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

Nº 12-001-XPB au catalogue
ISSN 0714-0045

Techniques d'enquête

Décembre 2012

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 ».

Ce produit est aussi disponible en version imprimée standard au prix de 30 $CAN l'exemplaire et de 58 $CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	10 $CAN	20 $CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste Statistique Canada
Finances
Immeuble R.-H.-Coats, 6e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

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)

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.
∞

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.
∞

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans *The ISI Web of kn owledge* (*Web of science*), *The Survey Statistician*, *Statistical Theory and Methods A bstracts* et *SRM Database of Social Research Methodology*, *Erasmus University*. On peut en trouver les référence s dans *Current Index to Statisti cs*, et *Journal Contents in Qu alitative Methods*. La revue est égale ment 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 désaisonnali sation, les étud es démographiques, l'intégration de données statisti ques, les méthodes d'estimation et d'analyse de d onnées et le développement de sy stèmes généralisés. Une importance particulière est acco rdée à l 'élaboration et à l'évaluation de méthodes qui ont été ut ilisées pour la collecte de données ou appliquées à des données réell es. Tous les a rticles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opi nions 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 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).

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 $ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux env ois à l'ex térieur du Canada : États-Unis 12 $ CA (6 $ × 2 exemplaires); autres pays, 20 $ CA (10 $ × 2 exemplaires). Un prix réduit est offert aux m embres de l'Am erican Statistical Association, l'Association Internationale de S tatisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.

Techniques d'enquête

Une revue éditée par Statistique Canada
Volume 38, numéro 2, décembre 2012

Table des matières

Article sollicité Waksberg

Articles réguliers

Communications brèves


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, p. 113
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 2014.

Ce numéro de *Techniques d'enquête* commence par le douzième article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé d'Elizabeth A. Martin (présidente), Mary E. Thompson, Steve Heeringa et J.N.K. Rao, d'avoir choisi Lars Lyberg comme auteur de l'article du prix Waksberg de cette année.

Communication sollicitée pour le prix Waksberg 2012

Auteur : Lars Lyberg

Lars Lyberg, Ph. D., est l'ancien chef du Département de la recherche-développement de Statistique Suède et est actuellement professeur émérite au Département de statistique de l'Université de Stockholm. Il a fondé le *Journal of Official Statistics* (*JOS*) et y a été rédacteur en chef pendant 25 ans. Il est rédacteur en chef de *Survey Measurement and Process Quality* (Wiley 1997) et corédacteur de *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (Wiley 2010), *Telephone Survey Methodology* (Wiley 1988) et *Measurement Errors in Surveys* (Wiley 1991). Il est coauteur de *Introduction to Survey Quality* (Wiley 2003). Il a présidé le groupe de travail sur la qualité du Système statistique européen ainsi que le comité organisateur de la première Conférence européenne sur la qualité des statistiques officielles, Q2001. Il a déjà présidé l'AISE ainsi que la Section des méthodes d'enquête de l'American Statistical Association. Il est membre de cette dernière association et de la Royal Statistical Society.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 115-142
Statistique Canada, N° 12-001-X au catalogue

La qualité des enquêtes

Lars Lyberg [1]

Résumé

La qualité des enquêtes est un concept multidimensionnel issu de deux démarches de développement distinctes. La première démarche suit le paradigme de l'erreur d'enquête totale, qui repose sur quatre piliers dont émanent les principes qui guident la conception de l'enquête, sa mise en œuvre, son évaluation et l'analyse des données. Nous devons concevoir les enquêtes de façon que l'erreur quadratique moyenne d'une estimation soit minimisée compte tenu du budget et d'autres contraintes. Il est important de tenir compte de toutes les sources connues d'erreur, de surveiller les principales d'entre elles durant la mise en œuvre, d'évaluer périodiquement les principales sources d'erreur et les combinaisons de ces sources après l'achèvement de l'enquête, et d'étudier les effets des erreurs sur l'analyse des données. Dans ce contexte, on peut mesurer la qualité d'une enquête par l'erreur quadratique moyenne, la contrôler par des observations faites durant la mise en œuvre et l'améliorer par des études d'évaluation. Le paradigme possède des points forts et des points faibles. L'un des points forts tient au fait que la recherche peut être définie en fonction des sources d'erreur et l'un des points faibles, au fait que la plupart des évaluations de l'erreur d'enquête totale sont incomplètes, en ce sens qu'il est impossible d'inclure les effets de toutes les sources. La deuxième démarche est influencée par des idées empruntées aux sciences de la gestion de la qualité. Ces sciences ont pour objet de permettre aux entreprises d'exceller dans la fourniture de produits et de services en se concentrant sur leurs clients et sur la concurrence. Ces idées ont eu une très grande influence sur de nombreux organismes statistiques. Elles ont notamment amené les fournisseurs de données à reconnaître qu'un produit de qualité ne peut pas être obtenu si la qualité des processus sous-jacents n'est pas suffisante et que des processus de qualité suffisante ne peuvent pas être obtenus sans une bonne qualité organisationnelle. Ces divers niveaux peuvent être contrôlés et évalués au moyen d'ententes sur le niveau de service, de sondages auprès des clients, d'analyses des paradonnées en recourant au contrôle statistique des processus et d'évaluations organisationnelles en se servant de modèles d'excellence opérationnelle ou d'autres ensembles de critères. À tous les niveaux, on peut rehausser la qualité en lançant des projets d'amélioration choisis selon des fonctions de priorité. L'objectif ultime de ces projets d'amélioration est que les processus concernés s'approchent progressivement d'un état où ils sont exempts d'erreur. Naturellement, il pourrait s'agir d'un objectif impossible à atteindre, mais auquel il faut tenter de parvenir. Il n'est pas raisonnable d'espérer obtenir des mesures continues de l'erreur d'enquête totale en se servant de l'erreur quadratique moyenne. Au lieu de cela, on peut espérer qu'une amélioration continue de la qualité par l'application des idées des sciences de la gestion ainsi que des méthodes statistiques permettra de minimiser les biais et d'autres problèmes que posent les processus d'enquête, afin que la variance devienne une approximation de l'erreur quadratique moyenne. Si nous y arrivons, nous aurons fait coïncider approximativement les deux démarches de développement.

Mots clés : Gestion de la qualité ; erreur d'enquête totale ; cadre de la qualité ; erreur quadratique moyenne ; variabilité des processus ; contrôle statistique des processus ; utilisateurs des données d'enquête.

1. Introduction

Le présent article a été rédigé en reconnaissance des apports uniques et du leadership de Joe Waksberg dans le domaine des techniques d'enquête. J'ai pris connaissance des travaux de Joe pour la première fois en lisant son article sur les erreurs de réponse dans les enquêtes sur les dépenses, rédigé en collaboration avec John Neter (Neter et Waksberg 1964). Entre autres, cet article m'a fait découvrir le phénomène cognitif appelé télescopage. Plus tard, j'ai eu l'occasion de travailler avec Joe à la préparation de la première conférence et monographie sur les méthodes d'enquête téléphonique en tant que membre du comité de rédaction (Groves, Biemer, Lyberg, Massey, Nicholls et Waksberg 1988). Nous avons également collaboré à la préparation de nombreuses conférences Morris Hansen dont les exposés ont été publiés dans le *Journal of Official Statistics* (JOS) durant mon mandat de rédacteur en chef. Joe lui-même a donné la sixième conférence, qui a été publiée dans le JOS (Waksberg 1998). Joe était un fantastique chef de file et c'est pour moi un grand honneur d'avoir été invité à rédiger cet article sur la qualité des enquêtes, sujet qui le préoccupait beaucoup.

Bon nombre de mes amis m'ont fait part de leurs opinions ou m'ont envoyé de la documentation en prévision du présent article. Je remercie tout spécialement Paul Biemer, Dan Kasprzyk, Fritz Scheuren, Dennis Trewin et Maria Bohata de leur aide.

La qualité des enquêtes est un concept vague, quoiqu'intuitif, ayant de nombreuses significations. Dans le présent article, je discute de certaines observations qui ont trait à l'élaboration et au traitement du concept au cours des soixante-dix dernières années et, dans le cas de certains développements, il m'est même possible de remonter à des origines encore plus lointaines. Toutefois, ma discussion porte en majeure partie sur les questions qui se posent aujourd'hui dans les organismes statistiques gouvernementaux. C'est au domaine de la statistique officielle

1. Lars Lyberg, Département de statistique, Université de Stockholm, 10691 Stockholm, Suède. Courriel : Lars.Lyberg@stat.su.se.

qu'appartiennent la plupart des exemples de qualité des enquêtes que j'expose.

La présentation de l'article est la suivante. La section 2 traite du paradigme de l'erreur d'enquête totale, y compris les typologies de l'erreur, le traitement des erreurs et la conception des enquêtes en tenant compte de toutes les sources d'erreur. La section 3 porte sur les approches de gestion de la qualité qui ont eu un effet important sur les organismes d'enquête depuis le début des années 1990. Cet effet s'est manifesté par des méthodes et des approches telles que la prise en compte de l'utilisateur ou du client, la discussion des coûts et des risques dans le cadre de la recherche sur les enquêtes, et la nécessité pour les organismes de continuer à s'améliorer. La section 4 fournit des exemples de projets d'amélioration de la qualité entrepris par les organismes d'enquête. La section 5 traite des difficultés que pose la mesure, directe ou indirecte, de la qualité au moyen d'indicateurs. Est également abordée la façon dont ces mesures doivent être communiquées aux utilisateurs ou aux clients. Enfin, la section 6 offre certaines réflexions quant à la manière dont les pratiques d'enquête *doivent* évoluer afin de mieux répondre aux besoins des utilisateurs. La dernière section est réservée à la bibliographie.

2. Le paradigme de l'erreur d'enquête totale

2.1 Bref historique de l'échantillonnage

Un certain nombre d'articles décrivent l'élaboration des premières méthodes d'échantillonnage. On constate dans ces premiers travaux une reconnaissance implicite ou explicite des problèmes de qualité, même s'ils sont masqués sous des termes tels que « erreurs » et « utilité de l'enquête » (Deming 1944). Les aperçus historiques que l'on trouve, par exemple, dans Kish (1995), Fienberg et Tanur (1996), et O'Muircheartaigh (1997) insistent tous sur le fait que, jusqu'à 1950, on a assisté au plein essor de la théorie de l'échantillonnage. Dans les années 1920, l'Institut international de statistique a accepté de promouvoir les idées sur l'échantillonnage représentatif proposées par Kiear (1897) et Bowley (1913). En 1934, Neyman a publié son article historique sur la méthode représentative. Plus tard, le principe de randomisation de Fisher (1935) a été appliqué à l'échantillonnage en agriculture et Neyman (1938) a élaboré l'échantillonnage par grappes, l'estimation par le ratio et l'échantillonnage à deux phases, et introduit le concept d'intervalle de confiance. Neyman a montré que l'erreur d'échantillonnage pouvait effectivement être mesurée en calculant la variance de l'estimateur. Bill Cochran, Frank Yates, Ed Deming, Morris Hansen et bien d'autres ont perfectionné les concepts de la théorie de l'échantillonnage. Hansen a dirigé un groupe de recherche au U.S. Census Bureau, où avait lieu à l'époque une grande partie des activités de recherche appliquée et d'élaboration de nouvelles théories. L'un des résultats remarquables des travaux du Census Bureau a été la production d'un manuel en deux volumes sur la théorie et les méthodes d'échantillonnage (Hansen, Hurwitz et Madow 1953). En fait, les progrès en théorie de l'échantillonnage étaient si importants à ce moment-là que Stephan (1948) a jugé bon de rédiger un article sur l'histoire des méthodes modernes d'échantillonnage.

Très tôt, on a reconnu qu'il pouvait exister d'autres erreurs d'enquête que celles attribuées à l'échantillonnage. Il existe des écrits sur les effets du libellé des questions, dont celui de Muscio (1917). La recherche sur la conception des questionnaires était assez intensive durant les années 1940. Mahalanobis (1946) s'est attaqué aux problèmes résultants des erreurs introduites par les enquêteurs sur le terrain chargés de recueillir les données agricoles en Inde, ce qui a donné une méthode d'estimation de ces erreurs. Cette méthode, appelée « interpénétration », peut être utilisée pour estimer ce que l'on appelle les variances corrélées introduites par les intervieweurs, les vérificateurs, les codeurs et les personnes qui supervisent ces groupes. Les sources d'erreur les plus importantes étaient certainement déjà connues autour de 1950. Deming a dressé une liste des sources d'erreur (1944) qui constitue la première typologie publiée des erreurs d'enquête, et Hansen et Hurwitz (1946) ont discuté du sous-échantillonnage des non-répondants pour essayer de fournir des estimations sans biais dans une situation présentant une non-réponse initiale. Cependant, sur le plan de la méthodologie, l'accent avait été mis jusque-là sur l'élaboration de la théorie de l'échantillonnage, ce qui est assez compréhensible. Il était en effet très important de pouvoir montrer qu'il était possible de réaliser des enquêtes en s'appuyant sur l'échantillonnage, et ce, dans diverses conditions. En 1950, il avait été démontré de manière assez satisfaisante que cela était effectivement faisable. Donc, il était temps de passer à d'autres questions et aux peaufinements.

Au début, l'emploi du terme qualité était limité avant tout au contrôle de la qualité, parfois au contrôle de la qualité des opérations d'enquête. Souvent, le contrôle de la qualité se résumait à la vérification et (ou) à l'estimation de la grandeur de l'erreur pour diverses opérations. On savait que les statistiques étaient affectées par d'autres erreurs que celles émanant de l'échantillonnage, mais la façon, liée à la qualité des processus, de réduire systématiquement ces erreurs et biais restait encore à établir (Deming 1944 ; Hansen et Steinberg 1956).

Il y a 60 ans d'ici, l'utilisateur était un joueur plutôt obscur, même si les éminents concepteurs des techniques d'enquête ne l'ignoraient pas du tout. Ainsi, Deming (1950)

soutenait que, jusqu'à ce que le but soit énoncé, il n'existait aucune bonne ou mauvaise façon d'entreprendre une enquête. Certains autres statisticiens ont fait des déclarations comparables. En fait, c'est vraiment l'utilisateur qui était caché derrière des termes tels que « problème lié au domaine spécialisé », « but de l'étude » ou « fonctions clés d'un système statistique ».

Même aujourd'hui, les concepts d'enquête et de qualité sont vagues. Comme l'on souligné Morganstein et Marker (1997), les définitions variées de la qualité nuisent aux travaux d'amélioration, de sorte que nous devons au moins essayer de faire la distinction entre les différentes définitions afin de déterminer à quoi elles servent. L'une des définitions citées le plus souvent est attribuée à Joseph Juran, à savoir que la qualité est une fonction directe de l'« adaptation à l'usage prévu ». En fait, déjà en 1944, Deming avait utilisé la phrase « fitness for purpose » (adaptation au but poursuivi), non pas pour définir la qualité, mais plutôt pour expliquer ce qui faisait la réussite d'un produit d'enquête.

Longtemps, la notion de « bonne » qualité était implicitement équivalente à une faible erreur quadratique moyenne (EQM), ce qui signifie que les données doivent être exactes et que l'exactitude d'une estimation peut être mesurée par l'EQM, qui est la somme de la variance et du carré du biais. Nous avons constaté que les statistiques fondées sur des sondages doivent aussi être utiles, ce qui a été désigné plus tard par le terme « pertinentes ». Nombre des dimensions actuelles de la qualité ne représentaient pas vraiment un sujet de préoccupation à l'époque. En outre, les utilisateurs étaient habitués à ce que la réalisation des enquêtes prenne du temps ; l'actualité des données était certes à l'ordre du jour, mais pas aussi explicitement qu'aujourd'hui. Le traitement des données d'un recensement prenait des années. Les utilisateurs étaient accoutumés à une technologie qui ne permettait d'offrir que des formes assez simples d'accessibilité. Donc, il était naturel pour les utilisateurs et les producteurs de faire en sorte avant tout que le problème statistique concorde raisonnablement avec le problème du secteur spécialisé et que l'EQM soit maintenue à un niveau acceptable. Cette EQM était, et est encore, équivalente dans de nombreux cas à la variance seulement, sans ajout d'un terme de carré du biais.

Avant de poursuivre, définissons ce qu'est une « enquête ». Une *enquête* est une étude statistique conçue pour mesurer les caractéristiques de la population afin de pouvoir estimer les paramètres de cette dernière. La proportion de chômeurs à un moment donné dans une population de personnes, ou le revenu total d'une entreprise ou d'un secteur d'activité durant une période donnée sont deux exemples de paramètres. Une enquête peut être définie comme une liste de conditions préalables (Dalenius 1985a). Selon Dalenius, une étude peut être catégorisée comme une enquête si les conditions préalables qui suivent sont satisfaites :

1. l'étude concerne un ensemble d'objets constituant une population ;
2. la population étudiée possède une ou plusieurs propriétés mesurables ;
3. le but de l'étude est de décrire la population au moyen d'un ou de plusieurs paramètres définis en fonction des propriétés mesurables, ce qui nécessite l'observation (d'un échantillon) de la population ;
4. pour arriver à observer la population, une base de sondage est nécessaire ;
5. un échantillon d'objets est sélectionné à partir de la base de sondage conformément à un plan d'échantillonnage qui spécifie un mécanisme probabiliste et une taille d'échantillon n (où n pourrait être égal à N, la taille de la population) ;
6. des observations sont faites sur l'échantillon conformément à un procédé de mesure (c'est-à-dire une méthode de mesure et une prescription concernant son utilisation) ;
7. un processus d'estimation fondé sur les mesures est appliqué pour calculer des estimations des paramètres lorsque l'on fait une inférence au sujet de la population étudiée d'après l'échantillon.

Cette définition énumère implicitement les sources particulières d'erreur présentes dans les travaux d'enquête. Pour chaque source, il existe un certain nombre de méthodes qui en minimisent les effets, mais mesurent également leur grandeur (Biemer et Lyberg 2003 ; Groves, Fowler, Couper, Lepkowski, Singer et Tourangeau 2009).

Les écarts par rapport à la définition reflètent des défauts de qualité. En outre, les écarts de ce genre sont fréquents. Dans certains plans de sondage, les probabilités de sélection sont inconnues ou l'estimateur de la variance choisi n'est pas nécessairement celui qui convient le mieux, étant donné le plan utilisé. Le fait que ces défauts posent problème ou non dépend du but de l'enquête.

2.2 Les composantes du paradigme de l'erreur d'enquête totale

Le paradigme de l'erreur d'enquête totale est un cadre théorique utilisé pour optimiser les enquêtes en minimisant la grandeur cumulée des erreurs provenant de toutes les sources, étant donné des contraintes budgétaires. En pratique, cela signifie que nous voulons minimiser l'erreur quadratique moyenne de certaines estimations fondées sur les données d'enquête, à savoir celles que les principales parties prenantes jugent les plus importantes. L'erreur quadratique moyenne, qui est la mesure utilisée le plus fréquemment pour évaluer le travail d'enquête, est égale à la

somme des variances et des termes de biais au carré provenant de chaque source connue d'erreur. Groves et Lyberg (2010) résument la situation du paradigme dans la pratique passée et contemporaine des enquêtes.

L'idée selon laquelle les enquêtes doivent être conçues en tenant compte de toutes les sources d'erreur émane des premiers ténors du domaine. Morris Hansen, Bill Hurwitz, Joe Waksberg, Leon Pritzker, Ed Deming et d'autres au U.S. Census Bureau, Leslie Kish à l'Université du Michigan, P.C. Mahalanobis à l'Institut statistique de l'Inde, et Tore Dalenius, à l'Université de Stockholm, étaient parmi les chefs de file de la recherche sur les enquêtes, mettant l'accent sur les erreurs et l'optimisation du plan de sondage. Ils étaient préoccupés par les limites inhérentes à la théorie de l'échantillonnage, car les erreurs non dues à l'échantillonnage risquaient de faire s'effondrer la théorie. Très pragmatiques, ils réfléchissaient beaucoup à la façon d'équilibrer les erreurs et aux coûts qu'entraîne leur traitement. Voyant des similarités entre une chaîne de montage d'usine (Deming et Geoffrey 1941) et la mise en œuvre de certains processus d'enquête, certains d'entre eux ont introduit des méthodes de contrôle tirées d'applications industrielles.

Dalenius (1967) s'est rendu compte qu'il n'existait pas encore de « formule de conception » pouvant fournir une solution optimale au problème. L'approche adoptée par Dalenius ainsi que par Hansen, Hurwitz et Pritzker (1967) consistait à minimiser tous les biais et à opter pour un scénario de variance minimale, pour que la variance devienne une approximation de l'EQM. Cela était censé se faire au moyen de schémas de vérification intense pour les productions en cours et d'études d'évaluation d'assez grande portée pour les futures productions. En 1969, inspiré par Hansen, Dalenius a présenté une communication portant sur la conception globale des enquêtes (*total survey design*), où le terme « totale » traduisait l'idée de prendre en compte toutes les sources d'erreur. Hansen, Hurwitz, Marks et Mauldin (1951), Hansen, Hurwitz et Bershad (1961), et Hansen, Hurwitz et Pritzker (1964) ont élaboré le modèle d'enquête du U.S. Census Bureau qui tenait compte de l'effet des intervieweurs, des codeurs, des vérificateurs et des chefs d'équipe, et permettait d'estimer leur contribution à l'erreur d'enquête totale. Ces schémas d'estimation, étoffés par Bailar et Dalenius (1969), consistaient en des variations de la répétition et de l'interpénétration. L'hypothèse était que l'estimation du biais était traitée par comparaison des estimations obtenues d'après les opérations ordinaires à celles obtenues au moyen des procédures privilégiées (qui ne pouvaient pas être utilisées à grande échelle pour des raisons financières, administratives ou pratiques). Aujourd'hui, ce genre d'approche est considéré comme étant la « norme de référence » (*gold standard*).

Il a été déclaré que pour bien concevoir une enquête il fallait contrôler de manière raisonnablement efficace l'erreur totale en spécifiant avec soin les procédures d'enquête, y compris des contrôles adéquats. Hansen, Deming et d'autres s'inquiétaient du coût des contrôles, mais, alors que le contrôle statistique des processus et l'échantillonnage d'acceptation avaient été mis en œuvre par un certain nombre d'organismes d'enquête, on parlait fort peu de l'amélioration continue des processus. Une part importante du travail relatif à la qualité concernait l'estimation des taux d'erreur, le contrôle des niveaux d'erreur des opérateurs individuels et la réalisation d'études d'évaluation à grande échelle qui prenait habituellement beaucoup de temps. Les utilisateurs ne participaient pas directement au processus de conception, mais dans le système statistique fédéral des États-Unis, ils exerçaient au moins une certaine influence sur la détermination des données qui devaient être recueillies et présentées. Dalenius (1968) fournit plus de 200 références concernant les utilisateurs et les conférences à l'intention des utilisateurs associées aux produits du système statistique fédéral des États-Unis.

Bien que Hansen, Dalenius et d'autres aient été les premiers à préconiser la conception globale des enquêtes, il était rare que les utilisateurs participent directement à la détermination finale des exigences concernant l'enquête. Assez souvent, un agent, un administrateur ou un statisticien jouait le rôle de spécialiste du domaine. Il y a plusieurs décennies, c'est comme cela que nous pensions aux utilisateurs. Leurs opinions comptaient, mais ils ne participaient pas vraiment aux prises de décisions. Cependant, au fond de nous-mêmes, nous savions qu'il ne s'agissait peut-être pas d'un modèle parfait et, à la fin des années 1970, Statistics Sweden a publié une brochure interne intitulée « Que faire si un client se présente à notre porte ».

L'approche fondamentale de conception proposée par Hansen, Dalenius et d'autres comprenait un certain nombre d'étapes, dont :

- la spécification d'un objectif idéal d'enquête ;
- l'analyse de la situation de l'enquête quant aux ressources en matière de budget, de méthodologie et d'information ;
- l'élaboration d'un petit nombre d'options de plan d'enquête ;
- l'évaluation des diverses options en se basant sur les déterminations préliminaires connexes de l'EQM et des coûts ;
- le choix de l'une des options ou d'une modification de l'une d'elles, ou la décision de ne pas procéder du tout à l'enquête ;
- l'élaboration du plan d'enquête administratif, y compris l'essai de faisabilité, un système de signalisation

concernant les processus (appelé aujourd'hui paradonnées), un document de conception et un plan B.

Les vues de Kish (1965) sur la conception des enquêtes différaient légèrement. Il favorisait les applications néo-bayésiennes en échantillonnage et la psychométrie prônée par certains collègues à l'Université du Michigan (Ericson 1969 ; Edwards, Lindman et Savage 1963). Par exemple, Kish aimait l'idée que des estimations au jugé des biais de mesure pourraient être combinées aux variances d'échantillonnage pour construire des estimations plus réalistes de l'erreur d'enquête totale. Quant au problème d'optimisation, il pensait qu'une approche polyvalente était économiquement favorable pour les enquêtes, mais qu'il pourrait être difficile de décider sur quoi fonder le plan de sondage. Si l'on arrive à désigner une statistique principale, celle-ci peut à elle seule déterminer le plan de sondage et s'il existe un petit nombre de ces statistiques, il est possible d'opter pour un plan de sondage de compromis ; par contre, si les statistiques sont trop disparates, il pourrait n'exister aucun plan de sondage raisonnable. Kish insiste aussi sur la nécessité d'obtenir des renseignements sur le plan de sondage au moyen d'enquêtes pilotes et de prétests afin de prendre plus facilement les décisions concernant ce plan. Il a constaté que le plan de sondage et les mesures pouvaient varier considérablement selon l'environnement, tandis que l'échantillonnage changeait moins. Il pourrait s'agir de l'une des raisons pour lesquelles l'échantillonnage peut être classé facilement parmi les théories et méthodes statistiques classiques, alors qu'il est plus difficile d'insérer le processus d'enquête dans une discipline particulière (Frankel et King 1996 dans leur entrevue avec Kish).

Comme les autres ténors de la conception des enquêtes, Kish insistait sur l'importance d'un faible biais, mais appréciait le fait que la réduction du terme de biais pourrait accroître l'erreur totale. Il avait à cœur d'arriver à un équilibre raisonnable entre les diverses sources d'erreur et la façon dont les structures d'erreur variaient sous diverses options de plan de sondage. Comme Hansen et ses collègues, Kish pensait que les renseignements pertinents devaient être enregistrés simultanément durant la mise en œuvre (de nouveau nous voyons le parallèle avec les paradonnées). Hansen et ses collègues s'inquiétaient vraiment de l'application de contrôles excessifs, mais inadéquats. Ils se sont rendu compte que certains contrôles devraient peut-être être relâchés en raison des améliorations limitées qui en découlaient et que le degré d'amélioration des estimations devait être vérifié avant de procéder à tout relâchement des contrôles. Ils ont également suggéré que l'on devrait peut-être compromettre la pertinence pour obtenir des mesures contrôlables ou s'abstenir de procéder à l'enquête. Tant Hansen et ses collègues que Kish défendaient vivement l'idée de mettre fin à la pratique voulant que l'erreur d'échantillonnage soit la seule erreur d'enquête mesurée.

Un examen de la situation actuelle porte à conclure que l'on ne dispose toujours pas d'une formule de conception des enquêtes. Il n'existe pour ainsi dire aucun manuel de planification, et la littérature sur la conception des enquêtes est par conséquent peu abondante, de même que celle sur les coûts (Groves 1989 est une exception). Aucune formule de conception n'est en vue. Depuis l'élaboration du modèle d'enquête du U.S. Census Bureau, plusieurs variantes ont fait leur apparition, certaines d'entre elles assez compliquées (Groves et Lyberg 2010). Une caractéristique commune est le fait qu'elles ont tendance à être incomplètes, c'est-à-dire qu'elles ne tiennent pas compte de toutes les sources d'erreur. Sur le plan statistique, l'attention se concentre surtout sur les composantes de la variance, en particulier la variance de l'erreur de mesure. Un certain nombre d'autres faiblesses sont associées au concept de l'erreur d'enquête totale. En premier lieu, la perspective des utilisateurs fait défaut et une vaste majorité d'entre eux ne sont pas à même de mettre en doute l'exactitude des données ni d'en discuter. Les structures et les interactions complexes des erreurs n'incitent pas les contrôles extérieurs et les contacts avec les utilisateurs ont souvent tendance à porter sur des questions moins techniques, telles que l'actualité, la comparabilité et le coût des données. Les utilisateurs ne sont pas vraiment au courant des niveaux réels d'exactitude et nous en savons fort peu quant à la façon dont ils perçoivent l'information sur les erreurs et comment y donner suite.

Comme l'a fait remarquer Biemer (2001), il existe un manque de mesures systématiques des composantes de l'EQM dans les organismes statistiques. Plusieurs bonnes raisons sont à l'origine de cette situation. À la complexité, qui a déjà été mentionnée, nous pouvons ajouter des facteurs tels que les coûts, le fait qu'il est presque impossible de publier ce genre d'information au moment où les données sont diffusées et le fait qu'il n'existe aucune mesure de l'erreur totale qui tient compte de toutes les sources d'erreur, faute d'une méthodologie appropriée ou parce qu'il est impossible d'exprimer certaines erreurs. Groves et Lyberg (2010) énumèrent certaines autres faiblesses du paradigme de l'erreur d'enquête totale. Par exemple, nous devons en savoir davantage sur l'interaction entre les variances et les biais. Il se peut qu'un accroissement de la simple variance de réponse aille de pair avec une réduction du biais de réponse, disons, quand nous comparons le mode d'interview à des options d'autoadministration du questionnaire. Récemment, West et Olson (2010) ont montré que la variance due à l'intervieweur peut résulter non seulement de l'effet individuel des intervieweurs sur les réponses recueillies dans le cadre de leurs tâches, mais aussi du fait que les intervieweurs réussissent individuellement à obtenir

la coopération de divers groupes de membres de l'échantillon.

Malgré toutes ses limites, le cadre de l'erreur d'enquête totale présente des points forts assez convaincants. Il fournit une décomposition taxonomique des erreurs, sépare la variance du biais et l'observation de la non-observation, et définit les diverses étapes du processus d'enquête. Il sert de fondement conceptuel au domaine de la méthodologie d'enquête, les sous-domaines étant définis par leur structure d'erreur connexe. Enfin, il permet de cerner les lacunes dans la littérature sur la recherche, puisque toute typologie montrera que certaines étapes et processus sont plus « populaires » que d'autres. Il suffit pour s'en convaincre de comparer les portées respectives de la littérature sur la collecte des données et de celle sur le traitement des données.

Il semble toutefois que le cadre de l'erreur d'enquête totale nécessite une extension dans des directions dont certaines avaient déjà été signalées il y a un demi-siècle. Nous avons besoin de directives afin de trouver un compromis entre la mesure de la taille des erreurs et l'obtention de processus davantage exempts d'erreur. La question que se pose Spencer (1985) est celle de savoir combien de ressources nous devrions consacrer à la mesure par opposition à l'amélioration de la qualité. Nous avons également besoin de certaines directives quant à la façon d'intégrer des notions supplémentaires dans le cadre, afin qu'il devienne un cadre de la qualité d'enquête totale plutôt qu'un cadre de l'erreur d'enquête totale (Biemer 2010). Par exemple, si l'« adaptation à l'usage prévu » est le fondement conceptuel dominant, comment pouvons-nous lancer des travaux de recherche englobant la variation de l'erreur associée à différents usages ? Cet aspect est discuté à la section suivante.

3. Principes de gestion de la qualité dans les organismes d'enquête

Au cours des années 1980 et au début des années 1990, certains organismes statistiques faisaient face à de fortes pressions financières et, dans certains cas, étaient simultanément critiqués s'ils n'accordaient pas suffisamment d'attention aux besoins des utilisateurs. Les gouvernements de la Suède, de l'Australie, de la Nouvelle-Zélande et du Canada, de même que l'administration Clinton aux États-Unis souhaitaient vivement accroître l'efficacité de leurs systèmes statistiques respectifs, ainsi que l'influence exercée par les utilisateurs sur ces systèmes. Il était naturel pour ces organismes de s'inspirer des théories et méthodes de la gestion (Drucker 1985), tout spécialement ce que l'on appelle la gestion de la qualité (Juran et Gryna 1988). Grâce à cette nouvelle littérature, il était possible d'étudier le rôle du client, les problèmes de leadership, la notion d'amélioration continue de la qualité et les divers outils susceptibles d'aider l'organisme statistique à s'améliorer. Les travaux de Deming (1986) ont influencé particulièrement les praticiens des enquêtes, car il insistait sur le rôle des statistiques dans l'amélioration de la qualité. Il faisait valoir vigoureusement l'idée que les statisticiens doivent diriger les travaux d'amélioration, puisqu'ils ont reçu une formation leur permettant de faire la distinction entre diverses formes de variation des processus. Selon lui, trop peu de chefs de file de la statistique conseillaient la haute direction des entreprises et il voulait que des statisticiens plus proactifs deviennent ce genre de chef de file. Il avait particulièrement à cœur de développer les idées de Shewhart au sujet des cartes de contrôle comme moyen de distinguer les divers types de variation, à savoir les variations ordinaires et les variations ayant une cause spéciale. Le cycle d'amélioration de Shewhart consistant à planifier-faire-contrôler-agir (*Plan-Do-Check-Act*) faisait également partie des réflexions de Deming sur la qualité (Shewhart 1939).

Naturellement, l'existence des principes de gestion remonte à des temps anciens. Juran (1995) donne une foule d'exemples de ceux qui étaient en place, par exemple, dans l'empire romain. Le savoir-faire des artisans et un système de guilde en étaient les éléments fondamentaux. Des méthodes existaient pour choisir les matières premières et les fournisseurs. Les procédés étaient inspectés et améliorés. Les travailleurs étaient formés et motivés, et les clients obtenaient des garanties. Toutes ces caractéristiques se retrouvent encore dans les systèmes de gestion d'aujourd'hui. Les développements plus contemporains comprennent les cadres de la qualité ou les modèles d'excellence opérationnelle, tels que la gestion de la qualité totale (GTQ), les normes de l'Organisation internationale de normalisation (ISO), les critères du prix de qualité Malcolm Baldrige, le modèle d'excellence de la European Foundation for Quality Management (EFQM), les Six Sigma, les Lean Six Sigma et le tableau de bord prospectif (*Balanced Scorecard*). Ces modèles ne sont pas entièrement différents les uns des autres. Ils ont souvent en commun un ensemble de valeurs et de critères d'excellence. Ils représentent plutôt une évolution naturelle que l'on peut constater dans toutes sortes de travaux.

Donc, on a assisté à l'adoption progressive des modèles de gestion de la qualité et des stratégies de qualité dans les organismes statistiques et à une fusion avec les concepts et les idées déjà appliqués par ces organismes. Mon calendrier personnel de cette évolution est le suivant (les lecteurs sont invités à produire des ensembles différents d'événements et de dates) :

1875 Taylor introduit ce qu'il appelle la gestion scientifique.

1900 à 1930	Les idées de Taylor sont appliquées, par exemple, aux chaînes de montage chez Ford et chez Mercedes Benz.
Années 1920	Fisher commence à élaborer des théories et des méthodes concernant les plans expérimentaux.
1924	Shewhart développe la carte de contrôle.
1940	Le U.S. War Department produit un guide pour l'analyse des données sur les processus.
1944	Deming présente la première classification des erreurs d'enquête.
1944	Dodge et Romig présentent la théorie et des tableaux pour l'échantillonnage d'acceptation.
1946	Deming part au Japon.
1950	Ishikawa propose le diagramme en arêtes de poisson comme outil pour déterminer les facteurs qui ont un profond effet sur le résultat du processus.
1954	Juran part au Japon.
1960	De nombreuses entreprises lancent un programme « zéro défaut ».
1960	Le U.S. Census Bureau élabore des programmes de contrôle de la qualité.
1961	Le U.S. Census Bureau lance son modèle d'enquête.
1965-1966	Kish et Slobodan Zarkovich commencent à parler de la qualité des données plutôt que des erreurs d'enquête.
Années 1970	De nombreux organismes statistiques fournissent des lignes directrices concernant la qualité.
1975	Lancement du cadre de la gestion de la qualité totale (GQT).
1976	Adoption par un organisme statistique du premier cadre de la qualité contenant plus de dimensions que la pertinence et l'exactitude.
1987 à 1989	Lancement de la norme ISO 9000, du prix Malcolm Baldrige, de la stratégie Six Sigma et des modèles de l'EFQM.
Années 1990	De nombreux organismes statistiques commencent à travailler avec des modèles d'amélioration de la qualité et d'excellence.
1997	Publication de la monographie sur la qualité des mesures et des processus d'enquête (*Survey Measurement and Process Quality*).
1998	Mick Couper introduit le concept des « paradonnées » en tant que sous-ensemble des données sur les processus.
2001	Le Leadership Group (LEG)-Qualité d'Eurostat organise la première conférence sur la gestion de la qualité en statistique officielle.
2007	Les notions d'architecture opérationnelle font leur apparition dans l'univers des enquêtes.

À partir du milieu des années 1990, les principes de gestion de la qualité ont eu un immense effet sur de nombreux organismes statistiques. Il ne s'agit pas nécessairement d'un accroissement de la qualité à tous les niveaux (personne n'a vérifié ce fait). Mais les principes se sont traduits dans la plupart des organismes par une prise de conscience de l'importance du maintien de bons contacts avec les utilisateurs et avec les clients, et une aspiration à devenir « le meilleur » ou « de niveau international ». La qualité est à l'ordre du jour.

3.1 Le concept de qualité

Au cours des dernières décennies, il est devenu évident que l'exactitude et la pertinence sont des éléments nécessaires, mais non suffisants pour évaluer la qualité des enquêtes. D'autres dimensions sont également importantes pour les utilisateurs. L'élaboration de cadres de la qualité des enquêtes s'est déroulée principalement au sein des organismes de statistique officielle et a été déclenchée par le progrès technologique rapide et d'autres développements sociétaux. Les technologies de pointe ont créé des possibilités et suscité des demandes de la part des utilisateurs au sujet de dimensions éventuelles de la qualité, telles que l'accessibilité, l'actualité et la cohérence, qui n'étaient tout simplement pas mises en relief auparavant. Les décisions prises par la société sont devenues plus complexes et de portée plus mondiale, ce qui s'est traduit par des demandes de statistiques harmonisées et comparables. Donc, des cadres de qualité permettant de faire face à toutes ces demandes étaient nécessaires. Plusieurs cadres de qualité ont été élaborés et chacun comprend un certain nombre de dimensions de la qualité. L'exactitude et la pertinence ne sont que deux de ces dimensions.

Par exemple, le cadre élaboré par l'OCDE (2011) comprend huit dimensions, à savoir la pertinence, l'exactitude, l'actualité, la crédibilité, l'accessibilité, l'intelligibilité, la cohérence et la rentabilité (tableau 1). Des cadres similaires ont été établis par Statistique Canada (Statistique Canada 2002 ; Brackstone 1999), et par Statistics Sweden (Felme, Lyberg et Olsson 1976 ; Rosén et Elvers 1999). Le système statistique fédéral des États-Unis met depuis longtemps l'accent sur l'élément d'exactitude (U.S. Federal Committee on Statistical Methodology 2001), mais il apprécie certainement d'autres dimensions. Peut-être voit-il celles-ci comme étant de nature moins statistique, mais nécessitant néanmoins une part du budget total d'enquête. Le Fonds monétaire international (FMI) a élaboré un cadre qui diffère de

ceux de l'OCDE, de l'Australian Bureau of Statistics, de Statistics Sweden et de Statistique Canada. Le cadre du FMI comprend un ensemble de conditions préalables et cinq dimensions de la qualité, à savoir l'intégrité, la rigueur méthodologique, l'exactitude et la fiabilité, l'utilité, et l'accessibilité (voir Weisman, Balyozov et Venter 2010).

Si l'exactitude est insuffisante, les autres dimensions sont sans pertinence, mais l'inverse est vrai également. Des données très exactes peuvent être inutiles si elles sont diffusées trop tard pour avoir une incidence sur les décisions importantes des utilisateurs ou si elles sont présentées de façon telle que l'utilisateur a de la difficulté à y avoir accès ou à les interpréter. En outre, les dimensions de la qualité sont souvent conflictuelles. Par conséquent, fournir un produit de qualité est un savant numéro d'équilibre dans lequel les utilisateurs informés jouent un rôle important. Des conflits existent habituellement entre l'actualité et l'exactitude, puisqu'il faut du temps pour obtenir des données exactes, grâce, par exemple, à un suivi à grande échelle des cas de non-réponse. Un autre conflit est celui qui survient entre la comparabilité et l'exactitude, puisque l'application de nouvelles méthodes plus exactes pourrait perturber les comparaisons au fil du temps (Holt et Jones 1998).

Tableau 1
Cadre de qualité de l'OCDE

Dimension	Description
Pertinence	Les statistiques sont pertinentes si les besoins de l'utilisateur sont satisfaits.
Exactitude	Degré de rapprochement entre la valeur finalement retenue et la valeur réelle, mais inconnue, dans la population.
Crédibilité	Le degré de confiance qu'ont les utilisateurs dans les produits de données en fonction de la perception qu'ils ont du fournisseur des données.
Actualité	Temps écoulé entre le moment où les données sont disponibles et le moment où a eu lieu l'événement ou le phénomène qu'elles décrivent.
Accessibilité	Facilité avec laquelle les données peuvent être localisées et consultées à l'intérieur des fonds de données.
Intelligibilité	Facilité avec laquelle l'utilisateur peut comprendre, utiliser et analyser correctement les données.
Cohérence	Reflète la mesure dans laquelle les produits de données sont reliés logiquement et mutuellement concordants.
Rentabilité	Une mesure des coûts et du fardeau imposé au fournisseur par rapport à la production.

Donc, de nombreux organismes ont adopté un concept de qualité multidimensionnel comprenant non seulement l'exactitude, mais aussi d'autres dimensions. Nous pourrions parler d'un vecteur de qualité dont les composantes varient légèrement d'un organisme à l'autre, tant en nombre qu'en contenu. Plusieurs problèmes sont associés à l'approche du vecteur de qualité.

En premier lieu, son élaboration n'a pas été précédée de communications avec les utilisateurs. Les producteurs de statistiques ont cru que les utilisateurs étaient intéressés par un ensemble particulier de dimensions, même s'il est évident que la vaste majorité d'entre eux pensent que les structures d'erreur sont trop difficiles à saisir et supposent que le producteur a la responsabilité de fournir les données les plus exactes possibles. Lorsque l'utilisateur ou le client a des exigences d'exactitude particulières, un dialogue plus approfondi peut s'établir entre eux. Selon les rares études qui ont examiné la façon dont les utilisateurs perçoivent l'information sur la qualité, les utilisateurs s'intéressent surtout aux dimensions qui sont faciles à comprendre, telles que l'actualité et les indicateurs qui paraissent simples, comme les taux de réponse. L'utilisateur veut que l'organisme statistique producteur soit crédible, ce qui se traduit par la capacité de produire des données contenant des erreurs faibles ou du moins connues, et de les livrer en temps opportun, de manière fiable et accessible. L'idée qu'il serait possible de produire une mesure de la qualité totale fondée sur des évaluations pondérées des différentes dimensions n'est pas raisonnable, même si Mirotchie (1993) soutient le contraire. Dans son article, il présente des arguments en faveur d'un ensemble normalisé d'indicateurs de la qualité et donne un exemple hypothétique d'indicateurs de la qualité de données d'indexation et calcule un indice réel (dans cet exemple, les indicateurs sont la précision, la non-réponse, la fiabilité, l'actualité et les résidus). Même s'il était possible d'élaborer un indicateur composite sous forme d'un indice, l'utilisateur voudrait savoir quels indicateurs ont contribué le plus à la valeur de l'indice. Du point de vue de l'utilisateur, la valeur de l'indice la moins favorable pourrait encore refléter une situation offrant le plus haut niveau de qualité. Il est rare qu'une faible exactitude puisse être compensée par de bonnes évaluations sur d'autres dimensions, pas même dans le cas, lors des élections, des sondages faits à la sortie de l'isoloir, pour lesquels l'actualité est indispensable. L'exactitude demeure nécessaire et il est généralement reconnu que tous les organismes dignes de confiance doivent satisfaire aux normes d'exactitude (Scheuren 2001 ; Kalton 2001 ; Brackstone 2001). Phipps et Fricker (2011) donnent un aperçu des cadres de qualité et de la littérature sur l'erreur d'enquête totale. Donc, nous pouvons convenir que la qualité des enquêtes est un concept multidimensionnel faisant intervenir plusieurs caractéristiques d'un produit ou service statistique.

3.2 Les répercussions du mouvement en faveur de la qualité sur les organismes statistiques

Simplement élargir le cadre de la qualité pour passer d'une ou deux dimensions à plusieurs d'entre elles ne suffit

pas à créer un environnement propice à la qualité. À la fin des années 1980 et au début des années 1990, de nombreux organismes statistiques se sont intéressés aux problèmes de qualité dépassant les aspects habituels de la qualité des données. Les questions concernant la satisfaction des clients, la communication avec les clients, la concurrence, la variabilité des processus, le coût de la mauvaise qualité, le gaspillage, les modèles d'excellence opérationnelle, les valeurs fondamentales, les pratiques exemplaires, l'assurance de la qualité, et l'amélioration continue de la qualité ont soudain fait partie des préoccupations quotidiennes de nombreux organismes.

Les organismes qui réussissent savent qu'il est nécessaire de s'améliorer continuellement (Kaizen) pour rester en activité et ils ont mis au point des mesures qui les aident à évoluer. Cela s'applique également aux producteurs de statistiques. Les changements qui sont censés améliorer le produit statistique sont déclenchés par les demandes des utilisateurs, par la concurrence des autres producteurs et par les valeurs des producteurs qui mettent l'accent sur l'amélioration continue en tant qu'environnement de fonctionnement général. Les mesures qui peuvent aider un organisme statistique à s'améliorer sont essentiellement les mêmes que pour les autres entreprises. Elles peuvent s'appuyer sur des modèles d'excellence opérationnelle, tels que celui de l'European Foundation for Quality Management (EFQM) (1999). Les valeurs fondamentales sur lesquelles repose le modèle de l'EFQM sont l'orientation résultats, l'orientation client, le leadership et la constance des objectifs, le management par les processus et les faits, le développement et l'implication des personnes, la formation continue, l'innovation et l'amélioration, le développement des partenariats, et la responsabilité sociale de l'organisme. Ce modèle a été adopté par le Système statistique européen (SSE) comme outil permettant aux instituts statistiques nationaux d'atteindre le niveau voulu de qualité organisationnelle. Le concept est qu'il n'est pas possible d'arriver à un produit de bonne qualité, selon les dimensions mentionnées (ou une autre définition de la qualité du produit), si l'organisme ne met pas en place de bons processus sous-jacents. On pourrait également soutenir que le moyen le plus efficace et le plus fiable d'obtenir un produit de bonne qualité est d'utiliser des processus de bonne qualité. Si nous considérons la qualité comme un concept à trois niveaux, elle peut être visualisée comme on le présente au tableau 2.

3.2.1 Qualité du produit

Les résultats qu'il est convenu de livrer sont appelés le produit. Il peut s'agir d'estimations, de jeux de données, d'analyses, de registres, de processus normalisés ou d'autre matériel d'enquête, tel que des bases de sondage et des questionnaires. La qualité du produit correspond au concept classique de qualité utilisé pour informer les utilisateurs ou les clients de la qualité du produit ou du service. Elle peut être mesurée et contrôlée par le degré de respect des spécifications et des exigences quant aux caractéristiques du produit qui forment les dimensions de qualité d'un cadre. Les mesures de l'exactitude et les marges d'erreur entrent dans cette catégorie. Sont également pertinentes les observations en vue de déterminer si les ententes de niveau de service établies avec les clients ont été respectées. En harmonie avec les principes de gestion de la qualité, il est également assez fréquent de réaliser des sondages sur la satisfaction des utilisateurs afin de découvrir ce que ceux-ci pensent des produits et services fournis.

Tableau 2
Qualité - Concept à trois niveaux*

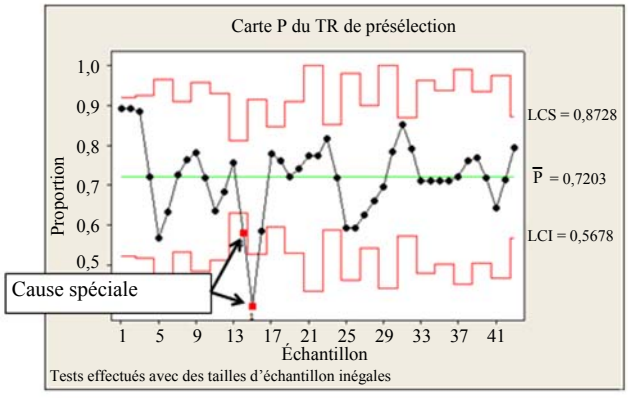
Niveau de qualité	Principaux intervenants	Instrument de contrôle	Mesures et indicateurs
Produit	Utilisateur, client	Spécifications du produit, ENS, études d'évaluation, cadres, normes	Cadres, conformité, EQM, sondage auprès des utilisateurs
Processus	Concepteur de l'enquête	CSP, cartes de contrôle, échantillonnage d'acceptation, analyse des risques, MMC, PON, paradonnées, listes de vérification, vérification	Variation au moyen de cartes de contrôle, analyse d'autres paradonnées, résultats des études d'évaluation périodique
Organisation	Organisme, propriétaire, société	Modèles d'excellence, ISO, CdP, examens, vérifications, autoévaluations	Scores, points forts et points faibles, sondages auprès des utilisateurs, sondages auprès du personnel

* ENS (Entente de niveau de service), CSP (contrôle statistique des processus), MMC (meilleures méthodes courantes), PON (procédure opérationnelle normalisée) et CdP (code de bonnes pratiques du SSE).

3.2.2 Qualité des processus

Tous les processus doivent être conçus de façon qu'ils fournissent ce qu'ils sont supposés produire. Cela signifie qu'une certaine perspective d'assurance de la qualité est nécessaire lorsque les processus sont définis. Ainsi, le processus d'interview implique qu'un certain nombre d'éléments doivent être en place pour que le processus livre ce qui est attendu. Ces éléments sont, par exemple un bon choix d'intervieweur, ainsi que la mise en place d'un programme de formation, d'un système de rémunération, ainsi que les activités de supervision et de rétroaction. Donc, nous nous employons à intégrer la qualité dans le processus par la voie de l'assurance de la qualité. Les activités de contrôle de la qualité ne sont utilisées que pour vérifier si le processus fonctionne comme il est prévu. Elles ne peuvent, à elles seules, être utilisées pour intégrer la qualité dans le processus. Cette vision des processus est discutée plus en détail à la section 4.4. La qualité des processus est mesurée et contrôlée par sélection, observation et analyses des principales variables de processus, ce que l'on appelle les données sur le processus ou paradonnées (Morganstein et Marker 1997 ; Couper 1998 ; Lyberg et Couper 2005). La théorie et les méthodes importées du domaine du contrôle statistique des processus peuvent aider le producteur à faire la distinction entre deux types de variation, à savoir la variation ordinaire et celle ayant une cause particulière. À condition que la variation totale soit contenue entre les limites supérieure et inférieure de contrôle associées aux cartes de contrôle choisies, le processus est déclaré être sous contrôle statistique et aucune amélioration n'est vraiment possible en essayant d'ajuster les résultats individuels. Des observations qui tombent en dehors des limites de contrôle (habituellement fixées à 3 sigma), sont une indication qu'il existe une cause spéciale de variation dont il faut s'occuper afin qu'après la correction, la variation soit ramenée à celle de cause ordinaire. La carte de contrôle P qui suit illustre une situation possible :

Carte de contrôle de processus avec valeurs extrêmes



Donc, l'enchaînement des actions est le suivant. D'abord, on recherche l'origine des causes spéciales afin de pouvoir éliminer ce type de variation. Après cela, le processus ne présente plus que la variation ordinaire. Si cette variation est jugée trop grande, le processus doit être modifié. Les types de changement nécessaires sont rarement évidents au départ. En effet, plusieurs changements sont parfois nécessaires pour réduire la variation du processus. Habituellement, il faut lancer un projet d'amélioration du processus et la littérature sur la gestion de la qualité propose un certain nombre d'outils utiles pour ce genre de projet. La plupart de ces outils sont empruntés à la statistique (cartes de contrôle, expériences, analyses de régression, diagramme de Pareto, nuages de points, stratification), mais il existe aussi des outils pour déterminer les causes probables du problème (diagrammes d'enchaînement du processus, séance de remue-méninges). Une opinion répandue est que les projets d'amélioration doivent être réalisés par les personnes qui travaillent avec le processus ou par des personnes très familiarisées avec ce dernier par d'autres moyens. Parfois, nous parlons de créer une équipe d'amélioration, à laquelle participe aussi le client. Dans tout projet d'amélioration, les changements proposés doivent être testés. Quand Shewhart a développé ses premières cartes de contrôle, il a également soutenu que le travail d'amélioration doit suivre une série d'opérations qu'il a appelées *Plan-Do-Check-Act* (planifier, faire, contrôler, agir). Cette séquence nous indique que tout changement qu'il est proposé d'apporter à un processus doit être testé pour voir s'il améliore réellement le processus. Dans la négative, un autre changement est fait et les tests sont répétés. Deming a appelé ce courant de pensée le cycle de Shewhart, mais puisqu'il a passé beaucoup de temps à le promouvoir, nombreux sont ceux qui ont fini par l'appeler cycle de Deming. Les changements recherchés pourraient être une réduction de la variation du processus, une réduction des coûts, ou un accroissement de la satisfaction des clients. La méthodologie des projets d'amélioration est décrite, par exemple, dans Joiner (1994), Box et Friends (2006), Breyfogle (2003) et Deming (1986).

Un autre moyen de vérifier la qualité du processus consiste à utiliser l'échantillonnage d'acceptation (Schilling et Neubauer 2009), qui peut être appliqué à des situations où les éléments du processus peuvent être groupés par lots. Les lots sont contrôlés et, en fonction du résultat de ce contrôle, il est décidé si le lot doit être approuvé ou retravaillé. Les plans d'échantillonnage d'acceptation garantissent une qualité sortante moyenne en ce qui a trait à, disons, le taux d'erreur, mais ne comportent aucune amélioration directe de la qualité. Il s'agit d'un instrument de contrôle qui convient pour des opérations telles que le codage, la vérification et le balayage optique, et ce, uniquement quand ces processus ne sont pas vraiment en contrôle statistique. La méthode a été

vivement critiquée par Deming (1986) et d'autres, mais peut représenter le seul moyen de contrôle disponible dans des situations où le roulement du personnel est élevé et que l'on ne dispose pas de suffisamment de temps pour attendre que les processus soient stables.

Les paradonnées globales (Scheuren 2001) sont des taux d'« erreurs » de différentes sortes. Les taux de non-réponse, les taux d'erreur de codage, les taux d'erreur de balayage optique et les taux d'erreur de listage en sont des exemples. Dans le cas de certaines opérations, les taux d'erreur sont calculés en recourant à la vérification, ce qui signifie que l'opération est répétée d'une certaine façon. Il en est ainsi de l'opération de codage. Pour d'autres opérations, le calcul peut être fondé sur un schéma de classification, comme pour le calcul des taux de non-réponse. Ces paradonnées globales nous renseignent sur le processus. Il s'agit de statistiques sur les processus, c'est-à-dire des sommaires de données. Un taux de non-réponse élevé signale des problèmes dans le processus de collecte des données et un taux élevé d'erreur de codage signale des problèmes dans le processus de codage. Partant de ces données sommaires, il est parfois possible de faire la distinction entre la variation dont la cause est ordinaire et celle dont la cause est spéciale, et de décider de la mesure à prendre.

Certains processus normalisés peuvent être contrôlés au moyen de simples listes de vérification. Ces dernières sont très efficaces parce qu'il est crucial que chaque étape du processus soit accomplie, et ce, dans le bon ordre (Morganstein et Marker 1997). La préparation au décollage effectuée par les pilotes d'avion en est un exemple. Peu importe le nombre de fois qu'ils ont décollé, sans liste de vérification, le jour viendra où ils oublieront un élément. Dans le domaine de la production de statistiques, l'échantillonnage est un processus de ce genre, même si les conséquences de l'oubli d'un élément sont moins graves. Il se pourrait fort bien qu'un organisme statistique possède un processus normalisé de sélection des échantillons et qu'une liste de vérification puisse être utilisée comme directive de travail et instrument de contrôle.

Une sorte de liste de vérification peut être utilisée dans les processus plus créatifs, tels que le processus de conception globale d'enquête. Il est impossible de normaliser ce processus, mais il est possible de dresser la liste d'un certain nombre d'étapes critiques qui doivent toujours être accomplies. La liste ne nous dit pas comment les accomplir. Elle sert juste à rappeler qu'une étape particulière ne doit pas être omise ni oubliée. Morganstein et Marker (1997) discutent de ce genre de liste de vérification et les appellent (ainsi que les listes de vérification plus simples) meilleures méthodes courantes (MMC). Ils décrivent le processus d'élaboration des MMC et la façon dont ces dernières peuvent être utilisées pour réduire la variation des processus dans les organismes statistiques. Ainsi, un organisme pourrait disposer de sept méthodes et systèmes d'imputation différents dans sa boîte à outils. Le maintien de ces sept systèmes coûte cher. Il est peu probable qu'ils soient tous aussi efficaces les uns que les autres. S'ils le sont, il n'est peut-être pas économiquement faisable de les retenir tous. Dans cette situation, une liste MMC qui décrit un plus petit nombre d'options pour l'organisme semble être une bonne idée. Elle pourrait être élaborée en créant une équipe d'amélioration comprenant les experts de l'imputation et certains clients. Les MMC sont censées être révisées lorsque de nouvelles connaissances sont acquises, ce qui implique qu'une date d'expiration est associée à chacune d'elle.

Dans un certain sens, les MMC sont naturellement des « pratiques exemplaires ». De nombreux organismes souhaitent mettre en œuvre et utiliser de telles pratiques. Morganstein et Marker offrent un processus pour élaborer ces pratiques exemplaires et les tenir à jour. Ce processus est utile pour un organisme s'il est possible de maintenir à un niveau minimum la variation de la conception des processus. Il devient alors facile de former le personnel et de modifier le processus quand il devient instable ou que de nouvelles méthodes sont mises au point. Par ailleurs, si les MMC et d'autres normes ne sont pas mises en application fermement au sein d'un organisme, leur usage ne sera pas répandu et l'investissement ne sera pas rentable.

3.2.3 Qualité organisationnelle

Les cadres sont responsables de la qualité au sens le plus large. C'est l'organisme qui assure le leadership, le perfectionnement du personnel, les outils permettant d'établir de bonnes relations avec la clientèle, les investissements et le financement. Le domaine de la gestion de la qualité nous a donné des modèles d'excellence opérationnelle qui peuvent nous aider à évaluer nos organismes statistiques de la même façon que d'autres entreprises le sont. Les deux principaux modèles d'excellence opérationnelle sont ceux du Baldrige National Quality Program et de l'European Foundation for Quality Management (EFQM).

Ces modèles consistent en une liste de critères à vérifier pour évaluer un organisme. Les sept principaux critères utilisés pour décerner le prix Malcolm Baldrige sont le leadership, la planification stratégique, l'orientation client et marché, l'information et l'analyse, l'orientation ressources humaines, la gestion des processus et les résultats opérationnels. Chaque critère comprend plusieurs sous-critères. Par exemple, l'orientation ressources humaines englobe les systèmes de travail, la formation et le perfectionnement des employés, et le bien-être et la satisfaction des employés. Les neuf critères sur lesquels s'appuie le modèle d'excellence de l'EFQM sont le leadership, les personnes, la stratégie, les partenariats et les ressources, les processus, produits et

services, les résultats personnes, les résultats clients, les résultats collectivité, et les résultats clés. Ces modèles peuvent être utilisés pour l'autoévaluation ou pour l'évaluation externe. L'organisme fournit une description de ce qui est en place en ce qui concerne chaque critère et il reçoit une note fondée sur cette description. Habituellement, les autoévaluations produisent des notes plus élevées que les évaluations externes. Il est très difficile d'obtenir une note élevée auprès d'évaluateurs externes, car les modèles sont très exigeants. Pour chaque critère, on demande à l'organisme s'il existe une bonne approche quelque part dans l'organisation. Cela est souvent le cas. La question suivante est celle de savoir dans quelle mesure cette bonne approche est répandue au sein de l'organisme. De nombreux organismes commencent à être en perte de vitesse à cette étape, car il y a peu de vrai dans la formule voulant que les bons exemples se répercutent dans toute l'organisation. Au contraire, les bonnes approches doivent habituellement être défendues énergiquement avant d'être acceptées au sein de l'organisme. La troisième question posée est celle de savoir si l'approche est évaluée périodiquement pour vérifier si elle produit les résultats attendus. C'est à ce stade que la plupart des organismes échouent. Leur stratégie consiste habituellement à épuiser une approche jusqu'à ce que les problèmes soient si importants qu'il faille la remplacer au lieu de l'ajuster. Évidemment, cette stratégie est perturbatrice et coûteuse, et ne reçoit pas une note élevée dans les évaluations de l'excellence. Le nombre maximal de points qui peut être obtenu avec ces modèles est égal à 1 000, et un gagnant obtient rarement plus de 450 à 600 points, ce qui est un indice qu'il y a matière à amélioration, même dans les organismes de calibre mondial.

Certains organismes statistiques ont utilisé des modèles d'excellence opérationnelle pour l'évaluation. L'office tchèque de statistique a été déclaré lauréat du Prix national de la qualité de la Tchécoslovaquie pour 2009 dans la catégorie Secteur public en se basant sur le modèle d'excellence de l'EFQM. Il a obtenu 464 points. Le Leadership Group-Qualité d'Eurostat a recommandé que les organismes statistiques nationaux européens utilisent le modèle de l'EFQM pour leurs travaux sur la qualité, et les organismes de la Finlande et de la Suède comptent parmi ceux qui l'ont fait. Depuis que le Leadership Group a publié son rapport en 2001 (voir Lyberg, Bergdahl, Blanc, Booleman, Grünewald, Haworth, Japec, Jones, Körner, Linden, Lundholm, Madaleno, Radermacher, Signore, Zilhao, Tzougas et van Brakel 2001), d'autres cadres et normes ont été élaborés. Le Système statistique européen a lancé son code de bonnes pratiques, qui compte un certain nombre de principes associés à des indicateurs. Pour certains de ces principes, cependant, les indicateurs constituent plutôt des éclaircissements. La liste de principes ressemble à d'autres listes qui ont été établies par l'ONU et d'autres organisations.

Les évaluations externes sont probablement plus fiables que les évaluations internes, et ce, pour plusieurs raisons. L'une est qu'il vous est difficile de critiquer vos pairs puisque vous devrez interagir avec eux dans l'avenir ou que votre propre produit ou service sera évalué par eux dans l'avenir. Les expériences vécues à Statistics Sweden et à Statistique Canada montrent que la capacité des autoévaluations à dégager les faiblesses importantes est limitée (voir la section 5.3).

3.2.4 Certaines conséquences particulières pour les organismes statistiques

La plupart des organismes statistiques adoptent les principes de gestion de la qualité à des degrés divers et avec plus ou moins de succès. Comme l'ont fait remarquer Colledge et March (1993), il est possible d'énumérer plusieurs obstacles à la mise en œuvre de ces principes. Un organisme gouvernemental peut avoir de la difficulté à motiver son personnel au moyen de primes monétaires, puisque la façon dont l'argent des contribuables peut être dépensé fait l'objet de restrictions. La diversité des utilisateurs et des produits complique le dialogue entre le fournisseur de services et l'utilisateur, et comme il est mentionné plus haut, ni les utilisateurs ni d'ailleurs les fournisseurs des données ne sont entièrement familiarisés avec tous les biais et autres problèmes de qualité présents dans la production de statistiques. L'effet des erreurs sur les utilisations des données peut varier et est souvent inconnu. La situation se complique encore davantage du fait que, contrairement à ce que connaissent la plupart des autres entreprises, les fournisseurs des organismes statistiques ne sont pas très enthousiastes. Les fournisseurs des autres entreprises sont payés, tandis que ceux des organismes statistiques, c'est-à-dire les répondants, qui reçoivent rarement un incitatif monétaire, doivent être motivés.

Par ailleurs, les organismes statistiques ont un énorme avantage quand il s'agit d'appliquer les principes de gestion de la qualité. Un organisme statistique sait comment recueillir et analyser les données qui orientent les efforts d'amélioration. L'une des pierres angulaires des concepts de gestion de la qualité est que les décisions doivent être fondées sur des données, et souvent, les entreprises qui ne bénéficient pas de l'appui des statisticiens ne sont pas au courant des problèmes de qualité des données qui peuvent avoir des répercussions sur les décisions qu'elles prennent. Néanmoins, dans l'ensemble, un organisme statistique n'est pas différent de toute autre entreprise et il lui est fort possible d'appliquer les concepts de gestion de la qualité afin d'améliorer tous les aspects de son travail.

4. Exemples de projet d'amélioration de la qualité dans les organismes statistiques

La présente section donne des exemples de projets entrepris par les organismes statistiques en raison de l'intérêt général pour la qualité que manifeste la société.

4.1 L'erreur d'enquête totale

L'aspect peut être le plus important qu'il convient de souligner est que le domaine de la recherche et du développement portant sur la conception et la mise en œuvre des enquêtes, l'échantillonnage et les erreurs non dues à l'échantillonnage, ainsi que les effets des erreurs sur l'analyse des données, demeure florissant. L'obtention de données entachées de faibles erreurs est l'objectif principal des organismes de bonne réputation, comme en témoigne la publication régulière de manuels sur la collecte des données, l'échantillonnage, la non-réponse, la conception des questionnaires, les erreurs de mesure et les études comparatives. De nouveaux manuels traitant de sujets tels que les enquêtes auprès des entreprises, la traduction du matériel d'enquête et les paradonnées sont en cours de rédaction en vue de combler les lacunes dans ces domaines. Des revues, dont le *Journal of Official Statistics, Techniques d'enquête* et *Survey Practice*, sont entièrement consacrées à des sujets liés à la production de statistiques au sens large. De nombreuses autres revues, telles que le *Public Opinion Quarterly*, le *Journal of the American Statistical Association* et le *Journal of the Royal Statistical Society*, consacrent beaucoup d'espace aux méthodes d'enquête. La *Wiley series in Survey Methodology* et les conférences connexes (sur les enquêtes par panel, les méthodes d'enquête téléphoniques (deux), les erreurs de mesure, la qualité des processus, les enquêtes-entreprises, la mise à l'essai et l'évaluation des questionnaires, la collecte de données d'enquête assistée par ordinateur, la non-réponse et les enquêtes comparatives) a eu beaucoup de succès et il en est de même des ateliers continus sur la non-réponse et l'erreur d'enquête totale. Donc, les idées quant aux sources d'erreur particulières et à leur traitement ne font pas défaut. Certains domaines, tels que les erreurs de spécification, les erreurs de traitement des données et l'effet des erreurs sur l'analyse des données, sont, certes, sous-étudiés, mais, dans l'ensemble, l'élargissement des connaissances sur les erreurs d'enquête suscite un véritable intérêt. Le défi tient à la communication de ces connaissances aux personnes qui travaillent dans les organismes statistiques et à l'élaboration de principes de conception qui peuvent être appliqués pour améliorer la production des statistiques. Une fracture évidente existe entre ce qui est connu grâce à la recherche et ce qui est connu et appliqué dans les organismes statistiques. Donc, il semble nécessaire de renforcer continuellement les capacités du personnel, en particulier parce que l'idée reçue voulant que les bons exemples se propagent comme des ondulations au sein des organismes statistiques et entre ceux-ci est un mythe. En effet, si cela se produisait vraiment, la qualité serait maintenant fantastique partout. Comme elle ne l'est pas, de nombreux organismes ont établi des programmes de formation de grande portée (Lyberg 2002).

4.2 Risque et gestion du risque

L'un des éléments de la gestion de la qualité qui a fait son entrée dans l'univers des enquêtes est le risque et la gestion de ce dernier. Eltinge (2011) parle même du risque d'enquête total (*Total Survey Risk*) comme alternative au paradigme de l'erreur d'enquête totale. L'identification et la gestion des risques est un volet important de la vérification interne moderne (Moeller 2005) et est peut-être le seul élément important absent des cadres de gestion de la qualité, tels que celui de l'EFQM. Une source d'erreur peut être jugée comme posant un plus grand risque qu'une autre et doit, par conséquent, être traitée avec plus de soin et de ressources. Par exemple, ne pas posséder de système efficace de contrôle de la divulgation statistique est considéré comme une situation très risquée. Historiquement, la divulgation illégale de données est très rare, mais lorsqu'elle a lieu, elle risque de saper toutes les tentatives ultérieures de collecte de données. Certaines décisions concernant la conception des enquêtes peuvent être considérées comme risquées. Par exemple, si nous choisissons une méthode de collecte des données qui n'est pas adaptée au sujet de l'enquête, nous pourrions obtenir des estimations qui s'écartent tellement de la vérité que les résultats seront inutiles. L'étude de comportements de caractère délicat par interview sur place ou par téléphone au lieu d'un questionnaire à remplir soi-même pourrait en être un exemple. Il existe également des risques techniques qui doivent être décelés et évalués. Ainsi, l'U.S. National Agricultural Statistical Service (Gleaton 2011) possède, comme de nombreux autres organismes, des plans de reprise après sinistre. Groves (2011) et Dillman (1996) discutent tous deux des visions différentes des risques qui pourraient émaner de la culture de production et de la culture de recherche au sein d'un organisme statistique. Le changement s'opère généralement lentement dans ces organismes, et ce, parfois, pour de bonnes raisons. Le changement pourrait aboutir à un échec, tel qu'une mise en œuvre infructueuse, des coûts importants et une réduction de la comparabilité des données. Donc, dans un certain sens, tant les producteurs que les utilisateurs des données ont tendance à hésiter à adopter les changements proposés par les chercheurs et par les innovateurs, ce qui pourrait être l'une des raisons de la lenteur avec laquelle les changements ont lieu. Il est courant de produire des mesures parallèles

pendant un certain temps afin de traiter les risques associés à la mise en œuvre d'une nouvelle méthode ou d'un nouveau système. Selon Groves (2011), la culture de la production et les utilisateurs ont eu le dernier mot au sujet de tout changement, du moins jusqu'à présent. Simultanément, l'innovation est désespérément nécessaire dans de nombreux systèmes de production et il existe des exemples d'organisations cloisonnées auxquelles il ne reste plus beaucoup de temps (avant de devoir changer), parce que les ressources pour maintenir leurs systèmes font tout simplement défaut. Donc, même en cas de résistance au changement, le manque de ressources et la concurrence feront en sorte que les organismes statistiques deviennent davantage axés sur les processus et plus efficaces. Réduire le nombre de systèmes et d'applications, et privilégier une plus grande normalisation semblent être l'une des voies d'avenir.

4.3 Le client/l'utilisateur

L'apparition des concepts de gestion de la qualité dans les organismes statistiques a rendu plus visibles les destinataires des produits et services statistiques. Les entreprises commerciales parlent toujours de leur client, tandis que les organismes gouvernementaux ont eu tendance à les appeler des utilisateurs. Quoi qu'il en soit, la prise en compte du fait que quelqu'un est censé utiliser les produits finaux ne semble pas avoir été évidente pour certains fournisseurs. Il faut admettre que l'utilisateur a été un interlocuteur depuis que l'industrie des enquêtes a vu le jour. Aux États-Unis, les conférences à l'intention des utilisateurs étaient déjà assez fréquentes il y a 50 ans (Dalenius 1968 ; Hansen et Voight 1967). Ainsi, durant six mois de 1965 à 1966, le U.S. Census Bureau a organisé 23 conférences à l'intention des utilisateurs à travers le pays et a aussi organisé des groupes consultatifs. De nombreux pays ont privilégié les contacts de nature consultative avec les utilisateurs. Les conférences à l'intention des utilisateurs ont encore lieu, mais l'apport des utilisateurs est maintenant complété par d'autres moyens, tels que des discussions publiques et des forums sur Internet. Il est rare que les utilisateurs aient participé directement à la planification et à la conception des enquêtes. Même lors des discussions au sujet de la qualité des données, les producteurs ont agi en tant que représentants des utilisateurs. Les cadres de la qualité en sont un bon exemple. Les dimensions de la qualité ont été définies en consultant très peu les utilisateurs. La littérature traitant de la façon dont ceux-ci perçoivent l'information sur la qualité est extrêmement limitée (Groves et Lyberg 2010). Qui plus est, nous ne savons pas si l'information sur la qualité que nous fournissons leur est utile (Dalenius 1985b). En fait, une supposition éclairée est que, souvent, elle ne l'est pas. Dans le cas de beaucoup d'enquêtes, les utilisateurs sont nombreux et parfois inconnus, et il n'est pas possible de prévoir leurs besoins d'information et d'analyse. Souvent, on peut isoler un ou quelques utilisateurs principaux avec lesquels communiquer, mais bon nombre de problèmes ayant trait à la conception et à la qualité des enquêtes sont tellement compliqués qu'une grande majorité d'utilisateurs s'attendent à ce que le fournisseur de services leur livre un produit contenant la plus petite erreur possible. Hansen et Voight ont déclaré que l'exactitude devrait être d'un niveau suffisant pour éviter les problèmes d'interprétation. Aujourd'hui, il semble exister un consensus voulant que les utilisateurs recherchent des produits et des services dans lesquels ils peuvent avoir confiance, ce qui signifie que le fournisseur de services doit être crédible. La plupart des utilisateurs n'ont pas la possibilité de vérifier les niveaux d'exactitude. Les aspects dont un utilisateur moyen peut discuter sont les questions ayant trait, par exemple, à l'exactitude, à l'accessibilité et à la pertinence. Des discussions détaillées au sujet de questions techniques et de problèmes de compromis de conception entre l'exactitude et la comparabilité sont plus difficiles à obtenir.

Au cours des dernières décennies, l'utilisateur a effectivement pris plus d'importance. Certains organismes élaborent avec un utilisateur ou un client important des ententes de niveau de service qui énumèrent les exigences concernant le produit ou service final afin qu'une vérification puisse avoir lieu au moment de la livraison. De nombreux organismes qui réalisent des enquêtes auprès des entreprises ont créé des unités qui communiquent continuellement avec les entreprises les plus grandes, puisque leur participation et leur fourniture de données exactes sont absolument essentielles au processus d'estimation (Willimack, Nichols et Sudman 2002). Les grandes entreprises ne sont pas des utilisateurs au sens strict. Il s'agit de fournisseurs importants ayant souvent un intérêt dans les résultats de l'enquête. Un autre outil de communication est le sondage sur la satisfaction des clients. La valeur de ce genre de sondage est limitée en raison du phénomène d'acquiescement et de la difficulté à trouver un répondant bien informé qui est prêt à répondre. En outre, de nombreux sondages sur la satisfaction des clients s'appuient sur l'autosélection, de sorte qu'ils n'ont aucune valeur inférentielle. Les résultats de ces sondages peuvent être vus seulement comme des listes de problèmes et de préoccupations dont font part certains clients. Cette information peut évidemment être fort utile, mais elle ne convient pas pour l'estimation. De nombreux organismes d'enquête réalisent maintenant des sondages auprès des utilisateurs en continu (Ecochard, Hahn et Junker 2008).

4.4 La vue du processus

L'approche de la gestion de la qualité a de nouveau mis en relief qu'il importe d'avoir une vue du processus dans la

production des statistiques. Considérer le processus de production comme une série d'actions ou d'étapes en vue d'atteindre un but particulier qui satisfait l'utilisateur mène à un produit de bonne qualité. La qualité du processus est évaluée en déterminant dans quelle mesure chaque étape satisfait à des exigences ou à des spécifications définies. Un moyen de contrôler la qualité du processus consiste à recueillir des données sur le processus qui peuvent varier avec chaque répétition de celui-ci. Les variables du processus qu'il est intéressant de surveiller sont celles qui ont un effet important sur le résultat final du processus. Donc, afin de vérifier la stabilité et la variation d'un processus, nous avons besoin de mécanismes pour cerner les variables clés et pour recueillir et analyser des données sur ces variables. La science de la gestion de la qualité nous a donné des outils tels que le diagramme en arrêtes de poisson d'Ishikawa pour déterminer quelles pourraient être les variables clés du processus. La méthodologie de contrôle statistique des processus nous a donné des outils pour faire la distinction entre la variation dont la cause est spéciale et celle dont la cause est ordinaire et déterminer comment traiter ces deux types de variation. Habituellement, nous nous servons de cartes de contrôle qui ont été développées au départ par Shewhart (Deming 1986 ; Mudryk, Burgess et Xiao 1996) pour faire ces distinctions. Ensuite, nous recourons de nouveau à des méthodes issues de la science de la gestion de la qualité pour ajuster le processus, au besoin. Les organigrammes, ou diagrammes de flux, les diagrammes de Pareto et d'autres moyens simples permettant à l'équipe de production de repérer les causes fondamentales des problèmes en sont des exemples Juran 1988).

Les données sur les processus ont été utilisées pour contrôler les processus employés dans la production de statistiques depuis les années 1940, d'abord au U.S. Census Bureau, puis à Statistique Canada et, dans une certaine mesure, dans d'autres organismes également. Les processus habituellement vérifiés comprenaient le codage, la saisie et l'impression des données et les données sur les processus étaient principalement des taux d'erreur. Certains contrôles des processus utilisés par le U.S. Census Bureau étaient tellement compliqués et coûteux que leur valeur a été mise en doute (Lyberg 1981), surtout parce que les boucles de rétroaction qui y étaient associées étaient inefficaces et ne visaient pas toujours à déterminer les causes fondamentales des erreurs. Il était courant de blâmer les opérateurs pour les problèmes causés par les systèmes et aucun accent n'était mis sur l'amélioration continue de la qualité. À l'époque, la réflexion était davantage axée sur la vérification et la correction.

Morganstein et Marker (1997) ont conçu un plan générique d'amélioration continue du processus qui peut être appliqué à la production de statistiques. Depuis les années 1980, ils avaient travaillé dans de nombreux organismes statistiques et constaté que, dans la plupart des cas, la réflexion au sujet de la qualité n'était pas très avancée. Ils ont fondé leur plan générique sur leurs expériences pratiques et sur les notions générales de gestion de la qualité, exposées entre autres par Juran (1988), Deming (1986), Box (1990), et Scholtes, Joiner et Streibel (1996). Ce plan compte essentiellement sept étapes :

- les caractéristiques critiques du produit sont précisées en collaboration avec l'utilisateur, en ce qui concerne les besoins généraux ainsi que les besoins plus uniques ;
- un schéma du déroulement du processus est élaboré par une équipe bien au courant. Le schéma doit comprendre la séquence des étapes du processus, les points de décision et les clients pour chaque étape ;
- les variables clés du processus sont identifiées parmi un ensemble plus grand de variables du processus ;
- la capacité de mesure est évaluée. Il est important que les décisions soient fondées sur de bonnes données et non pas simplement sur des données. Celles qui sont disponibles pourraient être inutiles. Il s'agit d'un domaine où les organismes statistiques sont avantagés par rapport à d'autres organismes. On ne doit pas tirer de conclusion au sujet de la stabilité du processus sans disposer d'information sur les erreurs de mesure. Avant tout, les données doivent permettre de quantifier l'amélioration ;
- la stabilité du processus est déterminée. Le schéma de variabilité des données sur le processus est analysé en utilisant des cartes de contrôle et d'autres outils statistiques ;
- la capacité du système est déterminée. Si la stabilité n'est pas atteinte après que la variation due à des causes spéciales a été éliminée, un effort d'amélioration est nécessaire. Des modifications peuvent être apportées au système lorsque la variation du processus est tellement grande que les spécifications, telles que les taux d'erreur minimaux ou les échéances de production, ne sont pas satisfaites. Des méthodes typiques en vue de réduire la variation sont l'élaboration et la mise en œuvre d'un nouveau programme de formation ou la mise en application d'une procédure opérationnelle normalisée. Cette dernière peut être une norme de processus, une norme relative aux meilleures méthodes courantes ou une simple liste de vérification ;
- la dernière étape du plan d'amélioration consiste à établir un système de surveillance permanente du processus. Nous ne pouvons pas nous attendre à ce que les processus restent stables au fil du temps. Pour de nombreuses raisons, une dérive s'amorce habituellement après un certain temps. Un système de surveillance facilite le suivi des nouvelles structures d'erreur, des nouvelles exigences des clients et des améliorations

possibles des méthodes et de la technologie, et permet de proposer des améliorations des processus.

