Nº 12-001-XIF au catalogue

Techniques d'enquête

2005

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1 800 263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements sur les commandes et les abonnements

Le produit nº 12-001-XIF au catalogue est publié deux fois par année sous format électronique au prix de 23 $CAN l'exemplaire et de 44 $CAN pour un abonnement annuel. Pour obtenir un exemplaire ou s'abonner, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Ce produit nº 12-001-XPB au catalogue 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 suivant s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	15 $CAN	30 $CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par

- Téléphone (Canada et États-Unis) 1 800 267-6677
- Télécopieur (Canada et États-Unis) 1 877 287-4369
- Courriel infostats@statcan.ca
- Poste Statistique Canada
  Division des finances
  Immeuble R.-H. Coats, 6e étage
  120, avenue Parkdale
  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, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. 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 dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada
Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

2005

Publication autorisée par le ministre responsable de Statistique Canada

Juillet 2005

Nº 12-001-XIF au catalogue, vol. 31, nº 1
ISSN 1712-5685

Nº 12-001-XPB au catalogue, vol. 31, nº 1
ISSN 0714-0045

Périodicité : semestriel

Ottawa

This publication is available in English upon request (Catalogue no. 12-001-XIE).

Note de reconnaissance

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

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée 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, Dr. M.P. Singh, singhmp@statcan.ca (Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 $ CA par année. Le prix n'inclus pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 $ CA (6 $ × 2 exemplaires); autres pays, 30 $ CA (15 $ × 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.

À Gordon J. Brackstone


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 31, numéro 1, juin 2005

Table des matières


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, Juin 2005
Vol. 31, N° 1, pp. 1-2
Statistique Canada, N° 12-001-XIF au catalogue

Dans ce numéro

Ce numéro de *Techniques d'enquête* est dédié à Gordon J. Brackstone qui a récemment pris sa retraite de Statistique Canada. Il était Statisticien en chef adjoint du Secteur de l'Informatique et la Méthodologie et a été président du Comité de direction de *Technique d'enquête* à partir de 1987. Son support continu à la revue a toujours été empreint de discernement et visiblement motivé par un désir constant de stimuler la poursuite de standards élevés de pratiques méthodologiques. De plus, il a lui-même produit des articles pour la revue. Nous sommes vraiment reconnaissants envers Gordon J. Brackstone.

Le présent numéro contient huit articles ordinaires traitant de divers sujets et trois communications brèves. Comme nous l'avons mentionné dans le dernier numéro de la revue, nous lançons une nouvelle section qui sera réservée à des communications brèves. Elle contiendra des articles courts, habituellement de quatre pages environ. Ces communications brèves pourraient être consacrées à la présentation de nouvelles idées sans les traiter complètement comme dans un article ordinaire, à des rapports brefs sur des travaux empiriques ou à des discussions ou des compléments d'autres articles publiés dans la revue.

Depuis quatre ans, le numéro de juin de *Techniques d'enquête* contient un article invité en l'honneur de Joseph Waksberg. À partir de cette année, cet article invité sera publié dans le numéro de décembre de la revue afin de mieux le sychroniser avec la présentation correspondante qui est faite au symposium annuel de Statistique Canada sur la méthodologie à l'automne. L'auteur Waksberg de cette année est J.N.K. Rao et son article sera sur « l'interaction entre la théorie et les méthodes d'échantillonnage : Une évaluation ».

Dans l'article d'ouverture du présent numéro, Winglee, Valliant et Scheuren présentent une nouvelle approche de simulation pour estimer les taux d'erreurs pour la sélection des seuils lors des couplages d'enregistrements. Pour chaque paire susceptible d'être un appariement vrai, il existe un vecteur de résultats des comparaisons qui détermine le poids d'appariement. Les auteurs supposent que chaque résultat de comparaison correspond à un modèle multinomial, et que la loi multinomiale diffère pour les appariements vrais et les non-appariements. Ils estiment les lois d'après un échantillon, puis les utilisent pour simuler les lois de probabilité des poids appariés pour les appariements vrais et les non-appariements. Ils illustrent la méthode au moyen d'une étude de cas en se servant de données provenant de la Medical Expenditure Panel Survey (MEPS) réalisée aux États-Unis.

Krewski, Dewanji, Wang, Bartlett, Zielinski et Mallick étudient les effets des erreurs de couplage d'enregistrements, aussi bien les résultats faussement positifs que faussement négatifs, sur les estimations du risque dans les études de cohorte. Ils montrent analytiquement comment les erreurs de couplage introduisent un biais et une variabilité supplémentaire dans les nombres observés et attendus de décès, ainsi que dans les estimations des ratios standardisés de mortalité et des coefficients de régression du risque relatif. Ils discutent des résultats dans leurs conclusions et soulignent les travaux qui devraient être réalisés dans ce domaine.

L'article rédigé par van den Brakel et Renssen traite du problème de la vérification d'hypothèses sous différentes mises en œuvre de l'enquête, comme des conceptions différentes du questionnaire, lorsqu'on utilise un plan d'échantillonnage complexe. Ils élaborent une théorie fondée sur le plan de sondage pour les cas où les diverses mises en œuvre de l'enquête sont affectées à des sous-échantillons au moyen de plans d'expérience en randomisation totale ou en blocs randomisés. La théorie s'appuie aussi sur des modèles de l'erreur de mesure. Les auteurs utilisent la statistique de Wald fondée sur le plan de sondage pour comparer les diverses mises en œuvre de l'enquête.

Tsuchiya aborde d'une manière intéressante l'ancien problème que soulèvent les questions délicates posées lors des enquêtes. Au lieu d'utiliser la technique de réponse aléatoire qui offre peu de contrôle au chercheur, il propose d'adapter la technique du dénombrement d'items au cas des questions délicates. La technique de dénombrement d'items consiste à présenter au répondant une liste de plusieurs phrases et de lui demander de choisir toutes celles qui s'appliquent à lui. Le chercheur construit la liste de deux façons : la première contient la phrase délicate, tandis que la deuxième ne la contient pas. Tsuchiya présente divers estimateurs pour cette technique et donne un exemple intéressant ayant trait au caractère national japonais.

