchantillonnage

Des méthodes de répliques peuvent être utilisées pour estimer la variance. Soit $w_i^{[k]}$ les k-ième poids de rééchantillonnage de sorte que

$$\hat{V}_{rep} = \sum_{k=1}^{L} c_k \left(\hat{Y}^{[k]} - \hat{Y} \right)^2$$

est convergent pour la variance de $\hat{Y} = \sum_{i \in A} w_i y_i$, où L est la taille des répliques, c_k est le k-ième facteur de réplication qui dépend de la méthode des répliques et du mécanisme d'échantillonnage, et $\hat{Y}^{[k]} = \sum_{i \in A} w_i^{[k]} y_i$ est la k-ième répétition de $\hat{Y}$. En estimation de la variance jackknife avec suppression d'un groupe, $L = n$ et $c_k = (n-1)/n$.

Pour appliquer la méthode des répliques en IFC, nous devons d'abord appliquer les poids de rééchantillonnage $w_i^{[k]}$ en (2.4) afin de calculer $\hat{\theta}^{[k]}$. Après avoir obtenu $\hat{\theta}^{[k]}$, nous utilisons les mêmes valeurs imputées pour calculer les poids de rééchantillonnage fractionnaires initiaux

$$w_{ij}^{*[k]} \propto w_j^{[k]} w_j^{-1} f\left(y_j \mid x_i; \hat{\theta}^{[k]} \right) \Big/ \left\{ \sum_{l \in A_R} w_l^{[k]} f\left(y_j \mid x_l; \hat{\theta}^{[k]} \right) \right\}, \tag{A.1}$$

où $\sum_{j \in A_R} w_{ij}^{*[k]} = 1$. La variance de $\hat{\eta}_{IFC}$, calculé en (3.4), est ensuite calculée comme suit :

$$\hat{V}_{rep} = \sum_{k=1}^{L} c_k \left(\hat{\eta}_{IFC}^{[k]} - \hat{\eta}_{IFC} \right)^2,$$

où $\hat{\eta}_{IFC}^{[k]}$ est obtenu en résolvant

$$\sum_{i \in A} w_i^{[k]} \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j \in A_R} w_{ij}^{*[k]} U\left(\eta; \mathbf{x}_i, y_j\right) \right\} = 0,$$

et $w_{ij}^{*[k]}$ est défini en (A.1).

Examinons maintenant l'estimation de la variance par rééchantillonnage de l'estimateur IFHD $\hat{\eta}_{IFHD}$ calculé en (3.8). Définissons $d_{ij} = 1$ si $j \in D_i$ et $d_{ij} = 0$ autrement. Il est à noter que $\hat{\eta}_{IFHD}$ est calculé en deux étapes. Dans la première étape, nous utilisons un échantillonnage PPT systématique où la probabilité de sélection est proportionnelle aux poids fractionnaires de la méthode IFC. Dans la deuxième étape, nous utilisons la méthode de pondération par calage en appliquant la contrainte (3.5) où $\sum_{j \in A_R} d_{ij} w_{ij,c}^* = 1$. Ainsi, les poids de rééchantillonnage fractionnaires sont eux aussi calculés en deux étapes. Premièrement, le poids de rééchantillonnage fractionnaire initial pour $w_{ij0}^* = 1/m$ est alors donné par

$$w_{ij0}^{*[k]} = \frac{d_{ij}\left(w_{ij}^{*[k]} / w_{ij}^*\right)}{\sum_{l \in A_R} d_{il}\left(w_{il}^{*[k]} / w_{il}^*\right)}, \tag{A.2}$$

où w_{ij}^* est le poids fractionnaire pour l'IFC défini en (2.6) et $w_{ij}^{*[k]}$ est le k-ième poids de rééchantillonnage fractionnaire pour l'IFC défini en (A.1). Deuxièmement, les poids de rééchantillonnage fractionnaires sont ajustés afin de respecter les contraintes de calage. L'équation de calage pour les poids de rééchantillonnage fractionnaires correspondant à (3.5) est alors

$$\sum_{i \in A} w_i^{[k]} \left\{ \left(1-\delta_i\right) \sum_{j \in D_i} w_{ij,c}^{*[k]} \mathbf{q}\left(\mathbf{x}_i, y_j\right) \right\} = \sum_{i \in A} w_i^{[k]} \left\{ \left(1-\delta_i\right) \sum_{j \in A_R} w_{ij}^{*[k]} \mathbf{q}\left(\mathbf{x}_i, y_j\right) \right\} \tag{A.3}$$

et $\sum_{j \in D_i} w_{ij,c}^{*[k]} = 1$. Nous pouvons utiliser la pondération par régression ou la pondération par entropie pour obtenir des poids de rééchantillonnage fractionnaires respectant les contraintes. Après avoir obtenu les poids de rééchantillonnage fractionnaires, nous calculons l'estimation par répliques $\hat{\eta}^{[k]}$ en résolvant

$$\sum_{i \in A} w_i^{[k]} \left\{ \delta_i U\left(\eta; x_i, y_i\right) + \left(1-\delta_i\right) \sum_{j \in A_R} w_{ijc}^{*[k]} U\left(\eta; x_i, y_j\right) \right\} = 0.$$

L'estimateur de la variance par rééchantillonnage de $\hat{\eta}$, calculé en (3.8) est donné par

$$\hat{V}_{rep}\left(\hat{\eta}\right) = \sum_{k=1}^{L} c_k \left(\hat{\eta}^{[k]} - \hat{\eta} \right)^2.$$

Comme $\hat{\eta}$ est une fonction lisse de $\hat{\theta}$, la convergence de $\hat{V}_{rep}(\hat{\eta})$ découle directement de l'argument standard de l'estimation de la variance par rééchantillonnage (Shao et Tu 1995).

A.2 Preuve de la formule (4.5)

Utilisons

$$\frac{g\left(y_j \mid x_i\right)}{g\left(y_j \mid x_k\right)} = \frac{f\left(y_j \mid x_i\right)}{f\left(y_j \mid x_k\right)} \exp\left(\varepsilon \Delta_{ik|j} - \kappa\left(x_i\right) + \kappa\left(x_k\right)\right)$$

où $\Delta_{ik|j} = z\left(x_i, y_j; \theta\right) - z\left(x_k, y_j; \theta\right)$. En nous fondant sur la linéarisation de Taylor et sur (4.4), nous avons

$$\frac{g\left(y_j \mid x_i\right)}{g\left(y_j \mid x_k\right)} \cong \frac{f\left(y_j \mid x_i\right)}{f\left(y_j \mid x_k\right)} \left\{1 + \varepsilon \Delta_{ik|j}\right\}.$$

Si nous connaissons la véritable densité, les poids fractionnaires corrects en (3.3) peuvent être exprimés comme suit :

$$\begin{aligned}
w_{ij,g}^{*} &\propto \frac{g\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k g\left(y_j \mid x_k\right)} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{g(y_j \mid x_k)}{g(y_j \mid x_i)}\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \exp\left(\varepsilon \Delta_{ki|j} - \kappa\left(x_i\right) + \kappa\left(x_k\right)\right)\right\}} \\
&\cong \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \left(1 + \varepsilon \Delta_{ki|j}\right)\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \sum_{k;\delta_k=1} w_k \left[\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \left\{z\left(x_k, y_j\right) - z\left(x_i, y_j\right)\right\}\right]} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \lambda^T I_\theta^{-1/2} \sum_{k;\delta_k=1} w_k \left(\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \left\{s\left(x_k, y_j\right) - s\left(x_i, y_j\right)\right\}\right)} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \lambda^T I_\theta^{-1/2} \sum_{k;\delta_k=1} w_k \frac{\partial}{\partial \theta} \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}}
\end{aligned}$$

$$\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{ \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \right\}} \left[1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\sum_{k;\delta_k=1} w_k \frac{\partial}{\partial \theta} \left\{ \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \right\}}{\sum_{k;\delta_k=1} w_k \left\{ \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \right\}} \right]$$

$$\propto \frac{f\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)} \left[1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\frac{\partial}{\partial \theta} \left\{ \sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \right\}}{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}} \right]$$

$$= \frac{f\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \left\{ \frac{1}{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}} \right\}$$

$$= a_{ij} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{ij},$$

où $a_{ij} = f\left(y_j \mid x_i\right) \Big/ \sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)$ et $a_{i+} = \sum_{j;\delta_j=1} a_{ij}$. Ainsi, $w_{ij,f}^* = a_{ij} / a_{i+}$ et

$$w_{ij,g}^* \cong \frac{a_{ij} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{ij}}{a_{i+} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{i+}}$$

$$= \frac{a_{ij}}{a_{i+}} \left(1 + \frac{\varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{ij}}{a_{ij}} \right) \left(\frac{a_{i+}}{a_{i+} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{i+}} \right)$$

$$\cong \frac{a_{ij}}{a_{i+}} \left(1 + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \log a_{ij} \right) \left(1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \log a_{i+} \right)$$

$$\cong \frac{a_{ij}}{a_{i+}} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{a_{ij}}{a_{i+}} \left(\frac{\partial}{\partial \theta} \log a_{ij} - \frac{\partial}{\partial \theta} \log a_{i+} \right)$$

$$= \frac{a_{ij}}{a_{i+}} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \left(\frac{a_{ij}}{a_{i+}} \right)$$

$$= w_{ij,f}^* + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \left(w_{ij,f}^* \right),$$

ce qui prouve (4.5).

A.3 Extension à un cas de données manquantes non ignorables

Nous considérons une extension de la méthode proposée à un cas de données manquantes non ignorables. Selon l'hypothèse des données manquantes non ignorables, le modèle conditionnel $f\left(y \mid x\right)$ et

le modèle de probabilité de réponse $P(\delta=1\,|\,\mathbf{x},y)$ sont nécessaires pour évaluer la fonction d'estimation prévue en (4.6). Soit le modèle de probabilité de réponse donné par $Pr(\delta_i=1\,|\,\mathbf{x}_i,y_i)=\pi(\mathbf{x}_i,y_i;\phi)$ pour certains ϕ ayant une fonction $\pi(\cdot)$ connue. Nous supposons que les paramètres sont identifiables, comme en ont discuté Wang, Shao et Kim (2013).

En IFP, selon Kim et Kim (2012), l'EMV $(\hat{\theta},\hat{\phi})$ peut être obtenu en résolvant

$$\sum_{i\in A} w_i\left\{\delta_i S(\theta;\mathbf{x}_i,y_i)+(1-\delta_i)\sum_{j=1}^{m} w_{ij}^{*} S\left(\theta;\mathbf{x}_i,y_i^{*(j)}\right)\right\}=0, \tag{A.4}$$

et

$$\sum_{i\in A} w_i\left\{\delta_i S(\phi;\mathbf{x}_i,y_i)+(1-\delta_i)\sum_{j=1}^{m} w_{ij}^{*} S\left(\phi;\mathbf{x}_i,y_i^{*(j)}\right)\right\}=0, \tag{A.5}$$

où $S(\theta;\mathbf{x},y)=\partial\log f(y\,|\,\mathbf{x};\theta)/\partial\theta$, $S(\phi;\mathbf{x},y)=\partial\log\pi(\mathbf{x},y;\phi)/\partial\phi$, et les poids fractionnaires sont donnés par

$$w_{ij}^{*}(\theta,\phi)=\frac{f\left(y_i^{*(j)}\,|\,\mathbf{x}_i;\theta\right)\left\{1-\pi\left(\mathbf{x}_i,y_i^{*(j)},\phi\right)\right\}\Big/h\left(y_i^{*(j)}\,|\,\mathbf{x}_i\right)}{\sum_{k=1}^{m}\left[f\left(y_i^{*(k)}\,|\,\mathbf{x}_i;\theta\right)\left\{1-\pi\left(\mathbf{x}_i,y_i^{*(k)},\phi\right)\right\}\Big/h\left(y_i^{*(k)}\,|\,\mathbf{x}_i\right)\right]}. \tag{A.6}$$

La solution de (A.4) et (A.5) peut être obtenue au moyen de l'algorithme EM. Dans l'algorithme EM, l'étape E calcule les poids fractionnaires en (A.6) en utilisant les valeurs paramétriques actuelles et l'étape M met à jour les valeurs paramétriques $\hat{\theta}^{(t+1)}$ et $\hat{\phi}^{(t+1)}$ en résolvant

$$\sum_{i\in A} w_i\left\{\delta_i S(\theta;\mathbf{x}_i,y_i)+(1-\delta_i)\sum_{j=1}^{m} w_{ij}^{*}\left(\hat{\theta}^{(t)},\hat{\phi}^{(t)}\right) S\left(\theta;\mathbf{x}_i,y_i^{*(j)}\right)\right\}=0,$$

et

$$\sum_{i\in A} w_i\left\{\delta_i S(\phi;\mathbf{x}_i,y_i)+(1-\delta_i)\sum_{j=1}^{m} w_{ij}^{*}\left(\hat{\theta}^{(t)},\hat{\phi}^{(t)}\right) S\left(\phi;\mathbf{x}_i,y_i^{*(j)}\right)\right\}=0.$$

Dans la méthode IFC proposée, les poids fractionnaires sont donnés par

$$\begin{aligned} w_{ij}^{*} &\propto f\left(y_j\,|\,\mathbf{x}_i,\delta_i=0;\theta,\phi\right)\Big/f\left(y_j\,|\,\delta_j=1\right)\\ &\propto f\left(y_j\,|\,\mathbf{x}_i;\theta\right)\left\{1-\pi\left(\mathbf{x}_i,y_j;\phi\right)\right\}\Big/f\left(y_j\,|\,\delta_j=1\right), \end{aligned}$$

où $\sum_{j;\delta_j=1} w_{ij}^{*}=1$. Parce que

$$\begin{aligned} f\left(y_j\,|\,\delta_j=1\right) &= \int \pi\left(\mathbf{x},y_j\right) f\left(y_j\,|\,\mathbf{x}\right) f(\mathbf{x})d\mathbf{x}\\ &\cong \sum_{k\in A} w_k \pi\left(\mathbf{x}_k,y_j\right) f\left(y_j\,|\,\mathbf{x}_k\right). \end{aligned} \tag{A.7}$$

Les poids fractionnaires peuvent être calculés à partir de

$$w_{ij}^{*} \propto \frac{f\left(y_j \mid \mathbf{x}_i;\theta\right)\left\{1-\pi\left(\mathbf{x}_i, y_j;\phi\right)\right\}}{\sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j;\phi\right) f\left(y_j \mid \mathbf{x}_k;\theta\right)}. \tag{A.8}$$

où $\sum_{j\in A_R} w_{ij}^{*} = 1$.

Nous pouvons donc utiliser l'algorithme EM suivant pour obtenir les estimations paramétriques désirées.

(Étape I) Pour chaque unité manquante $i \in A_M = \{i \in A; \delta_i = 0\}$, prendre m valeurs imputées comme $y_i^{(1)}, \ldots, y_i^{(m)}$ à partir de A_R, où $m = r$.

(Étape E) Les poids fractionnaires sont donnés par

$$w_{ij}^{*(t)} \propto \frac{f\left(y_j \mid \mathbf{x}_i, \hat{\theta}^{(t)}\right)\left\{1-\pi\left(\mathbf{x}_i, y_j;\hat{\phi}^{(t)}\right)\right\}}{\sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j;\hat{\phi}^{(t)}\right) f\left(y_j \mid \mathbf{x}_k;\hat{\theta}^{(t)}\right)}$$

et $\sum_{j=1}^{m} w_{ij}^{*(t)} = 1$.

(Étape M) Mettre à jour les paramètres $\hat{\theta}^{(t+1)}$ et $\hat{\phi}^{(t+1)}$ en résolvant les équations de score imputées suivantes :

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\theta;\mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j\in A_R} w_{ij}^{*(t)} S\left(\theta;\mathbf{x}_i, y_j\right) \right\} = 0,$$

et

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\phi;\mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j\in A_R} w_{ij}^{*(t)} S\left(\phi;\mathbf{x}_i, y_j\right) \right\} = 0.$$

Il est à noter que l'étape I n'a pas besoin d'être répétée dans l'algorithme EM. Une fois les estimations paramétriques finales obtenues, les poids fractionnaires sont calculés selon la formule en (A.8) et ils servent de probabilités de sélection pour l'IFHD lorsque le groupe d'imputation est de petite taille m. La méthode d'échantillonnage PPT systématique examinée à la section 3 peut aussi être utilisée pour obtenir l'IFHD.

Bibliographie

Andersen, P.K. et Gill, R.D. (1982). Cox's regression model for counting process: a large sample study. *The Annals of Statistics*, 10, 1100-1120.

Andridge, R.R. et Little, R.J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78, 40-64.

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.

Binder, D. et Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

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

Chen, J. et Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American Statistical Association*, 96, 260-269.

Copas, J.B. et Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society, Series B*, 63, 871-895.

Durrant, G.B. (2009). Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates. *International Journal of Social Research Methodology*, 12, 293-304.

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

Fuller, W.A. et Kim, J.K. (2005). Imputation hot deck pour le modèle de réponse. *Techniques d'enquête*, 31, 153-164.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of Statistics*, 29, 215-246.

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

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119-132.

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

Kim, J.K., Fuller, W.A. et Bell, W.R. (2011). Variance estimation for nearest neighbor imputation for U.S. census long form data. *Annals of Applied Statistics*, 5, 824-842.

Kim, J.K. et Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Chapman and Hall/CRC.

Kim, J.Y. et Kim, J.K. (2012). Parametric fractional imputation for nonignorable missing data. *Journal of the Korean Statistical Society*, 41, 291-303.

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

Pfeffermann, D. (2011). Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ? Comment le résoudre ? *Techniques d'enquête*, 37, 123-146.

Rao, J.N.K., Yung, W. et Hidiroglou, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *The Indian Journal of Statistics, Series A*, 64, 364-378.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schenker, N. et Welsh, A.H. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics*, 16, 1550-1566.

Shao, J. et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer.

Sung, Y.J. et Geyer, C.J. (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics*, 35, 990-1011.

Wang, C.-Y. et Pepe, M.S. (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society, Series B*, 62, 509-24.

Wang, S., Shao J. et Kim, J.K. (2013). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*. In press.

Wei, G.C. et Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699-704.

White, H. (1982). Maximum likelihood estimation of misspecifed models. *Econometrica*, 50, 1-25.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 257-269
Statistique Canada, n° 12-001-X au catalogue

Gains possibles lors de l'utilisation de l'information sur les coûts au niveau de l'unité dans un cadre assisté par modèle

David G. Steel et Robert Graham Clark[1]

Résumé

Quand nous élaborons le plan de sondage d'une enquête, nous essayons de produire un bon plan compte tenu du budget disponible. L'information sur les coûts peut être utilisée pour établir des plans de sondage qui minimisent la variance d'échantillonnage d'un estimateur du total pour un coût fixe. Les progrès dans le domaine des systèmes de gestion d'enquête signifient qu'aujourd'hui, il est parfois possible d'estimer le coût d'inclusion de chaque unité dans l'échantillon. Le présent article décrit l'élaboration d'approches relativement simples pour déterminer si les avantages pouvant découler de l'utilisation de cette information sur les coûts au niveau de l'unité sont susceptibles d'avoir une utilité pratique. Nous montrons que le facteur important est le ratio du coefficient de variation du coût sur le coefficient de variation de l'erreur relative des coefficients de coût estimés.

Mots-clés : Répartition optimale; plan optimal; plan de sondage; variance d'échantillonnage; coûts d'enquête.

1 Introduction

De simples modèles de coût linéaires ont été utilisés pour tenir compte de l'inégalité des coûts par unité dans les plans de sondage. Sous échantillonnage stratifié, il est parfois possible d'estimer un coefficient de coût par unité pour chaque strate. La répartition résultante de l'échantillon entre les strates est proportionnelle à l'inverse de la racine carrée des coûts par strate (Cochran 1977). Dans un plan de sondage à plusieurs degrés, les coûts d'inclusion des unités aux différents degrés de sélection peuvent être utilisés pour décider du nombre d'unités qu'il convient de sélectionner à chaque degré (Hansen, Hurwitz et Madow 1953).

Même si cette théorie est bien établie, l'utilisation de coûts inégaux n'est pas très répandue en pratique (Brewer et Gregoire 2009), peut-être à cause d'un manque d'information sur les coûts, et parce qu'une plus grande attention est accordée à la taille d'échantillon qu'au coût de dénombrement. Groves (1989) soutient que les modèles de coût linéaires sont irréalistes et que la modélisation mathématique des coûts peut faire oublier des décisions plus importantes, comme le choix du mode de collecte, du nombre d'appels de suivi et de la façon dont l'enquête interagit avec d'autres enquêtes que réalise l'organisme. Néanmoins, étant donné les pressions exercées sur les budgets d'enquête, il faut veiller à ce que le plan de sondage final reflète les coûts et la variance de manière rationnelle, sans être obnubilé par une optimalité formelle.

L'usage croissant d'ordinateurs pour la collecte des données permet de recueillir des renseignements sur les coûts plus nombreux et plus utiles pour les unités qui figurent dans les bases de sondage. Dans un programme d'enquêtes-entreprises mené par un institut national de statistique, la plupart des moyennes et

1. David G. Steel, National Institute for Applied Statistics Research Australia, Université de Wollongong, NSW Australie 2522. Courriel : dsteel@uow.edu.au; Robert Graham Clark, National Institute for Applied Statistics Research Australia, Université de Wollongong, NSW Australie 2522. Courriel : rclark@uow.edu.au.

grandes entreprises sont sélectionnées au moins tous les ans ou tous les deux ans pour participer à certaines enquêtes. Cela peut fournir des renseignements sur les coûts pour ces entreprises; par exemple, certaines d'entre elles peuvent avoir nécessité un suivi ou une vérification de grande portée lors d'une enquête antérieure. Des données directes sont moins susceptibles d'être disponibles pour les petites entreprises, mais des jeux de données sur les coûts pourraient être modélisés pour prédire les coûts probables.

Les plans de collecte de données adaptatifs ou dynamiques s'appuient sur les paradonnées (données sur les processus) recueillies durant une opération d'enquête et sur des données auxiliaires (provenant habituellement de sources administratives) pour créer la base de sondage, afin d'orienter les décisions courantes. Ces décisions peuvent porter sur le nombre de rappels, les répondants auprès desquels il faut effectuer un suivi, le ciblage des primes d'incitation, et le choix du mode de collecte pour les appels de suivi (Groves et Heeringa 2006). Dans un exemple discuté par Groves et Heeringa (2006), les intervieweurs ont classé les non-répondants comme ayant une faible ou une forte propension à répondre. La conversion en répondant étant moins coûteuse pour les membres de la seconde catégorie, une fraction d'échantillonnage plus élevée leur a été attribuée dans une deuxième phase de l'enquête. Plus récemment, Schouten, Bethlehem, Beullens, Kleven, Loosveldt, Luiten, Rutar, Shlomo et Skinner (2012, section 6) ont proposé de concevoir le suivi à la deuxième phase d'une enquête de manière à améliorer l'indicateur R de biais de non-réponse (défini dans Schouten, Cobben et Bethlehem 2009, ainsi que dans Schouten Shlomo et Skinner 2011). Peytchev, Riley, Rosen, Murphy et Lindblad (2010) soutiennent que les non-répondants probables devraient faire l'objet d'un protocole différent dès le début d'une enquête.

Donc, il existe en pratique des coûts par unité inégaux pour l'ensemble des unités avant l'échantillonnage, ou pour les non-répondants ciblés pour le suivi. Dans l'un et l'autre cas, la collecte et l'utilisation d'information sur les coûts entraînent une certaine dépense et une plus grande complexité. En outre, trouver un compromis efficace entre le coût et la variance ne représente qu'une partie du problème, car le biais de réponse doit également être pris en considération. Il est donc important de savoir si les avantages éventuels de l'utilisation de cette information en valent la peine, compte tenu surtout du fait que toute donnée sur les coûts est vraisemblablement imparfaite.

Le présent article décrit l'élaboration d'approximations relativement simples des gains d'efficacité découlant de l'utilisation d'information sur les coûts au niveau de l'unité dans un cadre assisté par modèle. La section 2 donne la notation et certaines expressions importantes. La section 3 traite du plan optimal lorsque les paramètres de coût sont connus. La section 4 offre une analyse de l'utilisation des coûts par unité estimés, et la section 5 présente des exemples. La section 6 offre une discussion.

2 Notation et critère objectif

Considérons une population finie, U, contenant N unités, qui consistent en valeurs Y_i pour $i \in U$. Un échantillon $s \in U$ doit être sélectionné en utilisant un plan d'échantillonnage à probabilités inégales avec une probabilité de sélection positive $\pi_i = P[i \in s]$ pour toutes les unités $i \in U$. On suppose qu'un vecteur de variables auxiliaires $\mathbf{x_i}$ est disponible pour l'ensemble de la population, ou pour toutes les unités $i \in s$ dont le total de population, $\mathbf{t_x} = \sum_{i \in U} \mathbf{x_i}$, est connu. Les variables auxiliaires pourraient être,

par exemple, l'industrie, la région et la taille dans une enquête-entreprises, ou l'âge, le sexe et la région dans une enquête-ménages.

Sous l'approche assistée par modèle (voir par exemple Särndal, Swensson et Wretman 1992), la relation entre la variable d'intérêt et les variables auxiliaires est traduite par un modèle, habituellement de la forme qui suit pour les sondages à un degré :

$$\left.\begin{array}{l} E_M\left[Y_i\right] = \beta^{\mathbf{T}}\mathbf{x_i} \\ var_M\left[Y_i\right] = \sigma^2 z_i \\ Y_i \text{ indépendant de } Y_j \text{ pout tout } i \neq j \end{array}\right\} \tag{2.1}$$

où E_M et var_M désignent l'espérance et la variance sous le modèle, β est un vecteur de paramètres de régression inconnus, σ^2 est un paramètre de variance inconnu, et $\mathbf{x_i}$ et z_i sont supposés connus pour tout $i \in U$. Soit E_p et var_p l'espérance et la variance sous échantillonnage probabiliste répété en maintenant fixes toutes les valeurs de population.

Un estimateur assisté par modèle de t_y d'usage très répandu est l'estimateur par la régression généralisée :

$$\hat{t}_y = \sum_{i \in s} \pi_i^{-1}\left(y_i - \hat{\beta}^{\mathbf{T}}\mathbf{x_i}\right) + \hat{\beta}^{\mathbf{T}}\mathbf{t_x} \tag{2.2}$$

où $\hat{\beta}$ peut être une estimation par les moindres carrés pondérés ou non pondérés des coefficients de régression de y_i sur $\mathbf{x_i}$ en utilisant les données d'échantillon. Des estimateurs peuvent aussi être construits pour des extensions non linéaires du modèle (2.1), mais en pratique, on utilise presque toujours le modèle linéaire.

La **variance anticipée** de $\hat{t}_y$ est définie par $E_M var_p\left[\hat{t}_y - t_y\right]$, et est approximée par

$$E_M var_p\left[\hat{t}_y\right] \approx \sigma^2 \sum_{i \in U}\left(\pi_i^{-1} - 1\right) z_i \tag{2.3}$$

pour les grands échantillons (Särndal et coll. 1992, formule 12.2.12, p. 451) sous le modèle (2.1). Les plans et les estimateurs assistés par modèle doivent minimiser $E_M var_p\left[\hat{t}_y\right]$ sous la contrainte d'une absence approximative de biais sous le plan, $E_p\left[\hat{t}_y\right] = t_y$. Même si le modèle est incorrect, l'estimateur (2.2) demeure approximativement sans biais sous le plan, mais sa variance anticipée en grand échantillon ne sera plus la plus faible possible. La variance anticipée a été utilisée pour motiver l'élaboration de plans de sondage assistés par modèle sous échantillonnage à un degré (Särndal et coll. 1992) et à deux degrés (Clark et Steel 2007; Clark 2009). Un avantage de l'utilisation de la variance anticipée tient au fait qu'elle ne dépend que des probabilités de sélection et d'un petit nombre de paramètres du modèle, qui peuvent être estimés approximativement durant la conception de l'échantillon. En revanche, $var_p\left[\hat{t}_y\right]$ dépend habituellement des valeurs de population de y_i et des probabilités conjointes de sélection, qui sont les unes et les autres difficiles à quantifier d'avance.

Le coût de dénombrement d'un échantillon est supposé être $C = \sum_{i \in s} c_i$ où c_i est le coût d'interrogation d'une unité particulière i. On suppose ordinairement que les valeurs de c_i sont connues.

Habituellement, on suppose aussi que c_i est constant pour toutes les unités de la population, ou constant à l'intérieur des strates. Sous la généralisation que c_i pourrait être différent pour chaque unité i, le coût C dépend de l'échantillon s particulier sélectionné. Le coût prévu est $E_p[C] = \sum_{i \in U} \pi_i c_i$. Le but est de minimiser la variance anticipée (2.3) sous une contrainte sur le coût de dénombrement prévu,

$$\sum_{i \in U} \pi_i c_i = C_f. \tag{2.4}$$

Il existe aussi des coûts fixes sur lesquels le plan de sondage n'a pas d'incidence et qui ne doivent pas être inclus ici.

Une notation est nécessaire pour les variances et les covariances de population. Considérons les paires (u_i, v_i), et soit $S_{uv} = N^{-1} \sum_{i \in U} (u_i - \bar{u})(v_i - \bar{v})$ leur covariance de population, et $S_u^2 = N^{-1} \sum_{i \in U} (u_i - \bar{u})^2$, la variance de population de u_i $(i = 1, \ldots, N)$. Soit $\bar{u}$ et $\bar{v}$ les moyennes de population de u_i et v_i. Le coefficient de variation de population de u_i est $C_u = S_u / \bar{u}$. La covariance relative de population de (u_i, v_i) est $C_{u,v} = S_{uv} / \bar{u}\,\bar{v}$. Un résultat utile est que

$$\sum_{i \in U} u_i v_i = N \bar{u}\,\bar{v} \left(1 + C_{u,v}\right). \tag{2.5}$$

3 Plan optimal avec paramètres de coût et de variance connus

3.1 Plan assisté par modèle optimal

Les valeurs de $(\pi_i : i \in U)$ qui minimisent (2.3) sous la contrainte (2.4) sont

$$\pi_i = C_f \frac{z_i^{1/2} c_i^{-1/2}}{\sum_{j \in U} z_j^{1/2} c_j^{1/2}} \propto z_i^{1/2} c_i^{-1/2} \tag{3.1}$$

et la variance anticipée résultante est

$$AV_{opt} = E_M var_p\left[\hat{t}_y\right] = \sigma^2 C_f^{-1} \left(\sum_{i \in U} c_i^{1/2} z_i^{1/2} \right)^2 - \sigma^2 \sum_{i \in U} z_i. \tag{3.2}$$

Cette expression peut être obtenue facilement en utilisant les multiplicateurs de Lagrange ou l'inégalité de Cauchy-Schwarz, et généralise Särndal et coll. (1992, résultat 12.2.1, p. 452) pour tenir compte des coûts inégaux. Une plus grande probabilité de sélection est attribuée aux unités dont la variance est plus élevée ou dont le coût est plus faible. Toutefois, dans (3.1), les racines carrées de z_i et c_i signifient que, dans de nombreuses enquêtes, les probabilités de sélection ne varient pas spectaculairement.

Dans le cas particulier de l'échantillonnage stratifié où $c_i = \bar{c}_h$ et $z_i = \bar{z}_h$ pour les unités i dans la strate h, (3.1) devient la répartition stratifiée optimale habituelle avec $\pi_i \propto \sqrt{\bar{z}_h / \bar{c}_h}$, de sorte que $n_h \propto N_h \sqrt{\bar{z}_h / \bar{c}_h}$.

Nous supposons que le dernier terme de (3.2), qui représente la correction pour population finie, est négligeable. L'application de (2.5) donne :

$$AV_{opt} \approx \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}\left(1+C_{\sqrt{c},\sqrt{z}}\right)^2}{\left(1+C_{\sqrt{c}}^2\right)\left(1+C_{\sqrt{z}}^2\right)} \tag{3.3}$$

où $C_{\sqrt{c}}$ et $C_{\sqrt{z}}$ désignent les coefficients de variation de population de $\sqrt{c_i}$ et $\sqrt{z_i}$, respectivement. Pour que nos résultats puissent être interprétés, nous supposons que les coûts par unité c_i et les variances σz_i ne sont pas reliés, de sorte que $C_{\sqrt{c},\sqrt{z}} = 0$. Cette hypothèse n'est pas toujours satisfaite en pratique, mais toute relation entre c_i et z_i sera propre à un échantillon particulier et pourrait être positive ou négative. Afin de dégager des principes généraux, il est logique d'ignorer ce genre de relation. En pratique, il est souvent raisonnable de supposer également que $C_{\sqrt{c}}$ et $C_{\sqrt{z}}$ sont faibles. Un développement en série de Taylor montre alors que $C_c^2 \approx 4C_{\sqrt{c}}^2$ et $C_z^2 \approx 4C_{\sqrt{z}}^2$. En regroupant ces approximations, (3.3) devient

$$AV_{opt} = \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}}{\left(1+\frac{1}{4}C_c^2\right)\left(1+\frac{1}{4}C_z^2\right)}. \tag{3.4}$$

Voir l'annexe pour les détails de ces calculs.

Ignorer les coûts

Si l'on fait abstraction des coûts, alors (3.1) fait penser que $\pi_i \propto z_i^{1/2}$. Pour faire des comparaisons pour le même coût prévu, C_f,

$$\pi_i = C_f \frac{z_i^{1/2}}{\sum_{j\in U} z_j^{1/2} c_j} \tag{3.5}$$

et la variance anticipée résultante est

$$AV_{nocosts} = \sigma^2 C_f^{-1}\left(\sum_{i\in U} z_i^{1/2}\right)\left(\sum_{i\in U} c_i z_i^{1/2}\right) - \sigma^2 \sum_{i\in U} z_i. \tag{3.6}$$

En effectuant des calculs similaires à ceux utilisés à la section 3.1, nous obtenons

$$AV_{nocosts} \approx \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}}{\left(1+\frac{1}{4}C_z^2\right)}. \tag{3.7}$$

Voir l'annexe pour des renseignements détaillés. La comparaison de (3.7) et de (3.4) montre que tenir compte des coûts dans le plan de sondage donne lieu à la division de la variance anticipée par $\left(1+(1/4)C_c^2\right)$.

4 Effet de l'utilisation de paramètres de coût estimés

En pratique, les coûts c_i ne sont pas connus précisément. Supposons qu'ils sont remplacés par les estimations $\hat{c}_i = b_i c_i$. L'utilisation de la variable auxiliaire et des coûts estimés dans les probabilités optimales implique que $\pi_i \propto z_i^{1/2} \hat{c}_i^{-1/2}$. Pour faire des comparaisons pour les mêmes coûts prévus,

$$\pi_i = C_f \frac{z_i^{1/2} \hat{c}_i^{-1/2}}{\sum_{j \in U} z_j^{1/2} \hat{c}_j^{-1/2} c_j}.$$

La variance anticipée résultante est

$$AV_{ests} = \sigma^2 C_f^{-1} \left(\sum_{i \in U} \hat{c}_i^{1/2} z_i^{1/2} \right) \left(\sum_{j \in U} z_j^{1/2} \hat{c}_j^{-1/2} c_j \right) - \sigma^2 \sum_{i \in U} z_i. \tag{4.1}$$

Si nous supposons que les valeurs de b_i ne sont pas reliées aux valeurs de c_i et z_i, alors

$$AV_{ests} = \sigma^2 C_f^{-1} \left(\sum_{i \in U} c_i^{1/2} z_i^{1/2} \right)^2 N^{-2} \left(\sum_{i \in U} b_i^{-1/2} \right) \left(\sum_{i \in U} b_i^{1/2} \right) - \sigma^2 \sum_{i \in U} z_i. \tag{4.2}$$

Voir l'annexe pour des renseignements détaillés. Si le coefficient de variation de b_i est faible, l'approximation par développement en série de Taylor donne $N^{-2} \sum b_i^{-1/2} \sum b_i^{1/2} \approx 1 + (1/4) C_b^2$. En appliquant cela, ainsi que les mêmes approximations qu'à la sous-section 3.1, (4.2) devient

$$AV_{ests} = \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z} \left(1 + \frac{1}{4} C_b^2 \right)}{\left(1 + \frac{1}{4} C_c^2 \right) \left(1 + \frac{1}{4} C_z^2 \right)}. \tag{4.3}$$

Voir l'annexe pour des renseignements détaillés.

La comparaison de (4.3) et (3.7) montre qu'utiliser des paramètres de coûts estimés au lieu de faire totalement abstraction des coûts a pour effet de multiplier la variance anticipée par $\left[1 + (1/4) C_b^2 \right] / \left[1 + (1/4) C_c^2 \right]$. Par conséquent, l'utilisation de l'information sur les coûts est valable à condition que $C_b < C_c$. Le coefficient de variation des facteurs d'erreur doit être plus faible que celui des coûts par unité réels sur l'ensemble de la population.

5 Exemples de modèles de coût

Les principales quantités qui déterminent l'utilité des données sur les coûts par unité sont C_b et C_c. Les plans optimaux établis en utilisant l'information sur l'inégalité des coûts n'étant pas très fréquents, la littérature sur les valeurs types de ces mesures est peu abondante. Divers facteurs peuvent donner lieu à des coûts inégaux, y compris les effets de mode de collecte, la géographie et la propension à répondre, et les publications traitant de ces questions donnent une vague idée des modèles de coût qui peuvent être appliqués en pratique.

L'utilisation d'un mode mixte d'interview est l'une des raisons pour lesquelles les coûts par unité peuvent être inégaux. Différents modes de collecte, comme l'interview sur place ou par téléphone assistée par ordinateur, l'envoi de questionnaires par la poste ou l'utilisation de questionnaires en ligne, ou l'interview en face à face, peuvent être utilisés pour obtenir les réponses auprès de différents répondants (Dillman, Smyth et Christian 2009). Cette approche peut être adoptée pour réduire les coûts ou pour améliorer le taux de réponse, mais il faut veiller à ce qu'elle n'introduise pas de biais dû aux effets de mode. Les effets de mode peuvent comprendre des effets de sélection (qui ne posent généralement pas problème) et des effets de mesure (qui entraînement habituellement un biais), et ces deux types d'effets sont souvent difficiles à isoler l'un de l'autre (Vannieuwenhuyze, Loosveldt et Molenbergs 2012). Les économies résultant de l'utilisation de modes mixtes pourraient éventuellement être amplifiées en incorporant les coûts selon le mode de collecte dans le plan de sondage, comme il est décrit dans le présent article. Groves (1989, p. 538) compare les coûts par répondant de l'interview téléphonique (38,00 $) et de l'interview sur place (84,90 $) de la population générale. Si l'on connaissait la préférence de toutes les unités figurant dans une base de sondage et que chaque mode était préféré par une moitié d'entre elle, cela impliquerait que $C_c = 0,38$. Greenlaw et Brown-Welty (2009) ont comparé l'utilisation de questionnaires papier et de questionnaires en ligne, et ont constaté des coûts par répondant de 4,78 $ et de 0,64 $, respectivement, dans le cadre d'une enquête auprès des membres d'une association professionnelle. Dans le cas d'une option à mode mixte, les deux tiers des répondants ont opté pour l'option de réponse en ligne. Si l'on connaît d'avance les préférences, alors $C_c = 0,76$.

Une autre raison de l'existence de coûts variables est que certains répondants sont plus difficiles à recruter que d'autres, et nécessitent un plus grand nombre de visites ou de rappels. Groves et Heeringa (2006, section 2.2) ont réalisé un essai d'enquête où les intervieweurs classaient les non-répondants au moment du premier contact comme étant susceptibles ou non de répondre. Lors du suivi, le taux de réponse a été de 73,7 % pour le premier groupe comparativement à 38,5 % pour le deuxième. Cela donne à penser que le coût par répondant serait au moins 1,9 fois plus élevé pour le deuxième groupe que pour le premier. (En fait, le ratio serait plus élevé, parce qu'un plus grand nombre de tentatives de suivi serait faites pour le groupe difficile.). Si les deux groupes contiennent chacun 50 % des répondants, alors $C_c = 0,31$.

La géographie est une autre source de différences de coût dans les enquêtes avec intervieweur. Dans le cas de l'Enquête sur la population active de l'Australie, les coûts ont été modélisés en utilisant une composante par îlot et une composante par logement (Hicks 2001, tableau 4.2.1 à la section 4.2) selon le type de région (15 types ont été définis). En supposant un échantillonnage constant de 10 logements par îlot, le coût par logement net varie de 4,98 $ dans le centre-ville de Sydney et de Melbourne à 6,71 $ dans les régions peu peuplées et autochtones. Bien qu'il s'agisse d'une différence de coût significative entre les types de régions, la grande majorité de la population se retrouve dans trois de ces types (zone habitée, croissance extérieure et grande ville) où les coûts par logement ne varient qu'entre 5,71 $ et 6,07 $. Par contre l'estimation de C_c est très faible, soit 0,054.

Le tableau 5.1 montre l'amélioration en pourcentage approximative de la variance anticipée lorsqu'on utilise l'information estimée sur les coûts pour différentes valeurs de C_c et C_b, certaines d'entre elles

étant suggérées par les exemples susmentionnés. Les valeurs négatives indiquent que le plan de sondage est moins efficace que si l'on fait complètement abstraction des coûts. Le tableau donne à penser que l'utilisation de l'information sur les coûts ne vaut la peine qu'en cas de variation respectable des coûts par unité; sinon, l'avantage est très faible et peut disparaître en cas d'une imprécision, même faible, des coûts estimés. Les enquêtes à mode de collecte mixte sont celles offrant le plus de possibilités d'exploitation des coûts par unité variables dans le plan de sondage, mais le risque d'un biais de mesure doit être évalué méticuleusement, en utilisant des méthodes telles que celles décrites dans Vannieuwenhuyze, Loosveldt et Molenberghs (2010), Vannieuwenhuyze et coll. (2012), Vannieuwenhuyze et Loosveldt (2013) et Schouten, Brakel, Buelens, Laan et Klaus (2013). Il serait peut-être même possible d'incorporer les effets de mode (ou l'incertitude au sujet des effets de mode) dans le plan optimal au moyen du modèle de variance, une approche qui pourrait être le sujet de futures études. Les constatations faites dans la présente étude font penser qu'une telle approche mérite d'être envisagée.

Tableau 5.1
Amélioration en pourcentage de la variance anticipée découlant de l'utilisation de l'information estimée sur les coûts comparativement à l'absence d'information sur les coûts.

Coefficient de variation des coûts par unité (C_c) (%)	Scénario possible	Coefficient de variation du facteur d'erreur (C_b) (%)			
		0	10	25	50
5		0,1	-0,2	-1,5	-6,2
10	Déplacement de l'intervieweur dû à l'éloignement	0,2	0,0	-1,3	-6,0
20		1,0	0,7	-0,6	-5,2
30	Propension à répondre	2,2	2,0	0,7	-3,9
40	Mode mixte (interview par téléphone/sur place)	3,8	3,6	2,3	-2,2
50		5,9	5,6	4,4	0,0
75	Mode mixte (papier/en ligne autoadministré)	12,3	12,1	11,0	6,8

6 Discussion

L'utilisation de coûts par unité inégaux peut améliorer l'efficacité des plans de sondage. Pour que les gains d'efficacité soit appréciables, les coûts par unité doivent varier considérablement. Même en l'absence d'erreur d'estimation, un coefficient de variation de 50 % peut n'entraîner qu'une amélioration de 6 % de la variance anticipée. Si ce coefficient de variation est de 75 %, comme cela peut se produire dans une enquête à mode de collecte mixte, la réduction de la variance anticipée (ou de la taille de l'échantillon pour une précision fixe) peut être supérieure à 12 %. Les coûts sont estimés avec une certaine erreur, ce qui réduit l'amélioration d'un facteur déterminé par la variation relative des erreurs relatives d'estimation des coûts au niveau individuel.

Annexe

A.1 Calculs détaillés

Lemme 1 : Soit u_i défini pour $i \in U$. Soit $u_i = \overline{u} + \theta e_i$, où $\sum_{i \in U} e_i = 0$ et θ est petit. Alors :

a. $\overline{\sqrt{u}} = \sqrt{\overline{u}} - \frac{1}{8}\theta^2 \overline{u}^{-3/2} S_e^2 + o\left(\theta^2\right)$.

b. $S_{\sqrt{u}}^2 = \frac{1}{4}\theta^2 \overline{u}^{-1} S_e^2 + o\left(\theta^2\right) = \frac{1}{4}\overline{u}^{-1} S_u^2 + o\left(\theta^2\right)$.

c. $N^{-2}\left(\sum_{i \in U} u_i^{1/2}\right)\left(\sum_{i \in U} u_i^{-1/2}\right) = 1 + \frac{1}{4}\theta^2 \overline{u}^{-2} S_e^2 + o\left(\theta^2\right) = 1 + \frac{1}{4} C_u^2 + o\left(\theta^2\right)$.

d. $C_{\sqrt{u}}^2 = \frac{1}{4}\theta^2 \overline{u}^{-2} S_e^2 + o\left(\theta^2\right) = \frac{1}{4} C_u^2 + o\left(\theta^2\right)$.

La notation $o\left(C_u^2\right)$ peut être utilisée à la place de $o\left(\theta^2\right)$, puisque $C_u^2 = \theta^2 C_e^2$, ce qui est fait dans la suite de l'annexe.

Preuve :

Nous commençons par écrire $\overline{\sqrt{u}}$ comme une fonction de θ :

$$\overline{\sqrt{u}} = N^{-1} \sum_{i \in U} \sqrt{u_i} = N^{-1} \sum_{i \in U} \sqrt{\overline{u} + \theta e_i}.$$

Si nous appelons cette fonction $g\left(\theta\right)$, la différenciation au point $\theta = 0$ donne $g(0) = \sqrt{\overline{u}}$, $g'(0) = 0$ et

$$g''(0) = -\frac{1}{4} N^{-1} \overline{u}^{-3/2} \sum_{i \in U} e_i^2 = -\frac{1}{4}\overline{u}^{-3/2} S_e^2.$$

D'où

$$\overline{\sqrt{u}} = g\left(\theta\right) = g(0) + g'(0)\theta + \frac{1}{2} g''(0)\theta^2 + o\left(\theta^2\right) = \sqrt{\overline{u}} - \frac{1}{8}\theta^2 \overline{u}^{-3/2} S_e^2 + o\left(\theta^2\right)$$

qui est le résultat a.

Le résultat b est prouvé en utilisant le résultat a :

$$\begin{aligned}
S_{\sqrt{u}}^2 &= N^{-1} \sum_{i \in U} \left(\sqrt{u_i}\right)^2 - \left(N^{-1} \sum_{i \in U} \sqrt{u_i}\right)^2 \\
&= \overline{u} - \left(\overline{\sqrt{u}}\right)^2 \\
&= \overline{u} - \left(\sqrt{\overline{u}} - \frac{1}{8}\theta^2 \overline{u}^{-3/2} S_e^2 + o\left(\theta^2\right)\right)^2 \\
&= \overline{u} - \left(\overline{u} + \frac{1}{64}\theta^4 \overline{u}^{-3} S_e^4 - \frac{1}{4}\theta^2 \overline{u}^{-1} S_e^2 + o\left(\theta^2\right)\right) \\
&= \frac{1}{4}\theta^2 \overline{u}^{-1} S_e^2 + o\left(\theta^2\right) = \frac{1}{4}\overline{u}^{-1} S_u^2 + o\left(\theta^2\right).
\end{aligned}$$

Pour obtenir c, nous commençons par écrire $N^{-1}\sum_{i\in U} u_i^{-1/2}$ comme une fonction $g()$ de θ et effectuons un développement en série de Taylor :

$$\begin{aligned} N^{-1}\sum_{i\in U} u_i^{-1/2} &= N^{-1}\sum_{i\in U}\left(\bar{u}+\theta e_i\right)^{-1/2} \\ &= g\left(\theta\right) = g(0)+g'(0)\theta+\frac{1}{2}g''(0)\theta^2+o\left(\theta^2\right) \\ &= \bar{u}^{-1/2}+0\theta+\frac{1}{2}\frac{3}{4}\bar{u}^{-5/2}N^{-1}\sum_{i\in U}e_i^2\theta^2+o\left(\theta^2\right) \\ &= \bar{u}^{-1/2}+\frac{3}{8}\bar{u}^{-5/2}S_e^2\theta^2+o\left(\theta^2\right). \end{aligned} \tag{A.1}$$

Notons que $N^{-1}\sum_{i\in U} u_i^{1/2} = \overline{\sqrt{u}}$. La multiplication de l'expression pour $\overline{\sqrt{u}}$ dans le résultat a par (A.1) donne

$$\begin{aligned} N^{-2}\left(\sum_{i\in U} u_i^{1/2}\right)\left(\sum_{i\in U} u_i^{-1/2}\right) &= \left\{\sqrt{\bar{u}}-\frac{1}{8}\theta^2\bar{u}^{-3/2}S_e^2+o\left(\theta^2\right)\right\}\left\{\bar{u}^{-1/2}+\frac{3}{8}\bar{u}^{-5/2}S_e^2\theta^2+o\left(\theta^2\right)\right\} \\ &= 1+\frac{1}{4}\bar{u}^{-2}S_e^2\theta^2+o\left(\theta^2\right) \\ &= 1+\frac{1}{4}C_u^2+o\left(\theta^2\right) \end{aligned}$$

qui est le résultat c.

Pour le résultat d, commençons par noter que $\overline{\sqrt{u}} = \sqrt{\bar{u}}+o\left(\theta\right)$ d'après le résultat a, et donc, un développement en série de Taylor d'ordre 1 donne

$$\left(\overline{\sqrt{u}}\right)^{-2} = \left(\sqrt{\bar{u}}\right)^{-2}+o\left(\theta\right) = \bar{u}^{-1}+o\left(\theta\right).$$

En combinant cela avec le résultat b, nous obtenons

$$\begin{aligned} C_{\sqrt{u}}^2 &= S_{\sqrt{u}}^2\left(\overline{\sqrt{u}}\right)^{-2} \\ &= \left\{\frac{1}{4}\theta^2\bar{u}^{-1}S_e^2+o\left(\theta^2\right)\right\}\left\{\bar{u}^{-1}+o\left(\theta\right)\right\} \\ &= \frac{1}{4}\theta^2\bar{u}^{-2}S_e^2+o\left(\theta^2\right) \\ &= \frac{1}{4}C_u^2+o\left(\theta^2\right) \end{aligned}$$

qui donne le résultat d.

Obtention de (3.3)

Pour le cas particulier où $u_i = v_i$, (2.5) devient

$$\sum_{i\in U} u_i^2 = N\overline{u}^2\left(1+C_u^2\right). \tag{A.2}$$

En appliquant (2.5), nous obtenons

$$\sum_{i\in U} c_i^{1/2} z_i^{1/2} = N\overline{\sqrt{c}}\ \overline{\sqrt{z}}\left(1+C_{\sqrt{c},\sqrt{z}}\right) \tag{A.3}$$

où $\overline{\sqrt{c}} = N^{-1}\sum_{i\in U}\sqrt{c_i}$ et $\overline{\sqrt{z}} = N^{-1}\sum_{i\in U}\sqrt{z_i}$. En utilisant (A.2), nous pouvons exprimer $\overline{\sqrt{c}}$ en fonction de $\overline{c}$:

$$\overline{c} = N^{-1}\sum_{i\in U} c_i = N^{-1}\sum_{i\in U}\left(\sqrt{c_i}\right)^2 = \left(\overline{\sqrt{c}}\right)^2\left(1+C_{\sqrt{c}}^2\right). \tag{A.4}$$

De même,

$$\overline{z} = \left(\overline{\sqrt{z}}\right)^2\left(1+C_{\sqrt{z}}^2\right). \tag{A.5}$$

En supposant que le dernier terme de (3.2) est négligeable, l'application de (A.3), (A.4) et (A.5) donne (3.3).

Obtention de (3.4)

Le lemme 1d implique que $C_{\sqrt{c}}^2 = (1/4)C_c^2 + o\left(C_c^2\right) \approx (1/4)C_c^2$ et $C_{\sqrt{z}}^2 = (1/4)C_z^2 + o\left(C_z^2\right) \approx (1/4)C_z^2$. Le résultat (3.4) découle de (3.3) en utilisant ces approximations, et en supposant que $C_{\sqrt{c},\sqrt{z}} = 0$.

Obtention de (3.7)

Premièrement, $\sum_{i\in U} c_i z_i^{1/2} = N\overline{c}\,\overline{\sqrt{z}}\left(1+C_{c,\sqrt{z}}\right)$, d'après (2.5), où $C_{c,\sqrt{z}}$ est la covariance relative dans la population entre les valeurs de $z_i^{1/2}$ et c_i. Nous supposons que les valeurs de c_i et z_i ne sont pas liées, de sorte que $C_{c,\sqrt{z}} = 0$. Nous supposons aussi que le deuxième terme de (3.6) est négligeable, ce qui correspond à une faible fraction d'échantillonnage. Donc, (3.6) devient :

$$AV_{nocosts} = \sigma^2 N^2 C_f^{-1}\overline{c}\left(\overline{\sqrt{z}}\right)^2. \tag{A.6}$$

Partant de (A.5) et du lemme 1d, nous obtenons

$$\left(\overline{\sqrt{z}}\right)^2 = \frac{\overline{z}}{1+C_{\sqrt{z}}^2} \approx \frac{\overline{z}}{1+(1/4)C_z^2}.$$

La substitution dans (A.6) donne (3.7).