Ce chapitre du livre de Morganstein et Marker a nettement influencé les travaux portant sur la qualité et la réflexion concernant les processus dans de nombreux organismes statistiques européens. L'intérêt pour ces questions s'est accru et certains organismes ont lancé leur propre système de gestion de la qualité dont l'amélioration des processus était un élément central.

Aux Joint Statistical Meetings de 1998, Mick Couper a donné un exposé sollicité sur la mesure de la qualité dans un environnement de collecte de données d'enquête assistée par ordinateur (CASIC). Il a mentionné que la nouvelle technologie produisait une foule de données secondaires susceptibles d'être utilisées pour améliorer le processus de collecte des données. Il a donné à ces données secondaires le nom de paradonnées, non pas dans son article, mais dans son exposé. Ce nom a été adopté très rapidement dans le monde des enquêtes et il était logique de définir la trilogie des données, des métadonnées et des paradonnées. Donc, nous disposions d'un terme pour les données au sujet des données (métadonnées) et un autre pour les données au sujet du processus (paradonnées). Les paradonnées sont manifestement des données sur le processus, mais très longtemps elles ont été limitées aux données au sujet du processus de collecte des données, alors que le terme utilisé dans de nombreux organismes statistiques européens était « données sur le processus » et tenaient compte de tous les processus d'enquête (Aitken, Hörngren, Jones, Lewis et Zilhao 2004). Récemment, on a assisté à un nouvel élargissement de la signification du concept. Kennickell, Mulrow et Scheuren (2009) nous rappellent ce qu'ils nomment macro-paradonnées, c'est-à-dire les données sur le processus global, tels que les taux de réponse, les taux de couverture, les taux de rejet au contrôle et les taux d'erreur de codage, qui ont toujours été des indicateurs de la qualité du processus dans les organismes statistiques. Lyberg et Couper (2005), Kreuter, Couper et Lyberg (2010), et Smith (2011) emploient aussi la signification plus inclusive des paradonnées, qui tient compte d'autres processus que la collecte des données. Il existe un risque que, comme celui de qualité, le concept de paradonnées soit utilisé exagérément. On trouve des exemples de discussions dans lesquelles toutes les données, sauf les estimations d'après les données d'enquête, sont considérées comme des paradonnées, ce qui, naturellement, n'a aucun sens.

Les paradonnées ont reçu un nom génial et elles sont nécessaires pour juger de la qualité du processus. Cependant, la prudence est de rigueur. On ne doit jamais recueillir des paradonnées qui ne sont pas reliées à la qualité du processus et il est important de savoir comment les analyser. Parfois, les méthodes de contrôle statistique des processus peuvent être appliquées, mais parfois d'autres techniques analytiques sont nécessaires. Par exemple, pour pouvoir contrôler la falsification des données par les intervieweurs, il se pourrait que l'on doive examiner plusieurs processus simultanément, mais que la théorie et la méthodologie pour appuyer cette analyse ne soit pas directement disponible.

L'usage élargi de microdonnées qui ont trait à des enregistrements individuels, telles que les données sur les touches frappées et les enregistrements marqués d'un indicateur d'imputation, découle de l'utilisation des nouvelles technologies. Les procédures de collecte de données modernes produisent d'énormes quantités de ces types de paradonnées, tout comme le font aussi les systèmes de codage manuel assisté par ordinateur et les systèmes de codage entièrement automatisé, ainsi que les systèmes de balayage optique des données. Il n'est pas logique de limiter le concept à la collecte des données.

La science de la gestion de la qualité nous a appris à prévenir les problèmes concernant les processus au lieu de les corriger au moment où ils se manifestent, et nous a fait découvrir qu'il est important de faire la distinction entre différents types de variation du processus, puisqu'ils nécessitent des interventions différentes, que toute intervention ou amélioration touchant le processus doit être fondée sur de bonnes données et des méthodes d'analyse appropriées et que même des processus stables finissent par dériver, de sorte que la surveillance doit être permanente.

4.5 Normalisation et outils similaires

Un moyen de maintenir la qualité du processus sous contrôle consiste à réduire la variation en favorisant l'utilisation de normes et de documents similaires. Colledge et March (1997) discutent de quatre catégories de documents.

- Une norme est un document qui doit être respecté presque sans exception. Les écarts par rapport à la norme sont déconseillés et requièrent l'approbation de la haute direction. Des mesures correctives doivent être prises lorsqu'une norme n'est pas entièrement satisfaite. Un organisme peut obtenir une certification de conformité à une norme. Il en est ainsi des normes ISO dont quelques-unes sont pertinentes pour les organismes statistiques.
- Une politique doit être appliquée sans exception. Par exemple, un organisme peut avoir une politique concernant l'utilisation de mesures incitatives pour accroître les taux de réponse.
- Plusieurs organismes ont élaboré des lignes directrices pour différents aspects de la production de statistiques. Habituellement, les lignes directrices peuvent être outrepassées s'il y a de « bonnes » raisons de le faire.

- Une pratique recommandée est une pratique privilégiée, mais il n'est pas obligatoire d'y adhérer.

Les catégories de cette classification ne sont certes pas mutuellement exclusives, surtout si l'on tient compte également des aspects linguistiques et culturels. Par exemple, en suédois, les politiques et les lignes directrices sont conceptuellement très proches. Wikipédia, encyclopédie libre mais fondée sur un consensus, dit que les politiques décrivent des normes, tandis que les lignes directrices décrivent les pratiques exemplaires qui permettent de suivre ces normes. Cette phrase contient trois des catégories mentionnées par Colledge et March. Le mieux est probablement de traiter ces divers types de documents de la même façon. Ils visent tous à améliorer la qualité en réduisant divers types de variations et nous ne devons pas nous attarder trop sur la façon dont ils sont appelés.

Bien que les normes aient occupé une place importante dans la méthodologie d'enquête depuis longtemps, leur rôle s'est accru depuis que les organismes statistiques ont commencé à s'intéresser à la gestion de la qualité. Les premières normes, comme celles de Hansen et coll. (1967) et du U.S. Bureau of the Census (1974), étaient axées sur la discussion de la présentation des erreurs dans les données. Au U.S. Census Bureau, toutes les publications doivent informer les utilisateurs que les données sont sujettes à erreur, que l'analyse pourrait être affectée par ces erreurs et que les erreurs d'échantillonnage estimées sont plus faibles que les erreurs totales. Dans le cas des grandes enquêtes, les erreurs non dues à l'échantillonnage doivent être traitées de manière plus détaillée, contrairement à ce qui se faisait dans le passé. De nombreux autres organismes statistiques ont adopté cette façon de penser. Par exemple, les cadres de la qualité mentionnés précédemment sont des extensions qui englobent d'autres dimensions de la qualité que l'exactitude. Le Système statistique européen a élaboré et lancé successivement ce qu'on a d'abord appelé les *Model Quality Reports* (rapports modèles sur la qualité), qui sont devenus aujourd'hui simplement la *Standard for Quality Reports* (norme pour les rapports sur la qualité) (Eurostat 2009a). La norme formule à l'intention des instituts nationaux de statistique européens (noter la complexité conceptuelle) des recommandations pour la préparation de rapports sur la qualité pour une gamme « complète » de processus statistiques et de leurs produits. La norme traite des dimensions fondamentales de la qualité, à savoir la pertinence, l'exactitude, l'actualité, l'accessibilité, la cohérence et la comparabilité.

Examinons certains exemples. En ce qui concerne l'erreur de mesure, qui fait partie de la composante d'exactitude, la norme dit qu'un rapport sur la qualité doit contenir l'information suivante :

- la détection et l'évaluation générale des principaux risques en ce qui a trait à l'erreur de mesure ;
- si elles sont disponibles, les évaluations fondées sur des comparaisons à des données externes, la répétition des interviews ou des expériences ;
- l'information sur les taux de rejet durant la vérification des données ;
- les efforts déployés pour concevoir et mettre à l'essai les questionnaires, l'information sur la formation des intervieweurs et d'autres travaux sur la réduction des erreurs ;
- les questionnaires utilisés qui doivent être annexés au rapport sous une forme ou l'autre.

En ce qui concerne l'actualité, la norme dit que les rapports doivent contenir l'information suivante :

- pour les diffusions de données annuelles ou moins fréquentes : la durée moyenne de production pour chaque diffusion de données ;
- pour les diffusions de données annuelles ou plus fréquentes : le pourcentage de diffusions effectuées à temps, selon les dates de diffusion planifiées ;
- les raisons des diffusions tardives.

La norme comprend aussi des sections sur la façon de communiquer l'information au sujet des compromis entre les diverses dimensions de la qualité, l'évaluation des besoins et les perceptions des utilisateurs, le rendement et le coût, le fardeau de réponse, ainsi que la confidentialité, la transparence et la sécurité. Bien qu'elles comprennent une section sur les besoins et les perceptions des utilisateurs, ces derniers n'ont manifestement pas participé à la préparation de la norme proprement dite. Nous n'en savons toujours que fort peu sur la façon dont les utilisateurs perçoivent et utilisent l'information au sujet de la qualité. La norme est appuyée par un manuel beaucoup plus détaillé concernant les rapports sur la qualité (Eurostat 2009b) et les deux documents s'articulent autour des 15 principes énumérés dans le Code de bonnes pratiques de la statistique européenne, qui constitue le cadre de qualité fondamental pour le Système statistique européen. Les principes du Code de bonnes pratiques ont trait à l'indépendance professionnelle, le mandat pour la collecte des données, l'adéquation des ressources, l'engagement sur la qualité, le secret statistique, l'impartialité et l'objectivité, une méthodologie solide, des procédures statiques adaptées, la limitation du fardeau imposé au répondant, le rapport coût-efficacité, la pertinence, l'exactitude et la fiabilité, l'actualité et la ponctualité, la cohérence et la comparabilité, et enfin, l'accessibilité et la clarté. Chaque principe est accompagné d'un ensemble d'indicateurs que les organismes individuels peuvent mesurer pour déterminer s'ils sont ou non en conformité avec le Code. Certains indicateurs sont vagues et de nature

très subjective, comme « l'étendue, la précision et le coût des statistiques européennes sont proportionnés aux besoins », tandis que d'autres sont plus spécifiques, tels que « un horaire standard de diffusion des statistiques européennes est porté à la connaissance du public ». Des examens par les pairs de la conformité à un ensemble limité de principes menés en utilisant une version antérieure du Code ont révélé, ce qui n'est pas surprenant, que de nombreux organismes statistiques nationaux en Europe ont de la difficulté à les respecter (Eurostat 2011a). Par conséquent, afin de faciliter la mise en œuvre du Code, on a élaboré un cadre de soutien, appelé cadre d'assurance de la qualité (CAQ) qui contient des directives plus spécifiques concernant les méthodes et les références (Eurostat 2011b). Ce cadre semble être un document fort utile, car ses références sont principalement des résumés de l'état des connaissances dans des domaines tels que l'échantillonnage, la conception de questionnaires, la vérification, et ainsi de suite qui encourage la conformité aux pratiques exemplaires courantes.

Le Code de bonnes pratiques présente de nombreuses similarités avec les Principes fondamentaux de la statistique officielle de l'ONU (de Vries 1999). Ces principes promeuvent aussi la coopération et la coordination internationales, qui constituent, en grande partie, un élément qui fait défaut dans le développement actuel de la production de statistiques (Kotz 2005). Même des pays voisins peuvent suivre des approches très différentes et posséder des niveaux de compétence en méthodologie très différents, et les différences sont parfois difficiles à expliquer. Nous savons par expérience que la collaboration en matière de développement est difficile à réaliser. Nous nous réunissons, nous parlons et nous ramenons des idées qui pourraient s'adapter à nos systèmes. En revanche, il est plus difficile de se mettre d'accord sur des approches communes. Une norme globale qui se rapporte à la production de statistiques est l'ISO 20252 – Études de marché, études sociales et d'opinion (Organisation internationale de normalisation 2006). Il s'agit d'une norme de procédure comprenant environ 500 exigences concernant les activités de recherche au sein d'un organisme. Il s'agit d'une norme minimale portant sur ce qu'il faut faire plutôt que sur la façon dont il faut faire les choses. Elle est appropriée pour les organismes qui réalisent des enquêtes et ils peuvent faire une demande de certification. En avril 2010, plus de 300 organismes du monde entier, dont la plupart étaient des entreprises de marketing, ont été certifiés. Un organisme statistique national (Uruguay) a été certifié en 2009 et Statistics Sweden prévoit obtenir la certification en 2013, mais ces organismes nationaux sont les seuls à s'être engagés sur cette voie. La norme porte sur le système de gestion de la qualité de l'organisme, ainsi que la gestion des éléments exécutifs de la recherche, la collecte des données, la gestion et le traitement des données, ainsi que la production de rapports sur les projets de recherche (Blyth 2012).

Les normes du système statistique fédéral des États-Unis sont axées sur la composante d'exactitude. Bien qu'il ne s'agisse pas officiellement d'une norme, le U.S. Federal Committee on Statistical Methodology (2001) propose diverses méthodes pour mesurer et communiquer les sources d'erreur dans les enquêtes. En 2002, l'Office of Management and Budget (OMB) des États-Unis a publié des lignes directrices sur la qualité de l'information (OMB 2002) dont l'objectif était d'assurer et de maximiser la qualité, l'objectivité, l'utilité et l'intégrité de l'information diffusée par les organismes fédéraux. L'OMB (2006a) a également émis des normes et des lignes directrices pour les enquêtes. Elles sont bâties d'une manière classique. Vient d'abord une norme telle que « les taux de réponse doivent être calculés en utilisant des formules normalisées pour mesurer la proportion de l'échantillon admissible qui est représentée par les unités répondantes dans chaque étude, à titre d'indicateur du biais de non-réponse possible. » Cette norme est suivie d'un certain nombre de lignes directrices indiquant comment faire les calculs nécessaires, tandis que la dernière de ces lignes directrices précise que « si le taux de non-réponse globale dépasse 20 %, une analyse du biais de non-réponse doit être effectuée pour voir si les données manquent entièrement au hasard. » Comme dans le cas des normes du SSE, les lignes directrices de l'OMB sont complétées par un document de soutien (OMB 2006b) pour faciliter le respect des normes.

Dans le système statistique fédéral décentralisé des États-Unis, la plupart des organismes ont produit des documents dans lesquels sont adaptées les lignes directrices de l'OMB. Par exemple, l'U.S. Census Bureau possède ses propres normes de qualité statistique dont le niveau de détails techniques est plus élevé que celui des documents de l'OMB. Chaque norme est décrite au moyen d'exigences et de sous-exigences, et le document fournit souvent des exemples très spécifiques d'études qui peuvent être réalisées. Le National Center for Health Statistics, le National Center for Education Statistics, et l'Energy information Administration sont d'autres exemples d'organismes américains dotés de normes concernant la qualité de l'information diffusée. Toutes ces normes peuvent être téléchargées à partir des sites Web de ces organismes.

Statistique Canada a émis des lignes directrices concernant la qualité depuis 1985. Elles sont similaires à celles du SSE puisqu'elles ne se limitent pas à mettre l'accent sur l'exactitude. Toutefois, elles sont nettement plus détaillées et contiennent un grand nombre de références. Une caractéristique particulière est que, pour certains processus, les lignes directrices prescrivent l'utilisation du contrôle

statistique des processus. Aucun autre organisme ne semble le faire.

L'édition la plus récente de ces lignes directrices est donnée dans Statistique Canada (2009).

De nombreux autres organismes statistiques de par le monde possèdent leurs propres normes de qualité. Elles sont parfois décrites comme des lignes directrices ou des normes et parfois, comme des systèmes de soutien opérationnel ou des cadres d'assurance de la qualité. Quoi qu'il en soit, le contenu et le style varient d'un organisme à l'autre, mais il faut que la variation soit gérable. Il devrait être possible d'arriver mondialement à un plus haut degré de normalisation, puisque cela s'est fait dans d'autres domaines, tels que les voyages aériens. Apted, Carruthers, Lee, Oehm et Yu (2011) discutent des divers moyens d'industrialiser le processus de production de statistiques à l'Australian Bureau of Statistics.

La question est de savoir si des normes internationales amélioreraient la qualité des enquêtes en général. Certains domaines dans lesquels des normes seraient avantageuses comprennent le calcul des indicateurs de qualité utilisés fréquemment, tels que les taux d'erreur et les effets de plan, ainsi que les pratiques exemplaires pour la traduction du matériel d'enquête, le traitement des enquêtés ne parlant pas la langue du pays, et la pondération pour tenir compte de la non-réponse. Il ne faut pas oublier que, quand une norme est émise, elle doit être mise à jour continuellement et qu'il est bien connu qu'elles sont parfois difficiles à appliquer. Si les normes sont exhaustives, le praticien peut se sentir écrasé et, par conséquent, les ignorer en grande partie, à moins que leur application ne soit rendue obligatoire et vérifiée.

4.6 Modèles de processus opérationnels statistiques

Au cours des dernières années, des concepts tels que les modèles de processus opérationnels et l'architecture opérationnelle ont été intégrés par certains organismes statistiques dans les travaux concernant la qualité. Afin de rendre les processus de production plus efficaces et plus souples, on peut les percevoir comme faisant partie d'un modèle d'architecture opérationnelle (Reedman et Julien 2010). Dans le domaine de la production statistique, un modèle générique du processus de production statistique est élaboré conjointement par la CEE-ONU, Eurostat et l'OCDE. Tout remaniement de système doit être dicté par les demandes des clients, les évaluations des risques et les nouveaux développements. Les principes architecturaux qui sous-tendent cette école de pensée sont résumés dans Doherty (2010), qui discute du renouvellement de l'architecture à Statistique Canada.

Voici quelques-uns des principes :

- La prise de décisions doit être optimale à l'échelle de l'organisme, ce qui implique la centralisation de l'informatique, du soutien méthodologique et du traitement des données.
- L'utilisation de services intégrés, tels que la collecte, la saisie et la diffusion des données, doit être optimisée.
- La réutilisation doit être maximisée et prévoir le plus petit nombre possible de processus opérationnels distincts et le plus petit nombre possible de systèmes informatiques.
- La boîte à outils de l'organisme doit être réduite au minimum.
- Le personnel doit bien connaître les outils et les systèmes.
- Les reprises, telles que les vérifications répétées, doivent être éliminées.
- Les activités de base doivent être le point de concentration, et le travail lié au processus de soutien doit être externalisé.
- Les activités de développement doivent être séparées des opérations continues.
- La collecte électronique des données doit être considérée comme le mode initial.
- Les obstacles structuraux, tels que le chevauchement ou le manque de clarté des mandats, doivent être éliminés.

Ces principes sont semblables à ceux que nous avons dégagés lorsque nous avons appliqué les principes de gestion de la qualité tirés des divers cadres et modèles d'excellence décrits plus haut. Les principes constituent une évolution de la décentralisation vers un mode de pensée plus intégré. De nombreux organismes statistiques sont conscients que le cloisonnement est une chose du passé et que le passage à une plus grande centralisation est nécessaire.

5. Mesure de la qualité

Donc, la qualité est un concept multidimensionnel et sa mesure est une tâche compliquée. Nous avons noté que la qualité des enquêtes peut être considérée comme un concept tridimensionnel associé au produit final, aux processus sous-jacents qui mènent au produit et à l'organisation qui fournit les moyens d'exécuter les processus et de livrer le produit ou le service avec succès. Il existe essentiellement deux façons de mesurer la qualité. L'une est l'estimation directe de l'erreur d'enquête totale ou de certaines composantes de cette dernière. L'autre consiste à mesurer des indicateurs de la qualité dans l'espoir qu'ils reflètent effectivement le concept proprement dit.

5.1 Estimations directes de l'erreur d'enquête totale

Les décompositions existantes de l'erreur quadratique moyenne décrite, par exemple, dans Hansen et coll. (1964), Fellegi (1964), Anderson, Kasper et Frankel (1979), Biemer et Lyberg (2003), Weisberg (2005), et Groves et coll. (2009) sont toutes incomplètes en ce sens qu'elles ne tiennent pas compte de toutes les sources d'erreur. Il est rarement possible de calculer directement l'EQM dans les situations pratiques d'enquête, parce que ce calcul nécessite en général une estimation des paramètres qui est essentiellement exempte d'erreur. Toutefois, il est possible d'obtenir une deuxième meilleure estimation de la vraie valeur des paramètres si des ressources sont disponibles pour recueillir des données par une méthode considérée comme la « norme de référence », mais qui n'est ni de coût abordable ni pratique dans des conditions normales d'enquête. Il s'agit de la méthode classique d'évaluation lorsque la valeur vraie des paramètres peut être définie de manière unique. Les méthodes considérées comme la norme de référence sont rarement exemptes d'erreur, mais elles peuvent, à des degrés variables, fournir de meilleures estimations, et l'écart entre l'estimation ordinaire et l'estimation de référence peut servir d'estimation du biais, méthode qui est utilisée dans les enquêtes postcensitaires (Nations Unies 2010). Souvent, l'évaluation porte sur une composante particulière de l'erreur, telle que le sous-dénombrement au recensement, le biais de non-réponse, la variance due à l'intervieweur ou la simple variance de réponse, puisque nous voulons de l'information non pas sur l'erreur d'enquête totale proprement dite, mais plutôt sur la contribution relative de la composante à l'erreur totale d'enquête afin de pouvoir cerner les causes fondamentales des problèmes et d'améliorer les processus pertinents. Les grandes études d'évaluation sont très rares, parce qu'elles sont exigeantes et que leur valeur est parfois mise en doute (Nations Unies 2010). Par ailleurs, des études d'évaluation régulières à plus petite échelle sont nécessaires pour obtenir des indices quant aux problèmes opérationnels et méthodologiques.

5.2 Indicateurs de qualité

La communication continue de l'erreur d'enquête totale est une tâche gigantesque qu'aucun organisme d'enquête n'entreprend. Ils fournissent plutôt des indicateurs ou des déclarations concernant la qualité. Par exemple, selon le manuel de production des rapports sur la qualité d'Eurostat (2009a), il convient de mesurer les indicateurs suivants :

- coefficient de variation ;
- taux de surcouverture ;
- taux de rejets à la vérification ;
- taux de réponses totales ;
- taux de réponses partielles ;
- taux d'imputation ;
- nombre d'erreurs ;
- ampleur moyenne des révisions.

Le thème commun ici est que ces éléments sommaires des paradonnées sont des indicateurs qui peuvent être calculés sans effectuer d'études spéciales. Le jeu d'indicateurs qui peuvent être calculés directement d'après les données d'enquête est, par définition, assez limité et leur valeur est douteuse. Par exemple, inclure la surcouverture, mais non la sous-couverture, simplement parce que la première peut être calculée directement d'après les données disponibles, n'est pas logique. C'est la sous-couverture qui pose le problème de couverture le plus important dans les enquêtes. Le manuel prescrit certes que le producteur évalue le biais possible (tant son signe que sa grandeur), mais il ne décrit pas clairement comment cela doit se faire. Il est demandé au producteur d'inclure les résultats des évaluations et des contrôles de la qualité, si cette information existe aussi. Les mesures du niveau d'effort pour des processus tels que la conception des questionnaires et la formation des codeurs seraient les bienvenues. Il n'existe aucun format normalisé de présentation de cette information qualitative et quantitative. De toute façon, la liste d'indicateurs clés est très restreinte si on la compare à la liste complète des principales sources d'erreur, et il est difficile de voir comment ces indicateurs sont perçus par les utilisateurs et comment ils peuvent être utilisés par le producteur pour améliorer le processus.

Le producteur a besoin d'une liste plus complète d'indicateurs pour pouvoir mesurer ou évaluer divers niveaux de qualité pour s'assurer que la mise en œuvre du plan d'enquête est maîtrisée ou être capable de mettre sur pied un projet d'amélioration de la qualité. Le plan d'enquête initial doit être modifié ou adapté durant la mise en œuvre afin de contrôler les coûts et de maximiser la qualité. Biemer (2010) discute de quatre stratégies de réduction des coûts et des erreurs en temps réel, à savoir l'amélioration continue de la qualité (ACQ), la collecte adaptative (Groves et Heeringa 2006), les Six Sigma (Breyfogle 2003), ainsi que la conception et la mise en œuvre totalement adaptatives.

Pour appliquer la stratégie d'amélioration continue de la qualité, il faut déterminer quelles sont les variables clés du processus ainsi que les caractéristiques de ce dernier qui sont essentielles à la qualité. Pour chaque caractéristique essentielle à la qualité, il faut élaborer des mesures fiables, en temps réel, du coût et de la qualité. Les mesures sont surveillées en permanence durant le processus et des interventions ont lieu afin de s'assurer que les coûts et la qualité demeurent dans les limites acceptables. La stratégie de collecte adaptative a été conçue pour réduire le biais de non-réponse dans les entrevues en personne. Elle comprend

trois phases. Durant la phase expérimentale, quelques options de plan de collecte sont mises à l'essai (par exemple, en ce qui concerne le niveau des mesures incitatives). Durant la phase principale de collecte de données, l'option choisie durant la phase expérimentale est mise en œuvre et se poursuit jusqu'à ce que soit atteinte la limite de capacité. Durant la phase de suivi des cas de non-réponse, des méthodes spéciales sont mises en œuvre pour réduire le biais de non-réponse et pour contrôler les coûts de la collecte des données. Ces méthodes comprennent le scénario d'échantillonnage double de Hansen-Hurwitz, l'augmentation des mesures incitatives et l'utilisation d'intervieweurs plus expérimentés. De nouveau, les efforts se poursuivent jusqu'à ce qu'une réduction supplémentaire du biais de non-réponse ne soit plus rentable. Le modèle des Six Sigma est le modèle d'excellence opérationnelle le plus élaboré, puisqu'il s'appuie fortement sur des méthodes statistiques. Il contient un jeu important de techniques et d'outils qui peuvent être utilisés pour contrôler et améliorer les processus. La conception et la mise en œuvre entièrement adaptatives combinent les caractéristiques de contrôle de l'amélioration continue de la qualité, de la collecte adaptative et du modèle Six Sigma, afin de surveiller simultanément les multiples sources d'erreur. Biemer et Lyberg (2012) donnent plusieurs exemples de caractéristiques essentielles à la qualité et de mesures pour divers processus d'enquête. Par exemple, dans le cas du processus de mesure, les attributs qui sont des caractéristiques essentielles à la qualité pourraient inclure les aptitudes à repérer et à corriger les questions d'enquête qui posent problème, à déceler et à contrôler les erreurs de réponse, et à minimiser les biais et les variances dus aux intervieweurs. Les mesures correspondantes pourraient inclure le nombre d'éléments de données manquants par question, les taux de refus selon la taille de l'entreprise, les résultats des mesures répétées, les contrôles douteux effectivement modifiés, et les résultats du travail sur le terrain par intervieweur. Les mesures peuvent être analysées en utilisant les méthodes de contrôle statistique des processus ou d'analyse de la variance. Différentes mesures connexes peuvent être affichées simultanément sous forme d'un tableau de bord. Par exemple, si l'une des caractéristiques essentielles à la qualité est la capacité de découvrir les intervieweurs qui trichent, nous pourrions créer un tableau de bord montrant la durée moyenne d'interview par intervieweur et la distribution de certaines caractéristiques de l'échantillon de nature sensible, également par intervieweur.

5.3 Autoévaluations et vérifications

Les principes de gestion de la qualité ont mené à l'introduction des concepts d'autoévaluation et de vérification dans la production de statistiques. Nous souhaitons vivement savoir ce que les utilisateurs, les clients, les propriétaires et d'autres parties prenantes pensent des produits et services fournis par l'organisme statistique. Un certain nombre d'outils sont disponibles pour ce genre d'évaluation. Nous avons déjà mentionné le sondage sur la satisfaction des clients. Les autres outils comprennent les sondages auprès des employés, les vérifications internes et les vérifications externes. Les sondages auprès des clients peuvent jeter de la lumière sur ce que les utilisateurs pensent des produits et services qui leur sont fournis. Ils peuvent servir à déterminer les besoins des utilisateurs et à cerner les caractéristiques du produit qui importent vraiment. Une autre série de questions pourraient avoir trait à l'image de l'organisme et à sa comparaison à celle d'autres organismes, qu'il s'agisse ou non de concurrents. Les sondages sur la satisfaction des clients sont très fréquents dans notre société. Souvent, il est impossible de les utiliser pour faire des inférences au sujet de la population cible d'utilisateurs, à cause de défauts méthodologiques et conceptuels. L'abondance de sondages sur la satisfaction, développés et mis en œuvre par des personnes ne possédant aucune formation officielle en méthodes d'enquête, contribue à l'accueil tiède qui leur est réservé dans des situations plus sérieuses donnant lieu à des erreurs de non-réponse et de mesure. Ainsi, le sondage sur la satisfaction des utilisateurs mené en 2007 par Eurostat comprenait deux sondages distincts. L'un, lancé sur la page Web d'Eurostat, avait pour population cible 3 800 utilisateurs inscrits. Seuls les utilisateurs inscrits qui sont entrés dans le site Web durant la période de collecte des données ont été exposés à la demande de participation au sondage, ce qui a donné un taux de réponse d'environ 5 %. Le second sondage, réalisé par courrier électronique, a été envoyé à plusieurs utilisateurs importants identifiés par Eurostat. Cet environnement plus contrôlé a produit un taux de réponse de 28 %. Ces sondages posent aussi des problèmes pour ce qui est d'identifier le répondant le plus approprié. Le choix du « mauvais » répondant au sein d'un organisme aboutira certainement à des résultats non éclairés et trompeurs.

Le type le plus simple d'autoévaluation est le questionnaire ou la liste de vérification remplie par le gestionnaire de l'enquête. Un exemple est offert par Statistics New Zealand. Il s'agit d'une liste de vérification comprenant un certain nombre d'indicateurs ou de déclarations, tels que « les besoins d'information sont évalués régulièrement en consultant les utilisateurs », « documentation utile et accessible », « production et surveillance régulières d'indicateurs de l'exactitude » et « respect des normes de présentation ». Le gestionnaire doit répondre par oui ou par non à chaque question et faire un commentaire s'il le juge nécessaire. Statistics Sweden utilisait un système similaire dont l'une des questions était « Comparativement à l'année dernière, la qualité globale de votre produit s'est-elle améliorée, a-t-elle

diminué ou est-elle restée la même ? ». Lorsque les résultats ont été compilés pour ces trois catégories pour l'ensemble de l'organisme, il s'est avéré qu'une très faible proportion de gestionnaires avait déclaré une baisse de qualité, une proportion un peu plus élevée, une amélioration de la qualité, tandis qu'une vaste proportion avait déclaré qu'il n'y avait pas eu de changement. Les gestionnaires n'avaient tout simplement pas les moyens appropriés d'évaluer la qualité globale. En outre, des quantificateurs vagues, tels que « régulièrement », « utile » et « respect des normes », sont une invitation à fournir des évaluations généreuses. De surcroît, la plupart des gestionnaires ne veulent pas faire mauvaise impression et le *statu quo* devient l'échappatoire parfaite. Statistics Sweden a fini par abandonner ce système d'évaluation. Il est possible d'accroître la valeur de ces évaluations en posant des questions supplémentaires pour obtenir des renseignements détaillés sur la façon dont les travaux liés à la qualité ont été menés et à quel moment. Certains organismes recourent à des équipes internes qui vérifient les produits importants. Julien et Royce (2007) décrivent une vérification de la qualité de neuf produits menée par Statistique Canada afin de repérer les points faibles et leurs causes fondamentales, ainsi que pour dégager les pratiques exemplaires. Des équipes d'examen constituées de gestionnaires adjoints ont été créées afin que chaque examinateur passe en revue trois programmes différents. Le principal point faible d'une telle approche est l'élément interne proprement dit. Chaque examinateur sait que son tour d'être soumis à un examen viendra tôt ou tard, ce qui risque de le freiner. Le problème est également interne en ce sens que les utilisateurs ne sont pas explicitement présents durant le processus d'examen. Toutefois, dans son programme général de vérification de la gestion de la qualité des données, Statistique Canada insiste beaucoup sur son système de liaison avec les utilisateurs (Julien et Born 2006), qui est l'un des cinq systèmes formant le cadre d'assurance de la qualité de l'organisme, les autres étant la planification intégrée, les méthodes et les normes, la diffusion et la production de rapports sur les programmes.

Une autre variante de l'autoévaluation est celle où elle précède une vérification externe. Statistics Netherlands (1997) décrit comment le Service des méthodes statistiques a été évalué par son personnel. L'évaluation a produit une liste de points faibles et de points forts qui a ensuite été examinée par une équipe externe. Habituellement, une vérification externe s'appuie pour l'évaluation sur certaines références, telles qu'un ensemble de règles, une norme ou un code de bonnes pratiques. La vérification aboutit alors à un certain nombre de recommandations concernant l'organisme ou le produit ou service en question.

Récemment, Statistics Sweden a élaboré un système général d'évaluation de l'erreur d'enquête totale. Le ministère des Finances de la Suède souhaite que les résultats des évaluations de la qualité permettent de suivre les améliorations de la qualité au fil du temps. Comme il faut évaluer la qualité d'un grand nombre d'enquêtes, de registres administratifs et d'autres programmes de l'organisme, il est nécessaire de disposer de certains indicateurs qui peuvent servir de mesures indirectes remplaçant les mesures réelles de la qualité. Parallèlement, le processus d'évaluation doit être complet, la communication des résultats doit être simple et les résultats doivent être crédibles. Pour chacune des sources d'erreur – spécification, base de sondage, non-réponse, mesure, traitement des données, échantillonnage, modélisation/estimation et révision –, huit produits clés ont été notés chacun comme étant mauvais, passable, bon, très bon ou excellent en ce qui concerne cinq critères. Ceux-ci étaient la connaissance des risques, la communication avec les utilisateurs, le respect des normes et des pratiques exemplaires, l'expertise disponible, et les réalisations en regard des plan d'atténuation des risques et(ou) d'amélioration. Les lignes directrices de notation variaient selon le critère. Voici celles appliquées pour la connaissance des risques :

Exemple de lignes directrices de notation – Connaissance des risques

Mauvais	Passable	Bon	Très bon	Excellent
La documentation interne sur le programme ne mentionne pas la source d'erreur comme un facteur de risque possible pour l'exactitude du produit.	La documentation interne sur le programme mentionne la source d'erreur comme un facteur de risque possible pour la qualité des données.	Un certain effort a été fait pour évaluer l'effet possible de la source d'erreur sur la qualité des données.	Des études ont été menées pour estimer les composantes pertinentes du biais et de la variance associés à la source d'erreur et elles sont bien décrites.	Il existe un programme permanent de recherche en vue d'évaluer toutes les composantes pertinentes de l'EQM associées à la source d'erreur, et leur incidence sur l'analyse des données. Le programme est bien conçu et axé sur les éléments appropriés, et fournit l'information requise pour faire face aux risques dus à cette source d'erreur.
	Mais : Aucun effort n'a été fait ou très peu d'effort a été fait pour évaluer ces risques.	Mais : Les évaluations n'ont pris en considération que des mesures indirectes (par exemple, les taux d'erreur) de l'effet, sans aucune évaluation des composantes de l'EQM.	Mais : Les études n'ont pas exploré les conséquences des erreurs sur divers types d'analyse des données, y compris les analyses de sous-groupes et de tendances et les analyses multivariées.	

Le processus d'évaluation a débuté par une auto-évaluation de chacun des huit produits clés. Les rapports de ces autoévaluations et d'autres documents pertinents ont été étudiés par des examinateurs externes qui ont ensuite rencontré les responsables des produits et leurs employés pour discuter des processus de production. Ensuite, les examinateurs ont présenté des évaluations détaillées et ont attribué une note à chaque produit. La procédure a permis de cerner d'importants domaines à améliorer, non seulement pour chaque produit, mais aussi pour l'ensemble des produits. Ce premier cycle d'évaluation a indiqué que l'erreur de mesure posait problème pour presque tous les produits clés. Comme toute autre approche de mesure ou d'indication de l'erreur d'enquête totale, celle-ci ne reflète pas vraiment l'erreur quadratique moyenne totale. Elle nécessite une description approfondie des processus et des améliorations apportées, et elle dépend fortement des compétences et des connaissances des examinateurs externes. Cette étude est présentée dans Biemer, Trewin, Japec, Bergdahl et Pettersson (2012).

5.4 Profils de qualité

Dans le cas des enquêtes continues, il est possible d'élaborer des profils de qualité. Ce genre de document contient tout ce que l'on sait de la qualité d'une enquête continue ou d'un autre produit statistique assemblé au cours de plusieurs années. Les profils de qualité n'existent que pour quelques grandes enquêtes, qui sont toutes sauf une, réalisées aux États-Unis, à savoir la Current Population Survey (Brooks et Bailar 1978), la Survey of Income and Program Participation (Jabine, King et Petroni 1990 ; Kalton, Winglee et Jabine 1998), la Schools and Staffing Survey (Kalton, Winglee, Krawchuk et Levine 2000), et l'American Housing Survey (Chakrabarty et Torres 1996). Fait exception la British Household Panel Survey (Lynn 2003). Le principal problème que pose un profil de qualité est son manque d'actualité, puisqu'il s'agit d'une compilation des résultats d'études de la qualité qui prennent souvent beaucoup de temps. L'objectif du profil de qualité est de cerner les domaines où ils existent des lacunes dans les connaissances sur les erreurs, afin de pouvoir apporter des améliorations. Kasprzyk et Kalton (2001) ainsi que Doyle et Clark (2001) passent en revue l'utilisation des profils de qualité aux États-Unis.

6. Et maintenant, où allons-nous ?

Les notions de gestion de la qualité ont été influentes dans de nombreux organismes statistiques. Des concepts tels que le leadership, la culture de la qualité, la prévention des problèmes, la clientèle, la concurrence, l'évaluation des risques, la réflexion au sujet du processus, l'amélioration, l'excellence opérationnelle et l'architecture opérationnelle sont des sujets abordés de plus en plus fréquemment par les dirigeants des organismes d'enquête, par exemple, Trewin (2001), Pink (2010), Fellegi (1996), Brackstone (1999), deVries (1999), Groves (2011), et Bohata (2011). Le monde des enquêtes semble s'engager dans une direction où la production de statistiques devient rationnelle et rentable, mais l'évolution est lente. Certains organismes ont commencé à se servir d'un modèle de gestion de la qualité à des fins d'autoévaluation et d'orientation. Le modèle d'excellence de l'EFQM est celui qui est recommandé aux instituts nationaux de statistique qui font partie du Système statistique européen et deux d'entre eux, ceux de la République tchèque et de la Finlande, ont même posé leur candidature pour l'obtention du Prix national de l'EFQM de leur pays. Certaines entreprises de marketing sont certifiées d'après la norme ISO 9001 de management de la qualité et d'autres sont certifiées d'après la norme ISO 20252 concernant les études de marché et les études sociales et d'opinion. Ce développement devrait aboutir à des améliorations de la qualité, mais nous ne pourrons pas vraiment en être certains tant que nous ne commencerons pas à recueillir des données pertinentes. Cependant, une chose est sûre. Certains clients préfèrent les fournisseurs de services qui sont certifiés, qui ont gagné des prix ou qui peuvent donner la preuve que leur travail est conforme à un cadre ou à un modèle de qualité. Rares sont les clients qui jugeraient qu'un tel attribut est négatif.

Les marges d'erreur que nous associons aux estimations sont habituellement trop étroites, puisqu'elles n'englobent pas toutes les sources de variation. Les estimations ponctuelles peuvent se situer hors des limites à cause des biais. Idéalement, il serait utile de pouvoir produire des estimations de l'erreur d'enquête totale plutôt que celles produites aujourd'hui. Toutefois, ce genre de progrès n'est pas réaliste. Nous ne sommes pas en mesure de produire ce genre d'estimations, même à l'occasion, pour des raisons de budget, de temps et de méthodologie. Cela nous laisse les indicateurs de l'erreur d'enquête totale et de ses composantes. Ces indicateurs n'ont qu'une valeur limitée pour les utilisateurs, qui ne savent que faire des taux de réponse, de la variance de réponse mesurée par répétition des interviews ou des taux de rejets au contrôle. Par contre, ils sont très utiles pour les producteurs des données d'enquête. Par exemple, des études par répétition des interviews permettent de déceler la falsification et les questions d'enquête pour lesquelles la réponse manque de cohérence. La majorité des utilisateurs apprécient la crédibilité du fournisseur de services et cette crédibilité tient en partie à la capacité de présenter des données exactes. Un autre aspect important de la crédibilité est la volonté qu'ont les fournisseurs d'évaluer leur propre qualité et de communiquer les résultats de ces

évaluations. Même si ces dernières révèlent des problèmes, il est préférable que ce soit le fournisseur qui les découvre plutôt que des entités externes. La plupart des utilisateurs ne souhaitent pas participer aux discussions au sujet des erreurs et des compromis entre les divers types d'erreurs, et ce, pour de bonnes raisons. Elles sont simplement trop techniques et obscures. Si nous admettons qu'un processus de bonne qualité est une condition préalable à un produit de bonne qualité, nous devrions améliorer progressivement les processus afin qu'ils s'approchent de la situation parfaite d'absence de biais. De cette façon, la variance d'une estimation devient une bonne approximation de l'erreur quadratique moyenne.

Malgré des discussions sans fin et une myriade de projets d'amélioration de la qualité des enquêtes, les pratiques n'ont guère changé (Lynn 2004 ; Pink, Borowik et Lee 2010 ; Groves 2011 ; Bohata 2011). Le manque de compétence au sein des organismes d'enquête est peut-être l'une des causes profondes de cette lenteur du changement. La recherche sur les enquêtes doit faire appel à de nombreuses théories et méthodologies, dont la statistique, la technologie de l'information, la gestion, la communication et les sciences du comportement. Ces dernières sont nécessaires pour déterminer les causes fondamentales des erreurs non dues à l'échantillonnage. Si l'on se contente de quantifier ces erreurs, aucune amélioration n'est possible. À l'heure actuelle, les programmes de formation insistent sur les erreurs d'échantillonnage, de non-réponse et de couverture, et sur l'estimation en présence de ces erreurs. D'autres processus et sources d'erreur, tels que la production de mesures et le traitement des données, ne se voient pas accorder autant d'importance. D'où une situation où les études sur l'erreur de mesure et sur l'erreur de traitement des données sont rares comparativement à celles sur disons, la non-réponse. Tant dans le camp des producteurs que dans celui des utilisateurs, la confusion est importante en ce qui concerne les concepts et les méthodes. Une autre cause de lenteur pourrait être la règle du consensus appliquée dans certains organismes lorsqu'il s'agit de prendre des décisions concernant les changements. Cette règle repose sur le compromis. L'avis de nombreuses parties prenantes est recueilli et une décision est habituellement prise en se basant sur le plus petit commun dénominateur, ce qui n'est jamais une bonne norme. En outre, arriver à ce compromis demande habituellement beaucoup de temps et de ressources. Cette approche est très éloignée du modèle planifier-faire-contrôler-agir.

La qualité des enquêtes n'est pas une entité absolue. Le mode uniformisé de communication de l'information sur la qualité en vigueur à l'heure actuelle ne convient pas, puisque chaque utilisateur définit l'adéquation à l'usage prévu. Les décisions concernant des dimensions de la qualité telles que l'actualité, la comparabilité et l'accessibilité doivent être prises en collaboration avec les utilisateurs principaux, tandis que le fournisseur du service est responsable d'offrir la meilleure exactitude possible, compte tenu des diverses contraintes.

Les discussions sur la qualité des enquêtes et l'adoption de stratégies de gestion de la qualité ont-elles abouti à de meilleures données ? Nous ne le savons pas. La qualité des enquêtes n'a pas été évaluée selon un mode avant-après. La tendance est à l'accroissement de la normalisation et de la centralisation, ce qui devrait s'avérer rentable, mais quand il s'agit de la qualité des données, certains indicateurs pointent dans la mauvaise direction. Par exemple, dans de nombreux pays, les taux de non-réponse augmentent et les propriétés des erreurs dues à la collecte en mode mixte, à la traduction du matériel d'enquête et à d'autres caractéristiques de conception ne sont pas entièrement comprises ou varient d'une culture à l'autre. Il n'existe pas de formule de conception, ce qui entraîne des prises de décisions boiteuses concernant les compromis et des difficultés à décider de l'intensité avec laquelle les contrôles de qualité doivent être appliqués. La quête de pratiques exemplaires persiste dans les organismes d'enquête, mais leur mise en œuvre est difficile et éparpillée. Un rehaussement généralisé du niveau des compétences s'impose manifestement. Un programme de perfectionnement international structuré à l'intention des fournisseurs de services est nécessaire, de même qu'une collaboration internationale systématique en vue de déterminer les meilleurs moyens de concevoir et de mettre en œuvre les enquêtes. Nous devons mieux servir les utilisateurs en leur fournissant des données dont l'erreur est faible. Nous pouvons pour cela combiner plus judicieusement nos connaissances de la statistique et des phénomènes cognitifs avec les principes de gestion de la qualité. La note vraiment positive est l'attitude extraordinairement favorable à l'amélioration de la qualité dont témoignent les organismes statistiques partout dans le monde.

Bibliographie

Aitken, A., Hörngren, J., Jones, N., Lewis, D. et Zilhao, M. (2004). *Handbook on improving quality by analysis of process variables*. Office for National Statistics, Royaume-Uni.

Anderson, R., Kasper, J. et Frankel, F. (1979). *Total Survey Error: Applications to Improve Health Surveys*. San Francisco : Jossey-Bass.

Apted, L., Carruthers, P., Lee, G., Oehm, D. et Yu, F. (2011). Industrialisation of statistical processes, methods and technologies. Document présenté à la réunion de l'Institut International de Statistique, Dublin.

Bailar, B., et Dalenius, T. (1969). Estimating the response variance components of the U.S. Bureau of the Census' Survey Model. *Sankhyā*, B, 341-360.

Biemer, P. (2001). Commentaire sur Platek et Särndal. *Journal of Official Statistics*, 17(1), 25-32.

Biemer, P. (2010). Overview of design issues: Total survey error. Dans *Handbook of Survey Research*, (Éds., P. Mardsen et J. Wright), Deuxième édition. Emerald Group Publishing Limited.

Biemer, P., et Lyberg, L. (2003). *Introduction to Survey Quality*. New York : John Wiley & Sons, Inc.

Biemer, P., et Lyberg, L. (2012). Short course on Total Survey Error. The Joint Program in Survey Methodology (JPSM), 16 au 17 avril, Washington, DC.

Biemer, P., Trewin, D., Japec, L., Bergdahl, H. et Pettersson, Å. (2012). A tool for managing product quality. Document présenté à la conférence de Q2012, Athènes.

Blyth, B. (2012). ISO 20252; Turning frameworks into best practice. Document présenté à la conférence de Q2012, Athènes.

Bohata, M. (2011). Fit-for-purpose statistics for evidence based policy making. Note, Eurostat.

Bowley, A.L. (1913). Working-class households in reading. *Journal of the Royal Statistical Society*, 76(7), 672-701.

Box, G. (1990). Good quality costs less? How come? *Quality Engineering*, 3, 1, 85-90.

Box, G., et Friends (2006). *Improving Almost Anything: Ideas and Essays*. New York : John Wiley & Sons, Inc.

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

Brackstone, G. (2001). Quelle est l'importance de l'exactitude ? *Recueil : Symposium 2001, La Qualité des Données d'un Organisme Statistique : Une Perspective Méthodologique*, Statistique Canada.

Breyfogle, F. (2003). *Implementing Six Sigma*. Deuxième édition. New York : John Wiley & Sons, Inc.

Brooks, C., et Bailar, B. (1978). An error profile: Employment as measured by the Current Population Survey. Document de travail 3, Office of Management and Budget, Washington, DC.

Chakrabarty, R., et Torres, G. (1996). American Housing Survey: A Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.

Colledge, M., et March, M. (1993). Quality management: Development of a framework for a statistical agency. *Journal of Business and Economic Statistics*, 11, 157-165.

Colledge, M., et March, M. (1997). Quality policies, standards, guidelines, and recommended practices. Dans *Survey Measurement and Process Quality*, (Éds., L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York : John Wiley & Sons, Inc.

Couper, M. (1998). Measuring Survey Quality in a CASIC Environment. Document présenté au Joint Statistical Meetings, American Statistical Association, Dallas, TX.

Dalenius, T. (1967). Nonsampling Errors in Census and Sample Surveys. Rapport N° 5 du projet de recherche Errors in Surveys. Stockholm University.

Dalenius, T.E. (1968). Official statistics and their uses. *Revue de l'Institut International de Statistique*, 26(2), 121-140.

Dalenius, T. (1969). Designing descriptive sample surveys. Dans *New Developments in Survey Sampling*, (Éds., N.L. Johnson et H. Smith), New York : John Wiley & Sons, Inc.

Dalenius, T. (1985a). *Elements of Survey Sampling*. Swedish Agency for Research Cooperation with Developing Countries. Stockholm, Suède.

Dalenius, T. (1985b). Relevant official statistics. *Journal of Official Statistics*, 1(1), 21-33.

Deming, E. (1944). On errors in surveys. *American Sociological Review*, 9, 359-369.

Deming, E. (1950). *Some Theory of Sampling*. New York : John Wiley & Sons, Inc.

Deming, E. (1986). *Out of the Crisis*. MIT.

Deming, W.E., et Geoffrey, L. (1941). On sample inspection in the processing of census returns. *Journal of the American Statistical Association*, 36, 215, 351-360.

De Vries, W. (1999). Are we measuring up…? Questions on the performance of national systems. *Revue Internationale de Statistique*, 67, 1, 63-77.

Dillman, D. (1996). Why innovation is difficult in government surveys (avec discussions). *Journal of Official Statistics*, 12, 2, 113-198.

Doherty, K. (2010). How business architecture renewal is changing IT at Statistics Canada. Document présenté au Meeting on the Management of Statistical Information Systems. Daejeon, Corée du Sud, 26 au 29 avril.

Doyle, P., et Clark, C. (2001). Quality profiles and data users. Document présenté à l'International Conference on Quality in Official Statistics (Q), Stockholm.

Drucker, P. (1985). *Management*. Harper Colophone.

Ecochard, P., Hahn, M. et Junker, C. (2008). User satisfaction surveys in Eurostat and in the European Statistical System. Document présenté à la conférence Q, Rome, Italie.

Edwards, W., Lindman, H. et Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.

Eltinge, J. (2011). Aggregate and systemic components of risk in total survey error models. Document présenté au ITSEW 2011, Québec, Canada.

Ericson, W. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*, Séries B, 195-233.

European Foundation for Quality Management (1999). *The EFQM Excellence Model*. Van Haren.

Eurostat (2009a). ESS Standard for Quality Reports. Eurostat.

Eurostat (2009b). ESS handbook for Quality Reports. Eurostat.

Eurostat (2011a). European statistics Code of Practice. Eurostat.

Eurostat (2011b). Quality assurance framework (QAF). Eurostat.

Fellegi, I. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.

Fellegi, I. (1996). Characteristics of an effective statistical system. *Revue Internationale de Statistique*, 64, 2, 165-197.

Felme, S., Lyberg, L. et Olsson, L. (1976). *Kvalitetsskydd av data*. (Protecting Data Quality.) Liber (en suédois).

Fienberg, S., et Tanur, J. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *Revue Internationale de Statistique*, 64, 237-253.

Fisher, R. (1935). *The Design of Experiments*. New York : Hafner.

Frankel, M., et King, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 1, 65-87.

Gleaton, E. (2011). Centralizing LAN services. Note, National Agricultural Statistics Service, U.S. Department of Agriculture.

Groves, R. (1989). *Survey Errors and Survey Costs*. New York : John Wiley & Sons, Inc.

Groves, R. (2011). The structure and activities of the U.S. Federal Statistical System: History and recurrent challenges. *The Annals of the American Academy of Political and Social Science*, 631, 163, Sage.

Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls, W. et Waksberg, J. (Éds.) (1988). *Telephone Survey Methodology*. New York : John Wiley & Sons, Inc.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. et Tourangeau, R. (2009). *Survey Methodology*, Deuxième édition. New York : John Wiley & Sons, Inc.

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

Groves, R., et Lyberg, L. (2010). Total survey error: Past, present and future. *Public Opinion Quarterly*, 74, 5, 849-879.

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

Hansen, M., Hurwitz, W. et Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 32e Session, 38, Partie 2, 359-374.

Hansen, M., Hurwitz, W. et Madow, W. (1953). *Sample Survey Methods and Theory*. Volumes I et II. New York : John Wiley & Sons, Inc.

Hansen, M., Hurwitz, W., Marks, E. et Mauldin, P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46,147-190.

Hansen, M., Hurwitz, W. et Pritzker, L. (1964). The estimation and interpretation of gross differences and simple response variance. Dans *Contributions to Statistics*, (Éd., C. Rao). Oxford : Pergamon Press, 111-136.

Hansen, M., Hurwitz, W. et Pritzker, L. (1967). Standardization of procedures for the evaluation of data: Measurement errors and statistical standards in the Bureau of the Census. Document présenté à la réunion de l'Institut International de Statistique, Sydney.

Hansen, M., et Steinberg, J. (1956). Control of errors in surveys. *Biometrics*, 462-474.

Hansen, M., et Voigt, R. (1967). Program guidance through the evaluation of uses of official Statistics in the United States Bureau of the Census. Document présenté à la réunion de l'Institut International de Statistique, Sydney.

Holt, T., et Jones, T. (1998). Quality work and conflicting policy objectives. *Proceedings of the 84th DGINS Conference*, 28 au 29 mai, Stockholm, Suède. Eurostat.

Jabine, T., King, K. et Petroni, R. (1990). Survey of Income and Program Participation (SIPP): Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.

Joiner, B. (1994). *Generation Management*. McGraw-Hill.

Julien, C., et Born, A. (2006). Quality management assessment at Statistics Canada. *Proceedings of the Q Conference*, Cardiff, Royaume-Uni.

Julien, C., et Royce, D. (2007). Quality review of key indicators at Statistics Canada. *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*, 1113-1120.

Juran, J.M. (1988). *Juran on Planning for Quality*. New York : Free Press.

Juran, J.M. (1995). *A History of Managing for Quality*. ASQC Quality Press.

Juran, J., et Gryna, F. (Éds.) (1988). *Juran's Quality Control Handbook*, 4e édition. McGraw-Hill.

Kalton, G. (2001). Quelle est l'importance de l'exactitude ? *Recueil : Symposium 2001, La Qualité des Données d'un Organisme Statistique : Une Perspective Méthodologique*, Statistique Canada.

Kalton, G., Winglee, M. et Jabine, T. (1998). *SIPP Quality Profile*. U.S. Bureau of the Census, 3e édition.

Kalton, G., Winglee, M., Krawchuk, S. et Levine, D. (2000). *Quality Profile for SASS Rounds 1-3: 1987-1995*. Washington, DC : U.S. Department of Education.

Kasprzyk, D., et Kalton, G. (2001). Quality profiles in U.S. Statistical Agencies. *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm 14 au 15 mai 2001, CD-ROM.

Kennickell, A., Mulrow, E. et Scheuren, F. (2009). Paradata or process modeling for inference. Document présenté à la Conference on Modernization of Statistics Production, Stockholm, Suède.

Kiear, A.N. (1897). The representative method of statistical surveys. *Kristiania Videnskaps-selskabets Skrifter: Historik-filosofiske Klasse*, (en norvégien), 4, 37-56.

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

Kish, L. (1 995). *The Hundred Years' Wars of Survey Sampling*. Centennial Representative Sampling, Rome.

Kotz, S. (2005). Reflections on early history of official statistics and a modest proposal for global coordination. *Journal of Official Statistics*, 21, 2, 139-144.

Kreuter, F., Couper, M. et Lyberg, L. (2010). The use of paradata to monitor and manage survey data collecti on. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Lyberg, L. (1981). *Control of the Coding Operation in Statistical Investigations: Some Contributions*. Thèse de doctorat, Stockholm University.

Lyberg, L. (2 002). Training of survey statisticians in government agencies-A review. Communication sollicitée présentée à l a réunion des Joint Stati stical Meetings, American Statistical Association, New York.

Lyberg, L., Bergdahl, M., Blanc, M., Booleman, M., Grünewald, W., Haworth, M., Japec, L., Jones, L., Körner, T., Linden, H., Lundholm, G., Madaleno, M., Radermacher, W., Signore, M., Zilhao, M.J., Tzougas, I. et van Brakel, R. (2001). Summary report from the Leadership Group (LEG) on Quality. Eurostat.

Lyberg, L., et Couper, M. (2005). The use of paradata in survey research. Communication sollicitée présenté à l a réunion de l'Institut International de Statistique, Sydney.

Lynn, P. (Éd.) (2003). *Quality Profile: British Household Panel Survey: Waves 1 to 10: 1991-2000*. Colchester : Institute for Social and Economic Research.

Lynn, P. (2004). Editorial: Measuring and communicating survey quality. *Journal of the Royal Statistical Society*, Séries A, 167.

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

Mirotchie, M. (1993). Data quality: A quest for standard indicators. *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 729-734.

Moeller, R. (2005). *Brink's Modern Internal Auditing*. Sixième édition. New York : John Wiley & Sons, Inc.

Morganstein, D., et Marker, D. (1997). Continuous qua lity improvement in statistical agencies. Dans *Survey Measurement and Process Quality*, (Éds., L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York : John Wiley & Sons, Inc., 475-500.

Mudryk, W., Burgess, M.J. et Xiao, P. (1996). Qua lity control of CATI operations in Statistics Canada, Note, Statistique Canada.

Muscio, B. (1917). The i nfluence of the form of a que stion. *The British Journal of Psychology*, 8, 351-389.

Nations Unies (2010). *Post Enumeration Surveys: Operational Guidelines*. Department of Economic and Social Affairs, Statistics Division.

Neter, J., et Waksberg, J. (1964). A study of re sponse errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 305, 18-55.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Neyman, J. (1938). *Lectures and Conferences on Mathematical Statistics and Probability*. U.S. Department of Ag riculture, Washington, DC.

OCDE (2011). Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities. OCDE.

O'Muircheartaigh, C. (1997). Measurement errors in surveys: A historical perspective. Dans *Survey Measurement and Process Quality*, (Éds., L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York : John Wiley & Sons, Inc., 1-25.

Organisation internationale de normalisation (2006). Market, Opinion and Social Research. ISO Standard N° 20252.

Phipps, P., et Fricker, S. (2011). Quality measures. Note, Office of Survey Methods Research, U.S. Bureau of Labor Statistics.

Pink, B., Borowik, J. et Lee, G. (2010). The case for an international statistical innovation program-Transforming national and international statistics systems. Document présenté au Collaboration Leaders Workshop, 19 au 23 avril, Sydney, Australie.

Platek, R., et Särndal, C.-E. (2001). Can a statistician deliver? *Journal of Official Statistics*, 17, 1, 1-20 et la discussion, 21-27.

Reedman, L., et Julien, C. (2010). Current and future applications of the generic statistical business process model at Statistics Canada. Document présenté au Q Conference, Helsinki.

Rosén, B., et Elvers, E. (1999). Quality concept for official statistics. *Encyclopedia of Statistical Sciences*, New York : John Wiley & Sons, Inc., mise à jour, Volume 3, 621-629.

Scheuren, F. (2001). Quelle est l'importance de l'exactitude ? *Recueil : Symposium 2001, La Qualité des Données d'un Organisme Statistique : Une Perspective Méthodologique*, Statistique Canada.

Schilling, E., et Neubauer, D. (2009). *Acceptance Sampling in Quality Control*, 2e éd. Chapman and Hall/CRC.

Scholtes, P., Joiner, B. et Streibel, B. (1996). *The Team Handbook*. Joiner Associates Inc.

Shewhart, W.A. (1939). *Statistical Methods from the Viewpoint of Quality Control*. U.S. Department of Agriculture, Washington, DC, États-Unis.

Smith, T. (2011). Report on the International Workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. NORC/University of Chicago.

Spencer, B. (1985). Optimal data quality. *Journal of the American Statistical Association*, 80, 564-573.

Statistique Canada (2002). Le cadre d'assurance de la qualité de Statistique Canada, N° au catalogue 12-586-XIE, Ottawa.

Statistique Canada (2009). Statistique Canada : Lignes directrices concernant la qualité, cinquième édition, Ottawa.

Statistics Netherlands (1997). A self assessment of the Department of Statistical Methods. Document de recherche N° 9747, Statistics Netherlands.

Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.

Trewin, D. (2001). Importance d'une culture de la qualité. *Recueil : Symposium 2001, La Qualité des Données d'un Organisme Statistique : Une Perspective Méthodologique*, Statistique Canada.

U.S. Bureau of the Census (1974). *Standards for Discussion and Presentation of Errors in Data*. U.S. Department of Commerce, Bureau of the Census.

U.S. Federal Committee on Statistical Methodology (2001). *Measuring and Reporting Sources of Errors in Surveys*, Statistical Policy, document de travail 31, Washington, DC : U.S. Office of Management and Budget.

U.S. Office of Management and Budget (2002). Guidelines for ensuring, and maximizing the quality, objectivity, utility, and integrity of information disseminated by Federal agencies. Federal register, 67, 36, 22 février.

U.S. Office of Management and Budget (2006a). *Standards and Guidelines for Statistical Surveys*. U.S. Office for Management and Budget.

U.S. Office of Management and Budget (2006b). Questions and answers when designing surveys for information collection. U.S. Office for management and Budget.

Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 2, 119-137.

Weisberg, H. (2005). *The Total Survey Error Approach*. The University of Chicago Press.

Weisman, E., Balyozov, Z. et Venter, L. (2010). IMF's data quality assessment framework. Document présenté à la Conference on Data Quality for International Organizations, Helsinkli, 6 au 7 mai.

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

Willimack, D., Nichols, E. et Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. Dans *Survey Nonresponse*, (Éds., R. Groves, D. Dillman, J. Eltinge et R. Little), 213-228.

Zarkovich, S. (1966). *Quality of Statistical Data*. Food and Agricultural Organization of the United Nations : Rome, Italie.

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 143-155
Statistique Canada, N° 12-001-X au catalogue

Collecte de données : expérience et leçons apprises au chapitre des questions de nature délicate dans une région éloignée de culture de la coca au Pérou

Jaqueline Garcia-Yi et Ulrike Grote [1]

Résumé

La coca est une plante indigène de la forêt tropicale humide amazonienne, dont on extrait la cocaïne, un alcaloïde illégal. Les agriculteurs considèrent comme délicates les questions concernant la superficie de leurs aires de culture de la coca dans les régions éloignées où cette plante est cultivée au Pérou. Par conséquent, ils ont tendance à ne pas participer aux enquêtes, à ne pas répondre aux questions de nature délicate ou à sous-déclarer la superficie de leurs aires individuelles de culture de la coca. La mesure exacte et fiable des aires de culture de la coca est une source de préoccupations politiques et stratégiques, ce qui fait que les méthodologistes d'enquête doivent déterminer comment encourager la déclaration honnête de données et la réponse aux questions de nature délicate concernant la culture de la coca. Parmi les stratégies d'enquête appliquées dans notre étude de cas figuraient l'établissement d'un rapport de confiance avec les agriculteurs, l'assurance de la confidentialité, la correspondance entre les caractéristiques des intervieweurs et celles des répondants, la modification de la présentation des questions de nature délicate et l'absence d'isolement absolu des répondants au cours de l'enquête. Les résultats de l'enquête ont été validés au moyen de données recueillies par satellite. Ils semblent indiquer que les agriculteurs ont tendance à sous-déclarer la superficie de leurs aires de culture de la coca dans une proportion de 35 % à 40 %.

Mots clés : Coca ; cocaïne ; questions de nature délicate ; déclaration incorrecte ; non-réponse ; Pérou.

1. Introduction

Au cours des 30 dernières années, on a utilisé de plus en plus les enquêtes pour explorer les sujets délicats (Tourangeau et Yan 2007). Par exemple, on a utilisé des données d'enquête pour examiner les comportements « socialement indésirables », comme la prévalence de la consommation de drogue illicite (par exemple Botvin, Griffin, Diaz, Scheier, Williams et Epstein 2000 ; Fergusson, Boden et Horwood 2008), les avortements illégaux (par exemple Johnson-Hanks 2002 ; Varkey, Balakrishna, Prasad, Abraham et Joseph 2000) ou la consommation d'alcool chez les adolescents (par exemple Strunin 2001 ; Zufferey, Michaud, Jeannin, Berchtold, Chossis, van Melle et Suris 2007). De telles enquêtes ont été couramment utilisées dans les recherches universitaires et l'analyse des politiques (Davis, Thake et Vilhena 2009), même si les questions de nature délicate ont toujours été perçues comme problématiques. Les réponses sont considérées comme sujettes aux erreurs et aux biais, parce que les répondants sous-déclarent constamment les comportements socialement indésirables (Barnett 1998 ; Tourangeau et Yan 2007). Les faibles taux de réponse présentent une préoccupation supplémentaire. Ceux qui sont sélectionnés pour une enquête peuvent simplement refuser d'y participer, ou ils peuvent y participer, mais refuser de répondre aux questions de nature délicate (Tourangeau et Yan 2007).

Des enquêtes récentes au niveau des ménages ont intégré des questions de nature délicate concernant la superficie des aires de culture de la coca (voir, par exemple, Ibanez et Carlsson 2010). La coca est un arbuste indigène de la forêt tropicale humide amazonienne en Amérique du Sud. On extrait la cocaïne de ses feuilles. La superficie des aires de culture de la coca représente 40 % en Colombie, 40 % au Pérou et 20 % en Bolivie de la superficie totale des aires de culture de la coca à l'échelle mondiale, soit 154 100 hectares (ONUDC 2011). Au Pérou et en Bolivie, les feuilles de cette plante sont utilisées traditionnellement à de nombreuses fins, depuis environ 3000 ans avant Jésus-Christ (Rivera, Aufderheide, Cartmell, Torres et Langsjoen 2005) jusqu'à nos jours. Ces utilisations traditionnelles comprennent principalement la mastication de la feuille de coca et l'absorption de tisane de feuilles de coca pour surmonter la fatigue, la faim et la soif, ainsi que pour soulager les symptômes du « mal de l'altitude » et les maux d'estomac, respectivement (Rospigliosi 2004). Depuis les années 1970, toutefois, la culture de la coca a atteint des sommets, en raison de son utilisation comme matière première pour la production de cocaïne (Caulkins, Reuter, Iguchi et Chiesa 2005). Le contenu en cocaïne de la feuille de coca est inférieur à 1 % et va de 0,13 % à 0,86 % (Holmstedt, Jaatmaa, Leander et Plowman 1977). Par conséquent, les trafiquants de narcotiques ont besoin de grandes quantités de feuilles de coca pour obtenir suffisamment d'alcaloïde pour sa commercialisation sur le marché illégal. En général, la culture de la coca pour le narcotrafic est une activité rentable. En fait, le revenu des agriculteurs qui cultivent de la coca est supérieur de 54 % à celui de ceux qui n'en cultivent pas (Davalos, Bejarano et Correa 2008).

1. Jaqueline Garcia-Yi, présidente de l'Agricultural and Food Economics Technical University of Munich Weihenstephaner Steig 22, 85350, Freising, Allemagne. Courriel : jaqueline.garcia-yi@tum.de ; Ulrike Grote, Professeur, Institute for Environmental Economics and World Trade, Leibniz University Hannover, Königsworther Platz 1, 30167 Hannover, Allemagne. Courriel : grote@iuw.uni-hannover.de.

Par conséquent, la recherche portant sur la coca est maintenant axée sur l'évaluation de la rentabilité de cette culture, par rapport à d'autres cultures commerciales (voir, par exemple, Gibson et Godoy 1993 ; Torrico, Pohlan et Janssens 2005). Différentes tentatives ont été faites pour remplacer la coca par d'autres cultures, mais il a été déterminé de façon générale que le remplacement de culture comme politique antidrogue a été un échec (ONUDC 2001). Les décideurs et les chercheurs ont reconnu qu'il existe des déterminants socioéconomiques pertinents autres que la rentabilité économique pour justifier la culture de la coca. Il s'agit notamment du capital social (Thoumi 2003), ainsi que des fonctions d'épargne et de réserve financière pour les dépenses importantes (Bedoya 2003 ; Mansfield 2006). Pour vérifier ces dernières hypothèses sur la superficie des aires de culture de la coca, on a besoin de bases de données exhaustives incluant des données sur les ménages.

La culture de la coca n'est pas illégale à proprement parler au Pérou (au cours des années 1990, le gouvernement péruvien avait comme principal objectif de « pacifier » le pays en contrôlant les groupes terroristes. Il a mis en œuvre ce que l'on appelle actuellement la « doctrine Fujimori ». L'hypothèse qui sous-tend cette doctrine est que la culture de la coca n'est pas de nature criminelle, mais plutôt attribuable à la pauvreté. Par conséquent, la doctrine Fujimori a fait en sorte de décriminaliser la culture de la coca, ce qui a diminué le besoin de protection des agriculteurs contre les associations terroristes et ce qui a par conséquent facilité la tâche du gouvernement dans sa lutte contre ces groupes violents (Obando 2006). Cela rend compte en partie de l'acceptation sociale des utilisations traditionnelles de la coca dans ce pays (ONUDC 2001). Ainsi, le cadre juridique actuel semble faciliter le narcotrafic parce que la coca utilisée pour le commerce illégal peut être cultivée sous prétexte de son utilisation à des fins traditionnelles (OICS 2009 ; Durand 2005). Par conséquent, Garcia et Antezana (2009) sont d'avis que certains agriculteurs vendent de la coca à des personnes qui semblent la négocier à des fins traditionnelles, mais qui sont plutôt des narcotrafiquants qui transforment les feuilles de coca à différents endroits, comme des petites villes à la frontière de la Bolivie.

Même si la culture de la coca n'est pas illégale, les régions où elle se pratique et qui sont perçues comme fournissant les narcotrafiquants (c'est-à-dire les régions ayant de grands champs de coca) peuvent être ciblées par le gouvernement pour la mise en œuvre de programmes d'éradication forcée (Obando 2006). L'éradication est susceptible d'entraîner de grandes pertes économiques pour les cultivateurs de coca, selon la superficie totale de leurs aires de culture individuelles. Ainsi, certains des agriculteurs peuvent être réticents à fournir des données sur le fait qu'ils cultivent ou non de la coca. On devrait aussi s'attendre à ce que certains des agriculteurs qui admettent cultiver de la coca ne déclarent pas la superficie totale de leurs aires de culture, parce qu'ils craignent que les grands champs de coca fassent l'objet d'une éradication.

Étant donné que la mesure exacte et fiable de la superficie des aires de culture de la coca suscite des préoccupations politiques et stratégiques, il est nécessaire pour les méthodologistes d'enquête de déterminer comment encourager la déclaration honnête de données et la réponse à des questions de nature délicate concernant la culture de la coca. Le présent article suggère et évalue un certain nombre de stratégies, en vue d'augmenter à la fois la déclaration et la fiabilité des données au niveau des ménages dans une région éloignée de culture de la coca au Pérou.

Même si le sujet du présent article est particulièrement lié à la culture de la coca, les leçons apprises concernant la conception de l'enquête et sa mise en œuvre pourraient servir de référence dans le cadre d'autres sujets de nature délicate, comme les questions liées à la santé (par exemple les mesures anticonceptionnelles et le comportement sexuel) ou des comportements indésirables (par exemple la consommation de drogue illégale) dans d'autres régions de différents pays.

L'article est structuré de la façon suivante : la section 2 décrit la collectivité du Pérou à l'étude, les stratégies particulières pour réduire la non-réponse et la déclaration incorrecte, ainsi que les leçons apprises de la collecte de données au moyen de questions de nature délicate dans la zone visée par la recherche. La section 3 présente les résultats de l'enquête liée à la culture de la coca et leur validation, tandis que la section 4 est constituée d'un sommaire des principaux résultats, suivis par la conclusion.

2. Collecte de données dans une collectivité cultivant la coca d'une région rurale du Pérou

La présente section décrit la collectivité cultivant la coca, ainsi que les principales stratégies de collecte des données appliquées dans notre étude, de même que les leçons apprises.

2.1 Description de la zone visée par la recherche

La zone visée par la recherche est située dans la vallée du cours supérieur du Tambopata, à la frontière avec la Bolivie, l'une des zones les plus éloignées et difficiles d'accès de la forêt tropicale humide amazonienne au Pérou (Bureau du Pérou de l'ONUDC 1999). Cette vallée est située dans le corridor de la conservation de la biodiversité de Vilcabamba-Amboro, à proximité des zones protégées nationales (voir la figure 1). L'ensemble de la population de la vallée du Tambopata est constituée d'immigrants, et plus particulièrement de descendants de la population indigène Aymara. Il s'agit d'un groupe ethnique autochtone originaire des régions des Andes et de l'Altiplano de l'Amérique du Sud. Au cours des années 1950, la plupart des agriculteurs étaient des immigrants saisonniers qui quittaient leurs terres de subsistance de l'Altiplano pendant trois à six mois par année et parcouraient les 320 km les séparant de la vallée du cours supérieur du Tambopata pour cultiver du

café sur des terres agricoles leur appartenant (Collins 1984). Au fil du temps, la plupart des agriculteurs sont devenus des colons permanents de la vallée et ont cultivé le café comme principale culture commerciale (*ibid*).

Avant 1989, la culture de la coca dans la vallée du cours supérieur du Tambopata était très peu répandue. La production de la coca à petite échelle se limitait à la consommation personnelle ou aux marchés locaux pour des usages traditionnels, comme la mastication de la coca par les agriculteurs et les mineurs des Andes. Après 1989, la culture de la coca s'est intensifiée, principalement dans la vallée avoisinante du cours supérieur de l'Inambari. Le changement ne semble pas avoir été le résultat d'une augmentation de la demande locale ou de la demande externe pour des usages traditionnels (Bureau du Pérou de l'ONUDC 1999). La coca provenant de ces vallées est considérée comme étant de mauvaise qualité en raison de son goût amer et est moins en demande pour la mastication traditionnelle que la coca de la région de Cuzco (Caballero, Dietz, Taboada et Anduaga 1998). Ces hausses ont par conséquent été liées à une augmentation de la demande liée au narcotrafic. Ces dernières années, des augmentations importantes de la culture de la coca dans la vallée du cours supérieur du Tambopata ont constamment été signalées par les Nations Unies, comme il est observé dans le tableau 1. La variation en pourcentage par année dans la vallée du cours supérieur du Tambopata est supérieure à la variation annuelle d'environ 4 % au niveau national.

Tableau 1
Culture de la coca dans la vallée du cours supérieur du Tambopata (2005-2008)*

Année	Hectares	Pourcentage de variation par rapport à l'année précédente
2005	253	-
2006	377	49,0
2007	863	128,9
2008	940	8,9

* Depuis 2009, les régions de culture de la coca de la vallée du cours supérieur du Tambopata sont agrégées avec celles de la vallée de l'Inambari dans les rapports de l'ONUDC. Par conséquent, il n'est pas possible d'estimer le pourcentage de variation par rapport à l'année précédente uniquement pour la vallée du Tambopata, ces dernières années.

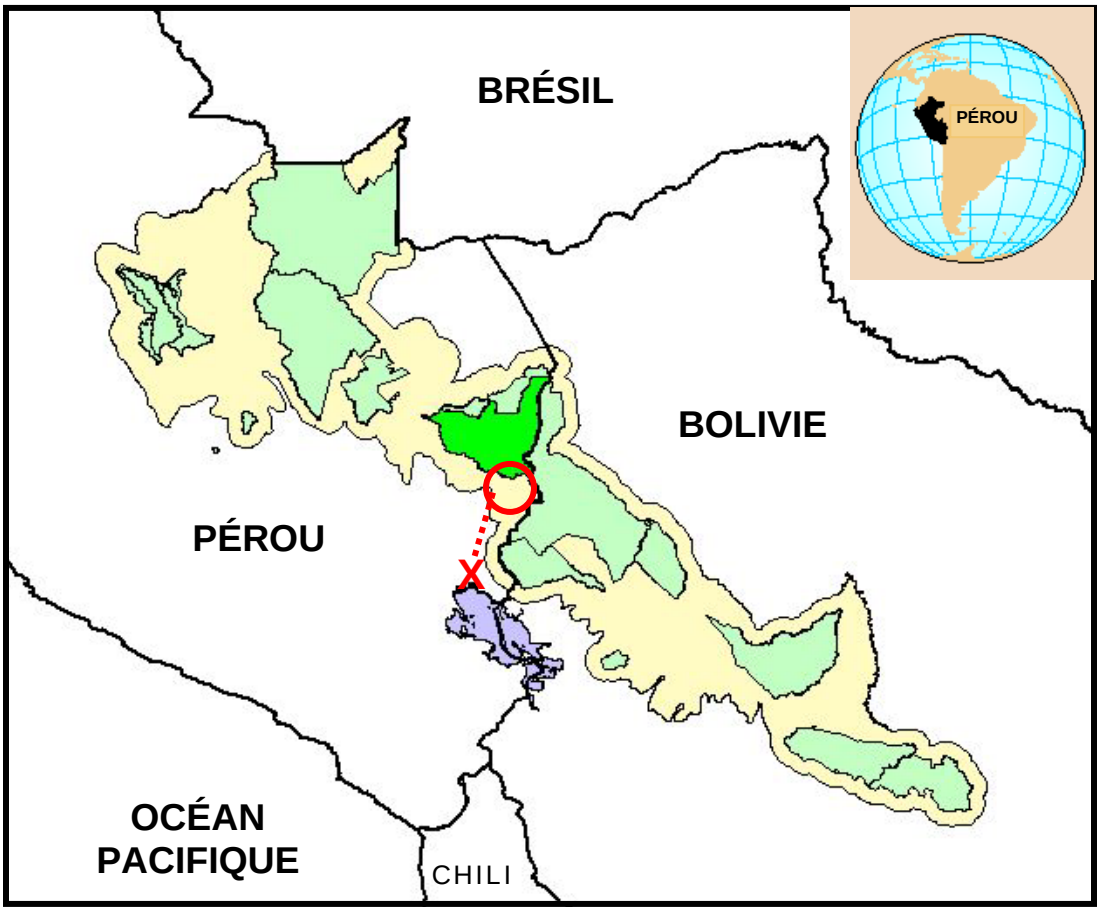
Source : Calcul de l'auteur à partir des données de l'ONUDC (2009).

La coca produite dans la vallée du cours supérieur du Tambopata et dans la vallée du cours supérieur de l'Inambari semble servir principalement à alimenter les associations de commerce transfrontalier entre les narcotrafiquants péruviens et boliviens. La Bolivie demeure le troisième producteur mondial en importance de cocaïne et est une zone de transit importante pour la cocaïne d'origine péruvienne (Département d'État des États-Unis 2009). Ces vallées constituent une région stratégique de production de la coca pour les narcotrafiquants, en raison de leur proximité avec une voie d'acheminement vers l'extérieur (Bureau du Pérou de l'ONUDC 1999). Les feuilles de coca ne sont pas toujours transformées en cocaïne sur les terres agricoles. Les narcotrafiquants semblent tirer parti des quantités importantes de feuilles de coca transportées vers les régions urbaines, apparemment pour les marchés d'utilisation traditionnelle, qu'ils achètent et transforment dans des installations cachées des régions urbaines près de la frontière de la Bolivie. Ainsi, le risque de se faire prendre par les autorités est réduit. À partir de la Bolivie, la cocaïne est acheminée vers le Brésil et l'Europe (Garcia et Antezana 2009).