Dans l'article de DiZio, Guarnera et Luzi, des modèles de mélanges finis sont utilisés pour déceler les erreurs dues à l'utilisation d'une unité incorrecte de mesure au stade de la collecte des données d'enquête. Dans un contexte multivarié et en supposant que les données suivent une loi normale multivariée, la méthode permet de déterminer quelles variables sont erronées pour une unité échantillonnée particulière. Les auteurs fournissent aussi des diagnostics pour établir l'ordre de priorité des cas qui doivent faire l'objet d'un examen manuel plus approfondi. La méthode proposée est illustrée au moyen d'un exemple portant sur des données simulées et d'un exemple portant sur des données réelles.

Chiu, Yucel, Zanutto et Zaslavsky présentent une méthode d'imputation multiple de variables contextuelles manquantes en vue de leur utilisation en analyse de régression. Pour chaque enregistrement dans lequel manque la variable, et pour un échantillon d'enregistrements complets, ils sélectionnent des cas appariés d'après un ensemble de variables d'appariement. L'échantillon d'enregistrements complets est alors utilisé pour estimer à un ajustement de la régression pour d'autres variables non incluses dans les variables d'appariement. Les variables contextuelles pour les enregistrements incomplets font ensuite l'objet d'une imputation multiple. Enfin, les auteurs décrivent une application à l'étude du cancer du côlon et du rectum et utilisent des simulations pour comparer leur approche à trois autres méthodes d'ajustement pour la non-réponse.

Nandram et Choi examinent l'important problème de la non-réponse non-ignorable lors de l'estimation d'une variable de l'état de santé pour petits domaines. Face a une situation où les estimateurs habituels sont biaisés parce que le nombre de non-répondants est trop élevé, ils essaient de tenir compte des différences par modélisation. Nandram et Choi utilisent deux modèles hiérarchiques bayésiens de la non-réponse non-ignorable, un modèle de sélection et un modèle de mélange de schémas d'observation, pour analyser les données sur la santé. Un élément important dans leur modélisation est l'intégration de l'opinion des médecins en ce qui concerne le comportement de non-réponse et la variable des résultats. Les résultats donnent un ajustement exact pour la non-réponse et une meilleure mesure de précision.

Park et Fuller proposent une méthode en vue de réduire la probabilité d'obtenir des poids d'estimation négatifs lorsqu'on utilise un estimateur par la régression. Leur méthode consiste à approximer d'abord les probabilités d'inclusion, sachant les estimations d'Horvitz-Thompson pour un vecteur de variables auxiliaires, puis à utiliser les probabilités d'inclusion approximatives conditionnelles comme poids initiaux dans un estimateur par la régression. Ils montrent que leur méthode donne de bons résultats dans une étude en simulation. Ils comparent aussi les poids obtenus par leur méthode à ceux obtenus par la programmation quadratique, le raking ratio, une procédure logit et la méthode du maximum de vraisemblance.

La première des trois communications brèves publiées dans le présent numéro, rédigée par Andersson et Thorburn, montre que l'estimateur par la régression optimal peut être exprimé sous forme d'un estimateur par calage avec une fonction de distance choisie convenablement. L'estimateur optimal résultant est asymptotiquement plus efficace que l'estimateur par la régression généralisée (GREG) habituel. Une petite étude en simulation illustre plusieurs situations où l'estimateur optimal est significativement plus efficace que l'estimateur GREG.

Lynn et Gabler étendent les résultats de Gabler, Hader et Lahiri (volume 25, 1999) à l'expression de l'effet de plan dû à la mise en grappes de Kish. Ils donnent une méthode pratique d'estimation de la quantité de Kish à l'étape du plan d'échantillonnage lorsque seul les nombres totaux d'observations et de grappes sont nécessaires.

Meza et Lahiri examinent les limites d'un critère de sélection standard du modèle de régression, c'est-à-dire la statistique de Mallows, quand on l'applique aux modèles de régression à erreur emboîtée. Ils montrent que, si une application directe de la statistique de Mallows peut produire des méthodes de sélection de modèle inefficaces, une transformation appropriée des données pourrait résoudre le problème.

Finalement, nous voudrions vous informer que Harold Mantel occupera dorénavant le nouveau poste de Rédacteur en chef délégué. Harold fait partie du Comité éditorial depuis 15 ans. Son dévouement à la revue a été notable et sa contribution continue au processus éditorial a été de première importance pour assurer le maintien de la haute qualité de *Techniques d'enquête*.

M.P. Singh

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 3-13
Statistique Canada, N° 12-001-XIF au catalogue

Une étude de cas en couplage d'enregistrements

M. Winglee, R. Valliant et F. Scheuren [1]

Résumé

Le couplage d'enregistrements est un processus qui consiste à apparier des enregistrements provenant de deux fichiers en essayant de sélectionner les paires dont les deux enregistrements appartiennent à une même entité. La démarche fondamentale consiste à utiliser un poids d'appariement pour mesurer la probabilité qu'un appariement soit correct et une règle de décision pour décider si une paire d'enregistrements constitue un « vrai » ou un « faux » appariement. Les seuils de poids utilisés pour déterminer si une paire d'enregistrements représente un appariement ou un non-appariement dépend du niveau de contrôle souhaité sur les erreurs de couplage. Les méthodes appliquées à l'heure actuelle pour déterminer les seuils de sélection et estimer les erreurs de couplage peuvent donner des résultats divergents, selon le type d'erreur de couplage et la méthode de couplage. L'article décrit une étude de cas reposant sur les méthodes existantes de couplage pour former les paires d'enregistrements, mais sur une nouvelle approche de simulation (SimRate) pour déterminer les seuils de sélection et estimer les erreurs de couplage. SimRate s'appuie sur la distribution observée des données dans les paires appariées et non appariées afin de générer un grand ensemble simulé de paires d'enregistrements, d'attribuer un poids d'appariement à chacune de ces paires d'après les règles d'appariement spécifiées et d'utiliser les courbes de distribution des poids des paires simulées pour estimer l'erreur.

Mots clés : Appariement de fichiers; taux d'erreurs de couplage; poids d'appariement; seuil de sélection; dossiers médicaux.