Obtention de (4.2)

Deux termes dans (4.1) se simplifient en utilisant (2.5). Premièrement,

$$\begin{aligned}\sum_{i\in U}\hat{c}_i^{1/2}z_i^{1/2} &= \sum_{i\in U}b_i^{1/2}c_i^{1/2}z_i^{1/2}\\ &= N\left(N^{-1}\sum_{i\in U}b_i^{1/2}\right)\left(N^{-1}\sum_{i\in U}c_i^{1/2}z_i^{1/2}\right)+C_{\sqrt{b},\sqrt{cz}}\end{aligned} \tag{A.7}$$

où $C_{\sqrt{b},\sqrt{cz}}$ est la covariance entre les valeurs de population de $b_i^{1/2}$ et $c_i^{1/2}z_i^{1/2}$. Deuxièmement,

$$\begin{aligned}\sum_{i\in U}z_i^{1/2}\hat{c}_i^{-1/2}c_i &= \sum_{i\in U}b_i^{-1/2}c_i^{1/2}z_i^{1/2}\\ &= N\left(N^{-1}\sum_{i\in U}b_i^{-1/2}\right)\left(N^{-1}\sum_{i\in U}c_i^{1/2}z_i^{1/2}\right)+C_{1/\sqrt{b},\sqrt{cz}}\end{aligned} \tag{A.8}$$

où $C_{1/\sqrt{b},\sqrt{cz}}$ est la covariance entre les valeurs de population de $b_i^{-1/2}$ et $c_i^{1/2}z_i^{1/2}$.

Si nous supposons que les valeurs de population de b_i ne sont pas reliées aux valeurs de c_i et z_i, de sorte que $C_{\sqrt{b},\sqrt{cz}} = C_{1/\sqrt{b},\sqrt{cz}} = 0$, et que nous introduisons (A.7) et (A.8) par substitution dans (4.1), alors nous obtenons (4.2).

Obtention de (4.3)

Nous pouvons exprimer (4.2) en fonction de AV_{opt} qui est défini dans (3.2), en supposant que le dernier terme de (3.2) est négligeable, ce qui correspond à une faible fraction d'échantillonnage :

$$AV_{ests} \approx AV_{opt}N^{-2}\sum_{i\in U}b_i^{-1/2}\sum_{i\in U}b_i^{1/2} \tag{A.9}$$

Le lemme 1c implique que

$$N^{-2}\sum_{i\in U}b_i^{-1/2}\sum b_i^{1/2} = 1+\frac{1}{4}C_b^2+o\left(C_b^2\right)\approx 1+\frac{1}{4}C_b^2.$$

L'introduction de cette expression et de (3.3) par substitution dans (A.9) donne (4.3).

Bibliographie

Brewer, K. et Gregoire, T.G. (2009). Introduction to survey sampling. Dans *Handbook of Statistics 29A: Sample Surveys: Design, Methods and Applications*, eds. Pfeffermann, D. and Rao, C.R., Amsterdam: Elsevier/North-Holland, pp. 9-37.

Clark, R.G. (2009). Sampling of subpopulations in two-stage surveys. *Statistics in Medicine*, 28, 3697-3717.

Clark, R.G. et Steel, D.G. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 63-82.

Cochran, W. (1977). *Sampling Techniques*. New York: Wiley, 3rd ed.

Dillman, D., Smyth, J. et Christian, L. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* John Wiley and Sons, 3rd ed.

Greenlaw, C. et Brown-Welty, S. (2009). A comparison of web-based and paper-based survey methods testing assumptions of survey mode and response cost. *Evaluation Review*, 33, 464-480.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

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.

Hansen, M., Hurwitz, W. et Madow, W. (1953). *Sample Survey Methods and Theory Volume 1: Methods and Applications*. New York: John Wiley and Sons.

Hicks, K. (2001). Cost and variance modelling for the 2001 redesign of the Monthly Population Survey. www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.037, *Australian Bureau of Statistics Methodology Advisory Committee Paper*.

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

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. et Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, 80, 382-399.

Schouten, B., Brakel, J.v.d., Buelens, B., Laan, J.v.d. et Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35(1), 107-121.

Schouten, B., Shlomo, N. et Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 231-253.

Vannieuwenhuyze, J.T. et Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: Three methods to disentangle selection and measurement effects. *Sociological Methods and Research*, 42, 82-104.

Vannieuwenhuyze, J.T., Loosveldt, G. et Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74, 1027-1045.

Vannieuwenhuyze, J. T., Loosveldt, G., et Molenberghs, G. (2012). A method to evaluate mode effects on the mean and variance of a continuous variable in mixed-mode surveys. *International Statistical Review*, 80, 306-322.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 271-292
Statistique Canada, n° 12-001-X au catalogue

Solutions optimales dans les problèmes de sélection contrôlée avec stratification à deux dimensions

Sun Woong Kim, Steven G. Heeringa et Peter W. Solenberger[1]

Résumé

Lorsqu'on envisage la stratification d'un échantillon en fonction de plusieurs variables, on se trouve souvent dans la situation où le nombre prévu d'unités de l'échantillon qui doivent être sélectionnées dans chaque strate est très petit et où le nombre total d'unités à sélectionner est plus petit que le nombre total de strates. Ces plans de sondage stratifiés sont représentés spécifiquement par des tableaux contenant des nombres réels, appelés problèmes de sélection contrôlée, et ne peuvent pas être résolus par les méthodes classiques de répartition. Depuis une soixantaine d'années, de nombreux algorithmes ont été examinés pour résoudre ces problèmes, à commencer par celui de Goodman et Kish (1950). Ceux qui ont été élaborés plus récemment sont particulièrement exigeants du point de vue informatique et trouvent toujours les solutions. Cependant, la question qui demeure sans réponse est celle de savoir dans quel sens les solutions d'un problème de sélection contrôlée obtenues au moyen de ces algorithmes sont optimales. Nous introduisons le concept général des solutions optimales, et nous proposons un nouvel algorithme de sélection contrôlée fondé sur des fonctions de distance type pour obtenir ces solutions. Cet algorithme peut être exécuté facilement par un nouveau logiciel basé sur SAS. La présente étude porte sur les plans de sondage avec stratification à deux dimensions. Les solutions de sélection contrôlée issues du nouvel algorithme sont comparées à celles obtenues au moyen des algorithmes existants, en se fondant sur plusieurs exemples. Le nouvel algorithme arrive à fournir des solutions robustes aux problèmes de sélection contrôlée à deux dimensions qui satisfont aux critères d'optimalité.

Mots-clés : Espérance au niveau de la cellule; échantillonnage probabiliste; fonction de distance; tableau optimal; problème de programmation linéaire; méthode du simplexe.

1 Introduction

Dans l'expression « sélection contrôlée (ou échantillonnage contrôlé) », le terme « contrôle » a un sens large. Dans leur article novateur, Goodman et Kish (1950, p. 351) définissaient la sélection contrôlée comme étant tout processus de sélection dans lequel, tout en maintenant la probabilité assignée à chaque unité, les probabilités de sélection de certaines combinaisons privilégiées de n sur N unités, ou de toutes les combinaisons privilégiées, sont plus grandes que dans l'échantillonnage aléatoire stratifié.

Le présent article porte sur les **contrôles** requis pour décider du nombre d'unités (p. ex. unités primaires d'échantillonnage, ou UPE) affectées à chaque cellule de strate dans un **plan de stratification à deux dimensions**, quand le nombre total d'unités à sélectionner est plus petit que le nombre de cellules de strate ou que le nombre prévu d'unités à sélectionner à partir de chaque cellule de strate est très petit. Cela suppose que, sachant les contraintes de précision et de coût, simplement réduire le nombre de cellules de strate ou augmenter le nombre d'unités échantillonnées n'est pas une solution appropriée pour le plan.

Ici, la **sélection contrôlée** s'entend de la procédure à deux degrés qui suit. En premier lieu, le **problème de sélection contrôlée** représenté par le tableau de nombres réels déterminé par le plan de stratification à deux dimensions est résolu au moyen d'un algorithme spécifié (ou d'une technique spécifiée). La solution du

1. Sun Woong Kim, Director, Survey & Health Policy Research Center, Professor, Department of Statistics, Dongguk University, 26, 3-Ga, Pil-Dong, Jung-Gu Seoul, South Korea 100-715. Courriel : sunwk@dongguk.edu; Steven G. Heeringa, Senior Research Scientist, Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106. Courriel : sheering@isr.umich.edu; Peter W. Solenberger, Applications Programmer Analyst Lead, Survey Research Center, Institute for Social Research, University of Michigan. Courriel : pws@isr.umich.edu.

problème comprend un ensemble de tableaux faisables avec répartition des unités de l'échantillon en nombres entiers non négatifs entre les cellules de chaque tableau, ainsi que les probabilités de sélection correspondant à chaque tableau. En deuxième lieu, l'un des tableaux de la solution est sélectionné aléatoirement en utilisant les probabilités assignées. Le nombre entier qui figure dans chacune des cellules du tableau de la solution choisi représente alors le nombre d'unités de l'échantillon qu'il faut allouer à la cellule correspondante de la stratification à deux dimensions. La clé de la sélection contrôlée est l'**algorithme** qui définit un ensemble de tableaux de solution qui réalise les **contrôles** pour résoudre le problème.

De nombreuses techniques de sélection contrôlée ont été élaborées depuis que Goodman et Kish (1950) ont décrit pour la première fois l'application de la sélection contrôlée au problème particulier du choix de 17 UPE pour représenter les États du Centre-Nord des États-Unis. Bryant, Hartley et Jessen (1960) ont proposé une méthode simple qui s'appliquait à un nombre limité de situations d'échantillonnage. Raghunandanan et Bryant (1971) ont généralisé cette méthode et Chernick et Wright (1983) ont proposé une alternative. Jessen (1970) a proposé deux méthodes appelées « méthode 2 » et « méthode 3 », toutes deux assez compliquées à mettre en œuvre et qui ne donnent parfois pas de solution. Jessen (1978, chapitre 11) a introduit un algorithme plus simple pour résoudre les problèmes de sélection contrôlée.

Hess, Riedel et Fitzpatrick (1975) ont donné une explication détaillée de la façon d'utiliser la sélection contrôlée pour tirer un échantillon représentatif d'hôpitaux du Michigan. Groves et Hess (1975) ont d'abord proposé un algorithme de calcul formel pour obtenir des solutions aux problèmes de sélection contrôlée avec stratification à deux et à trois dimensions. Heeringa et Hess (1983) ont publié la réponse à la question de Roe Goodman : Comment une solution informatique d'un problème de sélection hautement contrôlée se compare-t-elle à une solution manuelle? Leur réponse était que, pour un même plan de sondage, la sélection contrôlée générée par ordinateur mène généralement à des variances un peu plus élevées que la sélection contrôlée manuelle; mais puisque les différences de précision sont faibles et que la sélection contrôlée manuelle est laborieuse, la sélection contrôlée générée par ordinateur est privilégiée. Lin (1992) a amélioré l'algorithme de Groves et Hess (1975), et Heeringa (1998) a présenté le logiciel appelé « PCCONSEL » pour exécuter leur algorithme. Huang et Lin (1998) ont proposé un algorithme plus efficace, qui impose des contraintes additionnelles au problème de sélection contrôlée avec stratification à deux dimensions et s'exécute au moyen de tout progiciel standard de traitement du cheminement dans les réseaux (ou graphes). Hess et Heeringa (2002) résument les études portant sur la sélection contrôlée réalisées au cours de 40 années au *Survey Research Center* de la *University of Michigan*.

Adoptant une approche différente, Causey, Cox et Ernst (1985) ont proposé un algorithme qui appliquait aux problèmes de sélection contrôlée avec stratification à deux dimensions un modèle de transport basé sur la théorie proposée originalement dans un article précédent par Cox et Ernst (1982). Winkler (2001) a élaboré un algorithme de programmation par nombres entiers assez semblable à celui de Causey et coll. (1985). Deville et Tillé (2004) ont proposé un algorithme appelé méthode du cube.

En s'inspirant de Rao et Nigam (1990, 1992), Sitter et Skinner (1994) ont appliqué une méthode de programmation linéaire (PL) pour résoudre les problèmes de sélection contrôlée. Plus tard, Tiwari et Nigam (1998) ont proposé une méthode de PL qui réduit les probabilités de sélectionner des échantillons non privilégiés.

En résumé, une grande variété d'algorithmes pour la sélection contrôlée ont été étudiés et décrits dans la littérature. Ceux qui ont été élaborés le plus récemment sont particulièrement **exigeants du point de vue informatique**, puisqu'ils dépendent fortement de la disponibilité de logiciels et d'ordinateurs à haute vitesse.

Cependant, malgré cette évolution des algorithmes sur une période d'environ 60 ans, une question qui demeure encore sans réponse est celle de savoir dans quel sens les solutions d'un problème de sélection contrôlée obtenues au moyen de ces algorithmes sont **optimales**.

Dans le présent article, à la section 2, nous définissons le problème de sélection contrôlée à deux dimensions et réexaminons plusieurs problèmes de ce type qui ont été décrits dans la littérature antérieure. À la section 3, nous présentons les contraintes désirables. À la section 4, nous introduisons notre concept de **solutions optimales** des problèmes de sélection contrôlée. À la section 5, nous décrivons les faiblesses des algorithmes précédents. À la section 6, nous proposons un nouvel algorithme qui fait appel à l'approche de programmation linéaire (PL) pour obtenir des **solutions optimales** et à la section 7, nous présentons un nouveau **logiciel** accessible au public pour mettre en œuvre le nouvel algorithme de sélection contrôlée. À la section 8, pour montrer la **robustesse** du nouvel algorithme, nous l'appliquons à plusieurs exemples de problèmes de sélection contrôlée et comparons les résultats à ceux obtenus en utilisant les algorithmes existants. Enfin, nous présentons nos conclusions à la section 9.

2 Problèmes de sélection contrôlée

Afin de sélectionner un échantillon de n unités, considérons un plan de sondage avec stratification à deux dimensions comprenant la classification d'une population de N unités en fonction de deux critères possèdant R et C catégories, respectivement. Le problème de sélection contrôlée sous stratification à deux dimensions est défini par le tableau A de dimensions $R \times C$, qui est constitué de RC cellules contenant des nombres réels non négatifs a_{ij}, appelés **espérances de cellule**, représentant le nombre prévu d'unités qui doit être tiré dans chaque cellule ij. Le problème de sélection contrôlée à deux dimensions classique est décrit au tableau 2.1.

Tableau 2.1
Problème de sélection contrôlée de dimensions $R \times C$

a_{11}	a_{12}	. . . . .	a_{1C}	$a_{1.}$
a_{21}	a_{22}	. . . . .	a_{2C}	$a_{2.}$
.	.		.	.
.	.	. . a_{ij} . .	.	.
.	.		.	$a_{i.}$
.	.		.	.
a_{R1}	a_{R2}	. . . . .	a_{RC}	$a_{R.}$
$a_{.1}$	$a_{.2}$	. . $a_{.j}$ . .	$a_{.C}$	$a_{..}(=n)$

Les **espérances marginales** $a_{i.}$ et $a_{.j}$ désignent, respectivement, la somme des espérances de cellule dans chacune des catégories de la ligne i et dans chacune des catégories de la colonne j. D'où $a_{..}$ désigne la somme de toutes les espérances de cellule et est égal à la taille totale n de l'échantillon.

Bien que le tableau 2.1 prenne la forme d'un simple tableau à deux entrées, il convient de souligner qu'habituellement $n < RC$ et qu'en outre les a_{ij} peuvent être très petits (p. ex. souvent inférieurs à 1). Dans ces conditions, décider de la manière de répartir n unités entre les cellules, c'est-à-dire de la manière

d'obtenir un tableau de dimensions $R \times C$ avec les valeurs de cellules arrondies à un nombre entier non négatif pour chaque a_{ij} est un problème dont la résolution nécessite l'utilisation d'un algorithme.

Des problèmes de sélection contrôlée variés sont utilisés comme exemples dans la littérature. Le premier de ces exemples était le tableau de dimensions 17×4 décrit par Goodman et Kish (1950, p. 356), pour la répartition de 17 UPE entre 68 cellules données par 17 strates et 4 groupes d'États du Centre-Nord des États-Unis. Le tableau peut être formé comme il suit. Soit N_{ij} le nombre d'éléments de la population dans chaque cellule ij et soit $N_{i.}$ le nombre total d'éléments de la population dans chaque strate. Alors $a_{ij} = N_{ij} / N_{i.}$, où certaines valeurs de N_{ij} sont nulles et $0 \le a_{ij} < 1$. Tous les $a_{i.}$ sont égaux à l'entier 1, tandis que les $a_{.j}$ sont des sommes non entières des a_{ij} dans la colonne j. Le problème consiste donc à sélectionner une UPE par strate de l'échantillon (dimension i) et à contrôler simultanément la répartition entre les groupes d'États (dimension j). Au total, $n = 17$ UPE seront sélectionnées.

Les paragraphes qui suivent décrivent quatre problèmes supplémentaires que nous avons découverts dans la littérature et dont nous nous servirons pour la discussion et les évaluations comparatives présentées dans l'article.

Problème 2.1 : Jessen (1970)

Un problème de dimensions 3×3 faisant intervenir deux variables de stratification est donné par Jessen (1970, p. 779). Chaque cellule ij correspond à une UPE et $N = 9$. Un échantillon de taille $n = 6$ est tiré. $a_{ij} = nX_{ij} / X$, où X_{ij} est une « mesure de taille » pour l'UPE dans la cellule ij et $X = \sum_{i=1}^{R} \sum_{j=1}^{C} X_{ij}$. Notons que, dans ce problème, $0 < a_{ij} < 1$, et que $a_{i.}$ et $a_{.j}$ sont tous deux égaux à 2.

Problème 2.2 : Jessen (1978)

Une version étendue, de dimensions 4×4, du problème 2.1 est tirée de Jessen (1978, p. 375). Dans ce problème, $N = 16$ et $n = 8$. Comme dans le problème 2.1, $a_{i.}$ et $a_{.j}$ sont tous deux égaux à 2, mais $0 \le a_{ij} \le 1$.

Problème 2.3 : Causey et coll. (1985)

Causey et coll. (1985, p. 906) décrivent un problème de stratification à deux dimensions 8×3 conçu pour sélectionner 10 UPE, c'est-à-dire $n = 10$. Soit $X_{ijq_{ij}}$ $\left(q_{ij} = 1, \ldots, r_{ij}\right)$ une mesure de taille de l'UPE q_{ij} dans la cellule ij. Ici, $a_{ij} = n X_{ijq} / X_q$, où $X_{ijq} = \sum_{q_{ij}=1}^{r_{ij}} X_{ijq_{ij}}$ et $X_q = \sum_{i=1}^{R} \sum_{j=1}^{C} \sum_{q_{ij}=1}^{r_{ij}} X_{ijq_{ij}}$. Notons que, dans ce problème, $0 \le a_{ij} \le 2$, et la plupart des valeurs de $a_{i.}$ et $a_{.j}$ sont non entières.

Problème 2.4 : Winkler (2001)

Winkler (2001) fournit le problème de sélection contrôlée de dimensions 5×5 avec deux variables de stratification illustré au tableau 2.2.

L'objectif de la résolution de ce problème consiste à sélectionner $n = 37$ unités d'échantillon dans la population de taille $N = 1\,251$. La définition du problème débute par un tableau de dimensions 5×5 des

tailles de population des cellules N_{ij}, où certaines valeurs de N_{ij} sont assez petites. Les espérances marginales de ligne et de colonne, $a_{i.}$ et $a_{.j}$, sont des valeurs entières qui sont prédéterminées en utilisant l'information a priori sur la précision (p. ex. coefficients de variation).

Tableau 2.2
Problème de sélection contrôlée de dimensions 5×5

2,000	2,483	1,052	0,103	0,362	6
2,182	1,061	1,101	1,046	0,610	6
0,000	1,614	1,914	2,200	1,272	7
0,860	0,377	0,930	2,840	2,993	8
0,958	0,465	2,003	1,811	4,763	10
6	**6**	**7**	**8**	**10**	37

Source : Tableau 4, Annexe, Winkler (2001). Reproduit avec permission.

Les espérances de cellule, a_{ij}, sont obtenues en appliquant la procédure d'ajustement itératif généralisé de Dykstra (1985a, 1985b) et de Winkler (1990) au tableau initial. L'ajustement itératif généralisé est utilisé pour s'assurer que $a_{ij} < N_{ij}$ pour les cellules dont la valeur de N_{ij} est petite, les valeurs de $a_{i.}$ et $a_{.j}$ étant données. Notons que, dans le tableau 2.2, les a_{ij} sont données à la 3e décimale près, et que $0 \le a_{ij} < 5$.

La caractéristique commune à tous ces problèmes de sélection contrôlée est que, comme il est mentionné plus haut, le nombre total d'unités sélectionnées est plus petit que le nombre de cellules (sauf pour le problème 2.4, où $n = 37 > RC = 25$) et qu'un grand nombre de a_{ij} sont inférieures à 1. Les algorithmes utilisés pour résoudre ces problèmes doivent appliquer des contraintes strictes décrites à la section suivante. Comme il est indiqué à la section 4, la solution d'un problème de sélection contrôlée obtenue au moyen d'un algorithme comprend un ensemble de tableaux de dimensions $R \times C$ et les probabilités de sélection correspondant à chaque tableau.

3 Contraintes désirables

Chaque problème de sélection contrôlée de la forme illustrée au tableau 2.1 possède de nombreuses solutions en nombres entiers possibles. Soit B_k l'une de ces solutions, dont les entrées internes b_{ijk} sont le remplacement des nombres réels a_{ij} dans le problème de sélection contrôlée A par les nombres entiers non négatifs adjacents. L'entrée b_{ijk} est égale à $\lceil a_{ij} \rceil$ ou à $\lceil a_{ij} \rceil + 1$, où $\lceil \ \rceil$ est la fonction qui renvoie l'entier le plus grand. Si a_{ij} est un entier non négatif, $b_{ijk} = a_{ij}$ pour tout k. La même règle est appliquée aux espérances marginales. Comme l'ont souligné Jessen (1970) ainsi que Causey et coll. (1985), nous faisons principalement attention à B_k qui satisfait simultanément les **contraintes** suivantes pour tous i et j :

$$b_{ijk} \ge 0 \tag{3.1}$$

$$\left| b_{ijk} - a_{ij} \right| < 1 \tag{3.2}$$

$$\left| b_{i.k} - a_{i.} \right| < 1 \quad \text{et} \tag{3.3}$$

$$\left| b_{.jk} - a_{.j} \right| < 1, \tag{3.4}$$

où $b_{i.k} = \sum_{j=1}^{C} b_{ijk}$ est égal à $[a_{i.}]$ ou à $[a_{i.}]+1$, $b_{.jk} = \sum_{i=1}^{R} b_{ijk}$ est égal à $[a_{.j}]$ ou à $[a_{.j}]+1$, $\sum_{i=1}^{R} b_{i.k} = a_{..}$ et $\sum_{j=1}^{C} b_{.jk} = a_{..}$.

Considérons l'ensemble de **tous les tableaux possibles**, $\mathfrak{B} = \{B_k, k = 1,\ldots,L\}$, satisfaisant aux contraintes (3.1) à (3.4). Puisque a_{ij} est l'espérance de l'allocation de l'échantillon de chaque cellule de A, les **contraintes** (3.5) et (3.6) qui suivent appliquées sur b_{ijk} dans $B_k (\in \mathfrak{B})$ sont particulièrement importantes.

$$E\left(b_{ijk} \middle| i, j\right) = \sum_{B_k \in \mathfrak{B}} b_{ijk}\, p\left(B_k\right) = a_{ij}, \quad i = 1,\ldots,R, \text{ et } j = 1,\ldots,C \tag{3.5}$$

et

$$\sum_{B_k \in \mathfrak{B}} p\left(B_k\right) = 1, \tag{3.6}$$

où $p\left(B_k\right)$, qui dépend d'un algorithme spécifié pour résoudre le problème de sélection contrôlée, est la probabilité de sélection du tableau B_k et $p\left(B_k\right) \geq 0$.

Notons que (3.5) et (3.6) définissent une méthode d'**échantillonnage probabiliste** rigoureuse quand on sélectionne aléatoirement tout tableau dans $\mathfrak{B}$. Notons aussi que, puisque $\sum_{i=1}^{R}\sum_{j=1}^{C} E\left(b_{ijk} \middle| i, j\right) = a_{..} \sum_{B_k \in \mathfrak{B}} p\left(B_k\right) = a_{..}$, (3.5) implique (3.6) pour tout problème de sélection contrôlée tel que ceux décrits dans les problèmes 2.1 à 2.4. En outre, en guise d'illustration, lorsqu'elle est appliquée au problème 2.3, où $a_{ij} = n X_{ijq} / X_q$, la contrainte (3.5) donne

$$E\left(b_{ijk} / X_{ijq} \middle| i, j\right) = a_{ij} / X_{ijq} = n / X_q, \tag{3.7}$$

ce qui indique une équirépartition entre toutes les cellules.

4 Solutions optimales

Étant donné l'ensemble de L tableaux possibles dans $\mathfrak{B}$, considérons le sous-ensemble $\mathfrak{B}' (\subseteq \mathfrak{B})$ où

$$p\left(B_k\right) > 0.$$

Un **ensemble de solutions** d'un problème de sélection contrôlée A désigné par

$$\left\{\left(B_k, p\left(B_k\right)\right), B_k \in \mathfrak{B}'\right\}$$

est l'ensemble des tableaux qui possèdent les probabilités de sélection positives requises $\left(p\left(B_k\right) > 0\right)$. Cet ensemble de solutions, ou simplement une « solution » du problème de sélection contrôlée, est habituellement obtenu en se servant d'un algorithme pour appliquer les contraintes (3.1) à (3.6). Comme il est décrit dans l'introduction, depuis Goodman et Kish (1950), de nombreux algorithmes ont été élaborés pour trouver des solutions aux problèmes de sélection contrôlée.

Jusqu'à ce que Groves et Hess (1975) proposent un algorithme informatique, la plupart des solutions étaient obtenues manuellement selon un processus qui ressemble à la résolution d'un casse-tête mathématique. En outre, pour la plupart des problèmes, il se peut que les contraintes soient satisfaites par plus d'un ensemble de solutions. Depuis les années 1980, on a élaboré des algorithmes de sélection contrôlée, exigeants du point de vue informatique, qui s'appuient sur la théorie du transport, le cheminement dans les

réseaux, la programmation par nombres entiers et la programmation linéaire. Ces algorithmes dépendent parfois de logiciels hautement spécialisés ou peuvent être programmés pour être exécutés dans les grands systèmes logiciels.

Cependant, les solutions antérieures allant des algorithmes manuels aux algorithmes exigeants du point de vue informatique ont rarement été comparées empiriquement en appliquant un jeu normalisé de critères de performance. Par conséquent, nous commençons ici par décrire un concept appelé **ensembles de solutions optimaux**, ou plus simplement, **solutions optimales**.

Le problème de sélection contrôlée A ne comporte qu'un seul tableau, mais il pourrait exister de nombreux tableaux possibles dans $\mathfrak{B}$. En outre, un seul tableau B_k provenant de toute solution de A est choisi aléatoirement en appliquant $p\left(B_k\right)$ comme fondement de la sélection de l'échantillon stratifié. Donc, en général, nous pourrions définir une solution optimale comme étant celle qui satisfait les **exigences** suivantes (E1 et E2) :

E1. La solution est obtenue en se basant sur des mesures appropriées et objectives de la **proximité** de A par rapport à chaque tableau individuel B_k dans $\mathfrak{B}$.

E2. La solution maximise, dans la mesure du possible, les probabilités de sélection sur les **tableaux les plus proches** de A sous des mesures telles que celles mentionnées en E1.

La suite de la présente section porte sur la façon de spécifier E1 et E2 pour obtenir des solutions optimales. Premièrement, afin de définir la **proximité** dans E1, on peut considérer un nombre réel $d(B_k : A)$ représentant la distance entre A et B_k, où d est une fonction de distance qui satisfait les axiomes suivants :

(i) $d\left(B_k, A\right) > 0$ si $B_k \neq A$; $d\left(A, A\right) = 0$;

(ii) $d\left(B_k, A\right) = d\left(A, B_k\right)$;

(iii) $d\left(B_k, A\right) \leq d\left(B_k, B_k^{'}\right) + d\left(B_k^{'}, A\right)$ pour tout $B_k^{'} \in \mathfrak{B}$.

L'axiome (iii) porte le nom d'**axiome d'inégalité triangulaire**. Les fonctions de distance qui satisfont (i), (ii) et (iii) peuvent être définies en utilisant les deux RC-tuples ordonnés $\left(a_{11}, a_{12}, \cdots, a_{RC}\right)$ et $\left(b_{11k}, b_{12k}, \cdots, b_{RCk}\right)$ pour A et B_k. Nous commençons par définir la distance ordinaire ou **distance euclidienne** (**distance définie par la norme 2**) :

$$d_2\left(B_k, A\right) = \left[\sum_{i=1}^{R}\sum_{j=1}^{C}\left(b_{ijk} - a_{ij}\right)^2\right]^{\frac{1}{2}}, \quad k = 1, \ldots, L. \tag{4.1}$$

Cette fonction est probablement la mesure la plus connue pour définir la distance entre B_k et A.

Nous pouvons aussi définir la fonction appelée **distance de Chebyshev** (**distance définie par la norme infinie**) :

$$d_\infty\left(B_k, A\right) = \max\left\{\left|b_{ijk} - a_{ij}\right| : i = 1, \ldots, R,\ j = 1, \ldots, C\right\}, \quad k = 1, \ldots, L. \tag{4.2}$$

Ces fonctions de distance donnent naissance à deux espaces de distances distincts. En vertu de (3.2), pour tout B_k, les expressions qui suivent sont vérifiées.

$$0 \le d_2\left(B_k, A\right) < \left(RC\right)^{1/2} \tag{4.3}$$

et

$$0 \le d_\infty\left(B_k, A\right) < 1. \tag{4.4}$$

Par exemple, pour le tableau de dimensions 3×3 dans le problème 2.1 et le tableau de dimensions 8×3 dans le problème 2.3, $0 < d_2\left(B_k, A\right) < 3$ et $0 < d_2\left(B_k, A\right) < 4{,}9$, respectivement.

Deuxièmement, comme il est mentionné dans E2, en ce qui concerne les **tableaux les plus proches** de A sous de mesures telles que celles décrites en E1, considérons l'ensemble de tableaux dans $\mathfrak{B}$ possédant la valeur de d_2 ou d_∞ minimale par rapport à A. Soit $\mathfrak{B}_2 (\subseteq \mathfrak{B}')$ l'ensemble des tableaux ayant la valeur de d_2 minimale par rapport à A et $\mathfrak{B}_\infty (\subseteq \mathfrak{B}')$ l'ensemble des tableaux ayant la valeur de d_∞ minimale par rapport à A.

En supposant que tous les tableaux possibles dans $\mathfrak{B}$ sont connus, nous définissons les **tableaux optimaux** comme il suit.

Définition. Les tableaux compris dans $\mathfrak{B}_2 \cup \mathfrak{B}_\infty$ sont appelés **tableaux optimaux**.

Notons que, dans le nouvel algorithme pour la sélection contrôlée qui sera décrit à la section 6, d_2 ou d_∞ sont choisies en fonction des préférences. Nous évitons de définir l'intersection de $\mathfrak{B}_2$ et $\mathfrak{B}_\infty$ comme étant les tableaux optimaux, parce que cela pourrait exclure les autres tableaux non compris dans $\mathfrak{B}_2 \cap \mathfrak{B}_\infty$ ayant la même valeur de d_2 (d_∞) minimale. Nous illustrons ci-après le fait qu'il pourrait exister un très petit nombre de tableaux optimaux relativement au nombre total de tableaux possibles dans $\mathfrak{B}$ pour tout A. La façon de trouver tous les tableaux possibles sera décrite en détail aux sections 6 et 7.

Illustrations

Pour les problèmes 2.1 à 2.4, nous notons que $\mathfrak{B}_2 \subseteq \mathfrak{B}_\infty$. Donc, nous pouvons utiliser d_∞ seulement pour illustrer les tableaux optimaux.

1. Pour le problème 2.1, il existe six tableaux possibles satisfaisant (3.1), (3.2), (3.3) et (3.4). Autrement dit, $\mathfrak{B} = \left\{B_k, k = 1, \ldots, 6\right\}$, tel que donné dans le tableau 4.1. Il n'existe qu'un seul tableau optimal, B_2, ayant la valeur minimale de $d_\infty = 0{,}5$.

Tableau 4.1
Problème de sélection contrôlée de dimensions 3×3, tableau optimal avec $d_\infty = 0{,}5$ et les autres tableaux

A	B_1	B_2	B_3	B_4	B_5	B_6
0,8 0,5 0,7	0 1 1	1 0 1	1 1 0	0 1 1	1 0 1	1 1 0
0,7 0,8 0,5	1 0 1	1 1 0	0 1 1	1 1 0	0 1 1	1 0 1
0,5 0,7 0,8	1 1 0	0 1 1	1 0 1	1 0 1	1 1 0	0 1 1
d_∞	0,8	0,5	0,7	0,8	0,8	0,8

2. Pour le problème 2.2, il existe 30 tableaux possibles, et trois tableaux optimaux illustrés au tableau 4.2.

Tableau 4.2
Tableaux optimaux de dimensions 4×4 **avec** $d_\infty = 0,6$

0	0	1	1
1	1	0	0
0	0	1	1
1	1	0	0

0	1	1	0
1	0	0	1
0	0	1	1
1	1	0	0

0	1	1	0
1	0	1	0
1	0	0	1
0	1	0	1

3. Pour le problème 2.3, il existe 141 tableaux possibles. Il y a six tableaux optimaux, ayant tous la même distance $d_\infty = 0,6$. L'un d'eux est illustré au tableau 4.3.

Tableau 4.3
Un des six tableaux optimaux avec $d_\infty = 0,6$

1	2	0
1	0	1
0	0	0
1	1	0
1	1	0
0	0	1
0	0	0
0	0	0

4. Pour le problème 2.4, il existe 159 tableaux possibles et il n'y a qu'un seul tableau optimal donné au tableau 4.4.

Tableau 4.4
Tableau optimal de dimensions 5×5 **avec** $d_\infty = 0,517$

2	3	1	0	0
2	1	1	1	1
0	2	2	2	1
1	0	1	3	3
1	0	2	2	5

Par conséquent, en nous fondant sur la définition des tableaux optimaux, et sur le fait que d_2 et d_∞ satisfont les axiomes (i), (ii) et (iii), nous proposons les **spécifications** suivantes (S1 et S2) de E1 et E2 des solutions optimales :

S1. La solution est basée sur les valeurs de la distance $d_2\left(d_\infty\right)$ entre A et chaque tableau individuel B_k dans $\mathfrak{B}$.

S2. La solution maximise les probabilités de sélection des tableaux optimaux.

S1 et S2 représenteront les rudiments d'un nouvel algorithme présenté à la section 6, et à la section suivante, nous passons à la discussion des algorithmes antérieurs dans la perspective des solutions optimales.

5 Propriétés non optimales des méthodes existantes

Comme il est décrit à la section 4, les algorithmes pour la sélection contrôlée peuvent être répartis en deux classes, les algorithmes manuels d'avant les années 1980 et les algorithmes exigeants du point de vue informatique élaborés depuis. Pour les grands problèmes de sélection contrôlée comportant de nombreuses cellules, la préférence pourrait être donnée à la seconde classe d'algorithmes. Cependant, quand le problème est petit, la première classe peut être utilisée facilement sans devoir faire face à la complexité de la seconde. Par conséquent, nous ne dirions pas que la première est toujours inférieure à la seconde. Des critères plus objectifs sont nécessaires pour comparer ces deux classes d'algorithmes, et la solution optimale pourrait être considérée comme l'un des meilleurs critères pour comparer leurs forces et leurs faiblesses.

Comme il est discuté dans Jessen (1978, p. 375-376), les algorithmes décrits dans Jessen (1970) visent à minimiser le nombre de tableaux dans un ensemble de solutions $\mathfrak{B}'$, et l'algorithme de Jessen (1978) atteint assez facilement cet objectif comparativement à ceux de Jessen (1970). Donc, les algorithmes de cet auteur recherchent la « simplicité » dans la formulation d'une solution plutôt qu'une solution optimale.

L'algorithme de Causey et coll. (1985) peut donner une solution « partiellement » optimale. Outre le problème original, A, il crée séquentiellement un petit nombre de nouveaux problèmes de sélection contrôlée, puis en guise de solution, il ne trouve qu'un seul tableau $B_k (\in \mathfrak{B})$ comme étant **le plus proche** de chaque problème, en commençant par A. Chaque problème est considéré comme étant le problème de transport de Cox et Ernst (1982), qui est formé par la fonction objectif imitant le comportement de

$$\sum_{i=1}^{R}\sum_{j=1}^{C}\left|b_{ijk}-a_{ij}\right|^p,\ k=1,\ldots,L,\ 1\le p<\infty. \tag{5.1}$$

Notons que, comme la fonction (5.1) viole l'axiome de l'inégalité triangulaire (iii), ce n'est pas une fonction de distance. L'inclusion de la p^{e} racine est nécessaire pour qu'elle soit une fonction de distance. En outre, chaque $p\left(B_k\right)$ est calculée au moyen d'une formule simple. Étant donné les exigences d'optimalité données par E1 et E2, l'algorithme de Causey et coll. présente les faiblesses suivantes : 1) comme d'autres problèmes de sélection contrôlée sont ajoutés au problème original A, il est difficile d'obtenir la solution systématiquement basée sur la **proximité** entre le problème unique A et chaque B_k individuel dans $\mathfrak{B}$; 2) la maximisation des probabilités de sélection pour les **tableaux les plus proches** de A n'est pas garantie.

Winkler (2001) a présenté une modification de la méthode de Causey et coll. (1985). Au lieu d'utiliser le problème du transport, il a proposé un problème de programmation linéaire par nombres entiers donnant lieu à de légers changements de la probabilité $p\left(B_k\right)$. Néanmoins, l'algorithme de Winkler (2001) n'est pas exempt des faiblesses de la méthode de Causey et coll. (1985).

En adoptant une approche de problème de cheminement dans les réseaux, l'algorithme de Huang et Lin (1998) impose dans A les contraintes de sous-groupes additionnels soulevées par Goodman et Kish (1950). Cependant, tout comme ceux de Causey et coll. (1985) et de Winkler (2001), il n'atteint pas les objectifs E1 et E2, puisqu'il génère un nouveau réseau au lieu d'un nouveau problème de sélection contrôlée à chaque itération, qu'un $B_k \left(\in \mathfrak{B}\right)$ arbitraire est obtenu comme solution du réseau et que $p\left(B_k\right)$ est calculée au moyen d'une simple formule.

Par contre, les algorithmes de programmation linéaire (PL) proposés par Sitter et Skinner (1994) et par Tiwari et Nigam (1998) utilisent tous les tableaux possibles compris dans $\mathfrak{B}$. Notons que trouver tous ces tableaux est un enjeu important et que les $p\left(B_k\right)$ pour tous les tableaux possibles sont obtenues simultanément en exécutant le logiciel pour la PL une seule fois. L'idée essentielle qui sous-tend l'algorithme de Sitter et Skinner (1994) est l'utilisation d'une « fonction de perte » définie par

$$\sum_{i=1}^{R}\left(b_{i.k}-a_{i.}\right)^2+\sum_{j=1}^{C}\left(b_{.jk}-a_{.j}\right)^2. \tag{5.2}$$

En ce qui concerne E1 et E2, leur algorithme présente les inconvénients suivants : 1) la proximité entre A et B_k n'est pas bien traduite par la fonction de perte (5.2). Cela tient au fait qu'il ne s'agit pas d'une fonction de distance qui satisfait l'axiome (iii), car ce sont les totaux de marge qui sont utilisés plutôt que les entrées dans les cellules; 2) la fonction de perte (5.2) est sans pertinence pour la maximisation des probabilités de sélection sur les **tableaux les plus proches** de A dans les problèmes 2.1, 2.2 et 2.4, puisqu'elle est toujours nulle.

La méthode de PL de Tiwari et Nigam (1998) peut être utilisée pour réduire les probabilités de sélection des tableaux non privilégiés (p. ex. les tableaux ne contenant pas les UPE qui correspondent à la cellule $ij = 23$ dans le problème 2.1), qui sont déterminés au départ par les échantillonneurs. Pour les problèmes de sélection contrôlée avec valeurs marginales entières et sans tenir compte des tableaux non privilégiés, leur méthode donne les mêmes solutions que celles de Sitter et Skinner (1994).

À la section 8, nous comparerons au moyen de plusieurs exemples les solutions issues de ces méthodes antérieures à celles obtenues par la méthode que nous proposons à la section 6.

6 Méthode proposée

À la présente section, nous décrivons en détail un algorithme pour arriver aux spécifications S1 et S2 des solutions optimales décrites à la section 4.

6.1 L'algorithme

L'algorithme possède les **caractéristiques** suivantes : 1) il trouve une solution fondée directement sur les valeurs de la distance d_2 $\left(d_\infty\right)$ entre le problème de sélection contrôlée A et chaque tableau individuel B_k compris dans $\mathfrak{B}$; 2) il est exigeant du point de vue informatique, mais est facilement mis en œuvre par programmation linéaire; 3) il est applicable à tout type de problème de sélection contrôlée avec stratification à deux dimensions.

L'algorithme comporte les cinq étapes qui suivent :

Étape 1. Trouver l'ensemble de tous les tableaux possibles, $\mathfrak{B}$, qui satisfont les contraintes (3.1) à (3.4) pour un problème de sélection contrôlée donné A. En particulier, si A contient des espérances marginales non entières, trouver toutes les valeurs arrondies possibles de ces espérances marginales par entiers adjacents, qui satisfont (3.3) et (3.4). Ces valeurs entières marginales arrondies seront $[a_{i.}]$ ou $[a_{i.}]+1$ $\left([a_{.j}] \text{ ou } [a_{.j}]+1\right)$, tandis que les espérances marginales entières seront retenues, puisque $[a_{i.}]=a_{i.}$ $\left([a_{.j}]=a_{.j}\right)$. Ensuite, trouver tous les tableaux possibles satisfaisant (3.1) et (3.2) sous les valeurs entières marginales arrondies et les autres valeurs entières marginales.

Étape 2. Choisir $d_2^*\left(B_k, A\right)$ ou $d_\infty^*\left(B_k, A\right)$ (selon les préférences) et calculer la fonction de distance choisie pour chaque $B_k\left(\in \mathfrak{B}\right)$, où :

$$d_2^*\left(B_k, A\right) = d_2\left(B_k^*, A^*\right) = \left[\sum_{i=1}^{R}\sum_{j=1}^{C}\left(b_{ijk}^* - a_{ij}^*\right)^2\right]^{\frac{1}{2}} \tag{6.1}$$

$$d_\infty^*\left(B_k, A\right) = d_\infty\left(B_k^*, A^*\right) = \max\left\{\left|b_{ijk}^* - a_{ij}^*\right| : i=1,\ldots,R,\ j=1,\ldots,C\right\}. \tag{6.2}$$

Notons que, comme chacune des cellules ij du tableau du problème, A, recevra une allocation minimale égale à $[a_{ij}]$ avec certitude, les fonctions de distance ne doivent prendre en considération que la partie non entière de a_{ij} :

$$a_{ij}^* = a_{ij} - \left[a_{ij}\right], \tag{6.3}$$

et la différence entière (soit 0 ou 1) entre la taille d'échantillon allouée, b_{ijk}, pour la solution $k=1,\ldots,L$ et celle allouée avec certitude pour la ij^{e} cellule de A :

$$b_{ijk}^* = b_{ijk} - \left[a_{ij}\right]. \tag{6.4}$$

Étape 3. Selon la fonction de distance choisie à l'étape 2, construire le problème de PL qui suit consistant à minimiser la fonction objectif (6.5) ou (6.6), qui est une forme linéaire, sous les contraintes linéaires (6.7) et (6.8) :

Minimiser

$$FO_1 = \sum_{B_k \in \mathfrak{B}} d_2^*\left(B_k, A\right) p\left(B_k\right) \tag{6.5}$$

ou

$$FO_2 = \sum_{B_k \in \mathfrak{B}} d_\infty^*\left(B_k, A\right) p\left(B_k\right) \tag{6.6}$$

sous les contraintes

$$\sum_{B_k \in \mathfrak{B}} b_{ijk}^*\, p\left(B_k\right) = a_{ij}^*,\ i=1,\ldots,R,\ j=1,\ldots,C, \tag{6.7}$$

et

$$p\left(B_k\right) \geq 0,\ k=1,\ldots,L. \tag{6.8}$$

Étape 4. En utilisant un algorithme pour la PL, résoudre le problème de PL établi à l'étape 3 pour L variables inconnues

$$\left\{p\left(B_k\right), B_k \in \mathfrak{B}\right\}. \tag{6.9}$$

Étape 5. Obtenir l'ensemble de solutions $\left\{\left(B_k, p\left(B_k\right)\right), B_k \in \mathfrak{B}'\right\}$ pour A constitué de tableaux tels que $p\left(B_k\right) > 0$ dans l'ensemble de solutions du problème de PL obtenu à l'étape 4.

Quelques remarques utiles pour mettre en œuvre l'algorithme sont de rigueur.

Remarque 6.1. À l'étape 2, il convient de souligner que $\left[a_{ij}\right]$ dans (6.3) ou (6.4) indique le nombre d'unités qui doivent être sélectionnées avec certitude dans chaque cellule. Il convient aussi de souligner que

$$d_2^*\left(B_k, A\right) = d_2\left(B_k, A\right) \tag{6.10}$$

et

$$d_\infty^*\left(B_k, A\right) = d_\infty\left(B_k, A\right), \tag{6.11}$$

puisque $b_{ijk}^* - a_{ij}^* = b_{ijk} - a_{ij}$ en raison de (6.3) et (6.4).

Remarque 6.2. Outre le fait que d_2^* est le concept naturel de distance et que d_∞^* est la plus simple et la plus facile à calculer sous la norme, il existe un conseil sensé quant au choix de d_2^* ou d_∞^* à l'étape 2. Soit D_2 et D_∞ les ensembles de valeurs de distance pour tous les tableaux possibles calculés pour d_2^* et d_∞^*, respectivement. Posons que les tableaux ayant la même valeur de distance dans D_2 (D_∞) se trouvent dans le même groupe. Alors, logiquement, d_2^* regroupera les tableaux possibles en un grand nombre de groupes différents, où le nombre de groupes est plus grand que pour d_∞^*, en raison de (4.3) et (4.4). Par conséquent, si l'on utilise d_2^* dans le problème de PL, le nombre de tableaux dans $\mathfrak{B}$ tel que $p(B_k) > 0$ sera plus grand que si l'on utilise d_∞^*.

Remarque 6.3. Il est clair, d'après (6.5) et (6.6) faisant intervenir les valeurs de distance d_2^* ou d_∞^*, que la solution obtenue à l'étape 5 aboutit à la réalisation sûre de S1. En outre, la spécification S2 est obtenue efficacement en utilisant les contraintes linéaires (6.7) et (6.8).

Remarque 6.4. En construisant le problème de PL à l'étape 3, les contraintes pour les cellules pour lesquelles $a_{ij}^* = 0$ peuvent être omises dans (6.7). Par exemple, pour le problème de sélection contrôlée de dimensions 5×5 du problème 2.4, le nombre de contraintes nécessaires est 23, puisque $a_{ij}^* = 0$ dans deux cellules. En outre, la contrainte linéaire (3.6) n'est pas essentielle, parce qu'elle est implicite dans (6.7).

6.2 Utilisation de la méthode du simplexe

Le problème de PL construit à l'étape 3 avec le système de contraintes de RC équations en (6.7) pour L inconnues non négatives en (6.8) est la « forme classique » et ne requiert aucune transformation.

Si $RC < L$, le nombre d'équations est plus petit que le nombre d'inconnues. Par conséquent, il s'agit d'un problème de PL de forme classique et il peut toujours être résolu par la méthode du simplexe en transformant le système de RC contraintes en sa forme canonique. Pour transformer le système en sa forme canonique, nous pourrions choisir arbitrairement RC variables parmi les L variables comme **variables de base**, puis, en utilisant une opération pivot, essayer de mettre le système sous forme canonique, où chaque variable de base a un coefficient de 1 dans une équation et de 0 dans les autres, et chaque équation contient exactement une variable de base dont le coefficient est égal à 1.

En permettant que les $L - RC$ variables autres que les RC variables choisies comme variables de base prennent une valeur nulle dans le système sous forme canonique, nous obtenons la **solution faisable de base** initiale. Ensuite, en remplaçant exactement une variable de base, nous obtenons une autre solution faisable de base, et nous poursuivons ces étapes jusqu'à ce que la valeur minimale de la fonction objectif soit atteinte par une des solutions faisables de base. L'ensemble de ces solutions faisables de base du problème de PL est convexe. De nombreux progiciels sont disponibles pour appliquer la méthode du simplexe pour résoudre le problème de PL. Voir Dantzig (1963) ainsi que Thie et Keough (2008, chapitre 3) pour des renseignements détaillés sur la méthode du simplexe.

6.3 Les demandes informatiques du problème de PL

On peut soutenir que notre algorithme est coûteux en ressources informatiques pour les raisons suivantes :

a. Avant de résoudre le problème de PL, tous les tableaux possibles de résolution du problème de sélection contrôlée doivent être connus.

b. Le nombre d'inconnues dans le problème de PL, L, est égal au nombre total de tableaux possibles, lequel devient grand à mesure que RC, le nombre de cellules dans le problème de sélection contrôlée, augmente. D'où, il n'est pas déraisonnable que L puisse être aussi grand que le coefficient binomial

$$\binom{RC}{a_{..}^*}, \text{ où } a_{..}^* = a_{..} - \sum_{i=1}^{R}\sum_{j=1}^{C}\left[a_{ij}\right]. \tag{6.12}$$

c. Si RC est grand, il donne aussi un grand nombre de contraintes dans (6.7).

Sitter et Skinner (1994), ainsi que Tiwari et Nigam (1998) font aussi référence à ces inconvénients possibles en décrivant leurs algorithmes de PL. Toutefois, pour les raisons qui suivent, les fardeaux informatiques mentionnés aux points a, b et c pourraient ne pas être prohibitifs dans le contexte des **opérations réelles**.

Premièrement, trouver tous les tableaux possibles manuellement pourrait être difficile pour tout problème de sélection contrôlée comprenant un grand nombre de cellules, mais cette tâche est grandement simplifiée en utilisant un algorithme efficace et la puissance des ordinateurs modernes. En utilisant le logiciel décrit à la section suivante, les tableaux peuvent être obtenus facilement en quelques secondes, même dans le cas de problèmes relativement grands tels que les problèmes 2.3 et 2.4.

Deuxièmement, l'application de (6.12) aux problèmes 2.1 à 2.4 donne, respectivement, 84, 11 440, 10 626 et 4 457 400 tableaux. Cependant, les nombres réels pour L sont seulement 6, 30, 141 et 159, respectivement. Il en est ainsi parce que les espérances marginales des lignes ainsi que des colonnes sont

appariées simultanément et que certaines espérances de cellule sont nulles. Les nombres réels peuvent également être obtenus au moyen du logiciel décrit à la section suivante.

Troisièmement, bien que le nombre de contraintes dans le problème de PL soit d'autant plus élevé que RC est grand, les demandes informatiques peuvent dépendre de L ainsi que de RC, et plus particulièrement, du nombre de solutions faisables de base, possiblement désignées par

$$S = \binom{L}{RC}. \tag{6.13}$$

Par exemple, si $L = 1\,000$ et $RC = 100$, (6.13) donne 6,4E+139, qui est un nombre extrêmement grand. Dans ce cas, il est presque impossible de résoudre le problème de PL, puisque chaque solution faisable de base devrait être examinée. Toutefois, des cas de ce genre ne se produiraient pas en pratique. Selon Ross (2007, p. 221-224), quand $RC < L$, le **nombre de transitions nécessaires**, disons T, lorsqu'on chemine le long des solutions faisables de base en résolvant le problème de PL ayant une forme classique suit approximativement une loi normale de moyenne $E(T) = \log_e S$ et de variance $Var(T) = \log_e S$, où

$$\log_e S \approx RC\left[1 + \log_e\left\{\left(L/RC\right) - 1\right\}\right]. \tag{6.14}$$

Lorsqu'on applique cette théorie au cas de $L = 1\,000$ et $RC = 100$, l'approximation de la moyenne et de la variance de T par (6.14) donne 320, et l'intervalle de confiance (IC) à 95 % de T est (285, 355), chiffres qui sont plus petits que les bornes inférieure et supérieure prévues.

Tableau 6.1
Comparaison entre S et T

	Problème 2.1	Problème 2.2	Problème 2.3	Problème 2.4
L	6	30	141	159
RC^*	9	14	13	23
S	ND	1,5E+8	7,9E+17	3,1E+27
$E(T)$	ND	16	43	64
IC à 95 % de T	ND	(8, 24)	(30, 56)	(48, 80)

Nota : ND - non disponible

Le tableau 6.1 montre les résultats de la comparaison entre S et T pour les quatre problèmes considérés plus haut. Notons que, étant donné la remarque 6.4, dans (6.13) et (6.14), RC est remplacé par RC^*, c'est-à-dire le nombre obtenu en soustrayant le nombre de cellules pour lesquelles $a_{ij}^* = 0$ de RC. La théorie sur T n'est pas appliquée au problème 2.1 parce que $RC^* > L$.

Comme le montre le tableau, la moyenne ou les bornes de l'intervalle de confiance de T sont considérablement plus petites que S dans chaque problème. À la section 8, T dans le tableau 6.1 sera comparé au **nombre réel de transitions**, disons t.