La culture de la coca n'entraîne pas nécessairement une amélioration de la qualité de vie des agriculteurs de l'Amérique du Sud (Davalos et coll. 2008). Selon le dernier recensement de la population, les conditions de vie à San Pedro de Putina Punco (SPPP), le district situé au cœur de la vallée du cours supérieur du Tambopata, sont difficiles : 72 % des habitations sont des constructions en béton d'argile, 88 % ont un sol en terre battue, 16 % sont raccordées au réseau public d'électricité, 12 % sont raccordées au réseau public d'aqueduc et seulement 9 % ont accès aux égouts publics (INEI 2007). Cette situation est répandue dans les principales zones de culture de la coca du Pérou, dont 70 % des habitants continuent de vivre dans la pauvreté et 42 %, dans une situation de pauvreté extrême (Commission des stupéfiants 2005).

2.2 Stratégies de collecte des données et leçons apprises

En décembre 2007, on a mené une étude de faisabilité, afin de vérifier si les agriculteurs répondraient aux questions touchant la cocaïne. L'étude pilote pour la conception du questionnaire s'est tenue en mai 2008, et l'enquête finale a été menée entre juin et août 2008. L'étude de faisabilité, l'étude pilote et l'enquête finale étaient axées sur les agriculteurs vivant à San Pedro de Putina Punco (SPPP), un district de la vallée du cours supérieur de Tambopata, qui est situé au plus profond de la forêt tropicale humide. Tous les agriculteurs de la zone visée par la recherche produisent du café comme culture commerciale et certains complètent leur revenu par la culture de la coca. On compte cinq coopératives de producteurs de café à SPPP. Les agriculteurs doivent être membres de l'une de ces coopératives, pour pouvoir vendre leur café, en raison des restrictions touchant les intermédiaires. L'enquête finale a été menée uniquement auprès des membres de quatre de ces coopératives parce que la plupart des membres de la coopérative restante se trouvent à San Juan del Oro, un district à l'extérieur de la zone visée par la recherche.



Source : Auteur

Description de la carte :

- X Région de l'Altiplano
- O Vallée du cours supérieur du Tambopata
- – – Voie d'immigration
- Parc national Bahuaja Sonene
- Autres zones protégées
- Corridor de protection de la biodiversité Vilcabamba-Amboro
- Lac Titicaca

Figure 1 Carte de la zone visée par la recherche

L'enquête finale comprenait un questionnaire structuré axé sur la production agricole et le capital social. Le questionnaire comptait 15 sections :

1. Renseignements généraux au sujet de l'agriculteur et du ménage
2. Renseignements généraux au sujet des terres agricoles et de la superficie réservée au café
3. Activités économiques additionnelles
4. Données sur la certification biologique
5. Capital social cognitif et identité
6. Information et communication
7. Aspirations personnelles et attitudes face au risque
8. Capital social structurel
9. Chocs covariés et idiosyncrasiques
10. Capital humain
11. Réseaux sociaux
12. Traditions d'utilisation de la coca
13. Coûts détaillés de la production agricole
14. Accès à la main-d'œuvre
15. Questions additionnelles

Les éléments de l'enquête liés aux questions de nature délicate sont présentés à l'annexe 1.

La question sur la superficie de culture de la coca est une question de nature délicate pour les agriculteurs. Ceux qui cultivent des superficies importantes de coca craignent que les données fournies soient accessibles aux autorités responsables des programmes d'éradication. Ainsi, ils s'inquiètent parfois des conséquences possibles d'une réponse honnête, pour le cas où les données viendraient à la connaissance d'un tiers. Dans ces cas, l'anonymat doit être garanti aux agriculteurs. Ils peuvent aussi être tentés de fournir des réponses socialement désirables aux intervieweurs. La coca

est devenue un symbole important de la lutte de la population indigène pour l'autodétermination (Office of Technology Assessment 1993). Coca « oui », cocaïne « non » est le slogan des populations indigènes (Henman 1990), la formulation visant à faire une distinction claire entre les usages traditionnels (« coca ») et le narcotrafic (« cocaïne »). Ainsi, les usages traditionnels comme la mastication de la coca sont des symboles d'ethnicité (Allen 1981) et leur persistance pourrait être liée à un sentiment de nationalisme au Pérou (Henman 1990). En ce sens, on pourrait s'attendre à ce que les agriculteurs n'aient pas beaucoup de problèmes à indiquer qu'ils cultivent de la coca, à condition de pouvoir l'associer à des usages traditionnels. Par ailleurs, en raison de l'association entre les zones de production plus importantes et les activités illégales, il se peut que les producteurs de coca ne déclarent pas entièrement la superficie totale de leurs zones de production, afin de donner l'impression qu'ils la cultivent pour des usages traditionnels seulement.

Plusieurs stratégies peuvent contribuer à réduire les biais possibles liés à la nature délicate de la question, à la non-réponse partielle et à la non-réponse d'unités, ainsi qu'à la déclaration incorrecte délibérée. Ces stratégies comprennent les suivantes : assurance de la confidentialité ; sélection soigneuse du mode de collecte des données et du libellé de la question de nature délicate ; adaptation des caractéristiques et du comportement des intervieweurs (voir Coutts et Jann 2008 ; Tourangeau et Yan 2007). De plus amples renseignements sur la mise en œuvre de ces stratégies dans notre étude de cas figurent ci-après.

Établissement d'une relation de confiance et assurance de l'anonymat

Les agriculteurs des régions de culture de la coca ont tendance à ne pas faire confiance aux gens de l'extérieur. Dans cette région particulière, nous avons déterminé qu'ils font confiance aux directeurs des coopératives de café. Un des directeurs des coopératives de café a signé une lettre de présentation autorisant notre recherche sur la culture agricole. On a montré la lettre aux agriculteurs avant la tenue de l'enquête. Un essai pilote mené avec et sans la lettre de présentation a démontré que la lettre était importante pour réduire les refus de participation à l'enquête. Dans l'introduction à l'enquête, l'intervieweur a aussi indiqué que le directeur de la coopérative autorisait l'enquête parce qu'il s'attendait à ce que les résultats profitent aux membres. En outre, on a clairement dit aux agriculteurs, au début de l'enquête, que les données recueillies demeureraient confidentielles, et on a souligné l'objectif académique de recherche du questionnaire (voir l'annexe 1a). Cette assurance d'anonymat était courte et précise, afin de réduire les soupçons des agriculteurs, comme l'ont indiqué Singer, Hippler et Schwarz (1992). La culture de la coca a été traitée comme un comportement courant et ordinaire dans la région visée par la recherche, et une assurance de confidentialité longue et élaborée aurait pu susciter des réserves chez les agriculteurs, plutôt que d'alléger leurs soupçons. Un bref rappel de l'assurance de confidentialité a été inclus au milieu du questionnaire, avant les questions liées aux utilisations traditionnelles de la coca et avant la question de nature délicate sur la superficie consacrée à cette culture. Ce rappel se lisait comme suit : « Dans cette partie de l'enquête, nous vous poserons des questions concernant les usages et la culture de la coca. Veuillez vous rappeler que cette enquête est anonyme et qu'il n'y a pas de réponses correctes ou incorrectes » (voir l'annexe 1b). Cela fait suite à Willis (2005), qui mentionne qu'il est important de présenter des questions d'entrée en matière et une annonce du passage au sujet délicat, afin de réduire les réticences des répondants.

Mode de collecte des données

On a initialement envisagé d'avoir recours à des questionnaires administrés avec papier et crayon pour la collecte des données, afin de réduire le biais lié aux intervieweurs. Toutefois, au cours de l'étude de faisabilité, il est devenu évident que de nombreux agriculteurs, même ceux qui avaient dépassé le niveau primaire d'études (52 % de la population ; INEI 2007) n'étaient pas capables de lire sans effort. Les agriculteurs travaillent dans leurs champs presque toute la journée et n'ont pas beaucoup d'occasions de pratiquer leur habileté en lecture. De même, les interviews audio auto-administrées assistées par ordinateur (IAAAO), la méthode de prédilection pour la collecte de données sur des sujets délicats dans les pays développés (Mensch, Hewett et Erulkar 2003), dépassaient les limites de ce projet, en raison du manque d'équipement et d'alimentation électrique, ainsi que de connaissances en informatique dans la zone visée par la recherche. Il est probable que l'utilisation d'ordinateurs aurait augmenté l'anxiété et les soupçons concernant l'enquête, comme l'ont décrit Mensch et coll. (2003) pour l'Afrique. Par conséquent, les interviews sur place sont le mode de collecte qui a été sélectionné, et l'accent a été mis sur la sélection des intervieweurs, leur formation et leur comportement.

Sélection, formation et comportement des intervieweurs

Un des problèmes liés à la sélection des intervieweurs est l'absence de professionnels suffisamment scolarisés dans la zone visée par la recherche. Ainsi, un groupe de dix étudiants de l'université publique la plus proche, se trouvant à 16 heures de la zone visée par la recherche, ont été choisis comme intervieweurs. Tous les intervieweurs étaient de descendance Aymara ou Quechua ; on a tenté ainsi de faire correspondre partiellement les caractéristiques des intervieweurs et celles des répondants. On croyait que cela augmenterait la probabilité de participation, parce que l'appariement pouvait faire augmenter la confiance et la sympathie entre l'intervieweur et le répondant (Tourangeau et Yan 2007). Les intervieweurs se sont présentés comme des étudiants de l'université locale, et aucun renseignement additionnel n'a été fourni concernant une université ou une organisation à l'extérieur du pays finançant l'étude, afin

d'éviter les malentendus possibles et d'augmenter la confiance parmi les répondants. Au cours de l'étude pilote, certains agriculteurs ont indiqué des préoccupations concernant les programmes d'éradication de la coca financés à l'extérieur et, par conséquent, les références aux institutions externes ont été réduites au minimum. C'est donc dire que seuls des renseignements partiels ont été fournis aux répondants. Cela n'est pas courant, mais dans les circonstances particulières de l'étude, il n'y avait pas d'autre solution possible pour éviter les problèmes de sécurité potentiels.

En ce qui a trait à la formation, les intervieweurs ont d'abord participé à un atelier de deux jours à Puno, suivi par un atelier de trois jours dans la zone visée par la recherche. Le même groupe d'intervieweurs a aussi mené l'étude pilote, afin de tester les questions et le questionnaire, avec comme objectif de déterminer les problèmes de compréhension, de mémorisation, de jugement et d'acceptabilité de l'enquête, ainsi que de permettre la reformulation, l'élimination ou l'ajout de questions. L'étude pilote a aussi permis d'évaluer le rendement des intervieweurs et, dans certains cas, de déterminer les domaines nécessitant une formation sur mesure, selon la rétroaction concernant le rendement. Par exemple, au début, un des intervieweurs hésitait à poser les questions liées à la coca et cet intervieweur a obtenu un nombre plus élevé que la moyenne de non-réponses à la question de nature délicate. Après une formation adaptée, l'intervieweur a pu modifier son approche d'interview.

Présentation de la question de nature délicate

La présentation de la question présupposait le caractère délicat du comportement à l'étude, comme l'ont montré Tourangeau et Yan (2007). Par conséquent, on n'a pas demandé en premier aux agriculteurs s'ils avaient des terres consacrées à la culture de la coca, puis la superficie totale de leurs aires de culture de la coca. On a plutôt demandé directement aux agriculteurs d'indiquer la superficie totale de leurs aires de culture de la coca (« Quelle est la superficie de votre aire de culture de la coca en mètres ou en hectares ? »). Toutefois, il a été déterminé pendant l'étude pilote que les agriculteurs n'étaient pas à l'aise avec cette question et la sautaient ou se retiraient simplement de l'enquête. Par conséquent, le format de la question a été changé et on a utilisé un libellé indulgent à la place. On a posé aux agriculteurs la question suivante : « Combien de « petits arbustes de coca » avez-vous sur votre terre agricole ? » Ainsi, l'agriculteur pouvait répondre : « Seulement quelques-uns, j'ai… arbustes de coca ». Même si la différence est à peine perceptible, dans le premier cas, il était beaucoup plus difficile pour les agriculteurs de commencer leur réponse par « Seulement quelques-uns… ». Ainsi, grâce à la dernière question, il a été plus facile pour les agriculteurs d'ajouter des explications d'excuse à leur réponse, ce qui les a rendus plus détendus. Cette dernière présentation de la question de nature délicate avait aussi comme avantage d'utiliser de la terminologie que connaissent bien les Aymaras, qui utilisent couramment des diminutifs dans leurs conversations de tous les jours. Par ailleurs, le format de cette question pouvait impliquer indirectement que l'intervieweur s'attendait à ce que le répondant ait un petit nombre d'arbustes de coca, ce qui aurait probablement donné lieu à une sous-déclaration. Par conséquent, même si des non-réponses ont été évitées grâce à cette dernière question, on s'attendait à une certaine sous-déclaration.

Période de tenue de l'enquête et contexte de collecte des données

Les parcelles agricoles des agriculteurs sont disséminées dans la forêt tropicale humide montagneuse amazonienne au Pérou. Il était difficile de joindre les agriculteurs sur leurs terres pendant l'enquête. Par conséquent, pour mener l'enquête, nous avons principalement profité de la célébration de la fête de Saint-Pierre et des réunions de l'assemblée générale des coopératives, en juin et août 2008, respectivement, qui sont une occasion de rencontre des agriculteurs sur la place du village. La participation aux réunions de l'assemblée générale est obligatoire pour tous les membres des coopératives, ce qui fait que les répondants ciblés étaient présents à ces activités. La seule façon d'entrer sur la place du village ou d'en sortir est par une route non pavée. Afin de profiter de cette situation, l'enquête a été menée à partir d'une grande tente érigée sur la route non pavée ces journées-là. La tente comprenait dix cubicules, un pour chaque paire d'intervieweurs et de répondants. On n'a pas assuré une protection absolue de la vie privée, parce que pendant l'étude pilote, on a déterminé que les agriculteurs n'étaient pas à l'aise d'être les « seuls » à être interviewés ; ils préféraient en voir d'autres être interviewés en même temps qu'eux. Toutefois, les agriculteurs ne pouvaient pas entendre les réponses des autres. Comme tous les agriculteurs devaient utiliser la même route non pavée pour se rendre sur la place du village, peu importe leur provenance géographique, les biais géographiques potentiels qui, quant à eux, peuvent être liés à des variables importantes, comme la taille de la ferme et le revenu, ont probablement été réduits dans cette recherche.

Représentativité de l'échantillon

Une méthode d'échantillonnage de commodité a été utilisée, mais à la fin de l'enquête, nous avons demandé aux agriculteurs leur numéro d'enregistrement à la coopérative et nous avons utilisé les listes d'enregistrement pour inférer la représentativité de l'échantillon. Le numéro d'enregistrement à la coopérative fourni par les agriculteurs était inscrit sur une feuille de papier distincte et n'était pas joint au questionnaire du répondant. Les répondants ont été informés de cette procédure et ont pu en être témoins.

Les quatre coopératives visées par l'étude comptent 3 265 membres à SPPP. Le tableau 2 montre le nombre de répondants par coopérative. Le nombre de questionnaires remplis se chiffrait à 508. Au total, 12 répondants ont été exclus de l'échantillon parce que leur numéro

d'enregistrement à la coopérative était manquant. Dans deux cas, les agriculteurs ont refusé de fournir ces renseignements et, dans dix cas, les intervieweurs ont oublié de demander aux répondants leur numéro d'enregistrement à la fin de l'interview. Par conséquent, l'absence de ces renseignements a davantage été associée à une erreur de l'intervieweur qu'au refus de l'agriculteur de fournir ces renseignements.

Tableau 2
Nombre de répondants par coopérative

	Nombre total de membres des coopératives à SPPP	Taille de l'échantillon de l'enquête	Pourcentage des membres des coopératives interviewés (%)
Coopérative 1	756	106	14
Coopérative 2	911	138	15
Coopérative 3	887	138	16
Coopérative 4	711	114	16
Total	3 265	496	15

Source : Enquête de l'auteur.

Afin de vérifier la représentativité de l'échantillon, la répartition des numéros d'enregistrement aux coopératives obtenus auprès de l'échantillon de l'enquête a été comparée à la répartition des numéros d'enregistrement aux coopératives d'un échantillon aléatoire simple simulé, sans remise, tiré des listes des coopératives. Les listes des coopératives ont été classées par numéro d'enregistrement des membres, et les numéros d'enregistrement ont été associés à la date d'enregistrement des membres. Ainsi, la plupart des agriculteurs plus âgés ont des numéros d'enregistrement plus bas, et les agriculteurs plus jeunes, des numéros plus élevés. Malheureusement, les coopératives n'avaient pas d'autres renseignements sur les membres, comme la superficie totale des terres, ou encore les hectares consacrés au café ou à la coca, pouvant servir à sélectionner un échantillon aléatoire stratifié. Deux types de tests ont été utilisés pour la comparaison des échantillons : un test de la somme des rangs pour deux échantillons de Wilcoxon (Mann-Whitney) et un test pour l'égalité des fonctions de distribution pour deux échantillons de Kolmogorov-Smirnov. Le premier test sert à déterminer dans quelle mesure il est probable que les deux groupes proviennent de la même distribution, et repose sur le principe que les différences observées sont causées par une fluctuation du hasard. Le deuxième test est similaire au premier, mais il est aussi sensible aux différences dans l'emplacement et la forme des fonctions de distribution cumulative empirique des deux groupes. Les résultats des deux tests n'ont pas rejeté l'hypothèse nulle de l'égalité de la distribution entre l'échantillon de l'enquête et l'échantillon aléatoire simple simulé, au niveau de signification de 0,05. Ainsi, les résultats montrent que l'échantillon de l'enquête est équivalent à un échantillon aléatoire simple et, par conséquent, est représentatif de la population à l'étude.

3. Résultats de l'enquête et problèmes de validation

3.1 Résultats de l'enquête

Le taux de réponse à l'enquête se situe à environ 90 %, ce qui est bien au-dessus du taux de réponse minimum recommandé de 60 % (Punch 2003). Dans les 496 questionnaires remplis, 19 répondants (moins de 4 %) n'ont pas répondu aux questions liées à la coca. Lorsque l'on compare les statistiques descriptives des variables socioéconomiques, institutionnelles et liées à la coca, on note certaines différences significatives entre toutes les observations (sans les non-répondants) et les « non-répondants à la question de nature délicate » (voir l'annexe 2). Les non-répondants à la question de nature délicate étaient tous de sexe masculin, avec un pourcentage plus grand de descendance ethnique Aymara, et un plus grand nombre d'enfants. En outre, un pourcentage plus élevé d'entre eux utilisait la coca à des fins médicinales. Il est intéressant de constater qu'un nombre significativement plus élevé de non-répondants sont très réfractaires au risque (73,7 %), comparativement à tous les autres répondants (28,6 %). Cela pourrait indiquer une crainte possible de la part des « non-répondants à la question de nature délicate » que les intervieweurs divulguent les renseignements à des tiers. Le contexte du test d'aversion au risque suivi par Binswanger (1980) est présenté à l'annexe 1c.

Des statistiques descriptives comparatives de base des producteurs de coca et des non-producteurs sont présentées dans le tableau 3. Le nombre de questionnaires valides était de 477, si nous ne tenons pas compte des non-répondants à la question de nature délicate. Parmi eux, 64 % ont indiqué qu'ils étaient des producteurs de coca.

Il n'y a pas de différences statistiquement significatives en ce qui a trait aux caractéristiques socioéconomiques générales (âge, sexe, groupe ethnique, et nombre d'enfants) entre les producteurs de coca et les non-producteurs. La seule différence a été observée au chapitre de la scolarité. Les non-producteurs comptent plus d'années de scolarité que les producteurs. Chez les producteurs de coca, on retrouve une moins grande superficie de forêt au total et une moins grande superficie de forêt primaire, et davantage de terres en jachère que chez les non-producteurs, même si ces différences ne sont pas statistiquement significatives. Chez les producteurs de coca et les non-producteurs, on retrouve des superficies consacrées à la production de café et d'aliments comparables. Par contre, les producteurs de coca et les non-producteurs affichaient des différences statistiquement significatives dans les variables du capital social. Un plus grand nombre de non-producteurs que de producteurs trouvent important de respecter les lois nationales. Par ailleurs, un moins grand nombre de non-producteurs que de producteurs ont vu leur confiance à l'égard de leur voisin diminuer au cours des cinq dernières années et ont collaboré à des activités communautaires au cours de la dernière année.

Il existe un rapport statistiquement significatif entre la culture de la coca et les usages traditionnels. Un pourcentage plus élevé de producteurs de coca que de non-producteurs mastiquent la coca et l'utilisent comme médicament. Qui plus est, un plus grand nombre de producteurs de coca trouvent plus facile de vendre les feuilles de coca que les non-producteurs, dans le cas hypothétique où ils cultiveraient la coca à des fins commerciales.

Enfin, il est important de mentionner que le nombre moyen d'arbustes de coca est relativement faible, ce qui pourrait être attribuable à une sous-déclaration des zones de culture commerciale de la coca ou à la culture de la coca pour consommation propre seulement, ou les deux. Il n'est pas possible de faire de distinction entre ces deux scénarios, ce qui fait qu'il est plus facile pour les producteurs commerciaux de coca de se faire passer pour des producteurs de coca pour des usages traditionnels.

3.2 Problèmes de validation

On ne peut pas vérifier directement la validité des réponses individuelles parce qu'il existe peu de recherches empiriques antérieures sur ce sujet et pas d'autres sources pour confirmer les données. Toutefois, il est possible de fournir une comparaison brute entre les données de l'enquête et la superficie totale de production de coca comptabilisée par les organismes internationaux pour la vallée du cours supérieur du Tambopata, à partir de données recueillies par satellite. L'Office des Nations Unies contre la drogue et le crime (ONUDC 2009) indique que 940 hectares de coca étaient cultivés dans la vallée du cours supérieur du Tambopata en 2008. La densité habituelle de culture de la coca dans le cas des régions de producteurs de coca pour des usages traditionnels pourrait se situer entre 35 000 et 40 000 arbustes par hectare (ONUDC 2001) (au cours des années 1990, la densité de culture de la coca était plus faible, soit entre 20 000 et 25 000 arbustes par hectare (ONUDC 2009)). La densité de culture de la coca dans cette vallée particulière est relativement faible parce que les producteurs la partagent avec celle du café et des aliments de base, même si les rendements par arbuste ont augmenté ces dernières années (ONUDC 2009). Par conséquent, on s'attend à ce que le nombre total d'arbustes de coca dans cette vallée se situe entre 32,9 et 37,6 millions.

Tableau 3
Statistiques descriptives comparatives entre les producteurs de coca et les non-producteurs

Variable	Producteurs de coca	Non-producteurs de coca
Âge	42,5	41,7
	(12,7)	(12,5)
Hommes (%)	93,9	94,9
Aymara (%)	81,4	82,5
Nombre d'enfants	3,0	2,9
	(2,0)	(2,1)
Années de scolarité	8,2*	8,7*
	(3,3)	(3,3)
Superficie totale (hectares)	7,9	8,0
	(8,4)	(7,8)
Superficie consacrée au café (hectares)	2,2	2,2
	(2,0)	(1,4)
Superficie de forêt secondaire (zone en jachère)	1,6	1,4
	(2,4)	(2,1)
Superficie de forêt primaire (hectares)	3,9	4,2
	(7,5)	(7,0)
Superficie consacrées aux aliments de base (hectares)	0,5	0,5
	(0,7)	(0,6)
Aucune autre activité économique (%)	46,8	48,9
Aversion élevée au risque (%)	30,5	25,3
Importance de respecter les lois nationales (%)	81,9**	88,6**
Diminution de la confiance au cours des cinq dernières années (%)	19,3**	12,5**
Participation à des activités communautaires en 2007 (%)	92,0**	84,7**
Agriculteurs mastiquant la coca (%)	76,0***	53,1***
Agriculteurs utilisant la coca comme médicament (%)	81,7***	54,8***
Perception qu'il est facile de vendre les feuilles de coca (%)	26,4**	18,5**
Nombre d'arbustes de coca	3 093	-
	(6 710)	
Nombre d'observations	305	172

Les écarts-types sont entre parenthèses pour les variables continues.
Les moyennes pour les producteurs de coca et les non-producteurs de coca sont statistiquement différentes (test T avec variances inégales) au :
* niveau de signification de 0,1 ; ** niveau de signification de 0,05 ; *** niveau de signification de 0,01.
Source : Calculs de l'auteur.

Notre échantillon de 477 répondants (en excluant les agriculteurs qui n'ont pas indiqué leur numéro d'enregistrement à la coopérative et les non-répondants à la question de nature délicate) ont déclaré au total 960 000 arbustes de coca. Cet échantillon représente 14,6 % du total des 3 265 membres de coopératives à SPPP. Ainsi, si l'on extrapole en fonction du nombre total de membres de coopératives situées dans le district de SPPP, on obtient un total de 6,6 millions d'arbustes de coca. Par ailleurs, nous devons tenir compte du fait que la vallée du cours supérieur du Tambopata comprend aussi le district de San Juan del Oro, qui compte à peu près la même population que le district de SPPP (INEI 2007). Selon l'hypothèse très solide que les agriculteurs de SPPP se comportent de la même façon que les agriculteurs de San Juan del Oro, à tout le moins du point de vue de la culture de la coca, cela ferait doubler le nombre d'arbustes de coca pour l'ensemble de la vallée du cours supérieur du Tambopata, ce nombre atteignant environ 13,2 millions. Cette dernière estimation se situe entre 35 % et 40 % des 32,9 à 37,6 millions obtenus à partir des données recueillies par satellite de l'ONUDC. Elle se situe dans la fourchette attendue de déclaration pour les questions de nature délicate. Dans le cas des questions sur l'avortement, cette fourchette se situe entre 35 % et 59 % (Fu, Darroch, Henshaw et Kolb 1998), et pour l'utilisation des opiacées ou de la cocaïne, entre 30 % et 70 % (Tourangeau et Yan 2007).

4. Sommaire et conclusions

La coca, qui est la matière première pour la production de la cocaïne, est cultivée en Colombie, au Pérou et en Bolivie. Dans ces deux derniers pays, les usages traditionnels de la coca par les populations indigènes remontent à environ 3000 ans avant Jésus-Christ (Rivera et coll. 2005). Néanmoins, les questions aux agriculteurs sur l'étendue de leur culture de coca sont considérées comme délicates. Les producteurs de coca craignent les programmes d'éradication, même s'ils ne vendent pas de coca pour le narcotrafic, parce qu'il est difficile de faire une distinction entre les producteurs de coca à des fins commerciales, et ceux qui produisent pour leur propre consommation. Ainsi, les agriculteurs ont tendance à ne pas participer aux enquêtes, à ne pas répondre aux questions de nature délicate ou à sous-déclarer la superficie de leurs aires de culture de la coca, afin de réduire les possibilités d'identification, en vue d'une éradication possible.

Dans ce contexte, les procédures de collecte des données au niveau des ménages doivent être évaluées, ainsi que les stratégies pour réduire la non-réponse et la déclaration incorrecte. La plupart des stratégies utilisées dans notre zone de recherche au Pérou ont été fondées sur les pratiques exemplaires comprises dans des ouvrages publiés. Parmi les stratégies qui ont fonctionné dans notre cas figuraient l'établissement d'un lien de confiance avec les agriculteurs au moyen d'une lettre de présentation d'un directeur d'une coopérative de café, l'assurance de la confidentialité dès le départ et au milieu du questionnaire, la correspondance entre les caractéristiques ethniques des intervieweurs et des répondants, la formation des intervieweurs, afin de réduire leur réticence à poser des questions de nature délicate, la modification de la présentation de la question de nature délicate, en vue d'adopter un libellé familier et indulgent, et le non-respect de la protection absolue de la vie privée pour éviter que chaque agriculteur croit qu'il est le seul à être interviewé.

La validité des réponses individuelles des agriculteurs concernant l'étendue de leur culture de coca ne peut être vérifiée, parce que peu de recherches empiriques ont été effectuées à ce sujet, et qu'il n'y a pas d'autres sources au niveau des ménages pour confirmer ces données. Ainsi, la portée de la déclaration incorrecte a été évaluée à partir de données agrégées. Les résultats laissent supposer que les agriculteurs n'ont déclaré que de 35 % à 40 % de leur superficie réelle de culture de la coca. Toutefois, ces valeurs se situent à l'intérieur des fourchettes attendues pour les réponses aux questions de nature délicate. Du point de vue de la non-réponse à l'enquête et de la non-réponse à la question de nature délicate, les résultats étaient plus encourageants, indiquant des valeurs de 10 % et d'environ 4 %, respectivement.

Au moment de la tenue de l'enquête, nous avons principalement profité des célébrations et des assemblées générales des coopératives pour lesquelles les agriculteurs se réunissent dans le village, ceux-ci étant autrement très disséminés dans la forêt tropicale humide. L'enquête a suivi une méthode d'échantillonnage de commodité, mais il a été possible de vérifier la représentativité de l'échantillon parce que tous les agriculteurs sont enregistrés dans l'une des coopératives de la zone visée par la recherche. L'échantillon obtenu a été comparé à un échantillon aléatoire simple simulé sans remise, dans lequel chaque agriculteur avait la même probabilité d'être sélectionné au hasard à partir des listes de membres des coopératives. Il n'y avait pas de différences significatives dans les fonctions de distribution, ce qui fait que l'échantillon est équivalent à un échantillon aléatoire simple. Le principal inconvénient de cette approche est, qu'après l'interview, nous avons dû demander aux répondants leur numéro de membre de coopérative. Même si on a dit au répondant que ce numéro n'était pas joint à leur questionnaire, certains agriculteurs peuvent avoir eu des doutes à ce sujet et cela pourrait avoir eu des effets sur la crédibilité de l'assurance de la confidentialité dans les interviews suivantes, les agriculteurs s'étant passé le mot.

Par ailleurs, la comparaison des caractéristiques des non-répondants aux questions de nature délicate et du reste des non-répondants montre que les non-répondants sont très réfractaires au risque. Même si le nombre de non-répondants était faible (moins de 4 % de l'échantillon total), cela pourrait laisser supposer que la principale raison de la non-réponse partielle est la crainte des conséquences de la transmission des données à des tiers.

Les superficies consacrées à la culture de la coca déclarées par les agriculteurs étaient en moyenne très petites. Cela pourrait être une tentative de la part des producteurs commerciaux de coca de faire semblant de cultiver uniquement pour leur propre consommation. La culture de la coca pour des usages traditionnels n'a pas de connotation négative à proprement parler, étant donné qu'il s'agit d'un symbole de l'ethnicité et de la lutte de la population indigène pour l'autodétermination (Office of Technology Assessment 1993). Il n'est pas possible de faire de distinction entre les agriculteurs qui ont sous-déclaré l'étendue de leurs terres consacrées à la culture de la coca et ceux qui cultivent la coca pour leur propre consommation. Malheureusement, les producteurs commerciaux de coca peuvent profiter de cette situation et continuer de cultiver de la coca en prétextant des usages traditionnels.

Remerciements

La recherche a été financée par le BMZ (Ministère fédéral de la coopération économique et du développement, Allemagne) et le DAAD (Service d'échanges universitaires de l'Allemagne), ainsi que par le LA CEEP (Programme latino-américain et caribéen d'économie environnementale).

Annexe 1

Sections pertinentes du questionnaire

A) *Présentation :*

Bonjour/bonsoir. Je m'appelle _____ et j'étudie à _____ _____. Nous menons une enquête pour déterminer les risques et les vulnérabilités auxquels sont exposés les producteurs de café dans votre collectivité. Les directeurs des coopératives de café connaissent cette enquête et croient que les résultats pour raient profiter à la collectivité. Si vous décidez de répondre à notre questionnaire, vous pouvez sauter des questions ou vous retirer de l'étude à tout moment. Les données recueillies dans le cadre de cette enquête demeureront CONFIDENTIELLES et serviront uniquement à des fins UNIVERSITAIRES. Vos réponses et vos opinions son extrêmement importantes pour la coopérative et pour nous. Êtes-vous prêt à répondre à certaines questions ?

a) *Oui* (poursuivre)
b) *Non* (remercier le répondant, interrompre l'enquête et indiquer les caractéristiques de la personne dans la présentation 1)

B) *Questions relatives à la coca :*

Dans cette partie de l'enquête, nous vous poserons des questions concernant les usages et la culture de la coca. Veuillez vous rappeler que cette enquête est anonyme et qu'il n'y a pas de réponses correctes ou incorrectes.

Mastiquez-vous des feuilles de coca ?	a) Oui	b) Non
Utilisez-vous les feuilles de coca à des fins médicinales ?	a) Oui	b) Non
Vous croyez-vous obligé d'offrir des feuilles de coca à vos invités pendant les activités d'ayni et de minka ?	a) Oui	b) Non
Utilisez-vous les feuilles de coca pour des rituels ?	a) Oui	b) Non
Utilisez-vous des feuilles de coca pour le paiement de travailleurs de l'extérieur ?	a) Oui	b) Non
Utilisez-vous les feuilles de coca comme produit d'échange ou comme cadeau pour des amis ou parents ?	a) Oui	b) Non
Combien de petits arbustes de coca y a-t-il sur votre parcelle agricole ?	_____	

C) *Question sur l'aversion au risque :*

Ceci est un jeu. Avant de jouer, vous devez choisir l'une des options affichées ci-dessous. Puis, vous devez tirer à pile ou face. Si, par exemple, vous avez choisi l'option H et que je tire à pile ou face et que la pièce tombe sur face, vous ne gagnez pas d'argent du tout. Toutefois, si le résultat est pile, vous gagnez 200 soles. Par ailleurs, si vous avez choisi l'option A, vous recevrez 50 soles, peu importe si le résultat est pile ou face. Laquelle des options parmi celles qui précèdent choisissez-vous avant que je lance la pièce ?

OPTION	Si le résultat est face, vous gagnez :	Si le résultat est pile, vous gagnez :
A	50 soles	50 soles
B	45 soles	95 soles
C	40 soles	120 soles
D	35 soles	125 soles
E	30 soles	150 soles
F	20 soles	160 soles
G	10 soles	190 soles
H	0 sol	200 soles

Annexe 2

Statistiques descriptives comparatives entre toutes les observations et les non-répondants à la question de nature délicate

Variables	Toutes les observations [a]	Non-répondants à la question de nature délicate
Âge	42,2 (12,6)	45,9 (9,9)
Hommes (%)	94,3***	100***
Aymara (%)	81,8**	94,7**
Nombre d'enfants	3,0** (2,0)	4,1** (2,0)
Années de scolarité	8,4 (3,3)	7,5 (2,9)
Superficie totale (hectares)	7,9 (8,3)	6,8 (3,2)
Superficie consacrée au café (hectares)	2,2 (1,8)	2,5 (1,2)
Superficie de forêt secondaire (zone en jachère)	1,6 (2,3)	1,4 (1,1)
Superficie de forêt primaire (hectares)	4,0 (7,3)	2,9 (3,3)
Superficie consacrée aux aliments de base (hectares)	0,5 (0,7)	0,6 (0,6)
Aucune autre activité économique (%)	47,5	57,9
Aversion élevée au risque (%)	28,6***	73,7***
Importance de respecter les lois nationales (%)	84,3	89,5
Diminution de la confiance au cours des cinq dernières années (%)	16,8	26,3
Participation à des activités communautaires en 2007 (%)	89,4	89,5
Agriculteurs mastiquant la coca (%)	67,7	73,7
Agriculteurs utilisant la coca comme médicament (%)	72,0*	84,2*
Perception qu'il est facile de vendre les feuilles de coca (%)	23,6	27,8
Nombre d'observations	477	19

Les écarts-types sont entre parenthèses pour les variables continues.
a) Toutes les observations sans les non-répondants à la question de nature délicate.
Les moyennes pour les non-répondants sont statistiquement différentes de l'ensemble de l'échantillon (test T avec variances inégales) au :
* niveau de signification de 0,1 ; ** niveau de signification de 0,05 ; *** niveau de signification de 0,01.
Source : Calculs de l'auteur.

Bibliographie

Allen, C. (1981). To be Quechua: The symbolism of coca chewing in highland Peru. *American Ethnologist*, 8, 1, 157-171.

Barnett, J. (1 998). Sensitive questions and response effects: An evaluation. *Journal of Managerial Psychology*, 13, 1/2, 63-67.

Bedoya, E. (20 03). Estrategias productivas y el riesgo entre los cocaleros del val le de los ríos apurimac y ene. Dans *Amazonía: Procesos Demográficos y Am bientales*, (Éds., C. Aramburu et E. Bedoya), Consorcio de Investigación Económica y S ocial. Lima, Pérou.

Binswanger, H. (1980). Attitude towards risk: Experimental measurement in rural India. *American Journal of Economics*, 62, 395-407.

Botvin, G., Gri ffin, K., Diaz, T., Scheie r, L., Williams, C. et Epstein, J. (2000). Preventing illicit drug use in adolescents: Long-term follow-up data from a ra ndomized control trial of a sch ool population. *Addictive Behaviors*, 25, 5, 769-774.

Bureau du Pérou de l' ONUDC (1999). Desarrollo Alternativo del lnambari y Tambopata. Documento de Proyecto AD/PER/99/D96. Disponible au : http://www.onudd.org.pe/web/Html/Templates/proyectos.htm (accessible le 15 juin 2009).

Caballero, V., Dietz, E., Taboada, C. et Anduaga, J. (1998). Diagnostico Rural Participativo de las Cuencas Alto Inambari y Alto Tambopata Provincia de Sandia, Departamento de Puno. GTZ. Lima, Pérou.

Caulkins, J., Reuter, P., Iguchi, M. et Chiesa, J. (2005). How goes the War on Drugs? An Assessment of U.S. Drug Problems and Policy. RAND Drug Policy Research Center. États-Unis.

Collins, J. (1984). The maintenance of peasant coffee production in a peruvian valley. *American Ethnologist*, 11, 3, 413-438.

Commission des stupéfiants (2005). Alternative Development: A Global Thematic Evaluation. Rapport de synthèse final. Quarante-huitième session E/CN.7/2005/CRP.3. Autriche.

Coutts, E., et Jann, B. (2008). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). ETH Zurich Sociology, Document de travail, 3.

Davalos, L., Bejarano, A. et Correa, L. (2008). Disabusing cocaine: pervasive myths and enduring realities of a globalised commodity. *International Journal of Drug Policy*, 20, 5, 381-386.

Davis, C., Thake, J. et Vilhena, N. (2009). Social Desirability Biases in Self-Reported Alcohol Consumption and Harms. Addictive Behaviors. Article sous presse.

Durand, F. (2005). El Problema Cocalero y el Comercio Informal para Uso Tradicional. Debate Agrario 39. Lima, Pérou.

Département d'État des États-Unis (2009). International Narcotics Control Strategy Report. Volume I: Drug and Chemical Control. Bureau for International Narcotics and Law Enforcement Affairs. États-Unis.

Fergusson, D., Boden, J. et Horwood, L. (2008). The developmental antecedents of illicit drug use: Evidence from a 25-Year longitudinal study. *Drug and Alcohol Dependence*, 96, 165-177.

Fu, H., Darroch, J., Henshaw, S. et Kolb, E. (1998). Measuring the extent of abortion underreporting in the 1995 National Survey of Family Growth. *Family Planning Perspectives*, 30, 3, 128-138.

Garcia, J., et Antezana, J. (2009). Diagnostico de la Situación del Desvío de IQ al Narcotráfico. ConsultAndes and DEVIDA. Lima, Pérou.

Gibson, B., et Godoy, R. (1993). Alternatives to coca production in Bolivia: A computable general equilibrium approach. *World Development*, 21, 6, 1007-1021.

Henman, A. (1990). Tradicion y represion: Dos experiencias en america del sur. Dans *Coca, Cocaina y Narcotrafico. Laberinto en los Andes*, (Éds., Garcia – D. Sayan), Comision Andina de Juristas. Lima, Pérou.

Holmstedt, B., Jaatmaa, E., Leander, K. et Plowman, T. (1977). Determination of cocaine in some South American species of erythroxylum using mass fragmentography. *Phytochemistry*, 16, 1753-1755.

Ibanez, M., et Carlsson, F. (2010). A survey-based choice experiment on coca cultivation. *Journal of Development Economics*, 93, 2, 249-263.

INEI (2007). Censos Nacionales 2007: XI de Población y VI de Vivienda. Lima, Pérou.

Johnson-Hanks, J. (2002). The less er shame: Abortion among educated women in southern Cameroon. *Social Science & Medicine*, 55, 8, 1337-1349.

Mansfield, D. (2006). Development in Drug Environment: A Strategic Approach to Alternative Development. Article de discussion. Development Oriented Drug Control Program. GTZ. Allemagne.

Mensch, B., Hewett, P. et Erulkar, A. (2003). The reporting of sensitive behavior by adolescents: A methodological experiment in Kenya. *Demography*, 40, 2, 247-268.

Obando, E. (2006). U.S. Policy toward Peru: At odds for twenty years. Dans *Addicted to Failure. U.S. Security Policy in Latin America and the Andean Region*, (Éds., B. Loveman). Rowman & Littlefield Publishers Inc. États-Unis.

Office of Technology Assessment (1993). Alternative Coca Reduction Strategies in the Andean Region. U.S. Congress. OTA-F-556. Washington, États-Unis.

OICS (2009). Report on the International Narcotics Control Board for 2009. United Nations Publication. New York, États-Unis.

ONUDC (2001). Alternative Development in the Andean Area. The UNDCP Experience. Édition revisée. ODCCP Studies on Drugs and Crime. New York, États-Unis.

ONUDC (2009). Perú. Monitoreo de Cultivos de Coca 2008. Lima, Pérou.

ONUDC (2011). Perú. Monitoreo de Cultivos de Coca 2010. Lima, Pérou.

Punch, K. (2003). Survey research. The basics. *Sage Publications, Inc.* Royaume-Uni.

Rivera, M., Aufderheide, A., Cartmell, L., Torres, C. et Langsjoen, O. (2005). Antiquity of coca – Leaf chewing in the south central Andes: A 3000 year archaeological record of coca - Leaf chewing from Northern Chile. *Journal of Psychoactive Drugs*, 37, 4, 455-458.

Rospigliosi, F. (2004). Analisis de la Encuesta DEVIDA-INEI. Dans *El Consumo Traditional de la Hoja de Coca en el Peru*, (Éd., F. Rospigliosi). Instituto de Estudios Peruanos. Lima, Pérou.

Singer, E., Hippler, H. et Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research*, 4, 3.

Strunin, L. (2001). Assessing alcohol consumption: developments from qualitative research methods. *Social Science & Medicine*, 53, 2, 215-226.

Thoumi, F. (2003). Illegal Drugs, Economy, and Society in the Andes. Woodrow Wilson Center Press. Washington, États-Unis.

Torrico, J., Pohlan, H. et Janssens, M. (2005). Alternatives for the transformation of drug production areas in the chapare region, Bolivia. *Journal of Food, Agriculture and Development*, 3, 3-4, 167-172.

Tourangeau, R., et Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 5, 859-883.

Varkey, P., Balakrishna, P., Prasad, J., Abraham, S. et Joseph, A. (2000). The reality of unsafe abortion in a rural community in South India. *Reproductive Health Matters*, 8, 16, 83-91.

Willis, G. (2005). Cognitive interviewing. A tool for improving questionnaire design. *Sage Publications*, *Inc.* États-Unis.

Zufferey, A., Michaud, P., Jeannin, A., Berchtold, A., Chossis, I., van Melle, G. et Suris, J. (2007). Cumulative risk factors for adolescent alcohol misuse and its perceived consequences among 16 to 20 year old adolescents in Switzerland. *Preventive Medicine*, 45, 2-3, 233-239.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 157-170
Statistique Canada, N° 12-001-X au catalogue

Imputation pour la non-réponse non monotone dans le *Survey of Industrial Research and Development*

Jun Shao, Martin Klein et Jing Xu [1]

Résumé

Dans les études longitudinales, la non-réponse est souvent de nature non monotone. Dans le cas de la *Survey of Industrial Research and Development* (SIRD), il est raisonnable de supposer que le mécanisme de non-réponse dépend des valeurs antérieures, en ce sens que la propension à répondre au sujet d'une variable étudiée au point t dans le temps dépend de la situation de réponse ainsi que des valeurs observées ou manquantes de la même variable aux points dans le temps antérieurs à t. Puisque cette non-réponse n'est pas ignorable, l'approche axée sur la vraisemblance paramétrique est sensible à la spécification des modèles paramétriques s'appuyant sur la distribution conjointe des variables à différents points dans le temps et sur le mécanisme de non-réponse. La non-réponse non monotone limite aussi l'application des méthodes de pondération par l'inverse de la propension à répondre. En écartant toutes les valeurs observées auprès d'un sujet après la première valeur manquante pour ce dernier, on peut créer un ensemble de données présentant une non-réponse monotone ignorable, puis appliquer les méthodes établies pour la non-réponse ignorable. Cependant, l'abandon de données observées n'est pas souhaitable et peut donner lieu à des estimateurs inefficaces si le nombre de données écartées est élevé. Nous proposons d'imputer les réponses manquantes par la régression au moyen de modèles d'imputation créés prudemment sous le mécanisme de non-réponse dépendante des valeurs antérieures. Cette méthode ne requiert l'ajustement d'aucun modèle paramétrique sur la distribution conjointe des variables à différents points dans le temps ni sur le mécanisme de non-réponse. Les propriétés des moyennes estimées en appliquant la méthode d'imputation proposée sont examinées en s'appuyant sur des études en simulation et une analyse empirique des données de la SIRD.

Mots clés : Bootstrap ; modèle d'imputation ; régression à noyau ; ne manquant pas au hasard ; étude longitudinale ; dépendante des valeurs antérieures.

1. Introduction

Les études longitudinales, dans lesquelles des données sont recueillies auprès de chaque sujet échantillonné à plusieurs points dans le temps, sont très courantes dans des domaines de recherche tels que la médecine, la santé des populations, l'économie, les sciences sociales et les enquêtes par sondage. Habituellement, l'analyse statistique des données d'une enquête par sondage vise à estimer la moyenne d'une variable étudiée, ou à faire une inférence sur cette moyenne, à chaque point dans le temps. La non-réponse ou les données manquantes pour la variable étudiée représentent un obstacle sérieux à l'exécution d'une analyse statistique valide, parce que la propension à répondre peut dépendre directement ou indirectement de la valeur de la variable étudiée. La non-réponse est monotone si, quand une valeur manque à un point t dans le temps, toutes les futures valeurs au temps $s > t$ manquent. Nous nous concentrons sur la non-réponse non monotone, qui est fréquente dans les enquêtes longitudinales. Dans la *Survey of Industrial Research and Development* (SIRD) menée conjointement par le U.S. Census Bureau et la U.S. National Science Foundation (NSF), par exemple, une entreprise pourrait ne pas indiquer ses dépenses de recherche et de développement à l'année $t-1$, mais le faire à l'année t. Pour simplifier, nous faisons référence à la SIRD au temps présent tout au long de l'exposé, mais nous tenons à signaler qu'à partir de 2008, elle a été remplacée par la *Business R&D and Innovation Survey*.

Certaines méthodes de traitement de la non-réponse non monotone existantes peuvent être décrites brièvement comme il suit. L'approche paramétrique suppose des modèles paramétriques pour la propension à répondre ainsi que pour la distribution conjointe de la variable étudiée sur les divers points dans le temps (par exemple, Troxel, Harrington et Lipsitz 1998, Troxel, Lipsitz et Harrington 1998). Cependant, la validité de l'approche paramétrique dépend de la spécification correcte des modèles paramétriques. Vansteelandt, Rotnitzky et Robins (2007) ont proposé des méthodes sous certains modèles de la propension à répondre au temps t conditionnellement aux données observées antérieurement. Xu, Shao, Palta et Wang (2008) ont dérivé une procédure d'imputation sous les hypothèses que i) la propension à répondre au temps t dépend uniquement des valeurs de la variable étudiée au temps $t-1$ et ii) la variable étudiée à différents points dans le temps est une chaîne de Markov. Une autre approche, que nous appellerons censure, consiste à créer un ensemble de données présentant une « non-réponse monotone » en écartant toutes les valeurs observées de la variable étudiée auprès d'un sujet

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706. Courriel : shao@stat.wisc.edu ; Martin Klein, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C. 20233 ; Jing Xu, Department of Statistics, University of Wisconsin, Madison, WI 53706.

échantillonné après la première valeur manquante pour ce sujet. Les méthodes appropriées pour traiter la non-réponse monotone (par exemple Diggle et Kenward 1994, Robins et Rotnitzky 1995, Paik 1997) peuvent alors être appliquées à l'ensemble de données réduit. Cette approche peut être inefficace si de nombreuses données observées sont écartées. En outre, dans les applications pratiques, il n'est pas souhaitable d'éliminer des données observées.

Le but du présent article est de proposer une méthode d'imputation pour les données longitudinales en présence de non-réponse non monotone sous hypothèse de propension à répondre dépendante des valeurs antérieures décrite par Little (1995) : au point t dans le temps, la propension à répondre dépend des valeurs de la variable étudiée aux points dans le temps antérieurs à t. Cette hypothèse concernant la propension à répondre est plus faible que celle formulée dans Xu et coll. (2008) et diffère de celles figurant dans Vansteelandt et coll. (2007). Nous considérons une imputation qui ne requiert pas la spécification d'un modèle de la propension à répondre. L'imputation est utilisée fréquemment pour remplacer les valeurs manquantes dans les problèmes de sondage (Kalton et Kasprzyk 1986). Une fois que toutes les valeurs manquantes sont imputées, les estimations des paramètres sont calculées en se servant des moyennes estimées pour les données complètes en traitant les valeurs imputées comme des observations. La méthodologie d'imputation et d'estimation proposée, y compris une méthode bootstrap pour l'estimation de la variance, est présentée à la section 2. Afin d'examiner les propriétés en échantillon fini de la méthode proposée, nous présentons certains résultats de simulation à la section 3. Nous décrivons aussi une application de la méthode proposée à la SIRD. Enfin, à la dernière section, nous présentons certaines conclusions.

2. Méthodologie

Considérons l'approche assistée par modèle pour des données d'enquête échantillonnées à partir d'une population finie P. Nous supposons que la population P est divisée en un nombre fixe de classes d'imputation, qui sont habituellement des unions de certaines strates. Dans chaque classe d'imputation, la variable étudiée d'une unité de la population provient d'une superpopulation. Soit y_t la variable étudiée au point dans le temps t, $t = 1, ..., T$, $\mathbf{y} = (y_1, ..., y_T)$, δ_t l'indicateur précisant si y_t est observé, et $\boldsymbol{\delta} = (\delta_1, ..., \delta_T)$. Puisque l'imputation est effectuée indépendamment dans chaque classe d'imputation, pour simplifier la notation, nous supposons à la présente section qu'il n'existe qu'une seule classe d'imputation.

Dans tout l'exposé, nous considérons que la non-réponse est non monotone et nous supposons qu'il n'y a pas de non-réponse à la période de référence $t = 1$. La propension à répondre dépend des valeurs antérieures si

$$P(\delta_t = 1 \mid \mathbf{y}, \delta_1, ..., \delta_{t-1}, \delta_{t+1}, ..., \delta_T) = P(\delta_t = 1 \mid y_1, ..., y_{t-1}, \delta_1, ..., \delta_{t-1}), \quad t = 2, ..., T, \quad (1)$$

où P est calculée par rapport à la superpopulation. Quand la non-réponse est monotone, la propension à répondre dépendante des valeurs antérieures devient ignorable (Little et Rubin 2002), puisque nous observons toutes les valeurs antérieures ou savons avec certitude que y_t manque si la valeur manquait à la période $t - 1$, et nous pouvons utiliser une méthode d'imputation par la régression linéaire proposée par Paik (1997). Par contre, quand la non-réponse n'est pas monotone, la propension à répondre dépendante des valeurs antérieures n'est pas ignorable, parce que l'indicateur de réponse au temps t dépend statistiquement des valeurs antérieures de la variable étudiée, dont certaines pourraient ne pas être observées. Dans ce cas, la méthode de Paik ne s'applique pas.

2.1 Imputation pour des sujets dont la première valeur manquante a lieu au temps t

Soit $t > 1$ un point fixe dans le temps et $r + 1$ le point dans le temps auquel la première valeur manquante de $\mathbf{y}$ se produit. Quand $r + 1 = t$, c'est-à-dire un sujet dont la première valeur manquante se produit au temps t, la procédure d'imputation que nous proposons est la même que celle pour le cas de la non-réponse monotone (Paik 1997). Cependant, nous devons fournir une justification, puisque nous avons une propension à répondre différente. Nous montrons à l'annexe que, sous l'hypothèse (1),

$$E(y_t \mid y_1, ..., y_{t-1}, \delta_1 = \cdots = \delta_{t-1} = 1, \delta_t = 0) = E(y_t \mid y_1, ..., y_{t-1}, \delta_1 = \cdots = \delta_{t-1} = 1, \delta_t = 1) \quad t = 2, ..., T, \quad (2)$$

où E est l'espérance par rapport à la superpopulation. Désignons la quantité à la première ligne de (2) par $\phi_{t, t-1}(y_1, ..., y_{t-1})$, qui est l'espérance conditionnelle d'une valeur de y_t manquante sachant les valeurs observées $y_1, ..., y_{t-1}$. Si $\phi_{t, t-1}$ est connue, une valeur imputée naturelle pour y_t est $\phi_{t, t-1}(y_1, ..., y_{t-1})$. Cependant, $\phi_{t, t-1}$ est habituellement inconnue. Puisque $\phi_{t, t-1}$ ne peut pas être estimée par la régression de y_t sur $y_1, ..., y_{t-1}$ en se fondant sur des données provenant de sujets pour lesquels des valeurs de y_t manquent, nous devons utiliser (2), c'est-à-dire le fait que $\phi_{t, t-1}$ est égale à la quantité figurant à la deuxième ligne de (2), qui est l'espérance conditionnelle d'une valeur y_t observée sachant les valeurs observées $y_1, ..., y_{t-1}$, et qui peut être estimée par la régression de y_t sur $y_1, ..., y_{t-1}$ en utilisant les données provenant de tous les sujets pour lesquels on dispose de la valeur observée y_t et des valeurs observées $y_1, ..., y_{t-1}$. Notons que (2) est l'équivalent de (5) dans Xu et coll. (2008) sous l'hypothèse de dépendance à

l'égard de la dernière valeur observée, qui est plus forte que l'hypothèse de dépendance à l'égard des valeurs antérieures (1). Sous une hypothèse plus forte, nous arrivons à utiliser plus de données dans l'ajustement de la régression.

Supposons qu'un échantillon S soit sélectionné dans P selon un plan d'échantillonnage probabiliste donné. Pour chaque $i \in S$, $\boldsymbol{\delta}_i = (\delta_{i1}, ..., \delta_{iT})$ est observé, la variable étudiée y_{it} avec $\delta_{it} = 1$ est observée, et y_{it} avec $\delta_{it} = 0$ n'est pas observée, $t = 1, ..., T$. En ce qui concerne la superpopulation, $(\mathbf{y}_i, \boldsymbol{\delta}_i)$ possède la même distribution que $(\mathbf{y}, \boldsymbol{\delta})$ et les $(\mathbf{y}_i, \boldsymbol{\delta}_i)$ sont indépendants, où $\mathbf{y}_i = (y_{i1}, ..., y_{iT})$. Pour $t = 2, ..., T$, soit $\hat{\phi}_{t,\,t-1}$ l'estimateur par la régression de $\phi_{t,\,t-1}$ fondé sur les observations avec $\delta_{i1} = \cdots = \delta_{i(t-1)} = 1$. Une valeur manquante y_{it} lorsque l'on a les valeurs observées $y_{i1}, ..., y_{i(t-1)}$ est alors imputée par $\tilde{y}_{it} = \hat{\phi}_{t,\,t-1}(y_{i1}, ..., y_{i(t-1)})$.

À titre d'illustration, considérons le cas où $t = 3$ ou 4. Dans le tableau 1, la direction horizontale correspond aux points dans le temps et la direction verticale correspond à différents schémas de données manquantes, chaque schéma étant représenté par un vecteur de 0 et de 1, où 0 indique une valeur manquante et 1, une valeur observée. Pour $t = 3$ et $r = 2$, en tant que première des deux étapes, nous considérons le cas de données manquantes au temps 3 avec première occurrence de données manquantes au temps 3, c'est-à-dire le schéma (1,1,0). Conformément au modèle d'imputation (2), nous ajustons une régression en utilisant des données ayant un schéma (1,1,1) indiquées par + (utilisées comme variables explicatives) et × (utilisées comme réponses). Puis, les valeurs imputées (indiquées par ○) sont obtenues à partir de la régression ajustée en utilisant les données indiquées par ∗ comme variables explicatives. Pour $t = 4$ et $r = 3$, une imputation selon le schéma (1,1,1,0) peut être effectuée de manière similaire en utilisant des données présentant un schéma (1,1,1,1) pour ajuster la régression.

Quel type de régression pouvons-nous ajuster pour obtenir $\tilde{y}_{it}$? Nous montrons à l'annexe que, si (1) est vérifiée et que $E(y_t \mid y_1, ..., y_{t-1})$ est linéaire en $y_1, ..., y_{t-1}$ pour tout t en l'absence de non-réponse, alors

$$E(y_t \mid y_1, ..., y_{t-1}, \delta_1 = \cdots = \delta_{t-1} = 1) \quad \text{est linéaire en } y_1, ..., y_{t-1} \qquad (3)$$

et la régression linéaire sous l'approche assistée par modèle peut être utilisée pour estimer $\phi_{t,\,t-1}$. Si $E(y_t \mid y_1, ..., y_{t-1})$ n'est pas linéaire, l'une des méthodes décrites à la section 2.3 peut être appliquée.

Tableau 1
Illustration du processus d'imputation

	Étape 1 : $r = 2, t = 3$			Étape 2 : $r = 1, t = 3$		
	Temps			Temps		
Schéma	1	2	3	1	2	3
(1,0,0)				*		○
(1,1,0)	*	*	○	+		⊗
(1,1,1)	+	+	×			
(1,0,1)						

	Étape 1 : $r = 3, t = 4$				Étape 2 : $r = 2, t = 4$				Étape 3 : $r = 1, t = 4$			
	Temps				Temps				Temps			
Schéma	1	2	3	4	1	2	3	4	1	2	3	4
(1,0,0,0)									*			○
(1,1,0,0)					*	*		○	+			⊗
(1,1,1,0)	*	*	*	○	+	+		⊗	+			⊗
(1,0,1,0)									*			○
(1,0,0,1)												
(1,1,0,1)												
(1,0,1,1)												
(1,1,1,1)	+	+	+	×								

\+ : données observées utilisées dans l'ajustement de la régression comme variables explicatives.
× : données observées utilisées dans l'ajustement de la régression comme réponses.
⊗ : données imputées utilisées dans l'ajustement de la régression comme réponses.
∗ : données observées utilisées comme variables explicatives dans l'imputation.
○ : valeurs imputées.

2.2 Imputation pour des sujets dont la première valeur manquante a lieu au temps $r + 1 < t$

L'imputation pour un sujet dont la première valeur manquante se produit au temps $r + 1 < t$ est plus compliquée et diffère de celle applicable au cas de non-réponse monotone. En effet, quand $r+1 < t$ et que la non-réponse est monotone,

$$E(y_t | y_1, ..., y_r, \delta_1 = \cdots = \delta_r = 1, \delta_t = 0) = E(y_t \mid y_1, ..., y_r, \delta_1 = \cdots = \delta_r = 1, \delta_t = 1) \quad r = 1, ..., t-2, \quad t = 2, ..., T, \quad (4)$$

tandis que (4) n'est pas vérifiée quand la non-réponse est non monotone (voir la preuve à l'annexe). D'où, nous devons spécifier des modèles différents pour les sujets dont la première valeur manquante se produit à $r + 1 < t$. Nous montrons à l'annexe que, quand $r + 1 < t$,

$$E(y_t | y_1, ..., y_r, \delta_1 = \cdots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = E(y_t \mid y_1, ..., y_r, \delta_1 = \cdots = \delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) \quad r = 1, ..., t-2, \ t = 2, ..., T. \quad (5)$$

Expliquons maintenant comment utiliser (5) pour imputer les valeurs manquantes à un point dans le temps fixe t. Soit $\phi_{t,r}(y_1,...,y_r)$ la quantité à la première ligne de (5). Si $\phi_{t,r}$ est connue, y_t peut être imputée par $\phi_{t,r}(y_1,...,y_r)$. Sinon, elle doit être estimée en se fondant sur (5). Contrairement au modèle (2) ou (4), l'espérance conditionnelle à la deuxième ligne de (5) est conditionnelle à une valeur manquante y_t ($\delta_t = 0$), même si les valeurs $y_1, ..., y_r$ sont observées. Si nous exécutons l'imputation séquentiellement, conformément à $r = t-1, t-2, ..., 1$, alors, pour un temps donné $r < t-1$, les valeurs y_t manquantes pour les sujets dont la première valeur manquante a eu lieu au point dans le temps $r + 2$ sont déjà imputées par la méthode décrite à la présente section ou à la section 2.1. Nous pouvons ajuster une régression de la valeur imputée y_t sur les valeurs observées $y_1, ..., y_r$ en utilisant les données provenant de tous les sujets pour lesquels on dispose de la valeur y_t déjà imputée (utilisée comme réponses), des valeurs $y_1, ..., y_r$ observées (utilisées comme variables explicatives) et de $\delta_{r+1} = 1$. Une fois qu'un estimateur $\hat{\phi}_{t,r}$ est obtenu, une valeur manquante y_{it} avec une première valeur manquante au temps $r + 1$ est alors imputée par $\tilde{y}_{it} = \hat{\phi}_{t,r}(y_{i1}, ..., y_{ir})$.

Considérons de nouveau le cas où $t = 3$ ou 4 et le tableau 1. Après la première étape pour $t = 3$ décrite à la section 2.1, à la deuxième étape, nous imputons les valeurs manquantes avec $r = 1$ selon un schéma (1,0,0). Conformément au modèle d'imputation (5), nous ajustons une régression en utilisant des données présentant un schéma (1,1,0) indiquées par + (utilisées comme variables explicatives) et $\otimes$ (valeurs imputées antérieurement utilisées comme réponse). Alors, les valeurs imputées (indiquées par ○) sont obtenues d'après la régression ajustée en utilisant les données indiquées par $*$ comme variables explicatives. Pour $t = 4$, après la première étape décrite à la section 2,1, à la deuxième étape ($r = 2$), nous ajustons une régression en utilisant des données présentant un schéma (1,1,1,0) indiquées par + (utilisées comme variables explicatives) et $\otimes$ (valeurs imputées antérieurement utilisées comme réponses). Alors, les valeurs imputées (indiquées par ○) au temps $t = 4$ présentant un schéma (1,1,0,0) sont obtenues d'après la régression ajustée en utilisant les données indiquées par $*$ comme variables explicatives. À l'étape 3, pour $t = 4$, nous ajustons une régression en utilisant des données présentant les schémas (1,1,0,0) et (1,1,1,0) indiquées par + (utilisées comme variables explicatives) et $\otimes$ (valeurs imputées antérieurement utilisées comme réponses). Ensuite, les valeurs imputées (indiquées par ○) au temps $t = 4$ présentant les schémas (1,0,0,0) et (1,0,1,0) sont obtenues d'après la régression ajustée en utilisant les données indiquées par $*$ comme variables explicatives.

Bien qu'au temps t l'imputation doive être exécutée séquentiellement suivant $r = t-1, ..., 1$, l'imputation pour différents points dans le temps peut être effectuée dans n'importe quel ordre. On peut le constater en examinant l'exemple du tableau 1, où les valeurs imputées à $t = 3$ n'interviennent pas dans le processus d'imputation à $t = 4$ ou inversement, quoique certaines données observées seront utilisées à plusieurs reprises dans l'ajustement de la régression. Lorsque les données sont fournies en fonction du temps, il est naturel d'imputer les non-répondants dans l'ordre $t = 2, ..., T$.

Pourquoi pouvons-nous utiliser des valeurs imputées antérieurement comme réponses dans l'estimation de la fonction de régression $\phi_{t,r}$ quand $r < t-1$? Pour t et $r < t-1$ donnés, une valeur imputée antérieurement avec la première valeur manquante à $s+1 > r+1$ est un estimateur de

$$\tilde{y}_t = E(y_t | y_1,...,y_s, \delta_1 = \cdots = \delta_s = 1, \delta_{s+1} = 0, \delta_t = 0) = E(y_t \mid y_1,...,y_s, \delta_1 = \cdots = \delta_{s+1} = 1, \delta_t = 0).$$

En vertu de la propriété d'espérance conditionnelle et de (5),

$$E[E(y_t | y_1,...,y_s, \delta_1 = \cdots = \delta_{s+1} = 1, \delta_t = 0) \mid y_1,...,y_r, \delta_1 = \cdots = \delta_{r+1} = 1, \delta_t = 0] = E(y_t | y_1,...,y_r, \delta_1 = \cdots = \delta_{r+1} = 1, \delta_t = 0) = E(y_t | y_1,...,y_r, \delta_1 = \cdots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0). \quad (6)$$

D'où y_t et $\tilde{y}_t$ ont la même espérance conditionnelle, sachant $y_1,...,y_r, \delta_1 = \cdots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0$. Par conséquent, l'utilisation des valeurs imputées antérieurement comme réponses dans la régression produit un estimateur valide de $\phi_{t,r}$. Notons que les valeurs imputées antérieurement ne devraient pas être utilisées comme variables

explicatives dans la régression, car l'équation (6) n'est pas vérifiée si certaines des valeurs $y_1,...,y_s$ sont imputées.

Alors que l'on utilise toutes les données observées à n'importe quel point dans le temps t pour l'estimation de $E(y_t)$, on se sert de certaines données observées à un temps $< t$, mais pas de toutes pour l'imputation, afin d'éviter des biais sous non-réponse non ignorable. La situation est différente en cas de non-réponse ignorable, où habituellement toutes les données observées antérieures peuvent être utilisées dans l'imputation par la régression.

2.3 Régression pour l'imputation

Dans (5), les espérances conditionnelles dépendent non seulement de la distribution de $\mathbf{y}$, mais aussi de la propension à répondre. Même si $E(y_t \mid y_1,...,y_{t-1})$ est linéaire, les espérances conditionnelles figurant dans (5) ne le sont pas forcément, ce qui diffère du cas où $r+1=t$ considéré à la section 2.1. Le résultat (10) à l'annexe en est un exemple.

Si nous ne disposons pas d'un modèle paramétrique approprié pour $\phi_{t,r}$, nous pouvons appliquer la régression par la méthode du noyau, ou régression à noyau, non paramétrique donnée dans Cheng (1994) pour obtenir $\hat{\phi}_{t,r}$. Puisque la variable dépendante $(y_{i1},...,y_{ir})$ est multivariée quand $r \geq 2$, la régression à noyau présente toutefois une grande variabilité, à moins que le nombre de sujets échantillonnés dans la catégorie définie par $\delta_{i1} = \cdots = \delta_{i(r+1)} = 1$ soit très grand. Ce problème porte le nom de malédiction de la dimension.

Donc, nous considérons les options qui suivent sous l'hypothèse supplémentaire que la dépendance de δ_t à l'égard de $y_1,...,y_{t-1}$ a lieu par la voie d'une combinaison linéaire de $y_1,...,y_{t-1}$. C'est-à-dire

$$P(\delta_t = 1 \mid y_1,..., y_{t-1}, \delta_1,..., \delta_{t-1}) = \Psi\left(\sum_{l=1}^{t-1} \gamma_l^{\delta_1,...,\delta_{t-1}} y_l\right), \quad (7)$$

où $\gamma_l^{\delta_1,...,\delta_{t-1}}$, $l = 1, ..., t-1$, sont les paramètres inconnus qui dépendent de $\delta_1, ..., \delta_{t-1}$ et Ψ est une fonction inconnue dans l'intervalle $[0, 1]$. Sous (7), nous montrons à l'annexe que

$$\begin{aligned} E(y_t \mid z_r, \delta_1 = \cdots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) \\ = E(y_t \mid z_r, \delta_1 = \cdots = \delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) \\ r = 1,...,t-2, t = 2,...,T, \quad (8) \end{aligned}$$

où $z_r = \sum_{l=1}^{r} \gamma_{r,l}\, y_l$ et $\gamma_{r,l} = \gamma_l^{\delta_1,...,\delta_r}$ avec $\delta_1 = \cdots = \delta_r = 1$. Donc, pour imputer les valeurs des non-répondants, nous pouvons conditionner sur la combinaison linéaire z_r et utiliser (8) au lieu de conditionner sur $y_1, ..., y_r$ et d'utiliser (5).

Soit $\psi_{t,r}(z_r)$ la fonction définie à la deuxième ligne de (8). Soulignons que $\psi_{t,r}$ n'est pas nécessairement identique à $\phi_{t,r}$. S'il existe une forte relation linéaire entre y_t et $y_1, ..., y_r$, il se pourrait que $\psi_{t,r}$ soit approximativement linéaire de sorte que nous pouvons ajuster une régression linéaire pour obtenir un estimateur $\hat{\psi}_{t,r}$. En théorie, cette méthode contient un biais quand $\psi_{t,r}$ n'est pas linéaire. Si $\boldsymbol{\gamma}_r = (\gamma_{r,1}, ..., \gamma_{r,r})'$ est connu, nous pouvons appliquer une régression à noyau unidimensionnelle pour obtenir un estimateur $\hat{\psi}_{t,r}$, en utilisant l'indice unidimensionnel z_r. Puisque $\boldsymbol{\gamma}_r$ est inconnu, nous devons d'abord l'estimer par $\hat{\boldsymbol{\gamma}}_r$, puis obtenir $\hat{\psi}_{t,r}$ en exécutant la régression à noyau unidimensionnelle avec remplacement de $\boldsymbol{\gamma}_r$ par $\hat{\boldsymbol{\gamma}}_r$. Par exemple, on peut recourir à la régression inverse par tranches (Duan et Li 1991) pour obtenir $\hat{\boldsymbol{\gamma}}_r$. Cependant, ce genre de méthode non paramétrique est parfois inefficace. S'il existe une forte relation linéaire entre y_t et $y_1, ..., y_r$, nous pouvons appliquer la régression linéaire pour obtenir $\hat{\boldsymbol{\gamma}}_r$. Quoiqu'il en soit, nous utilisons $y_{i1}, ..., y_{ir}$ avec $\delta_{i1} = \cdots = \delta_{i(r+1)} = 1$ comme variables explicatives et les valeurs y_{it} imputées comme réponses dans tout type d'ajustement de la régression. Après avoir obtenu $\hat{\psi}_{t,r}$ et $\hat{\boldsymbol{\gamma}}_r = (\hat{\gamma}_{r,1}, ..., \hat{\gamma}_{r,r})'$, nous imputons une valeur y_{it} manquante par $\tilde{y}_{it} = \hat{\psi}_{t,r}(\hat{\gamma}_{r,1} y_{i1} + \cdots + \hat{\gamma}_{r,r} y_{ir})$.

Nous donnons à la méthode qui consiste simplement à appliquer la régression linéaire le nom de méthode d'imputation par la régression linéaire, et à celle qui consiste à appliquer la régression à noyau à l'indice z_r, le nom de méthode d'imputation par la régression à noyau avec indice unidimensionnel. L'un des avantages qu'a cette dernière sur l'imputation par la régression à noyau tient au fait que l'on n'effectue qu'une seule régression à noyau unidimensionnelle, ce qui évite la malédiction de la dimension et réduit la variabilité.

Ces méthodes peuvent aussi être appliquées au cas où $r = t-1$ si $E(y_t \mid y_1, ..., y_{t-1})$ n'est pas linéaire.

En théorie, des estimateurs tels que les moyennes estimées en se fondant sur l'imputation par la régression à noyau ou par la régression à noyau avec indice unidimensionnel sont asymptotiquement sans biais, mais ils ne sont peut-être pas meilleurs que ceux fondés sur l'imputation par la régression linéaire quand le nombre de sujets échantillonnés dans chaque catégorie (t, r) n'est pas très grand. Les propriétés des moyennes estimées en utilisant l'imputation par la régression linéaire, par la régression à noyau et par la régression à noyau avec indice unidimensionnel sont examinées par simulation à la section 3.

2.4 Estimation

Nous considérons l'estimation du total de population finie ou de la moyenne de y_t à chaque point t fixé dans le temps, ce qui est souvent l'objectif principal d'une étude par

sondage. À tout temps t, soit $\tilde{y}_{it} = y_{it}$ quand $\delta_{it} = 1$ et $\tilde{y}_{it}$ la valeur imputée en utilisant l'une des méthodes de la section 2 quand $\delta_{it} = 0$. Le total de population finie et la moyenne de y_t peuvent être estimés par

$$\hat{Y}_t = \sum_{i \in S} w_i \tilde{y}_{it} \qquad \text{et} \qquad \overline{Y}_t = \sum_{i \in S} w_i \tilde{y}_{it} \Big/ \sum_{i \in S} w_i \quad (9)$$

respectivement, où w_i est le poids de sondage construit de façon que, s'il n'y a pas de non-réponse, $\hat{Y}_t$ soit un estimateur sans biais du total de population finie au temps t sous le plan d'échantillonnage probabiliste. La moyenne de superpopulation de y_t peut alors être estimée par $\overline{Y}_t$. Notons que $\sum_{i \in S} w_i$ est un estimateur sans biais de la taille de la population finie N et, pour certains plans d'échantillonnage simples, il est exactement égal à N.

Les poids de sondage devraient également être utilisés dans l'ajustement de la régression pour l'imputation. Sous les mêmes conditions que celles données dans Cheng (1994), $\hat{Y}_t$ ou $\overline{Y}_t$ fondé sur l'imputation par la régression à noyau ou par la régression à noyau avec indice unidimensionnel est convergent et asymptotiquement normal à mesure que la taille de l'échantillon tend vers ∞. Les conditions requises et les preuves peuvent être consultées dans Xu (2007).

Si nous appliquons une méthode d'imputation par la régression linéaire telle qu'elle est exposée à la section 2.3, la moyenne estimée résultante au temps t peut être asymptotiquement biaisée. Ce biais est faible s'il est possible de bien approximer la fonction $\psi_{t,r}$ par une fonction linéaire dans l'étendue des valeurs des données. Par ailleurs, l'imputation par la régression à noyau ou par la régression à noyau avec indice unidimensionnel peut nécessiter un échantillon de beaucoup plus grande taille que l'imputation par la régression linéaire. Par conséquent, les propriétés globales de la moyenne estimée en utilisant l'imputation par la régression linéaire peuvent demeurer meilleures, comme l'indiquent les résultats de simulation présentés à la section 3.

2.5 Estimation de la variance

Afin d'évaluer l'exactitude statistique ou l'inférence, comme la construction d'un intervalle de confiance pour la moyenne de y_t au temps t, nous avons besoin des estimateurs de variance de $\hat{Y}_t$ ou $\overline{Y}_t$ fondés sur des données imputées. Étant donné la complexité de la procédure d'imputation, il est difficile d'obtenir des formules explicites pour la variance de $\hat{Y}_t$ ou $\overline{Y}_t$. Nous considérons alors la méthode du bootstrap (Efron 1979). Un bootstrap correct peut être obtenu en répétant le processus d'imputation dans chacun des échantillons bootstrap (Shao et Sitter 1996). Soit $\hat{\theta}$ l'estimateur examiné. Une procédure bootstrap peut être exécutée comme il suit.

1. Tirer de S un échantillon bootstrap sous forme d'échantillon aléatoire simple de même taille avec remise parmi l'ensemble de sujets échantillonnés.
2. Utiliser les poids de sondage, les indicateurs de réponse et les données observées provenant de l'ensemble de données original pour les unités de l'échantillon bootstrap pour former un ensemble de données bootstrap. Appliquer la procédure d'imputation proposée aux données bootstrap. Calculer l'analogue bootstrap $\hat{\theta}^*$ de $\hat{\theta}$.
3. Indépendamment, répéter B fois les étapes qui précèdent pour obtenir $\hat{\theta}^{*1}, ..., \hat{\theta}^{*B}$. La variance d'échantillon de $\hat{\theta}^{*1}, ..., \hat{\theta}^{*B}$ est l'estimateur de variance bootstrap pour $\hat{\theta}$.

Dans l'application, chaque $\hat{\theta}^{*b}$ peut être calculé en utilisant les b^{e} données bootstrap $(\mathbf{y}_i, \boldsymbol{\delta}_i, w_i^{*b})$, $i \in S$, où $w_i^{*b} = w_i$ est multiplié par le nombre de fois que l'unité i figure dans le b^{e} échantillon bootstrap. Notons que le même w_i^{*b} peut être utilisé pour toute les variables d'intérêt, plutôt que pour $\mathbf{y}_i$ seulement.

3. Résultats empiriques

Nous étudions $\hat{Y}_t$ ou $\overline{Y}_t$ dans (9) obtenu à chaque point t dans le temps en se fondant sur les méthodes d'imputation proposées. Nous considérons d'abord une simulation comprenant une population normale pour les y_t. Ensuite, nous présentons une application aux données de la SIRD. Pour examiner les propriétés des méthodes proposées pour la SIRD, nous terminons par une simulation portant sur une population créée en utilisant les données de la SIRD. Nous avons implémenté les méthodes d'imputation proposées en R (R Development Core Team 2009). Pour ajuster les régressions non paramétriques requises, nous utilisons la fonction *loess* de R avec les valeurs par défaut des paramètres, qui ajuste une surface polynomiale locale pour une ou plusieurs variables explicatives. Les régressions linéaires requises sont ajustées facilement en R en utilisant la fonction *lm*. Nos implémentations des méthodes proposées comprennent la vérification de l'erreur (par exemple veiller à ce que le nombre de points soit suffisant pour l'ajustement de la régression à chaque étape), ce qui est particulièrement important dans des conditions de bootstrap et de simulation où les méthodes d'imputation sont répétées de nombreuses fois et où chaque itération ne peut pas être examinée manuellement. Nous avons choisi comme option par défaut l'imputation d'une moyenne globale dans le cas où le nombre de points n'était pas suffisant pour ajuster une régression.

3.1 Résultats de simulation pour une population normale

Nous avons exécuté une étude par simulation en utilisant une population normalement distribuée $\mathbf{y}_1,...,\mathbf{y}_n$, $n = 2\,000$ et $T = 4$. Nous avons en outre utilisé une seule classe d'imputation et un plan d'échantillonnage aléatoire simple avec remise. Dans la simulation, les $\mathbf{y}_i$ ont été générés de manière indépendante à partir de la loi normale multivariée de vecteur de moyennes (1,33 ; 1,94 ; 2,73 ; 3,67) et de matrice de covariance ayant une structure AR(1) avec coefficient de corrélation de 0,7 et variance unitaire ; au temps $t = 1$, toutes les données ont été observées ; les données manquantes au temps $t = 2, 3, 4$ ont été générées conformément à

$$P(\delta_t = 1 \mid y_1,...,y_{t-1},\delta_1,...,\delta_{t-1}) = 1 - \Phi\left(0,6\left(1 - \sum_{j=1}^{t-1} y_j \gamma_j^{\delta_1,...,\delta_{t-1}}\right)\right)$$

où

$$\gamma_j^{\delta_1,\ ...,\ \delta_{t-1}} = \frac{j + (1-\delta_j)\,j}{\sum_{k=1}^{t-1}[k + (1-\delta_k)\,k]}, \quad j = 1, ..., t-1,$$

et Φ est la fonction de répartition normale centrée réduite. Les probabilités non conditionnelles des schémas de non-réponse sont données au tableau 2.