1. Introduction

La démarche fondamentale de couplage d'enregistrements établie par Newcombe, Kennedy, Axford et James (1959) et par Fellegi et Sunter (1969) repose sur l'utilisation d'un poids d'appariement pour évaluer la probabilité qu'un appariement soit correct et sur une règle de décision pour classer les paires d'enregistrements. La règle de décision optimale repose sur deux seuils de poids d'appariement pour la sélection (un seuil supérieur au-dessus duquel un couplage est traité comme un vrai appariement et un seuil inférieur sous lequel un couplage est traité comme un non-appariement). Le choix de ces seuils dépend du taux d'erreurs de couplage acceptable préétabli et de la nécessité de réduire au minimum le nombre de couplages de situation indéterminée entre les deux seuils. De nos jours, les praticiens du couplage informatisé utilisent souvent un seuil de sélection unique pour éviter l'intervention manuelle que requiert le traitement des couplages indéterminés. Habituellement, le système prend automatiquement les décisions concernant les couplages après qu'on l'ait « réglé » de façon à respecter le niveau d'erreurs préétabli. Le défi tient au fait que les méthodes courantes de détermination du seuil de sélection et d'estimation des erreurs de couplage peuvent produire des résultats divergents, selon le type d'erreur de couplage, le choix de l'espace de comparaison et la méthode d'estimation.

Le but du présent article est de partager nos connaissances avec les praticiens qui ont besoin d'une méthode pour orienter le choix des couplages et pour estimer l'erreur. Notre étude de cas porte sur des fichiers d'événements médicaux provenant de la Medical Expenditure Panel Survey (MEPS) réalisée aux États-Unis. La MEPS est conçue pour recueillir des données sur les frais médicaux auprès de répondants sélectionnés dans les ménages et auprès des prestateurs de soins médicaux. L'objectif est de combiner les données en provenance des deux sources pour produire des estimations annuelles de l'utilisation des services médicaux et des frais médicaux (pour d'autres renseignements sur la MEPS, consulter Agency for Healthcare Research et Quality 2001).

Nous discutons ici du couplage d'enregistrements portant sur trois ensembles de fichiers annuels d'événements médicaux, provenant de la MEPS de 1996, de la MEPS de 1997 et de la MEPS de 1998. Chaque ensemble comprend un fichier des ménages contenant les événements déclarés par les répondants des ménages pour une année particulière et un fichier des prestateurs de soins médicaux contenant les données sur les événements correspondants déclarés par les personnes ayant prodigué les soins aux répondants des ménages. Chaque année, environ 50 000 événements médicaux ont été déclarés pour près de 10 000 personnes et environ 15 000 unités personne-prestateur de soins, en moyenne.

1. M. Winglee, Westat, Statistical Group, 1650 Research Boulevard, Rockville, MD 20850-3195, États-Unis; R. Valliant, Joint Program for Survey Methodology, University of Maryland and University of Michigan; F. Scheuren, NORC, University of Chicago.

Nous utilisons deux options fondées sur un modèle pour estimer l'erreur de couplage. L'une repose sur la simulation pour obtenir une distribution des poids pour divers niveaux de concordance. Cette technique, appelée SimRate, commence par la génération des distributions des poids pour les paires d'enregistrements appariées et non appariées. Partant de ces distributions, SimRate peut alors fournir des estimations des taux d'erreurs de couplage pour divers seuils de sélection. Ces taux d'erreurs peuvent ensuite servir de guide pour apporter des corrections et pour évaluer le succès de l'opération de couplage. Nous comparons la méthode SimRate à une deuxième méthode de modélisation élaborée par Belin et Rubin (1995). Comme nous espérons le montrer, ces approches ont toutes deux leur place; chacune possède des points forts, comme l'illustre les comparaisons.

2. Modèles de mélange de lois et approche SimRate

La méthode d'estimation de l'erreur de couplage par modélisation de mélange de lois présentée dans Belin et Rubin (1995) possède plusieurs caractéristiques intéressantes. Elle est souple en ce sens que le processus d'établissement des poids n'a pas à être considéré directement. Par conséquent, cette méthode est applicable à de nombreuses méthodes de création des poids. Lorsqu'un modèle est spécifié, on peut examiner les taux d'erreurs pour un continuum de valeurs seuils potentielles et construire des bandes de confiance pour surveiller la précision des estimations de l'erreur (voir la section 7).

Toutefois, les modèles de mélange de lois ont leurs limites. La méthode fournit un taux particulier d'erreurs, à savoir la proportion d'enregistrements couplés qui représentent effectivement des non-appariements, mais il est impossible d'estimer les taux de résultats faussement positifs ou faussement négatifs, puisqu'on ne tient pas compte des paires non couplées. Le taux d'erreur estimé est un taux conditionnel qui dépend de l'ensemble de paires d'enregistrements qui ont été couplées. De surcroît, les paramètres du modèle peuvent être difficiles à estimer si l'on ne peut isoler les distributions des poids pour les ensembles de paires d'enregistrements appariées et non appariées (voir Winkler 1994).

L'une des hypothèses importantes qui sous-tend l'approche de Belin–Rubin est qu'il est possible de transformer les distributions des poids dans les ensembles de paires d'enregistrements appariées et non appariées de façon à les rendre normales. Cette hypothèse pose une difficulté réelle ici, puisque les poids transformés pourraient être loin de suivre une loi normale si la distribution des poids pour l'ensemble de paires d'enregistrements appariées ou de paires d'enregistrements non appariées est multimodale.

Une autre exigence essentielle est de disposer d'un ensemble de données d'apprentissage dont les caractéristiques sont semblables à celles qu'il faudra apparier. Faute de posséder un bon ensemble de données d'apprentissage, l'estimation des paramètres d'entrée du modèle de mélange de lois pourrait être médiocre, ce qui aurait une incidence sur les taux d'erreurs estimés finaux. Dans le cas de notre application, en utilisant des données annuelles sur des événements médicaux répétées sur trois années, les paramètres n'étaient pas stables au cours du temps. Cette instabilité nous a obligés à utiliser un ensemble d'apprentissage pour chaque année, ce qui rend l'approche de Belin–Rubin peu pratique pour notre application, en raison du coût et du temps requis.

L'approche par simulation, SimRate, offre, comme la modélisation de mélange de lois, la capacité d'examiner divers seuils, ce qui permet à l'utilisateur de surveiller à la fois la sensibilité et la spécificité de la règle de décision en vue de sélectionner les paires appariées. À condition de pouvoir modéliser raisonnablement le processus utilisé pour établir les poids d'appariement, il est possible d'appliquer des méthodes personnalisées d'attribution de poids, telles que celles utilisées pour la présente étude de cas. La méthode requiert la production de paires d'enregistrements en se fondant sur la distribution des caractéristiques des ensembles de paires appariées et non appariées. Un certain effort doit être déployé pour générer rationnellement les populations de paires. Dans le cadre de nos travaux, nous avons réussi à générer ces populations au moyen de modèles multinomiaux.