7 Logiciel

Afin de tirer parti de la puissance des ordinateurs modernes, nous avons développé un logiciel basé sur SAS, appelé SOCSLP (pour *Software for Optimal Control Selection Linear Programming*) pour exécuter

notre algorithme en vue de résoudre des problèmes de sélection contrôlée avec stratification à deux dimensions. La version récente peut être téléchargée à l'adresse : http://www.isr.umich.edu/src/smp/socslp.

En utilisant le logiciel, aucune contrainte n'est imposée au nombre de tableaux possibles qui peuvent être pris en considération pour la solution. Le nombre de ces tableaux et le nombre de contraintes qui peuvent être résolues dépendent de la capacité de mémoire et de l'espace disque disponible de l'ordinateur.

La méthode du simplexe révisée en deux phases, implémentée en utilisant la procédure SAS/OR LP, ou plus simplement « PROC LP », est utilisée pour résoudre le problème de PL. Une solution optimale unique au problème de PL s'obtient quand la fonction objectif est minimisée sous les contraintes données (6.7) au cours des phases 1 et 2 de PROC LP, sous l'hypothèse que toutes les variables inconnues sont non négatives (6.8).

Le logiciel produit beaucoup d'information, y compris l'ensemble de solutions du problème de sélection contrôlée. En outre, en choisissant une simple option dans le logiciel, un tableau peut être sélectionné aléatoirement à partir de l'ensemble de solutions, pour achever la sélection contrôlée. Le SOCSLP est disponible à l'heure actuelle pour les ordinateurs personnels, et des renseignements détaillés sont fournis dans le guide de l'utilisateur sur le site Web.

8 Comparaisons des algorithmes

En utilisant les quatre problèmes de sélection contrôlée mentionnés à la section 2, nous présentons certains résultats produits par les **deux méthodes** en utilisant d_2^* et d_∞^* dans le nouvel algorithme, et en comparant les solutions données par ces méthodes aux solutions générées sous les algorithmes décrits antérieurement par Jessen (1970), Jessen (1978), Causey et coll. (1985), Huang et Lin (1998) et Winkler (2001). Les solutions produites par les deux méthodes en utilisant d_2^* et d_∞^* ont été obtenues avec le SOCSLP, avec la version 9.2 de SAS/OR (2008). Les solutions de l'algorithme de Sitter et Skinner (1994) en utilisant la programmation linéaire ont également été obtenues en utilisant PROC LP de la version 9.2 de SAS/OR (2008). Les solutions pour les autres méthodes sont les résultats qui ont été publiés dans les articles originaux.

Les réponses à deux questions nous aident à comparer les algorithmes : 1) les solutions issues des nouvelles méthodes diffèrent-elles de celles fournies par les algorithmes antérieurs décrits à la section 5? 2) les solutions issues des nouvelles méthodes donnent-elles pour les tableaux optimaux des probabilités de sélection plus élevées que celles générées en utilisant les méthodes antérieures?

Avant de comparer les algorithmes, nous devons examiner les résultats du tableau 8.1 obtenus au moyen des deux méthodes. Dans le tableau, la méthode utilisant d_2^* et celle utilisant d_∞^* sont désignées par N_2 et N_∞, respectivement. Étant donné que, quand ils sont calculés avec d_2^* (d_∞^*), les tableaux ayant la même valeur de distance se trouvent dans le même groupe, il existera des groupes différents pour tous les tableaux possibles (voir la remarque 6.2). Soit G le nombre de groupes différents. En outre, soit FO la valeur réelle de la fonction objectif (6.5) ou (6.6) et t, le nombre réel de T, le nombre de transitions, présenté à la section 6.3. Ces valeurs sont toutes obtenues au moyen du SOCSLP, et t en particulier indique le nombre d'itérations aux phases 1 et 2 de PROC LP dans le logiciel.

Tableau 8.1
Résultats obtenus avec les nouvelles méthodes

	Problème 2.1		Problème 2.2		Problème 2.3		Problème 2.4	
	N_2	N_∞	N_2	N_∞	N_2	N_∞	N_2	N_∞
G	4	3	9	2	6	2	157	14
FO	1,336	0,620	1,689	0,640	1,582	0,720	1,661	0,701
t	2	2	8	6	18	15	43	41

Comme le montre le tableau, la plupart des valeurs de G sont beaucoup plus petites que celles de L, le nombre total de tableaux possibles donné dans le tableau 6.1, sauf dans le cas de la grande valeur de « 157 » pour le problème 2.4, qui découle simplement du fait que les valeurs de a_{ij} sont données à trois décimales près. Lorsqu'on utilise d_2^*, les valeurs de FO varient entre 1 et 2, tandis qu'elles sont toujours inférieures à 1 lorsqu'on utilise d_∞^*. La plupart des valeurs de t n'atteignent pas l'IC à 95 % de T indiqué au bas du tableau 6.1. Donc, les demandes de ressources informatiques réelles sont inférieures à celles prévues par la théorie.

Les solutions produites par différents algorithmes pour les trois premiers problèmes sont présentées par ordre dans les tableaux 8.2 à 8.4. Les résultats pour le problème 2.4 sont décrits simplement ci-dessous. (Le tableau des solutions de ce problème peut être obtenu sur demande.) Dans le tableau 8.2, la méthode de Sitter et Skinner (1994), et les méthodes 2 et 3 de Jessen (1970) sont désignées par SS, $J2$ et $J3$, respectivement. Les solutions pour $J2$ et $J3$ dans le tableau sont tirées de Jessen (1970, p. 782). Le tableau montre que toutes les méthodes, sauf la méthode 3 de Jessen (1970) donnent la même solution pour le tableau de dimensions 3 x 3 du problème 2.1. Dans les solutions communes, la probabilité de sélection des tableaux optimaux, désignée par $\sum_{B_k \in \mathfrak{B}_\infty} p(B_k)$, est de 0,5.

Tableau 8.2
Comparaison des solutions du problème 2.1

B_k	$p(B_k)$				
	N_2	N_∞	SS	$J2$	$J3$
0 1 1 1 0 1 1 1 0	0,2	0,2	0,2	0,2	0,1
1 0 1 * 1 1 0 0 1 1	0,5	0,5	0,5	0,5	0,4
1 1 0 0 1 1 1 0 1	0,3	0,3	0,3	0,3	0,2
0 1 1 1 1 0 1 0 1					0,1
1 0 1 0 1 1 1 1 0					0,1
1 1 0 1 0 1 0 1 1					0,1
Total	1,0	1,0	1,0	1,0	1,0
Total †	0,5	0,5	0,5	0,5	0,4

Nota : * - Tableau optimal
† - La somme des probabilités de sélection pour les tableaux optimaux.

Dans le tableau 8.3, la méthode de Jessen (1978) est désignée par JS. La solution pour JS présentée dans le tableau est tirée de Jessen (1978, p. 375-376). Comme le montre le tableau, les nouvelles méthodes en utilisant d_2^* et d_∞^* donnent la même solution pour le tableau de dimensions 4×4 du problème 2.2; cependant, la moitié seulement des tableaux figurant dans ces solutions concorde avec les tableaux figurant dans les solutions produites par les méthodes de Sitter et Skinner (1994) et Jessen (1978). En outre, les méthodes de Sitter et Skinner et de Jessen donnent une probabilité plus faible, égale à 0,6, aux tableaux optimaux, tandis que les nouvelles méthodes attribuent une probabilité plus élevée, égale à 0,8, aux tableaux.

Tableau 8.3
Comparaison des solutions du problème 2.2

B_k	$p(B_k)$			
	N_2	N_∞	SS	JS
0 0 1 1 0 1 0 1 1 1 0 0 1 0 1 0	0,2	0,2		
0 0 1 1 * 1 1 0 0 0 0 1 1 1 1 0 0	0,2	0,2	0,4	0,2
0 1 1 0 * 1 0 0 1 0 0 1 1 1 1 0 0	0,2	0,2		
0 1 1 0 * 1 0 1 0 1 0 0 1 0 1 0 1	0,4	0,4	0,2	0,4
0 1 1 0 0 0 1 1 1 1 0 0 1 0 0 1			0,2	
0 1 1 0 1 0 0 1 1 0 0 1 0 1 1 0			0,2	
0 1 1 0 1 0 0 1 0 1 0 1 1 0 1 0				0,2
0 0 1 1 0 1 0 1 1 0 1 0 1 1 0 0				0,2
Total	1,0	1,0	1,0	1,0
Total †	0,8	0,8	0,6	0,6

Voir la note du tableau 8.2.

Le problème 2.3, avec 141 tableaux possibles, est considérablement plus grand que les deux problèmes susmentionnés. Les solutions de ce problème sous les cinq méthodes sont comparées au tableau 8.4. Dans le tableau, les méthodes de Causey et coll. (1985) et de Huang et Lin (1998) sont désignées par CA et HU, respectivement. Les solutions pour CA et HU dans le tableau sont tirées de Causey et coll. (1985, p. 906) et de Huang et Lin (1998, figure 3), respectivement.

Tableau 8.4
Comparaison des solutions du problème 2.3

B_k	$p(B_k)$					B_k	$p(B_k)$					B_k	$p(B_k)$				
	N_2	N_∞	SS	CA	HU		N_2	N_∞	SS	CA	HU		N_2	N_∞	SS	CA	HU
0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 1 0 0 0 1 0 0 0	0,2	0,2	0,2			0 2 0 1 0 1 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1			0,11			0 2 0 1 0 1 0 0 0 2 0 0 1 0 0 0 1 0 0 0 1 0 0 1				0,2	
0 2 0 1 0 1 1 0 0 1 0 1 1 0 1 0 0 0 0 1 0 0 0 0	0,1	0,2	0,03			0 2 0 1 0 1 1 0 0 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0			0,03			0 2 0 1 0 1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 0 0				0,2	0,2
0 2 0 1 0 1 1 0 0 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0	0,1					0 2 0 1 0 1 1 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 1			0,03			0 2 0 2 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 0				0,2	
0 2 0 2 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 0 0 1	0,1					0 2 0 2 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 0			0,09			0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 0 0 0 1 0 0 0 1					0,2
0 2 0 2 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 1	0,1					0 2 0 2 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0			0,08			0 2 0 2 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0					0,2
1 2 0* 1 0 1 0 0 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 0	0,1		0,08			0 2 0 2 0 1 0 0 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 0			0,03								
1 2 0* 1 0 1 0 0 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 0	0,3	0,4	0,2	0,4	0,4	1 2 0 1 0 1 0 0 0 1 0 1 1 0 0 0 0 0 0 1 0 0 0 1			0,06								
0 2 0 2 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 1		0,2				1 2 0 1 0 1 0 0 0 1 0 1 1 1 0 0 1 0 0 0 0 0 0 0			0,06			Total Total †	1,0 0,4	1,0 0,4	1,0 0,28	1,0 0,4	1,0 0,4

Voir la note du tableau 8.2.

Nous notons que toutes ces méthodes fournissent des solutions différentes et qu'il y a chevauchement d'environ la moitié des tableaux entre les nouvelles méthodes et la méthode de Sitter et Skinner (1994). En outre, les solutions produites par les méthodes de Causey et coll. (1985) et de Huang et Lin (1998) sont assez

différentes de la solution de la méthode utilisant d_∞^*. La méthode utilisant d_2^* et la méthode de Sitter et Skinner répartissent les probabilités de sélection entre deux tableaux optimaux, tandis que les trois autres méthodes n'attribuent la probabilité qu'à un seul tableau optimal. La méthode de Sitter et Skinner semble être moins efficace pour sélectionner des tableaux optimaux, puisqu'elle donne la probabilité de 0,28 à ces derniers, tandis que les autres méthodes donnent une probabilité plus élevée, soit 0,4.

Les solutions du problème 2.4, qui est le plus grand des problèmes donnés, sont comparées sous quatre méthodes (N_2, N_∞, SS et la méthode de Winkler, 2001). Deux tableaux seulement, y compris un tableau optimal, sont les mêmes dans les solutions, et les deux nouvelles méthodes donnent les mêmes probabilités (0,127 et 0,483) à ces deux tableaux. Même si l'on compare la méthode utilisant d_∞^* aux méthodes de Sitter et Skinner (1994) et de Winkler (2001), les solutions de ces auteurs sont très différentes. En outre, les nouvelles méthodes donnent la même probabilité de sélection de 0,483 au tableau optimal, tandis que les méthodes antérieures donnent les probabilités plus faibles de 0,385 et 0,104, respectivement.

En résumé, il semble que les nouvelles méthodes réussissent à atteindre les spécifications S1 et S2 des solutions optimales. Notons que les nouvelles méthodes produisent systématiquement des probabilités de sélection plus élevées pour les tableaux optimaux et que les totaux de ces probabilités sont toujours les mêmes. Les solutions issues des nouvelles méthodes sont très différentes de celles obtenues en utilisant les méthodes antérieures lorsque les problèmes de sélection contrôlée ne sont pas petits. Cela implique que les solutions découlant des méthodes antérieures sont peut-être loin d'être optimales sous les critères S1 et S2 (E1 et E2).

9 Conclusion

Dans le présent article, nous avons présenté le concept des solutions optimales d'un problème de sélection contrôlée avec stratification à deux dimensions, et proposé un nouvel algorithme pour trouver ces solutions. L'algorithme a été implémenté facilement et avec succès dans le nouveau logiciel basé sur SAS (SOCSLP).

Puisqu'une solution optimale est une idée générale, elle pourrait être adoptée comme l'un des critères utiles pour comparer les différents algorithmes. Comme le montrent les comparaisons présentées plus haut, le nouvel algorithme donne des solutions aux grands problèmes de sélection contrôlée qui sont très différentes de celles obtenues en utilisant les méthodes publiées antérieures. Il est également susceptible de donner des probabilités de sélection plus élevées pour les tableaux optimaux que celles obtenues au moyen des méthodes antérieures.

Au vu des résultats obtenus pour les problèmes de sélection contrôlée à deux dimensions, nous nous attendons à ce que la méthode proposée contribue aussi à l'amélioration des propriétés des solutions des problèmes de sélection contrôlée avec stratification à trois dimensions ou plus.

Remerciements

Le présent article est dédié à I. Hess qui a consacré sa vie à l'étude de la sélection contrôlée. Les auteurs tiennent à remercier Jea-Bok Ryu de l'Université Chongju de leur avoir donné des idées et des conseils à un

premier stade de la présente étude. Nous remercions aussi deux examinateurs anonymes, le rédacteur et le rédacteur associé de leurs suggestions et commentaires précieux.

Bibliographie

Bryant, E.C., Hartley, H.O. et Jessen, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.

Causey, B.D., Cox, L.H. et Ernst, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

Chernick, M.R. et Wright, T. (1983). Estimation of a population mean with two-way stratification using a systematic allocation scheme. *Journal of Statistical Planning and Inference*, 7, 219-231.

Cox, L.H. et Ernst, L.R. (1982). Controlled rounding. *INFOR: Information Systems and Operational Research*, 20, 423-432.

Dantzig, G.B. (1963). *Linear Programming and Extensions*. Princeton University Press, Princeton, New Jersey.

Deville, J-C et Tillé, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91, 893-912.

Dykstra, R. (1985a). An iterative procedure for obtaining I-projections onto the intersection of convex sets. *Annals of Probability*, 13, 975-984.

Dykstra, R. (1985b). Computational aspects of I-projections. *Journal of Statistical Computation and Simulation*, 21, 265-274.

Goodman, R. et Kish, L. (1950). Controlled selection – a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Groves, R.M. et Hess, I. (1975). An algorithm for controlled selection. Dans *Probability Sampling of Hospitals and Patients*, *Second Edition*, (Eds., I. Hess, D.C. Ridel et T.B. Fitzpatrick), Health Administration Press, University of Michigan, Ann Arbor, USA.

Heeringa, S.G. (1998). PCCONSEL user guide. Dans *Controlled Selection Continued*, *2002 Edition*, (Eds., I. Hess et S.G. Heeringa), Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, USA.

Heeringa, S.G. et Hess, I. (1983). More on controlled selection. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 106-111.

Hess, I. et Heeringa, S.G. (2002). *Controlled Selection Continued.* Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, USA.

Hess, I., Ridel, D.C. et Fitzpatrick, T.B. (1975). *Probability Sampling of Hospitals and Patients*, *Second Edition*. Health Administration Press, University of Michigan, Ann Arbor, USA.

Huang, H.C. et Lin, T.K. (1998). On the two-dimensional controlled selection problem. Dans *Controlled Selection Continued*, *2002 Edition*, (Eds., I. Hess et S.G. Heeringa), Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, USA.

Jessen, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*, 65, 776-796.

Jessen, R.J. (1978). *Statistical Survey Techniques*. New York: John Wiley and Sons.

Lin, T.K. (1992). Some improvements on an algorithm for controlled selection. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 407-410.

Raghunandanan, K. et Bryant, E.C. (1971). Variance in multi-way stratification. *Sankhyā, Series A*, 33, 221-226.

Rao, J.N.K. et Nigam, A.K. (1990). Optimal controlled sampling designs. *Biometrika*, 77, 807-814.

Rao, J.N.K. et Nigam, A.K. (1992). "Optimal" controlled sampling: a unified approach. *International Statistical Review*, 60, 89-98.

Ross, S.M. (2007). *Introduction to Probability Models*. Burlington, MA: Academic Press.

SAS/OR (2008). *User's Guide: Mathematical Programming*. Version 9.2, Cary, NC: SAS Institute Inc.

Sitter, R.R. et Skinner, C.J. (1994). Stratification multidimensionnelle par programmation linéaire. *Techniques d'enquête*, 20 (1), 69-78.

Thie, P.R. et Keough, G.E. (2008). *An Introduction to Linear Programming and Game Theory*, *Third Edition*. Hoboken, New Jersey: John Wiley and Sons.

Tiwari, N. et Nigam, A.K. (1998). On two-dimensional optimal controlled selection. *Journal of Statistical Planning and Inference*, 69, 89-100.

Winkler, W.E. (1990). On Dykstra's iterative fitting procedure. *Annals of Probability*, 18, 1410-1415.

Winkler, W.E. (2001). Multi-way survey stratification and sampling. U.S. Census Bureau, *Statistical Research Division Report RRS 2001/01*. Disponible à : //www.census.gov/srd/www/byyear.html.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 293-312
Statistique Canada, n° 12-001-X au catalogue

Estimations composites harmonisées issues d'échantillons chevauchants pour les taux de croissance et les totaux

Paul Knottnerus[1]

Résumé

Lorsque les enquêtes mensuelles auprès des entreprises ne sont pas entièrement chevauchantes, il existe deux estimateurs différents du taux de croissance mensuelle du chiffre d'affaires, i) l'un fondé sur les totaux de population estimés mensuellement et ii) l'autre fondé purement sur les entreprises observées aux deux occasions dans la partie chevauchante des enquêtes correspondantes. Les estimations et les variances résultantes pourraient être assez différentes. Le présent article a pour but de proposer un estimateur composite optimal du taux de croissance, ainsi que des totaux de population.

Mots-clés : Enquêtes-entreprises; coefficient de variation; estimateur sous contraintes général; équations de Kalman; panels; variances.

1 Introduction

De nombreux pays réalisent mensuellement une enquête auprès des entreprises classées sous les principaux codes de la Classification type des industries (CTI) afin d'estimer le niveau du chiffre d'affaires mensuel et la variation de ce niveau comparativement au mois précédent ou au même mois un an plus tôt. Lorsqu'on échantillonne répétitivement une population, l'existence de diverses méthodes pour estimer la variation (relative) dans un panel donnant différents résultats est un facteur qui complique la situation, surtout si les échantillons sélectionnés à différentes occasions ne sont pas entièrement chevauchants.

Kish (1965), Tam (1984), Laniel (1987), Hidiroglou, Särndal et Binder (1995), Nordberg (2000), Berger (2004), Qualité et Tillé (2008), Wood (2008), ainsi que Knottnerus et Van Delden (2012) ont examiné divers estimateurs de la variation d'un paramètre dans différentes situations. L'objectif principal de la présente étude est d'obtenir des estimateurs de la variation relative d'un paramètre ainsi que des totaux de population correspondants qui sont en harmonie l'un avec l'autre et dont la variance est minimale. Le calcul des estimateurs composites harmonisés est fondé sur l'estimateur sous contraintes général, ou estimateur GR (pour *general restriction estimator*) de Knottnerus (2003). Särndal, Swensson et Wretman (1992, pages 370-378) proposent aussi des estimateurs composites des totaux et des variations (absolues), mais par étapes distinctes. En outre, la présente étude est axée sur des estimateurs des taux de croissance, car i) les utilisateurs des chiffres fournis par les enquêtes-enteprises pour un code CTI particulier s'intéressent souvent plus au taux de croissance qu'aux variations absolues, ii) en pratique, il pourrait exister des raisons liées aux techniques assistées par modèle d'examiner les taux de croissance (dans les modèles de régression, les variables auxiliaires expliquent souvent les différents taux de croissance plutôt que les différents niveaux des unités) et iii) les taux de croissance sont nécessaires pour construire un indice global du chiffre d'affaires (mensuel) pour chacun des principaux codes CTI. Par exemple, Smith, Pont et Jones (2003) décrivent la méthode des paires appariées pour mesurer une

1. Paul Knottnerus, Statistics Netherlands, CP 24500, 2490 HA La Haie, Pays-Bas. Courriel : pkts@cbs.nl.

variation d'un mois à l'autre, en se servant des réponses qui sont communes aux deux périodes. Ces auteurs appliquent la méthode pour calculer l'indice mensuel des ventes au détail (RSI pour *monthly retail sales index*).

La présentation de l'article est la suivante. La section 2 décrit brièvement deux méthodes d'estimation du taux de croissance du chiffre d'affaires total pour les entreprises possédant un code CTI donné. Deux exemples illustrent les différences parfois importantes entre les deux approches. La section 3 examine la question de savoir quelle méthode d'estimation doit être privilégiée et explique pourquoi l'écart entre les variances des deux estimateurs peut être si grand. La section 4 et la section 5 proposent un estimateur composite optimal pour diverses situations. La section 6 traite de certaines extensions de l'estimateur composite harmonisé ou estimateur AC (pour *aligned composite estimator*) des taux de croissance et des totaux. Enfin, la section 7 résume les principales conclusions et les questions nécessitant une étude plus poussée.

2 Deux estimateurs du taux de croissance du chiffre d'affaires total

Considérons une population de N entreprises $U = \{1, ..., N\}$, et supposons qu'il n'y a aucune création ni aucune disparition d'entreprise dans la population. Soit Y_i la valeur du chiffre d'affaires de la i^e entreprise durant un mois donné (disons t) et X_i la valeur du chiffre d'affaires de cette entreprise durant le mois $t-12$. Donc, les variables y et x concernent la même variable à deux occasions différentes. Désignons leurs totaux de population par Y et X, et leurs moyennes de population par $\bar{Y}$ et $\bar{X}$, respectivement. C'est-à-dire que $Y = \sum_{i \in U} Y_i$, $X = \sum_{i \in U} X_i$, $\bar{Y} = Y/N$ et $\bar{X} = X/N$. Soit s_1, s_2 et s_3 trois échantillons aléatoires simples mutuellement disjoints tirés de U sans remise. Définissons s_{12} et s_{23} par $s_{12} = s_1 \cup s_2$ et $s_{23} = s_2 \cup s_3$, respectivement. Désignons la taille de s_k par n_k $(k = 1, 2, 3, 12, 23)$ et les moyennes d'échantillon correspondantes par $\bar{y}_k$ et $\bar{x}_k$. Soit la variable x observée dans s_{12} à la première occasion et la variable y observée dans s_{23} à la deuxième occasion. Désignons les ratios de chevauchement par λ $(= n_2/n_{12})$ et μ $(= n_2/n_{23})$. Les estimateurs sous échantillonnage aléatoire simple (EAS) des totaux de population Y et X sont définis par $\hat{Y}_{EAS} = N\bar{y}_{23}$ et $\hat{X}_{EAS} = N\bar{x}_{12}$, respectivement.

Définissons le taux de croissance g du chiffre d'affaires total entre les deux occasions par $g = G - 1$ avec $G = Y/X$. Deux options existent pour estimer G. L'une des options standard (STN) est fondée sur les totaux estimés aux deux occasions, c'est-à-dire

$$\hat{G}_{STN} = \frac{\hat{Y}_{EAS}}{\hat{X}_{EAS}} = \frac{\bar{y}_{23}}{\bar{x}_{12}}; \tag{2.1}$$

voir Nordberg (2000), Qualité et Tillé (2008), et Knottnerus et Van Delden (2012). Soulignons que l'estimateur $\hat{g}_{STN} = \hat{G}_{STN} - 1$ pour g a la même variance que $\hat{G}_{STN}$. Pour une valeur de n suffisamment grande, cette variance peut être estimée approximativement par un développement en série de Taylor d'ordre 1 de $\hat{G}_{STN}$. C'est-à-dire

$$\begin{aligned}\operatorname{var}\left(\hat{G}_{STN}\right) &\approx \frac{1}{\bar{X}^2}\operatorname{var}\left(\bar{y}_{23}-G\bar{x}_{12}\right)\\ &= \frac{1}{\bar{X}^2}\left\{\operatorname{var}\left(\bar{y}_{23}\right)+G^2\operatorname{var}\left(\bar{x}_{12}\right)-2G\operatorname{cov}\left(\bar{y}_{23},\bar{x}_{12}\right)\right\}\\ &= \frac{1}{\bar{X}^2}\left\{\left(\frac{1}{n_{23}}-\frac{1}{N}\right)S_y^2+G^2\left(\frac{1}{n_{12}}-\frac{1}{N}\right)S_x^2-2G\left(\frac{\lambda\mu}{n_2}-\frac{1}{N}\right)S_{xy}\right\},\end{aligned} \tag{2.2}$$

où $S_y^2=\sum_U\left(Y_i-\bar{Y}\right)^2\Big/\left(N-1\right)$ est la variance de population ajustée des Y_i et S_x^2, celle des X_i tandis que $S_{xy}=\sum_U\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)\Big/\left(N-1\right)$ est leur covariance de population ajustée. Cochran (1977, page 153) propose une *règle de travail* afin d'utiliser le résultat en grand échantillon si la taille de l'échantillon dépasse 30 et que les coefficients de variation du numérateur et du dénominateur sont inférieurs à 10 %. Pour les (différentes) élaborations de l'expression du terme $\operatorname{cov}\left(\bar{y}_{23},\bar{x}_{12}\right)$ utilisé dans (2.2), voir Tam (1984) et Knottnerus et Van Delden (2012). Les (co)variances de population ajustées peuvent être estimées sans biais par les (co)variances d'échantillon; rappelons que les (co)variances d'échantillon s_{yk}^2 et s_{yxk} pour l'échantillon s_k $(k=1,2,3,12,23)$ sont définies comme étant

$$s_{yk}^2=\frac{1}{n_k-1}\sum_{i\in s_k}\left(Y_i-\bar{y}_k\right)^2$$

$$s_{yxk}=\frac{1}{n_k-1}\sum_{i\in s_k}\left(Y_i-\bar{y}_k\right)\left(X_i-\bar{x}_k\right).$$

Une autre option pour estimer G et g est fondée sur les entreprises observées aux deux occasions dans la partie chevauchante (OLP pour *overlap*) s_2. C'est-à-dire

$$\hat{G}_{OLP}=\frac{\bar{y}_2}{\bar{x}_2} \tag{2.3}$$

Pour une valeur suffisamment grande de n_2, l'approximation bien connue de la variance de cet estimateur est

$$\begin{aligned}\operatorname{var}\left(\hat{G}_{OLP}\right) &\approx \frac{1}{\bar{X}^2}\operatorname{var}\left(\bar{y}_2-G\bar{x}_2\right)\\ &= \frac{1}{\bar{X}^2}\left(\frac{1}{n_2}-\frac{1}{N}\right)S_{y-Gx}^2,\end{aligned} \tag{2.4}$$

où S_{y-Gx}^2 représente $S_y^2+G^2S_x^2-2GS_{xy}$; voir Cochran (1977, page 31). Afin de mieux percevoir les mérites des deux estimateurs $\hat{g}_{STN}$ et $\hat{g}_{OLP}$, considérons les exemples qui suivent.

Exemple 2.1. Les données utilisées dans cet exemple sont des observations de panel sur le chiffre d'affaires, en février 2011 et 2012, des supermarchés néerlandais appartenant à la strate 3 (catégorie de taille 3). La taille de la strate est $N=386$. En outre, $n_1=15$, $n_2=57$ et $n_3=17$. Pour les différents échantillons, nous avons (en milliers d'euros)

$$\bar{y}_{23}=97{,}2;\ \bar{x}_{12}=89{,}8;\ s_{y23}^2=3\ 781 \text{ et } s_{x12}^2=2\ 232.$$

Le coefficient de corrélation entre les Y_i et les X_i dans la population $\rho_{xy}\left(=S_{xy}/S_xS_y\right)$ est estimé d'après la partie chevauchante s_2 par $\hat{\rho}_{xy2}=s_{xy2}/s_{y2}s_{x2}=0{,}876.$ Afin d'éviter des estimations négatives de variance, Knottnerus et Van Delden (2012) proposent d'estimer le terme S_{xy} dans (2.2) par $\hat{S}_{xy}=\hat{\rho}_{xy2}s_{x12}s_{y23}=2\ 545.$ En introduisant les résultats susmentionnés par substitution dans (2.1) et (2.2), nous obtenons $\hat{g}_{STN}=0{,}082\ \left(=8{,}2\%\right)$ et $\text{vâr}\left(\hat{g}_{STN}\right)=0{,}00324.$ En émettant l'hypothèse de normalité et en utilisant $u_{0{,}975}=1{,}96,$ l'intervalle de confiance à 95 % est approximativement $I_{STN}^{95}\approx\left(-3{,}0\ \%;19{,}4\ \%\right).$ Par contre, pour la partie chevauchante s_2 nous obtenons les estimations

$$\bar{y}_2=102{,}2;\ \bar{x}_2=97{,}3\ \text{ et }\ \hat{g}_{OLP}=0{,}050\ \ (=5{,}0\ \%).$$

La substitution des mêmes estimations qu'auparavant à $\bar{X}$ et aux (co)variances des X_i et Y_i dans (2.4) donne $\text{vâr}\left(\hat{g}_{OLP}\right)=0{,}00166.$ Sous l'hypothèse de normalité, cela donne un plus petit intervalle de confiance à 95 % $I_{OLP}^{95}\approx\left(-3{,}0\ \%;13{,}0\ \%\right).$

Exemple 2.2. Dans les données de l'exemple 2.1, trois entreprises présentaient des valeurs de g extrêmes de -50 %, 133 % et -91 %. L'analyse plus poussée ou la correction de ces valeurs aberrantes dépassent le cadre du présent article. Cependant, pour illustrer une fois de plus la différence entre les estimateurs $\hat{g}_{STN}$ et $\hat{g}_{OLP,}$ nous omettons simplement ces entreprises, de sorte que $n_2=54$ au lieu de $n_2=57.$ Un premier résultat est que l'estimation $\hat{\rho}_{xy2}$ passe de 0,876 à 0,970. Cette dernière valeur est assez élevée en dépit du fait que le coefficient de variation des taux de croissance $g_i=\left(Y_i/X_i-1\right)$ est $cv_{g2}=s_{g2}/\bar{g}_2=4{,}1,$ ce qui témoigne de la persistance d'une assez forte volatilité des taux de croissance dans cet exemple. De surcroît, par analogie avec l'exemple précédent, nous obtenons $\hat{g}_{STN}=0{,}074\ \left(=7{,}4\ \%\right)$ avec $\text{var}\left(\hat{g}_{STN}\right)=0{,}00251$ et $\hat{g}_{OLP}=0{,}039\ \left(=3{,}9\ \%\right)$ avec $\text{var}\left(\hat{g}_{OLP}\right)=0{,}00039.$ Les intervalles de confiance à 95 % correspondants dans cet exemple légèrement modifié sont approximativement $I_{STN}^{95}\approx\left(\text{-}2{,}4\ \%;17{,}2\ \%\right)$ et $I_{OLP}^{95}\approx\left(0{,}1\ \%;7{,}7\ \%\right).$ Comparativement à l'exemple 2.1, la diminution de l'intervalle I_{OLP}^{95} a été relativement plus prononcée que celle de $I_{STN}^{95}.$

En outre, l'exemple 2.2 pourrait servir d'avertissement qu'il faut être prudent lorsqu'on utilise des moyennes d'échantillon telles que $\bar{y}_{23}$ et $\bar{x}_{12}$ pour estimer les taux de croissance, parce que ces estimations peuvent mener à un intervalle de confiance inutilement grand autour d'une estimation sous-optimale. À la section suivante, nous examinons de plus près la question de savoir quelles conditions peuvent donner lieu à un grand intervalle $I_{STN}^{95}.$

3 Raisons de l'obtention d'un grand intervalle I_{STN}^{95}

Afin de mieux comprendre la différence entre $\text{var}\left(\hat{g}_{OLP}\right)$ et $\text{var}\left(\hat{g}_{STN}\right),$ nous supposons que $n_{12}=n_{23}=n$ et $G,\ S_{xy}>0;$ d'où, $\lambda=\mu=n_2\ /\ n.$ Alors, en soustrayant (2.4) de (2.2), nous obtenons

$$\begin{aligned}\operatorname{var}\left(\hat{g}_{STN}\right)-\operatorname{var}\left(\hat{g}_{OLP}\right) &\approx \frac{1}{\bar{X}^2}\left\{2G\left(\frac{1}{n_2}-\frac{\lambda}{n}\right)S_{xy}-\left(\frac{1}{n_2}-\frac{1}{n}\right)\left(S_y^2+G^2S_x^2\right)\right\}\\ &=\frac{1}{\lambda n\bar{X}^2}\left\{2G\left(1-\lambda^2\right)S_{xy}-\left(1-\lambda\right)\left(S_y^2+G^2S_x^2\right)\right\}\\ &=\frac{1-\lambda}{\lambda n\bar{X}^2}\left(2G\lambda S_{xy}-S_{y-Gx}^2\right). \qquad (3.1)\end{aligned}$$

Autrement dit, $\operatorname{var}\left(\hat{g}_{OLP}\right)$ est plus petite que $\operatorname{var}\left(\hat{g}_{STN}\right)$ quand $\lambda > S_{y-Gx}^2 \big/ 2GS_{xy}$ à condition que $S_{xy} > 0$. En supposant que $S_y^2 = S_x^2$, Qualité et Tillé (2008) obtiennent un résultat similaire pour la variation *absolue* quand $\lambda > \left(1-\rho_{xy}\right)\big/\rho_{xy}$. Un examinateur anonyme a fait remarquer que $\lambda < \left(1-\rho_{xy}\right)\big/\rho_{xy}$ est une condition suffisante pour que $\operatorname{var}\left(\hat{g}_{OLP}\right) > \operatorname{var}\left(\hat{g}_{STN}\right)$ parce que (3.1) peut se réécrire sous la forme

$$\frac{\left(1-\lambda\right)GS_xS_y}{\lambda n\bar{X}^2}\left(2\lambda\rho_{xy}+2\rho_{xy}-\frac{S_y^2+G^2S_x^2}{GS_xS_y}\right)\le\frac{\left(1-\lambda\right)GS_xS_y}{\lambda n\bar{X}^2}\left(2\lambda\rho_{xy}+2\rho_{xy}-2\right)<0,$$

à condition que $\lambda < \left(1-\rho_{xy}\right)\big/\rho_{xy}$.

Si N est suffisamment grand, une contrainte plus faible peut être établie sous certaines hypothèses de modèle classiques. Supposons que les données satisfont le modèle $Y_i = BX_i + u_i$ avec $E(u_i)=0$, $E(u_i^2)=\sigma^2 X_i^{\delta}$ et $E(u_iu_j)=0$ $(i \ne j)$; rappelons que X_i n'est pas aléatoire dans ce contexte. Sous ce modèle, nous formulons les hypothèses (faibles) i) $G = S_{yx}\big/S_x^2$ et ii) $S_{y-Gx}^2 = S_y^2\left(1-\rho_{xy}^2\right)$. Pour justifier ces hypothèses, rappelons que, selon la théorie de la régression, $\hat{B} = S_{yx}\big/S_x^2$ peut être considéré comme l'estimateur sans biais, convergent de B issu d'une régression par les moindres carrés ordinaires (MCO) de Y_i sur X_i et une *constante* $(i=1,\ldots,N)$. En outre, l'estimateur MCO correspondant $\left(\bar{Y}-\hat{B}\bar{X}\right)$ pour la *constante* a une espérance nulle sous le modèle susmentionné, tandis que sa variance est d'ordre $1/N$. Donc, $0 = \operatorname{plim}\left(\bar{Y}-\hat{B}\bar{X}\right) = \operatorname{plim}\left\{\bar{X}\left(G-\hat{B}\right)\right\}$ quand $N \to \infty$ et, à condition que $\bar{X} > c > 0$ pour tout N, nous obtenons le résultat quelque peu contre-intuitif $\operatorname{plim}\left(G-\hat{B}\right)=0$. En fait, on peut montrer que

$$G = \bar{Y}/\bar{X} = \hat{B}\left[1+O_p\left(1/\sqrt{N}\right)\right] = \left(S_{yx}/S_x^2\right)\left[1+O_p\left(1/\sqrt{N}\right)\right]$$

quand $N \to \infty$. Cela justifie l'hypothèse (i); pour des explications plus détaillées, voir la fin de la présente section. De plus, $S_y^2\left(1-\rho_{xy}^2\right)$ peut être considérée comme la variance (non expliquée) des résidus de la régression par les MCO. Cependant, sous les hypothèses du modèle susmentionnées, les résidus sont asymptotiquement égaux à $Y_i - GX_i$, d'où découle la validité *approximative* de (ii). En outre, en notant que $S_y^2\rho_{xy}^2$ est la variance dite *expliquée* de la régression par les MCO susmentionnée, il découle de l'hypothèse (i) que $S_y^2\rho_{xy}^2 = \hat{B}^2S_x^2 \approx G^2S_x^2$. En combinant ce résultat avec les hypothèses (i) et (ii), nous pouvons réécrire (3.1) sous la forme

$$\begin{aligned}\operatorname{var}\left(\hat{g}_{STN}\right)-\operatorname{var}\left(\hat{g}_{OLP}\right) &\approx \frac{1-\lambda}{\lambda n\bar{X}^2}\left\{2G^2\lambda S_x^2-\left(1-\rho_{xy}^2\right)S_y^2\right\} \\ &\approx \frac{(1-\lambda)S_y^2}{\lambda n\bar{X}^2}\left(2\lambda\rho_{xy}^2-1+\rho_{xy}^2\right) \\ &= \frac{(1-\lambda)S_y^2}{\lambda n\bar{X}^2}\left\{\rho_{xy}^2\left(1+2\lambda\right)-1\right\}.\end{aligned} \tag{3.2}$$

D'où, $\operatorname{var}\left(\hat{g}_{OLP}\right)$ est plus grande que $\operatorname{var}\left(\hat{g}_{STN}\right)$ quand

$$\lambda<\left(1-\rho_{xy}^2\right)/2\rho_{xy}^2 \quad \left[>\left(1-\rho_{xy}\right)/\rho_{xy}\right]. \tag{3.3}$$

Donc pour, disons $\rho_{xy}=0{,}9$, $\operatorname{var}\left(\hat{g}_{OLP}\right)$ est, sous le modèle susmentionné pour une valeur suffisamment grande de N, plus grande que $\operatorname{var}\left(\hat{g}_{STN}\right)$ quand $\lambda<0{,}117$, et pour, disons $\rho_{xy}=0{,}75$, il en est ainsi quand $\lambda<0{,}389$. En outre, l'application de (3.2) aux données de l'exemple 2.1 avec $\lambda\approx 57/73=0{,}78$ et $\rho_{xy}=0{,}876$ donne comme différence approximative entre les deux variances la valeur de 0,0017, qui n'est pas très différente de la différence réelle de 0,0016 (=0,00324-0,00166) dans l'exemple. Pour l'exemple 2.2, si l'on prend $\lambda=54/70=0{,}77$ et $\rho_{xy}=0{,}970$, et que l'on applique (3.2), on obtient 0,00226 au lieu de 0,00212 (=0,00251-0,00039) dans l'exemple.

Sous les hypothèses susmentionnées, on peut également montrer que le ratio, disons Q, de $\operatorname{var}\left(\hat{g}_{OLP}\right)$ et $\operatorname{var}\left(\hat{g}_{STN}\right)$ peut être donné approximativement par

$$Q=\frac{\operatorname{var}\left(\hat{g}_{OLP}\right)}{\operatorname{var}\left(\hat{g}_{STN}\right)}\approx\left(\lambda^{-1}-f\right)\left(1-f+2(1-\lambda)\frac{\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}, \tag{3.4}$$

peu importe les valeurs de S_y^2 et S_x^2; f représente n/N. Pour une preuve de (3.4), voir l'annexe A.1. Partant de (3.4), on peut voir que Q et $\operatorname{var}\left(\hat{g}_{OLP}\right)$ tendent vers zéro quand ρ_{xy}^2 tend vers l'unité, à condition que N soit suffisamment grand et que $\lambda<1$.

Il convient de souligner qu'en pratique, les corrélations ρ_{xy} sont souvent assez élevées en raison de la nature même des données (Y_i, X_i). Autrement dit, une grande (petite) entreprise durant la période $(t-12)$ est, dans la plupart des cas, encore une grande (petite) entreprise après 12 mois; Knottnerus et Van Delden (2012, page 47) ont trouvé pour diverses strates une corrélation moyenne globale de 0,90 et une variance de 0,0074. Il semble donc que $\operatorname{var}\left(\hat{g}_{STN}\right)$ est plus affectée par une diminution de λ que $\operatorname{var}\left(\hat{g}_{OLP}\right)$, à moins que λ soit extrêmement faible parce que i) $\operatorname{var}\left(\hat{g}_{OLP}\right)=\operatorname{var}\left(\hat{g}_{STN}\right)$ quand $\lambda=1$ et ii) Q est grand quand ρ_{xy}^2 est grand. Par exemple, quand $\rho_{xy}=0{,}9$ et f = 0,1, une diminution de λ de 0,9 à 0,5 entraîne une diminution de Q qui passe de 0,58 à 0,37; rappelons que $Q=1$ quand $\lambda=1$. Cela souligne une fois de plus qu'il importe d'éviter l'érosion du panel lorsqu'on utilise l'estimateur $\hat{g}_{STN}$ quand N est grand.

Une question naturelle qu'il reste à poser est celle de savoir quand N est suffisamment grand. Pour y répondre, considérons la différence $\Delta\equiv\hat{B}-G$ et sa variance, disons σ_Δ^2. La différence Δ peut s'écrire

$$\begin{aligned}\Delta &= \frac{S_{xy}}{S_x^2} - \frac{\overline{Y}}{\overline{X}} = \frac{1}{N-1}\sum_{i\in U}\frac{X_i - \overline{X}}{S_x^2}Y_i - \frac{1}{N}\sum_{i\in U}\frac{Y_i}{\overline{X}} \\ &\approx \frac{1}{N}\sum_{i\in U}\left(\frac{X_i - \overline{X}}{S_x^2} - \frac{1}{\overline{X}}\right)Y_i \\ &= \frac{1}{N}\sum_{i\in U} M_i U_i \quad \left(M_i = \frac{X_i - \overline{X}}{S_x^2} - \frac{1}{\overline{X}}\right).\end{aligned}$$

À la deuxième ligne, nous supposons que $N >> 1$ et, à la dernière ligne, nous utilisons l'hypothèse du modèle $Y_i = BX_i + U_i$. Puis, en supposant que $\operatorname{var}(U_i) = \sigma^2 X_i^\delta$, nous obtenons

$$\sigma_\Delta^2 \equiv \operatorname{var}\left(\hat{B} - G\right) = \frac{\sigma^2}{N^2}\sum_{i\in U} M_i^2 X_i^\delta .$$

Cette variance peut être estimée par

$$\hat{\sigma}_\Delta^2 = \frac{\hat{\sigma}^2}{Nn_2}\sum_{i\in s_2}\hat{m}_i^2 X_i^{\hat{\delta}},$$

où

$$\hat{m}_i = \frac{X_i - \overline{x}_2}{s_{x2}^2} - \frac{1}{\overline{x}_2}, \qquad \hat{\sigma}^2 = \frac{1}{n_2 - 1}\sum_{i\in s_2}\left(Y_i - \frac{\overline{y}_2}{\overline{x}_2}X_i\right)^2 \Big/ X_i^{\hat{\delta}}$$

et $\hat{\delta}$ est une estimation tirée de la régression par les MCO

$$\ln\left(Y_i - \frac{\overline{y}_2}{\overline{x}_2}X_i\right)^2 = \alpha + \delta \ln X_i + w_i \qquad (i = 1,\ldots,n_2);$$

les unités pour lesquelles $Y_i = \overline{y}_2 X_i / \overline{x}_2$ sont omises. En se basant sur $\hat{\sigma}_\Delta^2$, on peut dire que N est suffisamment grand si le résultat de (3.1) n'est pas gravement affecté par le remplacement de G par $G + \hat{\sigma}_\Delta$. En outre, il ne faut pas perdre de vue que les relations dégagées pour une très grande valeur de N demeurent vraisemblablement une indication raisonnablement appropriée de ce qui pourrait se passer quand N n'est pas très grand.

4 Estimateur composite du taux de croissance

Partant d'un estimateur composite (COM) de la forme

$$\hat{g}_{COM} = k\hat{g}_{STN} + (1-k)\hat{g}_{OLP}, \tag{4.1}$$

il découle de la minimisation de $\operatorname{var}(\hat{g}_{COM})$ par rapport à k que

$$k = \frac{\operatorname{var}(\hat{g}_{OLP}) - \operatorname{cov}(\hat{g}_{OLP}, \hat{g}_{STN})}{\operatorname{var}(\hat{g}_{OLP}) + \operatorname{var}(\hat{g}_{STN}) - 2\operatorname{cov}(\hat{g}_{OLP}, \hat{g}_{STN})}; \tag{4.2}$$

voir aussi Särndal et coll. (1992, page 372). Soulignons que, par construction, $\operatorname{var}(\hat{g}_{COM})$ ne peut pas excéder $\min\{\operatorname{var}(\hat{g}_{STN}), \operatorname{var}(\hat{g}_{OLP})\}$.

En utilisant les formes linéarisées des estimateurs $\hat{g}_{OLP}$ et $\hat{g}_{STN}$, nous obtenons l'expression de leur covariance

$$\begin{aligned}\operatorname{cov}\left(\hat{g}_{OLP},\hat{g}_{STN}\right) &\approx \operatorname{cov}\left(\frac{\bar{y}_2 - G\bar{x}_2}{\bar{X}},\frac{\bar{y}_{23} - G\bar{x}_{12}}{\bar{X}}\right)\\ &= \frac{1}{\bar{X}^2}\left\{\operatorname{cov}\left(\bar{y}_2,\bar{y}_{23}\right) - G\operatorname{cov}\left(\bar{y}_2,\bar{x}_{12}\right) - G\operatorname{cov}\left(\bar{x}_2,\bar{y}_{23}\right) + G^2\operatorname{cov}\left(\bar{x}_2,\bar{x}_{12}\right)\right\}.\end{aligned}$$

En nous servant maintenant de certains résultats tirés de Knottnerus (2003, page 377)

$$\operatorname{cov}\left(\bar{y}_2,\bar{y}_{23}\right) = \operatorname{var}\left(\bar{y}_{23}\right) \quad \left[=\left(\frac{1}{n_{23}} - \frac{1}{N}\right)S_y^2\right]$$

$$\operatorname{cov}\left(\bar{x}_2,\bar{y}_{23}\right) = \operatorname{cov}\left(\bar{x}_{23},\bar{y}_{23}\right) \quad \left[=\left(\frac{1}{n_{23}} - \frac{1}{N}\right)S_{xy}\right],$$

nous obtenons

$$\operatorname{cov}\left(\hat{g}_{OLP},\hat{g}_{STN}\right) \approx \frac{1}{\bar{X}^2}\left\{\left(\frac{1}{n_{23}} - \frac{1}{N}\right)\left(S_y^2 - GS_{yx}\right) + \left(\frac{1}{n_{12}} - \frac{1}{N}\right)\left(G^2S_x^2 - GS_{yx}\right)\right\}. \tag{4.3}$$

En pratique, k peut être estimé en remplaçant toutes les (co)variances dans (4.2) par leurs estimations d'après l'échantillon, ce qui donne

$$\hat{k} = \frac{\widehat{\operatorname{var}}\left(\hat{g}_{OLP}\right) - \widehat{\operatorname{cov}}\left(\hat{g}_{OLP},\hat{g}_{STN}\right)}{\widehat{\operatorname{var}}\left(\hat{g}_{OLP}\right) + \widehat{\operatorname{var}}\left(\hat{g}_{STN}\right) - 2\widehat{\operatorname{cov}}\left(\hat{g}_{OLP},\hat{g}_{STN}\right)} \tag{4.4}$$

Pour illustrer cette approche, considérons l'exemple qui suit.

Exemple 4.1. Les données sont les mêmes que pour l'exemple 2.1. L'application des formules (2.1) à (2.4) et (4.3) à ces données produit

$$\begin{aligned}&\hat{g}_{STN} = 0{,}082\ (0{,}00254),\ \ \hat{g}_{OLP} = 0{,}050\ (0{,}00134),\ \text{ et}\\ &\widehat{\operatorname{cov}}\left(\hat{g}_{STN},\hat{g}_{OLP}\right) = 0{,}00097.\end{aligned}$$

Les variances sont mentionnées entre parenthèses. L'entrée de ces estimations dans (4.4) donne $\hat{k} = 0{,}191$ et, subséquemment, $\hat{g}_{COM} = 0{,}056\ (0{,}00127)$. Pour faciliter l'exposé, dans (4.4), toutes les (co)variances sont estimées d'après la partie chevauchante des échantillons s_2, y compris les estimations de G et de $\bar{X}$ dans (2.2), (2.4) et (4.3). En outre, en utilisant ces estimations, nous constatons que $\operatorname{var}\left(\hat{g}_{STN}\right) < \operatorname{var}\left(\hat{g}_{OLP}\right)$ et que $k > 0{,}5$ uniquement si $n_2 \le 12$ $\left(\lambda \le 0{,}167\right)$.

Par souci de complétude, nous donnons aussi un exemple pour l'estimateur composite du paramètre correspondant à la variation absolue (c.-à-d. $\bar{D} = \bar{Y} - \bar{X}$).

Exemple 4.2. Nous utilisons les mêmes données que dans l'exemple 2.1. Comme auparavant, toutes les estimations des (co)variances sont fondées sur s_2. Définissions D_i par l'expression $D_i = Y_i - X_i$. alors, nous avons deux estimateurs de la variation absolue,

$$\hat{\bar{D}}_{STN} = \bar{y}_{23} - \bar{x}_{12} = 7{,}35 \quad \text{et} \quad \hat{\bar{D}}_{OLP} = \bar{d}_2 = \bar{y}_2 - \bar{x}_2 = 4{,}89.$$

Pour les (co)variances de $\hat{\bar{D}}_{STN}$ et $\hat{\bar{D}}_{OLP}$, nous obtenons

$$\begin{aligned}
\hat{\text{var}}\left(\hat{\bar{D}}_{STN}\right) &= \hat{\text{var}}\left(\bar{y}_{23}\right) + \hat{\text{var}}\left(\bar{x}_{12}\right) - 2\,\hat{\text{cov}}\left(\bar{y}_{23}, \bar{x}_{12}\right) \\
&= \left(\frac{1}{n_{23}} - \frac{1}{N}\right) s_{y2}^2 + \left(\frac{1}{n_{12}} - \frac{1}{N}\right) s_{x2}^2 - 2\left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right) s_{xy2} = 23{,}58 \\
\hat{\text{var}}\left(\hat{\bar{D}}_{OLP}\right) &= \left(\frac{1}{n_2} - \frac{1}{N}\right) s_{y-x,2}^2 = 13{,}11
\end{aligned}$$

$$\begin{aligned}
\hat{\text{cov}}\left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right) &= \hat{\text{cov}}\left(\bar{y}_{23} - \bar{x}_{12}, \bar{y}_2 - \bar{x}_2\right) \\
&= \left(\frac{1}{n_{23}} - \frac{1}{N}\right)\left(s_{y2}^2 - s_{xy2}\right) - \left(\frac{1}{n_{12}} - \frac{1}{N}\right)\left(s_{xy2} - s_{x2}^2\right) = 9{,}46.
\end{aligned}$$

Par analogie avec (4.4), nous obtenons maintenant

$$\hat{k} = \frac{\hat{\text{var}}\left(\hat{\bar{D}}_{OLP}\right) - \hat{\text{cov}}\left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right)}{\hat{\text{var}}\left(\hat{\bar{D}}_{OLP}\right) + \hat{\text{var}}\left(\hat{\bar{D}}_{STN}\right) - 2\,\hat{\text{cov}}\left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right)} = 0{,}206$$

et, conséquemment, $\hat{\bar{D}}_{COM} = 5{,}40\ (12{,}37)$.

Notons que $\hat{g}_{COM}$ peut se réécrire

$$\begin{aligned}
\hat{g}_{COM} &= \hat{g}_{OLP} + \hat{k}\left(\hat{g}_{STN} - \hat{g}_{OLP}\right) \\
&\approx \hat{g}_{OLP} + k\left(\hat{g}_{STN} - \hat{g}_{OLP}\right),
\end{aligned}$$

où nous avons utilisé une approximation en série de Taylor d'ordre 1 de $\hat{g}_{COM}$. Par conséquent, nous pouvons faire abstraction du caractère aléatoire de l'estimateur $\hat{k}$ pour estimer $\text{var}\left(\hat{g}_{COM}\right)$. L'erreur ainsi introduite est d'ordre $1/n_2$ quand $n_2 \to \infty$ et $\hat{g}_{COM}$ est asymptotiquement sans biais. Rappelons que la procédure classique d'estimation de la variance de l'estimateur par le ratio ou de l'estimateur par la régression est fondée sur une approximation par développement en série de Taylor d'ordre 1 également.