Tableau 2
Probabilités des schémas de non-réponse dans l'étude en simulation (population normale)

	Schéma	Probabilité	
Monotone	(1,0,0,0)	0,062	
	(1,1,0,0)	0,043	total = 0,181
	(1,1,1,0)	0,076	
Intermittente	(1,0,0,1)	0,113	
	(1,0,1,0)	0,071	
	(1,0,1,1)	0,186	total = 0,494
	(1,1,0,1)	0,124	
Complète	(1,1,1,1)	0,325	

Pour les comparaisons, nous avons inclus neuf estimateurs de la moyenne de y_t, à savoir les moyennes d'échantillons fondées sur 1) les données complètes (utilisée comme norme de référence), 2) les répondants avec poids corrigés en supposant que la probabilité de réponse est la même dans chaque classe d'imputation, 3) la censure et l'imputation par la régression linéaire, qui consiste à écarter d'abord toutes les observations d'un sujet après la première valeur manquante pour créer un ensemble de données à « non-réponse monotone », puis à appliquer l'imputation par la régression linéaire comme il est décrit dans Paik (1997), 4) l'imputation par la régression à noyau proposée, 5) l'imputation par la régression linéaire proposée, 6) l'imputation par la régression à noyau avec indice unidimensionnel proposée, en utilisant la régression inverse par tranches pour obtenir $\hat{\boldsymbol{\gamma}}_r$; 7) l'imputation par la régression à noyau proposée dans Xu et coll. (2008) fondée sur la propension à répondre dépendant de la dernière valeur, 8) l'imputation par la régression linéaire fondée sur une régression des réponses au temps t sur les valeurs observées et imputées aux points dans le temps 1, ..., $t-1$ (en traitant les valeurs imputées comme des valeurs observées), 9) imputation par la régression linéaire fondée sur une régression des réponses au temps t sur les données observées pour les unités ayant le même schéma de données manquantes aux points dans le temps 1, ..., $t-1$.

La méthode (2) ne tient simplement pas compte des non-répondants et est donc biaisée et inefficace. Sous l'hypothèse de propension à répondre (1), les méthodes (7) à (9) sont également biaisées pour $t \geq 3$, parce que la méthode (7) requiert l'hypothèse de dépendance à l'égard de la dernière valeur qui est plus forte que l'hypothèse (1), la méthode (8) traite les valeurs imputées antérieurement comme des valeurs observées dans la régression, et la méthode (9) requiert la condition qui suit, qui n'est pas vérifiée sous (1) :

$$E(y_t \mid y_1,...,y_{t-1},\delta_1 = j_1,...,\delta_{t-1} = j_{t-1},\delta_t = 0) = E(y_t \mid y_1,...,y_{t-1},\delta_1 = j_1,...,\delta_{t-1} = j_{t-1},\delta_t = 1)$$

où $(j_1, ..., j_{t-1})$ est un schéma de données manquantes fixe. Enfin, comme il est discuté à la section 2.3, la méthode (5) est également biaisée pour $t \geq 3$, puisque la régression linéaire n'est pas un modèle exactement correct. Cependant, les méthodes (5), (8) et (9) peuvent quand même donner de bons résultats quand les biais ne sont pas importants, parce que l'emploi d'un modèle plus simple et de plus de données dans la régression pour l'imputation peut compenser la perte due à l'imputation biaisée. En outre, toute hypothèse concernant la propension à répondre peut n'être vérifiée qu'approximativement et il est souhaitable d'étudier empiriquement diverses méthodes dans toute application particulière.

Pour le cas où $r = t - 1$, nous appliquons l'imputation par la régression linéaire comme il est exposé à la section 2.1. Donc, les méthodes (3) à (6), (8) et (9) donnent toutes de bons résultats quand $t = 2$.

Le tableau 3 donne (sur la base de 1 000 exécutions de la simulation) le biais relatif et l'écart-type (E.-T.) de l'estimateur de la moyenne, la moyenne de $\widehat{\text{E.-T.}}_{\text{boot}}$, l'estimateur bootstrap de l'écart-type fondé sur 200 répliques bootstrap, et la probabilité de couverture de l'intervalle de confiance (IC) à 95 % approximatif obtenu en utilisant l'estimateur ponctuel $\pm 1,96 \times \widehat{\text{E.-T.}}_{\text{boot}}$. Les résultats du tableau 3 se résument comme il suit.

1. La moyenne d'échantillon calculée en ne tenant pas compte des données manquantes est clairement biaisée. Même si, pour $t = 4$, son biais relatif n'est que de 3,5 %, il donne lieu à une très faible probabilité de couverture de l'intervalle de confiance, parce

que l'écart-type de la moyenne estimée est également très faible.

2. L'estimateur bootstrap de l'écart-type donne de bons résultats dans tous les cas, même quand l'estimateur de la moyenne est biaisé.
3. Le biais de $\overline{Y}_t$ fondé sur la censure et l'imputation par la régression linéaire est négligeable, de sorte que la probabilité de couverture de l'intervalle de confiance apparenté est proche du niveau nominal de 95 % ; toutefois, son écart-type est grand quand $t = 3$ ou $t = 4$. L'inefficacité de cette méthode est manifestement due au fait d'écarter des données observées provenant de près de 50 % des sujets échantillonnés caractérisés par une non-réponse intermittente. Sa performance empire à mesure que t augmente.
4. Le biais relatif de $\overline{Y}_t$ fondé sur l'imputation par la régression à noyau proposée est compris entre 0,0 % et 0,5 %, mais il est suffisamment grand pour produire de mauvais résultats de couverture de l'intervalle de confiance apparenté au temps $t = 4$.
5. $\overline{Y}_t$ fondé sur l'imputation par la régression linéaire proposée possède un biais négligeable et une variance plus faible que celle de $\overline{Y}_t$ fondé sur la régression à noyau. La probabilité de couverture de l'intervalle de confiance apparenté est proche du niveau nominal de 95 %.
6. $\overline{Y}_t$ fondé sur l'imputation par la régression à noyau avec indice unidimensionnel est généralement bon, mais un peu moins que l'estimateur fondé sur l'imputation par la régression linéaire.
7. Le biais de $\overline{Y}_t$ fondé sur les méthodes (7) à (9) n'est pas négligeable quand $t = 3$ ou $t = 4$, ce qui entraîne de mauvaises propriétés de l'intervalle de confiance apparenté.

Tableau 3
Résultats de simulation pour l'estimation de la moyenne (population normale)

Méthode	Quantité	$t = 2$	$t = 3$	$t = 4$
Données complètes	Biais relatif	0 %	0 %	0 %
	E.-T.	0,0221	0,0223	0,0221
	$\widehat{\text{E.-T.}}_{\text{boot}}$	0,0223	0,0223	0,0224
	Couverture de l'IC	94,9 %	94,4 %	95,4 %
Répondants seulement	Biais relatif	12,8 %	6,8 %	3,5 %
	E.-T.	0,0282	0,0272	0,0248
	$\widehat{\text{E.-T.}}_{\text{boot}}$	0,0285	0,0267	0,0252
	Couverture de l'IC	0,0 %	0,0 %	0,2 %
Censure et imputation par la régression linéaire	Biais relatif	0,0 %	0,0 %	-0,1 %
	E.-T.	0,0275	0,0358	0,0418
	$\widehat{\text{E.-T.}}_{\text{boot}}$	0,0276	0,0354	0,0431
	Couverture de l'IC	95,1 %	94,6 %	95,6 %
Imputation par la régression à noyau proposée	Biais relatif	0,0 %	0,4 %	0,5 %
	E.-T.	0,0275	0,0288	0,0283
	$\widehat{\text{E.-T.}}_{\text{boot}}$	0,0276	0,0288	0,0288
	Couverture de l'IC	95,1 %	92,5 %	88,6 %
Imputation par la régression linéaire proposée	Biais relatif	0,0 %	0,1 %	0,0 %
	E.-T.	0,0275	0,0286	0,0279
	$\widehat{\text{E.-T.}}_{\text{boot}}$	0,0276	0,0287	0,0293
	Couverture de l'IC	95,1 %	93,8 %	95,7 %
Imputation par la régression à noyau avec indice unidimensionnel proposée	Biais relatif	0,0 %	0,4 %	0,4 %
	E.-T.	0,0275	0,0288	0,0279
	$\widehat{\text{E.-T.}}_{\text{boot}}$	0,0276	0,0288	0,0288
	Couverture de l'IC	95,1 %	92,5 %	91,7 %
Imputation par la régression à noyau avec dépendance à l'égard de la dernière valeur	Biais relatif	0,6 %	1,0 %	0,6 %
	E.-T.	0,0284	0,0310	0,0257
	$\widehat{\text{E.-T.}}_{\text{boot}}$	0,0288	0,0295	0,0263
	Couverture de l'IC	93,7 %	84,2 %	86,2 %
Imputation par la régression linéaire avec valeurs imputées antérieurement traitées comme observées	Biais relatif	0,0 %	1,6 %	0,8 %
	E.-T.	0,0275	0,0261	0,0241
	$\widehat{\text{E.-T.}}_{\text{boot}}$	0,0276	0,0260	0,0246
	Couverture de l'IC	95,1 %	59,7 %	76,0 %
Imputation par la régression linéaire fondée sur les données observées courantes et antérieures	Biais relatif	0,0%	1,6 %	0,8 %
	E.-T.	0,0275	0,0261	0,0242
	$\widehat{\text{E.-T.}}_{\text{boot}}$	0,0276	0,0261	0,0246
	Couverture de l'IC	95,1 %	59,0 %	76,1 %

Bien que la régression à noyau soit asymptotiquement valide, dans la présente étude en simulation, le nombre total de sujets est de 2 000 et, selon le tableau 2, les nombres moyens de points de données utilisés dans la régression à noyau sous les schémas $(t, r) = (4,1)$ et $(4,2)$ sont de 238 et 152, respectivement, ce qui pourrait ne pas suffire pour la régression à noyau et causer de légers biais dans l'imputation. En revanche, la régression linéaire est plus stable et donne de bons résultats pour une taille d'échantillon telle que 152. Bien qu'en théorie, l'imputation par la régression linéaire donne un biais, celui-ci peut être faible quand $E(y_t \mid y_1, ..., y_{t-1})$ est linéaire.

3.2 Application à la SIRD

La SIRD est une enquête annuelle menée auprès d'environ 31 000 entreprises susceptibles de faire de la recherche et du développement. La NSF parraine cette enquête dans le cadre d'un mandat exigeant qu'elle recueille, interprète et analyse des données sur les ressources en sciences et en génie aux États-Unis. L'enquête est menée conjointement par le U.S. Census Bureau et la NSF. Il est demandé aux entreprises qui participent à l'enquête de fournir des renseignements sur leurs dépenses totales en recherche et développement (R-D) durant l'année civile sur laquelle porte l'enquête. Chaque année, la SIRD est menée de manière déterministe auprès de certaines entreprises en les plaçant dans une strate à tirage complet, puisqu'elles représentent un pourcentage important de l'investissement monétaire total en R-D aux États-Unis. Les autres entreprises qui participent à l'enquête sont échantillonnées chaque année en utilisant un plan d'échantillonnage avec probabilité proportionnelle à la taille (PPT). Des mesures longitudinales sont disponibles pour le noyau d'entreprises qui sont échantillonnées avec certitudes et pour les entreprises de la strate à tirage partiel qui se trouvent être sélectionnées chaque année. En vue d'illustrer nos méthodes d'imputation, nous nous limitons à n'examiner que les entreprises qui ont été sélectionnées pour l'enquête chaque année de 2002 à 2005 ($T = 4$), et les entreprises qui ont fourni une réponse en 2002. Pour de la documentation sur la SIRD et des tableaux statistiques détaillés, nous renvoyons le lecteur au document intitulé *Research and Development in Industry: 2005*, qui peut être consulté à l'adresse http://www.nsf.gov/statistics/nsf10319. D'autres renseignements sur la *Business R&D and Innovation Survey* peuvent être consultés en ligne aux adresses http://bhs.dev.econ.census.gov/bhs/brdis/ et http://www.nsf.gov/statistics/srvyindustry/about/brdis/.

Nous divisons les données en deux classes d'imputation. L'une comprend toutes les entreprises contenues dans la strate à tirage complet chacune des quatre années ; la seconde comprend toutes les autres entreprises. Dans chaque classe d'imputation, les données sont de la forme $(\mathbf{y}_i, \boldsymbol{\delta}_i)$, $i = 1, ..., n$, où y_{it} représente les dépenses totales en R-D de l'entreprise i au temps $t = 1$ (2002), 2 (2003), 3 (2004), 4 (2005). Ici, la taille de l'échantillon est $n = 2\,309$ pour la classe des strates à tirage complet et $n = 1\,039$ pour la classe des strates à tirage partiel. La réponse manquante est non monotone et les pourcentages de réponses manquantes pour 2003, 2004 et 2005 étaient de 10,4 %, 14,0 % et 18,8 % pour la classe des strates à tirage complet et de 15,2 %, 20,7 % et 26,0 % pour la classe des strates à tirage partiel.

Le tableau 4 donne les totaux et les erreurs-types estimés en utilisant les méthodes (2) à (9) décrites à la section 3.1. Comme il est discuté à la fin de la section 2.1, dans chacune des méthodes d'imputation proposées, nous utilisons la régression linéaire quand $r + 1 = t$. Les erreurs-types présentées au tableau 4 ont été calculées par la méthode du bootstrap. Le tableau 4 donne aussi les totaux estimés obtenus quand les données manquantes sont remplacées par les valeurs établies par le Census Bureau afin de produire les tableaux de données publiés officiellement (ces tableaux sont disponibles en ligne à l'adresse http://www.nsf.gov/statistics/pubseri.cfm?seri_id=26). La méthode utilisée par le Census Bureau pour traiter les données manquantes en vue de produire ces tableaux de données publiés (que nous appelons « méthode courante ») était l'imputation par le ratio pour les entreprises pour lesquelles des données étaient disponibles pour les années précédentes, en utilisant des cellules d'imputation formées par le type d'industrie ; nous renvoyons le lecteur à Bond (1994) pour d'autres renseignements. Le tableau 4 donne aussi les dépenses totales en R-D estimées en se basant sur les répondants seulement sans correction de la pondération, qui montrent que ne pas tenir compte des données manquantes introduit un biais dans les estimations. Les méthodes (3) à (9) donnent des résultats comparables, vraisemblablement en raison de la forte dépendance linéaire des données, qui fait en sorte que des méthodes biaisées en théorie présentent un biais négligeable. Les totaux estimés fondés sur la méthode courante sont comparables à ceux fondés sur les méthodes proposées pour le cas des strates à tirage complet, mais diffèrent dans le cas des strates à tirage partiel. La méthode combinant la censure et la régression linéaire produit le même écart-type que les méthodes proposées, parce que le nombre de points de données écartés par censure n'est pas trop important. Dans la classe d'imputation des strates à tirage complet, 10 % seulement de l'échantillon présente un schéma de non-réponse intermittente et le pourcentage de cas complets est de 72 %. Dans la classe des strates à tirage partiel, 9 % seulement de l'échantillon présente un schéma de non-réponse intermittente et le pourcentage de cas complets est de 66 %.

Tableau 4
Estimations des dépenses totales en R-D (en milliers) d'après les données de la SIRD pour les années 2002 à 2005. L'erreur-type bootstrap (en milliers) entre parenthèses[1]

Méthode	Strates à tirage complet			Strates à tirage partiel		
	$t = 2$	$t = 3$	$t = 4$	$t = 2$	$t = 3$	$t = 4$
Imputation courante	154 066	156 754	168 015	2 694	2 790	2 782
	-	-	-	-	-	-
Répondants seulement sans correction des pondérations	149 502 (15 907)	148 300 (16 160)	159 822 (17 149)	2 448 (172)	2 553 (193)	2 419 (207)
Répondants seulement avec correction des pondérations	166 924 (17 728)	172 419 (18 720)	196 815 (21 045)	2 887 (199)	3 219 (237)	3 269 (273)
Censure et imputation par la régression linéaire	154 824 (15 888)	159 206 (16 394)	172 631 (17 470)	2 843 (189)	3 079 (208)	3 257 (246)
Imputation par la régression à noyau proposée	154 824 (15 888)	159 394 (16 414)	171 633 (17 603)	2 843 (189)	2 997 (199)	3 161 (290)
Imputation par la régression linéaire proposée	154 824 (15 888)	159 198 (16 383)	172 042 (17 247)	2 843 (189)	3 043 (203)	3 302 (250)
Imputation par la régression à noyau avec indice unidimensionnel proposée	154 824 (15 888)	159 394 (16 414)	171 494 (17 268)	2 843 (189)	2 997 (199)	3 254 (248)
Imputation par la régression à noyau avec dépendance à l'égard de la dernière valeur	154 688 (15 900)	158 768 (16 286)	170 606 (17 234)	2 831 (188)	2 983 (197)	3 177 (240)
Imputation par la régression linéaire, valeurs imputées antérieures traitées comme observées	154 824 (15 888)	159 401 (16 390)	172 600 (17 306)	2 843 (189)	3 098 (208)	3 257 (236)
Imputation par la régression linéaire fondée sur les données observées courantes et antérieures	154 824 (15 888)	160 205 (16 534)	172 452 (17 209)	2 843 (189)	3 168 (233)	3 273 (254)

[1] Avertissement : Les valeurs du tableau 4 ne représentent pas nécessairement des estimations nationales parce que nous avons appliqué certaines contraintes aux données afin de respecter notre cadre d'étude.

3.3 Résultats de simulation fondés sur la population de la SIRD

Une étude en simulation supplémentaire a été réalisée en utilisant une population créée à partir des données de la SIRD. La simulation a été exécutée indépendamment pour les classes d'imputation des strates à tirage complet et des strates à tirage partiel. Pour créer la population, nous partons des données de la SIRD en imputant les valeurs manquantes par la méthode courante utilisée pour la SIRD. Soit $\boldsymbol{\delta}_i$ le vecteur des indicateurs de réponse observés pour l'entreprise i et $\tilde{\mathbf{y}}_i$ le vecteur des valeurs observées ou imputées des dépenses totales en R-D de l'entreprise i au cours du temps, $i = 1, ..., n$. Pour la simulation, nous effectuons l'échantillonnage à partir d'une population fondée sur $\{(\tilde{\mathbf{y}}_i, \boldsymbol{\delta}_i), i = 1, ..., n\}$ comme il suit. Nous commençons par tirer un échantillon de taille n avec remise de $\tilde{\mathbf{y}}_1, ..., \tilde{\mathbf{y}}_n$, puis nous ajoutons un bruit aléatoire normal indépendant de moyenne 0 et d'écart-type 500 à chaque composante de chacun des vecteurs échantillonnés. Toute valeur négative résultante est remplacée par 0. Nous désignons ces dépenses totales en R-D simulées par $\mathbf{y}_1^*, ..., \mathbf{y}_n^*$, où n est défini de la même façon qu'à la section 3.2. Nous désignons les indicateurs de réponse simulés par $\boldsymbol{\delta}_1^*, ..., \boldsymbol{\delta}_n^*$.

Pour tout i et chaque $t = 2, 3, 4$, les δ_{it}^* sont des variables aléatoires binaires avec

$$P(\delta_{it}^* = 1 \mid y_{i1}^*, ..., y_{i,t-1}^*) = \frac{\exp(\beta_0^{(t)} + \beta_1^{(t)} y_{i,1}^* + \cdots + \beta_{t-1}^{(t)} y_{i,t-1}^*)}{1 + \exp(\beta_0^{(t)} + \beta_1^{(t)} y_{i,1}^* + \cdots + \beta_{t-1}^{(t)} y_{i,t-1}^*)}.$$

Les coefficients $\beta_0^{(t)}, \beta_1^{(t)}, ..., \beta_{t-1}^{(t)}$ sont fixes tout au long de la simulation et sont obtenus comme étant les coefficients estimés d'après un ajustement initial d'une régression logistique de δ_{it} sur $(\tilde{y}_{i1}, ..., \tilde{y}_{i,t-1})$ pour $i = 1, ..., n$.

Le tableau 5 donne les résultats des simulations pour les estimateurs du total fondés sur 1 000 exécutions et les méthodes (1) à (9) décrites à la section 3.1, où les quantités qui figurent dans le tableau sont définies à la section 3.1. Pour calculer le biais relatif, nous avons obtenu la valeur réelle du total grâce à une exécution préliminaire du modèle de simulation. Plusieurs conclusions tirées de la simulation de la population normale de la section 3.1 s'appliquent dans les conditions décrites ici. Voici un résumé de certaines constatations supplémentaires.

1. Contrairement aux observations dans les conditions de simulation d'une population normale, le total estimé fondé sur la censure et la régression linéaire possède un écart-type comparable à celui obtenu par les méthodes d'imputation proposées. Il en est ainsi parce que le nombre de points de données écartés par censure est faible dans le cas qui nous occupe. Les probabilités d'un schéma de réponse intermittente sont de 17 % et 19 % pour les classes des strates à tirage complet et à tirage partiel, respectivement. Dans la simulation de la population normale, ces probabilités s'approchaient de 50 % comme le montre le tableau 2.
2. Toutes les méthodes d'imputation proposées donnent des résultats relativement semblables. Comme nous l'avons mentionné antérieurement, l'imputation par la régression linéaire est généralement biaisée en théorie. Cependant, le biais est faible en raison de la forte dépendance linéaire des données.
3. La méthode (7) ne donne pas de bons résultats pour $t \geq 3$ pour les strates à tirage partiel, parce que l'hypothèse que la propension à répondre dépend de la dernière valeur observée ne tient pas.
4. Les méthodes (8) et (9) donnent de bons résultats, de nouveau en raison de la forte dépendance linéaire de données. Même si ces méthodes utilisent un plus grand nombre de données observées dans l'imputation par la régression, les résultats sont comparables à ceux de la méthode par la régression linéaire proposée.

Tableau 5
Résultats de simulation pour les estimations du total (en milliers) pour la population fondée sur la SIRD

Méthode	Quantité	Strates à tirage complet			Strates à tirage partiel		
		$t = 2$	$t = 3$	$t = 4$	$t = 2$	$t = 3$	$t = 4$
Données complètes	Biais relatif	0 %	0,1 %	0,1 %	0,2 %	0,0 %	0,4 %
	É.-T.	15 541	16 045	16 947	184	203	224
	$\widehat{\text{É.-T.}}_{\text{boot}}$	15 654	15 994	16 941	186	201	218
	Couverture de l'IC	94,0 %	94,0 %	94,3 %	94,3 %	93,7 %	93,9 %
Répondants seulement avec poids corrigés	Biais relatif	5 %	6,3 %	11,6 %	-1,1 %	1,1 %	-2,7 %
	É.-T.	16 870	17 858	20 032	191	220	244
	$\widehat{\text{É.-T.}}_{\text{boot}}$	16 917	17 915	20 048	192	219	234
	Couverture de l'IC	94,8 %	94,8 %	87,3 %	93,2 %	94,5 %	89,8 %
Censure et imputation par la régression linéaire	Biais relatif	0 %	0,4 %	0,5 %	0,4 %	0,1 %	-0,4 %
	É.-T.	15 582	16 272	17 247	191	214	238
	$\widehat{\text{É.-T.}}_{\text{boot}}$	15 654	16 145	17 195	194	214	236
	Couverture de l'IC	93,8 %	93,5 %	94,2 %	94,8 %	94,0 %	93,7 %
Imputation par la régression à noyau proposée	Biais relatif	0 %	0,2 %	-0,1 %	0,4 %	-0,3 %	-0,3 %
	É.-T.	15 582	16 130	17 098	191	205	246
	$\widehat{\text{É.-T.}}_{\text{boot}}$	15 654	16 072	17 231	194	204	262
	Couverture de l'IC	93,8 %	93,5 %	94,2 %	94,8 %	93,4 %	93,7 %
Imputation par la régression linéaire proposée	Biais relatif	0 %	0,2 %	0,0 %	0,4 %	0,0 %	-0,5 %
	É.-T.	15 582	16 130	16 955	191	206	229
	$\widehat{\text{É.-T.}}_{\text{boot}}$	15 654	16 072	16 964	194	206	224
	Couverture de l'IC	93,8 %	93,5 %	94,2 %	94,8 %	94,0 %	93,7 %
Imputation par la régression à noyau avec indice unidimensionnel proposée	Biais relatif	0 %	0,2 %	-0,1 %	0,4 %	-0,3 %	-0,9 %
	É.-T.	15 582	16 130	16 957	191	205	227
	$\widehat{\text{É.-T.}}_{\text{boot}}$	15 654	16 072	16 965	194	204	220
	Couverture de l'IC	93,8 %	93,5 %	94,3 %	94,8 %	93,4 %	93,1 %
Imputation par la régression à noyau avec dépendance à l'égard de la dernière valeur	Biais relatif	0 %	0,1 %	-0,3 %	0,0 %	-0,7 %	-0,7 %
	É.-T.	15 565	16 019	16 990	184	204	242
	$\widehat{\text{É.-T.}}_{\text{boot}}$	15 635	16 003	16 983	187	202	230
	Couverture de l'IC	93,8 %	93,7 %	94,0 %	93,9 %	92,7 %	91,1 %
Imputation par la régression linéaire avec valeurs imputées antérieures traitées comme observées	Biais relatif	0 %	0,2 %	0,0 %	0,4 %	0,6 %	-0,6 %
	É.-T.	15 582	16 120	16 952	191	210	231
	$\widehat{\text{É.-T.}}_{\text{boot}}$	15 654	16 065	16 954	194	210	225
	Couverture de l'IC	93,8 %	93,6 %	94,3 %	94,8 %	93,8 %	92,8 %
Imputation par la régression linéaire fondée sur les données observées courantes et antérieures	Biais relatif	0 %	0,2 %	0,0 %	0,4 %	0,6 %	-0,6 %
	É.-T.	15 582	16 117	16 945	191	213	241
	$\widehat{\text{É.-T.}}_{\text{boot}}$	15 654	16 062	16 954	194	211	254
	Couverture de l'IC	93,8 %	93,5 %	94,3 %	94,8 %	93,6 %	93,7 %

4. Conclusion

Nous considérons l'une des variables d'une étude longitudinale présentant une non-réponse non monotone. Sous l'hypothèse que la propension à répondre dépend des valeurs observées et non observées antérieures de la variable étudiée, nous proposons plusieurs méthodes d'imputation qui donnent des estimateurs sans biais ou presque sans biais du total ou de la moyenne de la variable étudiée à un point donné dans le temps. Nos méthodes ne requièrent l'ajustement d'aucun modèle paramétrique sur la distribution conjointe des variables aux divers points dans le temps ni les propensions à répondre. Elles sont fondées sur des modèles de régression sous divers schémas de non-réponse dérivés des propensions à répondre dépendantes des données antérieures. Trois méthodes de régression sont adoptées, à savoir la régression linéaire, la régression à noyau, et la régression à noyau avec indice unidimensionnel. La méthode d'imputation fondée sur la régression à noyau est asymptotiquement valide, mais elle requiert un grand nombre d'observations pour chaque schéma de non-réponse. L'imputation fondée sur la régression linéaire est asymptotiquement biaisée quand la relation linéaire n'est pas vérifiée, mais elle est plus stable et, par conséquent, peut demeurer meilleure que les méthodes fondées sur la régression à noyau.

La méthode de censure, qui consiste à écarter toutes les données observées auprès d'un sujet après la première valeur manquante chez ce dernier, peut donner de bons résultats quand le nombre de données écartées est faible ; sinon, elle peut être très inefficace, surtout quand T est grand. Pour l'analyse des données de la SIRD aux sections 3.2 et 3.3, la censure donne des résultats comparables à la méthode d'imputation par la régression linéaire proposée. Cependant, ces résultats sont fondés sur les données recueillies pour quatre années seulement et la censure pourrait aboutir à des estimateurs inefficaces si l'on considère les données recueillies pour un plus grand nombre d'années. Dans les applications, il pourrait être utile de comparer les estimateurs fondés sur la censure à ceux fondés sur les méthodes proposées.

Les estimateurs qui s'appuient sur les méthodes d'imputation par la régression linéaire (8) et (9) décrites à la section 3.1 sont en général asymptotiquement biaisés. Même s'ils donnent de bons résultats dans l'étude en simulation fondée sur la population de la SIRD, leurs propriétés sont médiocres dans les conditions des simulations de la section 3.1, mais elles sont bonnes sous l'imputation par la régression linéaire proposée.

Les résultats de la section 2 peuvent être étendus à la situation où chaque unité échantillonnée possède au temps t une covariable observée $\mathbf{x}_t$ sans valeur manquante. L'hypothèse (1) peut être modifiée comme il suit afin d'inclure les covariables :

$$P(\delta_t = 1 \mid \mathbf{y}, \mathbf{X}, \delta_1, ..., \delta_{t-1}, \delta_{t+1}, ..., \delta_T) = P(\delta_t = 1 \mid y_1, ..., y_{t-1}, \mathbf{X}, \delta_1, ..., \delta_{t-1}), \quad t = 2,...,T,$$

où $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_T)$. Les composantes manquantes de $\mathbf{y}_i$ peuvent être imputées en utilisant l'une des procédures des sections 2.1 à 2.3 en remplaçant $(y_{i1}, ..., y_{ir})$ par $(y_{i1}, ..., y_{ir}, \mathbf{X}_i)$. Après avoir imputé toutes les valeurs manquantes, nous pouvons aussi estimer la relation entre $\mathbf{y}$ et $\mathbf{X}$ en utilisant certaines approches d'usage répandu, comme celle des équations d'estimation généralisées. Certains détails figurent dans Xu (2007).

Nous supposons implicitement tout au long de l'exposé que les valeurs de y sont des variables continues sans aucune restriction. Quand les valeurs de y possèdent un ordre particulier ou sont des valeurs entières, les méthodes d'imputation par la régression proposée ne conviennent manifestement pas. De nouvelles méthodes pour cette situation doivent être élaborées.

Remerciements

Nous remercions Katherine Jenny Thompson et David L. Kinyon, tous deux du U.S. Census Bureau, ainsi que deux examinateurs et le rédacteur associé de leurs nombreux commentaires constructifs au sujet de l'article. L'étude a été financée en partie par une subvention de la NSF. Le présent article est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche en cours et de favoriser la discussion. Les opinions exprimées sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau.

Annexe

Preuve de (2) à (3). Soit $L(\xi)$ la distribution de ξ et $L(\xi \mid \zeta)$ la distribution conditionnelle de ξ sachant ζ. Soit $\mathbf{y}_t = (y_1, ..., y_t)$ et $\boldsymbol{\delta}_t = (\delta_1, ..., \delta_t)$. Alors, (2) ainsi que (3) découlent de $L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_t) = L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-1}) = L(\mathbf{y}_t, \boldsymbol{\delta}_{t-1}) / L(\mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-1}) = [L(\delta_{t-1} \mid \mathbf{y}_t, \boldsymbol{\delta}_{t-2}) / L(\delta_{t-1} \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-2})]$ $L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-2}) = L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-2}) = L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-3}) = \cdots = L(y_t \mid \mathbf{y}_{t-1})$, où les première et troisième égalités découlent de l'hypothèse (1).

Preuve de (5). En utilisant la même notation que dans la preuve de (2) et en posant que $\Delta_r = 1$ est l'indicateur de $\delta_1 = \cdots = \delta_r = 1$, nous avons $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = [L(\delta_{r+1} = 0 \mid y_t, \mathbf{y}_r, \Delta_r = 1, \delta_t = 0) / L(\delta_{r+1} = 0 \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)] L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$, qui est égal à $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$ en vertu de (1). De même, nous pouvons montrer que $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) = L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$. D'où, $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 1, \delta_t = 0)$ et le résultat (5) s'ensuit.

Un exemple dans le quel (4) n'est pas vérifiée. Pour montrer que (4) n'est pas vérifiée en général, il nou s suffit de donner un contre-exemple. Considérons $T = 3$. Posons que (y_1, y_2, y_3) suivent conjointement une loi normale caractérisée par $E(y_t) = 0$, $\text{var}(y_t) = 1$, $t = 1, 2, 3$, $\text{cov}(y_1, y_2) = \text{cov}(y_1, y_3) = \rho$, et $\text{cov}(y_2, y_3) = \rho^2$, où $\rho \neq 0$ est un paramètre. Supposons que y_1 est toujours observée et que $P(\delta_t = 0 \mid y_{t-1}) = \Phi(a_{t-1} + b_{t-1} y_{t-1})$, $t = 2, 3$, où a_t et b_t sont des paramètres, et Φ est la fonction de répartition de la loi normale centrée réduite. Alors $E(y_3 \mid y_2, y_1) = \rho y_2$, $E(y_2 \mid y_1) = \rho y_1$, et $E(y_3 \mid y_1) = \rho^2 y_1$. Notons que

$$
\begin{aligned}
&E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) = E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = 1) \\
&\quad = E(y_3 \mid y_1, \delta_3 = 0) \\
&\quad = \int y_3 L(y_3 \mid y_1, \delta_3 = 0)\, d y_3 \\
&\quad = \int y_3 \int L(y_3 \mid y_1, y_2, \delta_3 = 0) L(y_2 \mid y_1, \delta_3 = 0)\, d y_2\, d y_3 \\
&\quad = \iint y_3 L(y_3 \mid y_1, y_2) L(y_2 \mid y_1, \delta_3 = 0)\, d y_2\, d y_3 \\
&\quad = \int \Big(\int y_3 L(y_3 \mid y_2)\, d y_3 \Big) L(y_2 \mid y_1, \delta_3 = 0)\, d y_2 \\
&\quad = \rho \int y_2 L(y_2 \mid y_1, \delta_3 = 0)\, d y_2 \\
&\quad = \frac{\rho \int y_2 P(\delta_3 = 0 \mid y_2) L(y_2 \mid y_1)\, d y_2}{\int P(\delta_3 = 0 \mid y_2) L(y_2 \mid y_1)\, d y_2} \\
&\quad = \frac{\rho \int y_2 \Phi(a_2 + b_2 y_2) L(y_2 \mid y_1)\, d y_2}{\int \Phi(a_2 + b_2 y_2) L(y_2 \mid y_1)\, d y_2},
\end{aligned}
$$

où la première égalité est vérifiée parce que y_1 est toujours observée, la deuxième égalité est vérifiée parce que, sous (1), δ_2 et y_3 sont indépendants sachant y_1. Le dénominateur de l'expression précédente est égal à

$$
h(y_1) = \Phi\left(\frac{a_2 + b_2 \rho y_1}{\sqrt{1 + b_2^2 (1 - \rho^2)}} \right).
$$

En intégrant par parties, nous obtenons que

$$
\begin{aligned}
g(y_1) &= \int (y_2 - \rho y_1) \Phi(a_2 + b_2 y_2) L(y_2 \mid y_1)\, d y_2 \\
&= b_2 (1 - \rho^2) \int \Phi'(a_2 + b_2 y_2) L(y_2 \mid y_1)\, d y_2 \\
&= \frac{b_2^2 (1 - \rho^2)}{2\pi \sqrt{1 - \rho^2}} \int \exp\left\{ -\frac{(a_2 + b_2 \rho y_2)^2}{2} - \frac{(y_2 - \rho y_1)^2}{2(1 - \rho^2)} \right\} d y_2 \\
&= \frac{b_2 (1 - \rho^2)}{2\pi [1 + b_2^2 (1 - \rho^2)]} \exp\left\{ -\frac{(a_2 + b_2 \rho y_1)^2}{2[1 + b_2^2 (1 - \rho^2)]} \right\}.
\end{aligned}
$$

D'où,

$$
E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) = \rho^2 y_1 + \rho \frac{g(y_1)}{h(y_1)}. \quad (10)
$$

Cependant,

$$
\begin{aligned}
E(y_3 \mid y_1, \delta_1 = \delta_2 = 1) &= E(y_3 \mid y_1, \delta_1 = 1) \\
&= E(y_3 \mid y_1) = \rho^2 y_1.
\end{aligned}
$$

Cela montre que (4) n'est pas vérifiée dans ce cas particulier.

Preuve de (8). En utilisant la notation de la preuve de (2) et de (3), et en représentant le vecteur $(y_1, ..., y_{r-1}, y_{r+1}, ..., y_{t-1})$ de dimension $(t - 2)$ par $\mathbf{u}_{t,r}$, nous obtenons

$$
\begin{aligned}
&L(\delta_{r+1} = 1 \mid y_t, z_r, \Delta_r = 1, \delta_t = 0) \\
&\quad = \int L(\delta_{r+1} = 1 \mid y_t, z_r, \mathbf{u}_{t,r}, \Delta_r = 1, \delta_t = 0) \\
&\qquad L(\mathbf{u}_{t,r} \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)\, d \mathbf{u}_{t,r} \\
&\quad = \int L(\delta_{r+1} = 1 \mid y_1, ..., y_r, \Delta_r = 1) \\
&\qquad L(\mathbf{u}_{t,r} \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)\, d \mathbf{u}_{t,r} \\
&\quad = \int L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1) \\
&\qquad L(\mathbf{u}_{t,r} \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)\, d \mathbf{u}_{t,r} \\
&\quad = L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1) \\
&\qquad \int L(\mathbf{u}_{t,r} \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)\, d \mathbf{u}_{t,r} \\
&\quad = L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1),
\end{aligned}
$$

où la deuxième égalité découle de l'hypothèse (1) et du fait qu'il existe une fonction injective entre $(z_r, \mathbf{u}_{t,r})$ et $(y_1, ..., y_{t-1})$, et la troisième égalité découle de l'hypothèse (7). De même, $L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1, \delta_t = 0) = L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1)$ et, donc, $L(\delta_{r+1} = 1 \mid y_t, z_r, \Delta_r = 1, \delta_t = 0) = L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1, \delta_t = 0)$. Alors,

$$
\begin{aligned}
&L(y_t \mid z_r, \Delta_{r+1} = 1, \delta_t = 0) \\
&\quad = \frac{L(y_t, z_r, \Delta_{r+1} = 1, \delta_t = 0)}{L(z_r, \Delta_{r+1} = 1, \delta_t = 0)} \\
&\quad = \frac{L(\delta_{r+1} = 1 \mid y_t, z_r, \Delta_r = 1, \delta_t = 0)}{L(\delta_{r+1} = 1 \mid z_r, \Delta_r = 1, \delta_t = 0)} \frac{L(y_t, z_r, \Delta_r = 1, \delta_t = 0)}{L(z_r, \Delta_r = 1, \delta_t = 0)} \\
&\quad = L(y_t \mid z_r, \Delta_r = 1, \delta_t = 0).
\end{aligned}
$$

De même, $L(y_t \mid z_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t \mid z_r, \Delta_r = 1, \delta_t = 0)$. D'où, $L(y_t \mid z_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t \mid z_r, \Delta_{r+1} = 1, \delta_t = 0)$ et le résultat (8) s'ensuit.

Bibliographie

Bond, D. (1994). An evaluation of imputation methods for the Survey of Industrial Research and De velopment. *U.S. Bureau of the Census, Economic Statistical Methods and Programming Division Report Series*. 9404. Washington, DC.

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

Diggle, P., et Ken ward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society*, Séries C (*Applied Statistics*), 43, 49-93.

Duan, N., et Li, K.C. (1991). Sliced regression: A link-free regression method. *The Annals of Statistics*, 19, 505-530.

Efron, B. (1979). Bootstrap methods: Another look at the jac kknife. *Annals of Statistics*, 7, 1-26.

Kalton, G., et Kasprzyk, D. (1986). Le tra itement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1, 1-17.

Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Little, R.J., et Rubin, D.B. (2002). Statistical Analysis with Missing Data, deuxième édition. New York : John Wiley & Sons, Inc.

National Science Foundation, Division of Science Resources Statistics (2010). *Research and Development in Industry*: *2005*. Detailed Statistical Tables. Disponible au http ://www.nsf.gov/ statistics/nsf10319/.

Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92, 1320-1329.

R Development Core Team (2009). A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0.

Robins, J.M., et Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122-129.

Shao, J., et Si tter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Troxel, A.B., Harrington, D.P. et Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, 47, 425-438.

Troxel, A.B., Lipsitz, S.R. et Harrington, D.P. (1998). M arginal models for the analysis of longitudinal measurements with non-ignorable non-monotone missing data. *Biometrika*, 85, 661-672.

Vansteelandt, S., Rotnitzky, A. et Robins, J.M. (2007). Estimation of regression models for the mean of re peated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94, 841-860.

Xu, J. (2007). Methods for intermittent missing responses in longitudinal data. Thèse de doc torat, Department of Statistics, University of Wisconsin-Madison.

Xu, J., Shao, J., Palta, M. et Wang, L. (2008). Imputation pour les non-réponses non monotones dépendant de la dernière valeur dans les enquêtes longitudinales. *Techniques d'enquête*, 34, 2, 169-179.

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 171-180
Statistique Canada, N° 12-001-X au catalogue

Théorie concernant les estimateurs ajustés sur le score de propension dans les sondages

Jae Kwang Kim et Minsun Kim Riddles [1]

Résumé

La méthode d'ajustement sur le score de propension est souvent adoptée pour traiter le biais de sélection dans les sondages, y compris la non-réponse totale et le sous-dénombrement. Le score de propension est calculé en se servant de variables auxiliaires observées dans tout l'échantillon. Nous discutons de certaines propriétés asymptotiques des estimateurs ajustés sur le score de propension et dérivons des estimateurs optimaux fondés sur un modèle de régression pour la population finie. Un estimateur ajusté sur le score de propension optimal peut être réalisé en se servant d'un modèle de score de propension augmenté. Nous discutons de l'estimation de la variance et présentons les résultats de deux études par simulation.

Mots clés : Calage ; données manquantes ; non-réponse ; pondération.

1. Introduction

Considérons une population finie de taille N, où N est connu. Pour chaque unité i, y_i est la variable étudiée et $\mathbf{x}_i$ est le vecteur de dimension q de variables auxiliaires. Le paramètre d'intérêt est la moyenne de population finie de la variable étudiée, $\theta = N^{-1}\sum_{i=1}^{N} y_i$. Supposons que la population finie $\mathcal{F}_N = \{(\mathbf{x}_1', y_1), (\mathbf{x}_2', y_2), ..., (\mathbf{x}_N', y_N)\}$ est un échantillon aléatoire de taille N tiré d'une loi de superpopulation $F(\mathbf{x}, y)$. Supposons qu'un échantillon de taille n est tiré de la population finie selon un plan d'échantillonnage probabiliste. Soit $w_i = \pi_i^{-1}$ le poids d'échantillonnage, où π_i est la probabilité d'inclusion de premier ordre de l'unité i obtenue d'après le plan d'échantillonnage probabiliste. Sous réponse complète, la moyenne de population finie peut être estimée par l'estimateur d'Horvitz-Thompson (HT), $\hat{\theta}_{\mathrm{HT}} = N^{-1}\sum_{i\in A} w_i y_i$, où A est l'ensemble d'indices qui figurent dans l'échantillon.

En présence de données manquantes, l'estimateur HT $\hat{\theta}_{\mathrm{HT}}$ ne peut pas être calculé. Soit r la variable indicatrice de réponse, qui prend la valeur 1 si y est observée et la valeur 0 autrement. Conceptuellement, comme en ont discuté Fay (1992), Shao et Steel (1999), ainsi que Kim et Rao (2009), l'indicateur de réponse peut être étendu à la population entière sous la forme $\mathcal{R}_N = \{r_1, r_2, ..., r_N\}$, où r_i est une réalisation de la variable aléatoire r. L'estimateur sous cas complets (CC) $\hat{\theta}_{\mathrm{CC}} = \sum_{i\in A} w_i r_i y_i / \sum_{i\in A} w_i r_i$ converge alors en probabilité vers $E(Y \mid r = 1)$. À moins que le mécanisme de réponse soit tel que les réponses manquent entièrement au hasard en ce sens que $E(Y \mid r = 1) = E(Y)$, l'estimateur CC présente un biais. Pour corriger ce biais, si la probabilité de réponse

$$p(\mathbf{x}, y) = \Pr(r = 1 \mid \mathbf{x}, y) \tag{1}$$

est connue, on peut utiliser l'estimateur CC pondéré $\hat{\theta}_{\mathrm{CCP}} = N^{-1}\sum_{i\in A} w_i r_i y_i / p(\mathbf{x}_i, y_i)$ pour estimer θ. Il convient de souligner que $\hat{\theta}_{\mathrm{CCP}}$ est sans biais parce que $E\{\sum_{i\in A} w_i r_i y_i / p(\mathbf{x}_i, y_i) \mid \mathcal{F}_N\} = E\{\sum_{i=1}^{N} r_i y_i / p(\mathbf{x}_i, y_i) \mid \mathcal{F}_N\} = \sum_{i=1}^{N} y_i$.

Si la probabilité de réponse (1) est inconnue, on peut postuler pour cette dernière un modèle paramétrique $p(\mathbf{x}, y; \boldsymbol{\phi})$ indicé par $\boldsymbol{\phi} \in \Omega$ tel que $p(\mathbf{x}, y) = p(\mathbf{x}, y; \boldsymbol{\phi}_0)$ pour un certain $\boldsymbol{\phi}_0 \in \Omega$. Nous supposons qu'il existe un estimateur convergeant à la vitesse $\sqrt{n}$ $\hat{\boldsymbol{\phi}}$ de $\boldsymbol{\phi}_0$ tel que

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) = O_p(1), \tag{2}$$

où $g_n = O_p(1)$ indique que g_n est borné en probabilité. En utilisant $\hat{\boldsymbol{\phi}}$, nous pouvons obtenir la probabilité de réponse estimée par $\hat{p}_i = p(\mathbf{x}_i, y_i; \hat{\boldsymbol{\phi}})$, qui est souvent appelée score de propension (Rosenbaum et Rubin 1983). L'estimateur ajusté sur le score de propension (ASP) peut être construit comme

$$\hat{\theta}_{\mathrm{ASP}} = \frac{1}{N}\sum_{i\in A} w_i \frac{r_i}{\hat{p}_i} y_i. \tag{3}$$

L'estimateur ajusté sur le score de propension (3) est d'usage très répandu. De nombreux programmes d'enquête l'utilisent pour réduire le biais de non-réponse (Fuller, Loughin et Baker, 1994 ; Rizzo, Kalton et Brick, 1996). Rosenbaum et Rubin (1983) et Rosenbaum (1987) ont proposé d'utiliser l'approche d'ajustement sur le score de propension pour estimer les effets du traitement dans les études observationnelles. Little (1988) a passé en revue les méthodes d'ajustement sur le score de propension pour le traitement de la non-réponse totale dans les sondages. Duncan et Stasny (2001) ont utilisé l'approche d'ajustement sur le score de propension pour contrôler le biais de couverture

1. Jae Kwang Kim et Minsun Kim Riddles, Department of Statistics, Iowa State University, Ames (IA), États-Unis, 50011. Courriel : jkim@iastate.edu.

dans les enquêtes téléphoniques. Folsom (1991) et Iannacchione, Milne et Folsom (1991) ont utilisé un modèle de régression logistique pour estimer la probabilité de réponse. Lee (2006) a appliqué la méthode d'ajustement sur le score de propension à une enquête en ligne auprès d'un panel de volontaires. Durrant et Skinner (2006) ont utilisé l'approche d'ajustement sur le score de propension pour traiter l'erreur de mesure.

Malgré la popularité des estimateurs ajustés sur le score de propension, peu d'attention a été accordée à leurs propriétés asymptotiques dans la littérature sur les sondages. Kim et Kim (2007) ont utilisé un développement en série de Taylor pour obtenir la moyenne et la variance asymptotique des estimateurs ajustés sur le score de propension et discuté de l'estimation de leur variance. Da Silva et Opsomer (2006), ainsi que Da Silva et Opsomer (2009) ont considéré des méthodes non paramétriques pour obtenir des estimateurs ajustés sur le score de propension.

Dans le présent article, nous discutons des estimateurs ajustés sur le score de propension optimaux appartenant à la classe d'estimateurs de la forme (3) qui utilisent un estimateur $\hat{\phi}$ convergeant à la vitesse $\sqrt{n}$. Ces estimateurs sont asymptotiquement sans biais pour θ. La recherche d'estimateurs ajustés sur le score de propension à variance minimale parmi cette classe particulière d'estimateurs ajustés sur le score de propension est l'un des principaux sujets sur lesquels porte le présent article.

À la section 2, nous présentons les résultats principaux. À la section 3, nous proposons un estimateur ajusté sur le score de propension optimal utilisant un modèle de score de propension augmenté. À la section 4, nous discutons de l'estimation de la variance de l'estimateur proposé. À la section 5, nous présentons les résultats de deux études par simulation et à la section 6, nous formulons nos conclusions.

2. Résultats principaux

À la présente section, nous discutons de certaines propriétés asymptotiques des estimateurs ajustés sur le score de propension. Nous supposons que le mécanisme de réponse ne dépend pas de y, donc que

$$\Pr(r = 1 \mid \mathbf{x}, y) = \Pr(r = 1 \mid \mathbf{x}) = p(\mathbf{x}; \boldsymbol{\phi}_0) \qquad (4)$$

pour un certain vecteur $\boldsymbol{\phi}_0$ inconnu. La première égalité implique que les données manquent au hasard (MAR pour *missing-at-random*), car nous observons toujours $\mathbf{x}$ dans l'échantillon. Notons que la condition MAR fait partie des hypothèses du modèle de population. Dans la seconde égalité, nous supposons en outre que le mécanisme de réponse est connu jusqu'à un paramètre $\boldsymbol{\phi}_0$ inconnu. Le mécanisme de réponse est légèrement différent de celui considéré par Kim et Kim (2007), qui supposent qu'il a lieu sous un échantillonnage à deux phases classique et dépend de l'échantillon réalisé :

$$\Pr(r = 1 \mid \mathbf{x}, y, I = 1) = \Pr(r = 1 \mid \mathbf{x}, I = 1) = p(\mathbf{x}; \boldsymbol{\phi}_A^0). \quad (5)$$

Ici, I est la fonction indicatrice d'échantillonnage définie sur l'ensemble de la population. Autrement dit, $I_i = 1$ si $i \in A$ et $I_i = 0$ autrement. À moins que le plan d'échantillonnage soit non informatif en ce sens que les probabilités de sélection de l'échantillon sont corrélées à l'indicateur de réponse, même après conditionnement sur les variables auxiliaires (Pfeffermann, Krieger et Rinott, 1998), les deux mécanismes de réponse, (4) et (5), sont différents. Dans les sondages, l'hypothèse (4) est plus appropriée parce que la décision de répondre ou non à un sondage est laissée à la discrétion de la personne interrogée. Ici, la variable indicatrice de réponse r_i est définie sur l'ensemble de la population, comme il est discuté à la section 1.

Nous considérons une classe d'estimateurs convergeant à la vitesse $\sqrt{n}$ de $\boldsymbol{\phi}_0$ dans (4). En particulier, nous considérons une classe d'estimateurs qui peuvent s'écrire comme une solution de

$$\hat{\mathbf{U}}_h(\boldsymbol{\phi}) \equiv \sum_{i \in A} w_i \{r_i - p_i(\boldsymbol{\phi})\} \mathbf{h}_i(\boldsymbol{\phi}) = \mathbf{0}, \qquad (6)$$

où $p_i(\boldsymbol{\phi}) = p(\mathbf{x}_i; \boldsymbol{\phi})$ pour une certaine fonction $\mathbf{h}_i(\boldsymbol{\phi}) = \mathbf{h}(\mathbf{x}_i; \boldsymbol{\phi})$, une fonction lisse de $\mathbf{x}_i$ et du paramètre $\boldsymbol{\phi}$. Donc, la solution de (6) peut s'écrire comme étant $\hat{\boldsymbol{\phi}}_h$, qui dépend du choix de $\mathbf{h}_i(\boldsymbol{\phi})$. Toute solution $\hat{\boldsymbol{\phi}}_h$ de (6) est convergente pour $\boldsymbol{\phi}_0$ dans (4) parce que $E\{\hat{\mathbf{U}}_h(\boldsymbol{\phi}_0) | \mathcal{F}_N\} = E[\sum_{i=1}^N \{r_i - p_i(\boldsymbol{\phi}_0)\} \mathbf{h}_i(\boldsymbol{\phi}_0) \mid \mathcal{F}_N]$ est nulle sous le mécanisme de réponse donné en (4). Si nous laissons tomber les poids d'échantillonnage w_i dans (6), le paramètre estimé $\hat{\boldsymbol{\phi}}_h$ est convergent pour $\boldsymbol{\phi}_A^0$ dans (5) et l'estimateur ajusté sur le score de propension résultant est convergent uniquement si le plan d'échantillonnage est non informatif. Les estimateurs ajustés sur le score de propension obtenu à partir de (6) en utilisant les poids d'échantillonnage sont convergents que le plan d'échantillonnage soit ou ne soit pas non informatif. Selon Chamberlain (1987), tout estimateur convergeant à la vitesse $\sqrt{n}$ de $\boldsymbol{\phi}_0$ dans (4) peut s'écrire sous la forme d'une solution de (6). Donc, le choix de $\mathbf{h}_i(\boldsymbol{\phi})$ dans (6) détermine l'efficacité de l'estimateur ajusté sur le score de propension résultant.

Soit $\hat{\theta}_{\mathrm{ASP},h}$ l'estimateur ajusté sur le score de propension donné par (3) en utilisant $\hat{p}_i = p_i(\hat{\boldsymbol{\phi}}_h)$ où $\hat{\boldsymbol{\phi}}_h$ est la solution de (6). Pour discuter des propriétés asymptotiques de $\hat{\theta}_{\mathrm{ASP},h}$, supposons que nous avons une suite de populations finies et d'échantillons, comme dans Isaki et Fuller (1982), telle que $\sum_{i \in A} w_i \mathbf{u}_i - \sum_{i=1}^N \mathbf{u}_i = O_p(n^{-1/2}N)$ pour toute caractéristique de la population $\mathbf{u}_i$ avec les moments d'ordre quatre bornés. Nous supposons aussi que les poids

d'échantillonnage sont bornés uniformément. Autrement dit, $K_1 < N^{-1}nw_i < K_2$ pour tout i uniformément dans n, où K_1 et K_2 sont des constantes données. En outre, nous supposons qu'existent les conditions de régularité suivantes :

[C1] Le mécanisme de réponse satisfait (4), où $p(\mathbf{x}; \boldsymbol{\phi})$ est continue en $\boldsymbol{\phi}$ avec les dérivées première et seconde continues dans un ensemble ouvert contenant $\boldsymbol{\phi}_0$. Les réponses sont indépendantes dans le sens où $\mathrm{Cov}(r_i, r_j \mid \mathbf{x}) = 0$ pour $i \neq j$. En outre, $p(\mathbf{x}_i; \boldsymbol{\phi}) > c$ pour tout i pour une certaine constante donnée $c > 0$.

[C2] La solution de (6) existe et est unique presque partout. Dans (6), la fonction $\mathbf{h}_i(\boldsymbol{\phi}) = \mathbf{h}(\mathbf{x}_i; \boldsymbol{\phi})$ possède un moment de quatrième ordre borné. En outre, la dérivée partielle $\partial\{\hat{\mathbf{U}}_h(\boldsymbol{\phi})\}/\partial\boldsymbol{\phi}$ est non singulière pour tout n.

[C3] Dans (6), la fonction d'estimation $\hat{\mathbf{U}}_h(\boldsymbol{\phi})$ converge en probabilité vers $\mathbf{U}_h(\boldsymbol{\phi}) = \sum_{i=1}^{N}\{r_i - p_i(\boldsymbol{\phi})\}\mathbf{h}_i(\boldsymbol{\phi})$ uniformément en $\boldsymbol{\phi}$. En outre, la dérivée partielle $\partial\{\hat{\mathbf{U}}_h(\boldsymbol{\phi})\}/\partial\boldsymbol{\phi}$ converge en probabilité vers $\partial\{\mathbf{U}_h(\boldsymbol{\phi})\}/\partial\boldsymbol{\phi}$ uniformément en $\boldsymbol{\phi}$. La solution $\boldsymbol{\phi}_N$ de $\mathbf{U}_h(\boldsymbol{\phi}) = \mathbf{0}$ satisfait $N^{1/2}(\boldsymbol{\phi}_N - \boldsymbol{\phi}_0) = O_p(1)$ sous le mécanisme de réponse.

La condition [C1] énonce les conditions de régularité pour le mécanisme de réponse. La condition [C2] est la condition de régularité pour la solution $\hat{\boldsymbol{\phi}}_h$ de (6). Dans la condition [C3], certaines conditions de régularité sont imposées à la fonction d'estimation $\hat{\mathbf{U}}_h(\boldsymbol{\phi})$ proprement dite. Par [C2] et [C3], nous pouvons établir la convergence à la vitesse $\sqrt{n}$ de $\hat{\boldsymbol{\phi}}_h$ en (2).

Maintenant, le théorème qui suit traite de certaines propriétés asymptotiques de l'estimateur ajusté sur le score de propension $\hat{\theta}_{\mathrm{ASP},h}$.

Théorème 1 Si les conditions [C1] à [C3] sont vérifiées, sous la distribution conjointe du mécanisme d'échantillonnage et du mécanisme de réponse, l'estimateur ajusté sur le score de propension $\hat{\theta}_{\mathrm{ASP},h}$ satisfait

$$\sqrt{n}\,(\hat{\theta}_{\mathrm{ASP},h} - \tilde{\theta}_{\mathrm{ASP},h}) = o_p(1), \tag{7}$$

où

$$\tilde{\theta}_{\mathrm{ASP},h} = \frac{1}{N}\sum_{i\in A} w_i\left\{p_i\,\mathbf{h}_i'\,\gamma_h^* + \frac{r_i}{p_i}(y_i - p_i\,\mathbf{h}_i'\,\gamma_h^*)\right\}, \tag{8}$$

$\gamma_h^* = (\sum_{i=1}^{N} r_i\,\mathbf{z}_i\,p_i\,\mathbf{h}_i')^{-1}(\sum_{i=1}^{N} r_i\,\mathbf{z}_i\,y_i)$, $p_i = p(\mathbf{x}_i; \boldsymbol{\phi}_0)$, $\mathbf{z}_i = \partial\{p^{-1}(\mathbf{x}_i; \boldsymbol{\phi}_0)\}/\partial\boldsymbol{\phi}$, *et* $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i; \boldsymbol{\phi}_0)$. *De plus, si la population finie est un échantillon aléatoire tiré d'un modèle de surpopulation, alors*

$$V(\tilde{\theta}_{\mathrm{ASP},h}) \geq V_l \equiv V(\hat{\theta}_{\mathrm{HT}}) + \frac{1}{N^2}E\left\{\sum_{i\in A} w_i^2\left(\frac{1}{p_i} - 1\right)V(Y \mid \mathbf{x}_i)\right\}. \tag{9}$$

Dans (9), l'égalité tient quand $\hat{\boldsymbol{\phi}}_h$ satisfait

$$\sum_{i\in A} w_i\left\{\frac{r_i}{p(\mathbf{x}_i; \hat{\boldsymbol{\phi}}_h)} - 1\right\}E(Y \mid \mathbf{x}_i) = 0, \tag{10}$$

où $E(Y \mid \mathbf{x}_i)$ est l'espérance conditionnelle sous le modèle de superpopulation.

Preuve. Étant donné $p_i(\boldsymbol{\phi}) = p(\mathbf{x}_i; \boldsymbol{\phi})$ et $\mathbf{h}_i(\boldsymbol{\phi}) = \mathbf{h}(\mathbf{x}_i; \boldsymbol{\phi})$, définissons

$$\hat{\theta}(\boldsymbol{\phi}, \gamma) = N^{-1}\sum_{i\in A} w_i\left[p_i(\boldsymbol{\phi})\mathbf{h}_i'(\boldsymbol{\phi})\gamma + \frac{r_i}{p_i(\boldsymbol{\phi})}\{y_i - p_i(\boldsymbol{\phi})\mathbf{h}_i'(\boldsymbol{\phi})\gamma\}\right].$$

Puisque $\hat{\boldsymbol{\phi}}_h$ satisfait (6), nous avons $\hat{\theta}_{\mathrm{ASP}} = \hat{\theta}(\hat{\boldsymbol{\phi}}_h, \gamma)$ pour tout choix de γ. Nous voulons maintenant trouver un choix particulier de γ, disons γ^*, tel que

$$\hat{\theta}(\hat{\boldsymbol{\phi}}_h, \gamma^*) = \hat{\theta}(\boldsymbol{\phi}_0, \gamma^*) + o_p(n^{-1/2}). \tag{11}$$

Comme $\hat{\boldsymbol{\phi}}_h$ converge en probabilité vers $\boldsymbol{\phi}_0$, l'équivalence asymptotique (11) est vérifiée si

$$E\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\theta}(\boldsymbol{\phi}, \gamma^*) \mid \boldsymbol{\phi} = \boldsymbol{\phi}_0\right\} = \mathbf{0}, \tag{12}$$

en utilisant la théorie de Randles (1982). La condition (12) est vérifiée si $\gamma^* = \gamma_h^*$, où γ_h^* est défini en (8). Donc, (11) se réduit à

$$\hat{\theta}_{\mathrm{ASP},h} = \frac{1}{N}\sum_{i\in A} w_i\left\{p_i\,\mathbf{h}_i'\,\gamma_h^* + \frac{r_i}{p_i}(y_i - p_i\,\mathbf{h}_i'\,\gamma_h^*)\right\} + o_p(n^{-1/2}), \tag{13}$$

ce qui prouve (7). La variance de $\tilde{\theta}_{\mathrm{ASP},h}$ peut être calculée par

$$\begin{aligned}
V(\tilde{\theta}_{\mathrm{ASP},h}) &= V(\hat{\theta}_{\mathrm{HT}}) + \frac{1}{N^2}E\left\{\sum_{i\in A} w_i^2\left(\frac{1}{p_i} - 1\right)(y_i - p_i\mathbf{h}_i'\gamma_h^*)^2\right\} \\
&= V(\hat{\theta}_{\mathrm{HT}}) + \frac{1}{N^2}E\left[\sum_{i\in A} w_i^2\left(\frac{1}{p_i} - 1\right)\{y_i - E(Y|\mathbf{x}_i) \right. \\
&\qquad \left. + E(Y|\mathbf{x}_i) - p_i\mathbf{h}_i'\gamma_h^*\}^2\right] \\
&= V(\hat{\theta}_{\mathrm{HT}}) + \frac{1}{N^2}E\left\{\sum_{i\in A} w_i^2\left(\frac{1}{p_i} - 1\right)V(Y|\mathbf{x}_i)\right\} \\
&\quad + \frac{1}{N^2}E\left[\sum_{i\in A} w_i^2\left(\frac{1}{p_i} - 1\right)\{E(Y|\mathbf{x}_i) - p_i\mathbf{h}_i'\gamma_h^*\}^2\right],
\end{aligned} \tag{14}$$

où la dernière égalité découle du fait que y_i est conditionnellement indépendante de $E(Y \mid \mathbf{x}_i) - p_i \mathbf{h}_i' \gamma_h^*$, en conditionnant sur $\mathbf{x}_i$. Puisque le dernier terme de (14) est non négatif, l'inégalité en (9) est établie. En outre, si $E(Y \mid \mathbf{x}_i) = p_i \mathbf{h}_i' \boldsymbol{\alpha}$ pour un certain $\boldsymbol{\alpha}$, (10) est vérifiée et $E(\gamma_h^* \mid \mathbf{x}_i) = \boldsymbol{\alpha}$, en vertu de la définition de γ_h^*. Donc, $E(Y \mid \mathbf{x}_i) - p_i \mathbf{h}_i' \gamma_h^* = -p_i \mathbf{h}_i' \{\gamma_h^* - E(\gamma_h^* \mid \mathbf{x}_i)\} = o_p(1)$, ce qui implique que le dernier terme de (14) est négligeable.

Dans (9), V_l est la borne inférieure de la variance asymptotique des estimateurs ajustés sur le score de propension de la forme (3) satisfaisant (6). Tout estimateur ajusté sur le score de propension possédant la variance asymptotique V_l donnée en (9) est optimal puisqu'il atteint la borne inférieure de la variance asymptotique parmi la classe d'estimateurs ajustés sur le score de propension pour lesquels $\hat{\boldsymbol{\phi}}$ satisfait (2). La variance asymptotique des estimateurs ajustés sur le score de propension optimaux de θ est égale à V_l en (9). L'estimateur ajusté sur le score de propension utilisant l'estimateur du maximum de vraisemblance de $\boldsymbol{\phi}_0$ n'atteint pas nécessairement la borne inférieure de la variance asymptotique.

La condition (10) donne un moyen de construire un estimateur ajusté sur le score de propension optimal. Premièrement, nous avons besoin d'une hypothèse pour $E(Y \mid \mathbf{x})$, qui est souvent appelée modèle de régression du résultat. Si le modèle de régression du résultat est un modèle de régression linéaire de la forme $E(Y \mid \mathbf{x}) = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}$, un estimateur ajusté sur le score de propension optimal de θ peut être obtenu en résolvant

$$\sum_{i \in A} w_i \frac{r_i}{p_i(\boldsymbol{\phi})} (1, \mathbf{x}_i) = \sum_{i \in A} w_i (1, \mathbf{x}_i). \qquad (15)$$

La condition (15) est intéressante, car elle dit que l'estimateur ajusté sur le score de propension appliqué à $y = a + \mathbf{b}'\mathbf{x}$ mène à l'estimateur HT original. La condition (15) est appelée condition de calage dans le domaine des sondages. La condition de calage appliquée à $\mathbf{x}$ utilise complètement l'information que celui-ci contient si la variable étudiée est bien approximée par une fonction linéaire de $\mathbf{x}$. La condition (15) a également été utilisée dans Nevo (2003) et dans Kott (2006) sous le modèle de régression linéaire.

Si nous utilisons explicitement un modèle de régression pour $E(Y \mid \mathbf{x})$, il est possible de construire un estimateur qui possède la variance asymptotique (9) et n'est pas nécessairement un estimateur ajusté sur le score de propension. Par exemple, si nous supposons que

$$E(Y \mid \mathbf{x}) = m(\mathbf{x}; \boldsymbol{\beta}_0) \qquad (16)$$

pour une certaine fonction $m(\mathbf{x}; \cdot)$ connue jusqu'à $\boldsymbol{\beta}_0$, nous pouvons utiliser le modèle (16) directement pour construire un estimateur optimal de la forme

$$\hat{\theta}_{\text{opt}} = \frac{1}{N} \sum_{i \in A} w_i \left[m(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) + \frac{r_i}{p_i(\hat{\boldsymbol{\phi}})} \{y_i - m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})\} \right], \qquad (17)$$

où $\hat{\boldsymbol{\beta}}$ est un estimateur convergeant à la vitesse $\sqrt{n}$ de $\boldsymbol{\beta}_0$ dans le modèle de superpopulation (16) et $\hat{\boldsymbol{\phi}}$ est un estimateur convergeant à la vitesse $\sqrt{n}$ de $\boldsymbol{\phi}_0$ calculé par (6). Le théorème qui suit montre que l'estimateur optimal (17) atteint la borne inférieure en (9).

Théorème 2 Posons que les conditions du théorème 1 sont vérifiées. Supposons que $\hat{\boldsymbol{\beta}}$ satisfait $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$. Supposons que, dans le modèle de superpopulation (16), $m(\boldsymbol{x}; \boldsymbol{\beta})$ possède des dérivées partielles d'ordre un continues dans un ensemble ouvert contenant $\boldsymbol{\beta}_0$. Sous la distribution conjointe du mécanisme d'échantillonnage, du mécanisme de réponse et du modèle de superpopulation (16), l'estimateur $\hat{\theta}_{\text{opt}}$ en (17) satisfait

$$\sqrt{n}(\hat{\theta}_{\text{opt}} - \tilde{\theta}_{\text{opt}}^*) = o_p(1),$$

où

$$\tilde{\theta}_{\text{opt}}^* = N^{-1} \sum_{i \in A} w_i \left[m(\boldsymbol{x}_i; \boldsymbol{\beta}_0) + \frac{r_i}{p_i} \{y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta}_0)\} \right],$$

$p_i = p_i(\boldsymbol{\phi}_0)$, et $V(\tilde{\theta}_{\text{opt}}^*)$ est égal à V_l dans (9).

Preuve. Définissons $\hat{\theta}_{\text{opt}}(\boldsymbol{\beta}, \boldsymbol{\phi}) = N^{-1} \sum_{i \in A} w_i [m(\boldsymbol{x}_i; \boldsymbol{\beta}) + r_i p_i^{-1}(\boldsymbol{\phi}) \{y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta})\}]$. Notons que, dans (17), $\hat{\theta}_{\text{opt}}$ peut s'écrire $\hat{\theta}_{\text{opt}} = \hat{\theta}_{\text{opt}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}})$. Puisque

$$\frac{\partial}{\partial \boldsymbol{\beta}} \hat{\theta}_{\text{opt}}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \frac{1}{N} \sum_{i \in A} w_i \left\{ \breve{m}(\boldsymbol{x}_i; \boldsymbol{\beta}) - \frac{r_i}{p_i(\boldsymbol{\phi})} \breve{m}(\boldsymbol{x}_i; \boldsymbol{\beta}) \right\},$$

où $\breve{m}(\boldsymbol{x}_i; \boldsymbol{\beta}) = \partial m(\boldsymbol{x}_i; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$, et

$$\frac{\partial}{\partial \boldsymbol{\phi}} \hat{\theta}_{\text{opt}}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \frac{1}{N} \sum_{i \in A} w_i r_i \mathbf{z}_i(\boldsymbol{\phi}) \{y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta})\},$$

où $\mathbf{z}_i(\boldsymbol{\phi}) = \partial \{p_i^{-1}(\boldsymbol{\phi})\} / \partial \boldsymbol{\phi}$, nous avons $E[\partial \{\hat{\theta}_{\text{opt}}(\boldsymbol{\beta}, \boldsymbol{\phi})\} / \partial (\boldsymbol{\beta}, \boldsymbol{\phi}) \mid \boldsymbol{\beta} = \boldsymbol{\beta}_0, \boldsymbol{\phi} = \boldsymbol{\phi}_0] = \mathbf{0}$ et la condition de Randles (1982) est satisfaite. Donc,

$$\hat{\theta}_{\text{opt}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}) = \hat{\theta}_{\text{opt}}(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0) + o_p(n^{-1/2}) = \tilde{\theta}_{\text{opt}}^* + o_p(n^{-1/2})$$

et la variance de $\tilde{\theta}_{\text{opt}}^*$ est égale à V_l, la borne inférieure de la variance asymptotique.

L'optimalité (asymptotique) de l'estimateur donné par (17) est justifiée sous la distribution conjointe du modèle de réponse (4) et du modèle de superpopulation (16). Quand les deux modèles sont corrects, $\hat{\theta}_{\text{opt}}$ est optimal et son efficacité n'est pas affectée par le choix de $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}})$ à condition que ces derniers convergent à la vitesse $\sqrt{n}$. Robins, Rotnitzky et Zhao (1994) ont également préconisé l'utilisation de $\hat{\theta}_{\text{opt}}$ donné par (17) sous échantillonnage aléatoire simple.

Remarque 1 Si le modèle de réponse est correct mais que le modèle de superpopulation (16) n'est pas nécessairement correct, le choix de $\hat{\boldsymbol{\beta}}$ influence l'efficacité de l'estimateur optimal. Cao, Tsiatis et Davidian (2009) ont considéré l'estimation optimale quand seul le modèle de réponse est correct. En utilisant la linéarisation de Taylor, l'estimateur optimal (17) avec $\hat{\boldsymbol{\phi}}$ satisfaisant (6) est asymptotiquement équivalent à

$$\tilde{\theta}(\boldsymbol{\beta}) = \sum_{i\in A} w_i \left[m(\boldsymbol{x}_i ; \boldsymbol{\beta}) + \frac{r_i}{p_i}\{y_i - m(\boldsymbol{x}_i ; \boldsymbol{\beta})\} - \left(\frac{r_i}{p_i} - 1\right)\mathbf{c}_\beta' \, p_i \, \mathbf{h}_i \right],$$

où $\mathbf{c}_\beta$ est la limite de probabilité de $\hat{\mathbf{c}}_\beta = \{\sum_{i\in A} w_i \, r_i \, \mathbf{z}_i(\hat{\boldsymbol{\phi}}) \, \hat{p}_i \mathbf{h}_i'(\hat{\boldsymbol{\phi}})\}^{-1} \sum_{i\in A} w_i r_i \mathbf{z}_i(\hat{\boldsymbol{\phi}})\{y_i - m(\mathbf{x}_i ; \boldsymbol{\beta})\}$ et $\mathbf{z}_i(\boldsymbol{\phi}) = \partial\{p_i^{-1}(\boldsymbol{\phi})\} / \partial\boldsymbol{\phi}$. La variance asymptotique est alors égale à

$$V\{\tilde{\theta}(\boldsymbol{\beta})\} = V(\hat{\theta}_{\text{HT}}) + E\left[\sum_{i\in A} w_i^2 \frac{1 - p_i}{p_i}\{y_i - m(\boldsymbol{x}_i ; \boldsymbol{\beta}) - \mathbf{c}_\beta' \, p_i \, \mathbf{h}_i\}^2\right].$$

Donc, un estimateur optimal de $\boldsymbol{\beta}$ peut être calculé en trouvant le $\hat{\boldsymbol{\beta}}$ qui minimise

$$Q(\boldsymbol{\beta}) = \sum_{i\in A} w_i^2 \, r_i \frac{1 - \hat{p}_i}{\hat{p}_i^2}\{y_i - m(\boldsymbol{x}_i ; \boldsymbol{\beta}) - \hat{\mathbf{c}}_\beta' \, \hat{p}_i \, \mathbf{h}_i(\hat{\boldsymbol{\phi}})\}^2.$$

L'estimateur résultant est optimal par rapport au plan d'échantillonnage car il minimise la variance asymptotique sous le modèle de réponse.

3. Modèle de score de propension augmenté

À la présente section, nous considérons une estimation ajustée sur le score de propension optimale. Notons qu'en (17), l'estimateur optimal $\hat{\theta}_{\text{opt}}$ n'est pas nécessairement écrit sous la forme d'un estimateur ajusté sur le score de propension donnée par (3). Il l'est s'il satisfait $\sum_{i\in A} w_i \, r_i \, \hat{p}_i^{-1} \, m(\mathbf{x}_i ; \hat{\boldsymbol{\beta}}) = \sum_{i\in A} w_i \, m(\mathbf{x}_i ; \hat{\boldsymbol{\beta}})$. Donc, nous pouvons construire un estimateur ajusté sur le score de propension optimal en incluant $m(\mathbf{x}_i ; \hat{\boldsymbol{\beta}})$ dans le modèle du score de propension. Spécifiquement, étant donné $\hat{m}_i = m(\mathbf{x}_i ; \hat{\boldsymbol{\beta}})$, $\hat{p}_i = p_i(\hat{\boldsymbol{\phi}})$ et $\hat{\mathbf{h}}_i = \mathbf{h}_i(\hat{\boldsymbol{\phi}})$, où $\hat{\boldsymbol{\phi}}$ est obtenu à partir de (6), nous augmentons le modèle de réponse par

$$p_i^*(\hat{\boldsymbol{\phi}}, \boldsymbol{\lambda}) \equiv \frac{\hat{p}_i}{\hat{p}_i + (1 - \hat{p}_i)\exp(\lambda_0 + \lambda_1 \hat{m}_i)}, \qquad (18)$$

où $\boldsymbol{\lambda} = (\lambda_0, \lambda_1)'$ est le multiplicateur de Lagrange qui est utilisé pour intégrer la contrainte supplémentaire. Si $(\lambda_0, \lambda_1)' = \mathbf{0}$, alors $p_i^*(\hat{\boldsymbol{\phi}}, \boldsymbol{\lambda}) = \hat{p}_i$. La probabilité de réponse augmentée $p_i^*(\hat{\boldsymbol{\phi}}, \boldsymbol{\lambda})$ prend toujours une valeur comprise entre 0 et 1. Le modèle de probabilité de réponse augmenté (18) peut être obtenu en minimisant la distance de Kullback-Leibler $\sum_{i\in A} w_i \, r_i \, q_i^* \log(q_i^* / q_i)$, où $q_i^* = (1 - p_i^*)/p_i^*$ et $q_i = (1 - \hat{p}_i) / \hat{p}_i$, sous la contrainte $\sum_{i\in A} w_i (r_i / p_i^*)(1, \hat{m}_i) = \sum_{i\in A} w_i \, (1, \hat{m}_i)$.

En utilisant (18), l'estimateur ajusté sur le score de propension optimal est calculé comme

$$\hat{\theta}_{\text{ASP}}^* = \frac{1}{N}\sum_{i\in A} w_i \frac{r_i}{p_i^*(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}})} y_i, \qquad (19)$$

où $\hat{\boldsymbol{\lambda}}$ satisfait

$$\sum_{i\in A} w_i \frac{r_i}{p_i^*(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}})}(1, \hat{m}_i) = \sum_{i\in A} w_i \, (1, \hat{m}_i). \qquad (20)$$

Sous le modèle de réponse (4), on peut montrer que

$$\hat{\theta}_{\text{ASP}}^* = \frac{1}{N}\sum_{i\in A} w_i \left\{\hat{b}_0 + \hat{b}_1 \hat{m}_i + \frac{r_i}{\hat{p}_i}(y_i - \hat{b}_0 - \hat{b}_1 \hat{m}_i)\right\} + o_p(n^{-1/2}),$$

où

$$\begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix} = \left\{\sum_{i\in A} w_i \, r_i \left(\frac{1}{\hat{p}_i} - 1\right)\begin{pmatrix} 1 \\ \hat{m}_i \end{pmatrix}\begin{pmatrix} 1 \\ \hat{m}_i \end{pmatrix}'\right\}^{-1} \sum_{i\in A} w_i \, r_i \left(\frac{1}{\hat{p}_i} - 1\right)\begin{pmatrix} 1 \\ \hat{m}_i \end{pmatrix} y_i. \qquad (21)$$

En outre, en vertu de l'argument du théorème 1, nous pouvons établir que

$$\hat{\theta}_{\text{ASP}}^* = \frac{1}{N}\sum_{i\in A} w_i \left\{b_0 + b_1 \hat{m}_i + \gamma_{h2}' \, p_i \, \mathbf{h}_i + \frac{r_i}{p_i}\left(y_i - b_0 - b_1 \hat{m}_i - \gamma_{h2}' \, p_i \, \mathbf{h}_i\right)\right\} + o_p(n^{-1/2}),$$

où (b_0, b_1, γ'_{h2}) est la limite de probabilité de $(\hat{b}_0, \hat{b}_1, \hat{\gamma}'_{h2})$ avec

$$\hat{\gamma}_{h2} = \left\{\sum_{i\in A} w_i r_i \mathbf{z}_i(\hat{\boldsymbol{\phi}})\hat{p}_i \mathbf{h}'_i(\hat{\boldsymbol{\phi}})\right\}^{-1} \sum_{i\in A} w_i r_i \mathbf{z}_i(\hat{\boldsymbol{\phi}})(y_i - \hat{b}_0 - \hat{b}_1\hat{m}_i) \qquad (22)$$

et l'effet de l'estimation de $\boldsymbol{\phi}_0$ dans $\hat{p}_i = p(\mathbf{x}_i ; \hat{\boldsymbol{\phi}})$ peut être ignoré sans risque.

Notons que, sous le modèle de réponse (4), $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}})$ dans (19) converge en probabilité vers $(\boldsymbol{\phi}_0, \mathbf{0})$, où $\boldsymbol{\phi}_0$ est le paramètre réel dans (4). Donc, le score de propension donné par le modèle augmenté converge vers la probabilité de réponse réelle. Comme $\hat{\boldsymbol{\lambda}}$ converge vers zéro en probabilité, le choix de $\hat{\boldsymbol{\beta}}$ dans $\hat{m}_i = m(x_i ; \hat{\boldsymbol{\beta}})$ ne joue aucun rôle dans l'absence asymptotique de biais de l'estimateur ajusté sur le score de propension. Les variances asymptotiques varient pour divers choix de $\hat{\boldsymbol{\beta}}$.