3. Poids seuils et estimation de l'erreur de couplage

Plusieurs méthodes sont décrites dans la littérature pour sélectionner les appariements vrais et pour estimer les erreurs de couplage (par exemple, Bartlett, Krewski, Wang et Zielinski 1993; Armstrong et Mayda 1993; Belin 1993; Belin et Rubin 1995; Winkler 1992, 1995). Consulter Fellegi (1997) pour une vue d'ensemble de l'évolution du couplage d'enregistrements, Tepping (1968) et Larsen et Rubin (2001), pour d'autres méthodes de couplage et Scheuren (1983), pour une méthode de capture-recapture pour estimer l'erreur d'omission.

La comparaison des estimations obtenues selon les diverses approches se complique du fait que chacune a tendance à se concentrer sur des composantes différentes de l'erreur. En fait, les méthodes exposées dans la littérature sur le couplage d'enregistrements pour calculer des taux

d'erreurs de couplage manquent quelque peu de cohérence. Suit une illustration de ce problème.

Le tableau 1 représente un tableau de contingence 2×2 donnant les nombres de paires qui sont des appariements et des non-appariements réels et les nombres de couplages et de non-couplages déclarés par les systèmes de couplage. Les taux d'erreurs de couplage peuvent être estimés par rapport aux totaux réels figurant dans les colonnes. Dans le cas de la méthode de Fellegi et Sunter, l'estimation du taux d'erreurs de couplage qui sont des résultats faussement positifs est donnée par $\hat{\mu} = P(A_1 | U) = n_{12} / n_{\bullet 2}$ et celle du taux d'erreurs de couplage qui sont des résultats faussement négatifs, par $\hat{\lambda} = P(A_3 | M) = n_{21} / n_{\bullet 1}$ (voir aussi Armstrong et Mayda 1993). Il s'agit de taux que le programme SimRate est conçu pour estimer. Ils répondent à la question : « Parmi l'ensemble de paires qui sont des vrais appariements (ou des non-appariements), quelle proportion n'est pas identifiée correctement? ».

Tableau 1
Un tableau de contingence pour l'évaluation des erreurs de couplage

	Ensemble réel		
Ensemble déclaré	Appariement (M)	Non-appariement (U)	Total déclaré
Couplage (A_1)	n_{11} Vrai positif	n_{12} Faux positif	$n_{1\bullet}$
Non couplage (A_3)	n_{21} Faux négatif	n_{22} Vrai négatif	$n_{2\bullet}$
Total réel	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

Certaines évaluations des couplages ont également été fondées sur le calcul de taux par rapport aux totaux déclarés dans les lignes. Par exemple, Gomatam, Carter, Ariet et Mitchell (2002) utilisent le rapport $n_{12} / n_{1\bullet}$, qu'ils appellent pouvoir prédicteur positif du système de couplage. En revanche, d'autres dénomment cette expression taux de faux appariements (Belin et Rubin, 1995) ou taux de résultats faussement positifs déclarés (Bartlett et coll. 1993). Les taux calculés de cette manière répondent à la question : « Parmi les couplages (ou non-couplages) déclarés par le système, quelle proportion sont faux? » Il est important de répondre à ces deux questions lors de la sélection des paires qui sont des appariements vrais. C'est pourquoi il est séduisant d'utiliser à la fois SimRate et la méthode de Belin–Rubin, dans la mesure du possible.

4. Méthodes SimRate de simulation des distributions des poids pour estimer l'erreur de couplage

Quelle est la meilleure façon d'estimer les erreurs de couplage, compte tenu d'un budget et d'un délai limité, est une question à laquelle il est difficile de répondre. L'estimation exacte des erreurs de couplage devrait dépendre d'au moins deux facteurs, c'est-à-dire la possibilité qu'offrent les zones d'identification choisies de dépister sans ambiguïté les paires qui sont des appariements réels et la méthode de couplage utilisée. En regroupant ces deux facteurs, il est possible, dans des conditions données, de spécifier des catégories de couplage, d'estimer des probabilités de concordance et de déterminer des poids d'appariement.

Comme l'ont fait Newcombe et Kennedy (1962) et Jaro (1989), nous adoptons dans notre application, pour déterminer les distributions des poids, une méthode qui permet de tenir compte de ces facteurs. L'étape primordiale consiste à calculer le poids d'appariement et à classer toutes les configurations possibles de résultats concordants et non concordants des zones de comparaison selon le poids d'appariement. Puis, nous traçons la courbe de la fonction de distribution cumulative des poids pour les paires appariées et non appariées, et nous utilisons le tableau de poids résultant pour déterminer les seuils permettant d'atteindre les taux souhaités de résultats faussement positifs et faussement négatifs.

Une méthode idéale d'établissement de ces courbes consisterait à partir d'un ensemble de paires d'enregistrement pour lesquelles on connaît la situation réelle. Si nous en avions les ressources, nous pourrions utiliser un grand ensemble de paires réellement appariées, les classer par poids d'appariement et observer la proportion de ces paires au-dessus ou au-dessous d'un seuil donné. Pareillement, nous pourrions prendre un grand ensemble de paires que l'on sait être des paires réellement non appariées, les classer par poids, puis de nouveau calculer les proportions de part et d'autres du seuil. Les proportions de paires réellement appariées dont le poids est inférieur au seuil et de paires réellement non appariées dont le poids est supérieur au seuil représenteraient alors les estimations des taux d'erreurs associés à la façon dont l'algorithme d'appariement est appliqué.

Une méthode d'approximation de cette approche « idéale » (voir aussi Bartlett et coll. 1993) consiste à échantillonner les paires d'enregistrements et à procéder à un examen manuel pour déterminer lesquelles sont des appariements vrais. Après avoir repéré les appariements réels, nous pouvons y attacher les poids d'appariement provenant du système de couplage utilisé pour établir les distributions cumulatives des poids, tel que discuté plus haut. Naturellement, cette méthode souffre des contraintes bien connues de temps et d'autres ressources pour l'examen manuel et est rarement pratique dans le cas d'un grand échantillon.