En outre, sous les mêmes hypothèses que (3.4), on peut montrer que pour une valeur suffisamment grande de N,

$$k = \left(1 + \frac{2\lambda\rho_{xy}^2}{1 - \rho_{xy}^2}\right)^{-1}; \tag{4.5}$$

pour une preuve de (4.5), voir l'annexe A.1. Partant de (4.5), on peut constater que k est décroissant en λ. Nous obtenons donc le résultat quelque peu contre-intuitif selon lequel k est décroissant en λ tandis que, selon (3.1), le ratio Q dans (3.4) est une fonction convexe de λ; rappelons que $\text{var}\left(\hat{g}_{STN}\right) = \text{var}\left(\hat{g}_{OLP}\right)$ et, conséquemment, que $Q = 1$ pour $\lambda = 1$ et $\lambda = S_{y-Gx}^2 / 2GS_{xy}$.

5 Estimateurs composites harmonisés des taux de croissance et des totaux

Jusqu'à présent, nous n'avons examiné que les taux de croissance, parce qu'en pratique, l'estimation $\hat{X}_{EAS}$ du chiffre d'affaires 12 mois plus tôt peut être considérée comme plus ou moins fixe (c.-à-d. qu'elle ne peut plus varier). Quand X désigne le chiffre d'affaires total du mois $(t-1)$, il est probable que les chiffres du mois précédent peuvent encore être améliorés et modifiés. Dans ce cas, l'estimation initiale $\hat{X}_{EAS}$ peut être révisée également.

Avant d'examiner un estimateur composite multivarié des taux de croissance et des totaux, nous commençons par étudier un estimateur composite multivarié de la variation absolue et des moyennes ou des totaux de population correspondants; voir aussi l'exemple 4.2. Définissons l'estimateur vectoriel initial $\hat{\theta}_0$ par $\hat{\theta}_0 = \left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}, \bar{x}_{12}\right)'$. Désignons le vecteur de paramètres à estimer sous-jacent par $\theta = \left(\theta_1, \theta_2, \theta_3\right)'$. Soit V_0 la matrice de covariance de $\hat{\theta}_0$. En ce qui concerne θ, le problème est maintenant de trouver un estimateur composite harmonisé (AC) $\hat{\theta}_{AC}$ dont les éléments satisfont la contrainte a priori $\theta_1 - \theta_2 + \theta_3 = 0$ ou, de manière équivalente, $\bar{D} - \bar{Y} + \bar{X} = 0$ ou $D - Y + X = 0$. Bien qu'il n'y ait qu'une seule contrainte dans cette situation, nous traitons à la présente section le cas un peu plus général de m contraintes $\left(1 \le m \le 3\right)$. Lorsque les contraintes a priori sont de forme linéaire $c - R\theta = 0$, où R est une matrice de dimensions $m \times 3$ de rang m $(m \le 3)$, l'estimateur composite sans biais optimal de θ est égal à l'estimateur sous contraintes général (GR)

$$\hat{\theta}_{GR} = \hat{\theta}_0 + K\left(c - R\hat{\theta}_0\right) \tag{5.1}$$

$$K = V_0 R' \left(R V_0 R'\right)^{-1}$$

$$V_{GR} \equiv \operatorname{cov}\left(\hat{\theta}_{GR}\right) = \left(I_3 - KR\right) V_0, \tag{5.2}$$

où I_3 représente la matrice d'identité de dimensions 3×3. L'estimateur $\hat{\theta}_{GR}$ est optimal en ce sens que, si $\hat{\theta}_0$ suit une loi normale multivariée $N\left(\theta, V_0\right)$, la vraisemblance de $\hat{\theta}_0$ atteint sa valeur maximale, sous la contrainte $c - R\theta = 0$, pour $\theta_{\max} = \hat{\theta}_{GR}$. En outre, étant donné la forme $\hat{\theta}_K = \hat{\theta}_0 + K\left(c - R\hat{\theta}_0\right)$, on peut montrer que minimiser $\operatorname{tr}\left\{\operatorname{cov}\left(\hat{\theta}_K\right)\right\}$ par rapport à la matrice K de dimensions $3 \times m$ mène à (5.2). Rappelons que cela veut dire que, pour toute autre matrice K, la matrice de covariance correspondante $\operatorname{cov}\left(\hat{\theta}_K\right)$ excède V_{GR} d'une matrice semi-définie positive; voir Magnus and Neudecker (1988, pages 255-256). Pour des renseignements plus détaillés sur l'estimateur GR, voir Knottnerus (2003, pages 328-332). Afin d'illustrer comment (5.1) et (5.2) peuvent être utilisées pour obtenir un estimateur composite harmonisé $\hat{\theta}_{AC}$, considérons l'exemple qui suit qui concerne l'estimation de deux moyennes de population et de leur différence.

Exemple 5.1. Nous utilisons les mêmes données que dans les exemples 2.1 et 4.2. Le vecteur initial $\hat{\theta}_0 = \left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}, \bar{x}_{12}\right)'$ est donné par $\left(4{,}89; 97{,}19; 89{,}84\right)'$. Ces estimations ne satisfont pas la contrainte

$\theta_1 - \theta_2 + \theta_3 = 0;$ notons que $R = (1, -1, 1)$ et $c = 0.$ La plupart des éléments de V_0 ont déjà été discutés. Comme dans l'exemple 4.2, pour l'élément $\operatorname{cov}\left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}\right)$ nous obtenons

$$\begin{aligned}\operatorname{cov}\left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}\right) &= \operatorname{cov}\left(\bar{y}_2 - \bar{x}_2, \bar{y}_{23}\right) \\ &= \operatorname{var}\left(\bar{y}_{23}\right) - \operatorname{cov}\left(\bar{x}_{23}, \bar{y}_{23}\right).\end{aligned} \tag{5.3}$$

Chaque terme de (5.3) peut être estimé à partir de s_2 comme il est décrit plus haut. Les autres covariances figurant dans V_0 ont une forme similaire et peuvent être estimées de la même manière. Les estimations de la variance pour $\hat{\bar{D}}_{OLP}, \bar{y}_{23}$ et $\bar{x}_{12}$ sont 13,12; 38,79 et 22,92, respectivement. Ensuite, si nous appliquons (5.1) et (5.2) en remplaçant K par $\hat{K} = \hat{V}_0 R' \left(R \hat{V}_0 R'\right)^{-1}$, nous obtenons les estimations harmonisées composites AC suivantes

$$\hat{\bar{D}}_{AC} = 5{,}40\ (12{,}37), \quad \hat{\bar{Y}}_{AC} = 96{,}28\ (36{,}32), \quad \text{et} \quad \hat{\bar{X}}_{AC} = 90{,}88\ (19{,}75).$$

Les variances sont mentionnées entre parenthèses.

Ici, trois remarques sont de rigueur. Premièrement, $\hat{\bar{D}}_{COM}$ discuté à la section précédente peut également être calculé à partir de (5.1) et (5.2) en choisissant $\hat{\theta}_0 = \left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right)'$ sous la contrainte a priori $\theta_1 - \theta_2 = 0.$ Deuxièmement, par construction, l'estimateur $\hat{\bar{D}}_{AC}$ est égal à l'estimateur $\hat{\bar{D}}_{COM}$ et, par conséquent, ils ont la même variance. Troisièmement, si K était connue, l'estimateur AC serait sans biais. Mais, puisque K doit être remplacée par $\hat{K}$, l'estimateur AC $\hat{\theta}_{AC}$ est seulement asymptotiquement sans biais. La même remarque s'applique à l'estimateur $\left(I_3 - \hat{K}R\right)\hat{V}_0$ de $\operatorname{cov}(\hat{\theta}_{AC})$. Comme dans le cas de $\hat{\theta}_{COM}$ décrit à la section précédente, le biais de $\hat{\theta}_{AC}$ est d'ordre $O(1/n_2)$; pour la relation entre $\hat{\theta}_{AC}$ et l'estimateur par la régression, voir l'annexe A.3.

Dans le cas de m contrainte non linéaires, disons $c - R(\theta) = 0,$ une approximation en série de Taylor d'ordre 1 autour de $\theta = \hat{\theta}_0$ donne $c - R\left(\hat{\theta}_0\right) - D_R\left(\hat{\theta}_0\right)\left(\theta - \hat{\theta}_0\right) = 0$ ou, de manière équivalente,

$$c\left(\hat{\theta}_0\right) - D_R\left(\hat{\theta}_0\right)\theta = 0, \quad \text{où} \quad c\left(\hat{\theta}_0\right) = c - R\left(\hat{\theta}_0\right) + D_R\left(\hat{\theta}_0\right)\hat{\theta}_0. \tag{5.4}$$

$D_R(\theta)$ représente la matrice de dimensions $m \times 3$ des dérivées partielles de $R(\theta)$ $\left(\text{c.-à-d. } D_R(\theta) = \partial R(\theta)/\partial \theta'\right).$ Ensuite, une procédure itérative peut être exécutée en appliquant à plusieurs reprises (5.1) et (5.2) aux versions linéarisées mises à jour des contraintes non linéaires $c - R(\theta) = 0.$ Cela donne

$$\left.\begin{aligned}\hat{\theta}_h &= \hat{\theta}_0 + K_h \hat{e}_h; \\ \hat{e}_h &= c_h - D_h \hat{\theta}_0; \\ K_h &= V_0 D_h' \left(D_h V_0 D_h'\right)^{-1}; \\ \operatorname{cov}(\hat{\theta}_h) &= \left(I_3 - K_h D_h\right) V_0; \\ D_h &= D_R\left(\hat{\theta}_{h-1}\right); \\ c_h &= c - R\left(\hat{\theta}_{h-1}\right) + D_h \hat{\theta}_{h-1} \qquad (h = 1, 2, \ldots).\end{aligned}\right\} \tag{5.5}$$

Pour des renseignements plus détaillés, voir l'annexe A.2 et Knottnerus (2003, pages 351-354). Notons que la première équation peut être considérée comme une mise à jour de $\hat{\theta}_0$ plutôt que de $\hat{\theta}_{h-1}$. Il s'agit d'une différence importante par rapport aux célèbres équations de Kalman; voir Kalman (1960). Dans le présent contexte, les vecteurs $\hat{\theta}_{h-1}$ sont utilisés uniquement dans une procédure numérique en vue de trouver de nouvelles (meilleures) approximations en série de Taylor des contraintes non linéaires $c - R(\theta) = 0$ autour de $\theta = \hat{\theta}_{h-1}$ $(h = 1, 2, ...)$ jusqu'à ce que la convergence soit atteinte. En outre, notons que $\hat{e}_h$ peut être considéré comme un vecteur de dimension m des erreurs de contrainte lorsque l'on fait la substitution $\theta = \hat{\theta}_0$ dans les contraintes linéarisées autour de $\theta = \hat{\theta}_{h-1}$. Afin d'illustrer l'utilisation des équations de type Kalman dans (5.5) pour calculer les estimateurs composites harmonisés des taux de croissance et des totaux, considérons l'exemple suivant.

Exemple 5.2. Nous utilisons les mêmes données que dans l'exemple 4.1. Le vecteur initial $\hat{\theta}_0$ est maintenant défini par $\hat{\theta}_0 = \left(\hat{G}_{OLP}, \bar{y}_{23}, \bar{x}_{12}\right)'$ et est donné par $(1{,}050; 97{,}191; 89{,}840)'$. Ces estimations ne satisfont pas la contrainte a priori (non linéaire) $\theta_2 - \theta_1\theta_3 = 0$ $(m = 1)$. Tous les éléments de V_0 et leur estimation ont déjà été discutés. Pour la $(h+1)^{\text{e}}$ récursion, $R\left(\hat{\theta}_h\right)$ et la matrice D_{h+1} de dimensions 1×3 sont données par

$$R\left(\hat{\theta}_h\right) = \left(\hat{\theta}_{h2} - \hat{\theta}_{h1}\hat{\theta}_{h3}\right)$$
$$D_{h+1} = \left(-\hat{\theta}_{h3} \quad 1 \quad -\hat{\theta}_{h1}\right),$$

respectivement; $\hat{\theta}_{hk}$ est le k^{e} élément du vecteur $\hat{\theta}_h$ $(1 \le k \le 3)$. Rappelons que les termes V_0 et $\hat{V}_0$ restent invariants pendant toutes les récursions. La première récursion de (5.5) donne

$$\hat{\theta}_1 = (1{,}0544; 95{,}945; 91{,}000)'.$$

La contrainte (non linéaire) est presque satisfaite, c'est-à-dire que $R\left(\hat{\theta}_1\right) = -0{,}005$. La deuxième récursion donne les estimations composites harmonisées (AC) qui suivent.

$$\hat{G}_{AC} = 1{,}0544\ (0{,}00130), \quad \hat{\bar{Y}}_{AC} = 95{,}947\ (35{,}55), \quad \text{et} \quad \hat{\bar{X}}_{AC} = 90{,}998\ (19{,}85).$$

Les variances sont mentionnées entre parenthèses. L'erreur (absolue) de la deuxième contrainte ayant encore diminué, pour donner $R\left(\hat{\theta}_2\right) = -0{,}001$, nous avons arrêté les récursions. Étant donné la non-linéarité de la contrainte, les estimations de $\hat{G}_{AC}$ et de sa variance diffèrent légèrement de celles de $\hat{G}_{COM}$ et de sa variance dans l'exemple 4.1.

Il convient de souligner que, dans l'exemple 5.2, $\hat{G}_{AC}$ diffère peu de $\hat{G}_{OLP}$ (=1,050). Une méthode apparentée d'estimation des totaux est celle dite des paires appariées (PA); voir Smith et coll. (2003, pages 269-271). La méthode PA originale, qui est basée purement sur $\hat{G}_{OLP}$ (dans notre notation) entre les mois t et $t-1$, est utilisée par l'ONS pour estimer l'indice mensuel des ventes au détail. Dans une étude en simulation, les auteurs ont constaté que la méthode PA donne de bons résultats pour les taux de

croissance à court terme, mais que pour les périodes de plus de 15 mois, la performance devenait moins bonne en ce qui concerne le biais. Ce dernier pourrait être corrigé en procédant régulièrement à un calage sur les taux de croissance. Un autre inconvénient de la méthode PA semble être qu'il n'existe toujours pas de formule pour calculer la variance de l'estimateur PA. À la section suivante, nous décrivons une extension de l'estimateur AC en vue d'intégrer l'information auxiliaire dans la procédure d'estimation AC.

6 Extensions

À la présente section, nous discutons brièvement d'un certain nombre d'extensions de l'estimateur AC décrit à la section qui précède. Premièrement, nous examinons la situation où les estimateurs par la régression, disons $\hat{\bar{Y}}_{REG,k}$ et $\hat{\bar{X}}_{REG,k}$, sont utilisés au lieu des estimateurs EAS $(k = 2,\ 12 \text{ et } 23)$. Pour ne pas alourdir la notation, nous examinons le cas d'une seule variable explicative, disons z; la généralisation à un plus grand nombre de variables auxiliaires est facile. En outre, pour simplifier, nous supposons que les coefficients de régression estimés, désignés par b_{yz2} et b_{xz2}, sont issus de s_2. Afin de calculer les estimateurs composites harmonisés dans cette situation, il nous suffit d'évaluer les termes de (co)variance de la forme $\operatorname{cov}\left(\hat{\bar{Y}}_{REG,k}, \hat{\bar{X}}_{REG,l}\right)$ dans les différentes formules $(k,\ l = 2,\ 12 \text{ et } 23)$. Cette évaluation peut se faire comme il suit. Remplacer les Y_i et X_i dans les formules par les résidus (estimés) correspondants provenant d'une régression sur Z_i et une *constante*. C'est-à-dire

$$\operatorname{cov}\left(\hat{\bar{Y}}_{REG,k}, \hat{\bar{X}}_{REG,l}\right) = \operatorname{cov}\left(\bar{y}_k^*, \bar{x}_l^*\right), \tag{6.1}$$

où les variables résiduelles (estimées) Y_i^* et X_i^* sont définies par

$$Y_i^* = Y_i - \bar{y}_k - b_{yz2}\left(Z_i - \bar{z}_k\right) = Y_i - b_{yz2}Z_i + const.$$
$$X_i^* = X_i - \bar{x}_l - b_{xz2}\left(Z_i - \bar{z}_l\right) = X_i - b_{xz2}Z_i + const.$$

Le terme $\operatorname{cov}\left(\bar{y}_k^*, \bar{x}_l^*\right)$ dans le deuxième membre de (6.1) peut être calculé de la même manière que $\operatorname{cov}\left(\bar{y}_k, \bar{x}_l\right)$, dont nous avons discuté à la section précédente; voir aussi la formule (A.8) à l'annexe A.3 et rappelons que $\operatorname{var}\left(\hat{\bar{Y}}_{REG,k}\right) = \operatorname{cov}\left(\hat{\bar{Y}}_{REG,k}, \hat{\bar{Y}}_{REG,k}\right)$. En outre, la même approche peut s'appliquer lorsqu'on se sert des estimateurs par le ratio, tels que $\hat{\bar{Y}}_{R,k} = \bar{y}_k\bar{Z}/\bar{z}_k$ et $\hat{\bar{X}}_{R,l} = \bar{x}_l\bar{Z}/\bar{z}_l$. Dans ces conditions, les variables résiduelles Y_i^* et X_i^* doivent se lire

$$Y_i^* = Y_i - \frac{\bar{y}_2}{\bar{z}_2}Z_i \quad \text{et} \quad X_i^* = X_i - \frac{\bar{x}_2}{\bar{z}_2}Z_i.$$

Une autre option en vue de tenir compte d'une variable auxiliaire consiste à étendre le vecteur de paramètres θ et le jeu de contraintes a priori. Ainsi, dans l'exemple 5.1, le paramètre θ était défini

implicitement par $\theta = \left(\bar{D}, \bar{Y}, \bar{X}\right)'$. Lorsque la variable z est observée dans les échantillons 12 et 23, le nouvel estimateur étendu $\hat{\theta}_0$ est donné par

$$\hat{\theta}_0 = \left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}, \bar{x}_{12}, \bar{z}_{23}, \bar{z}_{12}, \bar{z}_2\right)'$$

et le jeu étendu de contraintes a priori est

$$\begin{aligned}
&\theta_2 - \theta_1 - \theta_3 = 0;\\
&\theta_4 - \theta_5 = 0;\\
&\theta_4 - \theta_6 = 0;\\
&\theta_4 = \bar{Z}.
\end{aligned}$$

Donc, la nouvelle contrainte c est $c = \left(0, 0, 0, \bar{Z}\right)'$. Cela permet d'améliorer encore davantage l'efficacité de $\hat{\theta}_0$.

Deuxièmement, une autre extension concerne les créations et les disparitions d'entreprises. Dans le cas des disparitions, la population de la période $t-12$ peut être divisée en deux (post)strates, l'une constituée des entreprises disparues à la période t et l'autre, des entreprises existantes aux périodes $t-12$ et t. Cette poststratification mène encore à un estimateur asymptotiquement sans biais pour la moyenne de population à la période t, à condition que de nouvelles entreprises ne soient pas créées. Afin de tenir compte des créations d'entreprises, il faut tirer un échantillon approprié de la strate des nouvelles entreprises, surtout si le nombre de ces dernières est important et s'il n'existe pas d'hypothèses réalistes quant au chiffre d'affaires total dans cette sous-strate durant le mois t.

Enfin, nous examinons la situation où une combinaison de données trimestrielles et semestrielles doit être analysée. Supposons que, durant les trimestres 2, 4 et 6, on tire des échantillons semestriels qui ne doivent pas nécessairement être les mêmes que les échantillons trimestriels obtenus pour ces trimestres. Afin d'expliquer l'estimateur AC dans ces conditions, considérons six estimations EAS trimestrielles consécutives pour les moyennes trimestrielles du chiffre d'affaires, disons $\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4, \bar{y}_5, \bar{y}_6$, et trois estimations EAS semestrielles pour les moyennes trimestrielles du chiffre d'affaires, disons $\bar{x}_2$, $\bar{x}_4$ et $\bar{x}_6$; notons que l'indice inférieur renvoie au trimestre d'observation et *non* à un jeu d'échantillons comme auparavant. En outre, supposons que l'on estime les ratios de croissance suivants : $G_{62} = Y_6/Y_2$, $H_{62} = X_6/X_2$ et $H_{64} = X_6/X_4$, ainsi que les totaux trimestriels et semestriels correspondants. Afin d'obtenir un jeu cohérent d'estimateurs pour les totaux (moyennes) et les taux de croissance, définissons par analogie avec l'approche adoptée à la section 5

$$\hat{\theta}_0 = \left(\hat{G}_{62,OLP}, \hat{H}_{62,OLP}, \hat{H}_{64,OLP}, \bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4, \bar{y}_5, \bar{y}_6, \bar{x}_2, \bar{x}_4, \bar{x}_6\right)'.$$

Le jeu correspondant de contraintes est

$$\begin{aligned}
\theta_9 - \theta_1\theta_5 &= \bar{Y}_6 - G_{62}\bar{Y}_2 = 0\\
\theta_{12} - \theta_2\theta_{10} &= \bar{X}_6 - H_{62}\bar{X}_2 = 0\\
\theta_{12} - \theta_3\theta_{11} &= \bar{X}_6 - H_{64}\bar{X}_4 = 0\\
\theta_4 + \theta_5 - \theta_{10} &= \bar{Y}_1 + \bar{Y}_2 - \bar{X}_2 = 0\\
\theta_6 + \theta_7 - \theta_{11} &= \bar{Y}_3 + \bar{Y}_4 - \bar{X}_4 = 0\\
\theta_8 + \theta_9 - \theta_{12} &= \bar{Y}_5 + \bar{Y}_6 - \bar{X}_6 = 0.
\end{aligned}$$

La matrice V_0 peut être estimée de la façon décrite aux sections 2 et 4.

7 Conclusion et discussion

À la présente section, nous résumons un certain nombre de conclusions et de questions qui requièrent une étude plus approfondie.

Lorsqu'on estime les totaux du chiffre d'affaires d'après des données de panel durant des mois tels que t et $t-12$, on peut distinguer deux estimateurs, $\hat{g}_{STN}$ et $\hat{g}_{OLP,}$, du taux de croissance entre les deux mois en question.

Si l'on utilise $\hat{g}_{STN}$, il faut être conscient qu'en pratique, $\text{var}(\hat{g}_{OLP})$ pourrait être beaucoup plus petite que $\text{var}(\hat{g}_{STN})$, surtout quand le chiffre d'affaires au mois $t-12$ et le chiffre d'affaires au mois t sont fortement corrélés et que les ratios de chevauchement λ et μ ne sont pas trop petits.

L'efficacité de $\hat{g}_{STN}$ et de $\hat{g}_{OLP}$ peut être améliorée en faisant appel à l'estimateur composite $\hat{g}_{COM}$ décrit à la section 4.

En appliquant les techniques des moindres carrés, on peut obtenir un estimateur vectoriel composite harmonisé $\left(\hat{g}_{AC}, \hat{Y}_{AC}, \hat{X}_{AC}\right)'$ qui satisfait à la contrainte non linéaire $\hat{Y}_{AC} = \left(1+\hat{g}_{AC}\right)\hat{X}_{AC}$ pour les totaux et les taux de croissance.

L'estimateur AC sous les contraintes *linéaires* peut être étendu de plusieurs façons : i) pour des contraintes non linéaires, ii) pour différents jeux de données, comme des données mensuelles, trimestrielles ou annuelles, iii) pour les créations et les disparitions d'entreprises, iv) pour les estimateurs par la régression et par le ratio, et v) pour des variables auxiliaires additionnelles.

Comme l'estimateur par la régression, l'estimateur AC est asymptotiquement sans biais. Cette remarque s'applique aussi à l'estimateur de la matrice de covariance $\left(I_k - \hat{K}R\right)\hat{V}_0$.

Il n'existe pas encore de réponse claire à la question de savoir dans quelle mesure les données recueillies dans le passé doivent être incluses chaque mois dans l'estimation $\hat{\theta}_0$ du vecteur. La réponse dépend i) de la politique et des règles de l'INS quant à la révision des chiffres déjà publiés, ii) du fait que, d'un point de vue théorique, la série de T estimations EAS mensuelles $\overline{y}_1, \overline{y}_2, ..., \overline{y}_T$ (incluse comme une composante dans $\hat{\theta}_0$) devrait avoir une longueur telle que la différence entre les deux estimateurs AC de $\overline{Y}_1$, disons $\hat{\overline{Y}}_{1AC}^{T}$ et $\hat{\overline{Y}}_{1AC}^{T+1}$, ne soit pas importante, et iii) de la taille des échantillons. Autrement dit, par analogie avec l'estimateur par la régression et, de manière équivalente, l'estimateur par calage, les tailles d'échantillon devraient être beaucoup plus grandes que le nombre de contraintes (de calage). Pour une étude par simulation de la variance de l'estimateur par la régression et du nombre de variables explicatives, voir Silva et Skinner (1997), et pour le lien entre l'estimateur par la régression et l'estimateur GR, voir l'annexe A.3 et Knottnerus (2003).

Dans le cas particulier de l'estimation de totaux et de variations mutuellement harmonisés, les travaux de recherche doivent se poursuivre afin de trouver i) la longueur optimale et pratique des séries mensuelles, trimestrielles, semestrielles et annuelles d'estimation EAS qu'il faut inclure dans le vecteur

$\hat{\theta}_0$ initial et ii) une règle donnant le nombre de contraintes relativement aux tailles d'échantillon, afin de trouver un estimateur AC $\hat{\theta}_{AC}$ plus efficace.

Remerciements

Les opinions exposées dans cet article sont celles de l'auteur et ne reflètent pas nécessairement la politique de Statistics Netherlands. L'auteur remercie Harm Jan Boonstra, Arnout van Delden, Sander Scholtus, le rédacteur associé et deux réviseurs anonymes pour leurs commentaires et corrections utiles.

Annexe

A.1 Preuves de (3.4) et (4.5)

La preuve de (3.4) est la suivante. Pour $n_{12} = n_{23} = n,$ la formule (2.2) peut se réécrire

$$\operatorname{var}\left(\hat{g}_{STN}\right) \approx \frac{1}{\bar{X}^2}\left\{\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-Gx}^2 + 2\left(\frac{1}{n}-\frac{\lambda}{n}\right)GS_{xy}\right\}. \tag{A.1}$$

La division de (2.4) par (A.1) donne

$$\begin{aligned} Q = \frac{\operatorname{var}\left(\hat{g}_{OLP}\right)}{\operatorname{var}\left(\hat{g}_{STN}\right)} &\approx \frac{\left(\frac{1}{\lambda n}-\frac{1}{N}\right)S_{y-Gx}^2}{\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-Gx}^2 + 2\left(\frac{1}{n}-\frac{\lambda}{n}\right)GS_{xy}} \\ &= \frac{\left(\lambda^{-1}-f\right)S_{y-Gx}^2}{\left(1-f\right)S_{y-Gx}^2 + 2\left(1-\lambda\right)GS_{xy}} \\ &= \left(\lambda^{-1}-f\right)\left(1-f+2\left(1-\lambda\right)\frac{GS_{xy}}{S_{y-Gx}^2}\right)^{-1} \\ &\approx \left(\lambda^{-1}-f\right)\left(1-f+2\left(1-\lambda\right)\frac{\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}. \end{aligned} \tag{A.2}$$

À la dernière ligne, nous nous sommes servi du fait que, sous les hypothèses du modèle mentionnées à la section 3, $GS_{xy} \approx \hat{B}^2 S_x^2 = \rho_{xy}^2 S_y^2$ et $S_{y-Gx}^2 \approx \left(1-\rho_{xy}^2\right)S_y^2,$ à condition que N soit suffisamment grand; voir aussi la façon d'obtenir (3.2).

Ensuite, sous les mêmes hypothèses, nous pouvons obtenir (4.5) comme il suit. Puisque $n_{12} = n_{23} = n,$ dans (4.3), la covariance peut être réécrite sous la forme

$$\operatorname{cov}\left(\hat{g}_{OLP}, \hat{g}_{STN}\right) \approx \frac{1}{\bar{X}^2}\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-Gx}^2. \tag{A.3}$$

En combinant (2.4), (A.1) et (A.3), dans (4.2), nous pouvons écrire k sous la forme

$$k \approx \frac{\left(\frac{1}{\lambda n}-\frac{1}{n}\right)S_{y-Gx}^2}{\left(\frac{1}{\lambda n}-\frac{1}{n}\right)S_{y-Gx}^2+2\left(\frac{1}{n}-\frac{\lambda}{n}\right)GS_{xy}}$$
$$=\left(1+\frac{2\lambda GS_{xy}}{S_{y-Gx}^2}\right)^{-1} \approx \left(1+\frac{2\lambda\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}.$$

Comme pour obtenir (A.2), nous avons utilisé à la dernière ligne le fait que $GS_{xy}\big/S_{y-Gx}^2 \approx \rho_{xy}^2\big/\left(1-\rho_{xy}^2\right)$.

A.2 Obtention de (5.5)

Dans le cas de m contraintes linéaires $c-R\theta=0$, on peut trouver la matrice K en minimisant

$$\min_K E\left[\left\{\theta-\hat{\theta}_0-K\left(c-R\hat{\theta}_0\right)\right\}'\left\{\theta-\hat{\theta}_0-K\left(c-R\hat{\theta}_0\right)\right\}\right];$$

voir Knottnerus (2003, page 330). La solution de ce problème de moindres carrés est donnée par

$$\begin{aligned} K &= E\left\{\left(\theta-\hat{\theta}_0\right)\left(c-R\hat{\theta}_0\right)'\right\}\left[\operatorname{cov}\left(c-R\hat{\theta}_0\right)\right]^{-1} \\ &= V_0R'\left(RV_0R'\right)^{-1}. \end{aligned} \tag{A.4}$$

Dans le cas de m contraintes non linéaires, la nouvelle expression à minimiser est

$$E\left[\left\{\theta-\hat{\theta}_0-K\left[c-R\left(\hat{\theta}_0\right)\right]\right\}'\left\{\theta-\hat{\theta}_0-K\left[c-R\left(\hat{\theta}_0\right)\right]\right\}\right].$$

Comme pour (A.4), on peut montrer que cette expression atteint son minimum pour

$$K = E\left\{\left(\theta-\hat{\theta}_0\right)\left[c-R\left(\hat{\theta}_0\right)\right]'\right\}\left[\operatorname{cov}\left\{c-R\left(\hat{\theta}_0\right)\right\}\right]^{-1}. \tag{A.5}$$

En introduisant la linéarisation de Taylor $R\left(\hat{\theta}_0\right)\approx R(\theta)+D_R(\theta)\left(\hat{\theta}_0-\theta\right)$ dans (A.5), nous obtenons l'approximation suivante, disons K_1, de K

$$\begin{aligned} K_1 &\approx V_0D_R'(\theta)\left[D_R(\theta)V_0D_R'(\theta)\right]^{-1} \\ &\approx V_0D_R'\left(\hat{\theta}_0\right)\left[D_R\left(\hat{\theta}_0\right)V_0D_R'\left(\hat{\theta}_0\right)\right]^{-1}. \end{aligned} \tag{A.6}$$

En supposant que $\hat{\theta}_0 \sim N\left(\theta,V_0\right)$, la première approximation de la solution du maximum de vraisemblance (MV) restreint, disons $\hat{\theta}_{MV}^{(1)}$, peut être calculée de la manière classique en utilisant les contraintes linéarisées

$$\hat{\theta}_{MV}^{(1)} = \hat{\theta}_0 + K_1\left\{c\left(\hat{\theta}_0\right)-D_R\left(\hat{\theta}_0\right)\hat{\theta}_0\right\}, \tag{A.7}$$

où $c\left(\hat{\theta}_0\right)$ est défini par (5.4). Si $\hat{\theta}_{MV}^{(1)}$ ne satisfait pas les contraintes non linéaires, $c-R(\theta)=0$, une meilleure approximation de K pourrait être obtenue en remplaçant $\hat{\theta}_0$ dans (A.6) par $\hat{\theta}_{MV}^{(1)}$ mis à jour qui

donne une nouvelle matrice K_2. À son tour, par analogie avec (A.7), K_2 donne une meilleure approximation ou mise à jour de $\hat{\theta}_0$, disons $\hat{\theta}_{MV}^{(2)}$,

$$\hat{\theta}_{MV}^{(2)} = \hat{\theta}_0 + K_2 \left\{ c\left(\hat{\theta}_{MV}^{(1)}\right) - D_R\left(\hat{\theta}_{MV}^{(1)}\right)\hat{\theta}_0 \right\},$$

où nous avons utilisé la linéarisation de Taylor des contraintes non linéaires autour de $\theta = \hat{\theta}_{MV}^{(1)}$. En répétant cette procédure, nous obtenons les récursions suivantes pour $\hat{\theta}_{MV}^{(h)}$ ou, en abrégé, $\hat{\theta}_h$

$$\hat{\theta}_h = \hat{\theta}_0 + K_h \left\{ c_h - D_h \hat{\theta}_0 \right\}$$

$$K_h = V_0 D_h' \left[D_h V_0 D_h' \right]^{-1} \qquad (h = 1, 2, \ldots).$$

Pour les définitions de c_h et D_h, voir la section 5; en pratique, V_0 doit être remplacée par son estimation $\hat{V}_0$. Par construction, pour chaque h, nous avons

$$\begin{aligned} 0 &= c\left(\hat{\theta}_{MV}^{(h-1)}\right) - D_R\left(\hat{\theta}_{MV}^{(h-1)}\right)\hat{\theta}_{MV}^{(h)} \\ &= c - R\left(\hat{\theta}_{MV}^{(h-1)}\right) + D_R\left(\hat{\theta}_{MV}^{(h-1)}\right)\hat{\theta}_{MV}^{(h-1)} - D_R\left(\hat{\theta}_{MV}^{(h-1)}\right)\hat{\theta}_{MV}^{(h)}; \end{aligned}$$

voir (5.4). D'où, quand $\hat{\theta}_{MV}^{(h)}$ converge vers la solution du maximum de vraisemblance (restreint) $\hat{\theta}_{MV}$, $c - R\left(\hat{\theta}_{MV}^{(h-1)}\right)$ converge vers zéro. En outre, en supposant que K_h converge vers, disons $\hat{K}_{MV}$, la matrice de covariance correspondante de $\hat{\theta}_{MV}$, disons V_{MV}, peut être approximée par

$$V_{MV} \approx \left\{ I_k - K D_R(\theta) \right\} V_0,$$

qui, si h est suffisamment grand, peut être estimée par $\hat{V}_{MV} = \left(I_k - K_h D_h\right)\hat{V}_0$; voir aussi Cramer (1986, page 38).

A.3 Estimateur par la régression comme estimateur GR

Supposons que Y_i et la variable auxiliaire Z_i, de moyenne de population $\overline{Z}$ connue, sont observées dans s_2. Afin d'appliquer l'estimateur GR à cette situation, définissons

$$\hat{\theta}_0 = \begin{pmatrix} \overline{y}_2 \\ \overline{z}_2 \end{pmatrix}, \quad V_0 = \operatorname{cov}\left(\hat{\theta}_0\right) = \left(\frac{1}{n_2} - \frac{1}{N}\right)\begin{pmatrix} S_y^2 & S_{yz} \\ S_{yz} & S_z^2 \end{pmatrix}.$$

La contrainte a priori est

$$0 = c - R\theta = \overline{Z} - (0,\ 1)\begin{pmatrix} \theta_y \\ \theta_z \end{pmatrix}.$$

L'application de (5.1) et (5.2) à ce cas donne l'estimateur GR suivant

$$\hat{\theta}_{GR} = \hat{\theta}_0 + K\left(c - R\hat{\theta}_0\right) = \begin{pmatrix} \overline{y}_2 \\ \overline{z}_2 \end{pmatrix} + K\left(\overline{Z} - \overline{z}_2\right)$$

$$K = V_0 R'\left(R V_0 R'\right)^{-1} = \begin{pmatrix} S_{yz} \\ S_z^2 \end{pmatrix}\frac{1}{S_z^2} = \begin{pmatrix} b_{yz} \\ 1 \end{pmatrix} \qquad \left(b_{yz} = S_{yz} / S_z^2\right)$$

$$V_{GR} = \left(I_2 - KR\right)V_0 = \begin{pmatrix} 1 & -b_{yz} \\ 0 & 0 \end{pmatrix} V_0.$$

Donc, en remplaçant b_{yz} par son estimation $b_{yz2} = s_{yz2} \big/ s_{z2}^2$, nous pouvons approximer le premier élément de $\hat{\theta}_{GR}$ par $\hat{\theta}_{GRy} \approx \bar{y}_2 + b_{yz2}\left(\bar{Z} - \bar{z}_2\right)$, qui correspond à l'estimateur par la régression bien connu, souvent désigné par $\hat{\bar{Y}}_{REG}$. Pour une valeur de n_2 suffisamment grande, la variance de $\hat{\bar{Y}}_{REG}$ peut être approximée par

$$\begin{aligned}
\operatorname{var}\left(\hat{\bar{Y}}_{REG}\right) &\approx \operatorname{var}\left(\hat{\theta}_{GRy}\right) = \left[V_{GR}\right]_{11} = \left(\frac{1}{n_2} - \frac{1}{N}\right)\left(S_y^2 - b_{yz}S_{yz}\right) \\
&= \left(\frac{1}{n_2} - \frac{1}{N}\right)S_e^2; \qquad \text{(A.8)} \\
S_e^2 &= \frac{1}{N-1}\sum_{i \in U}\left\{Y_i - \bar{Y} - b_{yz}\left(Z_i - \bar{Z}\right)\right\}^2;
\end{aligned}$$

rappelons que, selon la théorie de la régression, $b_{yz}S_{yz} = b_{yz}^2 S_z^2$ et $S_y^2 = b_{yz}^2 S_z^2 + S_e^2$. Dans (A.8), la variance peut être estimée par l'estimateur de variance bien connu

$$\hat{\operatorname{var}}\left(\hat{\bar{Y}}_{REG}\right) = \left(\frac{1}{n_2} - \frac{1}{N}\right)s_{\hat{e}2}^2, \quad \text{où} \quad s_{\hat{e}2}^2 = \frac{1}{n_2 - 1}\sum_{i \in s_2}\left\{Y_i - \bar{y}_2 - b_{yz2}\left(Z_i - \bar{z}_2\right)\right\}^2.$$

Des résultats semblables peuvent être obtenus pour plus d'une variable auxiliaire. Cela illustre de nouveau que, pour ce qui est du biais et de l'approximation de la variance, l'estimateur AC ressemble fortement à l'estimateur par la régression ou, de manière équivalente, l'estimateur par calage.

Bibliographie

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics*, 32, 451-467.

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

Cramer, J.S. (1986). *Econometric Applications of Maximum Likelihood Methods*. Cambridge: Cambridge University Press.

Hidiroglou, M.A., Särndal, C.E. et Binder, D.A. (1995). Weighting and estimation in business surveys. Dans le *Business Survey Methods*, (Eds., B.G. Cox et coll.). New York: John Wiley and Sons, Inc.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. Transactions ASME, *Journal of Basic Engineering*, 82, 35-45.

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

Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. New York: Springer-Verlag.

Knottnerus, P. et Van Delden, A. (2012). À propos de la variance des variations estimées d'après des panels rotatifs et des strates dynamiques. *Techniques d'enquête*, 38(1), 45-56.

Laniel, N. (1987). Variances for a rotating sample from a changing population. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 496-500.

Magnus, J.R. et Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: John Wiley and Sons, Inc.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

Qualité, L. et Tillé, Y. (2008). Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée. *Techniques d'enquête*, 34(2), 193-201.

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

Silva, P.L.D.N. et Skinner, C.J. (1997). Sélection des variables pour l'estimation par regression dans le cas des populations finies. *Techniques d'enquête*, 23(1), 25-35.

Smith, P., Pont, M. et Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994–2000. *Journal of the Royal Statistical Society D*, 52, 257-295.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, 288-289.

Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, 24, 53-78.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 313-351
Statistique Canada, n° 12-001-X au catalogue

L'estimation des flux bruts dans les enquêtes complexes avec non-réponse aléatoire

Andrés Gutiérrez, Leonardo Trujillo et Pedro Luis do Nascimento Silva[1]

Résumé

Les enquêtes par panel à renouvellement servent à calculer des estimations des flux bruts entre deux périodes consécutives de mesure. Le présent article examine une procédure générale pour estimer les flux bruts lorsque l'enquête par panel à renouvellement a été générée à partir d'un plan de sondage complexe avec non-réponse aléatoire. Une approche à pseudo-maximum de vraisemblance est envisagée dans le contexte d'un modèle à deux degrés de chaînes de Markov pour le classement des personnes dans les catégories de l'enquête et pour la modélisation de la non-réponse.

Mots-clés : Inférence fondée sur le plan; enquêtes par panel à renouvellement; flux bruts; chaînes de Markov.

1 Introduction

Des techniques d'enquête sont souvent utilisées pour estimer certains paramètres d'intérêt d'une population finie. L'inférence pour ces paramètres est basée sur la répartition de la probabilité induite par le plan d'échantillonnage utilisé pour obtenir l'échantillon de personnes. Dans la plupart des cas, pour les statistiques officielles, le plan d'échantillonnage envisagé est complexe en ce qu'il ne se base pas sur un échantillon aléatoire simple de la population.

Après avoir obtenu un échantillon probabiliste, il est parfois nécessaire d'examiner la classification des personnes dans l'échantillon dans différentes catégories d'une ou de plusieurs variables nominales. Cette classification peut être intégrée dans un tableau de contingence afin de résumer deux variables ou les variations temporelles d'une seule variable pendant deux périodes différentes. Toutefois, afin d'obtenir des estimations exactes, il n'est pas recommandé de faire fi du plan d'échantillonnage dans l'inférence pour les paramètres d'intérêt.

Un autre problème fréquent dans ce type d'enquête est la non-réponse pour certaines unités de l'échantillon, qui peut rarement être considérée comme aléatoire ou ignorable. Il est donc nécessaire d'envisager une approche qui pourrait neutraliser la non-réponse potentiellement non ignorable. Chen et Fienberg (1974), Stasny (1987) et récemment Lu et Lohr (2010) ont envisagé des modèles à deux degrés pour classer les personnes dans un échantillon pour deux périodes différentes avec non-réponse non ignorable. Cependant, cette approche faisait fi du plan d'échantillonnage, qui est complexe et informatif pour la plupart des enquêtes réalisées pour produire des statistiques officielles.

Cet article examine un scénario commun pour les enquêtes longitudinales, où le principal objectif consiste à estimer le nombre de personnes dans une population appartenant à plusieurs cellules dans un tableau de contingence en fonction des catégories d'une variable mesurée à deux moments donnés. Nous examinons également la modélisation de la non-réponse qui peut avoir une incidence sur les estimations si

1. Andrés Gutiérrez, Facultad de Estadística. Universidad Santo Tomás. Courriel : hugogutierrez@usantotomas.edu.co; Leonardo Trujillo, Department of Statistics, Universidad Nacional de Colombia. Courriel : ltrujilloo@unal.edu.co; Pedro Luis do Nascimento Silva, Instituto Brasileiro de Geografia e Estatística (IBGE). Courriel : pedro-luis.silva@ibge.gov.br.

elle n'est pas prise en compte. Les processus inférentiels sont liés au plan de sondage complexe utilisé pour recueillir l'information dans l'échantillon.

Par exemple, dans les enquêtes sur la population active, il est possible de trouver des classifications complexes en fonction de la situation d'activité des répondants pendant deux périodes consécutives d'observation et de mesure. L'objectif consiste à estimer le nombre de personnes qui pendant une période passée travaillaient et qui travaillent toujours pendant la période d'observation en cours. Un autre objectif possible est d'estimer le nombre de personnes qui étaient en chômage pendant la dernière période d'observation et qui le sont toujours pendant la période actuelle d'enquête, ou le nombre de personnes qui avaient un emploi au cours de la dernière période d'observation, et qui sont en chômage pendant la période en cours, ou vice versa. Pour cet exemple, toutes les entrées dans le tableau 1.1 sont considérées comme des paramètres d'intérêt. Soulignons que même dans le cadre d'un recensement, les chiffres du tableau 1.1 ne sont peut-être pas observables en raison de la non-réponse.

Tableau 1.1
Paramètres d'intérêt dans un tableau de contingence correspondant à une enquête sur la population active pendant deux périodes consécutives d'observation.

Période 1	**Période 2**			
	Avec emploi	**Chômeurs**	**Inactifs**	**Total**
Avec emploi	X_{11}	X_{12}	X_{13}	X_{1+}
Chômeurs	X_{21}	X_{22}	X_{23}	X_{2+}
Inactifs	X_{31}	X_{32}	X_{33}	X_{3+}
Total	X_{+1}	X_{+2}	X_{+3}	X_{++}

Kalton (2009) a déclaré que, pour ce qui est des totaux marginaux, il est possible d'estimer les flux nets au moyen d'une comparaison directe entre les deux périodes d'observation. Ensuite, il est possible de déterminer si le taux de chômage a crû ou diminué et dans quelle mesure. Par exemple, en faisant une comparaison avec la période 1, on constate qu'il y avait $X_{1+} = \sum_j X_{1j}$ personnes occupées, tandis que pendant la période 2, il y avait $X_{+1} = \sum_i X_{i1}$ personnes occupées. Néanmoins, on peut obtenir une analyse plus détaillée en analysant les flux bruts comme une décomposition des flux nets. Ainsi, si le taux de chômage a crû d'un point de pourcentage, il est possible de conclure si cette augmentation était attribuable au fait que 1 % des personnes occupées ont perdu leur emploi ou que 10 % des personnes occupées ont perdu leur emploi et 9 % des chômeurs ont trouvé un nouvel emploi. C'est possible lorsque l'on compare les valeurs X_{ij}.

De plus, étant donné que dans une enquête complexe, il est possible d'avoir des poids d'échantillonnage inégaux ainsi que la présence d'effets de grappe et de stratification, la fonction de vraisemblance des données de l'échantillonnage est difficile à trouver d'une manière analytique. Par conséquent, l'utilisation de la méthodologie classique du maximum de vraisemblance ne serait plus utile pour les données d'enquêtes complexes. Par ailleurs, les analyses standard doivent être modifiées afin de tenir compte des effets de la pondération et de l'effet de plan dû à l'enquête complexe, comme l'estimation pondérée des proportions, l'estimation de la variance fondée sur le plan de sondage et les corrections généralisées pour tenir compte des effets de plan de sondage (Pessoa et Silva 1998).

La section 2 examine les concepts statistiques de base utilisés dans cet article, comme les estimateurs d'enquête, la non-réponse et l'inférence de données catégoriques. La section 3 propose un modèle de superpopulation décrivant le comportement probabiliste du classement des personnes en fonction des catégories de la variable envisagée dans l'enquête. Il s'agit du modèle de chaîne de Markov à deux degrés. Certains concepts de base de l'estimation de la pseudo-vraisemblance sont également examinés à la section 3. Ensuite, à la section 4, nous proposons certains estimateurs pour les paramètres de modèle et les chiffres dans le tableau de contingence des flux bruts. Ces estimateurs sont sans biais par rapport au plan et les expressions mathématiques pour estimer leur variance apparaissent à la section 5. La section 6 considère une application empirique et une simulation Monte Carlo afin de mettre à l'essai la méthode proposée lorsque les données de l'enquête sont obtenues dans le contexte d'un plan de sondage simple et d'un complexe. Notre simulation démontre que d'autres approches méthodologiques donnent lieu à une estimation biaisée. La section 7 considère une application pratique pour les flux bruts de l'estimation pour l'enquête *Pesquisa Mensal de Emprego* (PME) au Brésil. À la section 8, nous décrivons les forces et les faiblesses de la méthode proposée. Toutes les preuves mathématiques sont présentées à l'annexe.

2 Motivation

2.1 Plans d'échantillonnage et estimateurs

Considérons une population finie comme un ensemble de N unités, où $N < \infty$, pour former l'univers de l'étude. N est connu comme la taille de la population. Chaque élément appartenant à la population peut être identifié par un indice k. Supposons que U soit l'ensemble des indices donné par $U = \{1,...,k,...,N\}$. La sélection d'un échantillon $s = \{k_1, k_2, \ldots, k_{n(s)}\}$ est effectuée selon un plan d'échantillonnage défini comme la répartition de la probabilité multivariée par rapport au soutien Q de manière à ce que $p(s) > 0$ pour chaque $s \in Q$ et

$$\sum_{s \in Q} p(s) = 1.$$

Selon un plan d'échantillonnage $p(\cdot)$, une probabilité d'inclusion est attribuée à chaque élément dans la population afin de désigner la probabilité que l'élément appartienne à l'échantillon. Pour l'élément k dans la population, cette probabilité est représentée par π_k et est connue comme la probabilité d'inclusion de premier ordre donnée par

$$\pi_k = Pr(k \in S) = Pr(I_k = 1) = \sum_{s \ni k} p(s)$$

où I_k est une variable aléatoire indiquant l'appartenance de l'élément k à l'échantillon, et le sous-indice $s \ni k$ indique la somme par rapport à tous les échantillons possibles renfermant l'élément k. De façon analogue, π_{kl} est appelé la probabilité d'inclusion de deuxième ordre et indique la probabilité que les éléments k et l appartiennent à l'échantillon et est donnée par

$$\pi_{kl} = Pr(k \in S; l \in S) = Pr(I_k = 1; I_l = 1) = \sum_{s \ni \mathrm{k,l}} p(s).$$

L'objectif de l'enquête-échantillon consiste à étudier une caractéristique d'intérêt y associée à chaque sous-section de la population et à estimer une fonction d'intérêt T, appelée paramètre :

$$T = f(y_1, \ldots, y_k, \ldots, y_N).$$

Cette approche inférentielle s'appelle l'inférence fondée sur le plan. Selon cette approche, les estimations des paramètres et de leurs propriétés dépendent directement de la mesure de la probabilité discrète liée au plan d'échantillonnage choisi et ne tiennent pas compte des propriétés de la population finie. De plus, les valeurs y_k sont prises comme l'observation pour la personne k pour la caractéristique d'intérêt y. De plus, y est considérée comme une quantité fixe au lieu d'une variable aléatoire.

Alors, l'estimateur Horvitz-Thompson (HT) peut être défini comme suit :

$$\hat{t}_{y,\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k$$

où $d_k = 1/\pi_k$ est la réciproque de la probabilité d'inclusion de premier ordre et s'appelle facteur d'expansion ou poids de base selon le plan. L'estimateur HT est sans biais pour la population totale $t_y = \sum_U y_k$, (en supposant que toutes les probabilités d'inclusion de premier ordre soient supérieures à zéro) et sa variance est donnée par :

$$Var\left(\hat{t}_{y,\pi}\right) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{2.1}$$

où $\Delta_{kl} = Cov\left(I_k, I_l\right) = \pi_{kl} - \pi_k \pi_l$. Si toutes les probabilités d'inclusion de deuxième ordre sont supérieures à zéro, un estimateur sans biais de (2.1) est donné par :

$$\widehat{Var}\left(\hat{t}_{y,\pi}\right) = \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Gambino et Silva (2009) suggèrent que dans une enquête-ménage, le principal intérêt est de se concentrer sur les caractéristiques pour des membres particuliers du ménage qui pourraient être liées aux variables sur la santé, aux variables sur la scolarité, au revenu et aux dépenses, à la situation d'emploi, etc. En général, les plans d'échantillonnage utilisés pour ce genre d'enquête sont complexes et utilisent des techniques comme la stratification, la mise en grappes ou les probabilités inégales de sélection. Certains des résultats d'enquêtes répétées tiennent compte de l'estimation du niveau à un moment particulier, de l'estimation des variations entre deux cycles d'enquête et de l'estimation des paramètres du niveau moyen au fil de plusieurs cycles répétés d'une enquête. Différents plans de rotation et la fréquence de l'enquête peuvent avoir une incidence considérable sur la précision des estimateurs.

2.2 Pseudo-vraisemblance

Certains auteurs, comme Fuller (2009), Chambers et Skinner (2003, p. 179), et Pessoa et Silva (1998, chapitre 5) considèrent le problème où l'estimation du maximum de vraisemblance est appropriée pour les échantillons aléatoires simples, comme c'est le cas dans Stasny (1987), mais pas pour les échantillons découlant d'un plan de sondage complexe. Selon cette classification, on suppose que la fonction de densité de la population est $f\left(y, \theta\right)$ où le paramètre d'intérêt est θ. Si l'on a accès à l'information pour la population au complet, au moyen d'un recensement, l'estimateur de maximum de vraisemblance de θ peut être obtenu en maximisant

$$L(\theta) = \sum_{k \in U} \log f(y_k, \theta)$$

par rapport à θ. Nous indiquerons θ_N comme la valeur maximisant la dernière expression. Les équations de vraisemblance pour la population sont données par

$$\sum_{k \in U} u_k(\theta) = 0.$$

Les valeurs u_k sont appelées *scores* et sont définies comme suit :

$$u_k(\theta) = \frac{\partial \log f(y_k, \theta)}{\partial \theta}.$$

L'approche de pseudo-vraisemblance considère que θ_N est le paramètre d'intérêt d'après l'information recueillie dans un échantillon complexe. Si $\sum_{k \in U} u_k(\theta)$ est considéré comme le paramètre d'intérêt, il est possible de l'estimer au moyen d'un estimateur linéaire pondéré

$$\sum_{k \in s} d_k u_k(\theta)$$

où d_k est un poids d'échantillonnage comme l'inverse de la probabilité d'inclusion de la personne k. Alors, il est possible d'obtenir un estimateur pour θ_N satisfaisant le système d'équations obtenu.