Sous le modèle de superpopulation (16), $\hat{b}_0 + \hat{b}_1\hat{m}_i \to E(Y \mid \mathbf{x}_i)$ en probabilité. Donc, l'estimateur ajusté sur le score de propension optimal (19) est asymptotiquement équivalent à l'estimateur optimal (17). L'introduction de $\hat{m}_i$ dans l'équation de calage pour atteindre l'optimalité est proche de l'esprit de la méthode de calage sur un modèle proposée par Wu et Sitter (2001).

4. Estimation de la variance

Nous abordons maintenant l'estimation de la variance des estimateurs ajustés sur le score de propension sous le modèle de réponse supposé. Singh et Folsom (2000) et Kott (2006) ont discuté de l'estimation de la variance pour certains types d'estimateurs ajustés sur le score de propension. Kim et Kim (2007) ont discuté de l'estimation de la variance quand l'estimateur ajusté sur le score de propension est calculé par la méthode du maximum de vraisemblance.

Nous considérons l'estimation de la variance de l'estimateur ajusté sur le score de propension de la forme (3) où $\hat{p}_i = p_i(\hat{\boldsymbol{\phi}})$ est construite en vue de satisfaire (6) pour une certaine fonction $\mathbf{h}_i(\boldsymbol{\phi}) = \mathbf{h}(\mathbf{x}_i ; \boldsymbol{\phi}, \hat{\boldsymbol{\beta}})$, où $\hat{\boldsymbol{\beta}}$ est obtenu en utilisant le modèle de superpopulation postulé. Soit $\boldsymbol{\beta}^*$ la limite de probabilité de $\hat{\boldsymbol{\beta}}$ sous le modèle de réponse. Notons que $\boldsymbol{\beta}^*$ n'est pas nécessairement égal à $\boldsymbol{\beta}_0$ dans (16), puisque nous ne supposons pas ici que le modèle de superpopulation postulé est spécifié correctement.

Si nous utilisons de l'argument pour la linéarisation de Taylor (13) utilisé dans la preuve du théorème 1, l'estimateur ajusté sur le score de propension satisfait

$$\hat{\theta}_{\mathrm{ASP}} = \frac{1}{N}\sum_{i\in A} w_i \eta_i(\boldsymbol{\phi}_0, \boldsymbol{\beta}^*) + o_p(n^{-1/2}), \qquad (23)$$

où

$$\eta_i(\boldsymbol{\phi}, \boldsymbol{\beta}) = p_i(\boldsymbol{\phi})\mathbf{h}'_i(\boldsymbol{\phi}, \boldsymbol{\beta})\gamma^*_h + \frac{r_i}{p_i(\boldsymbol{\phi})}\{y_i - p_i(\boldsymbol{\phi})\mathbf{h}'_i(\boldsymbol{\phi}, \boldsymbol{\beta})\gamma^*_h\}, \qquad (24)$$

$\mathbf{h}_i(\boldsymbol{\phi}, \boldsymbol{\beta}) = \mathbf{h}(\mathbf{x}_i ; \boldsymbol{\phi}, \boldsymbol{\beta})$ et γ^*_h est défini comme dans (8) avec $\mathbf{h}_i$ remplacé par $\mathbf{h}_i(\boldsymbol{\phi}_0, \boldsymbol{\beta}^*)$. Puisque $p_i(\hat{\boldsymbol{\phi}})$ satisfait (6) avec $\mathbf{h}_i(\boldsymbol{\phi}) = \mathbf{h}(\mathbf{x}_i ; \boldsymbol{\phi}, \hat{\boldsymbol{\beta}})$, l'expression $\hat{\theta}_{\mathrm{ASP}} = N^{-1}\sum_{i\in A} w_i \eta_i(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}})$ est vérifiée et, dans (23), la linéarisation peut être exprimée sous la forme $N^{-1}\sum_{i\in A} w_i \eta_i(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}}) = N^{-1}\sum_{i\in A} w_i \eta_i(\hat{\boldsymbol{\phi}}_0, \boldsymbol{\beta}^*) + o_p(n^{-1/2})$. Donc, si les $(\mathbf{x}_i, y_i, r_i)$ sont indépendantes et identiquement distribuées (IID), les $\eta_i(\boldsymbol{\phi}_0, \boldsymbol{\beta}^*)$ sont IID même si les $\eta_i(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}})$ ne le sont pas nécessairement. Comme les $\eta_i(\boldsymbol{\phi}_0, \boldsymbol{\beta}^*)$ sont IID, nous pouvons appliquer la méthode pour échantillon complet classique pour estimer la variance de $\hat{\eta}_{\mathrm{HT}} = N^{-1}\sum_{i\in A} w_i \eta_i(\boldsymbol{\phi}_0, \boldsymbol{\beta}^*)$, qui est asymptotiquement équivalente à la variance de $\hat{\theta}_{\mathrm{ASP}} = N^{-1}\sum_{i\in A} w_i \eta_i(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}})$. Voir Kim et Rao (2009).

Pour obtenir l'estimateur de variance, nous supposons que l'estimateur de variance $\hat{V} = N^{-2}\sum_{i\in A}\sum_{j\in A}\Omega_{ij} g_i g_j$ satisfait $\hat{V}/V(\hat{g}_{\mathrm{HT}} \mid \mathcal{F}_N) = 1 + o_p(1)$ pour un certain Ω_{ij} relié à la probabilité d'inclusion conjointe, où $\hat{g}_{\mathrm{HT}} = N^{-1}\sum_{i\in A} w_i g_i$ pour tout g avec un moment d'ordre deux fini et $V(g_{\mathrm{HT}} \mid \mathcal{F}_N) = N^{-2}\sum_{i=1}^N\sum_{j=1}^N \Omega_{N\cdot ij} g_i g_j$, pour un certain $\Omega_{N\cdot ij}$. Nous supposons aussi que

$$\sum_{i=1}^N |\Omega_{N\cdot ij}| = O(n^{-1}N). \qquad (25)$$

Pour obtenir la variance totale, nous considérons le *cadre inverse* de Fay (1992), Shao et Steel (1999) et Kim et Rao (2009). Dans ce cadre, la population finie est d'abord divisée en deux groupes, une population de répondants et une population de non-répondants. Connaissant la population, l'échantillon A est sélectionné selon un plan d'échantillonnage probabiliste. Donc, la sélection de la population de répondants à partir de la population finie complète est traitée comme l'échantillonnage de première phase et la sélection de l'échantillon de répondants à partir de la population de répondants est traitée comme l'échantillonnage de deuxième phase dans le cadre inverse. La variance totale de $\hat{\eta}_{\mathrm{HT}}$ peut s'écrire

$$V(\hat{\eta}_{\mathrm{HT}} \mid \mathcal{F}_N) = V_1 + V_2 = E\{V(\hat{\eta}_{\mathrm{HT}} \mid \mathcal{F}_N, \mathcal{R}_N) \mid \mathcal{F}_N\} + V\{E(\hat{\eta}_{\mathrm{HT}} \mid \mathcal{F}_N, \mathcal{R}_N) \mid \mathcal{F}_N\}. \qquad (26)$$

Dans (26), le terme de variance conditionnelle $V(\hat{\eta}_{\mathrm{HT}} \mid \mathcal{F}_N, \mathcal{R}_N)$ peut être estimé par

$$\hat{V}_1 = N^{-2} \sum_{i \in A} \sum_{j \in A} \Omega_{ij} \hat{\eta}_i \hat{\eta}_j, \qquad (27)$$

où $\hat{\eta}_i = \eta_i(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}})$ est défini dans (24) avec γ_h^* remplacé par un estimateur convergent tel que $\hat{\gamma}_h^* = \{\sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\boldsymbol{\phi}}) \hat{p}_i \hat{\mathbf{h}}_i'\}^{-1} \sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\boldsymbol{\phi}}) y_i$, et $\hat{\mathbf{h}}_i = \mathbf{h}(\mathbf{x}_i ; \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}})$. Pour montrer que $\hat{V}_1$ converge également vers V_1 dans (26), il suffit de montrer que $V\{n \cdot V(\hat{\eta}_{\mathrm{HT}} \mid \mathcal{F}_N, \mathcal{R}_N) \mid \mathcal{F}_N\} = o(1)$, ce qui découle de (25) et de l'existence du moment d'ordre quatre. Voir Kim, Navarro et Fuller (2006). Dans (26), le deuxième terme V_2 est

$$V\{E(\hat{\eta}_{\mathrm{HT}} \mid \mathcal{F}_N, \mathcal{R}_N) \mid \mathcal{F}_N\} = V\left(N^{-1} \sum_{i=1}^{N} \eta_i \mid \mathcal{F}_N \right) = \frac{1}{N^2} \sum_{i=1}^{N} \frac{1 - p_i}{p_i} (y_i - p_i \mathbf{h}_i^{*\prime} \gamma_h^*)^2,$$

où $\mathbf{h}_i^* = \mathbf{h}(\mathbf{x}_i ; \boldsymbol{\phi}_0, \boldsymbol{\beta}^*)$. Un estimateur convergent de V_2 peut s'obtenir sous la forme

$$\hat{V}_2 = \frac{1}{N^2} \sum_{i \in A} w_i r_i \frac{1 - \hat{p}_i}{\hat{p}_i^2} (y_i - \hat{p}_i \hat{\mathbf{h}}_i' \hat{\gamma}_h^*)^2, \qquad (28)$$

où $\hat{\gamma}_h^*$ est défini d'après (27). Donc,

$$\hat{V}(\hat{\theta}_{\mathrm{ASP}}) = \hat{V}_1 + \hat{V}_2, \qquad (29)$$

est un estimateur convergent de la variance de l'estimateur ajusté sur le score de propension défini dans (3) avec $\hat{p}_i = p_i(\hat{\boldsymbol{\phi}})$ satisfaisant (6), où $\hat{V}_1$ est donné par (27) et $\hat{V}_2$, par (28).

Notons que le premier terme de la variance totale est $V_1 = O_p(n^{-1})$, mais que le deuxième terme est $V_2 = O_p(N^{-1})$. Donc, quand la fraction d'échantillonnage nN^{-1} est négligeable, c'est-à-dire $nN^{-1} = o(1)$, le deuxième terme V_2 peut être ignoré et $\hat{V}_1$ est un estimateur convergent de la variance totale. Sinon, il faut tenir compte du deuxième terme V_2 pour pouvoir construire un estimateur de variance convergent comme dans (29).

Remarque 2 L'estimation de la variance de l'estimateur ajusté sur le score de propension optimal sous le modèle de score de propension augmenté (18) avec $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\lambda}})$ *satisfaisant (20) peut être dérivé de (29) en utilisant* $\hat{\eta}_i = \hat{b}_0 + \hat{b}_1 \hat{m}_i + \hat{\gamma}_{h2}' \hat{p}_i \hat{\mathbf{h}}_i + r_i \hat{p}_i^{-1} (y_i - \hat{b}_0 - \hat{b}_1 \hat{m}_i - \hat{\gamma}_{h2}' \hat{p}_i \hat{\mathbf{h}}_i)$ *où* $(\hat{b}_0, \hat{b}_1)$ *et* $\hat{\gamma}_{h2}$ *sont définis dans (21) et (22), respectivement.*

5. Études par simulation

5.1 Première étude

Deux études par simulation ont été exécutées pour étudier les propriétés de la méthode proposée. Dans la première simulation, nous avons créé une population finie de taille $N = 10\,000$ à partir de la loi normale multivariée suivante :

$$\begin{pmatrix} x_1 \\ x_2 \\ e \end{pmatrix} \sim N\left[\begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0{,}5 & 0 \\ 0{,}5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right].$$

La variable d'intérêt y a été construite sous la forme $y = 1 + x_1 + e$. Nous avons également généré des variables indicatrices de réponse r_i indépendamment à partir d'une loi de Bernoulli de paramètre

$$p_i = \frac{\exp(2 + x_{2i})}{1 + \exp(2 + x_{2i})}.$$

Partant de la population finie, nous avons utilisé l'échantillonnage aléatoire simple pour sélectionner deux échantillons de taille $n = 100$ et $n = 400$, respectivement. Nous avons utilisé $B = 5\,000$ échantillons Monte Carlo dans la simulation. Le taux de réponse moyen était de 69,6 % environ.

Pour calculer le score de propension, nous avons postulé un modèle de réponse de la forme

$$p(\mathbf{x} ; \boldsymbol{\phi}) = \frac{\exp(\phi_0 + \phi_1 x_2)}{1 + \exp(\phi_0 + \phi_1 x_2)} \qquad (30)$$

et un modèle de régression du résultat de la forme

$$m(\mathbf{x} ; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 \qquad (31)$$

pour obtenir les estimateurs ajustés sur le score de propension optimaux. Donc, les deux modèles étaient spécifiés correctement. Pour chaque échantillon, nous avons calculé quatre estimateurs de $\theta = N^{-1} \sum_{i=1}^{N} y_i$:

1. (ASP-EMV) : Estimateur ajusté sur le score de propension (3) avec $\hat{p}_i = p_i(\hat{\boldsymbol{\phi}})$ et $\hat{\boldsymbol{\phi}}$ étant l'estimateur du maximum de vraisemblance de $\boldsymbol{\phi}$.
2. (ASP-CAL) : Estimateur ajusté sur le score de propension (3) avec $\hat{p}_i$ satisfaisant la contrainte de calage (15) sur $(1, x_{2i})$.
3. (AUG) : Estimateur ajusté sur le score de propension augmenté (19).
4. (OPT) : Estimateur optimal (17).

Dans les estimateurs ajustés sur le score de propension augmenté, $\hat{\boldsymbol{\phi}}$ a été calculé par la méthode du maximum de vraisemblance. Sous le modèle (30), l'estimateur du maximum de vraisemblance de $\boldsymbol{\phi} = (\phi_0, \phi_1)'$ a été calculé en résolvant (6) avec $\mathbf{h}_i(\boldsymbol{\phi}) = (1, x_{2i})'$. Le paramètre (β_0, β_1) pour le modèle de régression du résultat a été calculé par régression de y sur x_1 par la méthode des moindres carrés ordinaires. En plus des estimateurs ponctuels, nous avons calculé les estimateurs de variance de ces estimateurs. Les estimateurs de variance des estimateurs ajustés sur le score

de propension ont été calculés en utilisant les pseudo-valeurs dans (24) et la fonction $\mathbf{h}_i(\boldsymbol{\phi})$ correspondant à chaque estimateur. Pour les estimateurs ajustés sur le score de propension augmenté, les pseudo-valeurs ont été calculées par la méthode de la remarque 2.

Le tableau 1 donne les biais, variances et erreurs quadratiques moyenne Monte Carlo des quatre estimateurs ponctuels, ainsi que les biais relatifs en pourcentage et les statistiques t Monte Carlo des estimateurs de variance des estimateurs ponctuels. Le biais relatif en pourcentage d'un estimateur de variance $\hat{V}(\hat{\theta})$ est calculé par $100 \times \{V_{\mathrm{MC}}(\hat{\theta})\}^{-1}[E_{\mathrm{MC}}\{\hat{V}(\hat{\theta})\} - V_{\mathrm{MC}}(\hat{\theta})]$, où $E_{\mathrm{MC}}(\cdot)$ et $V_{\mathrm{MC}}(\cdot)$ désignent l'espérance Monte Carlo et la variance Monte Carlo, respectivement. Au tableau 1, la statistique t est la statistique utilisée pour tester l'hypothèse que le biais de l'estimateur de variance est nul. Voir Kim (2004).

Les résultats de simulation présentés au tableau 1 nous permettent de tirer les conclusions suivantes.

1. Tous les estimateurs ajustés sur le score de propension sont asymptotiquement sans biais, parce que le modèle de réponse (30) est spécifié correctement. L'estimateur ajusté sur le score de propension utilisant la méthode de calage est légèrement plus efficace que celui utilisant l'estimateur du maximum de vraisemblance, parce que le dernier terme de (14) est plus petit pour la méthode de calage, car le prédicteur de $E(Y \mid \mathbf{x}_i) = \beta_0 + \beta_1 x_{1i}$ est mieux approximé par une fonction linéaire de $(1, x_{2i})$ que par une fonction linéaire de $(\hat{p}_i, \hat{p}_i x_{2i})$.
2. L'estimateur ajusté sur le score de propension augmenté est plus efficace que l'estimateur ajusté sur le score de propension direct (3). L'estimateur augmenté est construit en utilisant le modèle de régression spécifié correctement (31) et est donc asymptotiquement équivalent à l'estimateur ajusté sur le score de propension optimal (17).
3. Les estimateurs de variance sont tous approximativement sans biais. Les estimateurs de variance des estimateurs ajustés sur le score de propension présentent un biais modeste quand la taille de l'échantillon est faible ($n = 100$).

5.2 Deuxième étude

Dans la deuxième étude par simulation, nous avons poursuivi l'étude des estimateurs ajustés sur le score de propension avec un modèle de régression du résultat non linéaire sous un plan de sondage avec probabilités inégales. Nous avons créé deux populations finies stratifiées de (x, y) comprenant quatre strates ($h = 1, 2, 3, 4$), où les x_{hi} étaient des variables indépendantes tirées selon une loi normale $N(1, 1)$ et les y_{hi} étaient des variables dichotomiques prenant la valeur 1 ou 0 tirées selon une loi de Bernoulli de paramètre p_{1yhi} ou p_{2yhi}. Des probabilités différentes ont été utilisées pour ces deux populations, respectivement :

1. Population 1 (Pop1) :

$$p_{1yhi} = 1 / \{1 + \exp(0{,}5 - 2x)\}.$$

2. Population 2 (Pop2) :

$$p_{2yhi} = 1 / [1 + \exp\{0{,}25(x - 1{,}5)^2 - 1{,}5\}].$$

En plus de x_{hi} et y_{hi}, les variables indicatrices de réponse r_{hi} ont été tirées selon une loi de Bernoulli de paramètre $p_{hi} = 1 / \{1 + \exp(-1{,}5 + 0{,}7x_{hi})\}$. Les tailles des quatre strates étaient $N_1 = 1\,000$, $N_2 = 2\,000$, $N_3 = 3\,000$ et $N_4 = 4\,000$, respectivement. Dans chacune des deux populations finies, nous avons procédé au tirage d'un échantillon stratifié de taille $n = 400$ indépendamment sans remise, où un échantillon aléatoire simple de taille $n_h = 100$ a été tiré de chaque strate. Nous avons utilisé $B = 5\,000$ échantillons Monte Carlo dans cette simulation. Le taux de réponse était de 67 % environ.

Tableau 1
Biais, variance et erreur quadratique moyenne (EQM) Monte Carlo des quatre estimateurs ponctuels et biais relatif (BR) en pourcentage et statistique t (stat. t) des estimateurs de variance fondés sur 5 000 échantillons Monte Carlo

n	Méthode	$\hat{\theta}$			$V(\hat{\theta})$	
		Biais	Variance	EQM	BR(%)	Stat. t
100	(ASP-EMV)	-0,01	0,0315	0,0317	-2,34	-1,12
	(ASP-CAL)	-0,01	0,0308	0,0309	-3,56	-1,70
	(AUG)	0,00	0,0252	0,0252	-0,61	-0,30
	(OPT)	0,00	0,0252	0,0252	-0,21	-0,10
400	(ASP-EMV)	-0,01	0,00737	0,00746	0,35	0,17
	(ASP-CAL)	-0,01	0,00724	0,00728	0,29	0,14
	(AUG)	0,00	0,00612	0,00612	0,07	0,03
	(OPT)	0,00	0,00612	0,00612	-0,14	-0,07

Pour calculer le score de propension, nous avons postulé un modèle de réponse de la forme

$$p(x\,;\boldsymbol{\phi}) = \frac{\exp(\phi_0 + \phi_1 x)}{1 + \exp(\phi_0 + \phi_1 x)}$$

pour estimer les paramètres. Pour obtenir l'estimateur ajusté sur le score de propension augmenté, nous avons postulé pour la variable d'intérêt un modèle de la forme

$$m(x\,;\boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \qquad (32)$$

Donc, le modèle (32) est un modèle vrai sous (Pop1), mais non sous (Pop2).

Nous avons calculé quatre estimateurs :

1. (ASP-EMV) : estimateur ajusté sur le score de propension (3) en utilisant l'estimateur du maximum de vraisemblance de $\boldsymbol{\phi}$.
2. (ASP-CAL) : estimateur ajusté sur le score de propension (3) avec $\hat{p}_i$ satisfaisant la contrainte de calage (15) sur $(1, x)$.
3. (AUG-1) : estimateur ajusté sur le score de propension augmenté $\hat{\theta}^*_{\text{ASP}}$ (19) avec $\hat{\boldsymbol{\beta}}$ calculé par la méthode du maximum de vraisemblance.
4. (AUG-2) : estimateur ajusté sur le score de propension augmenté $\hat{\theta}^*_{\text{ASP}}$ (19) avec $\hat{\boldsymbol{\beta}}$ calculé par la méthode de Cao et coll. (2009) discutée à la remarque 1.

Nous avons considéré l'estimateur ajusté sur le score de propension augmenté (19) avec $\hat{p}_i = p_i(\hat{\boldsymbol{\phi}})$, où $\hat{\boldsymbol{\phi}}$ est l'estimateur du maximum de vraisemblance de $\boldsymbol{\phi}$. Le premier estimateur ajusté sur le score de propension augmenté (AUG-1) utilisait $\hat{m}_i = m(x_i\,;\hat{\boldsymbol{\beta}})$ avec $\hat{\boldsymbol{\beta}}$ obtenu par résolution de $\sum_{h=1}^{4}\sum_{i\in A_h} w_{hi}\, r_{hi}\{y_{hi} - m(x_{hi}\,;\boldsymbol{\beta})\}(1, x_{hi}) = \mathbf{0}$, où A_h est l'ensemble d'indices figurant dans l'échantillon de la strate h et w_{hi} est le poids d'échantillonnage de l'unité i dans la strate h.

Le tableau 2 donne les résultats de simulation pour chaque méthode. Pour chaque population, l'estimateur ajusté sur le score de propension augmenté montre une certaine amélioration de la variance comparativement à celui utilisant l'estimateur du maximum de vraisemblance de $\boldsymbol{\phi}$ ou l'estimateur par calage de $\boldsymbol{\phi}$. Sous (Pop1), puisque le modèle (32) est vrai, il n'existe essentiellement aucune différence entre les estimateurs ajustés sur le score de propension augmenté utilisant différentes méthodes d'estimation de $\boldsymbol{\beta}$. Cependant, sous (Pop2), où le modèle de régression du résultat supposé (32) est incorrect, l'estimateur ajusté sur le score de propension augmenté dans lequel $\hat{\boldsymbol{\beta}}$ est calculé par la méthode de Cao et coll. (2009) est un peu plus efficace, ce qui est en harmonie avec la théorie de la remarque 1. Les estimations de variance sont approximativement sans biais dans tous les cas dans l'étude par simulation.

6. Conclusion

Nous avons considéré le problème de l'estimation de la moyenne de y en population finie en présence de non-réponse en utilisant la méthode du score de propension. Nous avons calculé le score de propension à l'aide d'un modèle paramétrique de la probabilité de réponse et discuté de certaines propriétés asymptotiques des estimateurs ajustés sur le score de propension. En particulier, l'estimateur ajusté sur le score de propension optimal est établi en émettant une hypothèse supplémentaire au sujet de la distribution de y. Le score de propension pour l'estimateur ajusté sur le score de propension optimal peut-être obtenu à l'aide du modèle de score de propension augmenté présenté à la section 3. L'estimateur résultant reste convergent, même si le modèle de régression du résultat supposé ne tient pas.

Nous avons limité notre étude au mécanisme de données manquant au hasard dans lequel la probabilité de réponse ne dépend que de $\mathbf{x}$ qui est toujours observé. Si le mécanisme de réponse dépend également de y, l'estimation ajustée sur le score de propension devient plus difficile. L'estimation ajustée sur le score de propension quand les données ne manquent pas au hasard dépasse le cadre du présent article et sera le sujet d'une future étude.

Tableau 2
Biais, variance et erreur quadratique moyenne Monte Carlo des quatre estimateurs ponctuels et biais relatifs (BR) en pourcentage et statistique t des estimateurs de variance, fondés sur 5 000 échantillons Monte Carlo

Population	Méthode	$\hat{\theta}_{\text{ASP}}$			$V(\hat{\theta}_{\text{ASP}})$	
		Biais	Variance	EQM	BR (%)	Stat. t
Pop1	(ASP-EMV)	0,00	0,000750	0,000762	-1,13	-0,57
	(ASP-CAL)	0,00	0,000762	0,000769	-1,45	-0,72
	(AUG-1)	0,00	0,000745	0,000757	-1,73	-0,86
	(AUG-2)	0,00	0,000745	0,000757	-1,83	-0,91
Pop2	(ASP-EMV)	0,00	0,000824	0,000826	0,29	0,14
	(ASP-CAL)	0,00	0,000829	0,000835	-0,94	-0,46
	(AUG-1)	0,00	0,000822	0,000823	-0,71	-0,35
	(AUG-2)	0,00	0,000820	0,000821	-0,61	-0,30

Remerciements

Les travaux de recherche ont été financés en partie par une entente de coopération entre le Natural Resources Conservation Service du US Department of Agriculture et la Iowa State University. Les auteurs remercient F. Jay Breidt, trois examinateurs anonymes et le rédacteur associé de leurs commentaires constructifs.

Bibliographie

Cao, W., Tsiatis, A.A. et Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96, 723-734.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34, 305-334.

Da Silva, D.N., et Opsomer, J.D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *Canadian Journal of Statistics*, 34, 563-579.

Da Silva, D.N., et Opsomer, J.D. (2009). Pondération par la propension à répondre non paramétrique fondée sur la régression par polynômes locaux pour corriger la non-réponse aux enquêtes. *Techniques d'enquête*, 35, 2, 179-192.

Duncan, K.B., et Stasny, E.A. (2001). Utilisation des scores de propension pour contrôler le biais de couverture dans les enquêtes téléphoniques. *Techniques d'enquête*, 27, 2, 131-141.

Durrant, G.B., et Skinner, C. (2006). Utilisation de méthodes de traitement des données manquantes pour corriger l'erreur de mesure dans une fonction de distribution. *Techniques d'enquête*, 32, 1, 27-39.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistics Section*, American Statistical Association, 197-202.

Fuller, W.A., Loughin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la *Nationwide Food Consumption Survey* de 1987-1988. *Techniques d'enquête*, 20, 1, 79-89.

Iannacchione, V.G., Milne, J.G. et Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 637-642.

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

Kim, J.K. (2004). Finite sample properties of multiple imputation estimators. *The Annals of Statistics*, 32, 766-783.

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

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Kim, J.K., et Rao, J.N.K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160.

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

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-296.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business ss and Economic Statistics*, 21, 43-52.

Pfeffermann, D., Krieger, A.M. et Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Randles, R.H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, 10, 462-474.

Rizzo, L., Kalton, G. et Brick, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 1, 43-53.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387-394.

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

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Singh, A.C., et Folsom, R.E. (2000). Bias corrected estimating function approach for variance estimation adjusted for poststratification. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 610-615.

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

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 181-186
Statistique Canada, N° 12-001-X au catalogue

Évaluation de l'exactitude des modèles de propension à répondre dans les études longitudinales

Ian Plewis, Sosthenes Ketende et Lisa Calderwood [1]

Résumé

La question de la non-réponse dans les études longitudinales est abordée en évaluant l'exactitude des modèles de propension à répondre construits pour distinguer et prédire les divers types de non-réponse. Une attention particulière est accordée aux mesures sommaires dérivées des courbes de la fonction d'efficacité du receveur, ou courbes ROC (de l'anglais *receiver operating characteristics*), ainsi que des courbes de type logit sur rangs. Les concepts sont appliqués à des données provenant de la Millennium Cohort Study du Royaume-Uni. Selon les résultats, la capacité de faire la distinction entre les divers types de non-répondants et de les prévoir n'est pas grande. Les poids produits au moyen des modèles de propension à répondre ne donnent lieu qu'à de faibles corrections des transitions entre situations d'emploi. Des conclusions sont tirées quant aux possibilités d'intervention en vue de prévenir la non-réponse.

Mots clés : Études longitudinales ; données manquantes ; pondération ; scores de propension ; courbes ROC ; Millennium Cohort Study.

1. Introduction

Les exemples d'études ayant modélisé les prédicteurs des différents types de non-réponse et des raisons de la non-réponse dans les études longitudinales sont nombreux. De telles modélisations ont été rendues possibles grâce à la capacité de se servir de variables auxiliaires pour lesquelles des données ont été obtenues auprès des membres de l'échantillon avant (et après) les cycles auxquels ces membres non pas répondu. Notamment, Lepkowski et Couper (2002) proposent une analyse séparant les cas de refus des cas dont les répondants n'ont pu être localisés ou contactés ; Hawkes et Plewis (2006) font la distinction entre les non-répondants à une vague et les cas d'attrition dans la National Child Development Study menée au Royaume-Uni ; et Plewis (2007a) et Plewis, Ketende, Joshi et Hughes (2008) examinent la non-réponse aux deux premières vagues de la Millennium Cohort Study du Royaume-Uni. Le présent article porte sur la façon dont nous pouvons évaluer l'exactitude de ces modèles de propension à répondre (Little et Rubin 2002). Il s'appuie sur un cadre fréquemment utilisé en épidémiologie (Pepe 2003) et en criminologie (Copas 1999) pour évaluer les scores de risque, mais qui, autant que nous sachions, n'a pas été utilisé auparavant dans le domaine de la recherche des enquêtes. Les modèles de propension à répondre peuvent être utilisés pour construire des poids destinés à éliminer les biais des estimations, à faciliter les imputations, et à prédire les non-répondants aux vagues futures afin d'orienter les ressources pour le travail sur le terrain vers ces répondants qui, autrement, pourraient être perdus. Toutefois, l'exactitude des modèles de propension à répondre n'a pas reçu l'attention qu'elle aurait dû en ce qui concerne la capacité à faire la distinction entre les répondants et les non-répondants et à prédire la non-réponse. De bonnes estimations de l'exactitude peuvent être utilisées pour comparer l'efficacité de différents modèles de pondération et pour faciliter la répartition des ressources limitées réservées au travail sur le terrain afin de réduire la non-réponse.

La présentation de l'article est la suivante. Le cadre d'évaluation de l'exactitude est exposé à la section suivante. À la section 3, nous présentons la Millennium Cohort Study du Royaume-Uni et à la section 4, nous illustrons les méthodes au moyen de données provenant de cette étude. Enfin, à la section 5, nous présentons nos conclusions.

2. Modèles pour la prédiction de la non-réponse

Un modèle type de propension à répondre pour un résultat binaire (par exemple Hawkes et Plewis 2006) est donné par :

$$f(\pi_{it}) = \sum_p \beta_p x_{pi} + \sum_q \sum_k \gamma_q x^*_{qi,\,t-k} + \sum_r \sum_k \delta_r z_{ri,\,t-k} \qquad (1)$$

où

- $\pi_{it} = E(r_{it})$ est la probabilité que le sujet i ne réponde pas à la vague t ; $r_{it} = 0$ pour une réponse et 1 pour une non-réponse ; f est une fonction appropriée, telle que la fonction logit ou probit ;
- $i = 1, ..., n$ où n est la taille de l'échantillon observé à la première vague ;
- $t = 1, ..., T_i$ où T_i est le nombre de vagues pour lesquelles r_{it} est enregistré pour le sujet i ;

1. Ian Plewis, Social Statistics, University of Manchester, Manchester M13 9PL, Royaume-Uni. Courriel : ian.plewis@manchester.ac.uk ; Sosthenes Ketende et Lisa Calderwood, Centre for Longitudinal Studies, Institute of Education, Londres WC1H 0AL, Royaume-Uni.

- x_{pi} représente les caractéristiques fixes du sujet i mesurées à la vague un, $p = 0, ..., P$; $x_0 = 1$ pour tout i;
- $x^*_{qi,t-k}$ représente les caractéristiques variant avec le temps du sujet i, mesurées aux vagues $t - k$, $q = 1, ..., Q$, $k = 1, 2, ...,$ souvent k est égal à 1 ;
- $z_{ri,t-k}$ représente les caractéristiques variant avec le temps du processus de collecte des données, mesurées pour le sujet i aux vagues $t - k$, $r = 1, ..., R$, $k = 0, 1, ...,$ souvent k est égal à 1, mais peut être égal à 0 pour des variables telles que le nombre de contacts avant d'obtenir une réponse.

Le modèle (1) peut être étendu facilement à plus de deux catégories de réponse, comme {réponse, non-réponse à une vague, attrition}. D'autres approches sont également possibles. Par exemple, il est souvent plus commode de modéliser la probabilité de ne pas répondre seulement à la vague $t = t^*$ en ce qui a trait aux variables mesurées durant les vagues antérieures $t^* - k$, $k \geq 1$ ou, en l'absence de non-réponse à une vague de sorte que la courbe de non-réponse est monotone plutôt qu'arbitraire, de modéliser le temps écoulé jusqu'à l'attrition comme un processus de survie.

Les probabilités de réponses estimées p_i, pour $t = t^*$, sont dérivées des probabilités de non-réponse estimées en (1) et peuvent servir à produire des pondérations égales à l'inverse de la probabilité $g_i(=1/p_i)$. Ces pondérations sont très souvent utilisées (voir la section 4.2 pour un exemple) pour corriger le biais dû à la non-réponse sous l'hypothèse que les données manquantes sont dues au hasard (MAR pour *missing at random*), comme l'ont défini Little et Rubin (2002).

2.1 Évaluation de l'exactitude des prédictions

Une méthode très répandue pour évaluer l'exactitude des modèles tels que (1) consiste à estimer leur adéquation en utilisant une ou plusieurs statistiques pseudo-R^2 possibles. Les estimations du pseudo-R^2 ne sont pas particulièrement utiles dans le présent contexte, en partie parce qu'elles sont difficiles à comparer entre jeux de données, mais aussi parce qu'elles évaluent l'adéquation globale du modèle et ne font donc pas la distinction entre l'exactitude du modèle pour les répondants et pour les non-répondants pris séparément.

Comme le souligne Pepe (2003), l'exactitude possède deux composantes apparentées : la discrimination (ou classification) et la prédiction. Par discrimination, on entend les probabilités conditionnelles d'avoir un score de propension à répondre (s : le prédicteur linéaire provenant de (1)) supérieur à un seuil choisi (c) sachant qu'une personne est ou n'est pas un non-répondant. Par ailleurs, par prédiction, on entend les probabilités conditionnelles de devenir un non-répondant étant donné un score de propension supérieur ou inférieur au seuil.

De manière plus formelle, soit D et $\bar{D}$ la présence et l'absence du mauvais résultat (c'est-à-dire la non-réponse) et définissons + ($s > c$) et − ($s \leq c$) comme étant les tests positif et négatif dérivés du score de propension à répondre et de son seuil. Alors, pour la discrimination, nous nous intéressons à $P(+|D)$, la fraction de vrais positifs (FVP) ou sensibilité du test, et à $P(-|\bar{D})$, sa spécificité, qui est égale à un moins la fraction de faux positifs (1 − FFP). Pour la prédiction, par contre, nous nous intéressons à $P(D|+)$, la valeur prédictive positive (VPP) et à $P(\bar{D}|-)$, la valeur prédictive négative (VPN). Si la probabilité d'un test positif ($P(+) = \tau$) est la même que la prévalence du mauvais résultat ($P(D) = \rho$), les inférences au sujet de la discrimination et de la prédiction sont essentiellement les mêmes : la sensibilité est égale à la VPP et la spécificité est égale à la VPN. Cependant, généralement, {FVP, FFP, ρ} and {VPP, VPN, τ} communiquent des éléments d'information différents. La FVP peut être représentée graphiquement en fonction de la FFP pour tout seuil de score de risque c. On obtient ainsi la courbe de la fonction d'efficacité du receveur, ou courbe ROC (figure 1). Krzanowski et Hand (2009) discutent en détail de la façon d'estimer les courbes ROC. L'aire sous la courbe (ASC) – l'aire délimitée par la courbe ROC et par l'axe des x dans la figure 1 – est particulièrement intéressante et sa valeur peut varier de 1 (discrimination parfaite) à 0,5, c'est-à-dire l'aire sous la diagonale, ce qui implique l'absence de discrimination). L'ASC peut être interprétée comme la probabilité d'attribuer une paire de cas, un répondant et un non-répondant, aux catégories correctes, en se souvenant qu'une réponse devinée correspondrait à une probabilité de 0,5. Une transformation linéaire de l'ASC (= 2*ASC – 1) – parfois appelée coefficient de Gini et équivalent à l'indice de corrélation de rang D de Somer (Harrell, Lee et Mark 1996) – est fréquemment utilisée comme mesure plus naturelle que l'ASC, parce que sa valeur varie de 0 à 1.

Copas (1999) propose la courbe de type logit sur rangs comme alternative à la courbe ROC pour évaluer le pouvoir prédicteur d'un score de propension. Si le score de propension est calculé d'après une régression logistique, la courbe logit sur rangs est simplement la représentation graphique du prédicteur linéaire issu du modèle en fonction de la transformation logistique du rang proportionnel du score de propension. Plus généralement, il s'agit d'une représentation graphique de logit(p_i), où p_i est la probabilité estimée d'après toute forme de (1), c'est-à-dire $p(D|\mathbf{x}, \mathbf{x}^*, \mathbf{z})$, en fonction des logits des rangs proportionnels (r/n) où r est le rang du cas i ($i = 1, ..., n$) sur le score de propension. Cette relation est habituellement presque linéaire et sa pente – qui peut varier de zéro à un – est une

mesure du pouvoir prédicteur du score de propension. Selon Copas, la pente est plus sensible aux changements de spécification du modèle de propension à répondre et à ceux de la prévalence du résultat que ne l'est le coefficient de Gini. Une bonne estimation de la pente peut être obtenue en calculant les quantiles de la variable sur les axes des y et des x, puis en ajustant un simple modèle de régression.

La mesure dans laquelle les scores de propension à répondre permettent de distinguer les répondants des non-répondants est un indicateur de l'efficacité de tout ajustement statistique pour tenir compte des données manquantes. Un manque de pouvoir de discrimination donne à penser que des prédicteurs importants manquent dans le score de propension à répondre ou qu'une part importante du processus qui dicte l'existence des données manquantes est essentiellement aléatoire. La mesure dans laquelle les scores de propension à répondre prédisent si un cas sera un non-répondant aux vagues subséquentes – et de quel type de non-répondant il s'agira – est un indice du succès qu'aura toute intervention destinée à réduire la non-réponse.

3. La Millennium Cohort Study

L'échantillon de la première vague de la Millennium Cohort Study (MCS) réalisée au Royaume-Uni comprend 18 552 familles dans lesquelles est né un enfant au cours d'une période de 12 mois durant les années 2000 et 2001 et qui vivaient dans des circonscriptions électorales choisies du Royaume-Uni au moment où l'enfant était âgé de 9 mois. Le taux initial de réponse était de 72 %. Les secteurs où les proportions de familles noires et asiatiques sont élevées, les secteurs défavorisés et les trois plus petits pays du Royaume-Uni sont tous surreprésentés dans l'échantillon qui est structuré en grappes et stratifié de façon disproportionnée, comme l'a décrit Plewis (2007b). Les quatre premières vagues ont eu lieu lorsque les enfants membres de la cohorte étaient âgés (environ) de 9 mois, 3 ans, 5 ans et 7 ans. À la deuxième vague, 19 % de l'échantillon cible, qui n'inclut pas les enfants décédés et émigrés, ont été non productifs. Les cas non productifs étaient répartis de manière égale entre les cas de non-réponse à la vague et les cas d'attrition, et entre les cas de refus et les autres cas non productifs (pas localisés, pas contactés, *etc.*).

4. Analyse de la non-réponse

4.1 Exactitude de la discrimination et de la prédiction

Plewis (2007a) et Plewis et coll. (2008) montrent que les variables mesurées durant la première vague de la MCS qui sont associées à l'attrition durant la deuxième vague ne sont pas nécessairement associées à la non-réponse à une vague à ce moment-là (et inversement). Il en est de même des corrélats du refus et des autres cas non productifs. Le tableau 1 donne les estimations de l'exactitude d'après les modèles de propension à répondre. L'estimation du coefficient de Gini pour la non-réponse globale (0,38) est relativement faible : elle correspond à une ASC de 0,69 qui est la probabilité d'attribuer correctement (en se fondant sur leurs probabilités prédites) une paire de cas (un répondant et un non-répondant), ce qui indique que la discrimination entre les non-répondants et les répondants d'après les scores de propension à répondre n'est pas particulièrement bonne. Elle est légèrement meilleure pour les non-répondants à une vague que pour les cas d'attrition, et est nettement meilleure pour les autres cas non productifs que pour les refus. Ces estimations ont été obtenues au moyen de comparaisons par paire de chaque catégorie de non-réponse avec le fait d'être un répondant. Un tableau comparable se dégage lorsque l'on examine les pentes des courbes logit sur rangs, quoique celles-ci fassent ressortir plus clairement les différences de prédictivité pour les différents types de non-réponse et pour les raisons de la non-réponse.

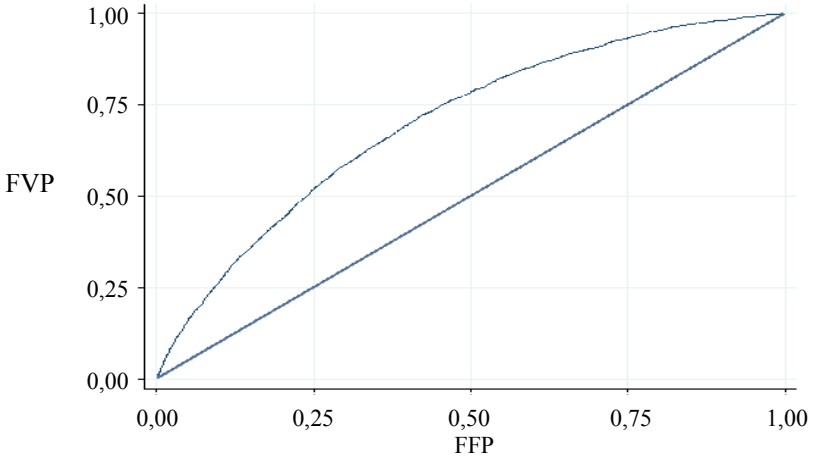


Figure 1 Courbe ROC

Tableau 1
Estimations de l'exactitude d'après les modèles de propension à répondre, deuxième vague de la MCS

Mesure de l'exactitude	Non-réponse globale[2]	Type de non-réponse[2]		Raison de la non-réponse[2]	
		Non-réponse à une vague	Attrition	Refus	Autres cas non productifs
ASC [1]	0,69	0,71	0,69	0,68	0,77
Gini [1]	0,38	0,42	0,39	0,37	0,53
Courbe logit sur rangs : pente[1]	0,45	0,51	0,44	0,40	0,63
Taille de l'échantillon	18 230	16 210	16 821	16 543	16 513

[1] ASC estimée sous l'hypothèse binormale (Krzanowski et Hand 2009) ; limites de confiance à 95 % pour a) ASC n'excédant pas ± 0,015, b) coefficient de Gini et pente de la courbe logit sur rangs n'excédant pas ± 0,03.
[2] Fondé sur une régression logistique, avec prise en compte du plan de sondage en utilisant les commandes *svy* de STATA avec la taille d'échantillon correspondant à la somme des cas productifs et des cas de non-réponse selon la catégorie.

La spécification correcte des modèles destinés à expliquer la non-réponse peut être difficile. De nouveaux candidats susceptibles d'être inclus dans un modèle peuvent apparaître après que l'on ait estimé ce dernier ainsi que les pondérations correspondantes par l'inverse de la probabilité, tandis que d'autres demeurent inconnus. Quelle est l'importance de l'effet que pourrait avoir sur les mesures de l'exactitude l'inclusion de nouvelles variables ? Ici, nous examinons les effets de l'ajout de trois nouvelles variables aux modèles de la MCS : i) le fait que les répondants donnent ou non leur consentement pour que leurs réponses à l'enquête soient appariées aux dossiers de santé à la première vague ; ii) un score de conditions dans le quartier calculé d'après les observations faites par l'intervieweur à la deuxième vague, et iii) le fait que, à la première vague, le répondant principal a ou non déclaré avoir voté aux dernières élections générales au Royaume-Uni. Les deux premières de ces variables n'étaient pas disponibles pour les analyses résumées au tableau 1 : le refus du consentement à la vague t pourrait être suivi par un refus global à la vague $t + 1$, et la non-réponse pourrait être plus fréquente dans les quartiers pauvres. La variable de vote est un indicateur de l'engagement social qui pourrait être associé à la probabilité de répondre. Comme le score des conditions dans le quartier n'a pas pu être obtenu pour les cas qui n'ont pas été localisés, nous utilisons cette variable uniquement dans le modèle où sont comparés les cas de refus aux cas productifs.

Le tableau 2 donne les résultats en utilisant la même méthode d'estimation qu'au tableau 1, ainsi que les niveaux correspondants de précision. Nous voyons (d'après les notes) que chacune des trois variables est associée à au moins un type de non-réponse. L'augmentation de l'exactitude de l'ASC est plus importante que celle à laquelle on s'attendrait par hasard ($p < 0,001$ sauf pour la non-réponse à une vague : $p > 0,06$), mais est faible excepté pour les refus, pour lesquels l'inclusion des trois nouvelles variables entraînent une différence : l'estimation du coefficient de Gini passe de 0,37 à 0,41 et la pente de la courbe logit sur rangs passe de 0,40 à 0,45 (quoique les données manquantes pour le score des conditions dans le quartier réduisent la taille de l'échantillon).

4.2 Utilisation des pondérations pour corriger la non-réponse

Bien que la non-réponse à la deuxième vague de la MCS soit systématiquement reliée à un certain nombre de variables mesurées durant la première vague ou après, nous avons constaté que la capacité du modèle à faire la distinction entre les catégories de non-réponse et à prédire ces dernières n'est pas très grande. Nous allons maintenant considérer l'effet que les pondérations produites à partir des modèles de propension à répondre ont sur une estimation longitudinale d'intérêt. Nous nous concentrons sur les transitions entre l'absence d'emploi et la possession d'un emploi entre les deux vagues. Comme le soutient Groves (2006), la clé en vue de résoudre le problème de biais causé par les données manquantes consiste à trouver des variables qui prédisent si un élément de données manquera et lesquelles de ces variables prédisant les données manquantes sont également reliées à la variable d'intérêt. Nous constatons que toutes les variables qui prédisent la non-réponse globale sont également associées au fait que la répondante principale travaille ou ne travaille pas à la deuxième vague, conditionnellement au fait qu'elle travaillait à la première vague, de sorte que nous devrions nous attendre à ce que l'application des poids de non-réponse réduise le biais. Les résultats, présentés au tableau 3, montrent que, comparativement à l'utilisation des poids de sondage uniquement, l'ajout des poids de non-réponse fondés sur le modèle qui sous-tend le tableau 1 produit de petites corrections des probabilités de transition estimées. Par contre, les variables de consentement et de vote n'ont aucun effet supplémentaire, ce qui est en harmonie avec l'augmentation marginale de l'exactitude indiquée au tableau 2.

Tableau 2
Estimation de l'exactitude pour les modèles de propension à répondre améliorés, deuxième vague de la MCS

Mesure de l'exactitude	Non-réponse globale[1]	Type de non-réponse		Raison de la non-réponse	
		Non-réponse à une vague[2]	Attrition[3]	Refus[4]	Autres cas non productifs[5]
ASC	0,70	0,72	0,71	0,70	0,77
Gini	0,41	0,44	0,41	0,41	0,54
Courbe logit sur rangs : pente	0,47	0,52	0,46	0,45	0,65
Taille de l'échantillon	18 148	16 177	16 745	15 656	16 443

(1) Inclut le consentement (Rapport de cotes (RC) = 2,1, e.-t.= 0,20) et le vote (RC= 1,4, e.-t. = 0,08).
(2) Inclut le vote seulement (RC = 1,4, e.-t. = 0,11), consentement pas important (t = 1,33 ; p > 0,18).
(3) Inclut le consentement (RC = 2,7, e.-t. = 0,26) et le vote (RC = 1,4, e.-t. = 0,09).
(4) Inclut le consentement (RC = 2,6, e.-t. = 0,32), le vote (RC = 1,3, e.-t. = 0,10) et le score du quartier (RC = 1,02, e.-t. = 0,014).
(5) Inclut le consentement (RC = 1,6, e.-t. = 0,20) et le vote (RC = 1,5, e.-t. = 0,11).

Tableau 3
Transitions d'emploi pondérées (erreurs-types), deuxième vague de la MCS

Variable	Poids de sondage seulement	Poids global[1]	Poids global[2]
Pas de changement	0,30 (0,0053)	0,30 (0,0056)	0,31 (0,0056)
Emploi → pas d'emploi	0,34 (0,0059)	0,35 (0,0059)	0,35 (0,0060)
Pas d'emploi → emploi	0,37 (0,0073)	0,35 (0,0073)	0,35 (0,0073)
Étendue du poids[3]	0,23 – 2,0	0,19 – 4,1	0,19 – 6,3
Taille de l'échantillon	14 891	14 796	14 733

(1) Fondé sur le produit des poids de sondage et des poids de non-réponse en utilisant le modèle qui sous-tend le tableau 1.
(2) Poids de non-réponse fondé sur un modèle qui inclut les variables de consentement et de vote.
(3) Tous les poids sont normalisés de manière que leur moyenne soit égale à un.

5. Discussion

Les méthodologistes d'enquête qui travaillent avec des données longitudinales sont confrontés depuis longtemps au problème de non-réponse. Presque toutes les études longitudinales souffrent d'une accumulation des cas de non-réponse au fil du temps. Il est fréquent, même pour des études bien réalisées et bien établies, d'obtenir des données pour moins de la moitié de l'échantillon cible. Par ailleurs, il est possible d'en savoir beaucoup sur les corrélats de différents types de non-réponse en s'appuyant sur les variables auxiliaires provenant de vagues antérieures. L'objectif principal du présent article était de présenter un moyen différent de réfléchir à l'utilité des approches qui s'appuient sur des modèles linéaires généraux à la fois pour construire des pondérations par l'inverse des probabilités et pour faciliter les imputations. Traiter les prédicteurs linéaires issus des modèles de régression comme des scores de propension à répondre, puis créer de courtes ROC offre des méthodes pour résumer l'information contenue dans ces scores afin de l'utiliser pour évaluer l'exactitude de la discrimination et de la prédiction pour différents types de non-réponse.

L'application de cette approche à la Millennium Cohort Study a montré que, même en utilisant une vaste gamme de variables explicatives, le pouvoir de discrimination demeure faible. L'une des implications de cette constatation est que certains cas de non-réponse découlent de facteurs circonstanciels, sans importance pris individuellement, qui peuvent raisonnablement être considérés comme le hasard. Notre étude étaye dans une certaine mesure cette hypothèse en ce sens que l'exactitude des modèles pour la non-réponse globale, la non-réponse à une vague et les autres cas non productifs (les deux dernières catégories étant reliées) a été peu modifiée par l'introduction des variables de vote et de consentement. Par ailleurs, ces variables (et le score de conditions du quartier) ont amélioré le pouvoir de discrimination entre les cas productifs, d'une part, et les cas d'attrition et de refus (qui sont aussi reliés), d'autre part. Néanmoins, le pouvoir de discrimination pour ces deux catégories demeurait plus faible que pour les autres types de non-réponse. Une deuxième implication éventuelle est que les modèles ne réalisent pas bien la discrimination parce que l'on a affaire à une répartition des données manquantes qui n'est pas due au hasard (NMAR pour *not missing at rando m*) au sens de Little et Rubin (2002). Autrement dit, il pourrait se produire après la vague précédente des changements de circonstances qui influencent la non-réponse durant la vague courante.

Les implications de nos constatations en ce qui concerne la prédiction sont qu'il pourrait être difficile de prédire quels cas deviendront des non-répondants avec un haut degré

d'exactitude. Pour être efficaces, les interventions pour prévenir la non-réponse dans les études longitudinales doivent être ciblées sur les cas les moins susceptibles de répondre, parce que ceux-ci sont probablement ceux qui diffèrent le plus des répondants et, par conséquent, constituent la source principale de biais. C'est dans cette situation que la méthode des courbes ROC peut être particulièrement utile, parce que, comme le montre Swets, Dawes et Monahan (2000), il est possible de déterminer le seuil optimal pour le score de propension à répondre fondé sur les coûts et les avantages de l'intervention d'après les taux de vrais et de faux positifs qu'implique le seuil. Une évaluation plus détaillée de ces questions dépasse le cadre du présent article, mais comprendrait l'examen d'interventions pour prévenir différents types de non-réponse, et les avantages des réductions éventuelles du biais et de la variabilité découlant d'un échantillon de plus grande taille et dont les caractéristiques sont plus proches de celles de l'échantillon cible.

Remerciements

La présente étude a été financée par l'Economic and Social Research Council du Royaume-Uni dans le cadre de la Survey Design and Measurement Initiative (réf. RES-175-25-0010).

Bibliographie

Copas, J. (1999). The effectiveness of risk scores: The logit rank plot. *Applied Statistics*, 48, 165-183.

Groves, R.M. (2006). Nonresponse rates and non-response bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Harrell, F.E. Jr., Lee, K.L. et Mark, D.B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361-387.

Hawkes, D., et Plewis, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Roy al Statistical Society A*, 169, 479-491.

Krzanowski, W.J., et Hand, D.J. (2009). *ROC Curves for Continuous Data*. Boca Raton, Fl. : Chapman and Hall/CRC.

Lepkowski, J.M., et Couper, M.P. (2002). Nonresponse in the second wave of longitudinal household surveys. Dans *Survey Nonresponse*, (Éds., R.M. Groves et coll.). New York : John Wiley & Sons, Inc.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2e Éd.). New York : John Wiley & Sons, Inc.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford : OUP.

Plewis, I. (2007a). Non-response in a birth cohort study: The case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, 10, 325-334.

Plewis, I. (Éd.) (2007b). *The Millennium Cohort Study: Technical Report on Sampling* (4e Éd.). Londres : Institute of Education, University of London.

Plewis, I., Ketende, S.C., Joshi, H. et Hughes, G. (2008). The contribution of residential mobility to sample loss in a birth cohort study: Evidence from the first two waves of the Millennium Cohort Study. *Journal of Official Statistics*, 24, 365-385.

Swets, J.A., Dawes, R.M. et Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Sciences in the Public Interest*, 1, 1-26.

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 187-203
Statistique Canada, N° 12-001-X au catalogue

Estimation des intervalles de confiance des paramètres de petit domaine avec rétrécissement des moyennes et des variances

Sarat C. Dass, Tapabrata Maiti, Hao Ren et Samiran Sinha [1]

Résumé

Nous proposons une nouvelle approche d'estimation sur petits domaines fondée sur la modélisation conjointe des moyennes et des variances. Le modèle et la méthodologie que nous proposons améliorent non seulement les estimateurs sur petits domaines, mais donnent aussi des estimateurs « lissés » des vraies variances d'échantillonnage. Le maximum de vraisemblance des paramètres du modèle est estimé au moyen de l'algorithme EM en raison de la forme non classique de la fonction de vraisemblance. Les intervalles de confiance des paramètres de petit domaine sont obtenus en adoptant une approche de la théorie de la décision plus générale que l'approche classique de minimisation de la perte quadratique. Les propriétés numériques de la méthode proposée sont étudiées au moyen d'études par simulation et comparées à celles de méthodes concurrentes proposées dans la littérature. Une justification théorique des propriétés effectives des estimateurs et intervalles de confiance résultants est également présentée.

Mots clés : Algorithme EM ; Bayes empirique ; modèles hiérarchiques ; échantillonnage réjectif ; variance d'échantillonnage ; estimation sur petits domaines.

1. Introduction

L'estimation sur petits domaines et les techniques statistiques qui s'y rapportent sont des sujets qui ont fait l'objet d'une attention croissante ces dernières années. De nombreux organismes, tant publics que privés, cherchent à obtenir des estimations sur petits domaines fiables pour prendre des décisions stratégiques utiles. La surveillance de la situation socioéconomique et de l'état de santé de divers groupes définis selon l'âge, le sexe et la race pour lesquels s'observent des tendances distinctes sur de petites régions géographiques est un exemple d'application pratique des techniques d'estimation sur petits domaines.

Il est aujourd'hui généralement reconnu que les estimations directes à partir de données d'enquête calculées pour les petits domaines ne sont d'ordinaire pas fiables parce que leurs erreurs-types et coefficients de variation sont très souvent grands. Il devient donc nécessaire d'obtenir de meilleures estimations, d'une plus grande précision. Des approches fondées explicitement ou implicitement sur un modèle sont élaborées pour relier des petits domaines et obtenir une plus grande précision par « emprunt d'information » à des domaines similaires. Cette technique d'estimation est également appelée estimation par rétrécissement, ou estimation à rétrécisseur, puisque les estimations directes sont « rétrécies » afin qu'elles se rapprochent de la moyenne globale. Les estimations directes d'après les données d'enquête et les variances d'échantillon sont les principaux ingrédients qui entrent dans la création des modèles d'estimation sur petits domaines de niveau agrégé. La stratégie de modélisation repose habituellement sur l'hypothèse que les variances d'échantillonnage sont connues, tandis qu'un modèle de régression linéaire approprié est utilisé pour les moyennes. Pour des renseignements détaillés sur ces développements, le lecteur est invité à consulter Ghosh et Rao (1994), Pfeffermann (2002) et Rao (2003). Les modèles habituels au niveau du domaine suscitent deux critiques importantes. Premièrement, en pratique, les variances d'échantillonnage sont des quantités estimées qui sont donc sujettes à d'importantes erreurs. Il en est ainsi parce qu'elles sont souvent fondées sur des tailles d'échantillon équivalentes à celles qui servent au calcul des estimations directes. Deuxièmement, en raison de l'hypothèse que les variances d'échantillonnage sont connues et fixes formulée dans les modèles d'estimation sur petits domaines classiques, l'incertitude que comporte l'estimation de la variance n'est pas prise en compte dans la stratégie d'inférence globale.

Des tentatives en vue de modéliser uniquement les variances d'échantillonnage ont été faites antérieurement ; voir, par exemple, Maples, Bell et Huang (2009), Gershunskaya et Lahiri (2005), Huff, Eltinge et Gershunskaya (2002), Cho, Eltinge, Gershunskaya et Huff (2002), Valliant (1987), et Otto et Bell (1995). Dans leurs articles, Wang et Fuller (2003) et Rivest et Vandal (2003) ont étendu l'estimation de l'erreur quadratique moyenne (EQM) asymptotique des estimateurs sur petits domaines au cas où l'on estime les variances d'échantillonnage au lieu de s'appuyer sur l'hypothèse classique que les variances sont connues. En outre, You et Chapman (2006) ont considéré la modélisation des variances d'échantillonnage avec inférence en appliquant des techniques d'estimation entièrement bayésiennes.

De nombreux praticiens ont jugé nécessaire de modéliser la variance. Les progrès les plus récents dans ce domaine

1. Sarat C. Dass et Tapabrata Maiti, Department of Statistics & Probability, Michigan State University. Courriel : maiti@stt.msu.edu ; Hao Ren, CTB/McGraw-Hill, 20 Ryan Ranch Rd, Monterey, CA 93940 ; Samiran Sinha, Department of Statistics, Texas A & M University.

sont résumés élégamment dans un article publié en 2008 par William Bell, du *United States Census Bureau*. Ce dernier a examiné minutieusement les conséquences de ces problèmes dans le contexte de l'estimation de l'EQM des estimateurs sur petits domaines fondés sur un modèle. Il a également donné des preuves numériques de l'estimation de l'EQM pour le modèle de Fay-Herriot (donné dans l'équation 1) quand il est supposé que les variances d'échantillonnage sont connues. Les progrès exposés jusqu'à présent dans la littérature traitant des petits domaines peuvent être considérés « grosso modo » comme étant i) le lissage des estimations directes des variances des erreurs d'échantillonnage pour obtenir des estimations des variances plus stables dont le biais est faible et ii) la prise en compte (partielle) de l'incertitude dans les variances d'échantillonnage en étendant le modèle de Fay-Herriot.

Manifestement, l'effort en vue de bien tenir compte des variances d'échantillonnage dans la modélisation de la moyenne a été faible, voire nul, comparativement au nombre d'études consacrées à la modélisation et à l'inférence des moyennes. Le développement systématique du « rétrécissement » des moyennes ainsi que des variances fait défaut dans la littérature traitant de l'estimation sur petits domaines. Autrement dit, nous aimerions exploiter la technique de l'« emprunt d'information » à d'autres petits domaines en vue d'« améliorer » les estimations de la variance, tout comme nous le faisons pour « améliorer » les estimations des moyennes de petits domaines. Nous proposons un modèle hiérarchique utilisant à la fois les estimations directes d'après les données d'enquête et les estimations des variances d'échantillonnage pour inférer les paramètres du modèle qui déterminent le système stochastique. Notre objectif méthodologique est d'élaborer l'estimation « par rétrécissement » double pour les moyennes ainsi que les variances de petit domaine, en exploitant la structure de la modélisation conjointe moyenne-variance afin que les estimateurs finaux soient plus précis. Des preuves numériques montrent l'efficacité du rétrécissement double appliqué aux estimations sur petits domaines de la moyenne si l'on prend pour critère l'EQM.

Une autre contribution importante du présent article est l'obtention d'intervalles de confiance pour les moyennes de petits domaines. La littérature relative à l'estimation sur petits domaines traite avant tout des estimations ponctuelles et de leurs erreurs-types ; pourtant, il est bien connu que la pratique classique consistant à utiliser [estimation ponctuelle $\pm\, q \times$ erreur-type], où q est la valeur seuil Z (normale standard) ou t, ne produit pas des probabilités de couverture exactes des intervalles ; voir Hall et Maiti (2006) et Chatterjee, Lahiri et Li (2008) pour plus de précisions. Les travaux antérieurs, qui sont fondés sur la technique du bootstrap, sont d'un usage limité en raison de l'estimation répétée des paramètres du modèle. Nous produisons des intervalles de confiance pour les moyennes dans une perspective de théorie de la décision. La construction des intervalles de confiance est facile à mettre en œuvre en pratique.

La présentation de la suite de l'article est la suivante. Le modèle hiérarchique proposé pour les moyennes et les variances d'échantillonnage est élaboré à la section 2. L'estimation des paramètres du modèle au moyen de l'algorithme EM est exposée à la section 3. La justification théorique des intervalles de confiance proposés et leurs propriétés de couverture sont présentées à la section 4. Une étude par simulation et un exemple fondé sur des données réelles sont présentés aux sections 5 et 6, respectivement. Enfin, une discussion et certaines conclusions sont présentées à la section 7. Une autre formulation du modèle pour les petits domaines, ainsi que des détails mathématiques sont donnés en annexe.

2. Modèle proposé

Supposons que l'on examine n petits domaines. Pour le i^{e} petit domaine, soit de (X_i, S_i^2) la paire comprenant l'estimation directe et la variance d'échantillonnage, pour $i = 1, 2, ..., n$. Soit $\boldsymbol{Z}_i = (Z_{i1}, ..., Z_{ip})^T$ le vecteur de p covariables disponibles à l'étape de l'estimation pour le i^{e} petit domaine. Nous proposons le modèle hiérarchique suivant :

$$\left.\begin{aligned} X_i \,|\, \theta_i, \sigma_i^2 &\sim \text{Normale}(\theta_i, \sigma_i^2) \\ \theta_i &\sim \text{Normale}(\boldsymbol{Z}_i^T \boldsymbol{\beta}, \tau^2) \end{aligned}\right\} \qquad (1)$$

$$\left.\begin{aligned} \frac{(n_i - 1)S_i^2}{\sigma_i^2} \,\Big|\, \sigma_i^2 &\sim \chi^2_{n_i - 1} \\ \sigma_i^{-2} &\sim \text{Gamma}(a, b), \end{aligned}\right\} \qquad (2)$$

indépendamment pour $i = 1, 2, ..., n$. Dans l'élaboration du modèle, n_i est la taille d'un échantillon aléatoire simple (EAS) tiré du i^{e} domaine, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ est le vecteur de dimension $p \times 1$ des coefficients de régression, et $\boldsymbol{B} \equiv (a, b, \boldsymbol{\beta}, \tau^2)^T$ est la série complète de paramètres inconnus dans le modèle. En outre, $\text{Gamma}(a, b)$ est la densité de probabilité Gamma dont les paramètres de forme et d'échelle a et b, respectivement, sont positifs, définis comme étant $f(x) = \{b^a \Gamma(a)\}^{-1} e^{-x/b} x^{a-1}$ pour $x > 0$, et 0 autrement. Le terme σ_i^2 inconnu est la variance réelle de X_i et est habituellement estimé par la variance d'échantillon S_i^2. On suppose généralement que les S_i^2 suivent une loi du khi-carré possédant $(n_i - 1)$ degrés de liberté (en raison de la normalité et de l'EAS), mais nous notons que sous des plans de sondage complexes, le nombre de degrés de liberté

doit être déterminé prudemment (par exemple, Maples et coll. 2009). Surtout, le rôle de taille d'échantillon dans l'estimation par rétrécissement de σ_i^2 est le suivant : l'estimation de σ_i^{-2} se rapproche davantage de la moyenne globale (ab) pour de faibles valeurs de n_i que pour des valeurs élevées. Donc, pour les variances, les tailles d'échantillon jouent le même rôle que la précision dans l'estimation par rétrécissement des moyennes de petit domaine. Nous notons que You et Chapman (2006) ont également envisagé un deuxième niveau de modélisation de la variance d'échantillonnage. Cependant, les hyperparamètres reliés à la loi a priori de σ_i^2 ne sont pas dictés par les données mais plutôt choisis de façon telle que la loi a priori soit vague. Donc, leur modèle peut être considéré comme la version bayésienne de modèle examiné dans Rivest et Vandal (2003) et dans Wang et Fuller (2003). Le deuxième niveau de modélisation de σ_i^{-2} dans (2) peut être étendu encore davantage à $\sigma_i^{-2} \sim \text{Gamma}(b, \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}_2)/b)$ de sorte que $E(\sigma_i^{-2}) = \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}_2)$ pour un autre jeu de p coefficients de régression $\boldsymbol{\beta}_2$ afin d'inclure l'information sur les covariables dans la modélisation de la variance.

Bien que notre modèle soit motivé par Hwang, Qiu et Zhao (2009), nous tenons à mentionner que Hwang et coll. (2009) ont considéré les moyennes et variances par rétrécissement dans le contexte de données micro vectorielles où ils ont préconisé une solution importante consistant à insérer un estimateur à rétrécisseur de la variance dans l'estimateur de la moyenne. L'estimateur par rétrécissement de la variance dans Hwang et coll. (2009) est une fonction de S_i^2 seulement et non de X_i ainsi que S_i^2 ; voir les remarques 2 et 3 à la section 2. Donc, l'inférence de la moyenne ne tient pas compte de toute l'incertitude dans l'estimation de la variance. En outre, leur modèle ne contient aucune information sur les covariables. L'étude par simulation décrite plus loin indique que notre méthode d'estimation donne de meilleurs résultats que celle de Hwang et coll. (2009).

Dans la formulation du modèle susmentionné, l'inférence pour le paramètre θ_i représentant la moyenne de petit domaine peut être faite en se basant sur la distribution conditionnelle de θ_i sachant toutes les données $\{(X_i, S_i^2, \boldsymbol{Z}_i), i = 1, ..., n\}$. Sous notre modèle, la distribution conditionnelle de θ_i est une distribution non standard qui ne possède pas de forme analytique et requiert donc des méthodes numériques, telles que la méthode de Monte Carlo et l'algorithme EM, pour l'inférence. Des renseignements détaillés sont fournis à la section suivante.

3. Méthodologie d'inférence

3.1 Estimation des paramètres inconnus au moyen de l'algorithme EM

En pratique, $\boldsymbol{B} \equiv (a, b, \boldsymbol{\beta}, \tau^2)^T$ est inconnu et doit être estimé d'après les données $\{(X_i, S_i^2, \boldsymbol{Z}_i), i = 1, 2, ..., n\}$. Nous proposons d'estimer $\boldsymbol{B}$ par la méthode du maximum de vraisemblance marginale : estimer $\boldsymbol{B}$ par $\hat{\boldsymbol{B}}$ où $\hat{\boldsymbol{B}}$ maximise la vraisemblance marginale $L_M(\boldsymbol{B}) = \prod_{i=1}^n L_{M,i}(\boldsymbol{B})$, où

$$L_{M,i} \propto \frac{\Gamma(n_i/2+a)}{\tau\Gamma(a)b^a} \int \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\} \psi_i^{-(n_i/2+a)} d\theta_i, \quad (3)$$

et

$$\psi_i \equiv \left\{0,5(X_i - \theta_i)^2 + 0,5(n_i - 1)S_i^2 + \frac{1}{b}\right\}. \quad (4)$$

La vraisemblance marginale L_M contient des intégrales qui ne peuvent pas être évaluées en forme analytique, de sorte que l'on doit recourir à des méthodes numériques pour sa maximisation. L'un de ces algorithmes est la procédure itérative EM (espérance-maximisation) qui est utilisée quand on a affaire à ce genre d'intégrales. L'algorithme EM comprend l'augmentation de la vraisemblance observée $L_M(\boldsymbol{B})$ présentant des données manquantes ; dans notre cas, les variables de l'intégration, θ_i, $i = 1, 2, ..., n$, constituent cette information manquante. Sachant $\boldsymbol{\theta} \equiv \{\theta_1, \theta_2, ..., \theta_n\}$, la log-vraisemblance ($\ell_c$) sous *données complètes* peut s'écrire

$$\begin{aligned}\ell_c(\boldsymbol{B}, \boldsymbol{\theta}) = \sum_{i=1}^n \Bigg[& \log\{\Gamma(n_i/2 + a)\} - \log\{\Gamma(a)\} \\ & - a\log(b) - 0,5\log(\tau^2) \\ & - \frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2} - (n_i/2 + a)\log(\psi_i) \Bigg],\end{aligned}$$

où l'expression de ψ_i est donnée par l'équation (4). Partant d'une valeur initiale de $\boldsymbol{B}$, disons $\boldsymbol{B}^{(0)}$, l'algorithme EM exécute itérativement une maximisation par rapport à $\boldsymbol{B}$. À la t^{e} étape, la fonction d'objectif maximisée est

$$\begin{aligned}Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)}) &= E(\ell_c(\boldsymbol{B}, \boldsymbol{\theta})) \\ &= \sum_{i=1}^n \Bigg[\log\{\Gamma(n_i/2 + a)\} - \log\{\Gamma(a)\} \\ &\quad - a\log(b) - 0,5\log(\tau^2) \\ &\quad - \frac{E(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2} - (n_i/2 + a)E\{\log(\psi_i)\} \Bigg].\end{aligned}$$

Dans $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$, l'espérance est prise par rapport à la distribution conditionnelle de chaque θ_i sachant les données, $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}^{(t-1)})$, ce qui est

$$\begin{aligned}\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) \propto & \\ \exp\{-0,5(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2/\tau^2\} & \psi_i^{-(n_i/2+a)}. \quad (5)\end{aligned}$$

L'une des difficultés ici est que les espérances ne sont pas disponibles sous une forme analytique. Donc, nous

recourons à une méthode de Monte Carlo pour évaluer l'expression. Supposons que R échantillons iid de θ_i soient disponibles, disons $\theta_{i,1}, \theta_{i,2}, ..., \theta_{i,R}$. Alors, chaque expression de la forme $E\{h(\theta_i)\}$ peut être approximée par la moyenne Monte Carlo

$$E\{h(\theta_i)\} \approx \frac{1}{R}\sum_{r=1}^{R} h(\theta_{i,k}). \tag{6}$$

Cependant, le tirage de nombres aléatoires de la distribution conditionnelle $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}^{(t-1)})$ n'est pas simple non plus, puisqu'il ne s'agit pas d'une densité standard. Les échantillons sont sélectionnés par la procédure d'acceptation-rejet (Robert et Casella 2004) : pour tirer un échantillon de la densité cible f, tirer un échantillon x de la loi instrumentale g, et l'accepter comme étant un échantillon tiré de f avec la probabilité $f(x)/\{M^* g(x)\}$, où $M^* = \sup_x \{f(x)/g(x)\}$. Un avantage de la méthode d'acceptation-rejet est que la densité cible f ne doit être connue que jusqu'à une constante de proportionnalité, ce qui est le cas pour $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}^{(t-1)})$ dans (5) ; étant donné la forme non standard de la densité, la constante de normalisation ne peut pas être obtenue sous une forme analytique. Pour l'algorithme d'acceptation-rejet, nous avons utilisé la densité normale $g(\theta_i) \propto \exp\{-0{,}5(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2/\tau^2\}$ comme loi instrumentale. La probabilité d'acceptation se calcule comme étant $[\{1/b + 0{,}5(n_i-1)S_i^2\}/\{1/b + 0{,}5(n_i-1)S_i^2 + 0{,}5(\theta_i - X_i)^2\}]^{n_i/2+a}$. Si l'on veut augmenter la probabilité d'acceptation, on peut choisir une meilleure loi instrumentale ou un algorithme différent (tel que les algorithmes d'échantillonnage réjectif adaptatif ou d'enveloppe d'acceptation-rejet), mais la loi instrumentale que nous avons choisie a donné des résultats satisfaisants dans les études que nous avons effectuées.

Le maximiseur de $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$ à la t^{e} étape peut être décrit explicitement. Les solutions pour $\boldsymbol{\beta}$ et τ^2 sont disponibles sous les formes analytiques suivantes

$$\boldsymbol{\beta}^{(t)} = \left(\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^T\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{Z}_i E(\theta_i)\right)$$

et

$$(\tau^2)^{(t)} = \frac{1}{n}\sum_{i=1}^{n} E(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2,$$

respectivement. En outre, $a^{(t)}$ et $b^{(t)}$ s'obtiennent en résolvant $S_a = \partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})/\partial a = 0$ et $S_b = \partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})/\partial b = 0$ par la méthode de Newton-Raphson où

$$S_a = \sum_{i=1}^{n} \frac{\partial}{\partial a}\log\{\Gamma(n_i/2 + a)\} - n\left\{\frac{\partial}{\partial a}\log\{\Gamma(a)\}\right\} - n\log(b) - \sum_{i=1}^{n} E\{\log(\psi_i)\}$$

et

$$S_b = -\frac{na}{b} + \sum_{i=1}^{n} \frac{(n_i/2 + a)}{b^2}\, E(\psi_i^{-1}).$$

Nous posons que $\boldsymbol{B}^{(t)} = (a^{(t)}, b^{(t)}, \boldsymbol{\beta}^{(t)}, (\tau^{(t)})^2)$ et procédons à la $(t+1)^{\text{e}}$ étape. Cette procédure de maximisation est répétée jusqu'à la convergence de l'estimation $\boldsymbol{B}^{(t)}$. L'EMV de $\boldsymbol{B}$ est $\hat{\boldsymbol{B}} = \boldsymbol{B}^{(\infty)}$ une fois que la convergence est établie.

3.2 Estimation ponctuelle et intervalle de confiance de θ_i

Selon la technique classique, nous posons que l'estimateur sur petits domaines de θ_i est

$$\hat{\theta}_i = E(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})\Big|_{\boldsymbol{B}=\hat{\boldsymbol{B}}}, \tag{7}$$

l'espérance de θ_i par rapport à la densité conditionnelle $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})$ où l'estimation du maximum de vraisemblance $\hat{\boldsymbol{B}}$ est « inséré » pour remplacer $\boldsymbol{B}$. L'estimation $\hat{\theta}_i$ est calculée numériquement en utilisant la procédure Monte Carlo (6) décrite à la section précédente. Dans la suite, le paramètre $\boldsymbol{B}$ inconnu sera remplacé par $\hat{\boldsymbol{B}}$ dans toutes les quantités dans lesquelles il intervient, même si nous continuons d'utiliser la notation $\boldsymbol{B}$ pour simplifier.

En outre, nous élaborons un intervalle de confiance pour θ_i fondé sur une théorie de la décision. Comme l'ont fait Joshi (1969), Casella et Hwang (1991), Hwang et coll. (2009), considérons la fonction de perte associée à l'intervalle de confiance C donnée par $(k/\sigma)L(C) - I_C(\theta)$, où k est un paramètre de mise au point indépendant des paramètres du modèle, $L(C)$ est la longueur de C et $I_C(\theta)$ est la fonction indicatrice prenant la valeur 1 ou 0 selon que $\theta \in C$ ou non. Notons que cette fonction de perte tient compte à la fois de la probabilité de couverture et de la longueur de l'intervalle ; la quantité positive (k/σ) sert de poids relatif de la longueur comparativement à la probabilité de couverture de l'intervalle de confiance. Si $k = 0$, la longueur de l'intervalle n'est pas prise en considération, de sorte que la valeur optimale de C est $(-\infty, \infty)$ avec une probabilité de couverture de 1. Par ailleurs, pour $k = \infty$, la probabilité de couverture est égale à 0, de sorte que la valeur optimale de C est un ensemble de points. Pour obtenir l'intervalle de confiance de Bayes de θ_i il faut minimiser la fonction de risque (la perte prévue) $E\{[(k/\sigma)L(C) - I_C(\theta)] \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})\}$. Le choix optimal de C est donné par

$$C_i(\boldsymbol{B}) = \{\theta_i : kE(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) < \pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})\}. \tag{8}$$

Puisque que $C_i(\boldsymbol{B})$ s'obtient en minimisant le risque a posteriori, on pourrait vouloir l'interpréter comme un

ensemble crédible bayésien. Cependant, à l'instar de Casella et Berger (1990, page 470), nous continuerons de donner à $C_i(\boldsymbol{B})$ le nom d'intervalle de confiance. Dans une perspective bayésienne empirique également, cette terminologie est plus appropriée. Nous montrerons à la section 3.3 comment le paramètre de mise au point k détermine le niveau de confiance de $C_i(\boldsymbol{B})$.

En supposant pour le moment que k est connu, nous suivons les étapes ci-après pour calculer $C_i(\boldsymbol{B})$. Les densités conditionnelles de σ_i^2 et θ_i sont données par

$$\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) \propto \frac{\exp\left[\dfrac{-0{,}5(X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{(\sigma_i^2 + \tau^2) - \left\{0{,}5(n_i - 1)S_i^2 + \dfrac{1}{b}\right\}\left(\dfrac{1}{\sigma_i^2}\right)}\right]}{(\sigma_i^2)^{(n_i-1)/2+a+1}(\sigma_i^2 + \tau^2)^{1/2}} \quad (9)$$

et (5), respectivement, expressions qui, comme nous l'avons mentionné plus haut, n'ont pas de forme analytique. Donc, comme dans le cas de θ_i, nous calculons $E(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})$ numériquement en utilisant la méthode Monte Carlo par approximation de la valeur prévue de la moyenne $1/N \sum_{k=1}^{N} 1/\sigma_{i,k}$, où $\sigma_{i,r}^2$, $r = 1, 2, ..., R$ sont les R échantillons tirés de la densité conditionnelle $\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})$. La procédure d'acceptation-rejet est utilisée pour tirer des nombres aléatoires de $\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})$ avec une loi instrumentale donnée par la loi Gamma inverse

$$\frac{\exp\left[-\left\{0{,}5(n_i - 1)S_i^2 + \dfrac{1}{b}\right\}\left(\dfrac{1}{\sigma_i^2}\right)\right]}{(\sigma_i^2)^{(n_i-1)/2+a+1}},$$

et la probabilité d'acceptation

$$\frac{\exp\left\{\dfrac{-0{,}5(X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{(\sigma_i^2 + \tau^2)}\right\}}{(\sigma_i^2 + \tau^2)^{1/2}} \times \exp(0{,}5) \times |X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}|.$$

L'étape suivante consiste à déterminer les valeurs des bornes de $C_i(\boldsymbol{B})$ en trouvant deux valeurs de θ_i qui satisfont l'équation $kE(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) - \pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) = 0$. Il faut pour cela que la constante de normalisation donnée en (5)

$$D_i = \int_{-\infty}^{\infty} \exp\{-0{,}5(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2/\tau^2\}\, \psi_i^{-(n_i/2+a)}\, d\theta_i$$

soit évaluée numériquement. Nous le faisons en procédant à l'intégration de Gauss-Hermite avec 20 nœuds.