Une autre méthode consiste à générer les distributions de poids cumulatives par simulation, approche qui est

l'élément central de SimRate. Pour l'expliquer de façon relativement détaillée, représentons une paire d'enregistrements par r et une zone de comparaison par $v (v = 1, \ldots, V$ zones). Dans notre application, les options pour les résultats des comparaisons comprenaient des catégories de concordance partielle et de résultats multiples, outre les catégories élémentaires de concordance totale et de non-concordance (voir aussi Newcombe 1988). Par conséquent, chaque zone v possède $i = 1, \ldots, c_v$ catégories de résultats. L'indicateur de résultat est $\mathbf{y}_{rv} = (y_{rv1}, \ldots, y_{rvc_v})$, un vecteur d'indicateurs montrant la catégorie dans laquelle rentre la paire *r*. Pour chaque zone, l'une des valeurs de y_{rvi} sera 1 et les autres, 0.

L'hypothèse théorique qui sous-tend l'approche SimRate est celle selon laquelle $\mathbf{y}_{rv}$ suit une loi multinomiale si la paire r est un appariement réel et une loi multinomiale différente si la paire est un non-appariement. Nous pouvons alors modéliser les vecteurs $\mathbf{y}_{rv}$ par une loi multinomiale de paramètre $\mathbf{m}_v = (m_{v1}, \ldots, m_{vc_v})$ si la paire est un appariement vrai et de paramètre $\mathbf{u}_v = (u_{v1}, \ldots, u_{vc_v})$ si elle est un non-appariement. Alors, la probabilité $m_{vi} = P$ (concordance de la catégorie *i* de la zone *v* dans la paire $r | r \in M$) est la probabilité conditionnelle de concordance pour la catégorie *I* de la zone *v*, sachant que la paire d'enregistrements *r* est comprise dans l'ensemble *M* de paires réellement appariées. Par contre, la probabilité $u_{vi} = P$ (concordance de la catégorie *i* de la zone *v* dans la paire $r | r \in U$) est la probabilité conditionnelle de concordance pour la catégorie *I* de la zone *v*, sachant que la paire d'enregistrements *r* est comprise dans l'ensemble *U* de paires réellement non appariées. En supposant que les variables d'appariement, $v = 1, \ldots, V$, sont indépendantes, nous pouvons spécifier la probabilité conjointe de $\mathbf{y}_r = (y_{r1}, \ldots, y_{rV})$ si une paire *r* est un appariement vrai sous la forme

$$\mathrm{P}(\mathbf{y}_r | r \in M) = \prod_{v=1}^{V} \prod_{i=1}^{c_v} m_{vi}^{y_{rvi}}.$$

La probabilité correspondante de la même configuration de données, si la paire est réellement un non-appariement, est

$$\mathrm{P}(\mathbf{y}_r | r \in U) = \prod_{v=1}^{V} \prod_{i=1}^{c_v} u_{vi}^{y_{rvi}}.$$

SimRate utilise des méthodes de simulation de Monte Carlo pour générer un grand nombre de réalisations de paires appariées et de paires non appariées en se fondant sur des estimations des probabilités m_{vi} et u_{vi}. Pour chaque paire simulée, l'application calcule un poids d'appariement w_r, qui est appliqué à une configuration particulière de données. Pour une réalisation donnée $\mathbf{y}_r$, un poids w_r est calculé pour la paire en additionnant les poids pour les catégories générées aléatoirement dans lesquelles la paire rentre. Le poids d'appariement w_r d'une paire d'enregistrements est habituellement estimé comme suit.

$$w_r = \log_2 \left[\frac{\prod_{v=1}^{V} \prod_{i=1}^{c_v} m_{vi}^{y_{rvi}}}{\prod_{v=1}^{V} \prod_{i=1}^{c_v} u_{vi}^{y_{rvi}}} \right].$$

Voir la section 6 sur les poids d'appariement utilisés pour la simulation.

La distribution cumulative de ces poids pour les paires appariées simulées est alors portée en graphique pour produire la courbe « Sim–*M* ». De la même façon, la distribution cumulative inverse des paires non appariées est représentée graphiquement pour produire la courbe « Sim–*U* » (voir la figure 1, à la section 8, pour un exemple de toutes les courbes de simulation utilisées dans l'étude). La proportion simulée de paires appariées dont les poids sont inférieurs au seuil est l'estimation du taux de résultats faussement négatifs. La proportion simulée de paires non appariées dont le poids est supérieur au seuil représente l'estimation du taux de résultats faussement positifs.

Cette approche requiert l'estimation empirique des distributions des variables d'appariement tant pour les paires réellement appariées que pour les paires réellement non appariées. Même si l'algorithme de détermination des poids repose sur l'hypothèse d'indépendance des variables d'appariement, les données réelles peuvent témoigner d'une dépendance. À condition de pouvoir générer des paires artificielles qui suivent raisonnablement la loi observée des données (en intégrant toute dépendance), alors cette méthode devrait produire des estimations appropriées des taux d'erreurs.

Dans notre étude de cas, nous avons modélisé des zones de données qui suivent des lois multinomiales indépendantes, mais cette hypothèse pourrait ne pas être raisonnable dans d'autres applications. Le concept de SimRate peut être appliqué à tout algorithme où sont utilisés des poids et un seuil d'acceptation aux fins de classification. Donc, d'autres méthodes que celle de Fellegi et Sunter (1969), comme celle de Belin et Rubin (1995), pourraient également être évaluées de cette façon. Si une méthode est nécessaire pour traiter des variables catégoriques dépendantes, les lois multinomiales multivariées décrites dans Johnson, Kotz et Balakrishnan (1997, chapitre 26) pourraient convenir. Cependant, dans des applications semblables à la nôtre, la méthode la plus simple pour tenir compte de la dépendance consiste à produire des classifications croisées des variables qui sont corrélées et à estimer les probabilités pour chaque cellule d'un tableau croisé. Par exemple, si deux variables possédant les catégories c_1 et c_2 sont associées, alors nous pouvons estimer la probabilité conjointe, p_{ij}, pour chaque cellule dans le tableau $c_1 * c_2$ et

utiliser ces probabilités dans la simulation. Des données peu nombreuses limiteront naturellement le nombre de cellules pour lesquelles la méthode est applicable. Cependant, en présence de données peu nombreuses, la pénalité en cas d'échec du modèle doit être faible.