Définition 2.1 *Un estimateur du maximum de pseudo-vraisemblance $\hat{\theta}_s$ pour θ_N correspond à la solution des équations de la pseudo-vraisemblance données par*

$$\sum_{k \in s} d_k u_k(\theta) = 0.$$

Au moyen de la méthode de linéarisation de Taylor, la variance asymptotique d'un estimateur du maximum de pseudo-vraisemblance fondé sur le plan de sondage est donnée par

$$V_p\left(\hat{\theta}_s\right) \approx \left[J(\theta_N)\right]^{-1} V_p\left[\sum_{k \in s} d_k u_k(\theta_N)\right]\left[J(\theta_N)\right]^{-1}$$

où $V_p\left[\sum_{k \in s} d_k u_k(\theta_N)\right]$ est la variance de l'estimateur pour la population totale des *scores* fondé sur le plan de sondage et

$$J(\theta_N) = \left.\frac{\partial \sum_{k \in U} u_k(\theta)}{\partial \theta}\right|_{\theta = \theta_N}.$$

Un estimateur pour $V_p\left(\hat{\theta}_s\right)$ est donné par

$$\hat{V}_p\left(\hat{\theta}_s\right) = \left[\hat{J}\left(\hat{\theta}_s\right)\right]^{-1} \hat{V}_p\left[\sum_{k \in s} d_k u_k\left(\hat{\theta}_s\right)\right]\left[\hat{J}\left(\hat{\theta}_s\right)\right]^{-1}$$

où $\hat{V}_p\left[\sum_{k \in s} d_k u_k\left(\hat{\theta}_s\right)\right]$ est un estimateur constant de la variance de l'estimateur du total de la population des *scores* et

$$\hat{J}\left(\hat{\theta}_s\right) = \left.\frac{\partial \sum_{k \in s} d_k u_k(\theta)}{\partial \theta}\right|_{\theta = \hat{\theta}_s}.$$

Alors, d'après Binder (1983), la répartition asymptotique de $\hat{\theta}_s$ est normale puisque

$$\hat{V}_p\left(\hat{\theta}_s\right)^{-1/2}\left(\hat{\theta}_s - \theta_N\right) \sim N(0,1).$$

Ces définitions offrent des renseignements de base solides pour la bonne inférence lorsque l'on utilise de gros échantillons comme c'est le cas dans les enquêtes sur la population active.

2.3 Non-réponse

Särndal et Lundström (2005) soutiennent que la non-réponse est un sujet qui intéresse de plus en plus les bureaux de statistique nationaux ces dernières décennies. De plus, dans la documentation sur les enquêtes par sondage, l'attention accordée à ce sujet s'est beaucoup intensifiée. La non-réponse est un problème fréquent associé à l'élaboration d'une enquête qui peut effriter considérablement la qualité des estimations.

Lohr (2000) décrit les différents types de mécanismes de non-réponse :

- Le mécanisme de non-réponse est ignorable lorsque la probabilité qu'une personne réponde à l'enquête ne dépend pas de la caractéristique d'intérêt. Soulignons que le qualificatif « ignorable » s'applique au modèle expliquant le mécanisme.

- Par ailleurs, le mécanisme de non-réponse est non ignorable lorsque la probabilité qu'une personne réponde à l'enquête dépend de la caractéristique d'intérêt. Par exemple, dans une enquête sur le travail, la possibilité de réponse peut dépendre de la classification de la population active selon l'activité des membres d'un ménage.

Lumley (2009, chapitre 9) analyse la non-réponse individuelle au moyen de données partielles pour un répondant qui envisage une approche fondée sur le plan rajustant le poids d'échantillonnage. Fuller (2009, chapitre 5) considère certaines techniques d'imputation pour le traitement de la non-réponse au moyen de modèles probabilistes et de poids d'échantillonnage. Särndal (2011) considère une approche fondée sur les données au moyen d'ensembles équilibrés afin d'assurer une haute représentativité des estimations. De la même manière, Särndal et Lundström (2010) proposent un ensemble d'indicateurs afin d'évaluer l'efficacité de l'information auxiliaire afin de contrôler le biais généré par la non-réponse. Särndal et Lundström (2005) donnent un grand nombre de références au sujet de la non-réponse. Ces références examinent deux principaux aspects complémentaires d'une enquête : prévention du problème de non-réponse (avant qu'il survienne) et techniques d'estimation afin de tenir compte de la non-réponse dans le processus d'inférence. Ce deuxième aspect s'appelle correction de la non-réponse.

3 Modèles de Markov pour les tableaux de contingence avec non-réponse

Considérons le problème des flux bruts des estimations entre deux périodes consécutives au moyen de données catégoriques obtenues d'une enquête par panel et dans un contexte de non-réponse. De plus, supposons que le résultat de chaque interview est la classification du répondant dans une des G catégories disjointes en paires possibles, et le but est d'estimer les flux bruts entre ces catégories au moyen de l'information de personnes qui ont été interviewées pendant deux périodes consécutives. Les personnes qui n'ont pas répondu pendant une ou deux périodes ou qui étaient exclues ou incluses pour une seule des deux périodes n'auront pas une classification définie parmi les catégories. Par conséquent, il y a un groupe de personnes avec une classification entre les deux périodes, un groupe de personnes qui ont l'information uniquement pour une des deux périodes et un groupe de personnes qui n'ont participé à aucune des deux périodes de l'enquête.

Pour les personnes qui ont répondu aux périodes $t-1$ et t, les données sur la classification peuvent être résumées dans une matrice de dimension $G \times G$. L'information disponible pour les personnes qui n'ont pas répondu à l'enquête pendant la période $t-1$ mais qui ont répondu pendant la période t peut être résumée dans un complément de colonne; l'information pour les personnes n'ayant pas répondu pendant la période t mais ayant répondu pendant la période $t-1$ peuvent être résumées dans un complément de ligne. Enfin, les personnes n'ayant répondu à aucune des deux périodes sont incluses dans une seule cellule en dénombrant les personnes ayant des données manquantes pendant les deux périodes.

La matrice complète est illustrée dans le tableau 3.1, où N_{ij} $(i, j = 1, \ldots, G)$ désigne le nombre de personnes dans la population ayant une classification i pendant la période $t-1$ et une classification j pendant la période t, R_i indique le nombre de personnes n'ayant pas répondu pendant la période t et ayant une classification i pendant la période $t-1$, C_j indique le nombre de personnes n'ayant pas répondu pendant la période $t-1$ et ayant une classification j pendant la période t, et M représente le nombre de personnes dans l'échantillon n'ayant pas répondu pendant les deux périodes. Il est important de mentionner que cette analyse ne tient pas compte de la non-réponse due à la rotation de l'enquête; elle tient compte uniquement des personnes appartenant à l'échantillon apparié en faisant fi des personnes n'ayant pas répondu parce qu'elles n'avaient pas été sélectionnées dans l'échantillon.

Tableau 3.1
Flux bruts pendant deux périodes consécutives

Période $t-1$	**Période** t				
	1	**2**	$\cdots$	G	**Complément de ligne**
1	N_{11}	N_{12}	$\ldots$	N_{1G}	R_1
2	N_{21}	N_{22}	$\ldots$	N_{2G}	R_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	N_{G1}	N_{G2}	$\ldots$	N_{GG}	R_G
Complément de colonne	C_1	C_2	$\ldots$	C_G	M

Le présent article tient compte des idées de Stasny (1987) et de Chen et Fienberg (1974), en ce qu'il tient compte de l'approche du maximum de vraisemblance dans les tableaux de contingence pour les données partiellement classées, ainsi que les données obtenues d'un processus à deux degrés comme suit :

1. Au premier degré (non observable), les personnes se trouvent dans les cellules d'une matrice $G \times G$ en fonction des probabilités d'un processus de chaîne de Markov. Supposons que η_i soit la probabilité initiale qu'une personne soit dans la catégorie i pendant la période $t-1$, où $\sum_i \eta_i = 1$, et que p_{ij} soit la probabilité de transition de la catégorie i à la catégorie j, où $\sum_j p_{ij} = 1$ pour chaque i.

2. Au deuxième degré (observable) du processus, chaque personne dans le cas ij peut être un non-répondant pendant la période $t-1$, perdant la classification par ligne; un non-répondant pendant la période t, perdant la classification par colonne; ou un non-répondant pendant les deux périodes, perdant les deux classifications.
   - Supposons que ψ soit la probabilité initiale qu'une personne dans le cas ij réponde pendant la période $t-1$.
   - Supposons que ρ_{RR} soit la probabilité de transition de la classification que la personne du cas ij ait répondu pendant la période $t-1$ et la période t.
   - Supposons que ρ_{MM} soit la probabilité de transition qu'une personne dans le cas ij soit un non-répondant pendant la période $t-1$ et devienne un non-répondant pendant la période t.

   Ces probabilités ne dépendent pas de l'état de classification de la personne.

Les données sont observées uniquement après le deuxième degré. L'idée est de faire des inférences pour les probabilités dans le processus de la chaîne de Markov produisant les données, mais aussi dans la chaîne générant le mécanisme de non-réponse. Dans le contexte de ce modèle à deux degrés, les probabilités correspondantes sont indiquées dans le tableau 3.2.

Tableau 3.2
Probabilités des flux bruts pendant deux périodes consécutives

Période $t-1$				**Période** t			
	1	**2**	$\cdots$	j	$\cdots$	G	**Complément de ligne**
1							
2							
$\vdots$							
i				$\{\eta_i p_{ij} \psi \rho_{RR}\}$			$\left\{\sum_j \eta_i p_{ij} \psi (1-\rho_{RR})\right\}$
$\vdots$							
G							
Complément de colonne			$\left\{\sum_i \eta_i p_{ij} (1-\psi)(1-\rho_{MM})\right\}$				$\sum_i \eta_i p_{ij} \sum_j (1-\psi) \rho_{MM}$

Ainsi, la fonction de vraisemblance pour les données observées dans ce modèle à deux degrés est proportionnelle à

$$\prod_i \prod_j \left[\psi \rho_{RR} \eta_i p_{ij}\right]^{N_{ij}} \times \prod_i \left[\sum_j \psi (1-\rho_{RR}) \eta_i p_{ij}\right]^{R_i} \times \prod_j \left[\sum_i (1-\psi)(1-\rho_{MM}) \eta_i p_{ij}\right]^{C_j} \times \left[\sum_i \sum_j (1-\psi) \rho_{MM} \eta_i p_{ij}\right]^{M}. \tag{3.1}$$

3.1 Paramètres d'intérêt

Les données sont observées uniquement après le deuxième degré, et l'objectif est de faire des inférences pour les probabilités dans la chaîne de Markov générant les données et la chaîne générant la non-réponse. Selon ce modèle à deux degrés, les probabilités de la matrice de données sont indiquées dans le tableau 3.2 et constituent certains des paramètres d'intérêt.

Par ailleurs, après le processus non observable, il faut tenir compte d'autres paramètres d'intérêt, comme suit. Supposons qu'il y ait une population finie U, ayant une classification de deux périodes pour toutes ses personnes. Il s'agit d'un processus non observable puisque, même lorsque les données du recensement sont obtenues, il ne serait pas possible d'avoir une classification complète puisque les personnes ne seraient pas toutes disposées à répondre. Compte tenu de ce processus non observable et en supposant qu'il y ait G classifications possibles dans chaque période, la répartition des flux bruts au niveau de la population est indiquée dans le tableau 3.3.

X_{ij} est le nombre d'unités de la population finie ayant la classification i pendant la période $t-1$ et la classification j pendant la période t ($i, j = 1, \ldots, G$). La taille de la population, N, doit satisfaire à l'expression :

$$N = \sum_i \sum_j X_{ij}.$$

Tableau 3.3
Flux bruts de la population (processus non observable) pendant deux périodes consécutives

Période $t-1$	**Période** t					
	1	**2**	$\cdots$	j	$\cdots$	G
1	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1G}
2	X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2G}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{iG}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
G	X_{G1}	X_{G2}	$\cdots$	X_{Gj}	$\cdots$	X_{GG}

Suivant le processus non observable de la dernière section, on suppose que le vecteur correspondant aux entrées dans le dernier tableau de contingence suit une répartition multinomiale avec un vecteur de

probabilité renfermant les valeurs $\{\eta_i p_{ij}\}_{i,j=1,\ldots,G}$. On suppose un modèle de superpopulation, où les chiffres du tableau de contingence sont considérés comme aléatoires. En ce qui concerne la notation, la mesure de la probabilité compte tenu de ces chiffres sera représentée par le sous-indice ξ. Alors, la probabilité de classification de la k-ième personne dans la cellule i, j est

$$\begin{aligned}
&P_\xi\left(k \text{ a la classification } i \text{ à } t\text{-}1 \text{ et la classification } j \text{ à } t\right)\\
&= P_\xi\left(k \text{ a la classification } i \text{ à } t\text{-}1\right)\\
&\times P_\xi\left(k \text{ a la classification } j \text{ à } t \mid k \text{ a la classification } i \text{ à } t\text{-}1\right)\\
&= \eta_i p_{ij}.
\end{aligned}$$

On considère X_{ij} comme une variable aléatoire et si la population finie comporte N personnes, sa valeur prévue en fonction du modèle est donnée par

$$E_\xi\left(X_{ij}\right) = N\eta_i p_{ij} = \mu_{ij}. \tag{3.2}$$

Soulignons que cette valeur prévue μ_{ij} est un des paramètres les plus importants à estimer dans cet article, puisqu'elle correspond à la valeur prévue des flux bruts de la population d'intérêt pendant deux périodes consécutives. Par ailleurs, il faut comprendre que μ_{ij} est un paramètre pour le modèle à deux degrés. De plus, les estimateurs pour η_i et p_{ij} sont interdépendants et déterminés par les estimations des paramètres définis au deuxième degré. Supposons que $\boldsymbol{\eta}$ soit le vecteur renfermant les paramètres η_i et que $\mathbf{p}$ soit le vecteur renfermant les paramètres p_{ij}, pour chaque $i, j = 1,\ldots,G$. Les paramètres finaux d'intérêt sont les suivants :

- les paramètres du modèle, déterminés par le vecteur

$$\boldsymbol{\theta} = \left(\psi', \rho'_{RR}, \rho'_{MM}, \boldsymbol{\eta}', \mathbf{p}'\right)' \;;$$

- le vecteur des valeurs prévues des chiffres de population est défini comme suit :

$$\boldsymbol{\mu} = \left(\mu_{11},\ldots,\mu_{ij},\ldots,\mu_{GG}\right)'.$$

4 Estimation des paramètres d'intérêt

Supposons que N_{ij} représente le nombre total de répondants pour la population d'intérêt ayant une classification i pendant la période $t-1$ et j pendant la période t. Supposons que R_i soit le nombre total de personnes dans la population n'ayant pas répondu pendant la période t mais ayant répondu pendant la période $t-1$ avec la classification i. Supposons que C_j représente le nombre total de personnes dans la population n'ayant pas répondu pendant la période $t-1$ mais ayant répondu pendant la période t avec la classification j et enfin, supposons que M représente le nombre de personnes appartenant à la population n'ayant répondu à aucune des deux périodes d'observation. Il s'ensuit que la taille totale de la population, N, doit respecter la contrainte suivante :

$$N = \sum_i \sum_j N_{ij} + \sum_j C_j + \sum_i R_i + M.$$

En définissant les caractéristiques d'intérêt suivantes, il est possible de déterminer les paramètres d'intérêt :

$$y_{1ik} = \begin{cases} 1, & \text{si la } k-\text{ième personne répond à la période } t-1 \text{ avec la classification } i; \\ 0, & \text{sinon.} \end{cases}$$

$$y_{2jk} = \begin{cases} 1, & \text{si la } k-\text{ième personne répond à la période } t \text{ avec la classification } j; \\ 0, & \text{sinon.} \end{cases}$$

Alors, le produit de ces quantités, défini par $y_{1ik} y_{2jk}$, correspond à une nouvelle caractéristique d'intérêt prenant la valeur un si la personne a répondu aux deux périodes et est classé dans le cas ij, ou zéro sinon. De plus,

$$N_{ij} = \sum_{k \in U} y_{1ik} y_{2jk}.$$

Définissez les caractéristiques dichotomiques suivantes :

$$z_{1k} = \begin{cases} 1, & \text{si la } k-\text{ième personne répond à la période } t-1; \\ 0, & \text{sinon.} \end{cases}$$

$$z_{2k} = \begin{cases} 1, & \text{si la } k-\text{ième personne répond à la période } t; \\ 0, & \text{sinon.} \end{cases}$$

Il s'ensuit que

$$\begin{aligned} R_i &= \sum_{k \in U} y_{1ik} \left(1 - z_{2k}\right) \\ C_j &= \sum_{k \in U} y_{2jk} \left(1 - z_{1k}\right) \\ M &= \sum_{k \in U} \left(1 - z_{1k}\right)\left(1 - z_{2k}\right). \end{aligned}$$

Supposons que w_k indique le poids pour la k-ième personne correspondant à une stratégie d'échantillonnage particulière (plan d'échantillonnage et estimateur) dans les deux vagues. Par conséquent, les expressions suivantes représentent les estimateurs des paramètres d'intérêt :

$$\begin{aligned} \hat{N}_{ij} &= \sum_{k \in S} w_k y_{1ik} y_{2jk} \\ \hat{R}_i &= \sum_{k \in S} w_k y_{1ik} \left(1 - z_{2k}\right) \\ \hat{C}_j &= \sum_{k \in S} w_k y_{2jk} \left(1 - z_{1k}\right) \\ \hat{M} &= \sum_{k \in S} w_k \left(1 - z_{1k}\right)\left(1 - z_{2k}\right) \end{aligned}$$

pour N_{ij}, R_i, C_j et M, respectivement. Soulignons qu'une estimation sans biais pour la taille de la population est donnée par

$$\hat{N} = \sum_i \sum_j \hat{N}_{ij} + \sum_j \hat{C}_j + \sum_i \hat{R}_i + \hat{M} = \sum_s w_k v_k$$

où

$$v_k = \sum_i y_{1ik} \sum_j y_{2jk} + \sum_j y_{2jk}\left(1 - z_{1k}\right) + \sum_i y_{1ik}\left(1 - z_{2k}\right) + \left(1 - z_{1k}\right)\left(1 - z_{2k}\right).$$

En tenant compte de la forme fonctionnelle de tous les paramètres d'intérêt, si nous constatons que la fonction de vraisemblance du modèle est proportionnelle à (3.1), nous obtenons le résultat suivant.

Résultat 4.1 *Le logarithme du rapport de vraisemblance pour les données observées au niveau de la population peut être réécrit comme suit* :

$$l_U = \sum_{k \in U} f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right) \tag{4.1}$$

où

$$\begin{aligned} & f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right) \\ & = \sum_i \sum_j y_{1ik} y_{2jk} \ln\left(\psi \rho_{RR} \eta_i p_{ij}\right) \\ & + \sum_i y_{1ik}\left(1 - z_{2k}\right) \ln\left(\sum_j \psi\left(1 - \rho_{RR}\right) \eta_i p_{ij}\right) \\ & + \sum_j y_{2jk}\left(1 - z_{1k}\right) \ln\left(\sum_i \left(1 - \psi\right)\left(1 - \rho_{MM}\right) \eta_i p_{ij}\right) \\ & + \left(1 - z_{1k}\right)\left(1 - z_{2k}\right) \ln\left(\sum_i \sum_j \left(1 - \psi\right) \rho_{MM} \eta_i p_{ij}\right) \end{aligned}$$

où $\mathbf{y}_1$ *est un vecteur renfermant les caractéristiques* y_{1ik}, $\mathbf{y}_2$ *est un vecteur renfermant les caractéristiques* y_{2jk}, $\mathbf{z}_1$ *est un vecteur renfermant les caractéristiques* z_{1k}, *et* $\mathbf{z}_2$ *est un vecteur renfermant les caractéristiques* z_{2k} *(pour chaque* $k = 1, \ldots, N$ *et* $i, j = 1, \ldots, G$*).*

Maintenant, afin d'obtenir des estimateurs des paramètres, il faut maximiser cette dernière fonction. Au moyen de techniques standard de maximum de vraisemblance, les équations de probabilité correspondantes sont données par

$$\sum_{k \in U} \mathbf{u}_k\left(\theta\right) = \mathbf{0}$$

où le vecteur $\mathbf{u}_k$, communément appelé *scores*, est défini par

$$\mathbf{u}_k\left(\theta\right) = \frac{\partial f_k\left(\theta\right)}{\partial \theta}.$$

De plus, comme il est inhabituel de sonder la population au complet, un échantillon probabiliste est sélectionné, et l'expression $\sum_{k \in U} \mathbf{u}_k\left(\theta\right)$ est considérée comme un paramètre de population. Ainsi, en considérant que $w_k = 1/\pi_k$ est le poids d'échantillonnage correspondant, un estimateur sans biais pour cette somme de *scores* est défini comme $\sum_{k \in S} w_k \mathbf{u}_k\left(\theta\right)$. La prochaine expression s'appelle pseudo-équation de vraisemblance et elle constitue une façon efficace de trouver des estimateurs pour les paramètres du modèle en tenant compte du poids d'échantillonnage :

$$\sum_{k \in S} w_k \mathbf{u}_k\left(\theta\right) = \mathbf{0}.$$

On présume que pour le modèle dans le présent article, la probabilité initiale qu'une personne réponde pendant une période $t-1$ est la même pour toutes les classifications possibles dans l'enquête. De plus, les probabilités de transition entre les répondants et les non-répondants ne dépendent pas de la classification des personnes dans l'enquête, ρ_{MM} et ρ_{RR}. Compte tenu de ces suppositions, les résultats suivants présumeront que l'estimation des probabilités du modèle de Markov tient compte du poids d'échantillonnage.

Résultat 4.2 *Sous les hypothèses du modèle, les estimateurs du maximum de pseudo-vraisemblance obtenus pour* ψ, ρ_{RR} *et* ρ_{MM} *sont données par*

$$\hat{\psi}_{mpv} = \frac{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j + \hat{M}}$$

$$\hat{\rho}_{RR,mpv} = \frac{\sum_i \sum_j \hat{N}_{ij}}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}$$

$$\hat{\rho}_{MM,mpv} = \frac{\hat{M}}{\sum_j \hat{C}_j + \hat{M}}$$

respectivement.

Résultat 4.3 *Sous les hypothèses du modèle, les estimateurs du maximum de pseudo-vraisemblance obtenus pour* η_i *et* p_{ij} *sont obtenus par itération jusqu'à la convergence des prochaines expressions*

$$\hat{\eta}_{i,mpv}^{(v+1)} = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j}$$

$$\hat{p}_{ij,mpv}^{(v+1)} = \frac{\hat{N}_{ij} + \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_j \hat{N}_{ij} + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}$$

respectivement. L'indice supérieur (v) *indique la valeur de l'estimation pour les paramètres d'intérêt à l'itération* v.

Les résultats qui précèdent offrent un cadre complet pour la mise en œuvre du modèle markovien à deux degrés afin de tenir compte du poids d'échantillonnage dans les enquêtes longitudinales. Une autre question d'intérêt consiste à déterminer comment choisir les valeurs initiales $\left\{\hat{\eta}_i^{(0)}\right\}$ et $\left\{\hat{p}_{ij}^{(0)}\right\}$. En général, n'importe quel ensemble de valeurs est valide s'il respecte les restrictions initiales :

$$\sum_i \hat{\eta}_i^{(0)} = 1$$

$$\sum_j \hat{p}_{ij}^{(0)} = 1.$$

Cependant, d'après les directives de Chen et Fienberg (1974) et compte tenu du cas hypothétique où toutes les personnes auraient répondu aux deux périodes, alors $M = 0, R_i = 0$ (pour chaque $i = 1,\ldots,G$) et $C_j = 0$ (pour chaque $j = 1,\ldots,G$) et leurs estimations d'échantillonnage sont également nulles. Par conséquent, et compte tenu des expressions des estimateurs obtenus, un choix judicieux est donné par

$$\hat{\eta}_i^{(0)} = \frac{\sum_j \hat{N}_{ij}}{\sum_i \sum_j \hat{N}_{ij}}$$

$$\hat{p}_{ij}^{(0)} = \frac{\hat{N}_{ij}}{\sum_j \hat{N}_{ij}}.$$

Enfin, cette approche itérative est souvent mise en œuvre pour les problèmes d'estimation par maximum de vraisemblance dans les tableaux de contingence. Toutefois, certaines approches pour l'intégration de modèles loglinéaires dans les tableaux de contingence pour les plans de sondage complexes se trouvent entre autres dans les travaux de Clogg et Eliason (1987), Rao et Thomas (1988), Skinner et Vallet (2010). Le prochain résultat offre une approche de l'estimation des flux bruts compte tenu du poids d'échantillonnage pendant les deux périodes d'intérêt.

Résultat 4.4 *Sous les hypothèses du modèle, un estimateur d'échantillonnage de* μ_{ij} *est*

$$\hat{\mu}_{ij,mpv} = \hat{N}\hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}.$$

5 Propriétés des estimateurs

D'après Cassel, Särndal et Wretman (1976), le but de la prise en compte d'une approche d'échantillonnage consiste à recueillir de l'information d'un sous-ensemble (échantillon) d'unités dans la population finie pour obtenir une conclusion pour l'ensemble de la population. Pendant ce processus, le statisticien doit composer avec les sources de hasard qui définissent le comportement stochastique complexe du processus inférentiel. Bien que cet article considère le plan d'échantillonnage comme une mesure de probabilité déterminant l'inférence pour les paramètres et le modèle, il faut comprendre que le modèle markovien proposé offre une autre mesure bien définie de la probabilité. Maintenant, nous obtenons certaines propriétés des estimateurs proposés à la dernière section.

L'objectif du présent article consiste à intégrer le poids d'échantillonnage dans le modèle proposé, et il est important d'obtenir des estimateurs à peu près sans biais en ce qui concerne la mesure de probabilité liée au plan d'échantillonnage pour $\boldsymbol{\theta}$ et $\boldsymbol{\mu}$. Les résultats suivants montrent certaines propriétés des estimateurs proposés considérés selon le plan de sondage complexe. En ce qui concerne la notation, la mesure de probabilité induite pour le plan d'échantillonnage sera représentée par le sous-indice p. Les résultats suivants fournissent les estimateurs du maximum de vraisemblance pour les paramètres d'intérêt lorsqu'au lieu d'obtenir un échantillon, la mesure est obtenue au moyen d'un recensement ou d'un dénombrement complet des personnes dans la population.

Résultat 5.1 *Supposons qu'il y ait un accès complet à l'ensemble de la population et que la fonction de logarithme du rapport de vraisemblance soit donnée par (4.1). Par conséquent, les estimateurs du maximum de vraisemblance, sous les hypothèses du modèle, sont les suivants*

$$\psi_U = \frac{\sum_i \sum_j N_{ij} + \sum_i R_i}{\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M}$$

$$\rho_{RR,U} = \frac{\sum_i \sum_j N_{ij}}{\sum_i \sum_j N_{ij} + \sum_i R_i}$$

$$\rho_{MM,U} = \frac{M}{\sum_j C_j + M}$$

$$\eta_{i,U}^{(v+1)} = \frac{\sum_j N_{ij} + R_i + \sum_j \left(C_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j} \tag{5.1}$$

$$p_{ij,U}^{(v+1)} = \frac{N_{ij} + \left(C_j \hat{\eta}_i^{(v)} p_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_j N_{ij} + \sum_j \left(C_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)} \tag{5.2}$$

où (5.1) et (5.2) doivent être itérés conjointement jusqu'à la convergence.

Résultat 5.2 *Sous les hypothèses du modèle, un estimateur de maximum de vraisemblance de* μ_{ij} *est*

$$\mu_{ij,U} = N \times \eta_{i,U} \times p_{ij,U}$$

où N *correspond à la taille de la population et* $\eta_{i,U}$ *et* $p_{ij,U}$ *sont définis par le dernier résultat, respectivement.*

Soulignons que $\boldsymbol{\theta}$ et $\boldsymbol{\mu}$ peuvent être définis comme des paramètres descriptifs de la population. D'après l'approche par inférence induite par la méthode du maximum de vraisemblance, il y a des estimateurs $\boldsymbol{\theta}_U = \left(\psi'_U, \boldsymbol{\rho}'_{RR,U}, \boldsymbol{\rho}'_{MM,U}, \boldsymbol{\eta}'_U, \mathbf{p}'_U \right)'$ et $\boldsymbol{\mu}_U = \left(\mu_{11,U}, \ldots, \mu_{ij,U}, \ldots, \mu_{GG,U} \right)'$ définis comme les paramètres descriptifs de la population correspondants qui font que $\boldsymbol{\theta}_{mpv}$ et $\boldsymbol{\mu}_{mpv}$ sont convergents en ce qui concerne le plan d'échantillonnage au sens de la définition 2 dans Pfeffermann (1993). Soulignons par ailleurs que $\boldsymbol{\theta}_U$ et $\boldsymbol{\mu}_U$ peuvent être calculés uniquement si l'on a accès à l'ensemble de la population finie.

D'après Pessoa et Silva (1998, p. 79), il est possible d'évaluer que sous certaines conditions de régularité, $\boldsymbol{\theta}_U - \boldsymbol{\theta} = o_p(1)$ et $\boldsymbol{\mu}_U - \boldsymbol{\mu} = o_p(1)$. De plus, comme dans de nombreuses enquêtes par sondage, la population et la taille de l'échantillon sont généralement grandes, un estimateur approprié de $\boldsymbol{\theta}_U$ est également un estimateur adéquat pour $\boldsymbol{\theta}$, et un estimateur approprié pour $\boldsymbol{\mu}_U$ sera un estimateur adéquat pour $\boldsymbol{\mu}$.

À la prochaine section, nous examinons les propriétés des estimateurs déjà proposés et nous parlerons de leur pertinence pour notre problème de recherche.

5.1 Propriétés des estimateurs de dénombrement

Résultat 5.3 *Les estimateurs* $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$, $\hat{M}$ *et* $\hat{N}$ *définis à la section 4 sont sans biais en ce qui concerne le plan d'échantillonnage.*

La preuve est très immédiate. Le coefficient de pondération w_k correspond à l'inverse de la probabilité d'inclusion π_k, associé à l'élément k. Tous les estimateurs sont de la catégorie Horvitz-Thompson et sont donc sans biais.

Résultat 5.4 *En supposant que* $w_k = 1/\pi_k$, *les variances correspondantes pour* $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$, $\hat{M}$ *et* $\hat{N}$, *sont données par*

$$Var_p\left(\hat{N}_{ij}\right) = \sum_U \sum_U \Delta_{kl} \frac{y_{1ik} y_{2jk}}{\pi_k} \frac{y_{1il} y_{2jl}}{\pi_l}$$

$$Var_p\left(\hat{R}_i\right) = \sum_U \sum_U \Delta_{kl} \frac{y_{1ik}\left(1 - z_{2k}\right)}{\pi_k} \frac{y_{1il}\left(1 - z_{2l}\right)}{\pi_l}$$

$$Var_p\left(\hat{C}_j\right) = \sum_U \sum_U \Delta_{kl} \frac{y_{2jk}\left(1 - z_{1k}\right)}{\pi_k} \frac{y_{2jl}\left(1 - z_{1l}\right)}{\pi_l}$$

$$Var_p\left(\hat{M}\right) = \sum_U \sum_U \Delta_{kl} \frac{\left(1 - z_{1k}\right)}{\pi_k} \frac{\left(1 - z_{1l}\right)}{\pi_l}$$

$$Var_p\left(\hat{N}\right) = \sum_U \sum_U \Delta_{kl} \frac{v_k}{\pi_k} \frac{v_l}{\pi_l}.$$

Les estimateurs sans biais pour ces variances, respectivement, sont donnés par

$$\widehat{Var}_p\left(\hat{N}_{ij}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_{1ik} y_{2jk}}{\pi_k} \frac{y_{1il} y_{2jl}}{\pi_l}$$

$$\widehat{Var}_p\left(\hat{R}_i\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_{1ik}\left(1 - z_{2k}\right)}{\pi_k} \frac{y_{1il}\left(1 - z_{2l}\right)}{\pi_l}$$

$$\widehat{Var}_p\left(\hat{C}_j\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_{2jk}\left(1 - z_{1k}\right)}{\pi_k} \frac{y_{2jl}\left(1 - z_{1l}\right)}{\pi_l}$$

$$\widehat{Var}_p\left(\hat{M}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{\left(1 - z_{1k}\right)}{\pi_k} \frac{\left(1 - z_{1l}\right)}{\pi_l}$$

$$\widehat{Var}_p\left(\hat{N}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{v_k}{\pi_k} \frac{v_l}{\pi_l}.$$

Par ailleurs, si w_k correspond aux poids de calage sur marges, alors tous les estimateurs envisagés sont sans biais asymptotique et des preuves sont données dans Deville et Särndal (1992). Leurs variances correspondantes sont données par Kim et Park (2010).

5.2 Propriétés des estimateurs de probabilités des modèles

Résultat 5.5 *L'approximation du premier degré de Taylor pour l'estimateur* ψ_{mpv}*, définie comme le résultat 4.2 qui précède, autour du point* $\left(N_{ij}, R_i, C_j, M\right)$ *et* $i, j = 1, \ldots, G$*, est donnée par l'expression*

$$\begin{aligned}\hat{\psi}_{mpv} &\cong \hat{\psi}_0 \\ &= \psi_U + a_1 \sum_i \sum_j \left(\hat{N}_{ij} - N_{ij}\right) + a_1 \sum_i \left(\hat{R}_i - R_i\right) \\ &\qquad + a_2 \sum_j \left(\hat{C}_j - C_j\right) + a_2 \left(\hat{M} - M\right)\end{aligned}$$

où

$$a_1 = \frac{\sum_j C_j + M}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M\right)^2}$$

$$a_2 = -\frac{\sum_i \sum_j N_{ij} + \sum_i R_i}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M\right)^2}.$$

Résultat 5.6 *L'approximation du premier degré de Taylor pour l'estimateur* $\hat{\rho}_{RR,mpv}$*, définie au résultat 4.2 plus haut, autour du point* $\left(N_{ij}, R_i\right)$ *et* $i, j = 1, \ldots, G$*, est donnée par l'expression*

$$\begin{aligned}\hat{\rho}_{RR,mpv} &\cong \hat{\rho}_{RR,0} \\ &= \rho_{RR,U} + a_3 \sum_i \sum_j \left(\hat{N}_{ij} - N_{ij}\right) + a_4 \sum_i \left(\hat{R}_i - R_i\right)\end{aligned}$$

où

$$a_3 = \frac{\sum_i R_i}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i\right)^2}$$

$$a_4 = -\frac{\sum_i \sum_j N_{ij}}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i\right)^2}.$$

Résultat 5.7 *L'approximation du premier degré de Taylor pour l'estimateur* $\hat{\rho}_{MM,mpv}$*, définie au résultat 4.2 plus haut, autour du point* $\left(C_j, M\right)$ *et* $j = 1, \ldots, G,$ *est donnée par l'expression*

$$\begin{aligned}\hat{\rho}_{MM,mpv} &\cong \hat{\rho}_{MM,0} \\ &= \rho_{MM,U} + a_5 \sum_j \left(\hat{C}_j - C_j\right) + a_6 \left(\hat{M} - M\right)\end{aligned}$$

où

$$a_5 = -\frac{M}{\left(\sum_j C_j + M\right)^2}$$
$$a_6 = -\frac{\sum_j C_j}{\left(\sum_j C_j + M\right)^2}.$$

Résultat 5.8 *Les estimateurs* $\hat{\psi}_{mpv}$, $\hat{\rho}_{MM,mpv}$ *et* $\hat{\rho}_{RR,mpv}$, *sont à peu près sans biais pour* ψ_U, $\rho_{MM,U}$, $\rho_{RR,U}$.

Résultat 5.9 *Les estimateurs* $\hat{\eta}_{i,mpv}$ *et* $\hat{p}_{ij,mpv}$, *sont à peu près sans biais pour* $\eta_{i,U}$ *et* $p_{ij,U}$.

Résultat 5.10 *Les variances approximatives pour les estimateurs* $\hat{\psi}_{mpv}$, $\hat{\rho}_{MM,mpv}$ *et* $\hat{\rho}_{RR,mpv}$, *sont données par*

$$AV_p\left(\hat{\psi}_{mpv}\right) = V_p\left(\sum_s \frac{E_k^{\psi}}{\pi_k}\right) = \sum_U \sum_U \Delta_{kl} \frac{E_k^{\psi}}{\pi_k} \frac{E_l^{\psi}}{\pi_l}$$
$$AV_p\left(\hat{\rho}_{RR,mpv}\right) = V_p\left(\sum_s \frac{E_k^{RR}}{\pi_k}\right) = \sum_U \sum_U \Delta_{kl} \frac{E_k^{RR}}{\pi_k} \frac{E_l^{RR}}{\pi_l}$$
$$AV_p\left(\hat{\rho}_{MM,mpv}\right) = V_p\left(\sum_s \frac{E_k^{MM}}{\pi_k}\right) = \sum_U \sum_U \Delta_{kl} \frac{E_k^{MM}}{\pi_k} \frac{E_l^{MM}}{\pi_l}$$

où

$$\begin{aligned} E_k^{\psi} &= a_1\left(2 - z_{2k}\right) + a_2\left(1 - z_{1k}\right)\left(2 - z_{2k}\right) \\ E_k^{RR} &= a_3 + a_4\left(1 - z_{2k}\right) \\ E_k^{MM} &= a_5\left(1 - z_{1k}\right) + a_6\left(1 - z_{1k}\right)\left(1 - z_{2k}\right). \end{aligned}$$

Résultat 5.11 *Les estimateurs sans biais pour les variances approximatives des estimateurs* $\hat{\psi}_{mpv}$, $\hat{\rho}_{MM,mpv}$ *et* $\hat{\rho}_{RR,mpv}$, *sont donnés par*

$$\hat{V}\left(\hat{\psi}_{mpv}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k^{\psi}}{\pi_k} \frac{e_l^{\psi}}{\pi_l}$$
$$\hat{V}\left(\hat{\rho}_{RR,mpv}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k^{RR}}{\pi_k} \frac{e_l^{RR}}{\pi_l}$$
$$\hat{V}\left(\hat{\rho}_{MM,mpv}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k^{MM}}{\pi_k} \frac{e_l^{MM}}{\pi_l}$$

respectivement, où

$$\begin{aligned} e_k^{\psi} &= \hat{a}_1\left(2 - z_{2k}\right) + \hat{a}_2\left(1 - z_{1k}\right)\left(2 - z_{2k}\right) \\ e_k^{RR} &= \hat{a}_3 + \hat{a}_4\left(1 - z_{2k}\right) \\ e_k^{MM} &= \hat{a}_5\left(1 - z_{1k}\right) + \hat{a}_6\left(1 - z_{1k}\right)\left(1 - z_{2k}\right) \end{aligned}$$

et

$$\hat{a}_1 = \frac{\sum_j \hat{C}_j + \hat{M}}{\left(\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j + \hat{M}\right)^2}$$

$$\hat{a}_2 = -\frac{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}{\left(\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j + \hat{M}\right)^2}$$

$$\hat{a}_3 = \frac{\sum_i \hat{R}_i}{\left(\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i\right)^2}$$

$$\hat{a}_4 = -\frac{\sum_i \sum_j \hat{N}_{ij}}{\left(\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i\right)^2}$$

$$\hat{a}_5 = -\frac{\hat{M}}{\left(\sum_j \hat{C}_j + \hat{M}\right)^2}$$

$$\hat{a}_6 = -\frac{\sum_j \hat{C}_j}{\left(\sum_j \hat{C}_j + \hat{M}\right)^2}.$$

Résultat 5.12 *Les variances approximatives pour les estimateurs* $\hat{\eta}_{i,mpv}$ *et* $\hat{p}_{ij,mpv}$ *sont données par*

$$AV_p\left(\hat{\eta}_{i,mpv}\right) = \frac{1}{\left(J_{\eta_i}\right)^2} \sum_U \sum_U \Delta_{kl} \frac{u_k\left(\eta_i\right)}{\pi_k} \frac{u_l\left(\eta_i\right)}{\pi_l}$$

$$AV_p\left(\hat{p}_{ij,mpv}\right) = \frac{1}{\left(J_{p_{ij}}\right)^2} \sum_U \sum_U \Delta_{kl} \frac{u_k\left(p_{ij}\right)}{\pi_k} \frac{u_l\left(p_{ij}\right)}{\pi_l}$$

où

$$u_k\left(\eta_i\right) = \frac{\sum_j y_{1ik} y_{2jk} + y_{1ik}\left(1 - z_{2k}\right)}{\eta_i} + \sum_j y_{2jk}\left(1 - z_{1k}\right) \frac{p_{ij}}{\sum_i \eta_i p_{ij}} + \left(1 - z_{1k}\right)\left(1 - z_{2k}\right) - 1$$

$$u_k\left(p_{ij}\right) = \frac{y_{1ik} y_{2jk}}{p_{ij}} + y_{1ik}\left(1 - z_{2k}\right) + y_{2jk}\left(1 - z_{1k}\right) \frac{\eta_i}{\sum_i \eta_i p_{ij}} + \left(1 - z_{1k}\right)\left(1 - z_{2k}\right)\eta_i$$
$$- \frac{1}{\hat{N}}\left(\sum_j \hat{N}_{ij} + \hat{R}_i + \hat{M}\eta_i + \sum_j \hat{C}_j \left(\frac{\eta_i p_{ij}}{\sum_i \eta_i p_{ij}}\right)\right)$$

$$J_{\eta_i} = -\frac{2}{\eta_i^2} \sum_U y_{1ik} + \frac{1}{\eta_i^2} \sum_U y_{1ik} z_{2k} - \sum_U \left(1 - z_{1k}\right) \sum_j \frac{y_{2jk} p_{ij}^2}{\left(\sum_i \eta_i p_{ij}\right)^2}$$

$$J_{p_{ij}} = -\frac{1}{p_{ij}^2} \sum_U y_{1ik} y_{2jk} - \frac{\eta_i^2}{\left(\sum_i \eta_i p_{ij}\right)^2} \sum_U y_{2jk}\left(1 - z_{1k}\right).$$

Résultat 5.13 *Les estimateurs sans biais pour les variances approximatives des estimateurs* $\hat{\eta}_{i,mpv}$ *et* $\hat{p}_{ij,mpv}$ *sont donnés par*

$$\hat{V}_p\left(\hat{\eta}_{i,mpv}\right)=\frac{1}{\left(\hat{J}_{\hat{\eta}_i}\right)^2}\sum_s\sum_s\frac{\Delta_{kl}}{\pi_{kl}}\frac{\hat{u}_k\left(\hat{\eta}_i\right)}{\pi_k}\frac{\hat{u}_l\left(\hat{\eta}_i\right)}{\pi_l}$$

$$\hat{V}_p\left(\hat{p}_{ij,mpv}\right)=\frac{1}{\left(\hat{J}_{\hat{p}_{ij}}\right)^2}\sum_s\sum_s\frac{\Delta_{kl}}{\pi_{kl}}\frac{\hat{u}_k\left(\hat{p}_{ij}\right)}{\pi_k}\frac{\hat{u}_l\left(\hat{p}_{ij}\right)}{\pi_l}$$

où

$$\hat{u}_k\left(\hat{\eta}_i\right)=\frac{\sum_j y_{1ik}y_{2jk}+y_{1ik}\left(1-z_{2k}\right)}{\hat{\eta}_i}+\sum_j y_{2jk}\left(1-z_{1k}\right)\frac{\hat{p}_{ij,mpv}}{\sum_i\hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}}+\left(1-z_{1k}\right)\left(1-z_{2k}\right)$$

$$\hat{u}_k\left(\hat{p}_{ij}\right)=\frac{y_{1ik}y_{2jk}}{\hat{p}_{ij,mpv}}+y_{1ik}\left(1-z_{2k}\right)+y_{2jk}\left(1-z_{1k}\right)\frac{\hat{\eta}_{i,mpv}}{\sum_i\hat{\eta}_{i,mpv}p_{ij,,mpv}}+\left(1-z_{1k}\right)\left(1-z_{2k}\right)\hat{\eta}_{i,mpv}$$

et

$$\hat{J}_{\hat{\eta}_i}=-\frac{2}{\hat{\eta}_{i,mpv}^2}\sum_U y_{1ik}+\frac{1}{\hat{\eta}_{i,mpv}^2}\sum_U y_{1ik}z_{2k}-\sum_U(1-z_{1k})\sum_j\frac{y_{2jk}\hat{p}_{ij,mpv}^2}{\left(\sum_i\hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}\right)^2}$$

$$\hat{J}_{\hat{p}_{ij}}=-\frac{1}{\hat{p}_{ij,mpv}^2}\sum_U y_{1ik}y_{2jk}-\frac{\hat{\eta}_{i,mpv}^2}{\left(\sum_i\hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}\right)^2}\sum_U y_{2jk}\left(1-z_{1k}\right).$$

5.3 Propriétés des estimateurs des flux bruts

Résultat 5.14 *Sous les hypothèses du modèle, l'approximation du premier degré de Taylor de l'estimateur des flux bruts donnée par* $\hat{\mu}_{ij}$ *et définie dans le résultat 4.4, autour du point* $(N,\eta_{i,U},p_{ij,U})$ *et* $i,j=1,\ldots,G,$ *est donnée par*

$$\begin{aligned}\hat{\mu}_{ij,mpv}&\cong\hat{\mu}_{ij,0}\\&=\mu_{ij,U}+a_7\left(\hat{N}_{ij}-N_{ij}\right)+a_8\left(\hat{\eta}_{i,mpv}-\eta_{i,U}\right)+a_9\left(\hat{p}_{ij,mpv}-p_{ij,U}\right)\end{aligned}$$

où

$$\begin{aligned}a_7&=\eta_{i,U}p_{ij,U}\\a_8&=N_{ij}p_{ij,U}\\a_9&=N_{ij}\eta_{i,U}.\end{aligned}$$

Résultat 5.15 *L'estimateur des flux bruts* $\hat{\mu}_{ij,mpv}$ *est à peu près sans biais pour* $\mu_{ij,U}$.

Résultat 5.16 *L'expression suivante évalue la variance approximative pour* $\hat{\mu}_{ij,mpv}$

$$AV_p\left(\hat{\mu}_{ij,mpv}\right)\cong a_7^2Var_p\left(\hat{N}_{ij}\right)+a_8^2AV_p\left(\hat{\eta}_{i,mpv}\right)+a_9^2AV_p\left(\hat{p}_{ij}\right).\tag{5.3}$$

Résultat 5.17 *Un estimateur approximativement sans biais pour la variance asymptotique dans (5.3) est donné par*

$$\hat{V}_p\left(\hat{\mu}_{ij,mpv}\right) = \hat{a}_7^2\hat{V}_p\left(\hat{N}_{ij}\right) + a_8^2\hat{V}_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2\hat{V}_p\left(\hat{p}_{ij}\right)$$

où

$$\hat{a}_7 = \hat{\eta}_{i,U}\,\hat{p}_{ij,U}$$

$$\hat{a}_8 = \hat{N}_{ij}\,\hat{p}_{ij,U}$$

$$\hat{a}_9 = \hat{N}_{ij}\hat{\eta}_{i,U}.$$

6 Application empirique

Nous examinons d'abord une approche empirique dans cette section, au moyen de simulations qui nous permettront d'évaluer certaines propriétés statistiques comme l'absence de biais et l'efficacité des estimateurs proposés. D'après la modélisation proposée par Stasny (1987), nous avons considéré une simulation à deux degrés comme suit :

- Répartition de toutes les personnes dans la population aux différentes cellules d'un tableau de contingence. Dans ce premier degré, nous allons définir les probabilités initiales η_i, p_{ij} et
- Processus de non-réponse à deux périodes consécutives. Dans ce deuxième degré, nous allons définir les probabilités initiales ψ, ρ_{RR} et ρ_{MM} .

Au premier degré, il a fallu présumer certaines conditions (processus non observable) où les probabilités de classification de groupe ont été établies à une période $t-1$ et les probabilités de classification conditionnelle pendant une période t. Ainsi, toutes les personnes dans la population étaient réputées classées dans une des trois catégories suivantes : E1, E2 et E3. Le vecteur d'état pendant une période t a été donné par

$$\boldsymbol{\eta} = \left(\eta_1, \eta_2, \eta_3\right)' = \left(0{,}9; 0{,}05; 0{,}05\right)'.$$

Ainsi, il y a une probabilité de classification dans E1 équivalant à 0,9 pour toute personne dans la population et des probabilités de classification dans E2 et E3 équivalant à 0,05. La matrice de transition de la période $t-1$ à la période t est donnée par

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}_1' \\ \mathbf{p}_2' \\ \mathbf{p}_3' \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} 0{,}80 & 0{,}15 & 0{,}05 \\ 0{,}30 & 0{,}60 & 0{,}10 \\ 0{,}10 & 0{,}10 & 0{,}80 \end{pmatrix}.$$

Nous avons présumé que la taille de la population était de $N = 100\ 000$ et que sa taille ne changerait pas aux deux périodes d'évaluation. Afin de classer les personnes aux périodes, nous avons utilisé la fonction R rmultinom (R Development Core Team 2012). Ainsi, la répartition des flux bruts en fonction de l'équation (3.2) serait donnée par les valeurs dans le tableau 6.1.

Tableau 6.1
Valeurs prévues selon le modèle ξ pour les flux bruts de la population pendant deux périodes consécutives

Période $t-1$	Période t		
	E1	**E2**	**E3**
E1	72 000	13 500	4 500
E2	1 500	3 000	500
E3	500	500	4 000

6.1 Méthodologie

Pour cet exercice empirique, nous avons considéré $L = 1\,000$ simulations par la méthode de Monte Carlo. Afin de répartir les personnes entre les répondants et les non-répondants pendant les deux périodes, nous avons utilisé la fonction rmultinom du langage R. Des variables dichotomiques y_{1ik}, y_{2jk}, z_{1k} et z_{2k} ont été créées au moyen de la fonction Domains de la bibliothèque TeachingSampling (Gutiérrez 2009).

Pour chaque exécution de la simulation, un échantillon de taille $n = 10\,000$ a été tiré. Nous avons considéré un échantillonnage aléatoire simple (SI) ainsi qu'un plan d'échantillonnage complexe donnant des probabilités d'inclusion inégales (π PS). Le comportement des différents estimateurs proposés sera évalué en fonction de leur biais relatif et de leur erreur quadratique moyenne relative, donnés par

$$BR = L^{-1}\sum_{l=1}^{L}\frac{\hat{\theta}_l - \theta}{\theta} \qquad \text{et} \qquad EQMR = \frac{\sqrt{L^{-1}\sum_{l=1}^{L}\left(\hat{\theta}_l - \theta\right)^2}}{\theta}.$$

respectivement. Dans les situations où le vecteur des probabilités d'inclusion était inégal, la fonction S.piPS dans la bibliothèque TeachingSampling a été utilisée afin de choisir un échantillon sans remplacement avec des probabilités d'inclusion proportionnelles à une caractéristique auxiliaire présumée connue et selon la répartition normale avec différents paramètres. La méthodologie proposée est comparée à deux autres estimateurs : un estimateur tenant compte de la forme fonctionnelle du modèle sans tenir compte du plan d'échantillonnage et un estimateur des flux bruts ne tenant pas compte du plan d'échantillonnage mais présumant que la non-réponse est ignorable.

Le premier estimateur, que nous appelons l'*estimateur fondé sur le plan*, correspond aux expressions aux résultats 4.2, 4.3 et 4.4. Le deuxième estimateur, que nous appellerons *estimateur basé sur le modèle*, correspond aux expressions au résultat 5.1, puisque les estimateurs du maximum de vraisemblance ne tiennent pas compte des poids d'échantillonnage. Enfin, le troisième estimateur, que nous appelons l'*estimateur naïf*, projette l'information au niveau de l'échantillon à la population et est donné par

$$\hat{\mu}_{ij,ING} = \frac{N}{\sum_i \sum_j N_{ij}} N_{ij}.$$

La probabilité de réponse pendant la période $t-1$ a été présumée comme $\psi = 0{,}8$. La probabilité de réponse pendant la période t pour les personnes ayant répondu à la période $t-1$ a été présumée $\rho_{RR} = 0{,}9$. Enfin, la probabilité de non-réponse pendant une période t pour les personnes n'ayant pas répondu pendant la période $t-1$ a été présumée $\rho_{MM} = 0{,}7$.

D'après le modèle ξ, les valeurs prévues des réponses sont données dans le tableau 6.2.

Tableau 6.2
Valeurs prévues d'après le modèle ξ pour la réponse aux deux périodes consécutives

Période $t-1$	Période t	
	Réponse	**Non-réponse**
Réponse	72 000	8 000
Non-réponse	6 000	14 000

En tenant compte de la dynamique des répondants dans les deux périodes et en présumant qu'il est possible de recueillir toute l'information sur la population au moyen d'un recensement, nous obtenons les classifications présentées dans le tableau 6.3 ci-après.