3.3 Choix de k

Nous choisissons pour le paramètre de mise au point k dans (8) l'expression

$$k = k(\boldsymbol{B}) = u_{i,0}\,\phi\left(t_{\alpha/2}\sqrt{\frac{n_i + 2a + 2}{n_i - 1}}\right) \quad (10)$$

où ϕ est la distribution normale standard, $t_{\alpha/2}$ est le $(1 - \alpha/2)^{e}$ centile de la distribution t avec $(n_i - 1)$ degrés de liberté, et $u_{i,0} = \sqrt{1 + \sigma_i^2/\tau^2}$. Puisque $u_{i,0}$ fait intervenir σ_i^2 qui est inconnue, une version estimée $\hat{u}_{i,0}$ s'obtient en introduisant l'estimation du maximum a posteriori

$$\hat{\sigma}_i^2 = \hat{\sigma}_i^2(\hat{\boldsymbol{B}}) = \arg\max_{\sigma_i^2} \pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})\Big|_{\boldsymbol{B}=\hat{\boldsymbol{B}}} \quad (11)$$

à la place de σ_i^2. En outre, nous remplaçons $\boldsymbol{B}$ par $\hat{\boldsymbol{B}}$ dans (11). Nous démontrons que la probabilité de couverture de $C_i(\hat{\boldsymbol{B}})$ avec ce choix de k s'approche de $1 - \alpha$. Les justifications théoriques sont présentées à la section 4.

3.4 Autres méthodes apparentées aux fins de comparaison

Nous donnons à notre méthode le nom de méthode I. Nous décrivons brièvement ci-dessous trois autres méthodes auxquelles nous la comparerons.

Méthode II : Wang et Fuller (2003) ont considéré le modèle d'estimation sur petits domaines de Fay-Herriot donné par (1). Leur principale contribution est la construction de la formule d'estimation de l'erreur quadratique moyenne pour les estimateurs sur petits domaines avec variances d'échantillonnage estimées. Ce faisant, ils ont construit deux formules désignées par $\widehat{\text{EQM}}_1$ et $\widehat{\text{EQM}}_2$. Pour nos comparaisons, nous utilisons $\widehat{\text{EQM}}_1$, qui a été dérivée en suivant l'approche de correction du biais de Prasad et Rao (1990). La différence fondamentale par rapport à notre approche est qu'ils n'ont pas lissé les variances d'échantillonnage, et n'ont tenu compte de l'incertitude que dans l'inférence au sujet des paramètres de petit domaine. La méthode d'estimation des paramètres, qui est fondée sur les moments pour tous les paramètres du modèle, diffère également de la nôtre.

Méthode III : Hwang et coll. (2009) ont considéré les modèles log-normal et Gamma inverse pour σ_i^{-2} dans (2) pour l'analyse des données micro vectorielles. Leur étude par simulation a montré que les propriétés des intervalles de confiance des estimateurs sur petits domaines étaient meilleures sous modèle log-normal que sous le modèle Gamma inverse. Nous avons donc modifié leur modèle log-normal afin d'ajouter des covariables et des tailles d'échantillon n_i inégales comme il suit :

$$\left.\begin{aligned} X_i|\theta_i,\sigma_i^2 &\sim \text{Normale}(\theta_i,\sigma_i^2) \\ \theta_i &\sim \text{Normale}(\boldsymbol{Z}_i^T\boldsymbol{\beta},\tau^2);\end{aligned}\right\} \quad (12)$$

$$\left.\begin{aligned} \log S_i^2 &= \log(\sigma_i^2)+\delta_i\ ;\ \delta_i \sim N(m_i,\sigma_{ch,i}^2) \\ \log(\sigma_i^{-2}) &\sim N(\mu_v,\tau_v^2),\end{aligned}\right\} \quad (13)$$

indépendamment pour $i = 1, 2, ..., n$. Notons que le modèle de la moyenne dans (12) est identique à celui figurant dans (1). Les quantités τ^2, m_i et $\sigma_{ch,i}^2$ sont supposées connues et sont données par $m_i = E[\log(\chi_{n_i-1}^2/(n_i-1))]$ et $\sigma_{ch,i}^2 = \text{Var}[\log(\chi_{n_i-1}^2/(n_i-1))]$. Donc, la taille d'échantillon n_i détermine la forme de la distribution χ^2 par la voie du paramètre de nombre de degrés de liberté mais surtout, comme nous l'avons mentionné plus haut, les tailles d'échantillon différentes expliquent différents degrés de rétrécissement du paramètre de variance réelle correspondant. Comme dans leur approche d'estimation, les paramètres μ_v et τ_v^2 inconnus du modèle sont estimés selon une méthode fondée sur le moment dans un cadre bayésien empirique donnant $\hat{\mu}_v$ et $\hat{\tau}_v^2$, respectivement. Notons que, dans Hwang et coll. (2009), des estimations sont obtenues en se basant sur le modèle hiérarchique pour σ_i^2 dans (13) *seulement*, sans se préoccuper de la modélisation (1) de la moyenne. Nous renvoyons le lecteur à la section 5 de leur article pour des renseignements détaillés sur l'estimation des hyperparamètres. Nous suivons la même procédure en utilisant uniquement (13) pour estimer μ_v et τ_v^2 dans le cas de tailles d'échantillon inégales.

La dérivation de l'estimation bayésienne de σ_i^2 est

$$\hat{\sigma}_{i,B}^2 = \exp\left[E\{\ln(\sigma_i^2)\mid \ln(S_i^2)\}\right]$$

$$= \left\{\frac{S_i^2}{\exp(m_i)}\right\}^{M_{v,i}} \exp\{\mu_v(1-M_{v,i})\}$$

où $M_{v,i} = \tau_v^2/(\tau_v^2+\sigma_{ch,i}^2)$ et avec insertion des estimations pour remplacer les quantités inconnues. La distribution conditionnelle de θ_i sachant (X_i, S_i^2), qui est donnée par

$$\pi(\theta_i\mid X_i,S_i^2) = \int_0^\infty \pi(\theta_i\mid X_i,S_i^2,\sigma_i^2)\,\pi(\sigma_i^2\mid X_i,S_i^2)\,d\sigma_i^2,$$

est approximée par $\pi(\theta_i\mid X_i,S_i^2) \approx \int_0^\infty \pi(\theta_i\mid X_i,S_i^2,\hat{\sigma}_{i,B}^2)\,\pi(\sigma_i^2\mid X_i,S_i^2)\,d\sigma_i^2 = \pi(\theta_i\mid X_i,S_i^2,\hat{\sigma}_{i,B}^2)$. Cela suggère l'estimateur bayésien approximatif des paramètres de petit domaine donné par

$$\hat{\theta}_i = E(\theta_i\mid X_i,\hat{\sigma}_{i,B}^2) = \hat{M}_i X_i + (1-\hat{M}_i)\boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}, \quad (14)$$

où $\hat{M}_i = \hat{\tau}_v^2/(\hat{\tau}_v^2+\hat{\sigma}_{i,B}^2)$. L'intervalle de confiance pour θ_i s'obtient sous la forme

$$C_i^H = \left\{\theta_i : \frac{|\theta_i-\hat{\theta}_i|}{\hat{M}_i\hat{\sigma}_{i,B}^2} < -2\ln\{k\sqrt{2\pi}\} - \ln(\hat{M}_i)\right\}. \quad (15)$$

À la section 3 de Hwang et coll. (2009), pages 269 à 271, l'intervalle C_i^H est apparié avec l'intervalle t à $100(1-\alpha)\,\%$ $[|\theta_i - X_i| < tS_i]$ pour obtenir l'expression de k comme $k \equiv k_i = \exp\{-t^2/2\}\exp\{m_i/2\}/(\sqrt{2\pi})$.

Méthode IV : Cette méthode comprend un cas particulier du modèle de Fay-Herriot donné en (1), mais avec l'estimation des paramètres du modèle empruntée à Qiu et Hwang (2007). Ces derniers ont considéré le modèle

$$\left.\begin{aligned} X_i\mid\theta_i,\sigma^2 &\sim \text{Normale}(\theta_i,\sigma^2) \\ \theta_i &\sim \text{Normale}(0,\tau^2),\end{aligned}\right\} \quad (16)$$

indépendamment pour $i = 1, 2, ..., n$, pour analyser des données micro vectorielles expérimentales. Dans le cas où les paramètres du modèle étaient connus, ils ont proposé l'estimateur ponctuel $\hat{\theta}_i = \hat{M}X_i, \hat{M} = (1-((n-2)\sigma^2/|X|^2))_+$ où a_+ désigne $\max(0,a)$ pour tout nombre a et $|X| = |(\sum_{i=1}^n X_i^2)^{1/2}$. L'intervalle de confiance pour θ_i est $\hat{\theta}_i \pm v_1(\hat{M})$, où $v_1^2(\hat{M}) = \sigma^2\hat{M}(q_1 - \ln(\hat{M}))$ avec q_1 désignant la valeur critique de la variable normale standard pour le niveau de confiance souhaité et $v_1(0) \equiv 0$. Ici, en vue de procéder à la comparaison avec notre méthode, nous modifions le premier niveau du modèle hiérarchique dans (16) comme il suit :

$$X_i = \boldsymbol{Z}_i^T\boldsymbol{\beta} + v_i + e_i$$

où $v_i \sim \text{Normale}(0,\tau^2)$ et $e_i \sim \text{Normale}(0,S_i^2)$ indépendamment pour $i = 1, 2, ..., n$, et S_i^2 est traité comme étant connu. Comme Qiu et Hwang (2007), nous estimons τ^2 par

$$\tilde{\tau}^2 = \frac{1}{n-p}\left[\sum_i \hat{u}_i^2 - \sum_i S_i^2\left\{1-\boldsymbol{Z}_i^T\left(\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T\right)^{-1}\boldsymbol{Z}_i^T\right\}\right]$$

et $\hat{\tau}^2 = \max(\tilde{\tau}^2, 1/n)$, où $\hat{u}_i = X_i - \boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T)^{-1}(\sum_{i=1}^n \boldsymbol{Z}_i X_i)$. Puis nous définissons $\hat{M}_{0i} = \hat{\tau}^2/(\hat{\tau}^2+S_i^2)$ et $\hat{M}_i = \max(\hat{M}_{0i}, M_1)$, où , dans la dernière expression, $\hat{M}_{0i}$ est tronqué par $M_{1i} = 1 - Q_\alpha/(n_i-2)$, et Q_α est le α^{e} quantile d'une distribution du khi-carré à n_i degrés de liberté. Cet $\hat{M}_i$ est utilisé dans la formule de l'intervalle de confiance de θ_i donnée plus haut. Quand nous avons appliqué cette méthode dans notre étude par simulation et notre analyse des données réelles, nous avons modifié le modèle afin de pouvoir utiliser les tailles d'échantillon inégales et l'information sur les covariables mentionnées plus haut.

Remarque 1. Hwang et coll. (2009) ont choisi k en prenant (15) égale à l'intervalle t fondé sur X_i seulement pour les paramètres de petit domaine θ_i. Notons que X_i est l'estimateur direct d'après les données d'enquête. Par conséquent, ce choix de k n'exerce aucun contrôle direct sur la probabilité de couverture de l'intervalle construit sous *estimation par rétrécissement*. Par ailleurs, notre choix proposé de k a été établi de manière à maintenir la couverture nominale sous, précisément, l'estimation par rétrécissement.

Remarque 2. Notons qu'en l'absence de toute hypothèse de modélisation hiérarchique, S_i et X_i sont indépendants car S_i^2 et X_i sont, respectivement, auxiliaire et la statistique exhaustive complète pour θ_i. Cependant, sous les modèles (1) et (2), la distribution conditionnelle de σ_i^2 et θ_i fait intervenir à la fois X_i et S_i^2, ce que l'on peut constater en examinant (5) et (9).

Remarque 3. Dans Hwang et coll. (2009), l'estimateur à rétrécisseur de σ_i^2 est fondé uniquement sur l'information au sujet de S_i^2, et non au sujet de X_i ainsi que S_i^2. L'estimateur bayésien de σ_i^2 est introduit par insertion dans l'expression de l'estimateur bayésien des paramètres de petit domaine. Donc, l'estimateur sur petits domaines de Hwang et coll. s'écrit sous la forme $E(\theta_i \mid X_i, \hat{\sigma}_{i,B}^2)$ dans (14) où $\hat{\sigma}_{i,B}^2$ est l'estimateur bayésien de σ_i^2. En raison de l'équation (9), l'estimateur à rétrécisseur de σ_i^2 dépend de $(X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2$ en plus de S_i^2 contrairement à l'estimateur de Hwang et coll. (2009). Nous pensons que cela pourrait être l'explication de la meilleure performance de notre méthode comparativement à celle de Hwang et coll. (2009).

Remarque 4. Comme nous l'avons mentionné plus haut, le nombre de degrés de liberté associés à la distribution χ^2 pour la variance d'échantillonnage ne doit pas être simplement $n_i - 1$, n_i étant la taille de l'échantillon pour le i^{e} domaine. Il n'existe aucun résultat théorique fiable pour déterminer le nombre de degrés de liberté quand le plan de sondage est complexe. Wang et Fuller (2003) ont approximé la distribution χ^2 par une distribution normale fondée sur l'approximation de Wilson-Hilferty. Si l'on connaît le plan de sondage exact, les lignes directrices basées sur la simulation de Maples et coll. (2009) pourraient être utiles. Pour produire des estimations au niveau du comté en se servant des données de l'American Community Survey, Maples et coll. (2009) ont suggéré d'estimer le nombre de degrés de liberté par $0{,}36 \times \sqrt{n_i}$.

4. Justification théorique

À la présente section, nous donnons la justification théorique du choix de k suivant l'équation (10). Comme dans Hwang et coll. (2009), la distribution conditionnelle de θ_i sachant X_i et S_i^2 peut être approximée par $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}) \approx \pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}, \hat{\sigma}_i^2)$, où $\hat{\sigma}_i^2$ est défini comme dans (11). De la même façon, nous approximons $E(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{B})$ par $E(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{B}) \approx \hat{\sigma}_i^{-1}$. Sur la base de ces approximations, nous avons $C_i(\boldsymbol{B}) \approx \tilde{C}_i(\boldsymbol{B})$ où $\tilde{C}_i(\boldsymbol{B})$ est l'intervalle de confiance de θ_i donné par $\tilde{C}_i(\boldsymbol{B}) = \{\theta_i : \pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}, \hat{\sigma}_i^2) \geq k\hat{\sigma}_i^{-1}\}$. De (1) il découle que la densité de probabilité conditionnelle $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}, \sigma_i^2)$ est normale de moyenne μ_i et de variance v_i, où μ_i et v_i sont donnés par les expressions

$$\mu_i = w_i X_i + (1 - w_i)\boldsymbol{Z}_i^T\boldsymbol{\beta},$$
$$v_i = \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)^{-1} = \sigma_i^2\left(1 + \frac{\sigma_i^2}{\tau^2}\right)^{-1}, \qquad (17)$$

et

$$w_i = \frac{1/\sigma_i^2}{(1/\sigma_i^2 + 1/\tau^2)}.$$

Maintenant, en choisissant

$$k = \hat{u}_0\phi\left(t_{\alpha/2}\sqrt{\frac{n_i + 2a + 2}{n_i - 1}}\right)$$

comme nous l'avons mentionné, l'intervalle de confiance $\tilde{C}_i(\boldsymbol{B})$ devient

$$\tilde{C}_i(\boldsymbol{B}) = \left\{\theta_i : \hat{u}_{0i}\frac{|\theta_i - \hat{\mu}_i|}{\hat{\sigma}_i} \leq t_{\alpha/2}\sqrt{\frac{n_i + 2a + 2}{n_i - 1}}\right\}, \quad (18)$$

où $\hat{\mu}_i$ est l'expression de μ_i dans (17) avec remplacement de σ_i^2 par $\hat{\sigma}_i^2$. Considérons maintenant le comportement de $\hat{\sigma}_i^2 \equiv \hat{\sigma}_i^2(\boldsymbol{B})$ quand τ^2 varie entre 0 et ∞. Quand $\tau^2 \to \infty$, $\hat{\sigma}_i^2$ converge vers

$$\hat{\sigma}_i^2(\infty) \equiv \hat{\sigma}_i^2(a, b, \boldsymbol{\beta}, \infty) = \frac{\frac{(n_i - 1)}{2}S_i^2 + \frac{1}{b}}{\frac{n_i - 1}{2} + a + 1} = \frac{(n_i - 1)S_i^2 + \frac{2}{b}}{n_i + 2a + 1}.$$

De même, quand $\tau^2 \to 0$, $\hat{\sigma}_i^2$ converge vers

$$\hat{\sigma}_i^2(0) \equiv \hat{\sigma}_i^2(a, b, \boldsymbol{\beta}, 0) = \frac{(X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2 + (n_i - 1)S_i^2 + \frac{2}{b}}{n_i + 2a + 2}.$$

Pour toute valeur intermédiaire de τ^2, nous avons $\min\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\} \leq \hat{\sigma}_i^2 \leq \max\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\}$. Donc, il

est suffisant de considérer les deux cas suivants : i) $\hat{\sigma}_i^2 \geq \hat{\sigma}_i^2(\infty)$, où il s'ensuit que $(n_i + 2a + 2)\hat{\sigma}_i^2 = (n_i + 2a + 1)\hat{\sigma}_i^2 + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2 + 2/b + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2$, et ii) $\hat{\sigma}_i^2 \geq \hat{\sigma}^2(0)$, où il s'ensuit que $(n_i + 2a + 2)\,\hat{\sigma}_i^2 = (X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2 + (n_i - 1)S_i^2 + 2/b \geq (n_i - 1)S_i^2$. Donc, dans les cas (i) ainsi que (ii),

$$(n_i + 2a + 2)\,\hat{\sigma}_i^2 \geq (n_i - 1)S_i^2. \tag{19}$$

Puisque $\theta_i - \mu_i \sim N(0, \sigma_i^2\tau^2/(\sigma_i^2 + \tau^2))$ et $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{n_i - 1}$, l'intervalle de confiance

$$D_i = \left\{\theta_i : u_{0i}\frac{|\theta_i - \mu_i|}{S_i} \leq t_{\alpha/2}\right\} \tag{20}$$

a une probabilité de couverture de $1 - \alpha$. Donc, si u_0 et μ_i sont remplacés par $\hat{u}_0$ et $\hat{\mu}_i$, il faut s'attendre à ce que l'intervalle de confiance résultant $\tilde{D}_i$, disons, ait une probabilité de couverture d'environ $1 - \alpha$. De (19), nous obtenons

$$P\{\tilde{C}_i(\boldsymbol{B})\} \geq P(\tilde{D}_i) \approx 1 - \alpha, \tag{21}$$

ce qui établit une borne inférieure approximative de $1 - \alpha$ pour le seuil de confiance de $\tilde{C}_i(\boldsymbol{B})$.

Dans (21), $\boldsymbol{B}$ était supposé fixe et connu. Quand $\boldsymbol{B}$ est inconnu, nous le remplaçons par l'estimation de son maximum de vraisemblance marginale $\hat{\boldsymbol{B}}$. Puisque l'expression (21) est vérifiée quelle que soit la valeur réelle de $\boldsymbol{B}$, la substitution de $\hat{\boldsymbol{B}}$ à $\boldsymbol{B}$ dans (21) comportera une erreur d'ordre $O(1/\sqrt{N})$, où $N = \sum_{i=1}^{n} n_i$. Comparativement à chaque n_i pris individuellement, ce groupement des n_i devrait réduire l'erreur de manière significative, de manière que $\tilde{C}_i(\hat{\boldsymbol{B}})$ soit suffisamment proche de $\tilde{C}_i(\boldsymbol{B})$ pour satisfaire la borne inférieure de $1 - \alpha$ dans (21).

5. Une étude par simulation

5.1 Conditions de simulation

Nous considérons les conditions de simulation dans lesquelles nous utilisons un sous-ensemble de configurations des paramètres emprunté à Wang et Fuller (2003). Chaque échantillon employé dans l'étude par simulation a été obtenu en suivant les étapes que voici. Premièrement, générer des observations en utilisant le modèle

$$X_{ij} = \beta + u_i + e_{ij},$$

où $u_i \sim N(0, \tau^2)$ et $e_{ij} \sim N(0, n_i\sigma_i^2)$, indépendamment pour $j = 1, ..., n_i$ et $i = 1, ..., n$. Alors, le modèle à effets aléatoires pour la moyenne de petit domaine, X_i, est

$$X_i = \beta + u_i + e_i, \quad \text{indépendamment pour } i = 1, ..., n,$$

où $X_i \equiv \bar{X}_{i\cdot} \equiv n_i^{-1}\sum_{j=1}^{n_i} X_{ij}$ et $e_i \equiv \bar{e}_{i\cdot} \equiv n_i^{-1}\sum_{j=1}^{n_i} e_{ij}$. Donc, $X_i \sim N(\theta_i, \sigma_i^2)$, où $\theta_i = \beta + u_i$, $\theta_i \sim N(\beta, \tau^2)$ et $e_i \sim N(0, \sigma_i^2)$. Nous avons estimé σ_i^2 en nous servant de l'estimateur sans biais

$$S_i^2 = (n_i - 1)^{-1} n_i^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2,$$

et il s'ensuit que $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{n_i - 1}$, indépendamment pour $i = 1, 2, ..., n$. Notons que le plan de simulation ne tenait pas compte de la modélisation des variances d'échantillonnage au deuxième niveau dans (2). Par conséquent, notre résultat indiquera une robustesse à l'erreur de spécification du modèle de variance.

Les étapes susmentionnées ont produit les données $(X_i, S_i^2), i = 1, ..., n$. Pour simplifier la simulation, nous ne choisissons aucune covariable $\boldsymbol{Z}_i$. À l'instar de Wang et Fuller (2003), nous donnons la valeur m à tous les n_i afin de faciliter la programmation. Cependant, nous choisissons quand même que les variances d'échantillonnage réelles soient inégales : la valeur d'un tiers des σ_i^2 est fixée à 1, celle d'un deuxième tiers est fixée à 4 et celle du dernier tiers est fixée à 16. Nous prenons $\beta = 10$ et trois valeurs différentes de $\tau^2 = 0{,}25$, 1 et 4. Nous avons choisi ces valeurs des paramètres en nous inspirant de Qiu et Hwang (2007). Pour chaque valeur de τ^2, nous avons généré 200 échantillons pour les deux combinaisons $(m, n) = (9, 36)$ et (18, 180).

Dans l'étude par simulation, nous comparons la méthode que nous proposons aux méthodes de Wang et Fuller (2003), Hwang et coll. (2009), et Qiu et Hwang (2007) que nous appelons méthodes I, II, III et IV, respectivement, en nous fondant sur le biais, l'erreur quadratique moyenne (EQM), la probabilité de couverture (PC) des intervalles de confiance et la longueur moyenne des intervalles de confiance (LMIC). Le tableau 1 donne les estimations des paramètres pour a, b, β et τ^2. Les résultats numériques indiquent que les estimations du maximum de vraisemblance des paramètres du modèle ont de bonnes propriétés ; les valeurs estimées de β et τ^2 sont proches des valeurs réelles, ce qui témoigne de bonnes propriétés de robustesse à l'erreur de spécification de la distribution au deuxième niveau de (2). L'obtention d'estimations statistiquement significatives pour a ainsi que b indique que les variances d'échantillonnage « rétrécies » sont intégrées dans la méthode proposée. Les tableaux 2, 3 et 4 donnent les moyennes des résultats numériques calculées sur les domaines qui, dans chaque groupe, ont les mêmes variances d'échantillonnage réelles. Les résultats des tableaux sont fondés sur 200 répliques.

Tableau 1
Résultats des simulations pour les paramètres du modèle, a (panneau supérieur gauche), b (panneau supérieur droit), β (panneau inférieur gauche) et τ^2 (panneau inférieur droit). Ici, É.-T. représente l'écart-type sur 200 répliques. Nous avons pris $\beta = 10$ et $\tau^2 = 0,25,\ 1$ et 4

	$n = 36,\ m = 9$		$n = 180,\ m = 18$			$n = 36,\ m = 9$		$n = 180,\ m = 18$	
τ^2	Moyenne	É.-T.	Moyenne	É.-T.	τ^2	Moyenne	É.-T.	Moyenne	É.-T.
		a					b		
0,25	1,0959	0,1540	1,0328	0,0442	0,25	0,3992	0,0983	0,4249	0,0323
1	1,0937	0,1555	1,0325	0,0445	1	0,4030	0,1012	0,4253	0,0326
4	1,0996	0,1577	1,0339	0,0450	4	0,3999	0,1017	0,4245	0,0328
		β					τ^2		
0,25	10,0071	0,3618	9,9951	0,1853	0,25	0,2558	0,0605	0,2575	0,0097
1	10,0142	0,3311	9,9970	0,1743	1	0,9418	0,3333	1,0426	0,1264
4	10,0282	0,4639	10,0048	0,2254	4	3,5592	1,3316	4,0817	0,5551

Tableau 2
Résultats des simulations pour la prédiction quand $\tau^2 = 0,25$. Ici, EQM, LMIC et PC représentent l'erreur quadratique moyenne, la longueur moyenne de l'intervalle de confiance et la probabilité de couverture de l'intervalle de confiance, respectivement

		$n = 36,\ m = 9$				$n = 180,\ m = 18$			
		Méthode				Méthode			
	σ_i^2	I	II	III	IV	I	II	III	IV
Biais	1	0,0048	0,0198	0,0272	0,0018	-0,0051	-0,0086	-0,0112	-0,0111
relatif	4	-0,0033	-0,0061	-0,0145	-0,0158	-0,0130	-0,0109	-0,0065	-0,0116
	16	0,0126	0,0370	0,0369	0,0096	-0,0046	-0,0045	-0,0080	-0,0061
EQM	1	0,3066	0,3890	0,6861	0,3805	0,2258	0,2680	0,4470	0,2922
	4	0,3281	0,5430	1,3778	0,7285	0,2595	0,3000	0,5805	0,3748
	16	0,3715	0,5240	1,6749	1,9316	0,2815	0,2850	0,4856	0,6383
LMIC	1	2,1393	2,5485	4,4906	3,0528	1,9220	1,6006	3,6466	2,4811
	4	2,2632	3,9574	6,8887	5,6842	2,0557	2,1524	5,2472	4,2160
	16	2,3221	4,5619	9,3335	11,1363	2,1046	2,3308	6,5273	7,8492
PC	1	0,9468	0,9770	0,9771	0,9708	0,9564	0,9710	0,9851	0,9631
	4	0,9468	0,9710	0,9829	0,9917	0,9555	0,9660	0,9967	0,9967
	16	0,9365	0,9660	0,9933	0,9975	0,9529	0,9610	0,9998	0,9999

Tableau 3
Résultats des simulations pour la prédiction quand $\tau^2 = 1$. Ici, EQM, LMIC et PC représentent l'erreur quadratique moyenne, la longueur moyenne de l'intervalle de confiance et la probabilité de couverture de l'intervalle de confiance, respectivement

		$n = 36,\ m = 9$				$n = 180,\ m = 18$			
		Méthode				Méthode			
	σ_i^2	I	II	III	IV	I	II	III	IV
Biais	1	-0,0152	0,0205	0,0255	0,0051	-0,0064	-0,0085	-0,0111	-0,0101
relatif	4	-0,0167	-0,0164	-0,0151	-0,0219	-0,0151	-0,0121	-0,0133	-0,0164
	16	-0,0323	0,0508	0,0515	0,0216	-0,0028	-0,0017	-0,0073	-0,0039
EQM	1	0,5645	0,6330	0,7238	0,6260	0,5288	0,5430	0,5673	0,6336
	4	0,8566	1,1100	1,5396	1,0992	0,8159	0,8770	0,9415	0,8948
	16	1,0482	1,3100	2,1059	2,3156	0,9786	1,0000	1,1024	1,1878
LMIC	1	3,4550	3,1822	4,4938	3,2117	3,1088	2,5094	3,6763	2,8676
	4	4,0321	5,8733	6,8984	5,7909	3,7844	4,2908	5,3323	4,5543
	16	4,4082	7,4286	9,3555	11,1555	4,1187	5,1590	6,6785	7,8937
PC	1	0,9704	0,9640	0,9762	0,9275	0,9660	0,9650	0,9786	0,8879
	4	0,9633	0,9560	0,9812	0,9808	0,9627	0,9680	0,9918	0,9740
	16	0,9533	0,9490	0,9912	0,9938	0,9613	0,9680	0,9974	0,9979

Tableau 4
Résultats des simulations pour la prédiction quand $\tau^2 = 4$. Ici, EQM, LMIC et PC représentent l'erreur quadratique moyenne, la longueur moyenne de l'intervalle de confiance et la probabilité de couverture de l'intervalle de confiance, respectivement

		$n = 36,\ m = 9$				$n = 180,\ m = 18$			
		Méthode				Méthode			
	σ_i^2	I	II	III	IV	I	II	III	IV
Biais	1	-0,0024	0,0248	0,0229	0,0180	-0,0084	-0,0098	-0,0122	-0,0106
relatif	4	-0,0343	-0,0310	-0,0210	-0,0340	-0,0110	-0,0092	-0,0174	-0,0132
	16	-0,0147	0,0702	0,0767	0,0467	0,0016	0,0024	-0,0059	0,0012
EQM	1	0,8822	0,8590	0,8579	1,0559	0,8359	0,8180	0,8541	0,8605
	4	2,0577	2,2900	2,1818	2,2422	2,0424	2,1000	2,0935	2,1130
	16	3,4516	3,7600	3,9267	3,8981	3,3153	3,3500	3,3939	3,3631
LMIC	1	4,6318	4,1936	4,5369	3,7677	4,0256	3,5346	3,9626	3,7499
	4	6,2015	10,9093	7,0376	6,4314	5,9000	9,0913	6,2217	6,1540
	16	7,7221	18,0039	9,6718	11,3341	7,4430	14,6665	8,3908	8,7537
PC	1	0,9791	0,9670	0,9733	0,9029	0,9674	0,9570	0,9600	0,9468
	4	0,9556	0,9670	0,9725	0,9496	0,9592	0,9610	0,9633	0,9573
	16	0,9510	0,9670	0,9796	0,9858	0,9573	0,9650	0,9718	0,9776

Comparaisons des biais : Dans la plupart des cas, les biais des quatre méthodes sont comparables. Il n'existe aucune preuve manifeste d'écarts significatifs entre elles pour ce qui est du biais. Une forte variance d'échantillonnage donne plus de poids à la moyenne de population par construction, ce qui rend l'estimateur plus proche de la moyenne au deuxième niveau. Par ailleurs, les méthodes I à III comprennent l'utilisation d'estimateurs à rétrécisseur des variances d'échantillonnage qui seraient donc inférieurs au maximum de l'ensemble des variances d'échantillonnage. Donc, les méthodes I à III ont tendance à présenter un biais un peu plus important. Cependant, en raison du rétrécissement des variances d'échantillonnage, on peut s'attendre à une amélioration de la variance des estimateurs qui, à son tour, réduit l'EQM. Parmi les méthodes I à III, la méthode I a donné de meilleurs résultats que les méthodes II et III, dont les propriétés étaient assez semblables. Le gain maximal en utilisant la méthode I au lieu de la méthode II est de 99 %.

Comparaison des EQM : En ce qui concerne l'EQM, la méthode I a donné systématiquement de meilleurs résultats que les trois autres dans tous les cas, sauf quand le ratio de σ_i^2 à τ^2 était le plus faible : $(\sigma_i^2 = 1) / (\tau^2 = 4) = 0{,}25$. Dans ce cas, la variance entre les petits domaines (variance du modèle) est beaucoup plus grande que la variance dans les domaines (variance d'échantillonnage). Lorsque notre méthode est utilisée pour estimer θ_i, l'information « empruntée » à d'autres domaines peut mal orienter l'estimation : la moyenne estimée de la loi Gamma pour σ_i^{-2} provenant du deuxième niveau de (2) est $\hat{a}\hat{b}$, qui est égale à 0,44 environ pour les deux combinaisons (m, n) correspondant à $(9, 36)$ et $(18, 180)$ (la valeur réelle est $ab = 0{,}4$). Donc, $E(\sigma_i^{-2} | X_i, S_i^2, \hat{B})$ est significativement plus petite que 1 en raison du rapprochement vers la moyenne pour le groupe pour lequel la valeur réelle est $\sigma_i^2 = 1$. En outre, puisque σ_i^2 est plus faible que τ^2, le poids de X_i devrait être beaucoup plus élevé comparativement à β, la moyenne globale. Cependant, étant donné la sous-estimation de σ_i^{-2} dans ce cas, l'estimateur résultant donne moins de poids à X_i, ce qui donne lieu à une EQM plus grande. Cependant, cette sous-estimation diminue pour les grandes tailles d'échantillon en raison de la cohérence des estimateurs de Bayes. Ce fait s'observe effectivement quand la taille d'échantillon passe de $n = 36$ à $n = 180$ pour $\sigma_i^2 = 1$ et $\tau^2 = 4$. Comparativement à la méthode II, la méthode I produit une amélioration dans la plupart des cas simulés ; le gain maximal est de 30 %, tandis que la seule perte observée est de 9 % pour la combinaison $\sigma_i^2 = 1$ et $\tau^2 = 4$ pour $n = 36$ et $m = 9$. De même, par rapport à la méthode III, le gain maximal donné par la méthode I est de 77 % et la seule perte est de 11 %, pour les mêmes spécifications de paramètres et de tailles d'échantillon.

Comparaisons des PC : Nous avons obtenu les intervalles de confiance au seuil de confiance de 95 %. Les méthodes I et III ne révèlent aucune sous-couverture, ce qui n'est pas étonnant étant donné la construction optimale de leurs intervalles de confiance. La méthode I produit le taux nominal de couverture plus fréquemment que n'importe quelle autre méthode. La méthode II présente une certaine sous-couverture, le taux pouvant être aussi faible que 82 %.

Comparaisons des LMIC : La méthode I produit en général des intervalles de confiance considérablement plus courts

que les autres méthodes. La méthode IV a produit des intervalles de longueur comparable à ceux des autres méthodes dans tous les cas sauf quand σ_i^2 était élevé, auquel cas les longueurs étaient considérablement plus grandes. L'intervalle de confiance proposée dans Qiu et Hwang (2007) n'a pas de bonnes propriétés en échantillon fini, particulièrement pour les petites valeurs de τ^2. Afin d'éviter un faible taux de couverture, ils ont proposé de tronquer $M_0 = \tau^2/(\tau^2 + \sigma_i^2)$ à l'aide d'un nombre positif $M_1 = 1 - Q_\alpha/(\nu - 2)$ pour σ_i^2 connu, où Q_α est le α^e quantile d'une distribution du khi-carré à ν degrés de liberté. Quand le ratio de la variance d'échantillonnage à la variance du modèle, σ_i^2/τ^2, est élevé, M_1 a tendance à être plus grand que M_0, ce qui donne le taux nominal de couverture, mais avec de plus grandes longueurs d'intervalle. Par exemple, dans le cas où $(\sigma_i^2, \tau^2) = (16, 0{,}25)$, la LMIC est de 11,13 pour la méthode IV, alors qu'elle est seulement de 2,78 et 4,56 pour les méthodes I et II, respectivement.

5.2 Étude de la robustesse

Afin d'étudier la robustesse de la méthode proposée aux écarts par rapport à l'hypothèse de normalité des erreurs, nous avons procédé à l'étude par simulation qui suit. Les données ont été générées comme précédemment, mais en tirant les e_{ij} d'une loi exponentielle double (loi de Laplace) et d'une loi uniforme. Les estimateurs des méthodes II et III ont eu peu d'effet. Cela pourrait tenir au fait que, dans ces méthodes, l'estimation des paramètres du modèle se fait par la méthode des moments. La méthode IV a produit de plus grandes valeurs du biais relatif, de l'EQM et de la LMIC, et une plus faible probabilité de couverture. L'EQM est systématiquement plus faible pour la méthode I que pour la méthode II. Quand $\tau^2 = 0{,}25$ et 1, la LMIC est plus petite pour la méthode I que pour la méthode II pour (n = 36, m = 9), mais le résultat inverse s'observe quand (n = 180, m = 18). Pour ce qui est de la PC, la méthode II produit une certaine sous-couverture (taux le plus faible égal à 80 %). Par contre, la méthode I ne produit aucune sous-couverture. Faute d'espace, nous présentons uniquement les résultats pour les paramètres a, b, β et τ^2 sous les erreurs laplaciennes (tableau 5).

6. Analyse de données réelles

Pour illustrer notre méthodologie, nous choisissons un exemple très souvent étudié. Le jeu de données, qui provient du U.S. Department of Agriculture, a été analysé pour la première fois par Battese (1988). Il s'agit de données sur les productions de maïs et de soja dans 12 comtés de l'Iowa. Les tailles d'échantillon pour ces domaines sont faibles, variant de 1 à 5. Faute d'espace, nous considérons uniquement le cas du maïs. Pour les modèles proposés, il faut nécessairement que l'on ait des tailles d'échantillon $n_i > 1$. Par conséquent, nous avons utilisé des données modifiées tirées de You et Chapman (2006) avec $n_i \geq 2$. Les nombres déclarés d'hectares consacrés à la culture du maïs (X_i), qui sont les estimations directes par sondage, sont présentés au tableau 6. Ce tableau donne aussi les variances d'échantillonnage qui sont calculées d'après les données originales sous l'hypothèse d'un échantillonnage aléatoire simple. L'écart-type d'échantillon varie fortement, de 5,704 à 53,999 (le coefficient de variation varie de 0,036 à 0,423). Deux covariables sont considérées dans le tableau 6 : Z_{i1}, le nombre moyen de pixels correspondant à du maïs et Z_{i2}, le nombre moyen de pixels correspondant à du soja, provenant des données de satellite LANDSAT.

Tableau 5
Résultats des simulations pour les paramètres du modèle, *a* (panneau supérieur gauche), *b* (panneau supérieur droit), β (panneau inférieur gauche) et τ^2 (panneau inférieur droit) quand les erreurs suivent une loi de Laplace. Ici, É.-T. représente l'écart-type sur 200 répliques. Nous avons pris $\beta = 10$ et $\tau^2 = 0{,}25$, 1 et 4

	$n = 36$, $m = 9$		$n = 180$, $m = 18$			$n = 36$, $m = 9$		$n = 180$, $m = 18$	
τ^2	Moyenne	É.-T.	Moyenne	É.-T.	τ^2	Moyenne	É.-T.	Moyenne	É.-T.
		a					*b*		
0,25	0,9624	0,1632	0,9471	0,0498	0,25	0,5793	0,1733	0,5279	0,0501
1	0,9628	0,1657	0,9476	0,0497	1	0,5816	0,1777	0,5275	0,0503
4	0,9689	0,1694	0,9487	0,0499	4	0,5758	0,1796	0,5263	0,0503
		β					τ^2		
0,25	9,9736	0,3775	9,9800	0,1773	0,25	0,2696	0,0882	0,2565	0,0074
1	9,9753	0,3709	9,9836	0,1662	1	1,0508	0,2501	1,0403	0,0668
4	9,9736	0,4835	9,9855	0,2161	4	3,9624	1,1719	4,1256	0,4201

Tableau 6
Données sur le maïs provenant de You et Chapman (2006)

Comté	n_i	X_i	Z_{1i}	Z_{2i}	$\sqrt{S_i^2}$
Franklin	3	158,623	318,21	188,06	5,704
Pocahontas	3	102,523	257,17	247,13	43,406
Winnebago	3	112,773	291,77	185,37	30,547
Wright	3	144,297	301,26	221,36	53,999
Webster	4	117,595	262,17	247,09	21,298
Hancock	5	109,382	314,28	198,66	15,661
Kossuth	5	110,252	298,65	204,61	12,112
Hardin	5	120,054	325,99	177,05	36,807

Les estimations de $\boldsymbol{B}$ sont les suivantes : $a = 1{,}707$, $b = 0{,}00135$, $\tau^2 = 90{,}58$ et $\boldsymbol{\beta} = (-186{,}0\ ;\ 0{,}7505\ ;\ 0{,}4100)$. La moyenne a priori estimée de $1/\sigma_i^2$ qui est la moyenne de la loi Gamma dont les paramètres sont a et b, est $ab = 0{,}002295$ dont la racine carrée est 0,048 (notons que $1/0{,}048 = 20{,}85$, valeur en harmonie avec l'intervalle de variation des écarts-types d'échantillon s'étendant de 5,704 à 53,999). Les estimations sur petits domaines et leurs intervalles de confiance sont résumés au tableau 7 et à la figure 1. Les estimations ponctuelles produites par les quatre méthodes sont comparables : les mesures sommaires comprenant la moyenne, la médiane et l'étendue des estimations des paramètres de petit domaine pour les méthodes I, II, III et IV sont (121,9 ; 124,1 ; 122,2 ; 122,6), (125,2 ; 120,4 ; 115,0 ; 114,5) et (23,1 ; 53,0 ; 58,4 ; 56,6), respectivement. Les distributions de $\hat{\theta}_i$ (représentées graphiquement en prenant en considération tous les i) sont résumées à la figure 2 qui révèle une différence significative de variabilité. La méthode I est celle dont la variabilité est la plus faible et qui est donc la meilleure en ce sens. En outre, le lissage des variances d'échantillonnage a de fortes répercussions sur la mesure de l'incertitude et donc de l'estimation de l'intervalle. La méthode proposée donne l'intervalle de confiance le plus court, en moyenne, comparativement à toutes les autres méthodes. Les méthodes II et III donnent des intervalles dont la borne inférieure est négative, ce qui paraît irréaliste, car la moyenne directe des superficies consacrées à la culture du maïs est positive et grande pour les 12 comtés [les intervalles de confiance bruts $(x_i \pm t_{0{,}025}\, S_i)$ ne contiennent non plus de valeur nulle pour aucun des domaines]. Il n'existe aucun soutien théorique pour les intervalles de confiance de la méthode II. Les méthodes II et III produisent des intervalles de confiance plus larges quand la variance d'échantillonnage est élevée. Par exemple, la taille d'échantillon pour les comtés de Franklin et de Pocahontas est de trois, mais les écarts-types d'échantillon sont de 5,704 et 43,406, respectivement. Alors que les intervalles de confiance sont comparables sous la méthode I, ils sont très différents sous les méthodes II et III. Il en est ainsi parce que, même si ces méthodes tiennent compte de l'incertitude dans les estimations de la variance d'échantillonnage, comme le lissage n'a pas été effectué en utilisant l'information provenant des estimations directes d'après l'enquête, les estimations de la variance d'échantillonnage sous-jacentes demeurent très variables (à cause de la petite taille d'échantillon). En fait, la variance de l'estimateur de variance (des estimations ponctuelles) est plus grande que celle obtenue lorsque l'on applique la méthode I. Cela est aussi confirmé par le fait que les écarts-types intuitifs des estimations sur petits domaines « lissées » (un quart de l'intervalle) sont plus faibles et moins variables sous la méthode I que sous les autres méthodes. Une autre caractéristique de notre méthode qui mérite d'être soulignée est que les largeurs des intervalles sont comparables pour les comtés pour lesquels la taille d'échantillon est la même. Cela pourrait être une indication que l'on obtient des estimateurs équi-efficaces pour des tailles d'échantillon équivalentes.

Choix du modèle : Afin de choisir le modèle le mieux ajusté, nous avons utilisé le critère d'information bayésien (BIC pour *Bayesian Information Criteria*) qui tient compte à la fois de la vraisemblance et de la complexité des modèles ajustés. Nous avons calculé le BIC pour les modèles utilisés dans les méthodes I et III (Hwang et coll. 2009). Ces deux modèles comprennent le même nombre de paramètres et ne diffèrent que par la façon dont ces paramètres sont estimés. Le BIC du modèle est égal à 210,025 pour la méthode I et à 227,372 pour la méthode III, ce qui témoigne de la supériorité de notre méthode. Nous n'avons pas pu calculer le BIC pour le modèle de Wang et Fuller (2003), car ils n'ont utilisé aucune fonction de vraisemblance explicite.

Tableau 7
Résultats de l'analyse des données sur le maïs. Ici, IC et LIC représentent l'intervalle de confiance et la longueur de l'intervalle de confiance, respectivement

Comté	$\hat{\theta}_i$	IC	LIC	$\hat{\theta}_i$	IC	LIC
		I : méthode proposée			II : Wang et Fuller (2003)	
Franklin	131,8106	104,085 ; 159,372	55,287	155,4338	124,151 ; 193,094	68,943
Pocahontas	108,7305	80,900 ; 136,436	55,536	102,3682	-38,973 ; 244,019	282,993
Winnebago	109,0559	81,430 ; 136,646	55,216	115,9093	-53,768 ; 279,314	333,083
Wright	131,6113	103,736 ; 159,564	55,828	131,0674	8,330 ; 280,263	271,932
Webster	113,1484	92,805 ; 133,348	40,543	109,4795	32,514 ; 202,675	170,161
Hancock	129,4279	111,781 ; 147,193	35,412	124,1028	56,750 ; 162,013	105,262
Kossuth	121,0071	103,451 ; 138,626	35,175	116,7147	68,049 ; 152,454	84,405
Hardin	130,2520	112,373 ; 148,114	35,741	137,7983	51,734 ; 188,373	136,638
		III : Hwang et coll. (2009)			IV : Qiu et Hwang (2007)	
Franklin	158,4677	128,564 ; 188,370	59,805	157,7383	146,999 ; 168,477	21,478
Pocahontas	100,1276	-44,039 ; 244,295	288,334	101,1661	19,444 ; 182,887	163,442
Winnebago	114,1473	0,065 ; 228,228	228,163	113,7746	56,263 ; 171,286	115,022
Wright	140,3717	-24,119 ; 304,862	328,982	143,2244	41,559 ; 244,889	203,330
Webster	115,7865	50,297 ; 181,275	130,978	115,2224	75,124 ; 155,320	80,196
Hancock	111,3087	66,213 ; 156,403	90,189	113,1766	83,691 ; 142,661	58,970
Kossuth	110,9585	74,366 ; 147,550	73,184	112,3239	89,520 ; 135,127	45,607
Hardin	126,6093	40,040 ; 213,178	173,137	123,9049	54,607 ; 193,202	138,594

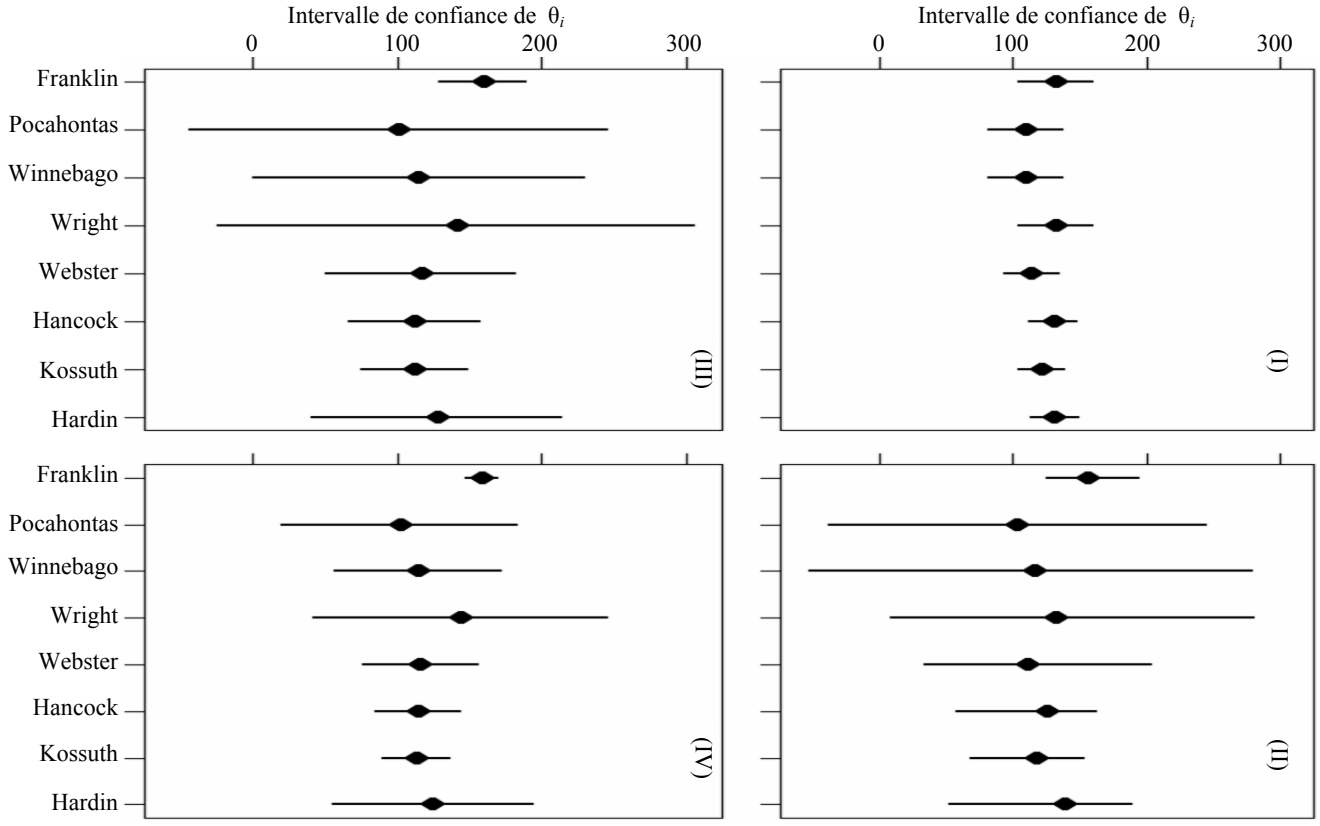


Figure 1 Estimation du nombre d'hectares consacrés au maïs. Pour chaque comté, la droite horizontale donne l'intervalle de confiance de $\hat{\theta}_i$, avec $\hat{\theta}_i$ marqué par le cercle, pour (I) la méthode proposée, (II) Wang et Fuller (2003), (III) Hwang et coll. (2009) et (IV) Qiu et Hwang (2007)

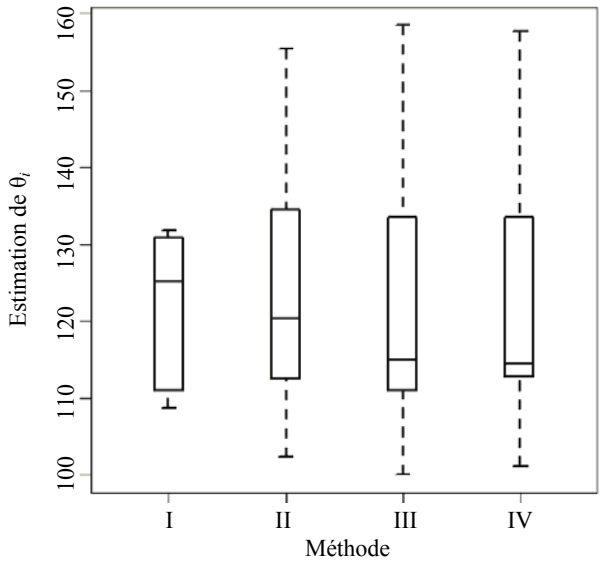


Figure 2 Boîtes à moustaches des estimations du nombre d'hectares consacrés au maïs p our chaque comté. (I) à (IV) sont les quatre méthodes correspondant à la figure I

7. Conclusion

Le présent article décrit la modélisation conjointe au niveau du domaine des moyennes et des variances pour l'estimation sur petits domaines. Il montre que les estimateurs sur petits domaines résultants sont plus efficaces que les estimateurs classiques obtenus en utilisant les modèles de Fay-Herriot qui ne rétrécissent que les moyennes. Bien que notre modèle soit le même que celui pris en considération dans Hwang et coll. (2009), notre méthode d'estimation diffère à deux égards, en ce qui concerne la détermination du paramètre de mise au point k et l'utilisation de $\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i)$ (qui dépend additionnellement de X_i), au lieu de $\pi(\sigma_i^2 \mid S_i^2, \boldsymbol{Z}_i)$ pour construire la distribution conditionnelle des paramètres θ_i de petit domaine. Nous avons démontré les propriétés de robustesse du modèle quand l'hypothèse que σ_i^2 est issue d'une loi Gamma inverse est violée. L'emprunt de l'information X_i pour estimer σ_i^2 ainsi que la robustesse à l'élicitation de la loi a priori démontre la supériorité de la méthode que nous proposons. Les valeurs des paramètres choisis dans l'étude par simulation diffèrent de celles utilisées dans l'analyse des données réelles. Cette dernière est présentée ici simplement en guise d'illustration. Notre objectif principal était d'élaborer la méthodologie de modélisation de la moyenne et de la variance, et de la comparer à certaines méthodes étroitement apparentée afin de montrer son efficacité. C'est pourquoi nous avons choisi de configurer les paramètres dans la simulation de la même façon que dans l'article traitant de l'estimation sur petits domaine bien connu de Wang et Fuller (2003).

L'obtention d'estimateurs améliorés de la variance d'échantillonnage est un produit secondaire de l'approche proposée. Nous avons fourni une technique d'estimation novatrice, qui est justifiée théoriquement et facile à utiliser. En ce qui concerne les calculs, la méthode est beaucoup plus simple que certaines méthodes concurrentes telles que les procédures MCMC bayésiennes ou les méthodes de ré-échantillonnage bootstrap. Notre méthode ne requiert qu'un seul échantillonnage à partir de la loi a posteriori durant l'estimation des paramètres du modèle, et les valeurs échantillonnées peuvent être utilisées par la suite à toute autre fin. Le logiciel peut être obtenu sur demande auprès des auteurs.

Remerciements

Les auteurs remercient deux examinateurs et le rédacteur associé de leurs commentaires constructifs qui leur ont permis d'améliorer considérablement l'article. L'étude a été financée en partie par les subventions SES 0961649, 0961618 et DMS 1106450 de la NSF.

Annexe

A. Obtention des distributions conditionnelles

Des équations (1) et (2) il découle que la distribution conjointe conditionnelle de $\{X_i, S_i^2, \theta_i, \sigma_i^2\}$, $\pi(X_i, S_i^2, \theta_i, \sigma_i^2 \mid a, b, \boldsymbol{\beta}, \tau^2)$ est

$$
\begin{aligned}
&\pi(X_i, S_i^2, \theta_i, \sigma_i^2 \mid \boldsymbol{Z}_i, \boldsymbol{B}) \\
&\quad = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2}\right\} \frac{1}{\Gamma\left(\frac{n_i-1}{2}\right) 2^{\frac{n_i-1}{2}}} \\
&\quad \times \left\{(n_i-1)\frac{S_i^2}{\sigma_i^2}\right\}^{\frac{n_i-1}{2}-1} \exp\left\{-\frac{(n_i-1)S_i^2}{2\sigma_i^2}\right\} \\
&\quad \times \left(\frac{n_i-1}{\sigma_i^2}\right) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\} \\
&\quad \times \frac{1}{\Gamma(a)b^a}\left(\frac{1}{\sigma_i^2}\right)^{a+1} \exp\left(-\frac{1}{b\sigma_i^2}\right) \\
&\quad \propto \exp\left[-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2} - \left\{\frac{(X_i-\theta_i)^2}{2} \right.\right. \\
&\qquad \left.\left. + \frac{(n_i-1)S_i^2}{2} + \frac{1}{b}\right\}\frac{1}{\sigma_i^2}\right] \\
&\quad \times \left(\frac{1}{\sigma_i^2}\right)^{\frac{n_i}{2}+a+1} \left(\frac{1}{\tau^2}\right)^{\frac{1}{2}} \frac{1}{\Gamma(a)b^a}.
\end{aligned}
$$

Par conséquent, les distributions conditionnelles de σ_i^2 et θ_i sachant les données et $\boldsymbol{B}$ sont

$$\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) = \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 \mid \boldsymbol{Z}_i, \boldsymbol{B})\, d\theta_i \propto \frac{1}{(\sigma_i^2)^{(n_i-1)/2+a+1}(\sigma_i^2+\tau^2)^{1/2}} \exp\left[-\frac{(X_i - \boldsymbol{Z}_i^T, \boldsymbol{\beta})^2}{2(\sigma_i^2+\tau^2)} - \left\{\frac{1}{2}(n_i-1)S_i^2 + \frac{1}{b}\right\}\left(\frac{1}{\sigma_i^2}\right)\right],$$

$$\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) = \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 \mid \boldsymbol{Z}_i, \boldsymbol{B})\, d\sigma_i^2 \propto \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)}$$

où ψ_i est définie dans l'équation (4).

B. Détails de l'algorithme EM

La maximisation de $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$ est effectuée en posant que les dérivées partielles par rapport à $\boldsymbol{B}$ sont nulles, c'est-à-dire

$$\frac{\partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})}{\partial \boldsymbol{B}} = 0. \tag{B.1}$$

Partant de l'expression de $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$ dans le corps du texte, nous obtenons des expressions explicites pour les dérivées partielles par rapport à chaque composante de $\boldsymbol{B}$. La dérive partielle correspondant à β est

$$\frac{\partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})}{\partial \beta} = \sum_{i=1}^{n} \frac{\int \boldsymbol{Z}_i \left(\frac{\theta_i - Z_i^T\beta}{\tau^2}\right) \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i}{\int \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i} = \sum_{i=1}^{n} E\left\{\boldsymbol{Z}_i\left(\frac{\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{\tau^2}\right)\right\}$$

où l'espérance est calculée par rapport à la distribution conditionnelle de θ_i, $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B})$. L'expression de la dérivée partielle correspondant à τ^2 est :

$$\frac{\partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})}{\partial \tau^2} = -\frac{n}{2\tau^2} + \sum_{i=1}^{n} \frac{\int \frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2(\tau^2)^2} \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i}{\int \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i} = -\frac{n}{2\tau^2} + \sum_{i=1}^{n} E\left\{\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2(\tau^2)^2}\right\}.$$

De même, pour a et b, nous obtenons les solutions en posant que $S_a = 0$ et $S_b = 0$, où S_a et S_b sont, respectivement, les dérivées partielles de $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$ par rapport à a et b en utilisant les expressions données dans le corps du texte. Ces équations sont résolues par la méthode de Newton-Raphson qui nécessite la matrice des dérivées secondes par rapport à a et b. Celles-ci sont données par les expressions suivantes :

$$S_{aa} = \sum_{i=1}^{n}\left[\log''\left\{\Gamma\left(\frac{n_i}{2}+a\right)\right\} - \log''\{\Gamma(a)\} + \mathrm{Var}\{\log(\psi_i)\}\right]$$

$$S_{ab} = \sum_{i=1}^{n}\left[-\frac{1}{b} + \frac{1}{b^2}E\left(\frac{1}{\psi_i}\right) - \left(\frac{n_i}{2}+a\right)\frac{1}{b^2} \mathrm{Cov}\left\{\frac{1}{\psi_i}, \log(\psi_i)\right\}\right],$$

et (B.2)

$$S_{bb} = \sum_{i=1}^{n}\left\{\frac{a}{b^2} - (n_i+2a)\frac{1}{b^3}E\left(\frac{1}{\psi_i}\right) + \left(\frac{n_i}{2}+a\right)\frac{1}{b^4} E\left(\frac{1}{\psi_i^2}\right) + \left(\frac{n_i}{2}+a\right)^2\frac{1}{b^4}\mathrm{Var}\left(\frac{1}{\psi_i}\right)\right\}$$

avec $S_{ba} = S_{ab}$. À la u^{e} étape, les mises à jour de a et b sont données par

$$\begin{bmatrix} a^{(u)} \\ b^{(u)} \end{bmatrix} = \begin{bmatrix} a^{(u-1)} \\ b^{(u-1)} \end{bmatrix} - \begin{bmatrix} S_{aa}^{(u-1)} & S_{ab}^{(u-1)} \\ S_{ba}^{(u-1)} & S_{bb}^{(u-1)} \end{bmatrix}^{-1} \begin{bmatrix} S_a^{(u-1)} \\ S_b^{(u-1)} \end{bmatrix}, \tag{B.3}$$

où l'indice supérieur $(u-1)$ sur S_{aa}, S_{ab}, S_{ba}, S_{bb}, S_a et S_b désigne ces quantités évaluées aux valeurs qu'avaient a et b à la $(u-1)^{\mathrm{e}}$ itération. Lorsque la procédure de Newton-Raphson converge, les valeurs de a et b à la t^{e} étape de l'algorithme EM sont fixées à $a^{(t)} = a^{(\infty)}$ et $b^{(t)} = b^{(\infty)}$.

C. Une autre formulation du modèle d'estimation sur petits domaines

Il est possible de réduire la largeur de l'intervalle de confiance $\tilde{C}(\boldsymbol{B})$ en se fondant pour l'estimation sur petits domaines sur un autre modèle hiérarchique qui présente une certaine élégance mathématique. Dans (19), le terme constant $n_i + 2a + 2$ devient $n_i + 2a$ dans cette autre formulation du modèle. Le modèle est donné par

$$X_i \mid \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \tag{C.1}$$

$$\theta_i \mid \sigma_i^2 \sim N(\boldsymbol{Z}_i\boldsymbol{\beta}, \lambda\sigma_i^2), \tag{C.2}$$

$$\frac{(n_i-1)S_i^2}{\sigma_i^2}\,\Big|\,\sigma_i^2 \sim \chi^2_{n_i-1}, \qquad \text{(C.3)}$$

$$\sigma_i^2 \sim \text{Inverse} - \text{Gamma}(a,b), \qquad \text{(C.4)}$$

indépendamment pour $i = 1, 2, ..., n$. Notons que, dans la formule susmentionnée, il est supposé que la variance conditionnelle de θ_i est proportionnelle à σ_i^2, tandis que la variance marginale est constante (en éliminant σ_i^2 par intégration en utilisant (C.4)). Dans (1) et (2), la variance de θ_i est une constante, τ^2, indépendante de σ_i^2, et il n'existe pour θ_i aucune structure conditionnelle dépendant de σ_i^2. L'ensemble de tous les paramètres inconnus dans le modèle hiérarchique courant est $\boldsymbol{B} = (a, b, \boldsymbol{\beta}, \lambda)$. La procédure d'inférence pour ce modèle est donnée ci-après. Le modèle repose essentiellement sur l'hypothèse que les effets réels de petit domaine ne sont pas identiquement distribués, même après avoir éliminé les variations connues.

C.1 Méthodologie d'inférence

En reparamétrisant la variance comme dans (C.2), on obtient certaines simplifications analytiques pour dériver les lois a posteriori de θ_i et σ_i sachant X_i, S_i^2 et $\boldsymbol{B}$. Nous avons

$$\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{B}) = GI\left(\frac{n_i}{2}+a, \left[\frac{(n_i-1)S_i^2}{2}+\frac{(X_i-\boldsymbol{Z}_i\boldsymbol{\beta})^2}{2(1+\lambda)}+\frac{1}{b}\right]^{-1}\right)$$

où $GI(a,b)$ représente la loi Gamma inverse dont les paramètres de forme et d'échelle sont a et b, respectivement. Sachant $\boldsymbol{B}$ et σ_i^2, la distribution conditionnelle de θ_i est

$$\pi(\theta_i \mid X_i, \sigma_i^2, \boldsymbol{B}) = \text{Normale}\left(\boldsymbol{Z}_i^T\boldsymbol{\beta}, \frac{\lambda\sigma_i^2}{1+\lambda}\right).$$

En éliminant σ_i^2 par intégration, on obtient la distribution conditionnelle de θ_i sachant X_i, S_i^2 et $\boldsymbol{B}$,

$$\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}) = \int_0^\infty \pi(\theta_i \mid X_i, \sigma_i^2, B)\,\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{B})\,d\sigma_i^2$$

$$\propto \left\{\frac{(1+\lambda)}{2\lambda}(\theta_i-\boldsymbol{Z}_i^T\boldsymbol{\beta})^2+\frac{\delta^2}{2}\right\}^{-(n_i+2a+1)/2}, \qquad \text{(C.5)}$$

où $\delta^2 = (n_i-1)\,S_i^2 + (X_i-\boldsymbol{Z}_i\boldsymbol{\beta})^2/(1+\lambda) + 2/b$. Nous pouvons réécrire (C.5) sous la forme

$$\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}) = \frac{\Gamma((n_i+1)/2+a)\sqrt{1+\lambda}}{\delta^*\Gamma(n_i/2+a)\sqrt{(n_i+2a)\lambda\pi}} \left\{1+\frac{(\theta_i-\mu_i)^2}{(n_i+2a)\delta^{*2}\lambda/(1+\lambda)}\right\}^{-(n_i+2a+1)/2}$$

qui peut être considéré comme une distribution t à échelle possédant n_i+2a degrés de liberté et le paramètre d'échelle $\delta^*\sqrt{\lambda/(1+\lambda)}$ avec $\delta^{*2} = \delta^2/(n_i+2a)$. D'où,

$$E(\sigma_i^{-1} \mid X_i, S_i, \boldsymbol{B}) = \frac{\Gamma((n_i+1)/2+a)(\delta^2/2)^{-\{(n_i+1)/2+a\}}}{\Gamma(n_i/2+a)(\delta^2/2)^{-(n_i/2+a)}} = \frac{\Gamma((n_i+1)/2+a)}{\Gamma(n_i/2+a)}\,\frac{\sqrt{2}}{\delta^*\sqrt{n_i+2a}}.$$

Dans ce contexte, en choisissant

$$k = k(\boldsymbol{B}) = \left\{1+\frac{t_{\alpha/2}^2}{n_i-1}\right\}^{-(n_i+2a+1)/2}\sqrt{\frac{1+\lambda}{\lambda}}\,\frac{1}{\sqrt{2\pi}},$$

l'intervalle de confiance donné par (8) se simplifie en

$$C_i(\boldsymbol{B}) \equiv \left\{\theta_i : \frac{|\,\theta_i-\mu_i\,|}{\sqrt{\frac{\lambda}{1+\lambda}\,\frac{(n_i+2a)\delta^{*2}}{n_i-1}}} \le t_{\alpha/2}\right\}. \qquad \text{(C.6)}$$

En utilisant les mêmes arguments qu'auparavant et en notant que $(n_i+2a)\delta^{*2} \ge (n_i-1)S_i^2$, nous avons $P\{C_i(\boldsymbol{B})\} \ge P(D_i) = 1-\alpha$, où D_i est l'intervalle de confiance donné par (20). Quand $\boldsymbol{B}$ est inconnu, nous le remplaçons par l'estimation de son maximum de vraisemblance marginale $\hat{\boldsymbol{B}}$. Nous nous attendons à ce que la technique de groupement donne une erreur suffisamment petite pour que $P\{C_i(\hat{\boldsymbol{B}})\} \approx P\{C_i(\boldsymbol{B})\} \ge 1-\alpha$.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satelite data. *Journal of the American Statistical Association*, 95, 28-36.

Bell, W. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. Rapport technique du U.S. Census Bureau.

Casella, G., et Hwang, J. (1991). Evaluating confidence sets using loss functions. *Statistica Sinica*, 1, 159-173.

Chatterjee, S., Lahiri, P. et Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics*, 36, 1221-1245.

Cho, M., Eltinge, J., Gershunskaya, J. et Huff, L. (2002). Evaluation of generalized variance function estimators for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 534-539.

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

Gershunskaya, J., et Lahiri, P. (2005). Variance estimation for domains in the U.S. current employment statistics program. *Proceedings of the Secti on on Survey Research Methods*, American Statistical Association, 3044-3051.

Ghosh, M., et Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 54-76.

Hall, P., et Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested-error regression models. *Annals of Statistics*, 34, 1733-1750.

Huff, L., Eltinge, J. et Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1519-1524.

Hwang, J., Qiu, J. et Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both mean and variances. *Journal of the Royal Statistical Society*, B, 71, 265-285.

Joshi, V. (1969). Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *The Annals of Mathematical Statistics*, 40, 1042-1067.

Maples, J., Bell, W. et Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the american community survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 5056-5067.

Otto, M., et Bell, W. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 160-165.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Revue Internationale de Statistique*, 70, 125-143.

Prasad, N., et Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Qiu, J., et Hwang, J. (2007). Sharp simultaneous intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, 63, 767-776.

Rao, J. (2003). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 145-169.

Rivest, L.-P., et Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*.

Robert, C., et Casella, G. (2004). *Monte Carlo Statistical Methods* (Deuxième édition).

Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82, 499-508.

Wang, J., et Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 1, 107-114.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 205-220
Statistique Canada, N° 12-001-X au catalogue

Indices de conditionnement et décompositions des variances pour le diagnostic de la colinéarité dans l'analyse de données d'enquête au moyen de modèles linéaires

Dan Liao et Richard Valliant [1]

Résumé

Les colinéarités entre les variables explicatives des modèles de régression linéaire affectent les estimations fondées sur des données d'enquête autant que celles fondées sur des données ne provenant pas d'enquêtes. Les effets indésirables sont des erreurs-types inutilement grandes, des statistiques *t* faussement faibles ou élevées et des estimations des paramètres de signe illogique. Les diagnostics de colinéarité disponibles ne conviennent généralement pas pour les données d'enquête, parce que les estimateurs de variance qui y sont intégrés ne tiennent pas compte correctement de la stratification, des grappes et des poids de sondage. Dans le présent article, nous élaborons des indices de conditionnement et des décompositions de variance pour diagnostiquer les problèmes de colinéarité dans des données provenant d'enquêtes complexes. Les diagnostics adaptés sont illustrés au moyen de données provenant d'une enquête sur les caractéristiques de l'état de santé.

Mots clés : Diagnostics pour données d'enquête ; multicolinéarité ; décomposition en valeurs singulières ; inflation de la variance.

1. Introduction

Lorsque les variables explicatives d'un modèle de régression sont corrélées entre elles, on parle de colinéarité. Les effets indésirables de cette dernière sont l'obtention d'erreurs-types inutilement grandes, de statistiques *t* faussement faibles ou élevées, et d'estimations des paramètres de signe illogique ou exagérément sensibles à de faibles variations des valeurs des données. Dans un plan expérimental, il peut être possible de créer des situations où les variables explicatives sont orthogonales les unes par rapport aux autres, mais il n'en va pas de même des données d'observation. Belsley (1991) a souligné que : [traduction] « [...] dans les sciences non expérimentales, ..., la colinéarité est une loi naturelle dans l'ensemble de données résultant des opérations incontrôlables du mécanisme de création des données, et est simplement une réalité douloureuse et inévitable. » Dans de nombreuses enquêtes, des données sur des variables fortement corrélées sont recueillies pour l'analyse. Peu d'analystes des données d'enquête échappent au problème de la colinéarité dans l'estimation par la régression, et l'existence de ce problème complique l'explication statistique précise des relations entre les variables explicatives et les réponses.

Alors que de nombreux diagnostics de régression existent pour les données ne provenant pas d'enquêtes, leur nombre est considérablement plus faible pour les données d'enquête. Les quelques articles existants se concentrent sur la détection des points influents et des groupes influents ayant des valeurs de données ou de poids de sondage anormaux. Elliot (2007) a élaboré des méthodes bayésiennes de troncature des poids des estimateurs par la régression linéaire et par la régression linéaire généralisée sous des plans avec probabilités d'inclusion inégales. Li (2007a, b) et Li et Valliant (2009, 2011) ont étendu une série de techniques diagnostiques classiques à la régression appliquée à des données d'enquête complexes. Leurs articles portent sur les résidus et les effets leviers, plusieurs diagnostics fondés sur la suppression de cas (DFBETA, DFBETAS, DFFIT, DFFITS et distance de Cook) et l'approche pas à pas ascendante (*forward search*). Alors que de nombreuses publications de statistique appliquée offrent des suggestions et des lignes directrices précieuses pour aider les analystes des données à diagnostiquer la présence de colinéarité (par exemple, Belsley, Kuh et Welsch 1980 ; Belsley 1991 ; Farrar et Glauber 1967 ; Fox 1986 ; Theil 1971), presque aucun de ces travaux de recherche ne traite des diagnostics de colinéarité lorsque les modèles sont ajustés en se servant de données d'enquête. Un article antérieur portant sur les problèmes de colinéarité dans le contexte des enquêtes est celui de Liao et Valliant (2012) qui ont adapté des facteurs d'inflation de la variance pour des modèles linéaires ajustés à des données d'enquête.

Supposons que le modèle structurel sous-jacent dans la superpopulation est $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. La matrice $\mathbf{X}$ est une matrice de dimensions $n \times p$ de variables explicatives où n est la taille de l'échantillon ; $\boldsymbol{\beta}$ est un vecteur de dimension $p \times 1$ de paramètres. Les termes d'erreur du modèle ont une structure de variance générale $\mathbf{e} \sim (0, \sigma^2\mathbf{R})$ où σ^2 est une constante inconnue et $\mathbf{R}$ est une matrice de covariance de dimensions $n \times n$ inconnue. Définissons $\mathbf{W}$ comme étant la matrice diagonale des poids de sondage.

1. Dan Liao, RTI International, 7011 3th Street, N.W., Suite 750, Washington DC, 20005. Courriel : dliao@rti.org ; Richard Valliant, University of Michigan et University of Maryland, Joint Program in Survey Methodology, 1218 Lefrak Hall, College Park, MD, 20742.

Nous supposons dans tout l'exposé que les poids de sondage sont construits de façon qu'ils puissent être utilisés pour estimer les totaux de population finie. L'estimateur par les moindres carrés pondérés par les poids de sondage (MCPPS) est donné par

$$\hat{\boldsymbol{\beta}}_{\text{PPS}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y} \equiv \mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y},$$

en supposant que $\mathbf{A} = \mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}$ est inversible. Fuller (2002) décrit les propriétés de cet estimateur. L'estimateur $\hat{\boldsymbol{\beta}}_{\text{PPS}}$ est modélisé sans biais pour $\boldsymbol{\beta}$ sous le modèle $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ que $\text{Var}_M(\mathbf{e}) = \sigma^2\mathbf{R}$ soit spécifiée correctement ou non, et est approximativement sans biais sous le plan pour le paramètre de recensement $\mathbf{B}_U = (\mathbf{X}_U^T\mathbf{X}_U)^{-1}\mathbf{X}_U^T\mathbf{Y}_U$, dans la population finie U de N unités. Les valeurs de population finie du vecteur de réponses et de la matrice des variables explicatives sont $\mathbf{Y}_U = (Y_1, ..., Y_N)^T$ et $\mathbf{X}_U = (\mathbf{X}_1, ..., \mathbf{X}_p)$ où $\mathbf{X}_k$ est le vecteur de dimension $N \times 1$ des valeurs de la covariable k.

La présentation de l'article est la suivante. À la section 2, nous passons en revue les résultats concernant les nombres de conditionnement et les décompositions de variance pour les moindres carrés ordinaires. À la section 3, nous les étendons afin de les adapter à l'estimation fondée sur des données d'enquête. À la section 4, nous donnons certains exemples numériques des techniques. À la section 5, nous présentons nos conclusions. Dans la plupart des dérivations, nous utilisons des calculs fondés sur un modèle car les formes des variances fondées sur un modèle sont utiles pour comprendre les effets de la colinéarité. Cependant, lorsque nous présentons des décompositions de variance, nous utilisons des estimateurs justifiés à la fois par un modèle et par le plan de sondage.

2. Indices de conditionnement et décompositions de variance dans l'estimation par les moindres carrés ordinaires

À la présente section, nous passons brièvement en revue les techniques de diagnostic de la colinéarité dans l'estimation par les moindres carrés ordinaires (MCO) fondées sur des indices de conditionnement et des décompositions de variance. À la section 3, nous étendrons ces méthodes au cas des données d'enquête complexes.

2.1 Valeurs propres et vecteurs propres de $\mathbf{X}^T\mathbf{X}$

S'il existe une relation de colinéarité exacte (parfaite) dans la matrice de données $\mathbf{X}$ de dimensions $n \times p$, nous pouvons trouver un ensemble de valeurs, $\mathbf{v} = (v_1, \ldots, v_p)$, non nulles, tel que

$$v_1\mathbf{X}_1 + \cdots + v_p\mathbf{X}_p = \mathbf{0}, \quad \text{ou} \quad \mathbf{X}\mathbf{v} = \mathbf{0}. \qquad (1)$$

Cependant, en pratique, si la matrice de données ne présente pas de colinéarité exacte, mais plutôt certaines quasi-dépendances, il est parfois possible de trouver un ou plusieurs vecteurs non nuls $\mathbf{v}$ tels que $\mathbf{X}\mathbf{v} = \mathbf{a}$ avec $\mathbf{a} \neq \mathbf{0}$, mais proche de $\mathbf{0}$. Ou bien, nous pourrions dire qu'une quasi-dépendance existe si la longueur du vecteur $\mathbf{a}$, $\| \mathbf{a} \|$, est petite. Pour normaliser le problème consistant à trouver l'ensemble de vecteurs $\mathbf{v}$ qui rend $\| \mathbf{a} \|$ petite, nous considérons uniquement les vecteurs $\mathbf{v}$ de longueur unitaire, c'est-à-dire tels que $\| \mathbf{v} \| = 1$. Belsley (1991) discute du lien des valeurs propres et des vecteurs propres de $\mathbf{X}^T\mathbf{X}$ avec le vecteur normalisé $\mathbf{v}$ et $\| \mathbf{a} \|$. La longueur minimale $\| \mathbf{a} \|$ est simplement la racine carrée positive de la plus petite valeur propre de $\mathbf{X}^T\mathbf{X}$. Le $\mathbf{v}$ qui produit le vecteur $\mathbf{a}$ de longueur minimale doit être le vecteur propre de $\mathbf{X}^T\mathbf{X}$ qui correspond à la plus petite valeur propre. Comme il est discuté à la section suivante, les valeurs propres et les vecteurs propres de $\mathbf{X}$ sont reliés à ceux de $\mathbf{X}^T\mathbf{X}$ et offrent certains avantages lorsque l'on examine la colinéarité.

2.2 Décomposition en valeurs singulières, nombre de conditionnement et indices de conditionnement

La décomposition en valeurs singulières (DVS) de la matrice $\mathbf{X}$ est très étroitement apparentée au système propre de $\mathbf{X}^T\mathbf{X}$, mais possède ses propres avantages. La matrice $\mathbf{X}$ de dimensions $n \times p$ peut être décomposée comme $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, où $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ et $\mathbf{D} = \text{diag}(\mu_1, \ldots, \mu_p)$ est la matrice diagonale des valeurs singulières (ou valeurs propres) de $\mathbf{X}$. Ici, les trois composantes de la décomposition sont des matrices très spéciales, possédant des propriétés hautement exploitables : $\mathbf{U}$ est de dimensions $n \times p$ (la même taille que $\mathbf{X}$) et est à colonnes orthogonales ; $\mathbf{V}$ est de dimensions $p \times p$ et est à colonnes et lignes orthogonales ; $\mathbf{D}$ est de dimensions $p \times p$, non négative et diagonale. Belsley et coll. (1980) ont estimé que la DVS de $\mathbf{X}$ offrait plusieurs avantages par rapport au système propre de $\mathbf{X}^T\mathbf{X}$, tant sur le plan des usages statistiques que de la complexité des calculs. Pour la prédiction, on se concentre sur $\mathbf{X}$ plutôt que sur la matrice à produit croisé $\mathbf{X}^T\mathbf{X}$, puisque $\hat{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$. En outre, les longueurs $\| \mathbf{a} \|$ des combinaisons linéaires (1) de $\mathbf{X}$ qui sont reliées à la colinéarité sont définies convenablement en fonction des racines carrées des valeurs propres de $\mathbf{X}^T\mathbf{X}$, qui sont les valeurs singulières de $\mathbf{X}$. Un deuxième élément à prendre en considération, étant donné la puissance de calcul actuelle, est que la décomposition en valeurs singulières de $\mathbf{X}$ évite le fardeau de calcul supplémentaire lié à la formation de $\mathbf{X}^T\mathbf{X}$, une opération qui comprend np^2

sommes et produits non nécessaires, ce qui peut entraîner une erreur de troncation inutile.