5. Couplage des enregistrements des événements médicaux de la MEPS

Le couplage des enregistrements des événements médicaux de la MEPS a été réalisé en utilisant cinq zones d'identification, à savoir la date de l'événement (année, mois, jour et jour de la semaine), les codes de problème médical, les codes d'intervention, le code de frais globaux et la durée (nombre de jours) de l'hospitalisation. Ces zones sont décrites plus en détail dans Winglee, Valliant, Brick et Machlin (2000). Nous nous sommes servis d'un échantillon d'apprentissage tiré de la MEPS de 1996 pour établir les règles d'appariement et les catégories de résultats, ainsi que pour estimer les probabilités de concordance pour chaque catégorie, en tenant compte des concordances partielles et des valeurs particulières. Nous avons répété les mêmes règles d'appariement chaque année, en apportant des corrections mineures aux paramètres d'appariement.

Pour l'ensemble d'apprentissage, nous avons utilisé le système de couplage Automatch (Matchware 1996) et l'algorithme d'appariement unique pour sélectionner les paires couplées. L'appariement « unique » consiste à coupler de façon optimale un enregistrement du fichier A à un seul enregistrement du fichier B (Jaro 1989). En outre, nous avons utilisé l'algorithme d'appariement multi-multivoque (plusieurs vers plusieurs) pour générer un échantillon aléatoire de paires non couplées en vue de faciliter l'estimation de l'erreur de couplage. Cependant, les méthodes d'estimation des taux d'erreurs, qui sont décrites plus loin, s'appliquent à tout logiciel qui met en œuvre des méthodes de couplage fondées sur des poids d'appariement. Elles ne sont pas particulières à Automatch.

Afin de déterminer le seuil de sélection pour la MEPS, nous avons fait un compromis entre l'obtention d'un taux élevé d'appariements réels et la limitation des erreurs de couplage par appariement incorrect. Un poids seuil élevé réduirait au minimum le nombre de résultats faussement positifs (appariements incorrects), au prix d'une réduction du taux d'appariements corrects et d'une perte de données précieuses recueillies auprès des prestateurs de soins médicaux. Par ailleurs, un seuil faible augmenterait le nombre de résultats faussement positifs et pourrait avoir sur la répartition des données sur les dépenses un effet dont on ne pourrait venir en bout qu'en recourant à des techniques analytiques spéciales et, même alors, avec incertitude seulement. Puisque les deux sources de données avaient fait des déclarations sur manifestement les mêmes événements médicaux pour les mêmes personnes au cours de la même période, notre stratégie a été de maintenir un taux d'appariements raisonnablement élevé et de procéder à un examen manuel d'un nombre limité de paires couplées douteuses après sélection afin d'évaluer l'effet analytique de leur acceptation erronée. Basé sur cette décision, le taux moyen d'appariement réel pour les fichiers annuels d'événements médicaux de la MEPS était d'environ 85 %.

La courbe *M* pour l'échantillon d'apprentissage tiré de la MEPS de 1996, annotée courbe « Tra–*M* » a été produite en appliquant les poids d'appariement aux paires qui étaient des appariements « réels » pour un échantillon aléatoire de 500 personnes ayant participé à la MEPS de 1996. Pour ces personnes, les fichiers examinés manuellement contenaient 2 507 événements déclarés par les répondants des ménages et 2 804 événements déclarés par les prestateurs de soins médicaux. Des gestionnaires de données chevronnés ont passé les événements en revue et ont sélectionné 1 501 paires. Nous avons considéré ces dernières comme étant des appariements vrais dans la présente évaluation. Nous avons attribué aux paires appariées manuellement les poids déterminés d'après notre spécification d'appariement pour générer une fonction de distribution cumulative.

Nous avons produit la courbe *U* pour l'échantillon d'apprentissage de 1996, annotée courbe « Tra–*U* », au moyen d'un échantillon aléatoire de paires constituant des non-appariements. Nous avons appliqué une méthode d'échantillonnage aléatoire simple avec remise pour sélectionner 500 événements dans chacun des fichiers d'appariement et un algorithme d'appariement multi-multivoque (plusieurs vers plusieurs) pour générer les 250 000 paires d'événements possibles. Pour ces ensembles de paires sélectionnées au hasard, les chances qu'il existe une paire correctement appariée sont négligeables; par conséquent, nous avons considéré l'ensemble complet comme formé de paires incorrectement appariées. Nous avons appliqué les poids d'appariement obtenus selon notre spécification d'appariement et tracé la courbe « Tra–*U* » égale à 1 moins la distribution cumulative des poids de ces paires. La figure 1 de la section 8 montre les courbes Tra–*M* et Tra–*U* pour la MEPS de 1996. Ces courbes ont été lissées au moyen d'une fonction lowess non paramétrique (Chamber, Cleveland, Kleiner et Tukey 1983) en S–PLUS 2000 (1999).

6. Application de SimRate à la MEPS

La méthode SimRate d'établissement des distributions des poids consiste à appliquer des méthodes de simulation de Monte Carlo pour générer des ensembles distincts de 10 000 paires appariées et non appariées simulées pour créer

les courbes de poids. Pour générer les distributions de poids « Sim–*M* », nous avons estimé les probabilités m_{vi} d'après des paires couplées déterminées au moyen d'un algorithme d'appariement unique. Nous avons utilisé le système de couplage « réglé » pour sélectionner des paires appariées d'après les fichiers annuels d'appariement de 1996 et totalisé les fréquences observées pour chaque catégorie de résultats pour chacune des cinq zones d'appariement. Nous avons alors utilisé la proportion de paires entrant dans la catégorie i de la zone v comme estimation $\hat{m}_{vi}$ de la probabilité m_{vi}.

Pour les paires incorrectement appariées et la courbe « Sim–*U* », nous avons estimé les probabilités u_{vi} au moyen du même échantillon de paires incorrectement appariées que celui utilisé pour créer la courbe « Tra–*U* ». La différence est que nous avons utilisé ces paires pour observer les fréquences relatives pour chaque catégorie de résultats pour chacune des cinq zones d'appariement parmi les paires non appariées. Nous avons utilisé la proportion de paires entrant dans la catégorie i de la zone v comme estimation $\hat{u}_{vi}$ de la probabilité u_{vi}.