Tableau 6.3
Valeurs prévues d'après le modèle ξ pour les flux bruts de la population (processus observable) pendant deux périodes consécutives

Période $t-1$	Période t			
	E1	**E2**	**E3**	**Complément de ligne**
E1	51 840	9 720	3 240	7 200
E2	1 080	2 160	360	400
E3	360	360	2 880	400
Complément de colonne	4 440	1 020	540	14 000

6.2 Résultats

6.2.1 Échantillonnage aléatoire simple : *estimateur fondé sur le plan* et *fondé sur le modèle*

Dans une première approche empirique, nous avons considéré un échantillonnage aléatoire simple sans remise comme plan d'échantillonnage. Ce plan d'échantillonnage induit des probabilités d'inclusion et des facteurs d'expansion uniformes. Selon ce scénario, les estimateurs fondés sur le plan et fondés sur le modèle sont les mêmes. Conformément à ce scénario, l'approche démontre une certaine robustesse d'après les valeurs des biais relatifs qui peuvent être considérées comme négligeables. On peut le constater dans les tableaux 6.4, 6.5 et 6.6.

Tableau 6.4
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) de l'estimateur proposé pour les flux bruts de la population

Période $t-1$	Période t		
	E1	**E2**	**E3**
E1	0,24 (0,094)	-0,35 (0,189)	-0,49 (0,474)
E2	-2,89 (0,158)	-1,89 (0,221)	2,00 (0,980)
E3	-0,63 (0,790)	4,54 (0,822)	-0,84 (0,569)

Tableau 6.5
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) des probabilités de transition p_{ij}

Période $t-1$	**Période** t		
	E1	**E2**	**E3**
E1	0,13 (0,284)	-0,39 (0,537)	-1,00 (3,225)
E2	1,70 (1,296)	-2,29 (0,569)	8,64 (0,347)
E3	-6,6 (3,415)	2,09 (1,992)	0,56 (0,158)

Tableau 6.6
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) des probabilités de classification initiales η_i

Période $t-1$		
η_1	η_2	η_3
-0,01 (0,094)	-1,42 (0,980)	1,74 (0,790)

De plus, le biais relatif en pourcentage pour la probabilité de réponse ψ était de -0,23 et l'erreur quadratique moyenne relative en pourcentage était de 0,221; pour la probabilité de réponse ρ_{RR}, le biais en pourcentage était de 0,055 et l'erreur quadratique moyenne relative en pourcentage était de 0,031; pour la probabilité de non-réponse ρ_{MM}, le biais en pourcentage était de -0,192 et l'erreur quadratique moyenne relative en pourcentage était de 0,189. Par ailleurs, le tableau 6.7 montre la valeur prévue empirique des flux bruts pour l'estimateur proposé, et on peut constater que les valeurs sont très proches de celles données dans le tableau 6.1.

Tableau 6.7
Valeurs prévues empiriques pour l'estimateur proposé des flux bruts de la population

Période $t-1$	**Période** t		
	E1	**E2**	**E3**
E1	72 085	13 444	4 454
E2	1 504	2 889	535
E3	474	519	4 092

6.2.2 Échantillonnage aléatoire simple : *estimateur naïf*

Conformément à ce scénario et étant donné que cet estimateur ne tient pas compte du processus de non-réponse, les valeurs des biais relatifs ne peuvent pas être considérées comme négligeables. On peut le constater dans le tableau 6.8.

Tableau 6.8
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) de l'estimateur naïf pour les flux bruts de la population

Période $t-1$	Période t		
	E1	**E2**	**E3**
E1	-1,21 (4,4)	**10,2** (60,7)	8,34 (25,2)
E2	-0,25 (38,8)	-7,51 (30,9)	1,33 (12,6)
E3	**13,7** (43,3)	-8,54 (46,1)	0,92 (6,9)

Le tableau 6.9 montre les valeurs empiriques prévues pour l'estimateur naïf; comparativement aux valeurs prévues pour le modèle présenté dans le tableau 6.1, ces valeurs ne sont même pas proches.

Tableau 6.9
Valeurs empiriques prévues pour l'estimateur naïf des flux bruts de la population

Période $t-1$	Période t		
	E1	**E2**	**E3**
E1	54 628	760	4 507
E2	1 506	2 079	1 175
E3	1 603	905	32 832

6.2.3 Probabilités d'inclusion inégales : *estimateur fondé sur le plan*

Dans un troisième scénario, nous avons considéré un plan d'échantillonnage qui induit des probabilités d'inclusion et des facteurs d'expansion inégaux. D'après ce scénario, les estimateurs proposés demeurent sans biais pour les flux bruts et pour les paramètres du modèle. Les biais relatifs sont présentés dans les tableaux 6.10, 6.11 et 6.12.

Tableau 6.10
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) de l'estimateur proposé pour les flux bruts de la population

Période $t-1$	Période t		
	E1	**E2**	**E3**
E1	-0,09 (0,8)	0,25 (3,6)	3,17 (7,9)
E2	0,72 (40,9)	-1,21 (27,2)	-4,62 (71,08)
E3	1,76 (20,4)	-3,19 (22,6)	-0,73 (7,2)

Tableau 6.11
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) des probabilités de transition p_{ij}

Période $t-1$	Période t		
	E1	**E2**	**E3**
E1	-0,05 (0,7)	0,115 (3,6)	0,47 (7,1)
E2	2,39 (36,0)	-0,13 (18,6)	-6,40 (69,1)
E3	1,15 (24,9)	-5,14 (21,7)	0,49 (3,7)

Tableau 6.12
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) des probabilités de classification initiales η_i

Période $t-1$		
η_1	η_2	η_3
-0,02 (1,1)	-0,70 (19,8)	1,13 (6,9)

Pour la probabilité de réponse ψ, le biais en pourcentage était -0,46 et l'erreur quadratique moyenne relative en pourcentage était de 0,6; pour la probabilité de réponse ρ_{RR}, le biais en pourcentage était de -0,21 et l'erreur quadratique moyenne relative en pourcentage était de 0,6; pour la probabilité de non-réponse ρ_{MM}, le biais en pourcentage était de 0,99 et l'erreur quadratique moyenne relative en pourcentage était de 1,8. Par ailleurs, le tableau 6.13 indique les valeurs prévues empiriques de l'estimateur proposé pour les flux bruts de la population, et ces valeurs sont très proches des valeurs fournies au tableau 6.1.

Tableau 6.13
Valeurs prévues empiriques de l'estimateur fondé sur le plan pour les flux bruts de la population

Période $t-1$	Période t		
	E1	**E2**	**E3**
E1	71 910	13 505	4 518
E2	1 523	2 972	470
E3	511	479	4 062

6.2.4 Probabilités d'inclusion inégales : *estimateur basé sur le modèle*

Un quatrième scénario considère un plan d'échantillonnage induisant des probabilités d'inclusion et des facteurs d'expansion inégaux de la même manière que le dernier scénario. Cependant, nous considérons les estimateurs qui ne tiennent pas compte du plan d'échantillonnage, mais seulement du modèle ξ. Conformément à ce scénario, les estimations sont biaisées pour les flux bruts et les paramètres du modèle, comme on peut le constater compte tenu des biais relatifs dans les tableaux 6.14, 6.15 et 6.16.

Tableau 6.14
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) de l'estimateur basé sur un modèle pour les flux bruts de la population

Période $t-1$	**Période** t		
	E1	**E2**	**E3**
E1	4,7 (6,1)	4,6 (8,9)	6,3 (10,5)
E2	**-89,0** (126,6)	**-89,5** (125,9)	**-88,4** (126,9)
E3	4,1 (23,8)	-3,7 (26,67)	5,3 (10,4)

Tableau 6.15
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) des probabilités de transition p_{ij}

Période $t-1$	**Période** t		
	E1	**E2**	**E3**
E1	0,03 (0,9)	-0,71 (4,1)	1,63 (8,6)
E2	2,77 (35,5)	-1,50 (19,6)	0,70 (70,6)
E3	4,00 (20,8)	**-14,6** (20,1)	1,33 (3,41)

Tableau 6.16
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) des probabilités initiales de classification η_i

Période $t-1$		
η_1	η_2	η_3
4,74 (6,48)	**-89,3** (126,7)	3,95 (11,9)

De même, pour la probabilité de réponse ψ, le biais relatif en pourcentage était de -0,77 et l'erreur quadratique moyenne relative était de 1,7; pour la probabilité de réponse ρ_{RR}, le biais relatif en pourcentage était de -0,53 et l'erreur quadratique moyenne relative était de 0,5; pour la probabilité de non-réponse ρ_{MM}, le biais relatif en pourcentage était de 0,11 et l'erreur quadratique moyenne relative était de 1,8. Par ailleurs, le tableau 6.17 montre les valeurs prévues empiriques pour l'estimateur fondé sur un modèle pour les flux bruts de la population (sans tenir compte du plan d'échantillonnage) et ces valeurs sont très loin des valeurs du tableau 6.1, en particulier pour la deuxième catégorie.

Tableau 6.17
Valeurs prévues empiriques pour l'estimateur basé sur un modèle pour les flux bruts de la population

Période $t-1$	**Période** t		
	E1	**E2**	**E3**
E1	75 438	14 039	4 790
E2	164	315	53
E3	540	443	4 213

6.2.5 Probabilités d'inclusion inégales : *estimateur naïf*

Dans un cinquième scénario, nous considérons un plan d'échantillonnage avec des probabilités d'inclusion et des facteurs d'expansion inégaux. Compte tenu de l'estimateur naïf, qui ne tient pas compte du plan d'échantillonnage ou du modèle du répondant, le tableau 6.18 montre le biais relatif pour chaque cas dans la matrice de flux bruts. Cet estimateur serait recommandable uniquement si la non-réponse était ignorable et que le plan d'échantillonnage correspondait à un plan d'échantillonnage aléatoire simple.

Tableau 6.18
Biais relatifs en pourcentage et erreurs quadratiques moyennes relatives en pourcentage (entre parenthèses) pour l'estimateur naïf des flux bruts de la population

Période $t-1$	**Période** t		
	E1	**E2**	**E3**
E1	-28,1 (34,7)	-27,6 (60,0)	-24,5 (41,1)
E2	**497,0** (629,2)	**570,2** (610,3)	**432,7** (686,2)
E3	**-40,5** (44,2)	**-37,0** (47,4)	**-33,0** (33,8)

Afin de faire une comparaison plus exacte, il serait possible de calculer les valeurs prévues des flux bruts et de les comparer au scénario actuel. Le tableau 6.19 montre les valeurs prévues empiriques pour l'estimateur naïf; comparativement aux valeurs prévues pour le modèle données dans le tableau 6.1, ces valeurs ne sont pas très proches et sont particulièrement médiocres pour les classifications de la deuxième catégorie.

Tableau 6.19
Valeurs prévues empiriques pour l'estimateur naïf des flux bruts de la population

Période $t-1$	**Période** t		
	E1	**E2**	**E3**
E1	51 755	9 849	3 297
E2	9 194	19 838	2 823
E3	279	295	2 665

7 Application réelle : estimation des flux bruts de la population pour l'enquête PME

La *Pesquisa Mensal de Emprego* (PME - enquête mensuelle brésilienne sur le travail) est une enquête qui fournit des indicateurs mensuels au sujet du marché du travail dans les principales régions métropolitaines du Brésil. Son principal objectif consiste à estimer la population active mensuelle et à évaluer les fluctuations et les tendances du marché du travail métropolitain. Il est également possible d'obtenir des indicateurs au sujet des effets de la conjoncture économique dans le marché du travail et de répondre à d'importants besoins en ce qui concerne la planification stratégique et le développement socio-

économique. Cette enquête est réalisée depuis 1980, certains changements méthodologiques ayant été apportés en 1982, 1988, 1993 et 2001 (IBGE 2007).

Cette section démontre l'utilisation des estimateurs proposés et présente les résultats finaux pour la PME. Nous allons considérer le panel P6 de cette enquête de novembre 2010 à février 2011, puis de novembre 2011 à février 2012. Pendant cette période d'observation, 21 374 interviews ont été réalisées auprès de différentes personnes. Nous avons choisi les deux premières mesures du panel (novembre et décembre 2010) afin de mettre en œuvre la procédure d'estimation proposée pour les flux bruts correspondants. En suivant un algorithme basé sur la bibliothèque TeachingSampling (Gutiérrez 2009), nous obtenons la classification au panel P6, pour les mois de novembre et de décembre 2010 donnés dans le tableau 7.1.

Tableau 7.1
Classification selon l'occupation et la réponse de l'échantillon pour le panel P6 de l'enquête PME

Novembre 2010	Décembre 2010				
	Employé	Chômeur	Inactif	Pas dans la population active	Complément de ligne
Employé	5 231	62	227	10	386
Chômeur	51	183	113	0	28
Inactif	235	93	4 200	12	281
Pas dans la population active	2	0	17	1 426	96
Complément de colonne	499	27	372	132	7 691

Cependant, étant donné que le panel P6 correspond à un échantillon probabiliste complexe des régions métropolitaines du Brésil, chaque personne dans le panel se représente elle-même, en plus d'autres personnes dans la population. Par conséquent, au moyen de la procédure d'estimation proposée dans cet article et des facteurs d'expansion correspondants de l'enquête, nous constatons que les valeurs estimées de la population pour le panel P6 correspondent à celles obtenues dans le tableau 7.2.

Tableau 7.2
Tableau de la contingence estimée pour la population indiquant le niveau d'activité et la non-réponse aux deux mesures considérées pour le panel P6 de l'enquête PME

Novembre 2010	Décembre 2010				
	Employé	Chômeur	Inactif	Pas dans la population active	Complément de ligne
Employé	2 162 635	20 602	76 303	3 074	160 768
Chômeur	16 233	80 169	37 786	0	11 504
Inactif	70 551	31 822	1 707 675	6 018	122 412
Pas dans la population active	958	0	7 035	566 530	38 171
Complément de colonne	205 033	9 293	136 146	53 640	3 076 388

Au moyen de la procédure d'estimation proposée dans cet article, nous avons calculé les flux bruts estimés de la population donnés dans le tableau 7.3. Les estimateurs correspondants sont non biaisés dans le contexte du plan complexe de l'enquête PME. Par conséquent, le nombre de personnes occupées pendant les deux périodes de mesure est estimé à 3 913 274, tandis que le nombre de personnes inactives pendant les deux périodes est estimé à 3 035 463.

Tableau 7.3
Flux bruts de la population estimés pour les deux périodes dans le cadre de l'enquête PME. Les coefficients de variation estimés, en pourcentage, sont entre parenthèses

Novembre 2010	Décembre 2010			
	Employé	Chômeur	Inactif	Pas dans la population active
Employé	3 913 274 (0,2)	36 570 (3,1)	136 102 (1,6)	5 573 (7,2)
Chômeur	29 776 (3,5)	144 253 (1,7)	68 320 (2,1)	0 (-)
Inactif	127 193 (1,6)	56 296 (2,3)	3 035 463 (0,3)	10 872 (6,5)
Pas dans la population active	1 727 (17,3)	0 (-)	12 496 (5,8)	1 022 836 (0,5)

Les estimations dans le dernier tableau plus haut découlent de la procédure d'estimation proposée dans cet article. Ensuite, nous examinons les paramètres estimés au premier degré du modèle, définis comme les probabilités de transition d'une catégorie à une autre pendant les deux périodes d'observation.

Tableau 7.4
Estimation des probabilités p_{ij} . Les coefficients de variation estimés, en pourcentage, sont entre parenthèses

Novembre 2010	Décembre 2010			
	Employé	Chômeur	Inactif	Pas dans la population active
Employé	0,9564 (0,1)	0,0089 (3,1)	0,0332 (1,6)	0,0013 (7,2)
Chômeur	0,1228 (3,4)	0,5952 (1,1)	0,2819 (2,0)	0 (-)
Inactif	0,0393 (1,5)	0,0174 (2,3)	0,9398 (0,1)	0,0033 (6,5)
Pas dans la population active	0,0016 (17,6)	0 (-)	0,0120 (5,8)	0,9862 (0,1)

Les probabilités initiales de classification pendant la première période d'intérêt sont indiquées dans le tableau 7.5. On peut remarquer que, pour cette enquête en particulier, les plus fortes probabilités de classification se trouvent pour les catégories de personnes occupées et inactives.

Tableau 7.5
Estimation des probabilités η_i . Les coefficients de variation estimés, en pourcentage, sont entre parenthèses

November 2010			
η_1	η_2	η_3	η_4
0,4757 (0,2)	0,0281 (1,2)	0,3755 (0,3)	0,1205 (0,5)

Enfin, la probabilité de réponse générale a été estimée à $\hat{\psi}_{mpv} = 0,595$ (avec un coefficient de variation de 0,1 %). Autrement dit, le taux de réponse se situe autour de 60 %. De plus, la probabilité de transition qu'un non-répondant à la première période soit encore un non-répondant la prochaine fois a été estimée à $\hat{\rho}_{MM,mpv} = 0,883$ (avec un coefficient de variation de 0,1 %). La probabilité de transition qu'un répondant à la première période soit encore un répondant par la suite a été estimée comme suit : $\hat{\rho}_{RR,mpv} = 0,934$ (avec un coefficient de variation de 0,1 %). D'une manière générale, il est possible de déclarer que l'état de réponse d'une personne pendant la première période ne change pas de façon significative la deuxième période.

8 Conclusions

Cet article a examiné un problème fréquent d'applications de l'échantillonnage. Au moyen des modèles en chaîne de superpopulation de Markov, une nouvelle méthodologie a été proposée, entraînant des estimateurs à peu près sans biais des flux bruts à différents moments pour le cas particulier des données provenant d'enquêtes complexes avec des poids d'échantillonnage inégaux. Les applications possibles de la méthodologie dans le présent article sont larges, notamment dans le cas des bureaux de statistique nationaux envisageant des enquêtes complexes. Les enquêtes sur la qualité de vie ou sur la population active s'intéressent habituellement à l'estimation des flux bruts. Toutefois, les extensions possibles de cette méthodologie pourraient être appliquées au secteur de la politique publique pour les évaluations d'impacts ayant une classification des répondants avant et après une intervention.

De plus, nous présentons une solution à un problème général, comme la non-réponse non ignorable. Des modèles où la non-réponse n'est pas différenciée pendant différentes périodes ou selon l'état de classification ont été envisagés. Cependant, dans certaines applications pratiques, il est possible que ce ne soit pas le cas.

L'approche de cet article considère que les poids déterminés par le plan d'échantillonnage pour les unités entre les deux périodes sont les mêmes. Dans le cadre de travaux plus poussés, on s'efforcera de considérer différents poids entre les vagues en envisageant une classification d'échantillonnage à deux phases ou une approche de calage sur marges à deux degrés. En effet, il serait intéressant de comparer le rendement de la méthodologie donné dans cet article à la méthode du calage sur marges. On pourrait considérer l'approche d'Ash (2005) et de Sikkel, Hox et de Leeuw (2008) pour calibrer en deux périodes, ainsi que l'approche de Särndal et Lundström (2005) pour traiter la non-réponse.

Des travaux plus poussés chercheront à élargir cette méthodologie pour des modèles en chaîne de Markov plus complexes afin de considérer différents poids d'échantillonnage. Une nouvelle définition des paramètres du modèle sera nécessaire. De plus, cette méthodologie pourrait être appliquée au cas des flux bruts dans plus de deux périodes lorsque les erreurs de classification sont prises en compte.

Remerciements

Les auteurs souhaitent remercier deux réviseurs anonymes de leurs commentaires constructifs au sujet d'une version précédente de l'article, qui ont donné lieu à la présente version améliorée. De plus, le

premier auteur tient à remercier l'Universidad Santo Tomas de son soutien financier pendant ses études doctorales. Cet article est le fruit de la thèse de doctorat d'Andrés Gutiérrez de l'Universidad Nacional de Colombia, sous la supervision des deux autres auteurs.

Annexe

A.1 Preuves mathématiques des résultats de l'article

Dans cette section, les preuves mathématiques de certains des résultats les plus importants de l'article sont incluses.

Preuve du résultat 4.1

Preuve. En prenant le logarithme de la fonction de vraisemblance, et en le définissant comme l, il s'ensuit que

$$\begin{aligned} l_U &= \ln\left(L_U\right) \\ &= \sum_i \sum_j N_{ij} \ln\left(\psi \rho_{RR} \eta_i p_{ij}\right) + \sum_i R_i \ln\left(\sum_j \psi\left(1-\rho_{RR}\right)\eta_i p_{ij}\right) \\ &\quad + \sum_j C_j \ln\left(\sum_i (1-\psi)(1-\rho_{MM})\eta_i p_{ij}\right) + M \ln\left(\sum_i \sum_j (1-\psi)\rho_{MM}\eta_i p_{ij}\right). \end{aligned}$$

Notons que $N_{ij} = \sum_{k\in U} y_{1ik} y_{2jk}$, $R_i = \sum_{k\in U} y_{1ik}\left(1-z_{2k}\right)$, $C_j = \sum_{k\in U} y_{2jk}\left(1-z_{1k}\right)$ et $M = \sum_{k\in U}\left(1-z_{1k}\right)\left(1-z_{2k}\right)$. Après avoir pris en compte la somme de la population totale, le résultat est finalement obtenu.

Preuve du résultat 4.2

Preuve. En commençant par la définition de la pseudo-vraisemblance et en tenant compte des hypothèses du modèle, il s'ensuit que

$$\begin{aligned} l_U &= \sum_{k\in U}\Bigg[\sum_i \sum_j y_{1ik} y_{2jk}\left[\ln(\psi) + \ln(\rho_{RR}) + \ln(\eta_i) + \ln(p_{ij})\right] \\ &\quad + \sum_i y_{1ik}\left(1-z_{2k}\right)\left[\ln(\psi) + \ln(1-\rho_{RR}) + \ln(\eta_i) + \ln\left(\sum_j p_{ij}\right)\right] \\ &\quad + \sum_j y_{2jk}\left(1-z_{1k}\right)\left[\ln(1-\rho_{MM}) + \ln(1-\psi) + \ln\left(\sum_i \eta_i p_{ij}\right)\right] \\ &\quad + \left(1-z_{1k}\right)\left(1-z_{2k}\right)\left[\ln(1-\psi) + \ln(\rho_{MM}) + \ln\left(\sum_i \sum_j \eta_i p_{ij}\right)\right]\Bigg] \\ &= \sum_{k\in U} f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right). \end{aligned}$$

Le *score* pour ψ peut être défini comme suit :

$$u_k(\psi) = \frac{\partial f_k(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2)}{\partial \psi}$$
$$= \frac{(1-\psi)\left(\sum_i \sum_j y_{1ik} y_{2jk} + \sum_i y_{1ik}(1-z_{2k})\right) - \psi\left(\sum_j y_{2jk}(1-z_{1k}) + (1-z_{1k})(1-z_{2k})\right)}{\psi(1-\psi)}.$$

Alors, pour ce paramètre, les équations de pseudo-vraisemblance sont données par

$$\sum_{k \in S} w_k u_k(\psi) = 0.$$

Pour la solution de ψ, on constate que

$$\hat{\psi}_{mpv} = \frac{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_{i\,i} + \sum_j \hat{C}_j + \hat{M}}.$$

Au moyen d'un processus analogue pour les paramètres restants, le résultat est obtenu.

Preuve du résultat 4.3

Preuve. D'abord, il faut savoir que l'estimation pour ces paramètres est assujettie aux restrictions $\sum_i \eta_i = 1$ et $\sum_j p_{ij} = 1$. Alors, le processus doit tenir compte de l'utilisation des multiplicateurs de Lagrange. La fonction à maximiser, y compris ces restrictions, peut être exprimée comme suit :

$$l_U + \lambda_1\left(\sum_i \eta_i - 1\right) + \lambda_2\left(\sum_j p_{ij} - 1\right).$$

Alors, le *score* correspondant pour η_i est défini par

$$u_k(\eta_i) = \frac{\partial f_k(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2)}{\partial \eta_i} + \frac{\partial \lambda_1\left(\sum_i \eta_i - 1\right)}{\partial \eta_i}$$
$$= \frac{\sum_j y_{1ik} y_{2jk} + y_{1ik}(1-z_{2k})}{\eta_i} + \sum_j y_{2jk}(1-z_{1k})\frac{p_{ij}}{\sum_i \eta_i p_{ij}} + (1-z_{1k})(1-z_{2k}) + \lambda_1.$$

La dernière étape tient compte des restrictions, puisque $\sum_i \sum_j \eta_i p_{ij} = \sum_i \eta_i \sum_j p_{ij} = \sum_i \eta_i = 1$. Alors, pour ce paramètre, les équations de pseudo-vraisemblance sont données par

$$\sum_{k \in S} w_k u_k(\eta_i) = 0.$$

Alors, après un peu d'algèbre, il s'ensuit que

$$\eta_i = \frac{\sum_j \sum_s w_k y_{1ik} y_{2jk} + \sum_s w_k y_{1ik}(1-z_{2k}) + \sum_j \sum_s w_k y_{2jk}(1-z_{1k})\left(\eta_i p_{ij} / \sum_i \eta_i p_{ij}\right)}{-\sum_s w_k (1-z_{1k})(1-z_{2k}) - \lambda_1 \sum_s w_k}.$$

Par ailleurs, en utilisant la restriction $\sum_i \eta_i = 1$ et en faisant la somme par rapport à i, il s'ensuit que

$$\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j = \left(-\sum_s w_k \left(1 - z_{1k}\right)\left(1 - z_{2k}\right) - \lambda_1 \sum_s w_k \right).$$

Alors, nous obtenons enfin que

$$\eta_i = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\hat{C}_j \eta_i p_{ij} \Big/ \sum_i \eta_i p_{ij} \right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j}.$$

Par ailleurs, afin de trouver l'estimateur du maximum de pseudo-vraisemblance de $\left\{ p_{ij} \right\}$, le *score* pour p_{ij} est défini comme suit :

$$\begin{aligned} u_k\left(p_{ij}\right) &= \frac{\partial f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right)}{\partial p_{ij}} + \frac{\partial \lambda_2 \left(\sum_i p_{ij} - 1 \right)}{\partial p_{ij}} \\ &= \frac{y_{1ik} y_{2jk}}{p_{ij}} + y_{1ik}\left(1 - z_{2k}\right) + y_{2jk}\left(1 - z_{1k}\right) \frac{\eta_i}{\sum_i \eta_i p_{ij}} + \left(1 - z_{1k}\right)\left(1 - z_{2k}\right)\eta_i + \lambda_2. \end{aligned}$$

Par conséquent,

$$p_{ij} = \frac{\sum_s w_k y_{1ik} y_{2jk} + \sum_s w_k y_{2jk} \left(1 - z_{1k}\right) p_{ij} \eta_i \Big/ \sum_i \eta_i p_{ij}}{-\sum_s w_k y_{1ik}\left(1 - z_{2k}\right) - \sum_s w_k \left(1 - z_{1k}\right)\left(1 - z_{2k}\right)\eta_i - \sum_s w_k \lambda_2}.$$

En utilisant la restriction $\sum_j p_{ij} = 1$ et en faisant la somme par rapport à j des deux côtés, il s'ensuit que

$$\begin{aligned} &\sum_j \hat{N}_{ij} + \sum_j \hat{C}_j \frac{p_{ij}\eta_i}{\sum_i \eta_i p_{ij}} \\ &= \left(-\sum_s w_k y_{1ik}\left(1 - z_{2k}\right) - \sum_s w_k \left(1 - z_{1k}\right)\left(1 - z_{2k}\right)\eta_i - \sum_s w_k \lambda_2 \right). \end{aligned}$$

Alors, il s'ensuit que

$$p_{ij} = \frac{\hat{N}_{ij} + \left(\hat{C}_j \eta_i p_{ij} \Big/ \sum_i \eta_i p_{ij} \right)}{\sum_j \hat{N}_{ij} + \sum_j \left(\hat{C}_j \eta_i p_{ij} \Big/ \sum_i \eta_i p_{ij} \right)}.$$

Maintenant, soulignons qu'il est impossible de résoudre la dernière expression pour $\left\{ p_{ij} \right\}$ de façon à ce que la solution soit une expression fermée. Il en va de même en ce qui concerne l'expression pour $\left\{ \eta_i \right\}$. Cependant, il est possible d'utiliser une approche itérative, qui s'est avérée avoir une convergence rapide des problèmes d'estimation du maximum de vraisemblance pour les tableaux de contingence. Cette approche présume que l'estimateur du maximum de pseudo-vraisemblance peut se trouver après une itération conjointe des expressions suivantes à l'étape $\left(v + 1\right)$, pour $v \geq 1$,

$$\hat{\eta}_{i,mpv}^{(v+1)} = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j}$$

$$\hat{p}_{ij,mpv}^{(v+1)} = \frac{\hat{N}_{ij} + \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_j \hat{N}_{ij} + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}.$$

Cette procédure itérative particulière a été utilisée au départ pour la formulation de modèles de vraisemblance imbriqués de Hocking et Oxspring (1971). Toutefois, elle semble également avoir été mise en œuvre par Blumenthal (1968), Reinfurt (1970), Chen et Fienberg (1974), Fienberg et Stasny (1983), Stasny (1987), Stasny (1988) et d'autres.

Preuve du résultat 5.5

Preuve. L'estimateur non linéaire $\hat{\psi}_{mpv}$, peut être exprimé comme une fonction des totaux estimés $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$ et $\hat{M}$ (où $i, j = 1, \ldots, G$). Alors,

$$\hat{\psi}_{mpv} = f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right).$$

Enfin, l'approximation du premier degré de Taylor au point $\left(\hat{N}_{ij} = N_{ij}, \hat{R}_i = R_i, \hat{C}_j = C_j, \hat{M} = M \right)$ est donnée par

$$\hat{\psi}_{mpv} = \psi_U + a_1 \sum_i \sum_j \left(\hat{N}_{ij} - N_{ij} \right) + a_1 \sum_i \left(\hat{R}_i - R_i \right) + a_2 \sum_j \left(\hat{C}_j - C_j \right) + a_2 \left(\hat{M} - M \right)$$

où

$$a_1 = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{R}_i} \right|_{\substack{\hat{N}_{ij} = N_{ij} \\ \hat{R}_i = R_i \\ \hat{C}_j = C_j \\ \hat{M} = M}} = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{N}_{ij}} \right|_{\substack{\hat{N}_{ij} = N_{ij} \\ \hat{R}_i = R_i \\ \hat{C}_j = C_j \\ \hat{M} = M}} = \frac{\sum_j C_j + M}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M \right)^2}$$

et

$$a_2 = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{C}_j} \right|_{\substack{\hat{N}_{ij} = N_{ij} \\ \hat{R}_i = R_i \\ \hat{C}_j = C_j \\ \hat{M} = M}} = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{M}} \right|_{\substack{\hat{N}_{ij} = N_{ij} \\ \hat{R}_i = R_i \\ \hat{C}_j = C_j \\ \hat{M} = M}} = -\frac{\sum_i \sum_j N_{ij} + \sum_i R_i}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M \right)^2}.$$

Preuve du résultat 5.8

Preuve. Pour calculer la valeur prévue conformément au plan d'échantillonnage, il s'ensuit que

$$\begin{aligned} AE_p\left(\hat{\psi}_{mpv}\right) &\cong E_p\left(\hat{\psi}_0\right) \\ &= \psi_U + a_1 \sum_i \sum_j \left(E_p\left(\hat{N}_{ij}\right) - N_{ij}\right) + a_1 \sum_i \left(E_p\left(\hat{R}_i\right) - R_i\right) \\ &\qquad + a_2 \sum_j \left(E_p\left(\hat{C}_j\right) - C_j\right) + a_2\left(E_p\left(\hat{M}\right) - M\right) \\ &= \psi_U . \end{aligned}$$

En suivant un processus semblable pour les estimateurs restants, on obtient le résultat. Cette preuve découle de l'application de la méthode de pseudo-vraisemblance qui induit les estimations sans biais pour les paramètres de population dans le modèle comme le prouve le corollaire 1 de Binder (1983, p. 291).

Preuve du résultat 5.10

Preuve. En supposant $\hat{\psi}_{mpv}$, en remplaçant les expressions pour $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$, $\hat{M}$ et en faisant quelques simplifications algébriques, on peut exprimer la variance approximative comme suit :

$$AV\left(\hat{\psi}_{mpv}\right) = Var\left(a_1 \sum_i \sum_j \hat{N}_{ij} + a_1 \sum_i \hat{R}_i + a_2 \sum_j \hat{C}_j + a_2 \hat{M}\right) = Var\left(\sum_{k \in S} \frac{E_k^{\psi}}{\pi_k}\right).$$

Initialement, nous avons

$$E_k^{\psi} = a_1 \sum_i \sum_j y_{1ik} y_{2jk} + a_1 \sum_i y_{1ik}\left(1 - z_{2k}\right) + a_2 \sum_j y_{2jk}\left(1 - z_{1k}\right) + a_2\left(1 - z_{1k}\right)\left(1 - z_{2k}\right).$$

Alors, sachant que $\sum_i \sum_j y_{1ik} y_{2jk} = \sum_i y_{1ik} = \sum_j y_{2jk} = 1$ et après un peu d'algèbre, il s'ensuit que

$$E_k^{\psi} = a_1\left(2 - z_{2k}\right) + a_2\left(1 - z_{1k}\right)\left(2 - z_{2k}\right).$$

Après un processus analogue pour $\hat{\rho}_{RR,mpv}$ et $\hat{\rho}_{MM,mpv}$, les autres expressions de la variance dans ce résultat sont obtenues.

Preuve du résultat 5.12

Preuve. On obtient la preuve en suivant l'expression (3.3) de Binder (1983) et en tenant compte de ce qui suit

$$J_{\eta_i} = \frac{\partial \sum_U u_k\left(\eta_i\right)}{\partial \eta_i}$$

$$J_{p_{ij}} = \frac{\partial \sum_U u_k\left(p_{ij}\right)}{\partial p_{ij}}.$$

De plus,

$$\frac{\partial u_k\left(\eta_i\right)}{\partial \eta_i} = -\frac{2y_{1ik} - y_{1ik} z_{2k}}{\eta_i^2} - \left(1 - z_{1k}\right) \sum_j \frac{y_{2jk} p_{ij}^2}{\left(\sum_i \eta_i p_{ij}\right)^2}$$

$$\frac{\partial u_k\left(p_{ij}\right)}{\partial p_{ij}} = -\frac{y_{1ik} y_{2jk}}{p_{ij}^2} - \frac{\eta_i^2}{\left(\sum_i \eta_i p_{ij}\right)^2} y_{2jk}\left(1 - z_{1k}\right).$$

Preuve du résultat 5.16

Preuve.

$$\begin{aligned} AV_p\left(\hat{\mu}_{ij,mpv}\right) &= a_7^2 Var_p\left(\hat{N}_{ij}\right) + a_8^2 AV_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 AV_p\left(\hat{p}_{ij}\right) \\ &\quad + 2a_7 a_8 Cov\left(\hat{N}_{ij}, \hat{\eta}_{i,mpv}\right) + 2a_7 a_9 Cov\left(\hat{N}_{ij}, \hat{p}_{ij}\right) 2a_8 a_9 Cov\left(\hat{\eta}_{i,mpv}, \hat{p}_{ij}\right) \\ &\cong a_7^2 Var_p\left(\hat{N}_{ij}\right) + a_8^2 AV_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 AV_p\left(\hat{p}_{ij}\right). \end{aligned}$$

Parce que

$$\begin{aligned} Cov\left(\hat{N}_{ij}, \hat{\eta}_{i,mpv}\right) &= E_p\left(\hat{N}_{ij}\hat{\eta}_{i,mpv}\right) - E_p\left(\hat{N}_{ij}\right) E_p\left(\hat{\eta}_{i,mpv}\right) \\ &\cong \hat{N}_{ij,U}\eta_{i,U} - \hat{N}_{ij,U}\eta_{i,U} = 0. \end{aligned}$$

Alors, il est possible d'obtenir ce qui suit :

$$E_p\left(\hat{N}_{ij}\hat{\eta}_{i,mpv}\right) \cong \hat{N}_{ij,U}, \eta_{i,U}$$

au moyen de la linéarisation de Taylor pour $\left(\hat{N}_{ij,U}, \eta_{i,U}\right)$. Les autres covariances sont obtenues de façon semblable.

Bibliographie

Ash, S. (2005). Calibration weights for estimators of longitudinal data with an application to the National Long Term Care Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*. American Statistical Association: Alexandria, VA, 2694–2699.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, 63, 542-551.

Cassel, C.M., Särndal, C.E. et Wretman, J.H. (1976). *Foundations of Inference in Survey Sampling*. New York: Wiley.

Chambers, R.L. et Skinner, C.J. (2003). *Analysis of Survey Data*. John Wiley and Sons, Chichester: UK.

Chen, T. et Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Clogg, C.C. et Eliason, S.R. (1987). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16, 8-44.

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

Fienberg, S.E. et Stasny, E.A. (1983). L'estimation des flux bruts mensuels de l'activité sur le marché du travail. *Techniques d'enquête*, 9(1), 85-110.

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

Gambino, J.G. et Silva, P.L. (2009). Sampling and estimation in household surveys. Dans D. Pfeffermann et C.R. Rao (Eds.), *Handbook of Statistics*. Vol. 29A. Sample Surveys: Design, Methods and Applications (pp. 407-439). Amsterdam: Elsevier.

Gutiérrez, H.A. (2009). TeachingSampling: Sampling designs and parameter estimation in finite population. R package version 2.0.1.

Hocking, R.R. et Oxspring, H.H. (1971). Maximun likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66, 65-70.

IBGE (2007). *Pesquisa Mensal de Emprego*. Vol. 23, 2nd edition.

Kalton, G. (2009). Designs for surveys over time. Dans D. Pfeffermann and C.R. Rao (Eds.), *Handbook of Statistics*. Vol. 29A. Sample Surveys: Design, Methods and Applications (pp. 89-108). Amsterdam: Elsevier.

Kim, J. K. et Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78, 21-39.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Lu, Y. et Lohr, S. (2010). L'estimation des flux bruts dans les enquêtes à base de sondage double. *Techniques d'enquête*, 36(1), 13-24.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R*. New York: Wiley.

Pessoa, D.G.C. et Silva, P.L. (1998). *Análise de Dados Amostrais Complexos*. São Paulo : Associação Brasileira de Estatística.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

R Development Core Team (2012). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0.

Rao, J.N.K. et Thomas, D.R. (1988). The analysis of cross-classified data from complex surveys. *Sociological Methodology*, 18, 213-269.

Reinfurt, D.W. (1970). The analyis of categorical data with supplemented margins including applications to mixed models. Thèse de doctorat non publiée. Department of Biostatistics. University of North Carolina.

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

Särndal, C.E. et Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* John Wiley and Sons, Chichester: UK.

Särndal, C.E. et Lundström, S. (2010). Plan d'estimation : détermination de vecteurs auxiliaires en vue de réduire le biais de non-réponse. *Techniques d'enquête*, 36(2), 141-156.

Sikkel, D., Hox, J. et de Leeuw, E. (2008). Using auxiliary data for adjustment in longitudinal research. Dans P. Lynn (Ed), *Methodology of longitudinal surveys*. New York: Wiley. Une version antérieure est disponible au http://www.iser.essex.ac.uk/ulsc/mols2006/programme/data/papers/Sikkel.pdf

Skinner, C.J. et Vallet, L.A. (2010). Fitting log-linear models to contingency tables from surveys with complex sampling designs: An investigation of the Clogg-Eliason approach. *Sociological Methods and Research*, 39, 83-108.

Stasny, E.A. (1987). Some Markov-chain models for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 3, 359-373.

Stasny, E.A. (1988). Modeling nonignorable nonresponse in categorical panel data with an example in estimating gross labor-flows. *Journal of Business and Economic Statistics*, 6, 207-219.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 353-366
Statistique Canada, n° 12-001-X au catalogue

Tests du khi-carré dans les enquêtes à base de sondage double

Yan Lu[1]

Résumé

Afin d'obtenir une meilleure couverture de la population d'intérêt et de réduire les coûts, un certain nombre d'enquêtes s'appuient sur un plan à base de sondage double, suivant lequel des échantillons indépendants sont tirés de deux bases de sondage chevauchantes. La présente étude porte sur les tests du khi-carré dans les enquêtes à base de sondage double en présence de données catégoriques. Nous étendons le test de Wald généralisé (Wald 1943), ainsi que les tests avec correction d'ordre un et correction d'ordre deux de Rao-Scott (Rao et Scott 1981) pour passer d'une enquête à base de sondage unique à une enquête à base de sondage double, et nous déterminons les distributions asymptotiques. Des simulations montrent que les deux tests avec correction de type Rao-Scott donnent de bons résultats, et il est donc recommandé de les utiliser dans les enquêtes à base de sondage double. Un exemple sert à illustrer l'utilisation des tests élaborés.

Mots-clés : Propriétés asymptotiques; tests du khi-carré; enquêtes à base de sondage double; test avec correction d'ordre un; test avec correction d'ordre deux; simulations.

1 Introduction

Une situation générale d'enquête à base de sondage double est illustrée à la figure 1.1, dans laquelle l'union de la base de sondage A et de la base de sondage B est désignée comme étant l'union de trois domaines non chevauchants, c'est-à-dire $A \cup B = a \cup ab \cup b$. Des échantillons probabilistes sont sélectionnés indépendamment de ces deux bases de sondage.

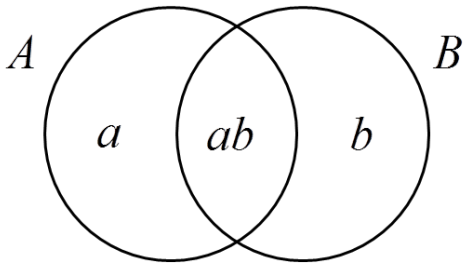

Figure 1.1 : Les bases de sondage A et B sont toutes deux incomplètes, mais chevauchantes

Une enquête à base de sondage double donne souvent une meilleure couverture de la population et permet de réaliser des économies importantes. Plusieurs méthodes d'analyse transversale des données issues d'enquêtes à base de sondage double sont décrites dans la littérature statistique; voir Hartley (1962, 1974), Fuller et Burmeister (1972), Skinner (1991), Skinner et Rao (1996), Lohr et Rao (2000, 2006), etc. Comme l'ont souligné Rao et Thomas (1988), il est souvent nécessaire de procéder à des analyses statistiques de données catégoriques en recherche sociologique quantitative. Le test du khi-carré de Pearson et le test du rapport de vraisemblance sont tous deux d'usage répandu lorsque l'on a affaire à des

1. Yan Lu, Département de mathématiques et de statistique, Université du Nouveau-Mexique, Albuquerque, NM, 87131-0001. Courriel : luyan@math.unm.edu.

données catégoriques. Ces méthodes reposent sur l'hypothèse que les données sont obtenues par échantillonnage aléatoire simple (EAS) d'une ou de plusieurs grandes populations. À l'heure actuelle, la plupart des enquêtes sont réalisées selon des plans de sondage complexes avec stratification et mise en grappes qui violent l'hypothèse d'EAS. Le test de Wald (Wald 1943) est l'une des méthodes les plus anciennes avancées pour évaluer la qualité de l'ajustement du modèle dans les plans de sondage complexes. Fay (1979, 1985) a proposé un test du khi-carré ajusté par la méthode du jackknife destiné à être utilisé dans les enquêtes complexes. Tant la méthode de Wald (1943) que celle de Fay (1979) nécessitent de posséder de l'information détaillée sur l'enquête pour pouvoir estimer la matrice de covariance. Or, cette information détaillée fait souvent défaut en pratique. Rao et Scott (1981, 1984) ont proposé des tests du khi-carré de la qualité de l'ajustement et de l'indépendance dans des tableaux à double entrée et à entrées multiples. Bedrick (1983), ainsi que Rao et Scott (1987) ont également étudié l'utilisation d'information limitée sur les effets de plan dans les cellules et dans les marges pour fournir des tests approximatifs. Thomas, Singh et Roberts (1996) ont décrit une étude Monte Carlo portant sur des procédures élaborées pour tester l'indépendance dans un tableau à double entrée.

Le problème de recherche examiné dans le présent article émane de l'analyse des données catégoriques dans les enquêtes à base de sondage double. Par exemple, une base de sondage double peut être constituée des répertoires en ligne des membres de l'*American Statistical Association* (ASA) et de l'*Institute for Mathematical Statistics* (IMS). Le domaine de chevauchement comprend les statisticiens qui sont membres des deux sociétés. On pourrait vouloir tester si le pourcentage de femmes en milieu universitaire est le même dans les trois domaines (domaine a: membres de l'ASA seulement; domaine ab: membres de l'ASA ainsi que de l'IMS; domaine b: membres de l'IMS seulement). Dans le contexte d'une enquête à base de sondage double, les tests présentent plus de difficultés que dans le cas d'une enquête à base de sondage unique, parce qu'il existe deux échantillons, pouvant chacun résulter d'un plan d'échantillonnage complexe et pouvant présenter un degré inconnu de chevauchement. Il est possible d'appliquer une constante de pondération fixe au domaine de chevauchement, disons 1/2, et de considérer l'union de l'échantillon A $(\mathcal{S}_A)$ et de l'échantillon B $(\mathcal{S}_B)$ comme un échantillon unique. Les tests du khi-carré pour base de sondage unique décrits dans la littérature, comme ceux de Rao et Scott (1981), peuvent alors être appliqués. Cependant, cette application est fondée sur l'hypothèse qu'un jeu de proportions dans les cellules ultimes existe pour le plan à base de sondage double, ce qui n'est pas nécessairement vrai. Dans le présent article, nous supposons que chaque domaine possède son propre jeu de proportions de cellule, l'estimateur de type Rao-Scott (1981) étant alors un cas particulier quand les jeux de proportions de cellule sont les mêmes dans les trois domaines. Nous étendons le test de Wald (1943) et les tests avec correction d'ordre un et correction d'ordre deux de Rao-Scott (Rao et Scott 1981) pour passer d'une enquête simple à une enquête à base de sondage double, et nous établissons les distributions asymptotiques.

La présentation de l'article est la suivante. À la section 2, nous donnons le contexte de l'étude. À la section 3, nous proposons plusieurs tests du khi-carré. À la section 4, nous présentons une petite étude de simulation des tests du khi-carré proposés sous une hypothèse simple. À la section 5, nous décrivons une étude basée sur un exemple réel. Enfin, nous présentons un résumé à la section 6.

2 Contexte

2.1 Tests du khi-carré dans une enquête à base de sondage unique

Considérons un tableau de fréquences à simple entrée avec k classes et les proportions de population finie associées $p_1, p_2, \cdots, p_k$ avec $\sum_{i=1}^{i=k} p_i = 1$. Soit $n_1, \cdots, n_k$ les fréquences de cellule observées dans un échantillon tombant dans chacune des k catégories avec $\sum_{i=1}^{k} n_i = n$. Sous un EAS, la statistique du khi-carré de Pearson pour vérifier l'hypothèse simple $H_0 : p_i = p_{0i}$, $(i = 1, \cdots, k)$ est donnée par

$$\tilde{X}^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_{0i}\right)^2}{np_{0i}}. \tag{2.1}$$

Pour les plans de sondage compliqués, $\tilde{X}^2$ fait intervenir des distributions non centrées. Il est naturel de considérer une statistique plus générale

$$X^2 = n\sum_{i=1}^{k} \frac{\left(\hat{p}_i - p_{0i}\right)^2}{p_{0i}}, \tag{2.2}$$

où $\hat{p}_i$ est un estimateur convergent de p_i sous un plan de sondage spécifié $p(s)$.

Soit $\hat{\mathbf{p}} = \left(\hat{p}_1, \cdots, \hat{p}_{k-1}\right)'$ le vecteur de dimension $k-1$ des proportions estimées avec $\hat{p}_k = 1 - \left(\hat{p}_1 + \cdots + \hat{p}_{k-1}\right)$; $\mathbf{p}_0$, le vecteur de dimension $k-1$ des proportions hypothétiques; $\mathbf{V}$, la matrice de covariance de dimensions $(k-1)\times(k-1)$ de $\hat{\mathbf{p}}$, et $\hat{\mathbf{V}}$, l'estimation de $\mathbf{V}$ obtenue à partir des données d'enquête. La statistique de Wald généralisée

$$X_W^2 = \left(\hat{\mathbf{p}} - \hat{\mathbf{p}}_0\right)' \hat{\mathbf{V}}^{-1} \left(\hat{\mathbf{p}} - \hat{\mathbf{p}}_0\right), \tag{2.3}$$

suit asymptotiquement la distribution de χ_{k-1}^2 sous $H_0 : p_i = p_{0i}$, $(i = 1, \cdots, k)$ pour une valeur suffisamment grande de n.

Rao et Scott (1981) ont montré que, sous H_0, X^2 en (2.2) suit asymptotiquement la distribution d'une somme pondérée $\delta_1 W_1 + \cdots + \delta_{k-1} W_{k-1}$ de $k-1$ variables aléatoires χ_1^2, W_i, $i = 1, 2, \cdots, k-1$. Les δ_i sont les valeurs propres d'une matrice d'effets de plan $\mathbf{P}^{-1}\mathbf{V}$, où $\mathbf{P}$ est la matrice de covariance correspondant à l'EAS quand H_0 est vraie, c.-à-d. $\mathbf{P} = n^{-1}\left(diag\left(\mathbf{p}_0\right) - \mathbf{p}_0\mathbf{p}_0'\right)$. Le résultat classique du test de Pearson se retrouve sous EAS. Soit $\hat{\delta}_i$ une estimation de δ_i et $\hat{\delta}. = \left(\sum_{i=1}^{k-1} \hat{\delta}_i\right) \Big/ \left(k-1\right)$, le test avec correction d'ordre un de Rao-Scott rapporte $X^2 \big/ \hat{\delta}.$ à χ_{k-1}^2. Lorsque la matrice de covariance entièrement estimée $\hat{\mathbf{V}}$ est connue, une meilleure approximation de la distribution asymptotique de X^2 consiste à faire correspondre le premier moment et le deuxième moment de la statistique de test à une distribution du χ^2. La statistique de test avec correction d'ordre deux de Rao-Scott (Rao et Scott 1981) considère $X_S^2 = X^2 \Big/ \left[\hat{\delta}.\left(1 + \hat{a}^2\right)\right]$. Cette statistique est approximativement une variable aléatoire khi-carré à $v = \left(k-1\right) \Big/ \left(1 + a^2\right)$ degrés de liberté, où $\hat{a}$ est une estimation de a avec $\hat{a}^2 = \sum_{i=1}^{k-1} \hat{\delta}_i^2 \Big/ \left[\left(k-1\right)\hat{\delta}.^2\right] - 1$,

et $\sum_{i=1}^{k-1}\hat{\delta}_i^2 = n^2\sum_{i=1}^{k}\sum_{j=1}^{k}\hat{\mathbf{V}}_{ij}^2 \big/ p_{0i}p_{0j}$. Si les effets de plan sont tous similaires, les corrections d'ordre un et d'ordre deux auront un comportement similaire. Sinon, la correction d'ordre deux donne presque toujours de meilleurs résultats.

2.2 Cadre des tests du khi-carré et de l'estimateur du pseudo maximum de vraisemblance dans les enquêtes à base de sondage double

La situation examinée à la présente section s'inspire de Hartley (1962) et de Lu et Lohr (2010). Supposons qu'il existe k catégories dans les deux enquêtes et que les mêmes quantités sont mesurées. Soit p_{id} la proportion dans la population de la catégorie i dans le domaine d (le domaine d peut être le domaine a, le domaine ab ou le domaine b), avec $\sum_{i=1}^{k} p_{id} = 1$. Soit N_a, N_{ab} et N_b les tailles des populations des trois domaines, respectivement, avec $N_a + N_{ab} = N_A$ et $N_b + N_{ab} = N_B$. Nous considérons le cas fréquent où N_{ab} est inconnue, tandis que N_A et N_B sont constantes. Par conséquent, $\sum_{i=1}^{k} p_{ia}N_a/N_A + \sum_{i=1}^{k} p_{iab}N_{ab}/N_A = 1$ et $\sum_{i=1}^{k} p_{ib}N_b/N_B + \sum_{i=1}^{k} p_{iab}N_{ab}/N_B = 1$ (voir la figure 2.1 pour l'illustration des proportions). Le vecteur de proportions $\mathbf{p} = \left(p_1, p_2, \cdots p_{k-1}\right)'$ pour l'union des deux bases de sondage est une fonction des paramètres p_{ia}, p_{iab}, p_{ib} et N_{ab}. Par exemple, une forme naturelle de p_i est

$$p_i = \frac{N_a}{N}p_{ia} + \frac{N_{ab}}{N}p_{iab} + \frac{N_b}{N}p_{ib}, \qquad \text{pour} \qquad i = 1, 2, \cdots, k-1, \tag{2.4}$$

où $N = N_A + N_B - N_{ab}$.

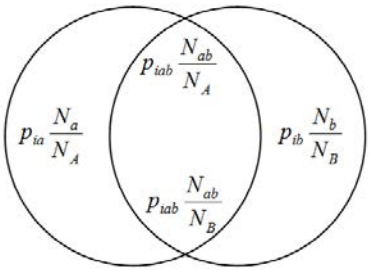

Figure 2.1 : Proportion de la population dans les domaines et les bases de sondage

Nous allons maintenant examiner brièvement l'estimateur du pseudo maximum de vraisemblance que nous utiliserons aux sections 4 et 5. Supposons que des échantillons aléatoires simples indépendants sont tirés des bases de sondage A et B, respectivement. La fonction de vraisemblance est

$$L\left(p_{ia}, p_{iab}, p_{ib}, N_{ab}\right) \propto \prod_i \left(p_{ia}\frac{N_a}{N_A}\right)^{x_{ia}} \times \prod_i \left(p_{iab}\frac{N_{ab}}{N_A}\right)^{x_{iab}^A} \times \prod_i \left(p_{ib}\frac{N_b}{N_B}\right)^{x_{ib}} \times \prod_i \left(p_{iab}\frac{N_{ab}}{N_B}\right)^{x_{iab}^B} \tag{2.5}$$

où x_{ia}, x_{ib} représentent les unités de la catégorie i dans le domaine a et dans le domaine b, respectivement; x_{iab}^A et x_{iab}^B représentent les unités de la catégorie i dans le domaine de chevauchement

ab qui sont échantillonnées au départ dans la base de sondage A et la base de sondage B, respectivement.