Le nombre de conditionnement de $\mathbf{X}$ est défini comme étant $\kappa(\mathbf{X}) = \mu_{\max} / \mu_{\min}$, où $\mu_{\max}$ et $\mu_{\min}$ sont les valeurs singulières maximale et minimale de $\mathbf{X}$. Les indices de conditionnement sont définis comme $\eta_k = \mu_{\max} / \mu_k$. Plus $\mu_{\min}$ est proche de zéro, plus $\mathbf{X}^T\mathbf{X}$ s'approche d'une matrice singulière. Empiriquement, si une valeur de κ ou η est supérieure à une valeur seuil de, disons, 10 à 30, deux colonnes ou plus de $\mathbf{X}$ présentent des liens moyens à forts. L'occurrence simultanée de plusieurs grandes valeurs de η_k est toujours un indice de l'existence de plus d'une quasi-dépendance.

L'une des questions associées à la DVS est celle de savoir s'il faut centrer les matrices $\mathbf{X}$ autour de leur moyenne. Marquardt (1980) maintient que le centrage des observations élimine le mauvais conditionnement non essentiel. En revanche, Belsley (1984) soutient que le centrage autour de la moyenne masque habituellement le rôle du terme constant dans toute quasi-dépendance sous-jacente. Un cas type est celui de la régression avec variables indicatrices. Par exemple, si le sexe est l'une des variables indépendantes dans une régression et que la plupart des cas étudiés sont des hommes (ou des femmes), la variable indicatrice de sexe peut présenter une forte colinéarité avec l'ordonnée à l'origine. Les discussions consécutives à Belsley (1984) illustrent les divergences d'opinions entre les praticiens (Wood 1984 ; Snee et Marquardt 1984 ; Cook 1984). Qui plus est, en analyse par régression linéaire, Wissmann, Toutenburg et Shalabh (2007) ont découvert que le choix de la catégorie de référence peut influencer le degré de multicolinéarité avec les variables indicatrices. Dans le présent article, nous ne centrons pas les matrices $\mathbf{X}$, mais nous illustrerons l'effet du choix de la catégorie de référence à la section 4.

Un autre problème associé au nombre de conditionnement est qu'il est affecté par l'échelle des mesures x (Steward 1987). En réduisant l'échelle de toute colonne de $\mathbf{X}$, le nombre de conditionnement peut être rendu arbitrairement grand. Cette situation est dénommée *conditionnement artificiellement mauvais*. Belsley (1991) propose de modifier l'échelle de chaque colonne de la matrice de plan $\mathbf{X}$ en utilisant la norme euclidienne de chaque colonne avant de calculer le nombre de conditionnement. Cette méthode est implémentée dans SAS et dans le progiciel *perturb* du logiciel statistique R (Hendrickx 2010). Les deux logiciels utilisent comme procédure standard la racine carrée de la moyenne quadratique de chaque colonne pour le changement d'échelle. Le nombre de conditionnement et les indices de conditionnement des matrices $\mathbf{X}$ mises à l'échelle sont appelés *nombre de conditionnement mis à l'échelle* et *indices de conditionnement mis à l'échelle* de la matrice $\mathbf{X}$. De même, les proportions découlant de la décomposition de la variance pertinentes pour la matrice $\mathbf{X}$ mise à l'échelle (dont nous discuterons à la section suivante) seront appelées *proportions de décomposition de la variance mises à l'échelle*.

2.3 Méthode de décomposition de la variance

Afin d'évaluer la mesure dans la quelle les quasi-dépendances (c'est-à-dire l'existence d'indices de conditionnement élevés pour $\mathbf{X}$ et $\mathbf{X}^T\mathbf{X}$) dégradent la variance estimée de chaque coefficient de régression, Belsley et coll. (1980) ont réinterprété et étendu les travaux de Silvey (1969) en décomposant la variance d'un coefficient en une somme de termes associés chacun à une valeur singulière. Dans la suite de la présente section, nous examinons les résultats des moindres carrés ordinaire (MCO) sous le modèle $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ et $\mathrm{Var}_M(\mathbf{Y}) = \sigma^2\mathbf{I}_n$ où $\mathbf{I}_n$ est la matrice identité de dimensions $n \times n$. À la section 3, ces résultats seront étendus aux moindres carrés pondérés par les poids de sondage. Rappelons que la matrice de variance-covariance sous le modèle de l'estimateur MCO $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ est $\mathrm{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}$. En utilisant la décomposition en valeurs singulières, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, $\mathrm{Var}_M(\hat{\boldsymbol{\beta}})$ peut s'écrire :

$$\mathrm{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2[(\mathbf{U}\mathbf{D}\mathbf{V}^T)^T(\mathbf{U}\mathbf{D}\mathbf{V}^T)]^{-1} = \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T \quad (2)$$

et le k^{e} élément diagonal dans $\mathrm{Var}_M(\hat{\boldsymbol{\beta}})$ est la variance estimée pour le k^{e} coefficient, $\hat{\beta}_k$. En utilisant (2), $\mathrm{Var}_M(\hat{\beta}_k)$ peut s'exprimer :

$$\mathrm{Var}_M(\hat{\beta}_k) = \sigma^2 \Sigma_{j=1}^{p} \frac{v_{kj}^2}{\mu_j^2} \quad (3)$$

où $\mathbf{V} = (v_{kj})_{p\times p}$. Soit $\phi_{kj} = v_{kj}^2 / \mu_j^2$, $\phi_k = \Sigma_{j=1}^{p}\phi_{kj}$ et $\mathbf{Q} = (\phi_{kj})_{p\times p} = (\mathbf{V}\mathbf{D}^{-1}) \cdot (\mathbf{V}\mathbf{D}^{-1})$, où $\cdot$ est le produit (par élément) de Hadamard. Les proportions de décomposition de la variance sont données par $\pi_{jk} = \phi_{jk} / \phi_k$, qui est la proportion de la variance du k^{e} coefficient de régression associée à la j^{e} composante de sa décomposition dans (3). Désignons la *matrice des proportions de décomposition de la variance* par $\boldsymbol{\Pi} = (\pi_{jk})_{p\times p} = \mathbf{Q}^T\bar{\mathbf{Q}}^{-1}$, où $\bar{\mathbf{Q}}$ est la matrice diagonale contenant les sommes de ligne de $\mathbf{Q}$ sur la diagonale principale et des 0 ailleurs.

Si le modèle est donné par $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathrm{Var}_M(\mathbf{Y}) = \sigma^2\mathbf{W}^{-1}$ et que l'on utilise la méthode des moindres carrés pondérés, alors $\hat{\boldsymbol{\beta}}_{\mathrm{MCP}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$ et $\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{MCP}}) = \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$. La décomposition en (3) est vérifiée avec $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$ décomposée comme $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Cependant, dans les applications d'enquête, la matrice de covariance de $\mathbf{Y}$ ne sera virtuellement jamais $\sigma^2\mathbf{W}^{-1}$ si

$\mathbf{W}$ est la matrice des poids de sondage. La section 3 décrit le cas plus raisonnable.

Dans la décomposition de la variance (3), toutes choses étant égales par ailleurs, une faible valeur singulière μ_j peut donner lieu à une grande composante de $\mathrm{Var}(\hat{\beta}_k)$. Toutefois, si v_{kj} est petit aussi, $\mathrm{Var}(\hat{\beta}_k)$ peut ne pas être affecté par une petite valeur μ_j. Un cas extrême est celui où $v_{kj} = 0$. Supposons que les k^{e} et j^{e} colonnes de $\mathbf{X}$ appartiennent à des blocs orthogonaux distincts. Soit $\mathbf{X} \equiv [\mathbf{X}_1, \mathbf{X}_2]$ avec $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$ et soit les décompositions en valeurs singulières de $\mathbf{X}_1$ et $\mathbf{X}_2$ données, respectivement, par $\mathbf{X}_1 = \mathbf{U}_1\mathbf{D}_{11}\mathbf{V}_{11}^T$ et $\mathbf{X}_2 = \mathbf{U}_2\mathbf{D}_{22}\mathbf{V}_{22}^T$. Puisque $\mathbf{U}_1$ et $\mathbf{U}_2$ sont les bases orthogonales pour l'espace couvert par les colonnes de $\mathbf{X}_1$ et de $\mathbf{X}_2$, respectivement, $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$ implique que $\mathbf{U}_1^T\mathbf{U}_2 = \mathbf{0}$ et $\mathbf{U} \equiv [\mathbf{U}_1, \mathbf{U}_2]$ est à colonnes orthogonales. La décomposition en valeurs singulières de $\mathbf{X}$ est simplement $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{U}_2^T$, avec

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{22} \end{bmatrix} \tag{4}$$

et

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{22} \end{bmatrix}. \tag{5}$$

Donc, $\mathbf{V}_{12} = \mathbf{0}$. Un résultat analogue s'applique manifestement à tout nombre de sous-groupes mutuellement orthogonaux. D'où, si toutes les colonnes de $\mathbf{X}$ sont orthogonales, tous les $v_{kj} = 0$ quand $k \neq j$ et $\pi_{kj} = 0$ également. Le fait que v_{kj} soit non nul est un signal que les prédicteurs k et j ne sont pas orthogonaux.

Puisqu'au moins un v_{kj} doit être non nul dans (3), cela implique qu'une forte proportion de toute variance peut être associée à une grande valeur singulière, même en l'absence de colinéarité. L'approche classique consiste à vérifier un indice de conditionnement élevé associé à une forte proportion des variances de deux coefficients ou plus lorsque l'on fait le diagnostic de colinéarité, puisqu'il faut qu'au moins deux colonnes de $\mathbf{X}$ entrent en jeu pour produire une quasi-dépendance. Belsley et coll. (1980) ont proposé de montrer la matrice $\mathbf{\Pi}$ et les indices de conditionnement de $\mathbf{X}$ dans un tableau de décomposition de la variance tel que celui qui suit. Si deux éléments ou plus de la j^{e} ligne de la matrice $\mathbf{\Pi}$ sont relativement grands et que son indice de conditionnement η_j associé est grand également, le signal est que des quasi-dépendances influencent les estimations par la régression.

Indice de	Proportions de la variance			
conditionnement	$\mathrm{Var}_M(\hat{\beta}_1)$	$\mathrm{Var}_M(\hat{\beta}_2)$	$\cdots$	$\mathrm{Var}_M(\hat{\beta}_p)$
η_1	π_{11}	π_{12}	$\cdots$	π_{1p}
η_2	π_{21}	π_{22}	$\cdots$	π_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
η_p	π_{p1}	π_{p2}	$\cdots$	π_{pp}

3. Adaptation aux moindres carrés pondérés par les poids de sondage

3.1 Indices de conditionnement et proportions de décomposition de la variance

Dans le cas des moindres carrés pondérés par les poids de sondage (MCPPS), nous sommes davantage intéressés par les relations de colinéarité entre les colonnes de la matrice $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$ qu'entre celles de $\mathbf{X}$, puisque $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{Y}}$. Définissons la décomposition en valeurs singulières de $\tilde{\mathbf{X}}$ comme étant $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, où les matrices $\mathbf{U}$, $\mathbf{V}$ et $\mathbf{D}$ diffèrent habituellement de celles de $\mathbf{X}$, en raison des poids de sondage inégaux.

Le nombre de conditionnement de $\tilde{\mathbf{X}}$ est défini comme étant $\kappa(\tilde{\mathbf{X}}) = \mu_{\max} / \mu_{\min}$, où $\mu_{\max}$ et $\mu_{\min}$ sont les valeurs singulières maximale et minimale de $\tilde{\mathbf{X}}$. Le nombre de conditionnement de $\tilde{\mathbf{X}}$ est également différent du nombre de conditionnement de la matrice de données $\mathbf{X}$ en raison des poids de sondage inégaux. Les indices de conditionnement sont définis comme

$$\eta_k = \mu_{\max} / \mu_k, \quad k = 1, \ldots, p \tag{6}$$

où μ_k est l'une des valeurs singulières de $\tilde{\mathbf{X}}$. Les indices de conditionnement et les nombres de conditionnement mis à l'échelle sont les indices de conditionnement et les nombres de conditionnement de la matrice $\tilde{\mathbf{X}}$ mise à l'échelle.

Basé sur les extrémums du ratio des formes quadratiques (Lin 1984), le nombre de conditionnement $\kappa(\tilde{\mathbf{X}})$ est borné dans l'intervalle :

$$\frac{w_{\min}^{1/2}}{w_{\max}^{1/2}}\,\kappa(\mathbf{X}) \;\leq\; \kappa(\tilde{\mathbf{X}}) \;\leq\; \frac{w_{\max}^{1/2}}{w_{\min}^{1/2}}\,\kappa(\mathbf{X}), \tag{7}$$

où $w_{\min}$ et $w_{\max}$ sont les poids de sondage minimal et maximal. Cette expression indique que, si les poids de sondage ne varient pas trop, le nombre de conditionnement dans les MCPPS ressemble à celui dans les MCO. En revanche, pour un échantillon présentant une grande gamme de poids de sondage, le nombre de conditionnement peut être très différent dans les MCPPS et les MCO. Quand le nombre de conditionnement des MCPPS est grand, il se peut que celui des MCO ne le soit pas. Dans le cas d'une dépendance linéaire exacte entre les colonnes de $\mathbf{X}$, les colonnes de $\tilde{\mathbf{X}}$ seront également linéairement dépendantes. Dans ce cas extrême, au moins une valeur propre de $\mathbf{X}$ sera nulle, et $\kappa(\mathbf{X})$ et $\kappa(\tilde{\mathbf{X}})$ seront toutes deux infinies. Comme dans les MCO, de grandes valeurs de κ ou de η_k, égales ou supérieures à 10, peuvent indiquer qu'il existe des dépendances moyennes à fortes entre deux colonnes ou plus de $\mathbf{X}$.

La variance sous le modèle de l'estimateur MCPPS des paramètres sous un modèle avec $\text{Var}_M(\mathbf{e}) = \sigma^2\mathbf{R}$ est donnée par :

$$\begin{aligned}\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{PPS}}) &= \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \\ &= \sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{G}, \qquad (8)\end{aligned}$$

où

$$\mathbf{G} = (g_{ij})_{p\times p} = \mathbf{X}^T\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \qquad (9)$$

est la matrice d'*effet de spécification incorrecte* (EFFSI) qui représente le facteur d'inflation nécessaire pour corriger les résultats standard afin de tenir compte de l'effet de la corrélation intra-grappe dans les données d'enquête en grappes et du fait que $\text{Var}_M(\mathbf{e}) = \sigma^2\mathbf{R}$ et non $\sigma^2\mathbf{W}^{-1}$ (Scott et Holt 1982).

En utilisant la décomposition en valeurs singulières de $\tilde{\mathbf{X}}$, nous pouvons réécrire $\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{PPS}})$ sous la forme

$$\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{PPS}}) = \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T\mathbf{G}. \qquad (10)$$

Le k^{e} élément diagonal dans $\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{PPS}})$ est la variance estimée du k^{e} coefficient, $\hat{\beta}_k$. En utilisant (10), on peut exprimer $\text{Var}_M(\hat{\beta}_k)$ comme :

$$\text{Var}_M(\hat{\beta}_k) = \sigma^2\,\Sigma_{j=1}^{p}\,\frac{v_{kj}}{\mu_j^2}\,\lambda_{kj} \qquad (11)$$

où $\lambda_{kj} = \Sigma_{i=1}^{p} v_{ij}g_{ik}$. Si $\mathbf{R} = \mathbf{W}^{-1}$, alors $\mathbf{G} = \mathbf{I}_p$, $\lambda_{kj} = v_{kj}$ et (11) se réduit à (3). Cependant, la situation est plus compliquée quand $\mathbf{G}$ n'est pas la matrice identité, c'est-à-dire quand le plan de sondage complexe affecte la variance d'un coefficient de régression estimé. Si les variables explicatives k et j sont orthogonales, $v_{kj} = 0$ pour $k \neq j$ et, dans (11), la variance dépend uniquement de la k^{e} valeur singulière et n'est pas affectée par les g_{ij} qui ne sont pas nuls. Si la variable explicative k et plusieurs variables explicatives j ne sont pas orthogonales, λ_{kj} reçoit la contribution de tous ces vecteurs propres, et des éléments hors diagonale de la matrice EFFSI, $\mathbf{G}$. Le terme λ_{kj} mesure alors à la fois la non-orthogonalité des x et les effets du plan de sondage complexe.

Par conséquent, nous pouvons définir des proportions de décomposition de la variance analogues à celles obtenues pour les MCO, mais leur interprétation est moins facile. Soit $\phi_{kj} = v_{kj}\lambda_{kj}/\mu_j^2$, $\phi_k = \Sigma_{j=1}^{p}\phi_{kj}$ et $\mathbf{Q} = (\phi_{kj})_{p\times p} = (\mathbf{V}\mathbf{D}^{-2})\cdot(\mathbf{V}^T\mathbf{G})^T$. Les proportions de décomposition de la variance sont $\pi_{jk} = \phi_{jk}/\phi_k$, qui représentent la proportion de la variance du k^{e} coefficient de régression associé à la j^{e} composante de sa décomposition dans (11). Désignons la matrice des proportions de décomposition de la variance par

$$\boldsymbol{\Pi} = (\pi_{jk})_{p\times p} = \mathbf{Q}^T\bar{\mathbf{Q}}^{-1}, \qquad (12)$$

où $\bar{\mathbf{Q}}$ est la matrice diagonale qui contient les sommes de ligne de $\mathbf{Q}$ sur la diagonale principale et des 0 ailleurs. L'interprétation des proportions dans (12) n'est pas aussi catégorique que pour les MCO en raison de l'effet de la matrice EFFSI. À la section 3.2, nous discutons plus en détail de l'interprétation dans le contexte de l'échantillonnage en grappes stratifié.

Comme pour la méthode de régression par les MCO, on peut créer un tableau de décomposition de la variance semblable à celui qui figure à la fin de la section 2. Si deux variables indépendantes ou plus sont colinéaires (ou « quasi dépendantes »), une valeur singulière devrait faire une contribution importante à la variance des estimations du paramètre associées à ces variables. Par exemple, le fait que les proportions π_{31} et π_{32} pour les variances de $\hat{\boldsymbol{\beta}}_{\text{PPS1}}$ et de $\hat{\boldsymbol{\beta}}_{\text{PPS2}}$ sont grandes indique que la contribution de la troisième valeur singulière aux deux variances est importante et que les première et deuxième variables indépendantes de la régression sont, dans une certaine mesure, colinéaires. Comme il est montré à la section 2.3, quand les k^{e} et j^{e} colonnes de $\mathbf{X}$ sont orthogonales, $v_{kj} = 0$ et la proportion de décomposition de la j^{e} valeur singulière π_{jk} sur $\text{Var}(\hat{\beta}_k)$ sera égale à 0.

Plusieurs cas particuliers méritent d'être mentionnés. Si $\mathbf{R} = \mathbf{W}^{-1}$ comme il est supposé dans les moindres carrés pondérés (MCP), alors $\mathbf{G} = \mathbf{I}$. La décomposition de la variance en (11) est de la même forme que l'expression (2) pour les MCO. Cependant, il serait inhabituel que $\mathbf{R} = \mathbf{W}^{-1}$ dans des données d'enquête puisque les poids de sondage ne sont généralement pas calculés en se fondant sur la structure de variance d'un modèle. Notons que $\mathbf{V}$ reste différente de celle pour les MCO et est une composante de la décomposition en valeurs singulières de $\tilde{\mathbf{X}}$ au lieu de $\mathbf{X}$. Un autre exemple est celui où $\mathbf{R} = \mathbf{I}$ et où les poids de sondage sont égaux, auquel cas on peut utiliser les résultats des MCO. Cependant, si les poids de sondage sont inégaux, même quand $\mathbf{R} = \mathbf{I}$, la décomposition de la variance donnée par (11) est différente de celle donnée par (2) dans les MCO, puisque $\mathbf{G} \neq \mathbf{I}$. À la section suivante, nous examinons certains modèles spéciaux qui tiennent compte des caractéristiques de la population, telles que les grappes et les strates, pour estimer cette décomposition de la variance.

3.2 Décomposition de la variance pour un modèle avec mise en grappes stratifiée

La variance sous le modèle de $\hat{\boldsymbol{\beta}}_{\text{PPS}}$ dans (8) contient la matrice $\mathbf{R}$ inconnue qu'il faut estimer. À la présente section, nous présentons un estimateur de $\hat{\boldsymbol{\beta}}_{\text{PPS}}$ approprié pour un modèle avec mise en grappes stratifiée. L'estimateur de variance possède une justification sous le modèle ainsi que sous le plan. Supposons que, dans un plan

d'échantillonnage stratifié à plusieurs degrés, il existe des strates $h = 1, ..., H$ dans la population, des grappes $i = 1, ..., N_h$ dans la strate h et des unités $t = 1, ..., M_{hi}$ dans la grappe hi. Nous sélectionnons les grappes $i = 1, ..., n_h$ dans la strate h et les unités $t = 1, ..., m_{hi}$ dans la grappe hi. Désignons l'ensemble de grappes échantillonnées dans la strate h par s_h et l'échantillon d'unités dans la grappe hi par s_{hi}. Le nombre total d'unités échantillonnées dans la strate h est $m_h = \sum_{i \in s_h} m_{hi}$, et le nombre total dans l'échantillon est $m = \sum_{h=1}^{H} m_h$. Supposons que les grappes sont sélectionnées avec des probabilités variables et avec remise dans les strates et indépendamment entre les strates. Le modèle que nous considérons est :

$$E_M(Y_{hit}) = \mathbf{x}_{hit}^T \boldsymbol{\beta}$$

$$h = 1, \ldots, H,\ i = 1, \ldots, N_h,\ t = 1, \ldots, M_{hi}$$

$$\mathrm{Cov}_M(\varepsilon_{hit}, \varepsilon_{hi't'}) = 0$$

$$\text{où} \quad \varepsilon_{hit} = Y_{hit} - \mathbf{x}_{hit}^T \boldsymbol{\beta}, \quad i \neq i'$$

$$\mathrm{Cov}_M(\varepsilon_{hit}, \varepsilon_{h'i't'}) = 0 \quad h \neq h'. \tag{13}$$

Nous supposons que les unités sont corrélées dans chaque grappe, mais il n'est pas nécessaire ici de spécifier la forme particulière des covariances pour l'analyse. L'estimateur $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$ du paramètre de régression peut s'écrire :

$$\hat{\boldsymbol{\beta}}_{\mathrm{PPS}} = \sum_{h=1}^{H} \sum_{i \in s_h} (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{Y}_{hi} \tag{14}$$

où $\mathbf{X}_{hi}$ est la matrice de dimensions $m_{hi} \times p$ des covariables pour les unités échantillonnées dans la grappe hi, $\mathbf{W}_{hi} = \mathrm{diag}(w_t)$, $t \in s_{hi}$, est la matrice diagonale des poids de sondage pour les unités dans la grappe hi et $\mathbf{Y}_{hi}$ est le vecteur de dimension $m_{hi} \times 1$ des variables réponses dans la grappe hi. La variance sous le modèle de $\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$ est :

$$\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}) = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{G}_{st} \tag{15}$$

où

$$\begin{aligned} \mathbf{G}_{st} &= \left[\sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{R}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi} \right] (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \\ &= \left[\sum_{h=1}^{H} \mathbf{X}_h^T \mathbf{W}_h \mathbf{R}_h \mathbf{W}_h \mathbf{X}_h \right] (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \end{aligned} \tag{16}$$

avec $\mathbf{R}_{hi} = \mathrm{Var}_M(\mathbf{Y}_{hi})$, $\mathbf{W}_h = \mathrm{diag}(\mathbf{W}_{hi})$, et $\mathbf{R}_h = \mathrm{Blkdiag}(\mathbf{R}_{hi})$, $\mathbf{W}_h = \mathrm{diag}(\mathbf{W}_{hi})$, $\mathbf{X}_h^T = (\mathbf{X}_{h1}^T, \mathbf{X}_{h2}^T, ..., \mathbf{X}_{h,n_h}^T)$, $i \in s_h$. L'expression (16) est un cas particulier de (9) avec $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T, ..., \mathbf{X}_H^T)$, où $\mathbf{X}_h$ est la matrice de dimensions $m_h \times p$ des covariables pour les unités échantillonnées dans la strate h, $\mathbf{W} = \mathrm{diag}(\mathbf{W}_{hi})$, pour $h = 1, ..., H$ et $i \in s_h$ et $\mathbf{R} = \mathrm{Blkdiag}(\mathbf{R}_h)$.

En nous inspirant du développement présenté dans Scott et Holt (1982, section 4), nous pouvons réécrire la matrice EFFSI, $\mathbf{G}_{st}$, pour un cas particulier de $\mathbf{R}_h$ de façon qu'il soit plus facile de comprendre les proportions de décomposition dans (12). Considérons le cas particulier de (13) avec

$$\mathrm{Cov}_M(\mathbf{e}_{hi}) = \sigma^2(1-\rho)\,\mathbf{I}_{m_{hi}} + \sigma^2 \rho \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T$$

où $\mathbf{I}_{m_{hi}}$ est la matrice identité de dimensions $m_{hi} \times m_{hi}$ et $\mathbf{1}_{m_{hi}}$ est un vecteur de m_{hi} valeurs 1. Dans ce cas,

$$\begin{aligned} \mathbf{X}_h^T \mathbf{W}_h \mathbf{R}_h \mathbf{W}_h \mathbf{X}_h &= (1-\rho) \mathbf{X}_h^T \mathbf{W}_h^2 \mathbf{X}_h \\ &\quad + \rho \sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{W}_{hi}^2 \mathbf{X}_{Bhi} \end{aligned}$$

où $\mathbf{X}_{Bhi} = m_{hi}^{-1} \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T \mathbf{X}_{hi}$. Supposons que l'échantillon est autopondéré, de sorte que $\mathbf{W}_{hi} = w \mathbf{I}_{m_{hi}}$. Après certaines simplifications, il s'ensuit que

$$\mathbf{G}_{st} = w[\mathbf{I}_p + (\mathbf{M} - \mathbf{I}_p)\,\rho]$$

où $\mathbf{I}_p$ est la matrice identité de dimensions $p \times p$ et $\mathbf{M} = (\sum_{h=1}^{H} \sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{X}_{Bhi})(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$. Donc, si l'échantillon est autopondéré et que ρ est très petit, alors $\mathbf{G}_{st} \approx w \mathbf{I}_p$ et $\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}})$ dans (15) sera approximativement la même que la variance sous MCO. S'il en est ainsi, les proportions de décomposition de la variance sous MCPPS seront comparables aux proportions sous MCO. Dans les problèmes de régression, ρ est souvent petit, puisqu'il s'agit de la corrélation des erreurs, $\varepsilon_{hit} = Y_{hit} - \mathbf{x}_{hit}^T \boldsymbol{\beta}$, pour différentes unités plutôt que pour les $\mathbf{Y}_{hit}$. Cela tient au fait que les effets de plan sont souvent plus faibles pour les coefficients de régression que pour les moyennes – phénomène qui a été constaté pour la première fois par Kish et Frankel (1974). Dans les applications où ρ est plus grand, les proportions de décomposition de la variance dans (12) demeurent utiles pour déceler la colinéarité, même si les écarts par rapport à l'hypothèse d'indépendance des termes d'erreur du modèle ont un effet sur elles.

Représentons les résidus au niveau de la grappe comme un vecteur, $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\boldsymbol{\beta}}_{\mathrm{PPS}}$. L'estimateur de (15) que nous avons considéré découlait au départ de considérations fondées sur le plan de sondage. Un estimateur par linéarisation, approprié quand les grappes sont sélectionnées avec remise, est donné par

$$\mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{PPS}}) = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \hat{\mathbf{G}}_L \tag{17}$$

avec l'effet de spécification incorrecte estimé comme étant

$$\hat{\mathbf{G}}_L = (\hat{g}_{ij})_{p \times p} =$$

$$\left[\sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i \in s_h} (\mathbf{z}_{hi}^* - \overline{\mathbf{z}}_h^*)(\mathbf{z}_{hi}^* - \overline{\mathbf{z}}_h^*)^T \right] (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}, \tag{18}$$

où $\bar{\mathbf{z}}_h^* = 1/n_h \sum_{i\in s} \mathbf{z}_{hi}^*$ et $\mathbf{z}_{hi}^* = \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{e}_{hi}$ avec $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\boldsymbol{\beta}}_{\text{PPS}}$, et la matrice de variance-covariance $\mathbf{R}$ peut être estimée par

$$\hat{\mathbf{R}} = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left[\text{Blkdiag}(\mathbf{e}_{hi} \mathbf{e}_{hi}^T) - \frac{1}{n_h} \mathbf{e}_h \mathbf{e}_h^T \right].$$

L'expression (17) est utilisée, entre autres, dans les progiciels Stata et SUDAAN. L'estimateur $\text{var}_L(\hat{\boldsymbol{\beta}}_{\text{PPS}})$ est convergent et approximativement sans biais sous un plan où les grappes sont sélectionnées avec remise (Fuller 2002). L'estimateur donné par (17) est également un estimateur approximativement sans biais sous le modèle de (15) (voir Liao 2010). Puisque l'estimateur $\text{var}_L(\hat{\boldsymbol{\beta}}_{\text{PPS}})$ est aussi disponible dans les progiciels, nous l'utiliserons dans les travaux empiriques présentés à la section 4.

En partant de l'expression (12) dérivée à la section 2, on peut écrire la matrice des proportions de décomposition de la variance $\boldsymbol{\Pi}$ pour $\text{var}_L(\hat{\boldsymbol{\beta}}_{\text{PPS}})$ sous la forme

$$\boldsymbol{\Pi} = (\pi_{jk})_{p\times p} = \mathbf{Q}_L^T \bar{\mathbf{Q}}_L^{-1} \qquad (19)$$

avec $\mathbf{Q}_L = (\phi_{kj})_{p\times p} = (\mathbf{V}\mathbf{D}^{-2}) \cdot (\mathbf{V}^T \hat{\mathbf{G}}_L)^T$ et $\bar{\mathbf{Q}}_L$ est la matrice diagonale contenant les sommes de ligne de $\mathbf{Q}_L$ sur la diagonale principale et des zéros ailleurs.

4. Exemples numériques

À la présente section, nous illustrons les mesures de colinéarité décrites à la section 3 et examinons leur comportement en utilisant des données sur l'apport alimentaire provenant de la National Health and Nutrition Examination Survey (NHANES) de 2007-2008.

4.1 Description des données

Les données sur l'apport alimentaire sont utilisées pour estimer les types et les quantités d'aliments et de boissons consommés durant la période de 24 heures (de minuit à minuit) qui précède l'entrevue, et pour estimer les apports d'énergie, de nutriments et d'autres composantes alimentaires provenant de ces aliments et boissons. La NHANES est réalisée selon un plan d'échantillonnage probabiliste complexe à plusieurs degrés ; certains sous-groupes de population sont suréchantillonnés afin d'accroître la fiabilité et la précision des estimations des indicateurs de l'état de santé pour ces groupes. Parmi les personnes qui ont répondu à l'interview sur place au centre d'examen mobile (CEM), environ 94 % ont fourni des renseignements complets sur les apports alimentaires. Les poids de sondage ont été construits en prenant les poids d'échantillon ajustés pour le CEM et en les rajustant en outre pour tenir compte de la non-réponse supplémentaire et de la différence de répartition selon le jour de la semaine pour la collecte des données sur les apports alimentaires. Ces poids sont plus variables que les poids produits pour le CEM. Le jeu de données utilisé dans notre étude est un sous-ensemble des données de 2007-2008 composé de femmes de 26 à 40 ans ayant répondu à l'enquête. Les observations comportant des valeurs manquantes pour les variables choisies ont été exclues de l'échantillon qui, en bout de ligne, contient 672 réponses complètes. Les poids finaux dans notre échantillon varient de 6 028 à 330 067, avec un ratio de 55 pour 1. Le National Center for Health Statistics des États-Unis recommande que le plan de sélection de l'échantillon s'approche de la sélection stratifiée avec remise de 32 UPE dans 16 strates, avec 2 UPE dans chaque strate.

4.2 Première étude : covariables corrélées

Dans une première étude empirique, nous avons considéré un modèle de régression linéaire de l'indice de masse corporelle (IMC) des participants à l'enquête. Les variables explicatives utilisées comprennent deux variables démographiques, l'âge et la race (Noir, non-Noir) de la personne, quatre variables binaires indiquant si la personne suit tout régime spécial, un régime pauvre en calories, un régime pauvre en lipides et un régime pauvre en glucides (la valeur est 1 si la personne suit le régime en question, et 0 autrement), et dix variables d'apport nutritionnel total quotidien, qui sont les quantités totales de calories (100 kcal), de protéines (100 g), de glucides (100 g), de sucre (100 g), de fibres alimentaires (100 g), d'alcool (100 g), de lipides totaux (100 g), d'acides gras saturés (100 g), d'acides gras monoinsaturés (100 g) et d'acides gras polyinsaturés (100 g). Les coefficients de corrélation entre ces variables sont présentés au tableau 2. Notons que les corrélations entre les variables d'apport nutritionnel total quotidien sont souvent grandes. Par exemple, les corrélations de l'apport total de lipides avec les apports totaux d'acides gras saturés, d'acides gras monoinsaturés et d'acides gras polyinsaturés sont de 0,85, 0,97 et 0,93.

Trois types de régression ont été spécifiés pour l'échantillon sélectionné afin de faire la démonstration des différents diagnostics. Des renseignements plus détaillés au sujet de ces trois types de régression et de leurs statistiques diagnostiques sont présentés au tableau 1.

TYPE1 : Régression par les moindres carrés ordinaires (MCO) avec estimation de σ^2 ; les statistiques diagnostiques sont obtenues par les méthodes classiques passées en revue à la section 2 ;

TYPE2 : Régression par les moindres carrés pondérés (MCP) avec estimation de σ^2 et en supposant que $\mathbf{R} = \mathbf{W}^{-1}$; les indices de conditionnement mis à l'échelle sont

estimés en utilisant (6) et les proportions de décomposition de la variance mises à l'échelle sont estimées en utilisant (12). Avec $\mathbf{R} = \mathbf{W}^{-1}$, elles correspondent aux décompositions de la variance produites par les progiciels standard en utilisant la méthode MCP et en spécifiant que les poids sont les poids de sondage ;

TYPE3 : Régression par les moindres carrés pondérés par les poids de sondage (MCPPS) avec estimation de $\hat{\mathbf{R}}$; les indices de conditionnement mis à l'échelle sont estimés en utilisant (6) ; les proportions de décomposition de la variance mises à l'échelle sont estimées en utilisant (12).

Les statistiques diagnostiques de ces régressions, y compris les indices de conditionnement et les proportions de décomposition de la variance mis à l'échelle, sont présentées aux tableaux 3, 4 et 5, respectivement. Pour rendre le tableau plus lisible, seules les proportions supérieures à 0,3 sont présentées. Les proportions inférieures à 0,3 sont représentées par des points. Soulignons que, dans la décomposition (12), certains termes peuvent être négatifs, de sorte que certaines « proportions » peuvent être supérieures à 1. Cela se produit dans cinq cas au tableau 5. Selon Belsley et coll. (1980), un indice de conditionnement de 10 indique que la colinéarité a un effet modéré sur les erreurs-types ; un indice de 100 indiquerait un effet important. Dans la présente étude, nous considérons qu'une valeur de l'indice de conditionnement mis à l'échelle supérieure à dix est assez grande et qu'une valeur supérieure à 30 est grande et remarquable. En outre, les grandes proportions de décomposition de la variance mises à l'échelle (supérieures à 0,3) associées à chaque grand indice de conditionnement mis à l'échelle sont utilisées pour repérer les variables touchées par une quasi-dépendance. La corrélation intragrappe des résidus est présentée à la dernière ligne du tableau 6 sous la colonne intitulée « Modèle original ». Dans le modèle utilisé pour les tableaux 3 à 5, $\rho = 0{,}0366$ tel qu'il est estimé d'après un modèle avec effets aléatoires pour les grappes. Comme nous l'avons mentionné à la section 3.2, quand ρ est petit et que l'échantillon est autopondéré, les proportions de décomposition sous MCPPS peuvent être interprétées de la même façon que celles sous MCO. Bien que l'échantillon de la NHANES ne soit pas équipondéré, ρ est petit dans cet exemple et les proportions de décomposition devraient encore fournir des renseignements utiles.

Dans les tableaux 3, 4 et 5, les méthodes de régression pondérées MCP et MCPPS utilisent la matrice de données pondérées par les poids de sondage $\tilde{\mathbf{X}}$ pour obtenir les indices de conditionnement, tandis que la méthode de régression non pondérée, MCO, utilise la matrice de données $\mathbf{X}$. La valeur la plus grande de l'indice de conditionnement mis à l'échelle pour les méthodes MCP et MCPPS est de 566, c'est-à-dire une valeur un peu plus faible que celle de 581 pour les MCO. Ces deux valeurs sont nettement plus grandes que 30 et indiquent donc une quasi-dépendance importante entre les variables explicatives dans les trois modèles de régression. Des nombres de conditionnement d'une telle grandeur impliquent que l'inverse de la matrice de plan, $\mathbf{X}^T\mathbf{W}\mathbf{X}$, peut être numériquement instable, c'est-à-dire que des faibles variations dans les données x pourraient entraîner des variations importantes dans les éléments de l'inverse.

Tableau 1
Modèles de régression et leurs statistiques de diagnostic de la colinéarité utilisés dans cette première étude expérimentale

Type	Méthode de régression	Matrice des poids $\mathbf{W}^a$	var($\hat{\boldsymbol{\beta}}$)	var($\hat{\boldsymbol{\beta}}_k$)	Matrice pour les indices de conditionnement[b]	Proportion π_{jk} de décomposition de la variance
TYPE1	MCO	$\mathbf{I}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$	$\sigma^2\Sigma_{j=1}^{p}\frac{u_{2kj}^2}{\mu_j^2}$ [c]	$\mathbf{X}^T\mathbf{X}$	$\frac{u_{2kj}^2}{\mu_j^2} / \Sigma_{j=1}^{p}\frac{u_{2kj}^2}{\mu_j^2}$
TYPE2	MCP	$\mathbf{W}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\sigma^2\Sigma_{j=1}^{p}\frac{u_{2kj}^2}{\mu_j^2}$ [d]	$\mathbf{X}^T\mathbf{W}\mathbf{X}$	$\frac{u_{2kj}^2}{\mu_j^2} / \Sigma_{j=1}^{p}\frac{u_{2kj}^2}{\mu_j^2}$
TYPE3	MCPPS	$\mathbf{W}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\hat{\mathbf{R}}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\sigma^2\Sigma_{j=1}^{p}\frac{u_{2kj}\Sigma_{i=1}^{p}\hat{g}_{ik}u_{2ij}}{\mu_j^2}$ [e]	$\mathbf{X}^T\mathbf{W}\mathbf{X}$	$\frac{u_{2kj}\Sigma_{i=1}^{p}\hat{g}_{ik}u_{2ij}}{\mu_j^2} / \Sigma_{j=1}^{p}\frac{u_{2kj}\Sigma_{i=1}^{p}\hat{g}_{ik}u_{2ij}}{\mu_j^2}$
			$\hat{\mathbf{R}} = \sum_{h=1}^{H}\frac{n_h}{n_h-1}\left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T) - \frac{1}{n_h}\mathbf{e}_h\mathbf{e}_h^T\right]$			

[a] Dans tous les modèles de régression, les paramètres sont estimés par : $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$.
[b] Les valeurs propres de cette matrice sont utilisées pour calculer les indices de conditionnement pour le modèle de régression correspondant.
[c] Les termes u_{2kj} et μ_j proviennent de la décomposition en valeurs singulières de la matrice des données $\mathbf{X}$.
[d] Les termes u_{2kj} et μ_j proviennent de la décomposition en valeurs singulières de la matrice des données pondérées $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$.
[e] Les termes u_{2kj} et μ_j proviennent de la décomposition en valeurs singulières de la matrice de données pondérées $\tilde{\mathbf{X}}$. Le terme $\hat{g}_{ik}$ est l'élément unitaire de la matrice de l'effet de spécification incorrecte $\hat{\mathbf{G}}$.

Tableau 2
Matrice des coefficients de corrélation de la matrice de données X

	Âge	Race noire	Tout régime	Régime pauvre en calories	Régime pauvre en lipides	Régime pauvre en glucides[a]	Calories	Protéines	Glucides	Sucre	Fibres	Alcool	Lipides totaux	Lipides saturés	Lipides mono.	Lipides poly.
Âge	1															
Race noire	.[b]	1														
Tout régime	.	.	1													
Régime pauvre en calories	.	.	*0,87*[c]	1												
Régime pauvre en lipides	.	.	.	.	1											
Régime pauvre en glucides	.	.	.	.	.	1										
Calories	.	.	.	.	.	.	1									
Protéines	.	.	.	.	.	.	*0,75*	1								
Glucides	.	.	.	.	.	.	*0,84*	0,45	1							
Sucre	.	.	.	.	.	.	*0,58*	.	*0,84*	1						
Fibres	.	.	.	.	.	.	*0,57*	*0,52*	*0,54*	.	1					
Alcool	.	.	.	.	.	.	.	.	.	.	.	1				
Lipides totaux	.	.	.	.	.	.	*0,86*	*0,72*	*0,54*	.	0,48	.	1			
Lipides sat.[d]	.	.	.	.	.	.	*0,74*	*0,56*	0,47	.	0,46	.	*0,85*	1		
Lipides mono.[e]	.	.	.	.	.	.	*0,83*	*0,68*	*0,51*	.	0,46	.	*0,97*	*0,82*	1	
Lipides poly.[f]	.	.	.	.	.	.	*0,81*	*0,71*	*0,51*	.	0,43	.	*0,93*	*0,63*	*0,87*	1

[a] Glucides.
[b] Coefficients de corrélation inférieur à 0,3 sont omises dans ce tableau.
[c] Coefficients de corrélation plus grand que 0,3 sont mises en italique dans ce tableau.
[d] Acides gras saturés totaux.
[e] Acides gras monoinsaturés totaux.
[f] Acides gras polyinsaturés totaux.

Tableau 3
Indices de conditionnement et proportions de décomposition de la variance mis à l'échelle : utilisation de TYPE1 : MCO

	Proportion mise à l'échelle de la variance de								
Indice de conditionnement mis à l'échelle	**Ordonnée à l'origine**	**âge**	**race noire**	**tout régime**	**régime pauvre en calories**	**régime pauvre en lipides**	**régime pauvre en glucides**	**calories**	**protéines**
1	.[a]	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	0,574	.	.
3	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	0,379	.	.	.
4	.	.	0,794	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.	.
11	.	.	.	0,842	0,820	.	.	.	.
12	.	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.
38	0,970	0,960	.	.	.	.	.	.	.
157	.	.	.	.	.	.	.	.	.
581	.	.	.	.	.	.	.	0,993	0,966

Indice de conditionnement mis à l'échelle	**Glucides**	**Sucre**	**Fibres alimentaires**	**Alcool**	**Lipides totaux**	**Lipides sat.[b]**	**Lipides mono.[c]**	**Lipides poly.[d]**
1	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.
26	.	0,633	.	.	.	.	.	.
38	.	.	.	.	.	.	.	.
157	.	.	.	.	0,304	0,866	0,890	0,904
581	0,988	.	0,482	0,986	0,696	.	.	.

[a] Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.
[b] Acides gras saturés totaux.
[c] Acides gras monoinsaturés totaux.
[d] Acides gras polyinsaturés totaux.

Tableau 4
Indices de conditionnement et proportions de décomposition de la variance mis à l'échelle : utilisation de TYPE2 : MCP

	Proportion mise à l'échelle de la variance de								
Indice de conditionnement mis à l'échelle	**Ordonnée à l'origine**	**âge**	**race noire**	**tout régime**	**régime pauvre en calories**	**régime pauvre en lipides**	**régime pauvre en glucides**	**calories**	**protéines**
1	.[a]	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	0,609	.	.
3	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	0,347	.	.	.
4	.	.	0,711	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.
11	.	.	.	0,902	0,878	.	.	.	.
13	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.
37	0,959	0,940	.	.	.	.	.	.	.
165	.	.	.	.	.	.	.	.	.
566	.	.	.	.	.	.	.	0,992	0,963

Indice de conditionnement mis à l'échelle	**Glucides**	**Sucre**	**Fibres alimentaires**	**Alcool**	**Lipides totaux**	**Lipides sat.[b]**	**Lipides mono.[c]**	**Lipides poly.[d]**
1	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.
26	.	0,630	.	.	.	.	.	.
37	.	.	.	.	.	.	.	.
165	.	.	.	.	0,342	0,871	0,909	0,919
566	0,987	.	0,486	0,981	0,658	.	.	.

[a] Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.
[b] Acides gras saturés totaux.
[c] Acides gras monoinsaturés totaux.
[d] Acides gras polyinsaturés totaux.

Tableau 5
Indices de conditionnement et proportions de décomposition de la variance mis à l'échelle : utilisation de TYPE3 : MCPPS

	Proportion mise à l'échelle de la variance de								
Indice de conditionnement mis à l'échelle	**Ordonnée à l'origine**	**âge**	**race noire**	**tout régime**	**régime pauvre en calories**	**régime pauvre en lipides**	**régime pauvre en glucides**	**calories**	**protéines**
1	.[a]	.	.	.	.	.	.	.	.
2	.	.	.	0,717	1,278	0,553	.	.	.
3	.	.	.	.	.	.	0,697	.	.
3	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.
7	0,766	1,686	0,461	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.
37	.	.	.	.	.	.	.	.	.
165	.	.	.	.	.	.	.	.	.
566	0,318	.	.	.	.	.	.	1,095	1,190

[a] Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.
[b] Acides gras saturés totaux.
[c] Acides gras monoinsaturés totaux.
[d] Acides gras polyinsaturés totaux.

Tableau 5 (suite)
Indices de conditionnement et proportions de décomposition de la variance mis à l'échelle : utilisation de TYPE3 : MCPPS

Indice de conditionnement mis à l'échelle	Glucides	Sucre	Fibres alimentaires	Alcool	Lipides totaux	Lipides sat.[b]	Lipides mono.[c]	Lipides poly.[d]
1	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.
26	.	0,379	.	.	.	.	.	.
37	.	.	.	.	.	.	.	.
165	.	.	.	.	.	0,651	0,749	0,615
566	1,008	1,509	0,740	1,036	0,805	0,486	.	0,390

[a] Les proportions de décomposition de la variance mises à l'échelle plus petites que 0,3 sont omises dans ce tableau.
[b] Acides gras saturés totaux.
[c] Acides gras monoinsaturés totaux.
[d] Acides gras polyinsaturés totaux.

Les valeurs des proportions de décomposition sous les MCO et les MCP sont très semblables et aboutissent à la détermination des mêmes variables explicatives comme étant éventuellement colinéaires. Les résultats pour les MCPPS diffèrent quelque peu, comme il est illustré plus bas. Dans le cas des MCO et des MCP, six variables d'apport alimentaire total quotidien – calories, protéines, glucides, alcool, fibres alimentaires et liquides totaux – interviennent dans la quasi-dépendance dominante qui est associée à l'indice de conditionnement mis à l'échelle le plus grand. Quatre variables d'apport quotidien de lipides – lipides totaux, acides gras saturés totaux, acides gras monoinsaturés totaux et acides gras polyinsaturés totaux – interviennent dans la quasi-dépendance secondaire qui est associée au deuxième plus grand indice de conditionnement mis à l'échelle. Les trois tableaux montrent aussi une quasi-dépendance modérée entre l'ordonnée à l'origine et l'âge. L'indice de conditionnement mis à l'échelle associé est égal à 38 pour les MCO, et à 37 pour les MCP et les MCPPS. Cependant, lorsque l'on utilise les MCPPS, le sucre, les acides gras saturés totaux et les acides gras polyinsaturés totaux semblent également intervenir dans la quasi-dépendance dominante comme le montre le tableau 5. Par ailleurs, seulement trois variables d'apport quotidien de lipides – acides gras saturés totaux, acides gras monoinsaturés totaux et acides gras polyinsaturés totaux – interviennent dans la quasi-dépendance secondaire associée au deuxième plus grand indice de conditionnement mis à l'échelle. Donc, lorsqu'on utilise les MCO ou les MCP, l'effet de la quasi-dépendance entre le sucre, les acides gras saturés totaux, les acides gras polyinsaturés totaux et les six variables d'apport nutritionnel total quotidien n'est pas aussi prononcé que dans le cas des MCPPS. Si l'on utilise les diagnostics conventionnels des MCO ou des MCP pour les MCPPS, on pourrait laisser passer cette quasi-dépendance.

Au lieu d'utiliser les indices de conditionnement et la méthode de décomposition de la variance mis à l'échelle (dans les tableaux 3, 4 et 5), un analyste pourrait essayer de déceler les colinéarités en examinant la matrice des coefficients de corrélation non pondérés au tableau 2. Même si la matrice des coefficients de corrélation montre que presque toutes les variables d'apport alimentaire total quotidien sont fortement ou moyennement corrélées par paires, elle ne peut pas être employée pour déceler fiablement les quasi-dépendances entre ces variables quand elles sont utilisées dans une régression. Par exemple, le coefficient de corrélation entre « tout régime » et « un régime pauvre en calories » est assez grand (0,73). Cette quasi-dépendance est associée à un indice de conditionnement mis à l'échelle de 11 (supérieur à 10, mais inférieur au seuil de 30) dans le cas des MCO et des MCP (présentés aux tableaux 3 et 4) et à un indice de conditionnement mis à l'échelle égal à 2 (inférieur à 10) dans le cas des MCPPS (présenté au tableau 5). L'effet de cette quasi-dépendance ne semble pas être très nuisible quelle que soit la méthode de régression utilisée. Par ailleurs, l'alcool est faiblement corrélé à toutes les variables d'apport nutritionnel total quotidien, mais intervient fortement dans la quasi-dépendance dominante présentée à la dernière ligne des tableaux 3 à 5.

Après avoir diagnostiqué les profils de colinéarité, la correction ordinaire consisterait à éliminer les variables corrélées, à réajuster le modèle et à réexaminer les erreurs types, les mesures de colinéarité et d'autres diagnostics. Il est conseillé d'omettre les $\mathbf{X}$ une à la fois en raison des interactions éventuellement complexes entre les variables explicatives. Dans le présent exemple, si l'apport total de

lipides est l'une des variables clés que l'analyste estime devoir garder, le sucre pourrait être abandonné pour commencer, suivi par les protéines, les calories, l'alcool, les glucides, les lipides totaux, les fibres alimentaires, les acides gras monoinsaturés totaux, les acides gras polyinsaturés totaux et les acides gras saturés totaux. D'autres remèdes contre la colinéarité pourraient comprendre la transformation des données ou l'utilisation de techniques spécialisées, telles que la régression ridge et la modélisation bayésienne mixte, qui requièrent de l'information (a priori) supplémentaire qui dépasse le cadre de la plupart des travaux de recherche et des évaluations.

Pour démontrer comment les diagnostics de colinéarité peuvent améliorer les résultats de régression dans le présent exemple, le tableau 6 représente les résultats de l'analyse de régression par les MCPPS des modèles originaux contenant toutes les variables explicatives et d'un modèle réduit contenant un moins grand nombre de ces variables. Dans le modèle réduit, toutes les variables d'apport alimentaire sont éliminées sauf l'apport de lipides totaux. Après réduction du nombre de variables corrélées posant problème, l'erreur-type de l'apport de lipides totaux n'est plus que le 46e de son erreur-type dans le modèle original. L'apport de lipides totaux devient significatif dans le modèle réduit. La réduction du nombre de variables corrélées semble avoir amélioré considérablement l'exactitude de l'estimation de l'effet de l'apport de lipides totaux sur l'IMC. Notons que les diagnostics de colinéarité ne fournissent pas une voie unique vers un modèle final. Le choix des variables explicatives particulières qui doivent être éliminées ou retenues peut varier selon l'analyste.

4.3 Deuxième étude : niveau de référence pour des variables catégoriques

Comme nous l'avons mentionné plus haut, lorsqu'on utilise des données ne provenant pas d'une enquête, les variables indicatrices peuvent aussi jouer un rôle important en tant que source éventuelle de colinéarité. Le choix du niveau de référence pour une variable catégorique peut avoir une incidence sur le degré de colinéarité des données. Plus précisément, choisir comme référence une catégorie dont la fréquence est faible et omettre ce niveau pour ajuster le modèle peut donner lieu à une colinéarité avec le terme d'ordonnée à l'origine. Ce phénomène se transpose à l'analyse des données d'enquête comme nous allons l'illustrer.

Tableau 6
Résultat de l'analyse de régression en utilisant TYPE3 : MCPPS

	Modèle original		Modèle réduit	
Variable	**Coefficient**	**E.-T.[a]**	**Coefficient**	**E.-T.**
Ordonnée à l'origine	24,14***[b]	2,77	24,20***	2,69
Âge	0,06	0,08	0,06	0,08
Race noire	3,19***	1,04	3,67***	0,98
Tout régime[c]	1,79	1,52	1,28	1,80
Régime faible en calories	4,09**	1,50	4,59**	1,69
Régime faible en lipides	3,67	2,86	3,87	3,76
Régime faible en glucides	0,46	3,51	0,87	3,86
Calories	-0,88	2,36		
Protéines	7,05	9,59		
Glucides	3,69	9,62		
Sucre	-0,31	1,11		
Fibres alimentaires	-14,52*	5,89		
Alcool	2,09	16,47		
Lipides totaux	29,34	31,37	1,47*	0,68
Acides gras saturés totaux	-15,90	20,18		
Acides gras monoinsaturés totaux	-22,40	23,01		
Acides gras polyinsaturés totaux	-27,69	21,10		
Coefficient ρ intra-grappe	0,0366		0,0396	

[a] erreur-type.
[b] valeur p : *, 0,05 ; **, 0,01 ; ***, 0,005.
[c] La catégorie de référence est « aucun régime » pour toutes les variables de régime étudiées.

Nous avons employé les quatre variables indicatrices concernant le régime utilisé dans l'étude précédente que nous désignons à la présente section par « tout régime » (RÉGIME), « régime pauvre en calories » (RÉGIMECAL), « régime pauvre en lipides » (RÉGIMELIP) et « régime pauvre en glucides » (RÉGIMEGLU). Le modèle considéré ici est le suivant :

$$\begin{aligned} IMC_{hit} &= \beta_0 + \beta_{noire} * noire_{hit} \\ &+ \beta_{LIPTOT} * LIP.TOT_{hit} \\ &+ \beta_{RÉGIME} * RÉGIME_{hit} \\ &+ \beta_{RÉGIMECAL} * RÉGIMECAL_{hit} \\ &+ \beta_{RÉGIMELIP} * RÉGIMELIP_{hit} \\ &+ \beta_{RÉGIMEGLU} * RÉGIMEGLU_{hit} + \varepsilon_{hit} \quad (20) \end{aligned}$$

où l'indice inférieur *hit* désigne la t^e unité dans l'UPE sélectionnée *hi*, noire est la variable indicatrice de race noire (noire = 1 et non-noire = 0), et LIP.TOT est la variable d'apport total quotidien de lipides. D'après le tableau des fréquences pondérées par les poids de sondage, 15,04 % des répondants suivent « tout régime », 11,43 % d'entre eux suivent un « régime pauvre en calories », 1,33 % suivent un « régime pauvre en lipides » et 0,47 % suivent un « régime pauvre en glucides ». Donc, suivre un régime est un événement relativement rare dans le présent exemple. Si nous choisissons le niveau majoritaire, « aucun régime », comme catégorie de référence pour quatre variables indicatrices de régime, nous ne nous attendons pas à une colinéarité importante entre les variables indicatrices et l'ordonnée à l'origine, parce que la plupart des valeurs des variables indicatrices seront nulles. Cependant, en ajustant le modèle (20), supposons qu'un analyste veuille voir l'effet de « aucun régime » sur l'IMC des répondants et inverse le niveau de référence de la variable RÉGIME dans le modèle (20) pour choisir « tout régime ». Ce changement peut causer une quasi-dépendance dans le modèle, parce que, dans $\mathbf{X}$, la colonne pour la variable RÉGIME sera presque égale à la colonne de valeurs 1 pour l'ordonnée à l'origine. L'étude empirique qui suit illustre l'effet de ce changement sur l'estimation des coefficients de régression et la façon dont nous devons diagnostiquer la gravité de la colinéarité résultante.

Les tableaux 7 et 8 présentent les résultats de l'analyse de régression du modèle (20) en utilisant les trois types de régression – MCO, MCP et MCPPS – énumérés au tableau 1. Le tableau 7 correspond à la modélisation des effets des facteurs de régime suivi sur l'IMC en traitant « aucun régime » comme la catégorie de référence pour les quatre variables de régime suivi, tandis que le tableau 8 correspond au changement du niveau de référence de la variable RÉGIME pour passer de « aucun régime » à « tout régime » et à la modélisation de l'effet de « aucun régime » sur l'IMC. Le choix du niveau de référence a une incidence sur le signe du coefficient estimé pour la variable RÉGIME, mais non sur sa valeur absolue ni sur son erreur-type. La grandeur de l'ordonnée à l'origine estimée et son erreur-type sont différentes dans les tableaux 7 et 8, mais les fonctions estimables, comme les prédictions, seront naturellement les mêmes pour l'un et l'autre ensemble de niveaux de référence. L'erreur-type de l'ordonnée à l'origine est environ trois fois plus grande lorsque la catégorie « tout régime » est le niveau de référence de la variable RÉGIME (tableau 8) que quand elle ne l'est pas (tableau 7).

Tableau 7
Résultat de l'analyse de régression : quand « aucun régime » est la catégorie de référence pour la variable RÉGIME dans le modèle

type de régression	ordonnée à l'origine	noire	lip. tot	tout régime	régime pauvre en calories	régime pauvre en lipides	régime pauvre en glucides
TYPE1	27,22***[a]	3,20***	0,95	3,03	1,75	2,75	-1,48
MCO	(0,61)[b]	(0,70)	(0,72)	(1,94)	(2,03)	(2,72)	(3,66)
TYPE2	26,13***	3,65***	1,44*	1,39	4,46*	3,86	0,94
MCP	(0,58)	(0,82)	(0,67)	(1,67)	(1,79)	(2,59)	(4,22)
TYPE3	26,13***	3,65***	1,44*	1,39	4,46**	3,86	0,94
MCPPS	(0,64)	(0,99)	(0,63)	(1,80)	(1,70)	(3,73)	(3,87)

[a] valeur p : *, 0,05 ; **, 0,01 ; ***, 0,005.
[b] Les erreurs-types sont indiquées entre parenthèses sous les estimations des paramètres.

Tableau 8
Résultat de l'analyse de régression : quand « tout régime » est la catégorie de référence pour la variable RÉGIME dans le modèle

type de régression	ordonnée à l'origine	noir	lip. tot	tout régime	régime pauvre en calories	régime pauvre en lipides	régime pauvre en glucides
TYPE1	30,25***[a]	3,20***	0,95	-3,03	1,75	2,75	-1,48
MCO	(2,00)[b]	(0,70)	(0,72)	(1,94)	(2,03)	(2,72)	(3,66)
TYPE2	27,52***	3,65***	1,44*	-1,39	4,46*	3,86	0,94
MCP	(1,71)	(0,82)	(0,67)	(1,67)	(1,79)	(2,59)	(4,22)
TYPE3	27,52***	3,65***	1,44*	-1,39	4,46**	3,86	0,94
MCPPS	(1,75)	(0,99)	(0,63)	(1,80)	(1,70)	(3,73)	(3,87)

[a] valeur p : *, 0,05 ; **, 0,01 ; ***, 0,005.
[b] Les erreurs-types sont indiquées entre parenthèses sous les estimations des paramètres.

Lorsqu'on choisit « aucun régime » comme catégorie de référence pour RÉGIME au tableau 9, les indices de conditionnement mis à l'échelle sont relativement petits et ne signalent aucune quasi-dépendance remarquable, quel que soit le type de régression. Seule la dernière ligne pour l'indice de conditionnement le plus grand est imprimée dans les tableaux 9 et 10. Souvent, la catégorie de référence d'une variable explicative catégorique est choisie de manière qu'elle ait une signification analytique. Dans le présent exemple, l'utilisation de « aucun régime » serait logique.

Au tableau 10, lorsque l'on choisit « tout régime » comme catégorie de référence par la variable RÉGIME, les indices de conditionnement mis à l'échelle augmentent et indiquent un degré modéré de colinéarité (indice de conditionnement supérieur à 10) entre les variables indicatrices des régimes suivis et l'ordonnée à l'origine. En utilisant le tableau des proportions de décomposition de la variance mise à l'échelle, pour les MCO et les MCP, les variables indicatrices pour « aucun régime » et « régime pauvre en calories » jouent un rôle dans la dépendance dominante avec l'ordonnée à l'origine ; par contre, pour les MCPPS, seule la variable indicatrice pour « aucun régime » joue un rôle dans la quasi-dépendance dominante avec l'ordonnée à l'origine et les trois autres variables de régime suivi sont nettement moins inquiétantes.

5. Conclusion

La dépendance entre les variables explicatives incluses dans un modèle de régression linéaire ajusté sur des données d'enquête affecte les propriétés des estimateurs des paramètres. Les problèmes sont les mêmes que ceux observés pour les données ne provenant pas d'enquêtes : les erreurs-types des estimateurs de pente peuvent être trop grandes et les pentes estimées peuvent avoir un signe illogique. Dans le cas extrême où une colonne de la matrice de plan est exactement une combinaison linéaire d'autres colonnes, les équations d'estimation ne peuvent pas être résolues. Les cas les plus intéressants sont ceux où les variables explicatives sont reliées, mais que la dépendance n'est pas exacte. Les diagnostics de colinéarité disponibles dans les routines des logiciels classiques ne conviennent pas entièrement pour les données d'enquête. Tous les diagnostics qui comportent une estimation de la variance doivent être modifiés pour tenir compte des caractéristiques de l'échantillon telles que la stratification, les grappes et la pondération inégale. Le présent article décrit l'adaptation des nombres de conditionnement et des décompositions de variance, qui peuvent être utilisés pour repérer les cas de dépendance non exacte, afin de les appliquer à l'analyse des données d'enquête.

Le nombre de conditionnement d'une matrice de plan pondérée par les poids de sondage $\mathbf{W}^{1/2}\mathbf{X}$ est égal au ratio de la valeur propre maximale à la valeur propre minimale de la matrice. La quasi singularité de la matrice $\mathbf{X}^T\mathbf{W}\mathbf{X}$ qui doit être inversée lorsque l'on ajuste un modèle linéaire est d'autant plus grande que le nombre de conditionnement est grand. Les valeurs élevées des nombres de conditionnement sont un symptôme de certains des problèmes numériques associés à la colinéarité. Les termes de la décomposition comprennent aussi les « effets de spécification incorrecte » si les erreurs du modèle ne sont pas indépendantes, comme cela serait le cas dans un échantillon en grappes. La variance de l'estimateur d'un paramètre de régression peut aussi s'écrire comme une somme de termes faisant intervenir les valeurs propres de $\mathbf{W}^{1/2}\mathbf{X}$. Les décompositions de la variance pour différents estimateurs des paramètres doivent être utilisées pour repérer les variables explicatives qui sont corrélées entre elles. Après avoir déterminé quelles variables explicatives sont colinéaires, un analyste peut décider si la colinéarité a des effets suffisamment importants sur un modèle ajusté pour justifier de prendre des mesures. La correction la plus simple consiste à éliminer une ou plusieurs variables explicatives, à réajuster le modèle, et à observer comment les estimations changent. Les outils que nous fournissons ici permettent de le faire d'une manière appropriée pour les modèles de régression pondérés par les poids de sondage.

Tableau 9
Indices de conditionnement mis à l'échelle les plus grands et proportions de décomposition de la variance qui y sont associées : quand « aucun régime » est la catégorie de référence pour la variable RÉGIME dans le modèle

Indice de conditionnement mis à l'échelle	Proportion mise à l'échelle de la variance de						
	Ordonnée à l'origine	Sexe	lip. tot	tout régime	régime pauvre en calories	régime pauvre en lipides	régime pauvre en glucides
TYPE1 : MCO 6	0,005	0,000	0,016	0,949	0,932	0,157	0,200
TYPE2 : MCP 6	0,013	0,008	0,020	0,938	0,926	0,189	0,175
TYPE3 : MCPPS 6	0,006	0,007	0,013	0,686	0,741	0,027	0,061

Tableau 10
Indices de conditionnement mis à l'échelle les plus grands et proportions de décomposition de la variance qui y sont associées : quand « tout régime » est la catégorie de référence pour la variable RÉGIME dans le modèle

Indice de conditionnement mis à l'échelle	Proportion mise à l'origine de la variance de						
	Ordonnée à l'origine	Sexe	lip. tot	tout régime	régime pauvre en calories	régime pauvre en lipides	régime pauvre en glucides
TYPE1 : MCO 17	*0,982*	0,001	0,034	*0,968*	*0,831*	0,155	0,186
TYPE2 : MCP 17	*0,982*	0,011	0,029	*0,968*	*0,820*	0,182	0,160
TYPE3 : MCPPS 17	*0,897*	0,018	-0,006	*0,971*	0,318	0,014	-0,019

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs dont les commentaires ont permis d'apporter d'importantes améliorations au texte. Ce travail de recherche a été partiellement financé par la U.S. National Science Foundation (subvention 0617081). Les opinions, découvertes et conclusions ou recommandations exprimées dans ce texte sont celles des auteurs et ne reflètent pas nécessairement celles de la National Science Foundation.

Bibliographie

Belsley, D.A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*, 38(2), 73-77.

Belsley, D.A. (1991). *Conditioning Diagnostics*, *Collinearity and Weak Data in Regression*. New York : John Wiley & Sons, Inc.

Belsley, D.A., Kuh, E. et Welsch, R.E. (1980). *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York : Wiley Interscience.

Cook, R.D. (1984). Comment on demeaning conditioning diagnostics through centering. *The American Statistician*, 2, 78-79.

Elliot, M.R. (2007). Réduction bayésienne des poids pour les modèles de régression linéaire généralisée. *Techniques d'enquête*, 33, 1, 27-40.

Farrar, D.E., et Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.

Fox, J. (1986). *Linear Statistical Models and Related Methods*, *with Applications to Social Research*. New York : John Wiley & Sons, Inc.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 1, 5-25.

Hendrickx, J. (2010). *perturb*: *Tools for evaluating collinearity*. R package version 2.04. Adresse URL http:// CRAN.R-project.org/package= perturb.

Kish, L., et Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society B*, 36(1), 1-37.

Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3341-3348.

Li, J. (2007b). Regression diagnostics for complex survey data. Thèse de doctorat non-publiée, University of Maryland.

Li, J., et Valliant, R. (2009). Matrice chapeau et effets de levier pondérés par les poids de sondage. *Techniques d'enquête*, 35, 1, 17-27.

Li, J., et Valliant, R. (2011). Detecting groups of influential observations in linear regression using survey data-adapting the forward search method. Festschrift for Ken Brewer. *Pakistan Journal of Statistics*, 27, 507-528.

Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Thèse de doctorat, University of Maryland.

Liao, D., et Valliant, R. (2012). Facteurs d'inflation de la variance dans l'analyse des données d'enquêtes complexes. *Techniques d'enquête*, 38, 1, 57-67.

Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.

Marquardt, D.W. (1980). Comment on "A critique on some ridge regression methods" par G. Smith et F. Campbell: "You should standardize the predictor variables in your regression models". *Journal of the American Statistical Association*, 75(369), 87-91.

Scott, A.J., et Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77(380), 848-854.

Silvey, S.D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society*, 31(3), 539-552.

Snee, R.D., et Marquardt, D.W. (1984). Collinearity diagnostics depend on the domain of prediction, and model, and the data. *The American Statistician*, 2, 83-87.

Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.

Theil, H. (1971). *Principles of Econometrics*. New York : John Wiley & Sons, Inc.

Wissmann, M., Toutenburg, H. et Shalabh (2007). Role of categorical variables in multicollinearity in the linear regression model. Rapport technique numéro 008, Department of Statistics, University of Munich. Disponible au http://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf.

Wood, F.S. (1984). Effect of centering on collinearity and interpretation of the constant. *The American Statistician*, 2, 88-90.

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 221-233
Statistique Canada, N° 12-001-X au catalogue

Inférence bayésienne pour les quantiles de population finie sous échantillonnage avec probabilités inégales

Qixuan Chen, Michael R. Elliott et Roderick J.A. Little [1]

Résumé

Le présent article décrit l'élaboration de deux méthodes bayésiennes d'inférence au sujet des quantiles de variables d'intérêt continues d'une population finie sous échantillonnage avec probabilités inégales. La première de ces méthodes consiste à estimer les fonctions de répartition des variables étudiées continues en ajustant un certain nombre de modèles de régression probit avec splines pénalisées sur les probabilités d'inclusion. Les quantiles de population finie sont alors obtenus par inversion des fonctions de répartition estimées. Cette méthode demande considérablement de calculs. La deuxième méthode consiste à prédire les valeurs pour les unités non échantillonnées en supposant qu'il existe une relation variant de façon lisse entre la variable étudiée continue et la probabilité d'inclusion, en modélisant la fonction moyenne ainsi que de la fonction de variance en se servant de splines. Les deux estimateurs bayésiens fondés sur un modèle avec splines donnent un compromis désirable entre la robustesse et l'efficacité. Des études par simulation montrent que les deux méthodes produisent une racine carrée de l'erreur quadratique moyenne plus faible que l'estimateur pondéré par les poids de sondage et que les estimateurs par le ratio et par différence décrits dans Rao, Kovar et Mantel (RKM 1990), et qu'ils sont plus robustes à la spécification incorrecte du modèle que l'estimateur fondé sur un modèle de régression passant par l'origine décrit dans Chambers et Dunstan (1986). Lorsque la taille de l'échantillon est petite, les intervalles de crédibilité à 95 % des deux nouvelles méthodes ont une couverture plus proche du niveau nominal que l'estimateur pondéré par les poids de sondage.

Mots clés : Analyse bayésienne ; fonction de répartition ; erreurs hétéroscédastiques ; régression avec splines pénalisées ; échantillons.

1. Introduction

Nous considérons l'inférence pour les quantiles d'une variable continue d'une population finie d'après un échantillon sélectionné avec probabilités inégales. Les quantiles de population finie sont habituellement estimés par les quantiles pondérés par les poids de sondage, c'est-à-dire un estimateur de type Horvitz-Thompson. Souvent, dans les sondages, la variable de plan de sondage (ici, la probabilité d'inclusion) ou une variable auxiliaire corrélée est mesurée sur des unités non échantillonnées, et cette information peut être utilisée pour accroître l'efficacité des estimateurs pondérés par les poids de sondage (Zheng et Little 2003 ; Chen, Elliott et Little 2010).

Les méthodes d'utilisation d'information auxiliaire pour estimer les fonctions de répartition en population finie ont fait l'objet d'études approfondies. Chambers et Dunstan (1986) ont proposé une méthode fondée sur un modèle et ont illustré leur approche au moyen d'un modèle de régression linéaire avec ordonnée à l'origine nulle pour une superpopulation. Dans la suite de l'exposé, nous donnons à cet estimateur le nom d'estimateur CD. Dorfman et Hall (1993) ont appliqué l'approche CD en remplaçant le modèle de régression linéaire par un modèle non paramétrique. Lombardía, González-Manteiga et Prada-Sánchez (2003, 2004) ont proposé une approximation par le bootstrap de ces estimateurs fondée sur le rééchantillonnage d'une version lissée de la distribution empirique des résidus. Kuk et Welsh (2001) ont également modifié l'approche CD pour résoudre la question des écarts par rapport au modèle en estimant la distribution conditionnelle des résidus sous forme d'une fonction de la variable auxiliaire. Rao, Kovar et Mantel (RKM 1990) ont démontré les avantages des estimateurs par le ratio et par différence fondés sur le plan de sondage par rapport à l'estimateur CD quand le modèle est mal spécifié. Wang et Dorfman (1996) ont proposé une moyenne pondérée des estimateurs CD et RKM. Kuk (1993) a proposé un estimateur à noyau qui combine la distribution connue de la variable auxiliaire avec une estimation par la méthode du noyau de la distribution conditionnelle de la variable étudiée sachant la valeur de la variable auxiliaire. Chambers, Dorfman et Wehrly (1993) ont proposé un estimateur fondé sur un modèle avec lissage par noyau, et Wu et Sitter (2001), ainsi que Harms et Duchesne (2006) ont proposé des estimateurs par calage.

La recherche sur l'utilisation d'information auxiliaire pour l'inférence au sujet des quantiles de population finie (définis comme étant l'inverse de la fonction de répartition) est plus limitée. Chambers et Dunstan (1986) ont discuté de l'estimation en prenant l'inverse de l'estimateur CD de la fonction de répartition, mais n'ont pas comparé les propriétés de cet estimateur des quantiles à d'autres options. Rao et coll. (1990) ont proposé de simples estimateurs par le ratio et par différence des quantiles qui étaient beaucoup

1. Qixuan Chen, professeur adjoint, Department of Biostatistics, Columbia University Mailman School of Public Health, 722 West 168 Street, New York, NY 10032. Courriel : qc2138@columbia.edu ; Michael R. Elliott et Roderick J.A. Little, professeurs, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109. Courriel : mrelliot@umich.edu et rlittle@umich.edu.

plus efficaces que l'estimateur pondéré par les poids de sondage quand la variable d'intérêt résultant de l'enquête était approximativement proportionnelle à la variable auxiliaire.

Nous supposons ici que l'on procède à un échantillonnage avec probabilités inégales où les probabilités d'inclusion sont connues pour toutes les unités de la population. Nous élaborons deux estimateurs bayésiens fondés sur un modèle avec splines des quantiles de population finie dans lequel sont intégrées les probabilités d'inclusion. La première méthode consiste à estimer la fonction de répartition d'un certain nombre de valeurs d'échantillon en utilisant des estimateurs prédictifs bayésiens avec splines pénalisées (Chen et coll. 2010). Nous estimons ensuite les quantiles de population finie en prenant l'inverse de la fonction de répartition prédictive. La deuxième méthode consiste à utiliser un estimateur prédictif bayésien avec splines pénalisées à deux moments, qui prédit les valeurs des unités non échantillonnées en se basant sur un modèle normal, dont la moyenne et la variance sont toutes deux modélisées au moyen de splines pénalisées sur les probabilités d'inclusion. Nous comparons la performance de ces deux nouvelles méthodes à celle de l'estimateur pondéré par les poids de sondage, de l'estimateur CD et des estimateurs par le ratio et par la différence de RKM, en réalisant des études par simulation sur des données générées artificiellement et sur des données d'enquêtes agricoles.

2. Estimateurs des quantiles

Soit s un échantillon aléatoire de taille n tiré avec probabilités inégales de la population finie de N unités identifiables selon les probabilités d'inclusion $\{\pi_i, i = 1, ..., N\}$ que l'on suppose être connues pour toutes les unités avant qu'un échantillon soit tiré. Soit Y une variable étudiée continue, pour laquelle les valeurs $\{y_1, y_2, ..., y_n\}$ sont observées dans l'échantillon aléatoire s. L'α-quantile de Y dans la population finie est défini comme étant :

$$\theta(\alpha) = \inf\left\{t;\ N^{-1}\sum_{i=1}^{N} \Delta(t - y_i) \geq \alpha\right\}, \quad (1)$$

où $\Delta(u) = 1$ quand $u \geq 0$ et $\Delta(u) = 0$ autrement. On estime souvent $\theta(\alpha)$ en utilisant l'α-quantile pondéré par les poids de sondage $\hat{\theta}(\alpha) = \inf\{t, \hat{F}_w(t) \geq \alpha\}$, où $\hat{F}_w(t)$ est la fonction de répartition pondérée par les poids de sondage donnée par

$$\hat{F}_w(t) = \sum_{i\in s} \pi_i^{-1}\Delta(t - y_i) \Big/ \sum_{i\in s} \pi_i^{-1}.$$

Woodruff (1952) a proposé une méthode de calcul des limites de confiance pour l'α-quantile pondéré par les poids de sondage. En premier lieu, on obtient une pseudo-population en pondérant chaque unité de l'échantillon par son poids de sondage ; on estime l'écart-type du pourcentage d'unités inférieur à l'α-quantile estimé ; on multiplie ensuite l'écart-type estimé par le centile z approprié, puis on l'ajoute et on le soustrait de α pour construire les limites de confiance pour le pourcentage d'unités inférieures à l'α-quantile estimé. Enfin, les valeurs de la variable étudiée correspondant aux limites de confiance du pourcentage d'unités inférieures à l'α-quantile estimé sont lues sur les unités pondérées de la pseudo-population rangées par ordre de taille. L'estimation de la variance du pourcentage d'unités de la pseudo-population dont la valeur est inférieure à l'α-quantile estimé est discutée dans Woodruff (1952). Sitter et Wu (2001) ont montré que les intervalles de Woodruff donnent de bons résultats, même dans les cas modérés à extrêmes des queues de la fonction de répartition. Une autre estimation de la variance a été établie par Francisco et Fuller (1991) en utilisant une version lissée de la version du test de signification en grand échantillon.

2.1 Approche fondée sur un modèle bayésien avec inversion de la fonction de répartition estimée

La fonction quantile de population finie est l'inverse de la fonction de répartition (FR) de population finie, définie comme étant $F(t) = N^{-1}\sum_{i=1}^{N} \Delta(t - y_i)$, où $\Delta(x) = 1$ quand $x \geq 0$ et $\Delta(x) = 0$ ailleurs. Nous pouvons estimer les quantiles de population finie en commençant par construire une estimation prédictive continue et strictement monotone de $F(t)$, en traitant $\Delta(t - y)$ comme une variable de résultat binaire et en appliquant des méthodes d'estimation des proportions en population finie.

En particulier, Chen et coll. (2010) ont proposé un estimateur prédictif bayésien avec splines pénalisées (PBSP) pour les proportions de population finie sous échantillonnage avec probabilités inégales. Ils font la régression de la variable étudiée binaire z sur les probabilités d'inclusion dans l'échantillon, en utilisant le modèle de régression probit avec splines pénalisées (2) avec m nœuds fixes présélectionnés :

$$\Phi^{-1}(E(z_i \mid \beta, b, \pi_i)) = \beta_0 + \sum_{k=1}^{p} \beta_k \pi_i^k + \sum_{l=1}^{m} b_l(\pi_i - k_l)_+^p,$$
$$b_l \sim N(0, \tau^2). \quad (2)$$

Des unités autoreprésentatives sont incluses en prenant $\pi_i = 1$. En supposant que les lois a priori pour β et τ^2 sont non informatives, ils ont simulé des tirages de z pour les unités non échantillonnées à partir des lois prédictives a posteriori. Un tirage à partir de la loi a posteriori de la proportion de population finie s'obtient alors en calculant la moyenne des unités échantillonnées observées et des tirages d'unités non échantillonnées. Le procédé est répété de nombreuses fois pour simuler la loi a posteriori de la proportion

de population finie. Des études par simulation ont indiqué que l'estimateur PBSP est plus efficace que l'estimateur pondéré par les poids de sondage et que l'estimateur par la régression généralisée de la proportion de population finie, avec une couverture des intervalles de confiance plus proches des niveaux nominaux.