Pour une paire appariée simulée, nous avons généré une réalisation de la variable aléatoire multinomiale y_{rv} pour chaque zone d'appariement. Par exemple, nous avons généré une configuration telle que (concordance pour la date de l'événement, concordance pour la durée de l'hospitalisation, concordance pour la gamme de codes de problème médical, concordance conjointe selon le type d'intervention, et concordance pour une valeur spécifique d'un indicateur de frais globaux) en utilisant les probabilités d'appariement $\hat{m}_{vi}$ pour chaque catégorie de résultats. De la même façon, pour chaque paire non appariée, nous avons généré une réalisation d'une catégorie pour chacune des cinq zones en utilisant les probabilités d'appariement incorrectes $\hat{u}_{vi}$ mentionnées plus haut.

Pour une réalisation donnée $\mathbf{y}_r$, nous avons calculé un poids w_r pour la paire en additionnant les poids pour les catégories générées aléatoirement dans lesquelles se classait la paire. Les poids réels utilisés dans notre simulation étaient des poids corrigés que nous avons spécifiés, plutôt que des poids définis directement par le logiciel d'appariement (voir Winglee et coll. 2000). Donc, nous avons simulé la façon dont l'appariement serait effectivement mis en œuvre. Pour cela, nous avons calculé le poids d'appariement pour les ensembles de 10 000 paires appariées et de 10 000 paires non appariées, et nous avons représenté graphiquement les fonctions simulées des poids d'appariement.

Le tableau 2 donne des exemples de catégories de concordance partielle utilisées pour l'appariement d'après la date de l'événement et les estimations de $\hat{m}_{vi}, \hat{u}_{vi}$ et w_r utilisées dans la simulation SimRate. Nous avons défini, en tout, 19 catégories de résultats pour l'appariement de la date de l'événement, 9 catégories pour la durée de l'hospitalisation, 27 catégories pour l'intervention médicale et 3 catégories pour le problème de santé, ainsi que pour les frais globaux. Par exemple, pour la catégorie de résultats Concordance exacte de la date de l'événement, l'estimation de $\hat{m}_{vi}$ était 0,69, ce qui signifie que la concordance de la date de l'événement était exacte pour 69 % des paires couplées. L'estimation de $\hat{u}_{vi}$ pour cette catégorie de résultats était 0,003, montrant que 0,3 % seulement des paires non couplées présentaient une concordance pour cette zone. Le poids d'appariement pour la concordance exacte de la date de l'événement était 8,52 et celui pour la non-concordance complète (différence de plus de deux semaines et jour de la semaine différent) était $-6{,}64$ (voir Winglee et coll. 2000 pour les poids d'appariement selon la zone d'appariement et la catégorie de résultats).

Tableau 2
Estimations des probabilités multinomiales pour les paires appariées ($\hat{m}_{vi}$) et pour les paires non appariées ($\hat{u}_{vi}$), et poids d'appariement (w_{vi}) pour la zone d'appariement Date de l'événement

Règle d'appariement pour la date de l'événement	$\hat{m}_{vi}$	$\hat{u}_{vi}$	w_{vi}
Manquante	0,031	0,046	0,00
Appariement exact	0,693	0,003	8,52
Décalage de +/– 1 jour	0,068	0,006	5,71
Décalage de +/– 3 jour	0,023	0,005	4,09
Décalage de +/– 5 jours	0,014	0,005	2,47
Décalage de +/– 7 jours	0,030	0,006	2,84
Appariement du jour de la semaine uniquement	0,014	0,034	− 3,64
Non-concordance	0,003	0,547	− 6,64

Pour notre étude de cas, nous avons choisi des zones d'appariement qui étaient approximativement indépendantes. Ainsi, nous n'avons observé aucune association fonctionnelle entre la date de l'événement médical et d'autres zones d'appariement, comme celles du problème médical et de la durée de l'hospitalisation. Pour des zones comme celles des indicateurs d'intervention chirurgicale, d'examen radiologique et d'analyses biologiques, nous avons utilisé des tests du chi-carré et constaté une certaine dépendance entre l'intervention chirurgicale et l'examen radiographique concurrents. Pour contourner cette situation, nous avons estimé des probabilités conjointes et spécifié des règles d'appariement en vue de traiter ces indicateurs d'intervention comme une zone d'appariement unique (voir la section 4 plus haut). Par conséquent, nous avons pu appliquer la loi multinomiale indépendante pour la simulation.

Le tableau 3 donne les résultats de l'estimation de l'erreur de couplage par la méthode SimRate et par celle des courbes d'apprentissage au poids seuil de $w = 1$ pour les MEPS de 1996, 1997 et 1998. La méthode SimRate a été

facile à répéter chaque année. Par contre, la répétition des courbes de poids établies manuellement dépendait en partie de l'examen manuel et nous ne disposions que d'un seul échantillon d'apprentissage fiable, c'est-à-dire celui obtenu pour 1996. Il convient de souligner que les paires couplées utilisées dans SimRate généreront naturellement un certain pourcentage de résultats faussement positifs et de résultats faussement négatifs, c'est-à-dire certaines paires considérées incorrectement comme étant appariées, d'une part, et non appariées, d'autre part. Donc, les probabilités $\hat{m}_{vi}$ calculées de cette façon pour les zones mentionnées peuvent comporter une erreur. Il aurait été préférable d'estimer les probabilités m à partir d'un ensemble « vrai » pour lequel nous étions certains que tous les appariements étaient corrects. Cependant, les ensembles d'apprentissage appariés manuellement que nous avons pu produire étaient trop petits pour donner des estimations stables pour toutes les catégories détaillées d'appariement et, qui plus est, la sélection manuelle est également imparfaite. Cette différence pourrait expliquer, du moins partiellement, les estimations légèrement plus élevées du taux d'erreurs global produites par SimRate comparativement à celles obtenues d'après les courbes de poids établies pour l'échantillon d'apprentissage.

Tableau 3
Méthodes des courbes de poids pour estimer les taux d'erreurs de couplage au poids seuil de 1, MEPS 1996 à 1998

Méthode	Taux d'erreurs	1996	1997	1998
Courbes de simulation	Faussement négatifs	5,2	6,5	5,8
SimRate	Faussement positifs	9,0	6,9	7,6
Courbe de l'échantillon	Faussement négatifs*	3,3	3,3	3,3
d'apprentissage	Faussement positifs **	5,5	6,4	5,7

* Les estimations établies d'après la courbe Tra–*M* de 1996 ont été utilisées pour les trois années.