Dans le cas des estimateurs pour enquêtes complexes, l'idée fondamentale consiste à formuler l'hypothèse de travail d'une distribution multinomiale issue d'une population finie pour donner la forme des estimateurs et à se servir d'un effet de plan pour ajuster les fréquences de cellule afin qu'elles reflètent le plan de sondage complexe. La fonction de pseudo vraisemblance est la suivante

$$L\left(p_{ia}, p_{iab}, p_{ib}, N_{ab}\right) \propto \prod_i \left(p_{ia}\frac{N_a}{N_A}\right)^{\frac{\tilde{n}_A}{N_A}\hat{X}_{ia}} \prod_i \left(p_{iab}\frac{N_{ab}}{N_A}\right)^{\frac{\tilde{n}_A}{N_A}\hat{X}^A_{iab}} \times \prod_i \left(p_{ib}\frac{N_b}{N_B}\right)^{\frac{\tilde{n}_B}{N_B}\hat{X}_{ib}} \prod_i \left(p_{iab}\frac{N_{ab}}{N_B}\right)^{\frac{\tilde{n}_B}{N_B}\hat{X}^B_{iab}}, \tag{2.6}$$

où l'effet de plan est défini comme étant $\left\{v\left(\hat{\theta}\right) \text{ d'une enquête complexe}\right\} \Big/ \left\{v\left(\hat{\theta}\right) \text{ d'un EAS de même taille}\right\}$, $\tilde{n}_A = n_A \Big/ \left(\text{effet de plan de } \hat{N}^A_{ab}\right)$, $\tilde{n}_B = n_B \Big/ \left(\text{effet de plan de } \hat{N}^B_{ab}\right)$, n_A et n_B sont les tailles observées de S_A et S_B, et $\hat{X}_{id}$ désigne les fréquences estimées conformément au plan de sondage. Les estimateurs du pseudo maximum de vraisemblance (PMV), obtenus en maximisant (2.6), sont $\hat{p}_{ia} = \hat{X}_{ia} \big/ \hat{N}_a$, $\hat{p}_{ib} = \hat{X}_{ib} \big/ \hat{N}_b$, et

$$\hat{p}_{iab} = \frac{\frac{\tilde{n}_A}{N_A}\hat{N}^A_{ab}\hat{p}^A_{iab} + \frac{\tilde{n}_B}{N_B}\hat{N}^B_{ab}\hat{p}^B_{iab}}{\frac{\tilde{n}_A}{N_A}\hat{N}^A_{ab} + \frac{\tilde{n}_B}{N_B}\hat{N}^B_{ab}}, \tag{2.7}$$

où $\hat{p}^A_{iab} = \hat{X}^A_{iab} \big/ \hat{N}^A_{ab}$ et $\hat{p}^B_{iab} = \hat{X}^B_{iab} \big/ \hat{N}^B_{ab}$, et $\hat{N}_{ab,PMV}$ est la plus petite racine de la fonction quadratique

$$\left[\tilde{n}_A + \tilde{n}_B\right]\hat{N}^2_{ab,PMV} - \left[\tilde{n}_A N_B + \tilde{n}_B N_A + \tilde{n}_A \hat{N}^A_{ab} + \tilde{n}_B \hat{N}^B_{ab}\right]\hat{N}_{ab,PMV} + \left[\tilde{n}_A \hat{N}^A_{ab} N_B + \tilde{n}_B \hat{N}^B_{ab} N_A\right] = 0. \tag{2.8}$$

Les estimateurs des proportions de la population sont

$$\hat{p}_{i,PMV} = \frac{\left(N_A - \hat{N}_{ab,PMV}\right)\hat{p}_{ia} + \hat{N}_{ab,PMV}\hat{p}_{iab} + \left(N_B - \hat{N}_{ab,PMV}\right)\hat{p}_{ib}}{N_A + N_B - \hat{N}_{ab,PMV}}. \tag{2.9}$$

Si l'on tire un EAS de chaque base de sondage et que $k = 1$, ces estimateurs PMV se réduisent à ceux présentés dans Skinner et Rao (1996).

3 Tests du khi-carré dans les enquêtes à base de sondage double

À la présente section, nous considérons le cas des tests du khi-carré dans les enquêtes à base de sondage double. Voici certaines hypothèses d'intérêt : une simple hypothèse $H_0 : q_{ia} = p_{ia}N_a/N_A = q^A_{ia0}$, $q^A_{iab} = p_{iab}N_{ab}/N_A = q^A_{iab0}$, $q^B_{iab} = p_{iab}N_{ab}/N_B = q^B_{iab0}$, $q_{ib} = p_{ib}N_b/N_B = p_{ib0}$ (à noter que les q_{ia}, etc., sont utilisés pour simplifier la notation); $H_0 : p_{i,PMV} = p_{i0,PMV}$, pour vérifier si l'estimateur PMV des proportions provenant de l'union des deux bases de sondage en (2.9) correspond à certaines valeurs spécifiques (à noter que p_i peut être estimé par d'autres méthodes);

$H_0 : p_{ia} = p_{iab} = p_{ib}$, pour vérifier si les proportions sont égales dans les trois domaines; ou $H_0 : p_{ij} = p_{i+}p_{+j}$, pour vérifier l'indépendance de la classification par ligne et de la classification par colonne.

Posons que $\boldsymbol{\eta} = \left(\mathbf{p}'_a N_a/N_A, \mathbf{p}'_{ab} N_{ab}/N_A, \mathbf{p}'_b N_b/N_B, \mathbf{p}'_{ab} N_{ab}/N_B\right)'$, $\mathbf{p}_a = \left(p_{1a}, p_{2a}, \cdots p_{ka}\right)'$, $\mathbf{p}_b = \left(p_{1b}, p_{2b}, \cdots p_{kb}\right)'$, $\mathbf{p}_{ab} = \left(p_{1ab}, p_{2ab}, \cdots p_{(k-1)ab}\right)'$, et que les h_i sont des fonctions continues. Une hypothèse d'intérêt plus général peut être énoncée comme il suit :

$$H_0 : h_i(\boldsymbol{\eta}) = 0, \quad i = 1, 2, \cdots r. \tag{3.1}$$

Soit η_j le j^{e} élément de $\boldsymbol{\eta}$ et soit $h(\boldsymbol{\eta}) = \left(h_1(\boldsymbol{\eta}), h_2(\boldsymbol{\eta}), \cdots h_r(\boldsymbol{\eta})\right)'$.

Supposons que $\partial h_i(\boldsymbol{\eta})/\partial \eta_j$ est continue dans le voisinage de $\boldsymbol{\eta}$ et que

$$\nabla = \frac{\partial h_i(\boldsymbol{\eta})}{\partial \eta_j} \tag{3.2}$$

est de plein rang. Formulons les hypothèses qui suivent.

A$_1$. Il existe une séquence de superpopulations $U_{A1} \subset U_{A2} \subset \ldots \subset U_{At} \subset \ldots$ telle qu'elle est définie dans Isaki et Fuller (1982).

A$_2$. Soit $\tilde{n}_A$ et $\tilde{n}_B$ tels que définis à la section 2, et supposons que $\tilde{n}_A$ et $\tilde{n}_B$ augmentent tous deux de manière que $\tilde{n}_A/\tilde{n}_B \to \gamma$ pour une certaine valeur $0 < \gamma < 1$.

A$_3$. Soit $\pi_{it}^A = p\left(\text{l'upe } i \text{ se trouve dans l'échantillon tiré de la base } A, \text{ en utilisant la population } U_{At}\right)$ et $\pi_{ijt}^A = p\left(\text{les upe } i \text{ et } j \text{ se trouvent dans l'échantillon tiré de la base } A, \text{ en utilisant la population } U_{At}\right)$ les probabilités d'inclusion et d'inclusion conjointe pour l'échantillon tiré de la base de sondage A en utilisant la population U_{At} et définissons π_{it}^B, π_{ijt}^B et U_{Bt} de la même façon pour la base de sondage B. Supposons qu'il existe des constantes c_1 et c_2 telles que

$$0 < c_2 < \pi_{it}^F < c_1 < 1 \tag{3.3}$$

pour tout i et toute superpopulation dans la séquence, où F désigne la base de sondage A ou la base de sondage B. Supposons aussi qu'il existe un α_t avec $\alpha_t = o(1)$ tel que

$$\pi_{it}^F \pi_{jt}^F - \pi_{ijt}^F \le \alpha_t \pi_{it}^F \pi_{jt}^F. \tag{3.4}$$

A$_4$. $N_{ab}/N \to \psi$ pour une certaine valeur de ψ entre 0 et 1.

Théorème 1. *Avec les hypothèses A_1 à A_4 établies préalablement, nous arrivons à la conclusion suivante : $\tilde{n}^{1/2}\mathbf{h}\left(\hat{\boldsymbol{\eta}}\right)$ est asymptotiquement normale de moyenne $\mathbf{0}$ et de variance asymptotique $\nabla\Sigma\nabla'$, où Σ est une matrice diagonale par bloc avec les blocs Σ_A et Σ_B, et $\tilde{n} = \tilde{n}_A + \tilde{n}_B$. Σ_A est la matrice de*

covariance asymptotique de $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_A$ *avec* $\hat{\boldsymbol{\eta}}_A = \left(\hat{\mathbf{p}}_a' \hat{N}_a / N_A, \hat{\mathbf{p}}_{ab}^{A\prime} \hat{N}_{ab} / N_A\right)'$, Σ_B *est la matrice de covariance asymptotique de* $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_B$ *avec* $\hat{\boldsymbol{\eta}}_B = \left(\hat{\mathbf{p}}_b' \hat{N}_b / N_B, \hat{\mathbf{p}}_{ab}^{B\prime} \hat{N}_{ab} / N_B\right)'$ *et* $\hat{\boldsymbol{\eta}} = \left(\hat{\boldsymbol{\eta}}_A', \hat{\boldsymbol{\eta}}_B'\right)'$.

Preuve. Les arguments donnés au théorème 1 dans Lu et Lohr (2010) montrent que $\hat{\boldsymbol{\eta}}$ converge vers $\boldsymbol{\eta}$ et que $\hat{\boldsymbol{\eta}}$ obéit au théorème de la limite central, car $\tilde{n}_A$ et $\tilde{n}_B$ augmentent tous deux de manière que $\tilde{n}_A / \tilde{n}_B \to \gamma$. Donc, puisque les échantillons S_A et S_B sont sélectionnés indépendamment, nous avons

$$\tilde{n}^{1/2}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) \xrightarrow{d} N\left(0, \Sigma\right).$$

$\mathbf{h}\left(\hat{\boldsymbol{\eta}}\right)$ converge vers $\mathbf{h}\left(\boldsymbol{\eta}\right)$ parce que $\hat{\boldsymbol{\eta}}$ converge vers $\boldsymbol{\eta}$. En utilisant la méthode delta, $\tilde{n}^{1/2}\mathbf{h}\left(\hat{\boldsymbol{\eta}}\right)$ suit asymptotiquement une distribution normale de moyenne $\mathbf{0}$ et de variance asymptotique $\nabla\Sigma\nabla'$.

En se basant sur le théorème 1, il s'ensuit immédiatement les résultats suivants.

Résultat 1. (Test de Wald étendu) Si un estimateur convergent de la variance Σ est disponible, en vertu du théorème 1, la statistique de Wald généralisée peut être formée comme il suit :

$$X_W^2 = \tilde{n}\mathbf{h}\left(\hat{\boldsymbol{\eta}}\right)' \left(\hat{\nabla}\hat{\Sigma}\hat{\nabla}'\right)^{-1} \mathbf{h}\left(\hat{\boldsymbol{\eta}}\right). \tag{3.5}$$

Cette statistique de test suit asymptotiquement une distribution $\chi^2(r)$ sous H_0 (voir l'équation 3.1), où r est le rang de ∇.

Comme nous l'avons mentionné plus haut, l'estimation de la variance Σ peut être instable ou ne pas avoir de forme explicite. Un moyen de modifier la statistique en (3.5) consiste à d'abord agir comme si l'échantillon était un échantillon aléatoire simple, puis à modifier la distribution de référence utilisée dans le test pour obtenir le niveau correct. L'équation (3.6) donne la statistique modifiée.

Résultat 2. Soit

$$X_{MW}^2 = \tilde{n}\mathbf{h}\left(\hat{\boldsymbol{\eta}}\right)' \left(\hat{\nabla}_0\hat{\mathbf{P}}_0\hat{\nabla}_0'\right)^{-1} \mathbf{h}\left(\hat{\boldsymbol{\eta}}\right), \tag{3.6}$$

où $\hat{\nabla}_0\hat{\mathbf{P}}_0\hat{\nabla}_0'$ peut être toute estimation de $\nabla\mathbf{P}\nabla$ qui est convergente quand H_0 est vérifiée. La matrice $\mathbf{P}_0$ est une matrice diagonale par bloc avec pour blocs diagonaux la matrice de covariance provenant de la base de sondage A et la matrice de covariance provenant de la base de sondage B quand H_0 est vérifiée et que l'échantillonnage est un EAS. Supposons que la matrice ∇ est de rang r sous l'hypothèse nulle $H_0 : \mathbf{h}(\boldsymbol{\eta}) = 0$. Alors, $X_{MW}^2 \approx \sum_1^r \lambda_{0i} W_i$, où les λ_i sont les valeurs propres de $\left(\nabla\mathbf{P}\nabla'\right)^{-1}\left(\nabla\Sigma\nabla'\right)$, les $W_1, \ldots, W_r$ sont des variables aléatoires indépendantes χ_1^2, et λ_{0i} est la valeur de λ_i sous H_0.

Résultat 3. (Correction d'ordre un de Rao-Scott étendue) Supposons que la matrice ∇ est de rang r. Soit X_{MW}^2 telle qu'elle est définie en (3.6). Sous l'hypothèse nulle $H_0 : \mathbf{h}(\boldsymbol{\eta}) = 0$, la statistique $X_{MW}^2 / \hat{\lambda}$. a

pour espérance r, où $\hat{\lambda}_. = \sum \hat{\lambda}_i / r$, $\hat{\lambda}_i$ est une estimation convergente de λ_i sous H_0. Par exemple, les $\hat{\lambda}_i$ pourraient être les valeurs propres de $\left(\hat{\nabla}_0 \hat{\mathbf{P}}_0 \hat{\nabla}_0{}'\right)^{-1} \left(\hat{\nabla}\hat{\Sigma}\hat{\nabla}'\right)$.

Résultat 4. (Correction d'ordre deux de Rao-Scott étendue) Supposons que la matrice ∇ est de rang r. Définissons

$$X_S^2 = \frac{X_{MW}^2}{\hat{\lambda}_.\left(1+\hat{a}^2\right)}$$

où $\hat{a}^2 = \sum_{i=1}^{k-1} \hat{\lambda}_i^2 \Big/ \left[(k-1)\hat{\lambda}_.^2\right] - 1$ est une estimation de la valeur de population a^2. Sous l'hypothèse nulle, X_S^2 suit asymptotiquement la même distribution que χ_v^2, une variable aléatoire khi-carré à $v = (k-1)\Big/\left(1+a^2\right)$ degrés de liberté.

4 Simulations

La présente section décrit une petite simulation effectuée pour étudier les tests du khi-carré proposés à la section 3 sous l'hypothèse simple H_0 : $q_{ia} = p_{ia}\, N_a / N_A = q_{ia0}^A$, $q_{iab}^A = p_{iab}\, N_{ab} / N_A = q_{iab0}^A$, $q_{iab}^B = p_{iab}\, N_{ab} / N_B = q_{iab0}^B$, $q_{ib} = p_{ib}\, N_b / N_B = p_{ib0}$ afin d'examiner leurs propriétés. Nous comparons les pourcentages des échantillons pour lesquels la valeur de la statistique de test dépasse la valeur critique au seuil nominal $(\alpha = 0{,}05)$. R (www.r-project.org) est utilisé pour exécuter l'étude de simulation et d'autres analyses.

Nous avons généré les données comme l'ont fait Skinner et Rao (1996), avec $\gamma_a = N_a / N$ et $\gamma_b = N_b / N$. Nous avons généré un échantillon en grappes tiré de la base de sondage A contenant n_p upe et m observations dans chaque upe, et un échantillon aléatoire simple de n_B observations tiré de la base de sondage B. Nous avons généré les réponses binaires en grappes pour l'échantillon tiré de la base de sondage A en générant des vecteurs de variables aléatoires normales multivariés corrélés, puis en utilisant la fonction probit pour convertir les réponses continues en réponses binaires. Après avoir généré l'échantillon, nous avons calculé les estimateurs PMV de $\mathbf{p}_{id}\, N_d / N_A$ et $\mathbf{p}_{id}\, N_d / N_B$ (voir la section 2.2). Nous nous sommes servis des proportions estimées pour calculer les statistiques des tests du khi-carré. Puis, nous avons comparé les pourcentages des échantillons pour lesquels les valeurs des statistiques de test dépassaient la valeur critique au seuil nominal sous différentes conditions.

La simulation a été exécutée en appliquant les paramètres suivants : 1) γ_a : 0,4; 2) γ_b : 0,2; 3) paramètre de mise en grappes ρ : 0,3; 4) tailles d'échantillon : n_p : 10, 30 ou 50; m : 3, 5 ou 10, n_B : 100, 300 ou 500; 5) exécution des simulations : 1 000 fois pour chaque paramétrisation et 100 fois pour estimer la matrice de variance-covariance V par la méthode du bootstrap. Toutes les simulations ont été effectuées en utilisant les paramètres probabilistes $\mathbf{p}_a$: $(0{,}3;0{,}1;0{,}2;0{,}4)$, $\mathbf{p}_{ab}$: $(0{,}3;0{,}1;0{,}1;0{,}5)$ et $\mathbf{p}_b$: $(0{,}4;0{,}1;0{,}1;0{,}4)$. Le tableau 4.1 donne les pourcentages des échantillons pour lesquels la valeur de la statistique de test dépasse la valeur critique.

Tableau 4.1
Comparaison des seuils de signification réels (%) entre les différents tests. X^2 est le test non corrigé; X^2_{FC} est le X^2 avec correction d'ordre un et X^2_{SC} est le X^2 avec correction d'ordre deux.

$\tilde{n}_p$	m	n_B	X^2	**Wald**	X^2_{FC}	X^2_{SC}
10	3	100	12,1	17,3	5,6	4,9
30	3	300	13,6	8,4	4,8	4,8
50	3	500	15,5	10,0	6,4	3,6
10	5	100	25,7	13,5	7,5	4,9
30	5	300	29,2	9,3	7,9	5,3
50	5	500	31,5	8,5	8,1	4,9
10	10	100	46,1	21,2	6,6	5,4
30	10	300	50,2	11,5	7,5	5,6
50	10	500	58,7	8,0	9,6	5,1

Le tableau 4.1 indique que l'application naïve du test X^2 non corrigé à des données provenant d'enquêtes complexes est dangereuse. Lorsque la taille et le nombre d'upe augmentent, le seuil de signification réel atteint même 62,2 %. Le test de Wald étendu ne donne pas de bons résultats puisque l'estimation de la variance peut être instable. Le test avec correction d'ordre un étendu donne un résultat acceptable avec un seuil de signification réel de 7 % environ. Le test avec correction d'ordre deux étendu atteint presque le seuil de signification nominal de 5 %, et est celui que nous recommandons d'utiliser dans une analyse de données catégoriques provenant d'une enquête à base de sondage double.

5 Application

À la présente section, nous donnons un exemple réel pour illustrer les propriétés des tests du khi-carré dans le contexte d'une enquête à base de sondage double. Nous considérons le test d'hypothèse $H_0: p_{ia} = p_{iab} = p_{ib}$ pour vérifier si les proportions sont égales dans les trois domaines.

5.1 Description des données et estimateurs PMV connexes

Les données (Lohr et Rao 2006) ont été recueillies à l'origine dans le cadre d'une enquête à trois bases de sondage auprès des statisticiens, en utilisant les répertoires en ligne des membres de l'American Statistical Association (ASA), de l'Institute for Mathematical Statistics (IMS) et de la Société statistique du Canada. Nous traitons l'union des répertoires en ligne des membres de l'ASA et des répertoires en ligne des membres de l'IMS comme une base de sondage double en utilisant la notation $A \cup B = a \cup ab \cup b$ (A: répertoires en ligne des membres de l'ASA; B: répertoires en ligne des membres de l'IMS; domaine a: membres de l'ASA mais pas de l'IMS; domaine ab: membres de l'ASA et aussi de l'IMS; domaine b: membres de l'IMS, mais pas de l'ASA). Notons que l'union de ces deux bases de sondage ne couvre pas la population entière de statisticiens. Bon nombre de statisticiens n'appartiennent ni à l'une ni à l'autre de ces sociétés, et certains refusent que leur nom figure dans les répertoires en ligne. Dans le jeu de données, l'information sur la profession est une variable catégorique à trois niveaux : milieu universitaire, industrie et administration publique. Nous combinons l'industrie et l'administration publique en un seul niveau que nous appelons milieu non universitaire. En combinant la

profession et le sexe, nous obtenons un tableau 2×2 contenant quatre cellules : femmes en milieu universitaire, femmes en milieu non universitaire, hommes en milieu universitaire et hommes en milieu non universitaire.

Au moment de la collecte des données, $15\,500$ personnes appartenaient à l'American Statistical Association (base de sondage A) et $4\,000$ personnes, à l'Institute for Mathematical Statistics (base de sondage B), de sorte que $N_A = 15\,500$ et $N_B = 4\,000$. De la base de sondage A a été tiré un échantillon en grappes stratifié de taille 500 pour lequel 378 observations contenaient l'information pour les deux variables (sexe et profession). Le plan comportait 26 strates construites par régions ou par États. En raison des restrictions d'accès aux dossiers, les grappes pour les grands États étaient formées des membres dont le nom de famille commençait par la même lettre de l'alphabet. La base de sondage A contient 173 upe. De la base de sondage B a été tiré un échantillon aléatoire simple de taille 140, pour lequel 102 enregistrements contenaient des renseignements valides pour les deux variables. Le total pondéré d'observations provenant de la base de sondage A est de $10\,976$. Nous supposons que les données manquent au hasard, de sorte que l'ajustement pour tenir compte de la non-réponse est effectué en appliquant une fraction de $15\,500/10\,976$. Le tableau 5.1 donne le nombre de statisticiens compris dans chaque cellule à l'intérieur de chaque domaine.

Tableau 5.1
Données observées dans le domaine a et dans le domaine ab provenant de la base de sondage A (ajustées par une fraction de 15 500/10 976) ainsi que les données observées dans le domaine b et le domaine ab provenant de la base de sondage B

	Domaine a		Domaine $ab \in A$		Domaine b		Domaine $ab \in B$	
	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes
Universitaire	2 425	4 969	302	1 488	10	41	10	33
Non universitaire	1 959	4 091	59	209	0	3	2	3

L'effet de plan estimé de la base de sondage A est 1,801209, de sorte que la taille effective d'échantillon correspondant à n_A est $\tilde{n}_A = 378/1{,}8 = 210$. La taille effective d'échantillon $n_{B,eff} = n_B = 102$. Les estimateurs PMV des proportions estimées en utilisant (2.6) et (2.9) sont donnés au tableau 5.2.

Tableau 5.2
Proportions estimées provenant des domaines et de l'union des deux bases de sondage

	Domaine a		Domaine ab		Domaine b		Base de sondage $A \cup B$	
	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes
Universitaire	0,180	0,370	0,186	0,701	0,185	0,759	0,182	0,452
Non universitaire	0,146	0,304	0,037	0,076	0	0,056	0,116	0,250

5.2 Test de l'équivalence des proportions dans les divers domaines

L'hypothèse à vérifier est que les proportions sont égales dans les trois domaines,

$$H_0 : p_{ia} = p_{iab} \quad \text{et} \quad p_{iab} = p_{ib},\ i = 1,2,3. \tag{5.1}$$

Dans cet exemple, p_{ia}, $i=1,2,3,4$ représente les proportions de femmes en milieu universitaire, de femmes en milieu non universitaire, d'hommes en milieu universitaire et d'hommes en milieu non universitaire parmi les membres de l'ASA, respectivement. Définissons de la même façon p_{iab} et p_{ib}. $\boldsymbol{\eta}$ (voir la section 3) se réduit à un vecteur de dimension 14×1

$$\boldsymbol{\eta}=(p_{1a}N_a/N_A, p_{2a}N_a/N_A, p_{3a}N_a/N_A, p_{4a}N_a/N_A, p_{1ab}N_{ab}/N_A, p_{2ab}N_{ab}/N_A, p_{3ab}N_{ab}/N_A,$$
$$p_{1b}N_b/N_B, p_{2b}N_b/N_B, p_{3b}N_b/N_B, p_{4b}N_b/N_B, p_{1ab}N_{ab}/N_B, p_{2ab}N_{ab}/N_B, p_{3ab}N_{ab}/N_B)'.$$

H_0 en (5.1) ne faisant intervenir que les simples paramètres p_{ia}, p_{iab}, p_{ib} et N_{ab}, un nouveau vecteur est introduit $\theta=\left(p_{1a}, p_{2a}, p_{3a}, p_{1ab}, p_{2ab}, p_{3ab}, N_{ab}/N_A, p_{1b}, p_{2b}, p_{3b}, p_{1ab}, p_{2ab}, p_{3ab}, N_{ab}/N_B\right)'$.

Soit $\Omega=\left(\partial h_i(\boldsymbol{\eta})/\partial\theta_j\right)$ et $\mathbf{D}(\theta)=\left(\partial\boldsymbol{\eta}/\partial\theta_j\right)$. $\mathbf{D}(\theta)$ s'avère être une matrice diagonale par bloc avec

$$\mathbf{D}_A=\begin{pmatrix}
\frac{N_a}{N_A} & 0 & 0 & 0 & 0 & 0 & -p_{1a} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & \frac{N_a}{N_A} & 0 & 0 & 0 & 0 & -p_{2a} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & \frac{N_a}{N_A} & 0 & 0 & 0 & -p_{3a} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
-\frac{N_a}{N_A} & -\frac{N_a}{N_A} & -\frac{N_a}{N_A} & 0 & 0 & 0 & -p_{4a} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & \frac{N_{ab}}{N_A} & 0 & 0 & p_{1ab} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & \frac{N_{ab}}{N_A} & 0 & p_{2ab} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_A} & p_{3ab} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_b}{N_B} & 0 & 0 & 0 & 0 & 0 & -p_{1b}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_b}{N_B} & 0 & 0 & 0 & 0 & -p_{2b}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_b}{N_B} & 0 & 0 & 0 & -p_{3b}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{N_b}{N_B} & -\frac{N_b}{N_B} & -\frac{N_b}{N_B} & 0 & 0 & 0 & -p_{4b}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_B} & 0 & 0 & p_{1ab}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_B} & 0 & p_{2ab}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_B} & p_{3ab}
\end{pmatrix}.$$

En notant la relation entre $\hat{p}_{iab}$ et $\hat{p}_{iab}^{A}$ et $\hat{p}_{iab}^{B}$ provenant de (2.7), on trouve que Ω est

$$\Omega = \begin{pmatrix} 1 & 0 & 0 & -\phi & 0 & 0 & 0 & 0 & 0 & 0 & -(1-\phi) & 0 & 0 & 0 \\ 0 & 0 & 0 & \phi & 0 & 0 & 0 & -1 & 0 & 0 & 1-\phi & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -\phi & 0 & 0 & 0 & 0 & 0 & 0 & -(1-\phi) & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi & 0 & 0 & 0 & -1 & 0 & 0 & (1-\phi) & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -\phi & 0 & 0 & 0 & 0 & 0 & 0 & -(1-\phi) & 0 \\ 0 & 0 & 0 & 0 & 0 & \phi & 0 & 0 & 0 & -1 & 0 & 0 & (1-\phi) & 0 \end{pmatrix},$$

où $\phi = N_B \tilde{n}_A / \left(N_B \tilde{n}_A + N_A \tilde{n}_B \right)$. Il est facile de montrer que $\nabla = \Omega\left(\mathbf{D}\right)^{-1}$ (rappelons que $\nabla = \partial h_i\left(\boldsymbol{\eta}\right) / \partial \eta_j$). $\hat{\Sigma}$ est estimée par une méthode du jackknife appliquée en supprimant chaque fois une upe provenant de la base de sondage A. Tous les résultats présentés à la section 3 peuvent être dérivés. Les valeurs propres de $\left(\hat{\nabla}_0 \hat{\mathbf{P}}_0 \hat{\nabla}_0'\right)^{-1} \left(\nabla \Sigma \nabla'\right)$ sont très proches l'une de l'autre, ce qui indique que le test avec correction d'ordre un donne les mêmes résultats que le test avec correction d'ordre deux. La statistique de Wald, la statistique avec correction d'ordre un et la statistique avec correction d'ordre deux donnent les valeurs de 81,48295, 72,31026 et 70,28581, respectivement. En comparant la valeur critique avec six degrés de liberté $\chi^2(6) = 12{,}95$, nous rejetons l'hypothèse nulle selon laquelle les proportions dans les cellules (femmes en milieu universitaire, femmes en milieu non universitaire, hommes en milieu universitaire et hommes en milieu non universitaire) sont les mêmes dans les trois domaines (membres de l'ASA seulement, membres de l'ASA et de l'IMS, et membres de l'IMS seulement).

6 Conclusion

Dans la présente étude, nous étendons le test de Wald (1943) et les tests avec correction d'ordre un et correction d'ordre deux de Rao-Scott (Rao et Scott 1981) d'une enquête à base de sondage unique à une enquête à base de sondage double, et nous établissons les distributions asymptotiques. Une étude de simulations limitée donne à penser que les tests avec correction d'ordre deux atteignent presque le seuil nominal. Bien que les résultats présentés ici se rapportent à des enquêtes à base de sondage double, les méthodes sont générales et pourraient être étendues à plus de deux bases de sondage. Notre étude est réalisée dans le contexte de l'échantillonnage; elle s'applique aussi à d'autres situations dans lesquelles des données provenant de deux sources indépendantes pourraient être combinées.

Remerciements

L'auteur remercie Dr. Sharon Lohr de ses conseils et commentaires précieux au sujet du manuscrit. L'auteur remercie aussi les examinateurs et le rédacteur associé de leurs commentaires très utiles et de leurs suggestions constructives.

Bibliographie

Bedrick, E.J. (1983). Adjusted chi-squared tests for cross-classified tables of survey data. *Biometrika*, 70, 591-595.

Fay, R.E. (1979). On adjusting the Pearson chi-square statistic for clustered sampling. Dans *ASA Proceedings of the Social Statistics Section*, 402-406. American Statistical Association.

Fay, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

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

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

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Series C*, 36 (3), 99-118.

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

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

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

Lu, Y. et Lohr, S. (2010). L'estimation des flux bruts dans les enquêtes à base de sondage double. *Techniques d'enquête*, 36(1), 13-24.

Rao, J.N.K. et Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K. et Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.

Rao, J.N.K. et Scott, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, 15, 385-397.

Rao, J.N.K. et Thomas, D.R. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, 18, 213-269.

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

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

Thomas, D.R., Singh, A. et Roberts, G. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review*, 64(3), 295-311.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54, 426-482.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 367-378
Statistique Canada, n° 12-001-X au catalogue

Méthodes d'estimation sur bases de sondage multiples dans le cadre de plans de sondage à deux degrés

Guillaume Chauvet et Guylène Tandeau de Marsac[1]

Résumé

Lorsqu'on s'intéresse à une population finie, il arrive qu'il soit nécessaire de tirer des échantillons dans plusieurs bases de sondage pour représenter l'ensemble des individus. Nous nous intéressons ici au cas de deux échantillons sélectionnés selon un plan à deux degrés, avec un premier degré de tirage commun. Nous appliquons les méthodes de Hartley (1962), Bankier (1986), et Kalton et Anderson (1986), et nous montrons que ces méthodes peuvent être appliquées conditionnellement au premier degré de tirage. Nous comparons également la performance de plusieurs estimateurs dans le cadre d'une étude par simulations. Nos résultats suggèrent que le choix d'un estimateur en présence de bases de sondage multiples se fasse de façon prudente, et qu'un estimateur simple est parfois préférable même s'il n'utilise qu'une partie de l'information collectée.

Mots-clés : Enquête à extension; estimateur de Hansen-Hurwitz; estimateur de Horvitz-Thompson; sondage à deux degrés.

1 Introduction

Lorsqu'on s'intéresse à une population finie, il arrive qu'aucune base de sondage ne la recouvre totalement et qu'il soit nécessaire de tirer des échantillons dans deux bases de sondage (ou plus) pour représenter l'ensemble des individus. Pour mettre en commun ces échantillons, de nombreuses méthodes d'estimation sur bases de sondage multiples ont été proposées (Hartley 1962; Bankier 1986; Kalton et Anderson 1986; Mecatti 2007; Rao et Wu 2010); voir également les articles de revue de Lohr (2009, 2011), et les articles référencés, pour un panorama complet. Notons que la méthode de Mecatti (2007) s'inspire des travaux de Lavallée (2002, 2007) sur la méthode généralisée du partage des poids. À la section 2, nous présentons différentes méthodes d'estimation pour des bases de sondage multiples.

À la section 3, nous nous intéressons au cas de deux échantillons sélectionnés selon un plan à deux degrés, avec un premier degré de tirage commun. Ce cadre correspond aux enquêtes de l'INSEE avec extension : un premier échantillon de logements est sélectionné dans les communes de l'échantillon-maître (Bourdalle, Christine et Wilms 2000), et un second échantillon est sélectionné et enquêté dans les communes du même échantillon-maître afin de cibler une sous-population spécifique. On dispose de deux mesures d'enquêtes provenant de deux échantillons indépendants au deuxième degré du plan de sondage. Nous appliquons des méthodes d'estimation sur bases de sondage multiples pour la mise en commun de ces deux échantillons. Nous montrons que les estimateurs étudiés peuvent dans ce contexte être calculés conditionnellement au premier degré de tirage, ce qui simplifie leur calcul notamment pour l'estimateur optimal de Hartley (1962). À la section 4, nous comparons les performances de ces estimateurs dans le cadre d'une étude par simulations. Nous concluons à la section 5.

1. Guillaume Chauvet, ENSAI (CREST), Campus de Ker Lann, Bruz – France. Courriel : chauvet@ensai.fr. Guylène Tandeau de Marsac, INSEE, Direction Régionale de Lille – France. Courriel : guylene.tandeau-de-marsac@insee.fr.

2 Estimation pour des bases de sondage multiples

On considère une population finie U sur laquelle est définie une variable d'intérêt y de valeur y_k pour l'individu k. Si on sélectionne dans U un échantillon S avec des probabilités d'inclusion π_k, l'estimateur $\hat{Y} = \sum_{k \in S} \pi_k^{-1} y_k$ proposé par Narain (1951) et Horvitz et Thompson (1952) est sans biais pour le total $Y = \sum_{k \in U} y_k$ si toutes les probabilités π_k sont strictement positives.

Nous nous intéressons au cas où la population est entièrement couverte par deux bases de sondage chevauchantes U_A et U_B. En utilisant les notations de Lohr (2011), soient $a = U_A \setminus U_B$ le domaine couvert par U_A seulement; $b = U_B \setminus U_A$ le domaine couvert par U_B seulement; $ab = U_A \cap U_B$ le domaine couvert à la fois par U_A et U_B. On sélectionne dans U_A un échantillon S^A avec des probabilités d'inclusion $\pi_k^A > 0$. Pour tout domaine $d \subset U_A$, le sous-total $Y_d = \sum_{k \in d} y_k$ est estimé sans biais par $\hat{Y}_d^A = \sum_{k \in S_A} d_k^A y_k 1(k \in d)$ avec $d_k^A = \left(\pi_k^A\right)^{-1}$. On sélectionne dans U_B un échantillon S^B avec des probabilités d'inclusion $\pi_k^B > 0$. Pour tout domaine $d \subset U_B$, le sous-total Y_d est estimé sans biais par $\hat{Y}_d^B = \sum_{k \in S_B} d_k^B y_k 1(k \in d)$ avec $d_k^B = \left(\pi_k^B\right)^{-1}$. L'objectif est de combiner les échantillons S^A et S^B pour obtenir une estimation de Y aussi précise que possible.

2.1 Estimateur de Hartley

Hartley (1962) propose la classe d'estimateurs sans biais

$$\hat{Y}_\theta = \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B, \tag{2.1}$$

avec θ un paramètre à déterminer. Le choix $\theta = 1/2$ conduit à donner aux échantillons S^A et S^B le même poids pour l'estimation sur le domaine intersection ab. Hartley (1962) propose de choisir le paramètre qui minimise la variance de $\hat{Y}_\theta$. Cela conduit à

$$\theta_{opt} = \frac{Cov\left(\hat{Y}_a^A + \hat{Y}_{ab}^B + \hat{Y}_b^B, \hat{Y}_{ab}^B - \hat{Y}_{ab}^A\right)}{V\left(\hat{Y}_{ab}^B - \hat{Y}_{ab}^A\right)}, \tag{2.2}$$

que l'on peut réécrire sous la forme

$$\theta_{opt} = \frac{V\left(\hat{Y}_{ab}^B\right) + Cov\left(\hat{Y}_{ab}^B, \hat{Y}_b^B\right) - Cov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A\right)}{V\left(\hat{Y}_{ab}^A\right) + V\left(\hat{Y}_{ab}^B\right)} \tag{2.3}$$

quand les échantillons S^A et S^B sont indépendants. Comme le remarque Lohr (2007), le coefficient optimal θ_{opt} peut ne pas être compris entre 0 et 1 si un terme de covariance présent dans (2.3) est grand. Supposons pour simplifier que $Cov\left(\hat{Y}_{ab}^B, \hat{Y}_b^B\right) = 0$, ce qui est le cas si b et ab sont utilisés comme strates dans la sélection de S^B. Alors $\theta_{opt} > 1$ si et seulement si $Cov\left(\hat{Y}^A, \hat{Y}_{ab}^A\right) < 0$. Dans le cas où S^A est sélectionné par sondage aléatoire simple, ce sera par exemple le cas si dans U_A les faibles valeurs de la variable y sont concentrées dans le domaine ab.

En pratique, les termes de variance et de covariance sont inconnus et doivent être remplacés par des estimateurs, ce qui introduit une variabilité supplémentaire. Un autre inconvénient est que le paramètre

optimal dépend de la variable d'intérêt considérée. Si des estimateurs optimaux sont calculés pour différentes variables d'intérêt, les estimations peuvent souffrir d'une incohérence interne (Lohr 2011).

2.2 Estimateur de Kalton et Anderson

Une classe plus générale d'estimateurs s'obtient en remarquant que le total Y peut se réécrire

$$Y = Y_a + \sum_{k \in ab} \theta_k y_k + \sum_{k \in ab} \left(1 - \theta_k\right) y_k + Y_b,$$

avec θ_k un coefficient propre à l'individu k. Kalton et Anderson (1986) proposent le choix $\theta_k = \left(d_k^A + d_k^B\right)^{-1} d_k^B$, qui conduit à l'estimateur

$$\hat{Y}_{KA} = \sum_{k \in S^A} d_k^A m_k^A y_k + \sum_{k \in S^B} d_k^B m_k^B y_k \tag{2.4}$$

avec d'une part $m_k^A = 1$ si $k \in a$ et $m_k^A = \theta_k$ si $k \in ab$, d'autre part $m_k^B = 1$ si $k \in b$ et $m_k^B = 1 - \theta_k$ si $k \in ab$. Les poids d'estimation sont les mêmes quelle que soit la variable d'intérêt, ce qui assure la cohérence interne des estimations; en revanche, l'estimateur de Kalton et Anderson est moins efficace que l'estimateur optimal de Hartley pour une variable d'intérêt donnée. Notons qu'il s'agit d'un estimateur de type Hansen-Hurwitz (1943), qui peut se réécrire sous la forme $\hat{Y}_{KA} = \sum_{k \in U} \left[W_k / E\left(W_k\right)\right] y_k$ en notant $W_k = 1\left(k \in S^A\right) + 1\left(k \in S^B\right)$ le nombre de fois où l'unité k est sélectionnée dans l'échantillon réunion $S^A \cup S^B$. On a en particulier $E(W_k) = \pi_k^A + \pi_k^B$.

2.3 Estimateur de Bankier

Bankier (1986) propose d'utiliser un estimateur de type Horvitz-Thompson, en calculant les probabilités d'inclusion dans l'échantillon réunion

$$\pi_k^{HT} \equiv P\left(k \in S^A \cup S^B\right) = \pi_k^A + \pi_k^B - Pr\left(k \in S^A \cap S^B\right).$$

Si les échantillons S^A et S^B sont indépendants, on obtient $\pi_k^{HT} = \pi_k^A + \pi_k^B - \pi_k^A \pi_k^B$ et l'estimateur

$$\hat{Y}_{HT} = \sum_{k \in S^A \cup S^B} \frac{y_k}{\pi_k^{HT}} = \sum_{k \in S^A \cap a} \frac{y_k}{\pi_k^A} + \sum_{k \in S^B \cap b} \frac{y_k}{\pi_k^B} + \sum_{k \in \left(S^A \cup S^B\right) \cap ab} \frac{1}{\pi_k^A + \pi_k^B - \pi_k^A \pi_k^B} y_k. \tag{2.5}$$

3 Estimation avec un premier degré de tirage commun

Nous étudions ici le cas de deux échantillons sélectionnés selon un plan à deux degrés, avec un premier degré de tirage commun. La population U est partitionnée pour obtenir une population $U_I = \{u_1, \ldots, u_M\}$ de M unités primaires d'échantillonnage. Au premier degré, on sélectionne un échantillon S_I d'unités primaires d'échantillonnage (UPE) avec une probabilité de tirage π_{Ii} pour une UPE u_i. Au second degré, dans chaque unité primaire d'échantillonnage $u_i \in S_I$, on sélectionne : un échantillon S_i^A dans

$u_i^A \equiv u_i \cap U_A$, avec une probabilité de sélection (conditionnelle) $\pi_{k|i}^A > 0$ pour $k \in u_i^A$; un échantillon S_i^B dans $u_i^B \equiv u_i \cap U_B$, avec une probabilité de sélection (conditionnelle) $\pi_{k|i}^B > 0$ pour l'unité $k \in u_i^B$. Nous faisons les hypothèses suivantes, habituelles pour un tirage à deux degrés : le second degré de tirage au sein de l'unité primaire d'échantillonnage u_i ne dépend que de i; entre deux unités primaires d'échantillonnage $u_i \neq u_j \in S_I$, les échantillons S_i^A et S_j^A (respectivement, S_i^B et S_j^B) sont indépendants conditionnellement à S_I (propriété d'indépendance). Nous supposons également qu'au sein de chaque unité primaire d'échantillonnage $u_i \in S_I$, les sous-échantillons S_i^A et S_i^B sont indépendants conditionnellement à S_I.

Pour un domaine $d_1 \subset U_A$, le sous-total Y_{d_1} est estimé par $\hat{Y}_{d_1}^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{d_1,i}^A$ avec $d_{Ii} = \left(\pi_{Ii}\right)^{-1}$ le poids de sondage de l'unité primaire d'échantillonnage u_i, $\hat{Y}_{d_1,i}^A = \sum_{k \in S_i^A} d_{k|i}^A y_k 1\left(k \in d_1\right)$ l'estimateur du sous-total $Y_{d_1,i} = \sum_{k \in u_i} y_k 1\left(k \in d_1\right)$ sur $d_1 \cap u_i$, et $d_{k|i}^A = \left(\pi_{k|i}^A\right)^{-1}$ le poids de sondage de k dans u_i^A. Pour un domaine $d_2 \subset U_B$, le sous-total Y_{d_2} est estimé par $\hat{Y}_{d_2}^B = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{d_2,i}^B$ avec $\hat{Y}_{d_2,i}^B = \sum_{k \in S_i^B} d_{k|i}^B y_k 1\left(k \in d_2\right)$ l'estimateur du sous-total $Y_{d_2,i}$ et $d_{k|i}^B = \left(\pi_{k|i}^B\right)^{-1}$ le poids de sondage de k dans u_i^B. On obtient en particulier les estimateurs

$$\hat{Y}_{ab}^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{ab,i}^A \text{ où } \hat{Y}_{ab,i}^A = \sum_{k \in S_i^A} d_{k|i}^A y_k 1\left(k \in ab\right), \tag{3.1}$$

$$\hat{Y}_b^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{b,i}^A \text{ où } \hat{Y}_{b,i}^A = \sum_{k \in S_i^A} d_{k|i}^A y_k 1\left(k \in b\right), \tag{3.2}$$

$$\hat{Y}_{ab}^B = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{ab,i}^B \text{ où } \hat{Y}_{ab,i}^B = \sum_{k \in S_i^B} d_{k|i}^B y_k 1\left(k \in ab\right). \tag{3.3}$$

3.1 Estimateur de Hartley

L'estimateur de Hartley donné en (2.1) peut se réécrire sous la forme

$$\hat{Y}_\theta = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{\theta,i} \tag{3.4}$$

avec $\hat{Y}_{\theta,i} = \hat{Y}_{a,i}^A + \theta \hat{Y}_{ab,i}^A + \left(1-\theta\right) \hat{Y}_{ab,i}^B + \hat{Y}_{b,i}^B$ l'estimateur de Hartley du sous-total Y_i sur l'unité primaire d'échantillonnage u_i. On obtient $E\left(\hat{Y}_\theta \mid S_I\right) = \sum_{i \in S_I} d_{Ii} Y_i$, puis

$$V\left(\hat{Y}_\theta\right) = V\left(\sum_{i \in S_I} d_{Ii} Y_i\right) + EV\left(\hat{Y}_\theta \mid S_I\right). \tag{3.5}$$

Dans (3.5), le premier terme du membre de droite ne dépend pas de θ. L'estimateur optimal de Hartley peut donc se calculer en minimisant seulement le second terme. On obtient :

$$\theta_{opt|S_I} = \frac{EV\left(\hat{Y}_{ab}^B \mid S_I\right) + ECov\left(\hat{Y}_{ab}^B, \hat{Y}_b^B \mid S_I\right) - ECov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A \mid S_I\right)}{EV\left(\hat{Y}_{ab}^A \mid S_I\right) + EV\left(\hat{Y}_{ab}^B \mid S_I\right)}, \tag{3.6}$$

que l'on peut estimer par

$$\hat{\theta}_{opt} = \frac{\hat{V}\left(\hat{Y}_{ab}^{B}\right) + \widehat{Cov}\left(\hat{Y}_{ab}^{B}, \hat{Y}_{b}^{B}\right) - \widehat{Cov}\left(\hat{Y}_{a}^{A}, \hat{Y}_{ab}^{A}\right)}{\hat{V}\left(\hat{Y}_{ab}^{A}\right) + \hat{V}\left(\hat{Y}_{ab}^{B}\right)} \tag{3.7}$$

en remplaçant chaque terme de variance et de covariance par un estimateur sans biais conditionnellement au premier degré.

3.2 Estimateur de Kalton et Anderson

Avec le plan de sondage considéré, on a $d_k^A = d_{Ii} d_{k|i}^A$ pour toute unité $k \in u_i^A$, et $d_k^B = d_{Ii} d_{k|i}^B$ pour toute unité $k \in u_i^B$. L'estimateur de Kalton et Anderson donné en (2.4) peut donc se réécrire

$$\hat{Y}_{KA} = \sum_{i \in S_I} d_{Ii} \hat{Y}_{KA,i} \tag{3.8}$$

avec $\hat{Y}_{KA,i} = \sum_{k \in S^A} d_{k|i}^A m_{k|i}^A y_k + \sum_{k \in S^B} d_{k|i}^B m_{k|i}^B y_k$ l'estimateur de Kalton et Anderson du sous-total Y_i, où

$$m_{k|i}^A = \begin{cases} 1 & \text{si } k \in a \cap u_i, \\ \dfrac{d_{k|i}^B}{d_{k|i}^A + d_{k|i}^B} & \text{si } k \in ab \cap u_i, \end{cases} \quad \text{et} \quad m_{k|i}^B = \begin{cases} 1 & \text{si } k \in b \cap u_i, \\ \dfrac{d_{k|i}^A}{d_{k|i}^A + d_{k|i}^B} & \text{si } k \in ab \cap u_i. \end{cases}$$

3.3 Estimateur de Bankier

Avec le plan de sondage considéré, on a $\pi_k^{HT} = \pi_{Ii}\left(\pi_{k|i}^A + \pi_{k|i}^B - \pi_{k|i}^A \pi_{k|i}^B\right)$ pour tout $k \in u_i$. L'estimateur de Bankier donné en (2.5) peut donc se réécrire

$$\hat{Y}_{HT} = \sum_{i \in S_I} d_{Ii} \hat{Y}_{HT,i} \tag{3.9}$$

avec $\hat{Y}_{HT,i} = \sum_{k \in S_i^A \cup S_i^B} \left(y_k / \pi_{k|i}^{HT}\right)$ l'estimateur de Bankier pour le sous-total Y_i, et $\pi_{k|i}^{HT} = \pi_{k|i}^A$ si $k \in a$, $\pi_{k|i}^{HT} = \pi_{k|i}^B$ si $k \in b$, $\pi_{k|i}^{HT} = \pi_{k|i}^A + \pi_{k|i}^B - \pi_{k|i}^A \pi_{k|i}^B$ si $k \in ab$.

Chacun des trois estimateurs étudiés s'obtient donc en appliquant la méthode d'estimation UPE par UPE, conditionnellement au premier degré. Ce résultat est particulièrement intéressant pour la méthode optimale de Hartley, puisque l'estimateur du coefficient optimal donné en (3.7) ne nécessite que des estimateurs de variance conditionnels au premier degré.

4 Étude par simulations

Nous utilisons des populations artificielles proposées par Saigo (2010). Nous générons deux populations, contenant chacune $M = 200$ unités primaires d'échantillonnage regroupées en $H = 4$ strates U_{Ih} de taille $M_h = 50$. Chaque unité primaire d'échantillonnage u_{hi} contient $N_{hi} = 100$ unités

secondaires. Dans chaque population, nous générons pour chaque unité primaire d'échantillonnage $u_{hi} \in U_{Ih}$:

$$\mu_{hi} = \mu_h + \sigma_h v_{hi} \tag{4.1}$$

où les valeurs μ_h et σ_h sont celles utilisées par Saigo (2010). Le terme σ_h^2 permet de contrôler la dispersion entre les unités primaires d'échantillonnage. Les v_{hi} sont générés de façon iid selon une loi normale centrée réduite $N(0,1)$. Pour chaque unité $k \in u_{hi}$, nous générons ensuite la valeur y_k selon le modèle

$$y_k = \mu_{hi} + \left\{\rho^{-1}\left(1-\rho\right)\right\}^{0,5} \sigma_h v_k, \tag{4.2}$$

où les v_k sont générés de façon iid selon une loi normale centrée réduite. Le terme de variance dans le modèle (4.2) permet d'obtenir un coefficient de corrélation intra-grappes approximativement égal à ρ. En particulier, plus le coefficient ρ est grand, moins les valeurs y_k sont dispersées dans les unités primaires d'échantillonnage. Nous utilisons $\rho = 0,2$ pour la population 1 et $\rho = 0,5$ pour la population 2, ce qui traduit une moindre dispersion de la variable y dans la population 2. La base de sondage U_A correspond à l'ensemble des unités secondaires, et la partie correspondante de u_{hi} est $u_{hi}^A = u_{hi}$, de taille $N_{hi}^A = N_{hi}$. On génère pour chaque unité secondaire k une valeur u_k selon une loi uniforme sur $[0,1]$. La base de sondage U_B correspond aux unités secondaires k telles que $u_k \leq 0,5$, et la partie correspondante de u_{hi} est $u_{hi}^B = u_{hi} \cap U_B$ de taille N_{hi}^B. On se trouve donc dans la situation où $ab = U_B$ et $b = \varnothing$. Le cadre retenu dans les simulations correspond à celui des enquêtes auprès des ménages de l'INSEE, avec extension pour cibler une sous-population spécifique. Pour ces enquêtes, un échantillon S_I de communes (ou de regroupements de communes) est tout d'abord sélectionné au premier degré. Un sous-échantillon S_i^A de logements est ensuite sélectionné dans chaque $u_i \in S_I$; l'échantillon réunion $S^A = \bigcup_{u_i \in S_I} S_i^A$ représente la population entière de logements $U_A = U$. Un second sous-échantillon S_i^B de logements est ensuite sélectionné au sein d'une sous-population de chaque $u_i \in S_I$, afin de cibler une sous-population spécifique U_B (par exemple, logements situés dans une Zone Urbaine Sensible); l'échantillon réunion $S^B = \bigcup_{u_i \in S_I} S_i^B$ ne représente que la sous-population ciblée U_B.