Nous employons l'approche PBSP n fois pour estimer $F(t)$ à chacune des valeurs échantillonnées de y, $t = \{y_1, y_2, ..., y_n\}$. Cet estimateur ne tient pas compte du fait que nous estimons une fonction de répartition complète, et il ne s'agit pas nécessairement d'une fonction monotone. En outre, l'interpolation linéaire des n fonctions de répartition estimées peut mener à une mauvaise estimation de la fonction de répartition de la population finie. Pour contourner ces deux problèmes, nous ajustons une courbe de régression cubique lisse aux n fonctions de répartition estimées en imposant des contraintes de monotonicité (Wood 1994). Nous désignons la fonction de répartition estimée résultante par $\hat{F}(t)$. L'estimateur fondé sur un modèle bayésien de $\theta(\alpha)$, obtenu par inversion de la fonction de répartition (FR), est alors défini comme il suit :

$$\hat{\theta}_{\text{inv-FR}}(\alpha) = \inf\{t;\ \hat{F}(t) \geq \alpha\}. \qquad (3)$$

Nous ajustons également deux autres courbes de régression lisse monotone aux limites supérieures et inférieures des intervalles de crédibilité (IC) à 95 % de ces fonctions de répartition estimées, désignées par $\hat{F}_U(t)$ et $\hat{F}_L(t)$. Afin de réduire le temps de calcul dans nos études par simulation, nous estimons uniquement la fonction de répartition à $k < n$ points présélectionnés de l'échantillon.

L'idée fondamentale qui sous-tend cette approche est illustrée graphiquement à la figure 1. Supposons qu'un échantillon de taille 100 est tiré d'une population finie. Nous choisissons 20 observations dans l'échantillon et estimons les fonctions de répartition respectives et les IC à 95 % associés en utilisant l'estimateur PBSP. À la figure 1(a), nous représentons les estimations PBSP pour ces 20 observations par des points noirs et les limites inférieure et supérieure de l'IC à 95 %, par des signes « - » que nous relions par un trait plein. À la figure 1(b), nous ajoutons trois courbes de prédiction lisses monotones en utilisant un trait plein noir pour l'estimation ponctuelle et des traits pointillés noirs pour les limites supérieure et inférieure des IC à 95 %.

À la figure 1(c), nous traçons à travers le graphique une droite horizontale passant par la valeur α sur l'axe des y. Nous lisons x_A, x et x_B respectivement sur l'axe des x de façon telle que $\hat{F}_L(x_A) = \alpha$, $\hat{F}(x) = \alpha$ et $\hat{F}_U(x_B) = \alpha$. Alors, x est l'estimation bayésienne avec inversion de la fonction de répartition de $\theta(\alpha)$. Si l'IC à 95 % de la fonction de répartition $F(\cdot)$ est construit en divisant en parties égales les queues de la distribution a posteriori, l'intervalle formé par x_A et x_B est un IC à 95 % de $\theta(\alpha)$. La preuve en est la suivante : si α est la limite inférieure de l'IC à 95 % de $F(x_A)$, seulement 2,5 % des tirages de $F(x_A)$ dans la distribution a posteriori sont plus petits que α. C'est-à-dire que

$$\Pr(F^{-1}(\alpha) > F^{-1}(F(x_A))) \equiv \Pr(\theta(\alpha) > x_A) = 0,025.$$

De même si α est la limite supérieure de l'IC à 95 % de $F(x_B)$, $\Pr(\theta(\alpha) < x_B) = 0,975$. Par conséquent, il y a une probabilité de 95 % que $\theta(\alpha)$ soit compris entre x_A et x_B dans la distribution a posteriori, étant donné l'échantillon.

Cette approche fondée sur un modèle bayésien avec inversion de la fonction de répartition permet d'éviter de fortes hypothèses de modélisation et peut être appliquée à des distributions normales ou asymétriques. L'estimation de la fonction de répartition à chacune des n unités échantillonnées permet d'utiliser complètement l'information fournie par l'échantillon, mais requiert d'importants calculs ; l'estimation de la fonction de répartition à $k < n$ valeurs réduit le temps de calcul au prix d'une certaine perte d'efficacité. Dans l'approche classique, les quantiles de population sont estimés par inversion de la fonction de répartition empirique non lissée. Nous recommandons d'ajuster une courbe de régression cubique lisse aux fonctions de répartition estimées avant d'inverser la fonction de répartition estimée résultante. Les estimations résultantes des quantiles sont plus efficaces, parce que la courbe lisse exploite l'information provenant de toutes les données. Des simulations dont les résultats ne sont pas présentés ici donnent à penser que la courbe de la fonction de répartition estimée en se basant sur un sous-ensemble bien choisi de k unités échantillonnées est similaire à celle estimée en se basant sur la totalité des unités échantillonnées, mais le temps de calcul est réduit considérablement.

Nous suggérons de choisir le sous-ensemble de k points de données à intervalles égaux dans le milieu de la distribution, et à intervalles plus fréquents dans les extrémités afin d'améliorer l'estimation de la fonction de répartition dans les queues. Par exemple, dans notre étude par simulation avec un échantillon de taille 100, nous avons estimé les fonctions de répartition à 20 points : les 3 valeurs les plus faibles, les 3 valeurs les plus grandes et 14 autres points uniformément espacés dans le milieu de l'échantillon rangé par ordre de valeur.

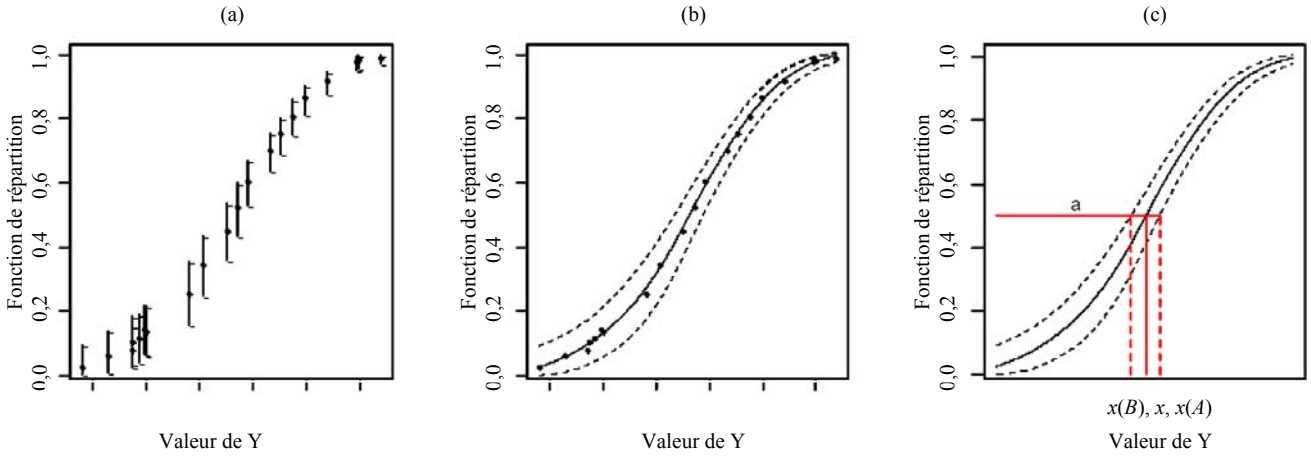


Figure 1 Approche fondée sur un modèle bayésien avec inversion de la fonction de répartition (FR) pour estimer les fonctions de répartition de population finie et les quantiles associés, illustrée en utilisant un échantillon de taille 100 tiré d'une population finie. (a) La méthode PBSP est utilisée pour estimer les fonctions de répartition de la population finie à vingt points de l'échantillon ; les points représentent les estimateurs PBSP et les signes moins représentent les limites supérieure et inférieure des IC à 95 %. (b) Trois modèles de régression cubiques lisses monotones sont ajustés sur les estimateurs PBSP, les limites supérieures et les limites inférieures, respectivement ; la courbe en trait plein représente les fonctions de répartition continues prédictives et les deux courbes en trait interrompu représentent les IC à 95 % des fonctions de répartition. (c) L'estimation ponctuelle et l'IC à 95 % de l' α- quantile de population sont obtenus en inversant la fonction de répartition estimée ; x est l'estimation ponctuelle et x(B) et x(A) sont les limites inférieure et supérieure de l'IC à 95 %

2.2 Approche prédictive bayésienne avec deux moments modélisés par splines pénalisées

Nous considérons d'autres estimateurs des quantiles de population finie de la forme :

$$\tilde{\theta}(\alpha) = \inf\left\{t;\ N^{-1}\left(\sum_{i\in s}\Delta(t - y_i) + \sum_{j\notin s}\Delta(t - \hat{y}_j)\right) \geq \alpha\right\}, \quad (4)$$

où $\hat{y}_j$ est la valeur de la j^{e} unité non échantillonnée prédite par une régression sur les probabilités d'inclusion $\{\pi_i\}$. Un modèle normal de base pour un résultat continu repose sur l'hypothèse d'une fonction moyenne linéaire en $\{\pi_i\}$, c'est-à-dire :

$$Y_i \overset{\text{ind}}{\sim} N(\beta_0 + \beta_1\pi_i, c_i\sigma^2), \quad (5)$$

avec des constantes c_i connues pour modéliser la variance non constante. Ce modèle donne une estimation biaisée de $\theta(\alpha)$ si la relation n'est pas linéaire. Pour estimer les totaux de population finie, Zheng et Little (2003, 2005) ont remplacé dans (5) la fonction moyenne linéaire par une spline pénalisée, et ont supposé que $c_i = \pi_i^{2k}$ pour une certaine valeur connue de k. Des simulations ont donné à penser que leur estimateur fondé sur un modèle du total de population finie donne de meilleurs résultats que l'estimateur pondéré par les poids de sondage, même quand la structure de variance est mal spécifiée.

Pour l'estimation des quantiles au lieu du total, il est important de spécifier correctement la structure de la variance afin d'éviter un biais. Par conséquent, nous étendons le modèle avec spline pénalisée de Zheng et Little (2003) en modélisant la moyenne ainsi que la variance en utilisant des splines pénalisées. Le modèle avec deux moments modélisés par splines pénalisées peut s'écrire (Ruppert, Wand et Carroll 2003, page 264) :

$$Y_i \overset{\text{ind}}{\sim} N(\text{SPL}_1(\pi_i, k), \exp(\text{SPL}_2(\pi_i, k'))),$$

$$\text{SPL}_1(\pi_i, k) = \beta_0 + \sum_{k=1}^{p}\beta_k\pi_i^k + \sum_{l=1}^{m_1} b_l(\pi_i - k_l)_+^p,$$

$$b_l \overset{\text{iid}}{\sim} N(0, \tau_b^2),$$

$$\text{SPL}_2(\pi_i, k') = \alpha_0 + \sum_{k=1}^{p}\alpha_k\pi_i^k + \sum_{l=1}^{m_2} \nu_l(\pi_i - k'_l)_+^p,$$

$$\nu_l \overset{\text{iid}}{\sim} N(0, \tau_\nu^2). \quad (6)$$

Dans (6), la moyenne et le logarithme de la variance sont modélisés par des splines pénalisées (SPL_1) et (SPL_2) sur $\{\pi_i\}$. La modélisation du logarithme de la variance fait en sorte que les estimations de la variance soient positives. Nous permettons des nombres (m_1, m_2) et des emplacements (k, k') différents des nœuds pour les deux splines.

Ruppert et coll. (2003) ont proposé une approche itérative pour estimer les paramètres de (6). Ils ont d'abord supposé que SPL_2 était connue et ont ajusté un modèle linéaire mixte pour estimer les paramètres dans SPL_1. Ils ont calculé le carré de la différence entre Y et SPL_1, qui suivait une loi gamma de paramètre de forme ½ et de paramètre d'échelle $2\mathrm{SPL}_2$. Ils ont ensuite ajusté un modèle linéaire mixte généralisé pour les carrés des différences afin d'estimer les paramètres dans SPL_2. Ils ont itéré les procédés susmentionnés jusqu'à ce que les estimations des paramètres convergent. Cette approche itérative est simple à mettre en œuvre. Cependant, ici, notre objectif n'est pas d'estimer les paramètres, mais d'obtenir des prédictions bayésiennes de Y pour les unités non échantillonnées afin de pouvoir utiliser (4) pour estimer les quantiles.

Crainiceanu, Ruppert, Carroll, Joshi et Goodner (2007) ont élaboré une méthodologie inférencielle bayésienne pour (6). Ils ont constaté que la mise en œuvre de la méthode MCMC en utilisant des pas de Metropolis-Hastings multivariés est instable avec de mauvaises propriétés de mélange. Ils ont proposé d'ajouter des termes d'erreur à la deuxième spline pour rendre les calculs plus faisables, en remplaçant l'échantillonnage à partir de lois conditionnelles complètes complexes par de simples pas de Metropolis-Hastings univariés. Cette idée peut s'exprimer comme

$$Y_i \overset{\text{ind}}{\sim} N(\mathrm{SPL}_1(\pi_i, k), \sigma_\varepsilon^2(\pi_i)),$$

$$\log(\sigma_\varepsilon^2(\pi_i)) \overset{\text{iid}}{\sim} N(\mathrm{SPL}_2(\pi_i, k'), \sigma_A^2).$$

Nous avons utilisé une loi a priori $N(0, 10^6)$ pour les paramètres à effets fixes β et α, et une loi a priori gamma inverse propre $\mathrm{IGamma}(10^{-6}, 10^{-6})$ pour les composantes de la variance τ_b^2 et τ_v^2. Nous avons fixé les valeurs de $\sigma_A^2 = 0{,}1$. Les lois conditionnelles a posteriori complètes sont décrites en détail dans Crainiceanu et coll. (2007).

La loi a posteriori de l'α-quantile de population finie est simulée en générant un grand nombre D de tirages et en utilisant l'estimateur prédictif de la forme

$$\tilde{\theta}^{(d)}(\alpha) = \inf\left\{t;\ N^{-1}\left(\sum_{i\in s}\Delta(t - y_i) + \sum_{j\notin s}\Delta(t - \hat{y}_j^{(d)})\right) \geq \alpha\right\},$$

où $\hat{y}_j^{(d)}$ est un tirage à partir de la loi prédictive a posteriori de la j^{e} unité non échantillonnée de la variable résultat continue. La moyenne de ces tirages simule l'estimateur prédictif bayésien avec deux moments modélisés par splines pénalisées (PB2SP) de l'α-quantile de population finie,

$$\hat{\theta}_{\mathrm{PB2SP}}(\alpha) = D^{-1}\sum_{d=1}^{D}\tilde{\theta}^{(d)}(\alpha).$$

L'intervalle de crédibilité à 95 % bayésien pour l'α-quantile de population dans les simulations est formé en divisant également la queue de la distribution entre les points finaux supérieur et inférieur.

3. Étude par simulation

3.1 Étude par simulation avec données artificielles

Nous avons d'abord simulé une superpopulation de taille $M = 20\,000$. La variable de taille X dans la superpopulation prend 20 000 valeurs entières consécutives allant de 710 à 20 709. Puis, nous avons tiré une population finie de taille $N = 2\,000$ de cette superpopulation par échantillonnage systématique avec probabilité proportionnelle à la taille (ppt) où la probabilité était proportionnelle à l'inverse de la variable de taille. Par conséquent, dans la population finie, la distribution de la variable de taille est asymétrique avec étalement à droite. La variable résultat étudiée Y a été tirée d'une loi normale de moyenne $f(\pi)$ et de variance d'erreur égale à 0,04 (erreur homoscédastique) ou π (erreur hétéroscédastique). Trois structures de moyenne $f(\pi)$ ont été simulées : pas d'association entre Y et π (NULL) $f(\pi) = 0{,}5$, une association linéaire (LINUP) $f(\pi) = 6\pi$, et une association non linéaire (EXP) $f(\pi) = \exp(-4{,}64 + 52\pi)$. Pour chacune des six conditions de simulation, nous avons généré un millier de répliques de la population finie et nous avons tiré de chaque population un échantillon ppt systématique ($n = 100$) avec x comme variable de taille ; donc $\pi_i = nx_i / \sum_{j=1}^{N} x_j$. Les nuages de points de Y en fonction de π pour ces six populations sont présentés à la figure 2.

Nous avons comparé les propriétés de l'estimateur bayésien avec fonction de répartition inverse et de l'estimateur bayésien PB2SP à cinq autres approches :

a) PS, l'estimateur pondéré par les poids de sondage défini par inversion de $\hat{F}_w$;
b) PS lisse, l'estimateur pondéré par les poids de sondage lisse. Une courbe de régression cubique lisse a été ajustée à $\hat{F}_w$ et désignée par $\tilde{F}_w$. L'estimateur pondéré par les poids de sondage lisse est alors défini comme $\tilde{\theta}_w = \inf\{t;\ \tilde{F}_w \geq \alpha\}$;
c) CD, l'estimateur de Chambers et Dunstan (1986), en supposant le modèle suivant : $Y_i = \beta\pi_i + \sqrt{\pi_i}U_i$, où U_i est une variable aléatoire dont les valeurs sont indépendantes et identiquement distribuées de moyenne nulle ;
d) Ratio, l'estimateur par le ratio de RKM (1990) donné par $\{\hat{\theta}_y(\alpha) / \hat{\theta}_x(\alpha)\} \times \theta_x(\alpha)$, où $\hat{\theta}_y(\alpha)$ et $\hat{\theta}_x(\alpha)$ désignent respectivement les estimations pondérées par les poids de sondage pour Y et la

variable de taille X, et $\theta_x(\alpha)$ est le quantile de population connu de X ;

e) Diff, l'estimateur par la différence de RKM (1990) donné par $\hat{\theta}_y(\alpha) + \hat{R} \times \{\theta_x(\alpha) - \hat{\theta}_x(\alpha)\}$, où $\hat{R}$ est l'estimation pondérée par les poids de sondage de Y/X.

Les sept estimateurs pour les 10^e, 25^e, 50^e, 75^e et 90^e centiles de la population finie ont été comparés pour ce qui est du biais empirique et de la racine carrée de l'erreur quadratique moyenne (REQM). Étant donné la complexité de l'estimation de la variance des estimateurs CD et RKM, nous avons comparé seulement la largeur moyenne et le taux de non-couverture de l'intervalle de confiance/crédibilité (IC) à 95 % pour les deux estimateurs fondés sur un modèle bayésien et l'estimateur pondéré par les poids de sondage. Pour l'IC à 95 %, nous avons utilisé la méthode de Woodruff pour l'estimateur pondéré par les poids de sondage, la méthode illustrée à la figure 1(c) pour l'estimateur bayésien avec fonction de répartition inverse et la probabilité a posteriori de 95 % du quantile avec queues égales pour l'estimateur PB2SP. Nous avons utilisé des splines cubiques avec 15 nœuds également espacés.

Les tableaux 1 et 2 montrent le biais empirique et la REQM pour les trois distributions normales avec erreurs homoscédastiques et erreurs hétéroscédastiques, respectivement. Dans l'ensemble, le biais empirique dans l'estimation des cinq quantiles est semblable lorsque l'on utilise les estimateurs bayésiens, les deux estimateurs pondérés par les poids de sondage et les deux estimateurs fondés sur le plan de sondage de RKM. Par contre, l'estimateur CD produit un grand biais et une grande REQM dans tous les scénarios, sauf LINUP avec erreur hétéroscédastique, où le modèle sous-jacent de l'estimateur est spécifié correctement. Les deux estimateurs fondés sur un modèle bayésien produisent des racines carrées de l'erreur quadratique moyenne plus petites que les autres estimateurs, et cet accroissement de l'efficacité est important dans certains scénarios, en particulier lorsque l'on utilise l'estimateur PB2SP. Par l'application d'une courbe de régression cubique lisse à la fonction de répartition empirique estimée pondérée par les poids de sondage, l'estimateur pondéré par les poids de sondage lisse produit un gain d'efficacité par rapport aux estimateurs pondérés par les poids de sondage classiques, mais la REQM demeure plus grande que pour l'estimateur bayésien avec fonction de répartition inverse. Les comparaisons des trois estimateurs fondés sur le plan de sondage donnent à penser qu'aucun de ces estimateurs ne domine uniformément les deux autres. En particulier, l'estimateur pondéré par les poids de sondage a une plus petite REQM que les estimateurs par différence et par le ratio de RKM pour les cinq quantiles dans la population NULL et pour les quantiles inférieurs dans les populations LINUP et EXP ; par ailleurs, les estimateurs RKM ont une plus petite REQM pour les quantiles supérieurs dans les populations LINUP et EXP.

Le tableau 3 donne la largeur moyenne et le taux de non-couverture de l'IC à 95 % pour les deux estimateurs fondés sur un modèle bayésien et l'estimateur pondéré par les poids de sondage. Dans l'ensemble, les deux estimateurs fondés sur un modèle bayésien donnent de plus courtes largeurs moyennes de l'IC à 95 % que l'estimateur pondéré par les poids de sondage. Le taux de couverture de l'IC à 95 % est comparable pour les trois estimateurs, excepté quand α est égal à 0,1, auquel cas l'IC à 95 % de l'estimateur PB2SP possède la largeur moyenne la plus courte et une très bonne couverture, tandis que l'estimateur pondéré par les poids de sondage présente une sous-couverture importante. Cette situation est due au fait que la méthode de Woodruff pour estimer la variance de l'estimateur pondéré par les poids de sondage est fondée sur une hypothèse de grand échantillon, alors qu'ici l'échantillonnage ppt fait qu'un petit nombre seulement de cas sont échantillonnés dans la queue inférieure de la distribution.

Bien que l'estimateur pondéré par les poids de sondage se comporte de manière comparable aux estimateurs fondés sur un modèle bayésien avec splines pour ce qui est du biais empirique global, le biais conditionnel des estimations varie considérablement à mesure qu'augmente la moyenne d'échantillon des probabilités d'inclusion. À l'exemple de Royall et Cumberland (1981), nous avons classé les estimations provenant des 1 000 échantillons en fonction de la moyenne d'échantillon des probabilités d'inclusion et nous les avons réparties en 20 groupes de 50, puis nous avons calculé le biais empirique pour chaque groupe. La figure 3 donne le biais conditionnel des deux estimateurs bayésiens et de l'estimateur pondéré par les poids de sondage pour le 90^e centile dans le cas « EXP + erreur homoscédastique ». La figure 3 montre une tendance linéaire du biais dans l'estimateur pondéré par les poids de sondage à mesure qu'augmente la moyenne d'échantillon des probabilités d'inclusion, tandis que le biais de groupe des deux estimateurs fondés sur un modèle bayésien avec splines est moins affecté par cette moyenne. Des constatations comparables sont faites pour d'autres scénarios.

3.2 Étude par simulation avec les données de l'enquête sur les exploitations agricoles à grande échelle

L'estimateur PB2SP repose sur l'hypothèse que la variable résultat suit une loi normale, après conditionnement sur les probabilités d'inclusion. Puisque l'approche fondée sur un modèle bayésien avec fonction de répartition inverse ne comporte pas d'hypothèse de normalité, nous pourrions

nous attendre à ce qu'elle donne de meilleurs résultats que l'approche PB2SP lorsque l'hypothèse de normalité est violée. Cela motive une comparaison de l'estimateur pondéré par les poids de sondage et de l'estimateur bayésien avec fonction de répartition inverse pour des données ne suivant pas une loi normale.

La population considérée ici est définie par 398 exploitations agricoles à grande échelle (qui produisent des céréales, des bovins, des moutons et de la laine) ayant une superficie agricole de 6 000 hectares ou moins qui ont participé à l'Australian Agricultural and Grazing Industries Survey de 1982 réalisée par l'Australian Bureau of Agricultural and Resource Economics (ABARE 2003). La variable Y est le total des recettes monétaires agricoles. Nous avons tiré 1 000 échantillons systématiques ppt de taille égale à 100 en prenant la superficie agricole, X, comme variable de taille, de sorte que les grandes exploitations agricoles sont plus susceptibles que les autres d'être sélectionnées dans l'échantillon. La figure 4 donne le nuage de points de Y en fonction de la variable de taille X pour ces exploitations, chaque cercle plein représentant un échantillon ppt sélectionné. Ce graphique montre que la variation de Y augmente à mesure que X augmente. En outre, la distribution de Y est étalée vers la droite étant donné X. Nous avons réalisé une étude par simulation en utilisant ces données sur les exploitations agricoles à grande échelle pour comparer les deux estimateurs fondés sur un modèle bayésien avec splines à l'estimateur pondéré par les poids de sondage.

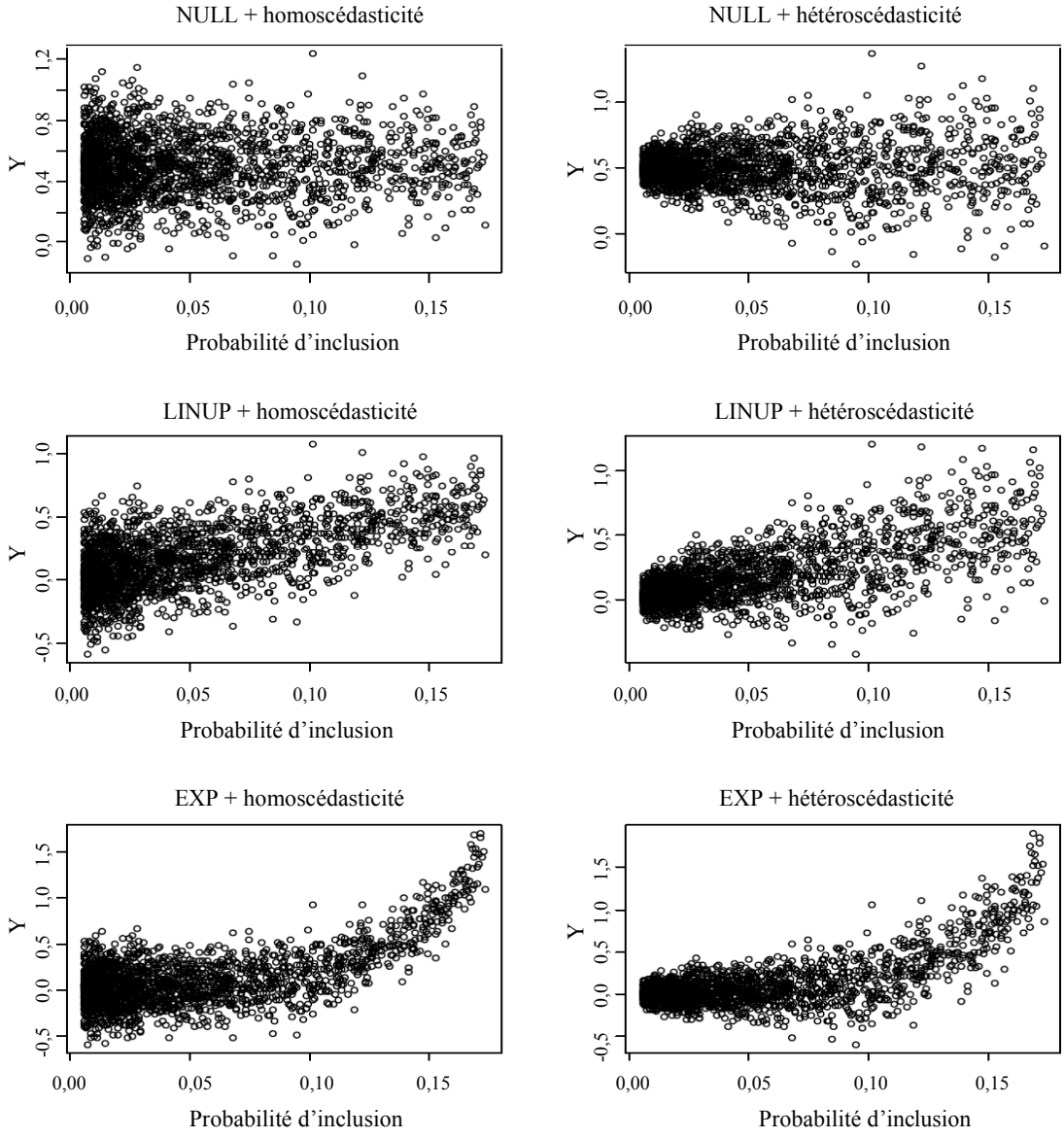


Figure 2 Nuages de points de Y en fonction des probabilités d'inclusion pour les six populations finies artificielles de taille égale à 2 000

Tableau 1
Comparaisons du biais empirique et de la racine carrée de l'erreur quadratique moyenne × 10^3 de $\theta(\alpha)$ pour α = 0,1, 0,25, 0,5, 0,75 et 0,9 : scénarios avec erreurs homoscédastiques

	Biais empirique					REQM empirique				
	0,1	**0,25**	**0,5**	**0,75**	**0,9**	**0,1**	**0,25**	**0,5**	**0,75**	**0,9**
NULL										
FR inverse	-6	-3	-1	-1	-5	46	37	36	37	45
PB2SP	-5	-1	1	2	6	41	33	31	34	42
PS	-5	-3	-1	-4	-6	54	41	39	41	50
PS lisse	-7	-4	-1	-2	-5	50	39	37	38	47
CD	-197	-272	-265	-108	168	203	274	266	115	189
Ratio de RKM	3	25	33	16	6	77	125	159	112	79
Diff de RKM	-5	-1	6	14	14	58	58	94	122	113
LINUP										
FR inverse	-15	-3	-2	-1	-2	70	49	39	34	33
PB2SP	-3	-1	1	4	7	56	43	35	31	29
PS	-15	-3	-3	-2	-6	77	57	48	44	42
PS lisse	-14	-5	-2	-1	-4	72	53	45	42	41
CD	101	35	-37	-49	1	104	38	39	53	31
Ratio de RKM	-23	-9	2	5	-0.2	95	67	53	51	40
Diff de RKM	-15	-4	-4	-0.2	-2	77	55	45	43	38
EXP										
FR inverse	-8	0.4	4	7	4	60	45	41	43	49
PB2SP	-10	-6	-3	0.3	13	52	40	35	36	36
PS	-9	-3	-2	-2	-8	65	49	46	50	72
PS lisse	-12	-5	-2	-1	-2	62	47	43	46	68
CD	92	54	14	19	61	96	57	21	31	75
Ratio de RKM	-17	-11	1	3	-5	87	65	50	53	55
Diff de RKM	-9	-4	-2	-2	-7	65	49	47	47	59

Tableau 2
Comparaisons du biais empirique et de la racine carrée de l'erreur quadratique moyenne × 10^3 de $\theta(\alpha)$ pour α = 0,1, 0,25, 0,5, 0,75 et 0,9 : scénarios avec erreurs hétéroscédastiques

	Biais empirique					REQM empirique				
	0,1	**0,25**	**0,5**	**0,75**	**0,9**	**0,1**	**0,25**	**0,5**	**0,75**	**0,9**
NULL										
FR inverse	-9	-8	-2	4	1	30	24	22	24	31
PB2SP	-6	-6	1	7	7	25	21	19	23	27
PS	-4	-3	-2	-1	-5	34	26	23	26	35
PS lisse	-4	-5	-2	1	-4	34	26	23	26	35
CD	-298	-325	-253	-46	270	302	327	255	60	288
Ratio de RKM	8	31	32	16	5	81	143	154	94	57
Diff de RKM	-5	-1	6	17	16	44	54	87	113	97
LINUP										
FR inverse	-11	-1	5	2	-3	32	24	24	29	35
PB2SP	-10	-1	7	3	1	29	22	22	24	30
PS	-5	-1	-0.1	-1	-4	31	28	33	45	51
PS lisse	-11	-3	2	-0.4	-5	32	26	30	44	50
CD	10	7	6	7	11	20	13	13	20	32
Ratio de RKM	-7	-3	2	3	1	36	29	30	35	41
Diff de RKM	-5	-2	-1	1	-0.2	32	27	28	33	41
EXP										
FR inverse	-8	-3	5	7	-3	30	23	23	30	48
PB2SP	-11	-7	2	6	7	28	23	20	25	36
PS	-3	-3	-2	1	-2	30	26	26	41	84
PS lisse	-8	-5	1	2	-5	30	23	24	39	86
CD	18	16	35	84	68	27	21	38	88	81
Ratio de RKM	-5	-6	-1	2	-0.1	36	31	27	32	62
Diff de RKM	-3	-3	-2	1	-0.1	32	28	28	31	67

Tableau 3
Comparaisons de la largeur moyenne et du taux de non-couverture de l'IC à 95 % × 10^3 de $\theta(\alpha)$ pour α = 0,1, 0,25, 0,5, 0,75 et 0,9

	Largeur moyenne de l'IC à 95 %					Taux de non-couverture de l'IC à 95 %				
	0,1	**0,25**	**0,5**	**0,75**	**0,9**	**0,1**	**0,25**	**0,5**	**0,75**	**0,9**
					Erreurs homoscédastiques					
NULL										
FR inverse	199	156	141	152	184	46	35	44	38	67
PB2SP	178	134	118	134	177	52	55	61	59	50
PS	195	164	151	167	237	112	65	46	40	38
LINUP										
FR inverse	257	207	157	139	141	61	45	37	46	52
PB2SP	230	167	134	123	121	58	54	44	57	59
PS	248	231	188	179	187	119	60	42	41	39
EXP										
FR inverse	234	184	163	177	234	59	44	47	40	42
PB2SP	217	157	132	144	156	54	59	55	53	60
PS	231	199	175	210	402	106	64	47	40	40
					Erreurs hétéroscédastiques					
NULL										
FR inverse	146	104	90	101	137	42	43	38	38	47
PB2SP	107	89	79	89	107	38	49	37	68	65
PS	146	101	91	113	169	80	60	51	37	42
LINUP										
FR inverse	131	107	104	124	154	70	31	36	42	40
PB2SP	125	97	87	93	116	47	35	50	58	52
PS	141	110	133	184	219	138	69	41	50	42
EXP										
FR inverse	131	99	99	134	242	63	49	34	40	41
PB2SP	116	92	84	98	139	57	55	40	63	59
PS	135	100	106	186	378	111	65	46	45	34

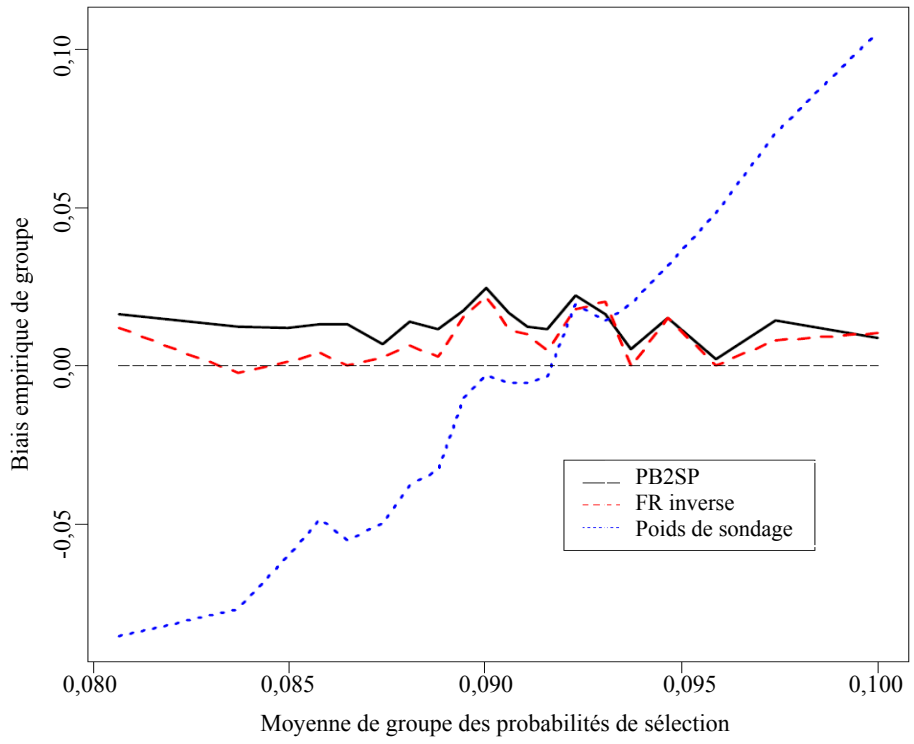


Figure 3 Variation du biais empirique des trois estimateurs pour le 90e centile dans le cas « EXP + homoscédasticité »

Le tableau 4 donne les résultats des simulations. L'approche bayésienne avec fonction de répartition inverse donne en général un biais empirique et une REQM plus petits, et une plus courte largeur moyenne de l'IC à 95 % que l'estimateur pondéré par les poids de sondage. L'IC à 95 % de l'approche bayésienne avec fonction de répartition inverse possède aussi une couverture plus proche du niveau de confiance nominal que l'estimateur pondéré par les poids de sondage quand α est égal à 0,1 et à 0,25. Cependant, dans la queue supérieure avec $\alpha = 0{,}90$, le taux de non-couverture pour l'approche bayésienne avec fonction de répartition inverse est plus élevé que le niveau nominal de 0,05, tandis que l'IC de Woodruff de l'estimateur pondéré par les poids de sondage a de bonnes propriétés. Ces résultats sont en harmonie avec ceux de Sitter et Wu (2001) selon lesquels les intervalles de Woodruff ont de bonnes propriétés même dans les parties moyennes à extrêmes des queues de la distribution. Puisque l'hypothèse de normalité conditionnelle n'est pas raisonnable ici, l'estimateur PB2SP est biaisé et l'IC à 95 % a une mauvaise couverture.

4. Discussion

L'usage des estimateurs des quantiles de population finie pondérés par les poids de sondage est très répandu chez les praticiens des sondages. Bien qu'ils soient faciles à calculer et puissent fournir des inférences valides en grand échantillon, les estimateurs pondérés avec intervalle de confiance de Woodruff peuvent être inefficaces et donner une mauvaise couverture des intervalles de confiance pour les échantillons de taille petite à modérée. Les estimateurs fondés sur un modèle peuvent améliorer l'efficacité des estimations quand le modèle est spécifié correctement, mais produisent des estimations biaisées s'il est mal spécifié. Pour trouver un compromis entre la robustesse et l'efficacité, nous avons considéré des estimateurs fondés sur des modèles avec splines. Pour l'estimation des quantiles d'une variable étudiée continue, nous pouvons estimer des fonctions de répartition fondées sur le modèle puis inverser ces fonctions pour obtenir les quantiles, ou modéliser directement la variable résultat étudié sur les probabilités d'inclusion. Dans le présent article, nous proposons deux estimateurs des quantiles fondés sur un modèle bayésien avec splines. La première méthode est celle de l'estimateur bayésien avec fonction de répartition (FR) inverse, obtenue en inversant les estimations fondées sur un modèle avec splines des fonctions de répartition. La deuxième méthode est celle de l'estimateur PB2SP, estimé en supposant que la variable résultat étudié continue suit une loi normale dont la fonction moyenne et la fonction variance sont toutes deux modélisées au moyen de splines.

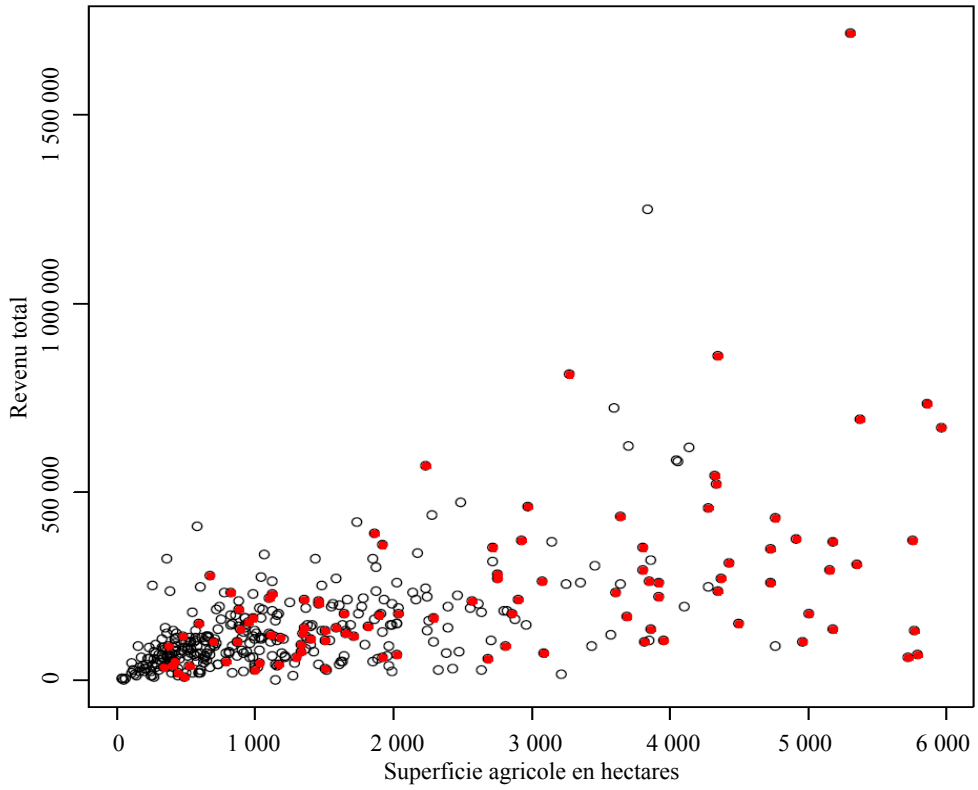


Figure 4 Nuage de points des données sur les exploitations agricoles à grande échelle avec les cercles pleins représentant chacun un échantillon ppt

Tableau 4
Biais empirique × 10^{-2}, racine carrée de l'erreur quadratique moyenne × 10^{-2}, largeur moyenne de l'IC à 95 % × 10^{-2}, et taux de non-couverture de l'IC à 95 % × 10^{3} de $\theta(\alpha)$ pour $\alpha = 0{,}1$, 0,25, 0,5, 0,75 et 0,9 : données sur les exploitations agricoles à grande échelle

	0,1	0,25	0,5	0,75	0,9
			Biais empirique		
FR inverse	8	14	10	-22	-60
PB2SP	-110	-125	-63	-12	88
PS	20	-19	-17	-21	-61
			REQM empirique		
FR inverse	117	117	108	164	256
PB2SP	113	141	124	140	206
PS	132	173	167	226	350
			Largeur moyenne de l'IC à 95 %		
FR inverse	402	443	501	697	906
PB2SP	170	327	539	726	964
PS	285	468	615	864	1 589
			Taux de non-couverture de l'IC à 95 %		
FR inverse	96	53	26	52	90
PB2SP	670	258	42	8	17
PS	220	121	68	42	44

Les simulations donnent à penser que les deux estimateurs fondés sur un modèle bayésien avec splines donnent de meilleurs résultats que l'estimateur pondéré par les poids de sondage, les estimateurs par le ratio et par différence fondés sur le plan de sondage, ainsi que l'estimateur CD fondé sur un modèle lorsque le modèle supposé est incorrect. Les nouvelles méthodes donnent toutes deux des racines de l'erreur quadratique moyenne plus petites qu'il n'y ait pas d'association ou qu'il y ait une association linéaire ou une association non linéaire entre le résultat de l'enquête et la probabilité d'inclusion. Dans certains scénarios, l'accroissement de l'efficacité obtenu en utilisant les deux méthodes bayésiennes est considérable. Lorsque l'hypothèse de normalité du résultat étudié sachant les probabilités d'inclusion est vérifiée, l'estimateur PB2SP produit une REQM plus petite et un intervalle de crédibilité plus court que l'approche avec fonction de répartition inverse. En outre, les deux estimateurs fondés sur un modèle bayésien sont robustes à l'erreur de spécification tant de la fonction moyenne que de la fonction variance. En revanche, l'estimateur fondé sur un modèle CD est biaisé et inefficace quand la fonction moyenne ou la fonction variance est mal spécifiée. Enfin, les méthodes fondées sur un modèle bayésien ont l'avantage de permettre de calculer plus facilement l'IC à 95 % et l'inférence fondée sur les lois a posteriori des paramètres. Cette caractéristique est intéressante, parce que l'estimation de la variance pour les autres estimateurs fondés sur le plan de sondage peut être compliquée. La méthode d'estimation de la variance de Woodruff pour l'estimateur pondéré par les poids de sondage donne de bons résultats quand une fraction importante des données est sélectionnée à partir de la population finie, même dans les parties moyenne à extrême des queues de la fonction de répartition. Cependant, lorsque les données provenant de la population sont peu nombreuses, la méthode de Woodruff a tendance à sous-estimer la couverture de l'intervalle de confiance, alors que les deux méthodes bayésiennes donnent une couverture de ces intervalles plus proche du niveau nominal.

Les trois estimateurs fondés sur le plan de sondage ont un biais empirique global comparable à celui des deux estimateurs fondés sur un modèle bayésien avec splines. Toutefois, la variation du biais de l'estimateur pondéré par les poids de sondage présente une tendance linéaire lorsqu'augmente la moyenne d'échantillon des probabilités d'inclusion. En l'absence d'association entre le résultat étudié et la probabilité d'inclusion, les estimateurs par le ratio et par différence donnent un biais et une REQM relativement plus grands que l'estimateur pondéré par les poids de sondage. Cependant, dans certains scénarios de simulation, les estimateurs par le ratio et par différence produisent une REQM plus petite que l'estimateur pondéré par les poids de sondage. La comparaison entre l'estimateur pondéré par les poids de sondage classique et l'estimateur pondéré par les poids de sondage lisse laisse entendre que l'ajustement d'une courbe cubique lisse pour la fonction de répartition pondérée par les poids de sondage peut améliorer l'efficacité, mais que l'estimateur pondéré par les poids de sondage lisse continuera d'avoir une REQM plus grande que l'estimateur bayésien avec fonction de répartition inverse.

Pour les données dont la distribution est normale, nous recommandons d'utiliser l'estimateur PB2SP de préférence aux autres, en raison du biais plus petit, de la REQM plus petite, et de la meilleure couverture et de la plus courte largeur de l'intervalle de confiance. L'estimateur PB2SP et son intervalle de probabilité a posteriori de 95 % sont faciles à obtenir en utilisant l'algorithme proposé par Crainiceanu et coll. (2007), qui offre aussi l'avantage d'un temps de calcul relativement court.

L'estimateur PB2SP peut être biaisé quand l'hypothèse de normalité conditionnelle ne tient pas. Une option dans ce cas consiste à transformer le résultat étudié afin que l'hypothèse de normalité conditionnelle devienne plus raisonnable. L'estimateur PB2SP peut être appliqué aux données transformées et les tirages à partir des lois a posteriori des unités non échantillonnées sont de nouveau transformés pour revenir à l'échelle originale avant d'estimer les quantiles d'intérêt.

Dans nos simulations avec des données non normales, l'approche bayésienne avec fonction de répartition inverse demeurait plus efficace que l'estimateur pondéré par les poids de sondage. L'amélioration de la couverture de l'intervalle de confiance était limitée aux situations où la taille de l'échantillon est petite, avec une méthode de détermination de l'IC de Wood ruff donnant de bons résultats quand l'hypothèse de grand échantillon est vérifiée. Donc, pour les données ne suivant pas une loi normale pour lesquelles il n'existe aucune transformation évidente en vue d'améliorer la normalité, nous ne recommandons pas l'approche bayésienne avec fonction de répartition inverse quand l'échantillon est de grande taille. Étant donné les bonnes propriétés de l'estimateur PB2SP dans les conditions de normalité, l'extension à examiner lors de futurs travaux consisterait à relâcher l'hypothèse de normalité dans nos approches proposées.

Nous utilisons la probabilité d'inclusion comme variable auxiliaire ici. Lorsqu'il n'existe qu'une seule variable auxiliaire pertinente, peu importe que l'on modélise la probabilité d'inclusion ou la variable auxiliaire. Par contre, s'il existe plus d'une variable auxiliaire pertinente, la probabilité d'inclusion est la variable auxiliaire principale qui doit être modélisée correctement, puisque la spécification incorrecte du modèle reliant le résultat étudié à la probabilité d'inclusion entraîne un biais. Lorsque d'autres variables auxiliaires sont observées pour toutes les unités de la population finie, nos estimateurs bayésiens peuvent tous deux être étendus facilement afin d'inclure les covariables auxiliaires supplémentaires en ajoutant des termes linéaires pour ces variables dans le modèle avec splines pénalisées correspondant.

Un examinateur a proposé une approche pondérée de rechange fondée sur la loi de Dirichlet, qui est facile à calculer, mais n'utilise pas les variables auxiliaires connues dans les unités non échantillonnées. Une autre possibilité consiste à redéfinir l'estimateur CD au moyen du modèle avec splines que nous avons utilisé pour définir l'estimateur PB2SP. Plus précisément, au lieu de supposer que le modèle de régression passe par l'origine, un modèle avec splines est ajusté aux moments d'ordres un et deux de la loi conditionnelle de la variable résultat étudiée sachant la probabilité d'inclusion. L'estimateur CD fondé sur les splines devrait donner des résultats comparables à ceux de l'estimateur PB2SP, et sa variance peut être estimée en utilisant des méthodes de rééchantillonnage.

Dans le contexte de la statistique officielle, les méthodes décrites dans le présent article illustrent les avantages éventuels d'un changement de paradigme pour passer de méthodes fondées sur le plan de sondage à la modélisation bayésienne en vue de produire des inférences ayant de bonnes propriétés fréquentistes. Nos collègues spécialistes de la statistique fondée sur l'échantillonnage probabiliste ont deux grandes objections à ce point de vue.

Premièrement, l'idée d'une approche exagérément fondée sur un modèle – pire encore, bayésienne – des enquêtes probabilistes est mal acceptée, quoique nous mettions ici l'accent sur des méthodes bayésiennes ayant de bonnes propriétés de randomisation. Selon nous, les méthodes probabilistes classiques fondées sur le plan ne fournissent pas l'approche globale nécessaire pour traiter les problèmes complexes qui se posent de plus en plus souvent en statistique officielle. Des choix judicieux de modèles bien calés sont nécessaires pour s'y attaquer. En accordant de l'attention aux caractéristiques du plan de sondage et en choisissant des lois a priori objectives, on peut obtenir des inférences bayésiennes exemptes de subjectivité, et comme les hypothèses de modélisation sont explicites, elles peuvent être critiquées et perfectionnées. Voir Little (2004, 2012) pour une discussion plus approfondie de ces points.

La deuxième objection est que les méthodes bayésiennes requièrent des calculs trop compliqués pour le secteur de la statistique officielle qui doit calculer correctement et produire rapidement un grand nombre de statistiques régulières. Il est vrai qu'à l'heure actuelle, le calcul bayésien peut sembler rébarbatif aux statisticiens habitués à de simples statistiques pondérées et à des méthodes d'estimation de la variance par rééchantillonnage. Dans un article défendant vigoureusement les approches bayésiennes, Sedransk (2008) mentionne que les difficultés pratiques de calcul sont un inhibiteur. Nous convenons que du travail reste à faire pour répondre à cette objection, mais nous ne pensons pas que le problème soit insurmontable. La recherche sur les méthodes de calcul bayésien a connu une véritable explosion ces dernières décennies, tout comme la capacité de calcul. Des modèles bayésiens ont été ajustés pour résoudre des

problèmes de très grande portée et très complexes, dans certains cas nettement plus complexes que ceux qui se posent habituellement dans le secteur de la statistique officielle.

Remerciements

Nous remercions M. Philip Kokic de la Commonwealth Scientific and Industrial Research Organisation, de nous avoir fourni les données sur les exploitations agricoles à grande échelle (*broadacre farms*). Nous remercions aussi un rédacteur associé et les examinateurs de leurs commentaires constructifs au sujet de la version originale du présent article.

Bibliographie

ABARE (2003). Australian farm surveys report 2003. Canberra.

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of American Statistical Association*, 88, 268-277.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2010). Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 36, 1, 25-37.

Crainiceanu, C.M., Ruppert, D., Carroll, R.J., Joshi, A. et Goodner, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic error. *Journal of Computational and Graphical Statistics*, 16, 265-288.

Dorfman, H., et Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1474.

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

Harms, T., et Duchesne, P. (2006). De l'estimation des quantiles par calage. *Techniques d'enquête*, 32, 1, 41-57.

Kuk, A.Y.C. (1993). A kernel method for estimating finite population functions using auxiliary information. *Biometrika*, 80, 385-392.

Kuk, A.Y.C., et Welsh, A.H. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society*, Séries B, 63, 277-292.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, DOI: 10.1198/016214504000000467. {70}, 99, 546-556.

Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (avec discussion et réplique). *Journal of Official Statistics*, 28, 309-334.

Lombardía, M.J., González-Manteiga, W. et Prada-Sánchez, J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference*, 116, 367-388.

Lombardía, M.J., González-Manteiga, W. et Prada-Sánchez, J.M. (2004). Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimate of a finite population distribution function. *Journal of Nonparametric Statistics*, 16, 63-90.

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

Royall, R.M., et Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Ruppert, D., Wand, M. P. et Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge, Royaume-Uni : Cambridge University Press.

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24, 495-506.

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

Wang, S., et Dorfman, A.H. (1996). A new estimator for the finite population distribution function. *Biometrika*, 83, 639-652.

Wood, S.N. (1994). Monotonic smoothing splines fitted by cross validation SIAM. *Journal on Scientific Computing*, 15, 1126-1133.

Woodruff, R. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

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

Zheng, H., et Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zheng, H., et Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Techniques d'enquête, décembre 2012
Vol. 38, N° 2, pp. 235-238
Statistique Canada, N° 12-001-X au catalogue

Imputation multiple dans le cas de données de recensement

Satkartar K. Kinney [1]

Résumé

L'un des avantages de l'imputation multiple est qu'elle permet aux utilisateurs des données de faire des inférences valides en appliquant des méthodes classiques avec des règles de combinaison simples. Toutefois, les règles de combinaison établies pour les tests d'hypothèse multivariés échouent quand l'erreur d'échantillonnage est nulle. Le présent article propose des tests modifiés utilisables dans les analyses en population finie de données de recensement comportant de multiples imputations pour contrôler la divulgation et remplacer des données manquantes, et donne une évaluation de leurs propriétés fréquentistes par simulation.

Mots clés : Populations finies ; données manquantes ; test de signification ; données synthétiques.

1. Introduction

L'imputation multiple a été proposée au départ pour traiter la non-réponse dans les grandes enquêtes complexes (Rubin 1987). Depuis, plusieurs autres usages ont été suggérés, dont le contrôle de la divulgation statistique et la correction de l'erreur de mesure. L'un des attraits de l'imputation multiple tient au fait que l'on peut appliquer des méthodes classiques à chaque ensemble de données imputé, puis utiliser de simples règles de combinaison, qui varient selon l'application. Voir Reiter et Raghunathan (2007) pour une revue détaillée des différentes règles et applications. Les règles existantes de combinaison sous imputation multiple ont été établies pour des échantillons aléatoires et des modèles de superpopulation (Deming et Stephan 1941). Dans les analyses de données de recensement en population finie, où la variance d'échantillonnage est nulle, les règles de combinaison applicables aux paramètres à estimer univariés peuvent encore l'être en tant que cas particulier ; par contre, les tests d'hypothèse échouent pour les paramètres multivariés.

Motivé par l'utilisation de l'imputation multiple pour produire des données partiellement synthétiques (Rubin 1993 ; Little 1993) pour la base de données longitudinales sur les entreprises du U.S. Census Bureau (Kinney, Reiter, Reznek, Miranda, Jarmin et Abowd 2011), c'est-à-dire un recensement économique, le présent article décrit l'élaboration d'un test multivarié pour populations finies applicable à des données partiellement synthétiques et son extension à l'imputation de données manquantes. Les extensions à d'autres applications d'imputation multiple devraient être simples.

La présentation de l'article est la suivante. La section 2 décrit le cas de données partiellement synthétiques et la section 3, l'extension aux données manquantes. Enfin, la section 4 décrit les simulations en vue d'évaluer les règles de combinaison pour le cas des données manquantes ainsi que celui des données partiellement synthétiques.

2. Données partiellement synthétiques

Pour créer des ensembles de données partiellement synthétiques, on remplace certaines valeurs des données confidentielles par m tirages indépendants à partir de leur loi prédictive a posteriori. Pour une population finie de taille N, soit $Z_j = 1$, $j = 1, ..., N$ indiquant que l'unité j a été sélectionnée pour le remplacement par imputation de n'importe laquelle de ses valeurs observées. Les imputations ne devraient être effectuées qu'à partir de la loi prédictive a posteriori de ces unités avec $Z_j = 1$. Pour simplifier, dans le présent article, nous supposons que $Z_j = 1$, $j = 1, ..., N$. Soit $Y = (y_1, ..., y_d)$ la matrice de variables confidentielles dont les valeurs seront remplacées par des imputations et X la matrice de variables dont les valeurs ne seront pas remplacées. Représentons par $D_{\text{rec}} = (X, Y)$ un recensement des N unités contenant des données confidentielles et supposons que toutes les unités sont entièrement observées, c'est-à-dire qu'il n'existe aucune valeur manquante. Soit $Y_{\text{rep}}^{(i)}, i = 1, ..., m$ la i^{e} imputation de Y, et soit $D_{\text{syn}}^{(i)} = (X, Y_{\text{rep}}^{(i)})$. L'ensemble $D_{\text{syn}} = \{D_{\text{syn}}^{(i)}, i = 1, ..., m\}$ est celui qui est diffusé aux membres du public.

Toute procédure d'imputation appropriée extraite de l'abondante littérature sur l'imputation multiple peut être utilisée pour générer D_{syn} à partir de D_{rec}. Les méthodes pour population finie proposées ici peuvent être appliquées que l'on ait ou non supposé qu'une population finie a été utilisée pour générer D_{syn}. Sous une hypothèse de population finie, puisque les données sont entièrement observées (recensement), les paramètres du modèle d'imputation seraient considérés comme étant connus et fixes. Voir Reiter et Kinney (2012) pour une illustration de la façon d'obtenir des inférences valides à partir d'échantillons aléatoires

1. Satkartar K. Kinney, National Institute of Statistical Sciences, Research Triangle Park, NC 27709, États-Unis. Courriel : saki@niss.org.

partiellement synthétiques générés à l'aide de paramètres de modèle d'imputation fixes ainsi qu'aléatoires. Les simulations (non présentées) confirment qu'il en est de même dans le cas d'une population finie.

Un analyste qui a accès à D_{syn} mais non à D_{rec} peut obtenir des inférences valides pour une grandeur scalaire ou vectorielle Q en utilisant les quantités suivantes :

$$\bar{Q}_m = \frac{1}{m}\sum_{i=1}^{m} Q^{(i)} \tag{2.1}$$

$$\bar{U}_m = \frac{1}{m}\sum_{i=1}^{m} U^{(i)} \tag{2.2}$$

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m} (Q^{(i)} - \bar{Q}_m)\,(Q^{(i)} - \bar{Q}_m)' \tag{2.3}$$

où $Q^{(i)}$, $i = 1, ..., m$ est l'estimation ponctuelle de Q obtenue à partir de $D_{\text{syn}}^{(i)}$, $U^{(i)}$ est la variance estimée de Q, et B_m est la variance d'échantillon des $Q^{(i)}, i = 1, ..., m$.

En l'absence de variance d'échantillonnage, les règles de combinaison pour la grandeur scalaire Q établie par Reiter (2003) peuvent être appliquées comme cas particulier où $\bar{U}_m = 0$. La simplification résultante signifie que les approximations de Reiter (2003) ne sont pas nécessaires et que la loi a posteriori exacte sous la théorie normale multivariée est $(Q \mid D_{\text{syn}}) \sim t_{m-1}(\bar{Q}_m, B_m / m)$. Cependant, pour un vecteur Q, le test d'hypothèse de Reiter (2005) dépend de l'hypothèse que B_∞ est proportionnelle à $\bar{U}_\infty$, c'est-à-dire que la proportion de l'information remplacée par des imputations est la même pour toutes les composantes de Q, de sorte qu'une hypothèse différente est nécessaire pour le cas où $\bar{U}_\infty = 0$.

2.1 Test multivarié proposé

À la présente section, un test de remplacement est établi en se fondant sur l'hypothèse plus forte que $B_\infty = r_\infty I$, pour une quantité scalaire r_∞ et une matrice identité I de dimension k. Autrement dit, la variance entre imputations est constante pour toutes les composantes de Q, et B_∞ est supposée diagonale. Tant dans le test de Reiter (2005) que dans le test proposé, on calcule la moyenne sur l'ensemble des composantes de la variance, de sorte que le test est moyennement robuste à cette hypothèse ; cependant, la validité de la randomisation diminue quand les estimations de Q, $\bar{Q}^{(i)}$, $i = 1, ..., m$, sont fortement corrélées. Cet aspect est évalué à l'aide de simulations à la section 4.3. Des tests comparables fondés sur l'hypothèse que $B_\infty \propto \bar{U}_\infty$ perdent, on le sait, de la puissance quand l'hypothèse n'est pas satisfaite (Li et coll. 1991).

Le test proposé pour l'hypothèse $H_0 : Q = Q_0$ est effectué en supposant que la statistique de test

$$S_c = \frac{(Q_0 - \bar{Q}_m)'(Q_0 - \bar{Q}_m)}{k r_c}$$

suit une loi $F_{k,k(m-1)}$, où $r_c = 1 / m \;\; tr(B_m) / k$.

Sous l'hypothèse que $B_\infty = r_\infty I$, la valeur p bayésienne est donnée par

$$\begin{aligned}\int P(\chi_k^2 > (Q_0 - \bar{Q})' T_\infty^{-1} (Q_0 - \bar{Q}) \mid D_{\text{syn}}, B_\infty) \\ P(B_\infty \mid D_{\text{syn}})\, dB_\infty \quad (2.4)\end{aligned}$$

$$\begin{aligned}&= \int P\left(\chi_k^2 > \frac{(Q_0 - \bar{Q})' I (Q_0 - \bar{Q})}{r_\infty / m} \mid D_{\text{syn}}, r_\infty\right) \\ &\qquad P(r_\infty \mid D_{\text{syn}})\, dr_\infty \\ &= \int P\left(\frac{\chi_k^2}{k} \cdot \frac{r_\infty}{m r_c} > S_c \mid D_{\text{syn}}, r_\infty\right) \\ &\qquad P(r_\infty \mid D_{\text{syn}})\, dr_\infty. \quad (2.5)\end{aligned}$$

Donc, l'hypothèse de proportionnalité réduit le nombre de paramètres de variance à estimer qui passe de $k(k-1)/2$ à 1, et permet une approximation analytique de l'intégrale en (2.4). Comme $\bar{U}_\infty = 0$, les calculs sont plus simples que dans Reiter (2005). Pour achever l'intégration, nous avons besoin de la distribution de $(r_\infty \mid D_{\text{syn}})$. En étendant le cas scalaire donné dans Reiter (2003), la distribution d'échantillonnage de $Q^{(i)}$, l'estimation de Q obtenue à partir de $D_{\text{syn}}^{(i)}$, est donnée par $(Q^{(i)} \mid Q_{\text{rec}}, B_\infty) \sim N(Q_{\text{rec}}, B_\infty)$. Sous l'hypothèse de proportionnalité, l'expression devient $(Q^{(i)} \mid Q_{\text{rec}}, r_\infty) \sim N(Q_{\text{rec}}, r_\infty I)$. En utilisant des lois a priori diffuses et en appliquant la théorie normale multivariée classique pour les matrices de covariance d'échantillon, nous obtenons

$$(m-1)\frac{\sum_{i=1}^{m}(Q^{(i)}-\bar{Q}_m)(Q^{(i)}-\bar{Q}_m)'}{(m-1)r_\infty} \mid D_{\text{syn}} \sim \text{Wish}(m-1, I).$$

En prenant la trace de chaque membre de l'équation et en intégrant sur r_∞ dans (2.5), nous obtenons une valeur p bayésienne de

$$P\left(\frac{\chi_k^2}{k}\frac{k(m-1)}{\chi_{k(m-1)}^2} > S_c \mid D_{\text{syn}}\right) = P(F_{k,k(m-1)} > S_c \mid D_{\text{syn}}).$$

3. Données manquantes

L'extension au cas des données manquantes est simple. Quand $\bar{U}_\infty = 0$, les règles de combinaison (Rubin 1987) pour les paramètres à estimer scalaires q se simplifient de

façon que $(q \mid D_{\text{com}}) \sim N(\bar{q}_m, (1+1/m)B_m)$, où D_{com} est le jeu de m ensembles de données complétés. Comme à la section 2, les tests de Rubin (1987) et de Li, Raghunathan et Rubin (1991) pour les composantes multivariées dépendent de l'hypothèse que $B_\infty \propto \bar{U}_\infty$, et donc, quand $\bar{U}_\infty = 0$, nous déterminons un test sous l'hypothèse que $B_\infty = r_\infty I$.

En suivant des méthodes de calcul semblables à celles de la section 2.1, la valeur p bayésienne obtenue pour tester l'hypothèse $H: Q = Q_0$ avec Q de dimension k est $P(F_{k,k(m-1)} > S_q \mid D_{\text{com}})$, où

$$S_q = \frac{(Q_0 - \bar{Q}_m)'(Q_0 - \bar{Q}_m)}{kr_q},$$

et $r_q = (1+1/m)\, tr(B_m)/k$.

4. Étude par simulation

À la présente section, des exemples de simulations simples illustrent la validité analytique des règles de combinaison proposées, d'abord pour le cas de données partiellement synthétiques, puis pour celui de données manquantes. Enfin, la robustesse des tests à l'hypothèse de proportionnalité est évaluée.

Pour une population de $N = 50\,000$, $X = (X_1, ..., X_{20})$ est tiré d'une loi normale multivariée de moyenne nulle et de matrice de covariance où chaque élément diagonal est égal à 1 et chaque élément hors diagonale est égal à 0,5. Y est tiré d'une loi normale centrée réduite. Pour chacune des 5 000 itérations, une nouvelle population finie est générée et m imputations sont tirées pour $m \in \{2, 5, 10\}$. Les tests d'hypothèse proposés sont effectués pour $H_0: Q = Q_0$, où Q est le vecteur des coefficients de régression, à l'exclusion de l'ordonnée à l'origine, de la régression de Y sur X, et est de dimension k, $k \in \{2, 5, 20\}$, et Q_0 est la valeur réelle de Q déterminée à partir de la population finie (X, Y). Puisque l'hypothèse nulle H_0 est vraie par conception, elle devrait être rejetée dans 100α % des cas, pour un seuil de signification de $\alpha = 0{,}05$.

Des scénarios d'échantillonnage aléatoire sont également simulés aux fins de comparaison. À chaque itération, un échantillon aléatoire de taille $s = 50\,000$ d'une population infinie est généré à partir de la distribution décrite plus haut, avant de générer les m données manquantes et imputations synthétiques. La même hypothèse $H_0: Q = Q_0$ est testée, où Q_0 est le vecteur de valeurs de population réelles. Les règles de combinaison pour les tests d'hypothèse sont celles de Reiter (2005) dans le cas des données synthétiques et celles de Li et coll. (1991) et de Rubin (1987) dans le cas des données manquantes.

4.1 Imputation de données partiellement synthétiques

Soit Y la variable réponse confidentielle et X, les variables explicatives non remplacées. Alors, Y_{syn} est générée en effectuant m tirages indépendants à partir de la loi prédictive a posteriori $f(Y \mid X)$ sous l'hypothèse d'un modèle linéaire normal, en utilisant toutes les données disponibles.

Le tableau 1 donne les taux de rejet au niveau nominal de 5 % pour le test d'hypothèse proposé pour des quantités à estimer à composantes multiples, et montre qu'ils sont proches du seuil de signification de 0,05, ainsi que de ceux obtenus sous échantillonnage aléatoire. Ces résultats semblent indiquer que les règles de combinaison proposées pour les données de population ont de bonnes propriétés fréquentistes. Les taux de rejet obtenus en appliquant les règles établies pour les échantillons aléatoires (Reiter 2005) à des populations finies, lesquels étaient relativement élevés, habituellement égaux à 1, dans les simulations exécutées ne sont pas présentés.

Tableau 1
Comparaison des taux de rejet au niveau nominal de 5 % pour les tests sur des données partiellement synthétiques

	$k = 2$	$k = 5$	$k = 20$
Données de recensement			
$m = 2$	0,048	0,065	0,052
$m = 5$	0,048	0,061	0,057
$m = 10$	0,051	0,067	0,055
Échantillonnage aléatoire			
$m = 2$	0,067	0,062	0,060
$m = 5$	0,054	0,052	0,050
$m = 10$	0,047	0,049	0,049

4.2 Données manquantes

Des simulations analogues à celles exécutées sur les données synthétiques ont été effectuées dans le cas des données manquantes. Les valeurs manquantes de Y ont été imputées à partir de la loi prédictive a posteriori $f(Y_{\text{obs}} \mid X)$ sous l'hypothèse d'un modèle linéaire normal. Les données manquantes ont été simulées comme si elles manquaient entièrement au hasard, avec $P(R_l = 1) = 0{,}3$, $l = 1, ..., s$, où R est une variable indicatrice de l'absence de données.

Le tableau 2 donne les taux de rejet au niveau nominal de 5 % pour le test d'hypothèse proposé pour des grandeurs à estimer à composantes multiples, et montre qu'ils sont proches de 0,05, ainsi que des valeurs obtenues sous échantillonnage aléatoire. Ces résultats semblent indiquer que les règles de combinaison proposées pour des données de population donnent des inférences valides.

Tableau 2
Comparaison des taux de rejet au niveau nominal de 5 % pour les tests sur des données de recensement complétées

	$k = 2$	$k = 5$	$k = 20$
Données de recensement			
$m = 2$	0,052	0,061	0,053
$m = 5$	0,048	0,063	0,051
$m = 10$	0,048	0,058	0,054
Échantillonnage aléatoire			
$m = 2$	0,061	0,056	0,053
$m = 5$	0,056	0,052	0,052
$m = 10$	0,048	0,050	0,051

4.3 Robustesse

L'hypothèse que $B_\infty \propto r_\infty I$ est étonnante à première vue, et il est peu probable qu'elle soit exactement vraie. À la présente section, nous évaluons l'effet de fortes corrélations entre les composantes de Q. Même si des corrélations moyennement fortes étaient présentes dans les simulations précédentes, ici, nous augmentons la grandeur de la variance entre imputations, ce qui accroît les écarts le long de la diagonale de B ainsi que la distance par rapport à zéro des éléments hors diagonale de B.

Les simulations sont configurées comme précédemment, pour le cas d'une population finie, avec $k = 5$ et $m = 5$. À chaque itération, la population est générée de la même façon qu'auparavant, excepté que nous prenons $Y = (1, 2, 5, 10, 20, 0, ..., 0)\,(X_1, X_2, ..., X_{20})' + \eta$, $\eta \sim N(0, 100)$ et $X_2 = c \cdot X_1 + \varepsilon, c \in \{1/2, 1, 5\}$ et $\varepsilon \sim N(0, 1)$. Des valeurs croissantes de c donnent des corrélations de plus en plus fortes. La grande variance de η induit des valeurs plus grandes et plus variables des éléments de B.

Le tableau 3 montre que, bien que les tests aient de bonnes propriétés même sous des violations moyennement fortes de l'hypothèse de proportionnalité, leur performance diminue à mesure qu'augmente la force des corrélations. En maintenant l'hypothèse que Q représente un vecteur de coefficients de régression, l'existence d'une corrélation aussi forte peut également être un signe de multicollinéarité dans le modèle utilisé, de sorte que les analystes en présence d'une forte corrélation entre les $\overline{Q}^{(i)}$ pourraient vouloir prendre des mesures en vue de réduire la multicollinéarité avant d'appliquer les tests proposés. Si les variables sont de grandeur très différente, une normalisation en vue de les rééchelonner réduira les écarts entre les Q.

Tableau 3
Évaluation des tests sous violations d'hypothèse, $k = 5$, $m = 5$

	$c = 1/2$	$c = 1$	$c = 5$
Données synthétiques	0,059	0,083	0,145
Données manquantes	0,051	0,083	0,136

Remerciements

Une partie des travaux susmentionnés, financée par la bourse ITR-0427889 de la NSF, a été effectué pendant que l'auteur était étudiant à la Duke University, sous la supervision de Jerry Reiter, dont l'aide a été fort appréciée. En outre, les commentaires d'examinateurs anonymes ont été très utiles.

Bibliographie

Deming, W.E., et Stephan, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 213, 45-49.

Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S. et Abowd, J.M. (2011). Toward unrestricted public-use business microdata: The Longitudinal Business Database. *Revue Internationale de Statistique*, 79, 3, 362-384.

Li, K.H., Raghunathan, T.E. et Rubin, D.B. (1991). Large-sample significance levels from multiply-imputed data using moment-based statistics and an *F* reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Reiter, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 2, 203-211.

Reiter, J.P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131, 365-377.

Reiter, J.P., et Kinney, S.K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. Rapport technique, National Institute of Statistical Sciences.

Reiter, J.P., et Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.

AVERTISSEMENT

Statistique Canada cessera de publier une version imprimée de la revue *Techniques d'enquête.* Ce plus récent numéro (décembre 2012 – volume 38, numéro 2) sera le dernier disponible en version imprimée. Veuillez noter que la version électronique de *Techniques d'enquête* demeurera disponible gratuitement sur le site internet de Statistique Canada, www.statcan.gc.ca.

Notre prochain numéro sera diffusé en juin 2013 en version électronique selon nos mêmes normes rigoureuses quant au contenu.

Vous pouvez vous inscrire sous « Mon compte » sur le site internet de Statistique Canada pour recevoir un avis par courriel lors de la publication des prochains numéros de la revue.

CORRIGENDUM

James Chipperfield et John Preston,
« Bootstrap efficace pour les enquêtes-entreprises », vol. 33, n^{o} 2 (Décembre 2007), 187-193.

À la section 4.2 de cet article, sous l'équation

$$\mathrm{Var}\left(\hat{v}_{\mathrm{boot}}\right) = \mathrm{Var}_{s}\left(E_{*}\left[\hat{v}_{\mathrm{boot}}\middle|s\right]\right) + E_{s}\left(\mathrm{Var}_{*}\left[\hat{v}_{\mathrm{boot}}\middle|s\right]\right),$$

l'exposé contient cinq mentions du terme

$$\mathrm{Var}_{s}\left(E_{*}\left[\hat{v}_{\mathrm{boot}}\middle|s\right]\right).$$

Pour que l'exposé soit correct, ces cinq mentions doivent être remplacées par

$$E_{s}\left(\mathrm{Var}_{*}\left[\hat{v}_{\mathrm{boot}}\middle|s\right]\right).$$

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 2012.

S.R. Amer, *RTI International*
T. Asparouhov, *Mplus*
M. Barron, *NORC*
W. Bell, *U.S. Census Bureau*
E. Berg, *National Agricultural Statistical Services*
P. Biemer, *RTI*
I. Bilgen, *NORC*
C. Bocci, *Statistics Canada*
J. van den Brakel, *Statistics Netherlands*
M. Brick, *Westat, Inc.*
R. Bruni, *University of Rome, La Sapienza*
C.-T. Chao, *National Cheng-Kung University, Taiwan*
G. Chauvet, *CREST-ENSAI*
J. Chipperfield, *Australian Bureau of Statistics*
G. Datta, *University of Georgia*
M. Davern, *NORC*
T. DeWaal, *Statistics Netherlands*
D. Dolson, *Statistics Canada*
S. Eckman, *Institute for Employment Research, Germany*
S. Er, *Instanbul University*
E. Escobar, *University of Southampton*
V. Estevao, *Statistics Canada*
O.P. Fischer, *U.S. Census Bureau*
J. Gambino, *Statistics Canada*
N. Ganesh, *NORC at University of Chicago*
T.I. Garner, *U.S. Bureau of Labor Statistics*
J. Garrett, *Knowledge Networks, Inc.*
C. Goga, *Université de Bourgogne*
M. Graf, *Office fédéral de la Statistique, Suisse*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
D. Kasprzyk, *NORC at the University of Chicago*
C. Kennedy, *Abt SRBI*
M.G.M. Khan, *University of the South Pacific, Fiji*
J.-K. Kim, *Iowa State University*
P. Kott, *RTI*
P. Lavallée, *Statistics Canada*
F. Li, *Duke University*
J. Li, *Westat Inc.*
P. Lugtig, *Utrecht University*
P. Lynn, *University of Essex*
D. Malec, *National Center for Health Statistics*
H. Mantel, *Statistics Canada*
I. Molina, *Universidad Carlos III de Madrid*
R. Münnich, *Economic and Social Statistics Dept. Univ. of Trier, Germany*
J. Oleson, *University of Iowa*
A.J. O'Malley, *Harvard Medical School*
J. Opsomer, *Colorado State University*
V. Parsons, *National Center for Health Statistics*
D. Pfeffermann, *Hebrew University*
F. van de Pol, *Statistics Netherlands*
N.G.N. Prasad, *University of Alberta*
L. Qualité, *Université de Neuchâtel*
T. Raghunathan, *University of Michigan*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
R. Rodriguez, *U.S. Census Bureau*
K. Rust, *Westat, Inc.*
E. Saleh, *Carleton University*
F. Scheuren, *NORC*
A. Scott, *University of Auckland*
J. Sedransk, *Case Western Reserve University & University of Maryland*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
R. Sigman, *Westat Inc.*
A. Singh, *NORC*
C. Skinner, *London School of Economics*
P.A. Smith, *Office for National Statistics*
P.W.F. Smith, *University of Southampton*
N. Thomas, *Pfizer*
R. Thomas, *Carleton University*
K.J. Thompson, *U.S. Census Bureau*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
V. Toepoel, *Tilburg University*
M. Torabi, *University of Manitoba*
V. Vehovar, *University of Ljubljana*
J. Vermunt, *Tilburg School of Social and Behavioral Sciences*
M. de Toledo Vieira, *Universidade Federal de Juiz de Fora, Brazil*
J. Wagner, *University of Michigan*
K. Wolter, *NORC*
C. Wu, *University of Waterloo*
C. Yu, *Iowa State University*
W. Yung, *Statistique Canada*
E. Zanutto, *National Analysts Worldwide*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2012 : Céline Ethier de la Division de la recherche et de l'innovation en statistique, Christine Cousineau de la Division des méthodes d'enquêtes auprès des ménages, Nick Budko et Annette Everett de la Division des méthodes d'enquêtes auprès des entreprises, Anne-Marie Fleury de la Division des opérations et de l'intégration, Roberto Guido, Liliane Lanoie, Darquise Pellerin, Joseph Prince, Jacqueline Luffman, Suzanne Bélair, Janice Burtt, Jeff Campbell, Kathy Charbonneau et Fadi Salibi de la Division de la diffusion.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

ANNONCES

Demande de candidatures pour le prix Waksberg 2014

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 2014 au Symposium de Statistique Canada qui se tiendra à l'automne de 2014. L'article paraîtra dans un numéro de *Techniques d'enquête* (publication prévue pour décembre 2014).

L'auteur de l'article Waksberg 2014 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 2013 au président du comité, Steve Heeringa (sheering@isr.umich.edu).

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**, Sujet de l'article à l'étude.

Membres du comité de sélection de l'article Waksberg (2012-2013)

Steve Heeringa, *University of Michigan* (Président)
Cynthia Clark*, USDA*
Louis-Paul Rivest, *Université de Laval*
J.N.K. Rao, *Carleton University*

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)

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 28, No. 2, 2012

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 28, No. 3, 2012

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 40, No. 2, June/juin 2012

CONTENTS TABLE DES MATIÈRES

Volume 40, No. 3, September/septembre 2012

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de finaliser votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) *etc*.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0 ; l, 1).

3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, *etc.* à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

6. Communications brèves

6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.