Dans chacune des deux populations ainsi constituées, on pratique plusieurs échantillonnages concurrents; le tableau 4.1 présente pour chaque population les huit combinaisons possibles de tailles d'échantillon par strate aux premier et second degré, ainsi que les valeurs μ_h et σ_h. Au premier degré on sélectionne indépendamment dans chaque strate U_{Ih} : soit un échantillon S_{Ih} de $m_h = 5$ unités primaires d'échantillonnage par sondage aléatoire simple; soit un échantillon S_{Ih} de $m_h = 25$ unités primaires d'échantillonnage par sondage aléatoire simple. Au second degré, on sélectionne dans chaque $u_{hi} \in S_{Ih}$: soit un échantillon S_{hi}^A de taille $n_{hi}^A = 10$ par sondage aléatoire simple dans u_{hi}^A; soit un échantillon S_{hi}^A de taille $n_{hi}^A = 40$ par sondage aléatoire simple dans u_{hi}^A. Au second degré, on sélectionne également dans chaque $u_{hi} \in S_{Ih}$: soit un échantillon S_{hi}^B de taille $n_{hi}^B = 5$ par sondage aléatoire simple dans u_{hi}^B; soit un échantillon S_{hi}^B de taille $n_{hi}^B = 20$ par sondage aléatoire simple dans u_{hi}^B. On note également $f_{hi}^A = \left(N_{hi}^A\right)^{-1} n_{hi}^A$ et $f_{hi}^B = \left(N_{hi}^B\right)^{-1} n_{hi}^B$ les taux de sondage dans u_{hi}^A et u_{hi}^B.

Tableau 4.1
Paramètres utilisés dans chaque strate pour générer les deux populations et sélectionner les échantillons

	Tailles d'échantillon par strate			Paramètres							
				Strate 1		Strate 2		Strate 3		Strate 4	
	m_h	n_{hi}^A	n_{hi}^B	μ_h	σ_h	μ_h	σ_h	μ_h	σ_h	μ_h	σ_h
Population 1	5 ou 25	10 ou 40	5 ou 20	200	20	150	15	120	12	100	10
Population 2	5 ou 25	10 ou 40	5 ou 20	200	10	150	7,5	120	6	100	5

Pour chaque échantillon, on calcule l'estimateur de Hartley donné en (3.4) avec soit $\theta = 1/2$ (HART1), soit pour valeur de θ l'estimateur du coefficient optimal donné en (3.7) (HART2), avec

$$\hat{V}\left(\hat{Y}_{ab}^A\right) = \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{1h}}\left(N_{hi}^A\right)^2 \frac{1-f_{hi}^A}{n_{hi}^A\left(n_{hi}^A-1\right)} \sum_{k\in S_{hi}^A}\left\{y_k 1(k\in ab) - \bar{y}_{ab;S_{hi}^A}\right\}^2,$$

$$\hat{V}\left(\hat{Y}_{ab}^B\right) = \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{1h}}\left(N_{hi}^B\right)^2 \frac{1-f_{hi}^B}{n_{hi}^B\left(n_{hi}^B-1\right)} \sum_{k\in S_{hi}^B}\left\{y_k 1(k\in ab) - \bar{y}_{ab;S_{hi}^B}\right\}^2,$$

$$\widehat{Cov}\left(\hat{Y}_a^A, \hat{Y}_{ab}^A\right) = \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{1h}}\left(N_{hi}^A\right)^2 \frac{1-f_{hi}^A}{n_{hi}^A\left(n_{hi}^A-1\right)} \sum_{k\in S_{hi}^A}\left\{y_k 1(k\in a) - \bar{y}_{a;S_{hi}^A}\right\}\left\{y_k 1(k\in ab) - \bar{y}_{ab;S_{hi}^A}\right\},$$

en notant $\bar{y}_{d;V}$ la moyenne de la variable $y_k 1(k\in d)$ sur une partie $V\subset U$. Pour chaque échantillon, on calcule également l'estimateur de Kalton et Anderson (KALT) donné en (3.8), l'estimateur de Bankier (BANK) donné en (3.9), et l'estimateur de Horvitz-Thompson $\hat{Y}^A$ basé sur le seul échantillon S^A (HTA). La procédure d'échantillonnage est répétée $10\ 000$ fois. Pour mesurer le biais d'un estimateur $\hat{Y}$, nous calculons son biais relatif de Monte Carlo

$$BR_{MC}\left(\hat{Y}\right) = \frac{E_{MC}\left(\hat{Y}\right) - Y}{Y} \times 100$$

avec $E_{MC}\left(\hat{Y}\right) = (1/10\ 000)\sum_{b=1}^{10\ 000}\hat{Y}_{(b)}$, et $\hat{Y}_{(b)}$ la valeur de l'estimateur $\hat{Y}$ pour l'échantillon b. Pour mesurer la variabilité de $\hat{Y}$, nous calculons son erreur quadratique moyenne de Monte Carlo

$$EQM_{MC}\left(\hat{Y}\right) = \frac{1}{10\ 000}\sum_{b=1}^{10\ 000}\left(\hat{Y}_{(b)} - Y\right)^2.$$

Les résultats sont donnés dans le tableau 4.2. Comme l'a souligné un arbitre, les performances de l'estimateur HTA ne dépendent pas de la taille d'échantillon n_{hi}^B choisie. Par souci de cohérence, nous indiquons donc dans le tableau 4.2 les résultats obtenus dans les simulations avec $n_{hi}^B = 5$ uniquement. À tailles d'échantillon m_h et n_{hi}^A identiques, les mêmes résultats sont reportés dans le cas $n_{hi}^B = 20$.

Tous les estimateurs sont virtuellement sans biais. L'estimateur HART2 donne les meilleurs résultats en termes d'erreur quadratique moyenne, comme on pouvait s'y attendre. L'estimateur HTA donne des résultats quasiment équivalents. Ce résultat s'explique par le fait que le coefficient optimal est proche de 1 (dans les simulations, $\hat{\theta}_{opt}$ est compris entre $0{,}80$ et $1{,}06$), et que dans ce cas la formule (2.1) montre que

les estimateurs HART2 et HTA sont très proches : nous présentons en annexe des conditions générales sous lesquelles cette propriété est approximativement vérifiée. Parmi les trois autres estimateurs, HART1 donne les meilleurs résultats, avec une erreur quadratique moyenne plus faible ou équivalente à celle de KALT et BANK dans 11 cas sur 16.

Tableau 4.2
Biais relatif et erreur quadratique moyenne de cinq estimateurs

				HART1		HART2		KALT		BANK		HTA	
Pop.	m_h	n_{hi}^A	n_{hi}^B	**BR**	**EQM**	**BR**	**EQM**	**BR**	**EQM**	**BR**	**EQM**	**BR**	**EQM**
				(%)	$\times 10^9$	**(%)**	$\times 10^9$	**(%)**	$\times 10^9$	**(%)**	$\times 10^9$	**(%)**	$\times 10^9$
1	5	10	5	0,05	7,76	0,01	5,70	0,05	7,79	0,06	8,56	0,04	5,75
1	5	10	20	0,01	7,57	-0,05	5,57	0,03	11,36	0,04	12,75	0,04	5,75
1	5	40	5	0,01	5,01	-0,02	4,51	-0,02	4,57	-0,02	4,81	-0,02	4,52
1	5	40	20	0,00	4,65	-0,01	4,33	0,00	4,66	0,00	5,22	-0,02	4,52
1	25	10	5	-0,03	1,19	-0,02	0,78	-0,03	1,20	-0,02	1,34	-0,01	0,78
1	25	10	20	-0,01	1,17	0,00	0,78	-0,03	1,94	-0,03	2,22	-0,01	0,78
1	25	40	5	0,00	0,62	0,01	0,51	0,00	0,52	0,00	0,57	0,01	0,51
1	25	40	20	0,02	0,58	0,01	0,51	0,02	0,58	0,02	0,68	0,01	0,51
2	5	10	5	0,00	3,59	0,01	1,15	0,00	3,56	0,02	4,38	0,01	1,15
2	5	10	20	0,00	3,60	-0,02	1,15	0,00	7,38	0,00	8,76	0,01	1,15
2	5	40	5	0,00	1,48	0,01	1,07	0,00	1,13	0,01	1,35	0,01	1,07
2	5	40	20	0,00	1,49	-0,01	1,09	0,00	1,49	0,00	2,03	0,01	1,07
2	25	10	5	0,00	0,63	0,00	0,14	0,00	0,63	0,00	0,78	0,00	0,14
2	25	10	20	0,00	0,62	0,00	0,13	0,00	1,38	0,00	1,67	0,00	0,14
2	25	40	5	0,00	0,20	0,00	0,12	0,00	0,13	0,00	0,18	0,00	0,12
2	25	40	20	0,00	0,20	0,00	0,12	0,00	0,20	0,01	0,31	0,00	0,12

Pour chaque estimateur, toutes choses égales par ailleurs, l'erreur quadratique moyenne est plus faible dans la population 2 que dans la population 1. Ce résultat provient du fait que la variance due au premier degré de tirage, qui est la même pour chaque estimateur et vaut

$$V\left(\sum_{i\in S_I} d_{Ii}Y_i\right) = \sum_{h=1}^{H} M_h^2\left(\frac{1}{m_h}-\frac{1}{M_h}\right)S_{Y;U_{Ih}}^2, \tag{4.3}$$

est plus grande dans la population 1 : le terme de dispersion $S_{Y;U_{Ih}}^2 = \left(M_h - 1\right)^{-1}\sum_{u_i\in U_{Ih}}\left(Y_i - \overline{Y}_{U_{Ih}}\right)^2$ augmente avec σ_h^2 et, dans une moindre mesure, augmente quand ρ diminue. L'erreur quadratique moyenne diminue pour chaque estimateur quand le nombre m_h d'unités primaires d'échantillonnage tirées dans chaque strate augmente, car dans ce cas le terme de variance commun donné en (4.3) diminue. De façon analogue, l'erreur quadratique moyenne diminue pour chaque estimateur quand n^A augmente, car dans ce cas la variance due au second degré de tirage diminue. Pour les estimateurs HART1 et HART2, l'erreur quadratique moyenne est stable quand n^B augmente, et de façon plus surprenante pour les estimateurs KALT et BANK l'erreur quadratique moyenne augmente quand n^B augmente. Ce résultat quelque peu contre-intuitif est dû à la conjonction de deux faits. D'une part, la contribution de l'échantillon S^B à la variance due au second degré de tirage est faible : l'augmentation de n^B peut diminuer cette variance, mais même dans ce cas la réduction globale de variance est marginale. D'autre part, dans le cas

des estimateurs KALT et BANK, la contribution de l'échantillon S^A à la variance due au second degré de tirage augmente quand n^B augmente.

Dans le cas de KALT, l'estimateur peut se réécrire

$$\hat{Y}_{KA} = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{i \in S_{Ih}} \hat{Y}_{KA,i}$$

avec

$$\hat{Y}_{KA,i} = \frac{1}{f_{hi}^A} \sum_{k \in S_i^A} m_{k|i}^A y_k + \frac{1}{f_{hi}^A + f_{hi}^B} \sum_{k \in S_i^B} y_k \quad \text{et} \quad m_{k|i}^A = \begin{cases} 1 & \text{si } k \in a \cap u_i, \\ \dfrac{f_{hi}^A}{f_{hi}^A + f_{hi}^B} & \text{si } k \in ab \cap u_i. \end{cases} \tag{4.4}$$

Dans (4.4), la dispersion de la variable $m_{k|i}^A$ (donc celle de $m_{k|i}^A y_k$) augmente quand le facteur $f_{hi}^A \big/ \left(f_{hi}^A + f_{hi}^B\right)$ s'éloigne de 1. Or, ce facteur est proche de 1 quand f_{hi}^B est faible devant f_{hi}^A (donc n^B petit devant n^A), mais s'en éloigne quand n^B augmente. Notons que la variance (conditionnelle à S_I) du second terme de $\hat{Y}_{KA,i}$ est égale à

$$V\left(\frac{1}{f_{hi}^A + f_{hi}^B} \sum_{k \in S_i^B} y_k \,\middle|\, S_I \right) = \left(N_{hi}^A\right)^2 N_{hi}^B \times \frac{n_{hi}^B \left(N_{hi}^B - n_{hi}^B\right)}{\left(N_{hi}^B n_{hi}^A + N_{hi}^A n_{hi}^B\right)^2} \times S_{u_{hi}^B}^2$$

avec $S_{u_{hi}^B}^2 = \left(N_{hi}^B - 1\right)^{-1} \sum_{k \in u_{hi}^B} \left(y_k - \bar{y}_{u_{hi}^B}\right)^2$. Cette variance ne décroît pas forcément quand n_{hi}^B augmente. Par exemple, l'un des cas considéré dans les simulations correspond à $N_{hi}^A = 100$, $N_{hi}^B \simeq 50$ et $n_{hi}^A = 40$. Dans ce cas, le terme $n_{hi}^B \left(N_{hi}^B - n_{hi}^B\right) \big/ \left(N_{hi}^B n_{hi}^A + N_{hi}^A n_{hi}^B\right)^2$ atteint sa valeur maximale pour $n_{hi}^B = 11$.

Dans le cas de BANK, l'estimateur peut se réécrire

$$\hat{Y}_{HT} = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{i \in S_{Ih}} \hat{Y}_{HT,i}$$

avec

$$\hat{Y}_{HT,i} = \sum_{k \in S_i^A \cup S_i^B} \frac{y_k}{\pi_{k|i}^{HT}} \quad \text{et} \quad \pi_{k|i}^{HT} = \begin{cases} f_{hi}^A & \text{si } k \in a, \\ f_{hi}^A + f_{hi}^B \left(1 - f_{hi}^A\right) & \text{si } k \in ab. \end{cases} \tag{4.5}$$

Dans (4.5), la dispersion de la variable $\pi_{k|i}^{HT}$ augmente quand le facteur $f_{hi}^B \left(1 - f_{hi}^A\right)$ augmente. Or, ce facteur est proche de 0 quand n_{hi}^B (et donc f_{hi}^B) est faible, mais s'accroît quand n_{hi}^B augmente.

5 Conclusion

Nous avons étudié les estimateurs de Hartley (1962), de Kalton et Anderson (1986) et de Bankier (1986) pour mettre en commun les échantillons issus de deux vagues d'enquête. Nous avons plus particulièrement étudié le cas où un échantillon représente la population entière (échantillon complètement représentatif), alors que le second n'en représente qu'une partie (échantillon partiellement représentatif).

Dans le cadre considéré dans les simulations (voir également l'annexe pour un cadre plus général), l'utilisation de l'échantillon partiellement représentatif ne permet pas de gagner en précision : si sa taille augmente, la précision des estimateurs de la classe de Hartley reste stable ou s'améliore légèrement, alors que la précision des estimateurs de Kalton et Anderson et de Bankier se dégrade. L'estimateur optimal de Hartley lui-même, bien que plus complexe à calculer, offre une précision qui n'est que légèrement améliorée par rapport à l'estimateur de Horvitz-Thompson classique calculé sur l'échantillon complètement représentatif. Bien que notre étude par simulations soit limitée, ces résultats suggèrent d'être prudents dans le choix d'un estimateur en présence de bases de sondage multiples, et qu'un estimateur simple est parfois préférable, même s'il n'utilise qu'une partie de l'information collectée.

Remerciements

Les auteurs remercient un éditeur associé et un arbitre pour leur lecture attentive et leurs remarques qui ont permis d'améliorer significativement l'article, et David Haziza pour des discussions utiles.

Annexe

A.1 Comparaison entre l'estimateur optimal de Hartley et l'estimateur de Horvitz-Thompson

Nous reprenons le cadre et les notations de la section 4 : les échantillons S^A et S^B sont sélectionnés selon un plan à deux degrés avec un premier degré de tirage commun. On utilise un sondage aléatoire simple stratifié au premier degré, et un sondage aléatoire simple au second degré dans chaque unité primaire d'échantillonnage. La base de sondage U_A correspond à la population entière, alors que la base de sondage U_B ne recouvre qu'une partie de la population.

Dans le cas de l'estimateur optimal de Hartley, la formule (3.6) donne

$$\theta_{opt|S_I} = \frac{EV\left(\hat{Y}_{ab}^B \mid S_I\right) - ECov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A \mid S_I\right)}{EV\left(\hat{Y}_{ab}^B \mid S_I\right) + EV\left(\hat{Y}_{ab}^A \mid S_I\right)}.$$

Après un peu de calcul, nous obtenons

$$EV\left(\hat{Y}_{ab}^A \mid S_I\right) = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{u_{hi} \in U_{Ih}} \left(N_{hi}\right)^2 \frac{1 - f_{hi}^A}{n_{hi}^A} \left\{ \frac{N_{hi}^B - 1}{N_{hi} - 1} S_{u_{hi}^B}^2 + \frac{N_{hi}^B \left(N_{hi} - N_{hi}^B\right)\left(\bar{y}_{u_{hi}^B}\right)^2}{N_{hi}\left(N_{hi} - 1\right)} \right\}, \tag{A.1}$$

$$-ECov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A \mid S_I\right) = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{u_{hi} \in U_{Ih}} \left(N_{hi}\right)^2 \frac{1 - f_{hi}^A}{n_{hi}^A} \left\{ \frac{N_{hi}^B \left(\bar{y}_{u_{hi}^B}\right)\left(N_{hi}\bar{y}_{u_{hi}} - N_{hi}^B \bar{y}_{u_{hi}^B}\right)}{N_{hi}\left(N_{hi} - 1\right)} \right\}$$

avec $\bar{y}_{u_{hi}} = \left(N_{hi}\right)^{-1} \sum_{k \in u_{hi}} y_k$, $\bar{y}_{u_{hi}^B} = \left(N_{hi}^B\right)^{-1} \sum_{k \in u_{hi}^B} y_k$ et $S_{u_{hi}^B}^2 = \left(N_{hi}^B - 1\right)^{-1} \sum_{k \in u_{hi}^B} \left(y_k - \bar{y}_{u_{hi}^B}\right)^2$.

L'estimateur de Horvitz-Thompson basé sur le seul échantillon S^A et l'estimateur optimal de Hartley coïncident si le coefficient $\theta_{opt|S_I}$ est égal à 1, ce qui est le cas si $EV\left(\hat{Y}_{ab}^A \mid S_I\right) = -ECov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A \mid S_I\right)$. Cette condition sera en particulier vérifiée si dans (A.1) les termes entre accolades coïncident pour chaque unité primaire d'échantillonnage u_{hi}. On aura donc $\theta_{opt|S_I} \simeq 1$ si

$$\forall \quad u_{hi} \in U_I \quad \frac{N_{hi}\left(N_{hi}^B - 1\right)}{N_{hi}^B} \frac{S_{u_{hi}^B}^2}{\overline{y}_{u_{hi}^B}\left(N_{hi}\overline{y}_{u_{hi}} - N_{hi}^B \overline{y}_{u_{hi}^B}\right)} + \frac{\left(N_{hi} - N_{hi}^B\right)\overline{y}_{u_{hi}^B}}{N_{hi}\overline{y}_{u_{hi}} - N_{hi}^B \overline{y}_{u_{hi}^B}} \simeq 1. \tag{A.2}$$

Supposons que la valeur moyenne de y soit approximativement la même dans les bases U_A et U_B pour chaque unité primaire d'échantillonnage, c'est-à-dire que $\forall\ u_{hi} \in U_I \quad \overline{y}_{u_{hi}^B} \simeq \overline{y}_{u_{hi}}$. Alors la condition (A.2) sera approximativement vérifiée si $\forall\ u_{hi} \in U_I \quad cv_{u_{hi}^B}^2$ est proche de 0, avec $cv_{u_{hi}^B} = \sqrt{S_{u_{hi}^B}^2} \Big/ \overline{y}_{u_{hi}^B}$.

En résumé, l'estimateur de Horvitz-Thompson basé sur le seul échantillon S^A et l'estimateur optimal de Hartley seront proches si au sein de chaque unité primaire d'échantillonnage u_{hi} : (a) la valeur moyenne de y est peu différente entre les deux bases, et (b) la variable y est faiblement dispersée au sein de u_{hi}^B. Dans les simulations, la condition (a) est approximativement respectée car la répartition des individus entre les bases de sondage U_A et U_B se fait complètement aléatoirement; la condition (b) est approximativement respectée avec des valeurs de $cv_{u_{hi}^B}^2$ variant de $0{,}02$ à $0{,}10$ pour la population 1, et de $0{,}001$ à $0{,}005$ pour la population 2.

Bibliographie

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

Bourdalle, G., Christine, M. et Wilms, L. (2000). Échantillons maître et emploi. *Série INSEE Méthodes*, 21, p. 139-173.

Hansen, M.H. et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, p. 333-362.

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

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, p. 663-685.

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

Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles (Belgique) et Éditions Ellipses (France).

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

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

Lohr, S.L. (2009). Multiple frame surveys. Dans *Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, Eds., D. Pfeffermann et C.R. Rao. Amsterdam: North Holland, Vol. 29A, p. 71-88.

Lohr, S.L. (2011). Autres plans de sondage : échantillonnage avec bases de sondage multiples chevauchantes. *Techniques d'enquête*, Vol.37 no.2, p. 213-232.

Mecatti, F. (2007). Un estimateur à base de sondage unique fondé sur la multiplicité pour les sondages à bases multiples. *Techniques d'enquête*, Vol.33 no.2, p. 171-178.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, p. 169-175.

Rao, J.N.K. et Wu, C. (2010). Pseudo-empirical likelihood inference for dual frame surveys. *Journal of the American Statistical Association*, 105, p. 1494-1503.

Saigo, H. (2010). Comparing four bootstrap methods for stratified three-stage sampling. *Journal of Official Statistics*, Vol. 26, No. 1, 2010, p. 193–207.

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 379-387
Statistique Canada, n° 12-001-X au catalogue

Combinaison de l'information de plusieurs enquêtes complexes

Qi Dong, Michael R. Elliott et Trivellore E. Raghunathan[1]

Résumé

Le présent document décrit l'utilisation de l'imputation multiple pour combiner l'information de plusieurs enquêtes de la même population sous-jacente. Nous utilisons une nouvelle méthode pour générer des populations synthétiques de façon non paramétrique à partir d'un bootstrap bayésien fondé sur une population finie qui tient systématiquement compte des plans d'échantillonnage complexes. Nous analysons ensuite chaque population synthétique au moyen d'un logiciel standard de données complètes pour les échantillons aléatoires simples et obtenons une inférence valide en combinant les estimations ponctuelles et de variance au moyen des extensions de règles de combinaison existantes pour les données synthétiques. Nous illustrons l'approche en combinant les données de la *National Health Interview Survey* (NHIS) de 2006 et de la *Medical Expenditure Panel Survey* (MEPS) de 2006.

Mots-clés : Populations synthétiques; répartition prédictive a posteriori; bootstrap bayésien; échantillonnage inverse.

1 Introduction

Il arrive souvent que les organismes d'enquête tirent de multiples échantillons à partir de populations similaires et recueillent des variables semblables, parfois même en utilisant la même base de sondage. Par exemple, la *National Health Interview Survey* (NHIS) et la *National Health and Nutrition Examination Survey* (NHANES) sont toutes deux réalisées par le *National Center for Health Statistics* des États-Unis. Ces deux enquêtes ciblent la population non institutionnalisée des États-Unis et leurs questions se recoupent considérablement. En combinant l'information provenant de diverses enquêtes, nous espérons obtenir une inférence plus exacte pour la population que si nous utilisions les données d'une seule enquête.

L'une des plus grandes difficultés liées à une telle combinaison d'information concerne la compatibilité de diverses sources de données. Les enquêtes peuvent utiliser différents plans d'échantillonnage ou modes de collecte des données, ce qui peut donner lieu à diverses propriétés d'erreur d'échantillonnage et d'erreur non due à l'échantillonnage. Au lieu de compiler directement les données de plusieurs enquêtes à partir d'une analyse simple, nous devons corriger pour tenir compte des écarts entre les données et les rendre comparables.

Différentes méthodes pour la combinaison de données tirées de deux enquêtes différentes ont été proposées dans la documentation sur les méthodes d'enquête (Hartley 1974; Skinner et Rao 1996; Lohr et Rao 2000; Elliott et Davis 2005; Raghunathan, Xie, Schenker, Parsons, Davis, Dodd et Feuer 2007; Schenker, Gentleman, Rose, Hing et Shimizu 2002; Schenker et Raghunathan 2007; Schenker, Raghunathan et Bondarenko 2009). Les derniers travaux de Raghunathan et coll. (2007) et Schenker et coll. (2009) appliquaient des approches fondées sur un modèle. Le principe pour l'approche fondée sur un

1. Qi, Dong, Google, Inc., 1R4A, Quad 5, Google Inc, 399 N. Whisman Road, Mountain View, CA 94043. Courriel : qdong@google.com; Michael R. Elliott, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109 et Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. Courriel : mrelliot@umich.edu; Trivellore E. Raghunathan, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109 et Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. Courriel : teraghu@umich.edu.

modèle est d'intégrer un modèle d'imputation aux données de meilleure qualité et d'utiliser le modèle intégré pour imputer les valeurs dans les autres échantillons de qualité inférieure. Dans la mesure où le modèle d'imputation est correctement défini, cette approche peut tirer profit des forces des différentes sources de données et améliorer l'inférence statistique. Cependant, comme suggéré par Reiter, Raghunathan et Kinney (2006), lorsque l'échantillon est recueilli au moyen de plans d'échantillonnage complexes, le fait de ne pas tenir compte de ces caractéristiques peut entraîner des estimations biaisées de la perspective fondée sur le plan. Cependant, il est très difficile de tenir pleinement compte des caractéristiques du plan d'échantillonnage complexe en pratique. Par exemple, Raghunathan et coll. (2007) et Schenker et coll. (2009) ont utilisé une méthode simplifiée pour tenir compte de la stratification et de la mise en grappe. Raghunathan et coll. (2007) ont utilisé un concept rudimentaire de l'effet de plan et Schenker et coll. (2009) ont utilisé des scores de propension pour créer des sous-groupes de correction pour la modélisation.

Ici, nous proposons une nouvelle méthode pour combiner plusieurs enquêtes; cette méthode tient compte des caractéristiques du plan d'échantillonnage complexe dans chaque enquête. La population non observée dans chaque enquête sera traitée comme des données manquantes à imputer. Le modèle d'imputation tiendra compte des caractéristiques du plan complexe au moyen d'une nouvelle méthode de génération de la population synthétique non paramétrique (Dong, Elliott et Raghunathan 2014). Pour chaque enquête, les données observées et la population non observée imputée produisent des populations synthétiques multiples. Une fois la population complète obtenue, les caractéristiques du plan d'échantillonnage complexe, comme la stratification, la mise en grappe et la pondération, seront inutiles dans l'analyse, et les populations synthétiques peuvent être traitées comme des échantillons aléatoires simples équivalents. Enfin, l'estimation de la quantité de population d'intérêt sera calculée à partir de chaque population synthétique et sera combinée d'abord à chaque enquête individuelle, puis à plusieurs enquêtes.

Le présent document procède comme suit : la section 2 résume la génération de la population synthétique tout en tenant compte des caractéristiques du plan d'échantillonnage complexe au moyen de l'approche non paramétrique. La section 3 décrit la méthodologie pour produire les estimations combinées à partir de ces populations synthétiques multiples. À la section 4, nous appliquons la méthode proposée pour combiner la NHIS de 2006 et la *Medical Expenditure Panel Survey* (MEPS) afin d'estimer les taux de couverture de l'assurance-maladie de la population des États-Unis. La section 5 conclut par une discussion et des orientations pour les recherches à venir.

2 Production de populations synthétiques à partir des données d'une seule enquête en tenant compte des plans d'échantillonnage complexes

Dong et coll. (2014) ont poursuivi les travaux relatifs au bootstrap bayésien pour la population finie afin d'élaborer une approche non paramétrique de la génération de distributions prévisionnelles postérieures. Voici un résumé de l'algorithme pour tirer les populations synthétiques $l,\ l = 1,...,L$ pour les plans d'échantillonnage en grappes stratifiés ayant des probabilités inégales de sélection :

1. Utilisez le bootstrap bayésien (BB) (Rubin 1981) pour tenir compte de la stratification et de la mise en grappe. Tirez un échantillon aléatoire simple avec remise (EASAR) de taille m_h à partir des grappes c_h dans chaque strate $h = 1,...,H$ et calculez les poids de rééchantillonnage bootstrap pour chacune des observations n_{hi} dans chaque grappe comme suit : $w^{*(l)} = \left\{ w_{hi}^{*(l)},\ h = 1,\ldots,H,\ i = 1,\ldots,c_h,\ k = 1,\ldots,n_{hi} \right\}$, où $w_{hik}^{*} = w_{hik}\left(\left(1 - \sqrt{\left(m_h / c_h - 1\right)}\right) + \sqrt{\left(m_h / c_h - 1\right)}\left(c_h / m_h\right) m_{hi}^{*}\right)$ et m_{hi}^{*} indique le nombre de fois qu'une grappe i, $i = 1,\ldots,c_h$ est sélectionnée. Pour que tous les poids de rééchantillonnage soient non négatifs, $m_h \le \left(c_h - 1\right)$; ici et ci-après, nous supposons que $m_h = \left(c_h - 1\right)$.

2. Utilisez le bootstrap bayésien pour la population finie (BBPF) (Lo 1986; Cohen 1997) pour des probabilités de sélection inégales pour tenir compte des probabilités inégales de sélection. Pour chaque grappe i dans la strate h où la taille de la population est de N_{hi}, tirez un échantillon de taille $N_{hi} - n_{hi}$, indiquée par $\left(y_1^{*},\ldots,y_{N_{hi}-n_{hi}}^{*}\right)$, en tirant y_{hik}^{*} des données relatives à une grappe $\left(y_1,\ldots,y_{n_{hi}}\right)$ avec la probabilité $\dfrac{w_{hik}^{*} - 1 + l_{hik,j-1} * \left(N_{hi} - n_{hi}\right)/n_{hi}}{N_{c_H} - n_{c_H} + \left(j-1\right) * \left(N_{hi} - n_{hi}\right)/n_{hi}}$, où w_{hik}^{*} est le poids de rééchantillonnage de l'unité k dans la grappe i de la strate h, et $l_{hik,j-1}$ est le nombre de sélections bootstrap de y_{hik} dans $y_1^{*},\ldots,y_{j-1}^{*}$. Créez la population BBPF $y_1,\ldots,y_{n_{hi}}, y_1^{*},\ldots,y_{N_{hi}-n_{hi}}^{*}$.

3. Produisez F échantillons BBPF pour chaque échantillon BB, représenté par $S_{l1},\ldots,S_{lF}, l = 1,\ldots,L.$ Rassemblez les F échantillons BBPF afin de produire une population synthétique, S_l. (parce que $N = \sum_h \sum_i N_{hi}$ pourrait avoir une taille déraisonnablement grande, la production d'un échantillon de taille $k * n$ pour une grande valeur de k est suffisante).

3 Règle de combinaison pour les populations synthétiques d'enquêtes multiples

Supposons que $Q = Q\left(Y\right)$ est la quantité de population d'intérêt selon l'ensemble de variables Y qui sont recueillies au cours de plusieurs enquêtes : par exemple, une moyenne de population, une proportion ou un total, un vecteur des coefficients de régression, etc. Par souci de simplicité de l'exposition, nous supposons que Q est scalaire. Supposons qu'au moyen des données d'une seule enquête s, nous créons L populations synthétiques, $S_l^{(s)}$, $l = 1,\ldots,L$, à partir de la méthodologie résumée à la section 2. Désignons $Q_l^{(s)}$ comme l'estimation correspondante de la quantité de population Q obtenue de la population synthétique l générée au moyen des données de l'enquête s (soulignons que cette estimation peut être obtenue à partir d'une hypothèse d'échantillonnage aléatoire simple). Dong et coll. (2014)

démontrent que conformément à des suppositions asymptotiques raisonnables (taille d'échantillon suffisante pour la quantité d'échantillons d'intérêt répartie suivant une distribution normale, populations synthétiques conformes au plan d'enquête),

$$Q \mid S_1^{(s)},\ldots,S_L^{(s)} \overset{\bullet}{\sim} t_{L-1}\left(\bar{Q}_L^{(s)},\left(1+L^{-1}\right)B_L^{(s)}\right) \tag{3.1}$$

où $\bar{Q}_L^{(s)} = L^{-1}\sum_{l=1}^{L} Q_l^{(s)}$ est la moyenne de Q à l'étendue des populations synthétiques L et $B_L^{(s)} = (L-1)^{-1}\sum_{l=1}^{L}\left(Q_l^{(s)} - \bar{Q}_L^{(s)}\right)^2$ est la variance entre les étapes d'imputation. Le résultat suit immédiatement à partir de la section 4.1 de Raghunathan, Reiter et Rubin (2003), et est fondé sur les règles standard de combinaison de l'imputation multiple de Rubin (1987). La variance moyenne de l'imputation « interne » est de zéro, puisque la population complète est synthétisée; par conséquent, la variance a posteriori de Q dépend entièrement de la variance entre les étapes d'imputation.

La règle de combinaison obtenue à (3.1) ne donnera pas nécessairement une inférence valide pour les paramètres d'intérêt pour plusieurs enquêtes, puisque les modèles pour générer les populations synthétiques pour les enquêtes multiples pourraient être différents. Par conséquent, une nouvelle règle pour combiner les estimations dans plusieurs enquêtes doit être élaborée.

3.1 Approximation normale lorsque L est grand

Supposons que $\bar{Q}_L^{(s)}$ et $B_L^{(s)}$ soient l'estimateur combiné de la quantité de population d'intérêt et sa variance pour l'enquête s obtenue au moyen des formules de combinaison pour les populations synthétiques $\mathrm{S}_{syn}^{(s)} = \left\{\mathrm{S}_l^{(s)}, l = 1,\ldots,L\right\}$, $s = 1,\ldots,S$ dans le contexte d'une seule enquête. Lorsque L est grand, nous avons

$$Q \mid S_{syn}^{(1)},\ldots,S_{syn}^{(S)} \overset{\bullet}{\sim} N\left(\bar{Q}_L, B_L\right) \tag{3.2}$$

où $\bar{Q}_L = \sum_{s=1}^{S}\left(\bar{Q}_L^{(s)}/B_L^{(s)}\right)\Big/\sum_{s=1}^{S}\left(1/B_L^{(s)}\right)$ et $B_L = 1\Big/\sum_{s=1}^{S}\left(1/B_L^{(s)}\right)$. L'équation (3.2) suit immédiatement à partir des résultats bayésiens standard, en supposant que 1) la vraie variance de $\bar{Q}_L^{(s)}$, B_s, peut être estimée par $B_L^{(s)}$ obtenue à partir des populations synthétiques comme à la section 3, c.-à-d. $\left(\bar{Q}_L^{(s)} \mid Q, B_s\right) = \left(\bar{Q}_L^{(s)} \mid Q, B_L^{(s)}\right) \sim N\left(Q, B_L^{(s)}\right)$, 2) chaque enquête est indépendante et 3) Q a une répartition a priori non informative $\pi\left(Q \mid B_L^{(s)}\right) \propto 1$.

3.2 Répartition corrigée en fonction de T pour un L petit ou modéré

Pour un L, petit ou modéré, la répartition a posteriori de Q s'estime mieux comme suit

$$Q \mid S_{syn}^{(1)},\ldots,S_{syn}^{(S)} \overset{\bullet}{\sim} \mathrm{t}_{\nu_L}\left(\bar{Q}_L, \left(1+L^{-1}\right)B_L\right) \tag{3.3}$$

où $\bar{Q}_L$ et B_L sont définis comme à 3.1, et les degrés de liberté $\vartheta_L = (L-1) \Big/ \sum_{s=1}^{S} \left(\left(1/b_L^{(s)} \right) \Big/ \sum_{s=1}^{S} \left(1/b_L^{(s)} \right) \right)^2$. Dong (2012) fournit davantage de détails, qui suivent les recherches de Raghuanthan et coll. (2003) ayant servi à déterminer indirectement les résultats pour L grand.

4 Estimations combinées de la couverture d'assurance-maladie de la NHIS, la MEPS et la BRFSS

Les données de la NHIS et de la MEPS de 2006 sont des échantillons probabilistes à plusieurs degrés qui comprennent la stratification, la mise en grappe et le suréchantillonnage de certaines sous-populations (p. ex. les Noirs, les Hispaniques et les Asiatiques au cours des dernières années). Pour des motifs de confidentialité, les vraies strates et les UPE sont supprimées. La NHIS est publiée avec 300 pseudo-strates et deux pseudo-UPE par strate; la MEPS, qui est un sous-échantillon de ménages qui participent à la NHIS, est publiée avec 203 pseudo-strates et jusqu'à trois pseudo-UPE par strate (Ezzati-Rice, Rohde et Greenblatt 2008; National Center for Health Statistics 2007). Dans le cadre de la NHIS et de la MEPS, on demande à un adulte sélectionné au hasard dans chaque ménage s'il a une assurance-maladie et, dans l'affirmative, s'il s'agit d'une assurance privée ou publique. Nous considérons cette répartition trinomiale de la couverture d'assurance dans l'ensemble de la population adulte, ainsi que dans les sous-populations composées d'hommes, d'Hispaniques, de Blancs non hispaniques et de Blancs non hispaniques gagnant de 25 000 $ à 35 000 $ par année. Nous supprimons les cas comportant des données manquantes et nous étudions les cas complets. Nous obtenons ainsi 20 147 et 20 893 cas pour les données de la NHIS et de la MEPS, respectivement.

La BRFSS de 2006 est obtenue au moyen de la composition aléatoire de numéros à partir d'un échantillonnage par liste, stratifié par état. Bien que ce genre de plans évite la mise en grappe, une probabilité inégale de sélection est introduite, parce que la taille de l'échantillon est à peu près égale dans chaque état; en outre, un seul adulte est échantillonné par ménage. Contrairement à la NHIS et à la MEPS, la BRFSS se contente de demander si la personne a une assurance ou pas, ce qui nous permet de calculer uniquement la proportion de répondants qui ne sont pas couverts par une assurance. Nous supprimons les cas ayant des valeurs manquantes aux questions et nous nous concentrons sur notre simulation des cas complets. Il y a 294 559 cas complets dans les données de la BRFSS de 2006.

Nous générons les populations synthétiques pour les trois enquêtes à partir de 200 échantillons BB, chacun comportant 10 échantillons BBPF de taille $5n$ $(B = 200,\ F = 10,\ k = 5)$. Nous produisons alors les estimations combinées des taux de couverture par une assurance-maladie des personnes au moyen de la méthode de combinaison d'enquêtes susmentionnée. Étant donné que pour les trois enquêtes, nous savons si les personnes ont une assurance ou pas, nous pouvons combiner la NHIS, la BRFSS et la MEPS afin d'estimer la proportion de personnes non assurées. Cependant, la BRFSS ne demande pas aux personnes quel type d'assurance elles ont (publique ou privée). Pour ces proportions, nous pouvons seulement combiner la NHIS et la MEPS. Les résultats sont résumés au tableau 4.1. Les estimations de la variance pour l'estimateur combiné sont bien plus petites que celles qui ont été obtenues à partir des données réelles. Plus précisément, la précision des estimations obtenues de la NHIS est accrue de 43 % en

moyenne, l'augmentation la plus prononcée de 98 % étant obtenue par la combinaison de la NHIS et de la MEPS. Les gains de précision pour la MEPS sont encore plus importants. L'augmentation moyenne de la précision pour la MEPS est de 101 %, l'augmentation la plus prononcée étant de 202 %. La précision est accrue davantage lorsque nous combinons les trois enquêtes. Par exemple, pour la proportion de personnes qui ne sont pas assurées, en moyenne, la précision est quintuplée pour la NHIS, multipliée par 1,5 pour la BRFSS et par 4,2 pour la MEPS. Autrement dit, les gains de précision grâce à l'utilisation de l'information de plusieurs enquêtes peuvent être considérables, et plus nous combinons d'information, plus les gains de précision seront importants.

5 Discussion

Dans cet article, nous proposons une nouvelle façon de combiner de l'information de plusieurs enquêtes complexes. Nous appliquons la nouvelle méthode pour combiner de l'information au sujet de la couverture d'assurance-maladie dans le cadre de la NHIS, de la MEPS et de la BRFSS de 2006. Les résultats indiquent que l'estimation combinée est plus précise que les estimations des enquêtes individuelles. Comme l'ont démontré les travaux précédents (Dong et coll. 2014), peu d'information se perd en ce que les propriétés d'échantillonnage des inférences de la population synthétique et de l'échantillon réel sont très semblables. Par conséquent, lorsque nous combinons les estimations de trois échantillons, l'estimation combinée est considérablement plus efficace que les estimations des enquêtes individuelles. (Soulignons que cette application sert principalement à titre d'exemple; des inférences semblables pourraient être faites en calculant les estimations fondées sur le plan et les variances pour chacune des enquêtes, puis en appliquant la règle de combinaison dans (3.2) dans les estimations fondées sur le plan.)

Cette nouvelle méthode de combinaison d'enquêtes offre deux avantages par rapport à la méthodologie existante. D'abord, l'approche utilisée ici pour générer des populations synthétiques, décrite en détail dans Dong et coll. (2014), tient compte du plan de sondage complexe de façon non paramétrique en extrapolant la méthodologie du bootstrap bayésien de la population finie. Étant donné que les populations synthétiques obtenues peuvent être analysées comme des échantillons aléatoires simples, l'information d'autres enquêtes peut être utilisée pour tenir compte des erreurs non dues à l'échantillonnage et/ou pour imputer les variables manquantes. Un autre avantage de cette méthode est qu'elle n'a pas de limite du nombre d'enquêtes à combiner, dans la mesure où les enquêtes ont la même population sous-jacente. La méthode proposée qui tient compte des caractéristiques du plan d'échantillonnage complexe peut être appliquée à chaque enquête indépendamment. Une fois l'information manquante imputée, quel que soit le nombre d'enquêtes à combiner, il nous suffit de combiner les estimations de chaque enquête au moyen de la règle de combinaison décrite dans le présent document. Un dernier avantage de l'approche proposée est la capacité des populations synthétiques générées par la méthode non paramétrique de conserver les valeurs manquantes aux questions dans les données réelles. Cette méthode pourrait combler une lacune dans la zone visée par l'imputation multiple, parce que les méthodes d'imputation existantes ne prennent pas en compte habituellement les caractéristiques du plan d'échantillonnage complexe dans les données et imputent les valeurs manquantes comme s'il s'agissait d'échantillons aléatoires simples. Nous envisageons cette application dans les travaux à venir.

Tableau 4.1
Estimations individuelles et combinées pour la NHIS, la MEPS et la BRFSS de 2006

Domaine	Types	Données réelles (plan complexe)			Estimations combinées	
		NHIS	BRFSS	MEPS	NHIS et MEPS	NHIS, BRFSS et MEPS
Population complète	**Proportion**					
	Régime privé	0,746		0,735	0,741	
	Régime public	0,075		0,133	0,094	
	Non assuré	0,179	0,154	0,132	0,152	0,153
	Variance					
	Régime privé	2,46E-05		2,78E-05	1,61E-05	
	Régime public	6,29E-06		1,44E-05	5,35E-06	
	Non assuré	1,84E-05	3,32E-06	1,41E-05	9,80E-06	2,55E-06
Hommes	**Proportion**					
	Régime privé	0,740		0,735	0,738	
	Régime public	0,060		0,101	0,074	
	Non assuré	0,200	0,167	0,164	0,181	0,172
	Variance					
	Régime privé	3,32E-05		3,87E-05	2,06E-05	
	Régime public	6,82E-06		1,53E-05	5,72E-06	
	Non assuré	2,94E-05	8,88E-06	2,64E-05	1,51E-05	5,61E-06
Hispaniques	**Proportion**					
	Régime privé	0,494		0,506	0,5014	
	Régime public	0,096		0,161	0,1157	
	Non assuré	0,410	0,371	0,334	0,3684	0,3689
	Variance					
	Régime privé	1,24E-04		1,73E-04	9,76E-05	
	Régime public	2,57E-05		8,03E-05	2,66E-05	
	Non assuré	1,23E-04	7,18E-05	1,19E-04	8,71E-05	3,79E-05
Blancs non hispaniques	**Proportion**					
	Régime privé	0,805		0,788	0,796	
	Régime public	0,062		0,116	0,081	
	Non assuré	0,134	0,1059	0,096	0,113	0,107
	Variance					
	Régime privé	2,99E-05		3,35E-05	1,97E-05	
	Régime public	8,20E-06		1,81E-05	6,86E-06	
	Non assuré	2,02E-05	2,15E-06	1,51E-05	1,02E-05	1,90E-06
Blancs non hispaniques ayant un revenu de [25 000 $ à 35 000 $)	**Proportion**					
	Régime privé	0,827		0,813	0,821	
	Régime public	0,039		0,079	0,053	
	Non assuré	0,134	0,173	0,108	0,122	0,154
	Variance					
	Régime privé	1,0E-04		1,39E-04	7,74E-05	
	Régime public	2,82E-05		6,31E-05	2,52E-05	
	Non assuré	7,24E-05	2,78E-05	8,92E-05	5,14E-05	1,93E-05

Bibliographie

Cohen, M.P. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 635-638.

Dong, Q. (2012). Unpublished PhD thesis, University of Michigan.

Dong Q., Elliott, M.R. et Raghunathan T.E. (2014). Une méthode non paramétrique de production de populations synthétiques qui tient compte des caractéristiques des plans de sondage complexes. *Techniques d'enquête*, 40 (1), 33-52.

Elliott, M.R. et Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Journal of the Royal Statistical Society C*: Applied Statistics, 54, 595-609.

Ezzati-Rice, T.M., Rohde, F. et Greenblatt, J. (2008). Sample design of the medical expenditure panel survey household component, 1998–2007. *Methodology Report No. 22*. Agency for Healthcare Research and Quality, Rockville, MD. Consulté au http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr22/mr22.pdf, février 2014.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *The Indian Journal of Statistics*, C, 38, 99-118.

Lo, A.Y. (1986). Bayesian statistical inference for sampling a finite population. *Annals of Statistics*, 14, 1226-1233.

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

National Center for Health Statistics (2007). Data file documentation, National Health Interview Survey, 2006 (machine readable data file and documentation). *National Center for Health Statistics*, Centers for Disease Control and Prevention, Hyattsville, Maryland. Consulté au: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2006/srvydesc.pdf, février 2014.

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

Reiter, J.P., Raghunathan, T.E. et Kinney, S.K. (2006). L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes. *Techniques d'enquête*, vol. 32, 161-168.

Rubin, D.B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9, 131-134.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schenker, N., Gentleman, J.F., Rose, D, Hing, E. et Shimizu, I.M. (2002). Combining estimates from complementary surveys: A case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports*, 117, 393-407.

Schenker, N. et Raghunathan, T.E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, 26, 1802-1811.

Schenker, N., Raghunathan, T.E. et Bondarenko, I. (2009). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine*, 29, 533-545.

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

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2014.

T. Adams, *U.S. Census Bureau*
S. Adeshiyan, *U.S. Energy Information Administration*
R. Andridge, *Ohio State University*
P. Ardilly, *INSEE, France*
T. Asparouhov, *Muthén & Muthén*
J. Aston, *University of Warwick*
R. Bautista, *NORC at the University of Chicago*
J.-F. Beaumont, *Statistique Canada*
E. Benhin, *Statistique Canada*
Y.G. Berger, *University of Southampton*
J. Bethlehem, *Statistics Netherlands/Leiden University*
C. Bocci, *Statistique Canada*
J. Breidt, *Colorado State University*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Census Bureau*
R. Chambers, *Centre for Statistical and Survey Methodology*
S. Chaudhury
G. Chauvet, *ENSAI, France*
R. Clark, *NIASRA, University of Wollongong*
G. Datta, *University of Georgia*
T. DeMaio, *U.S. Census Bureau*
J. Dever, *Research Triangle Institute*
J. Eltinge, *U.S. Bureau of Labour Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
D. Haziza, *Université de Montréal*
S. Heeringa, *ISR, U. of Michigan*
K.A. Henry, *Statistics of Income, Internal Revenue Service*
B. Hulliger, *U. of Applied Sciences Northwestern Switzerland*
F. Hutchinson, *Cancer Research Center*
D. Judkins, *Abt Associates*
K. Kadraoui, *Université Laval*
R.J. Karunamuni, *University of Alberta*
D. Kasprzyk, *NORC at the University of Chicago*
J.-K. Kim, *Iowa State University*
P.S. Kott, *RTI International*
P. Lahiri, *JPSM*, *University of Maryland*
P. Lavallée, *Statistique Canada*
L. Lee, *NORC at the University of Chicago*
P. Linde, *Statistics Denmark*
S. Lohr, *Westat*
P. Lynn, *University of Essex*
D.J. Malec, *National Center for Health Statistics*
D. Marker, *Westat*
J. Montaquila, *Westat*
B. Nandram, *Worcester Polytechnic Institute*
J. Opsomer, *Colorado State University*
D. Pfeffermann, *Hebrew University*
F. Picard, *Statistique Canda*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
E. Robison, *U.S. Bureau of Labor Statistics*
T. Savitsky, *Bureau of Labor Statistics*
J. Shao, *University of Wisconsin*
F.J. Scheuren, *National Opinion Research Center*
P.L.D.N. Silva, *Escola Nacional de Ciências Estatísticas*
C. Skinner, *LSE*
P. Smith, *Office for National Statistics*
D. Steel, *University of Wollongong*
D. Sverchkov, *U.S. Bureau of Labor Statistics*
N. Thomas, *Statisctics Reasearch and Consulting Center, Pfizer*
M. Thompson, *University of Waterloo*
M. Torabi, *Mplus*
D. Toth, *U.S. Bureau of Labor Statistics*
R. Valliant, *University of Maryland*
J. van den Brakel, *Statistics Netherlands*
F. Verret, *Statistique Canada*
B.T. West, *ISR, University of Michigan - Ann Arbor*
K.M. Wolter, *National Opinion Research Center*
C. Wu, *University of Waterloo*
D. Yang, *Bureau of Labor Statistics*
Y. You, *Statistique Canada*
Z. Yu, *U. of Wisconsin Madison*
W. Yung, *Statistique Canada*
A. Zaslavsky, *Harvard University*
S.Z. Zangeneh, *Vaccine and Infectious Disease Division*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2014 : Joana Bérubé de la Division des méthodes d'enquête auprès des entreprises; l'équipe de la Division de la diffusion, en particulier Éva Demers-Brett, Chantal Chalifoux, Jacqueline Luffman, Kathy Charbonneau, Lucie Gauthier, Daniel Piché, Jasvinder Jassal, Joseph Prince et Darquise Pellerin; Céline Ethier et Nick Budko de la Division de la recherche et de l'innovation en statistique, de même que nos partenaires de la Division des communications.

ANNONCES

Demande de candidatures pour le prix Waksberg 2016

La revue *Techniques d'enquête* a mis sur pied une série annuelle de communications sollicitées en l'honneur de Joseph Waksberg, en reconnaissance des contributions qu'il a faites à la méthodologie d'enquête. Chaque année, un éminent statisticien d'enquête est choisi pour rédiger un article où il examine l'évolution et l'état actuel d'un thème important du domaine de la méthodologie d'enquête. L'article reflète le mélange de théorie et de pratique caractéristique des travaux de Joe Waksberg.

Le lauréat du prix Waksberg recevra une prime en argent et présentera la communication sollicitée Waksberg 2016 au Symposium de Statistique Canada qui se tiendra à l'automne de 2016. L'article paraîtra dans un numéro de *Techniques d'enquête* (publication prévue pour décembre 2016).

L'auteur de l'article Waksberg 2016 sera choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*. Les candidatures ou les suggestions de thèmes doivent être envoyées avant le 28 février 2015 au président du comité, Louis-Paul Rivest (Louis-Paul.Rivest@mat.ulaval.ca).

Les gagnants et articles précédents du prix Waksberg sont

2001 Gad **Nathan**, « Méthodes de téléenquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir ». *Techniques d'enquête*, vol. 27, 1, 7-34.
2002 Wayne A. **Fuller**, « Estimation par régression appliquée à l'échantillonnage ». *Techniques d'enquête*, vol. 28, 1, 5-25.
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.
2004 Norman M. **Bradburn**, « Comprendre le processus de question et réponse ». *Techniques d'enquête*, vol. 30, 1, 5-16.
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.
2006 Alastair **Scott**, « Études cas-témoins basées sur la population ». *Techniques d'enquête*, vol. 32, 2, 137-147.
2007 Carl-Erik **Särndal**, « La méthode de calage dans la théorie et la pratique des enquêtes ». *Techniques d'enquête*, vol. 33, 2, 113-135.
2008 Mary E. **Thompson**, « Enquêtes internationales : motifs et méthodologies ». *Techniques d'enquête*, vol. 34, 2, 145-157.
2009 Graham **Kalton**, « Méthodes de suréchantillonnage des sous-populations rares dans les enquêtes sociales ». *Techniques d'enquête*, vol. 35, 2, 133-152.
2010 Ivan P. **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.
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.
2012 Lars **Lyberg**, « La qualité des enquêtes ». *Techniques d'enquête*, vol. 38, 2, 115-142.
2013 Ken **Brewer**, « Trois controverses dans l'histoire de l'échantillonnage ». *Techniques d'enquête*, vol. 39, 2, 275-289.
2014 Constance F. **Citro**, « Des modes multiples pour les enquêtes à des sources de données multiples pour les estimations ». *Techniques d'enquête*, vol. 40, 2, 151-181.
2015 Robert M. **Groves**, Sujet de l'article à l'étude.

Membres du comité de sélection de l'article Waksberg (2014-2015)

Louis-Paul Rivest, *Université Laval* (Président)
J.N.K. Rao, *Carleton University*
Kirk Wolter, *National Opinion Research Center*
Tommy Wright, *U.S. Bureau of the Census*

Présidents précédents :

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)
Elizabeth A. Martin (2010 - 2011)
Mary E. Thompson (2011 - 2012)
Steve Heeringa (2012 - 2013)
Cynthia Clark (2013 - 2014)

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 30, No. 2, 2014

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 30, No. 3, 2014

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 42, No. 3, September/septembre 2014

CONTENTS TABLE DES MATIÈRES

Volume 42, No. 4, December/décembre 2014