** Les estimations d'après la courbe Tra–*U* de 1996 ont été produites au moyen d'échantillons de 500 enregistrements provenant de chaque fichier d'appariement et d'un total de 250 000 paires non appariées. Les estimations pour 1997 et pour 1998 ont été produites au moyen d'autres courbes Tra–*U* fondées sur des échantillons de 1 000 enregistrements provenant de chaque fichier d'appariement et un total de 1 000 000 de paires non appariées.

7. Application des modèles de mélange de lois à la MEPS

Dans leur approche par modélisation d'un mélange de lois, Belin et Rubin (1995) considèrent la distribution des poids d'appariement observés à partir d'un système automatisé de couplage d'enregistrements comme étant un mélange de poids pour les vrais appariements et les faux appariements. En principe, la méthode du modèle de mélange de lois possède deux caractéristiques intéressantes qui conviennent pour la MEPS. En premier lieu, l'application répétée de la méthode peut se faire efficacement. Lorsqu'on dispose d'estimations paramétriques globales des paramètres transformés et du ratio des variances des deux distributions, on peut les appliquer à des données similaires pour l'estimation. Puisque le couplage des enregistrements de la MEPS est réalisé annuellement, des estimations globales calculées d'après des échantillons d'apprentissage antérieurs pourraient, en théorie, être appliquées à l'estimation de l'erreur de couplage lors d'années ultérieures si l'on ne dispose pas d'échantillons pour l'examen manuel.

Le deuxième avantage est que le modèle de mélange de lois peut s'appuyer sur plusieurs ensembles d'estimations de paramètres provenant de divers échantillons d'apprentissage et refléter les variations. Cette caractéristique est particulièrement séduisante dans le cas de la MEPS, parce que l'examen manuel est un processus complexe, qui n'est pas forcément toujours exact. Donc, une autre solution consiste à considérer les paires sélectionnées par le système informatique comme étant des appariements vrais et à les utiliser pour produire un ensemble d'estimations des paramètres de rechange. Ce processus peut également être répété en utilisant les échantillons d'apprentissage provenant de plus d'une année.

Dans notre application de l'approche de Belin–Rubin, nous avons utilisé les mêmes échantillons d'apprentissage provenant de la MEPS de 1996, ainsi qu'un deuxième échantillon d'apprentissage de même taille provenant de l'enquête de 1997. Suivant l'exemple de Belin–Rubin, nous avons appliqué la méthode du modèle de mélange de lois en utilisant des paires reconnues manuellement comme étant des appariements vrais ou faux produites par un système d'appariement biunivoque (un vers un) (il convient de souligner que ce genre de système produit assez peu de paires représentant de faux appariements pour l'estimation). Nous avons calculé les estimations fondées sur le modèle pour la MEPS de 1996 et la MEPS de 1997 en supposant que la sélection manuelle était correcte et, pour tester le comportement du modèle, nous avons calculé un deuxième ensemble d'estimations en supposant que les paires sélectionnées par le système informatique comme étant des appariements étaient les paires correctes.

L'application de la méthode comportait deux procédures, à savoir, la procédure de Box et Cox (1964) pour l'estimation globale des paramètres et la procédure de calage (Belin et Rubin 1995) pour ajuster un modèle de mélange de lois en vue d'estimer le taux d'erreurs. Avant d'appliquer la méthode de Box–Cox, nous avons rééchelonné les poids entre 1 et 1 000. La transformation de Box–Cox discutée par Belin et Rubin (1995) était

$$\Psi(w_r) = \frac{w_r^{\gamma} - 1}{\gamma \overline{w}^{\gamma-1}}$$

où w_r est le poids d'appariement pour la paire r, $\overline{w}$ est la moyenne géométrique des poids w_r, et γ est un paramètre qui dépend du fait que la paire appartient à l'ensemble de paires appariées ou non appariées.

Pour que la méthode du modèle de mélange de lois soit efficace, les poids transformés doivent suivre une loi approximativement normale. La distribution des poids non observés obtenus au moyen de nos données indiquait une bimodalité et pratiquement aucun chevauchement entre les poids d'appariement des paires appariées et non appariées (Belin–Rubin 1995 ont également observé une bimodalité). Par exemple, l'application de leur procédure de transformation aux paires sélectionnées par le système pour la MEPS de 1996 a donné les estimations des paramètres $\overline{w} = 585{,}7$ et $\gamma = 1{,}15$ pour les paires vraiment appariées et $\overline{w} = 113{,}1$ et $\gamma = 0{,}48$ pour les paires faussement appariées. Cependant, le rapprochement des poids transformés vers la loi normale était assez faible. Puisque les poids d'appariement sont égaux au logarithme d'un produit, c'est-à-dire à la somme des logarithmes des termes de ce produit, nous pourrions espérer que les poids suivent une loi normale si la somme compte un grand nombre de composantes. Cependant, nous ne disposions que de cinq zones pour procéder à l'appariement. Le petit nombre de zones pourrait expliquer, en partie, l'écart de nos données transformées par rapport à la loi normale.

Le tableau 4 donne les résultats de l'application du modèle de mélange de lois de Belin–Rubin aux données de la MEPS de 1996. Il contient les taux de faux appariements estimés d'après le modèle, l'intervalle de confiance à 95 % du taux estimé et le taux de faux appariements réels observés au poids seuil de 1. En considérant les paires déterminées par examen manuel comme étant des appariements vrais, une estimation fondée sur le modèle du taux prévu d'appariements faux au seuil de $w = 1$ était de 9,1 %, avec un intervalle de confiance à 95 % variant de 6,0 à 12,2. Par contre, le taux réel observé de faux appariements était de 14,5 %, valeur plus élevée que la borne supérieure de l'intervalle de confiance à 95 %. Il convient de souligner qu'il s'agit des taux de la forme $n_{12} / n_{1\bullet}$ du tableau 1, qui ne sont pas les mêmes que les taux estimés par la méthode SimRate et par la méthode des courbes de poids.

Puisque l'examen manuel n'est pas forcément toujours exact, une option, aux fins d'évaluation, consiste à traiter les paires couplées sélectionnées par le système informatique comme étant les paires vraiment appariées et à les utiliser pour la modélisation. Sous cette hypothèse, l'estimation fondée sur le modèle du taux d'erreurs prévu est de 0,9 et l'intervalle de confiance à 95 % varie de 0,6 à 1,2. Le taux réel observé dans ce cas, c'est-à-dire 0 %, est un résultat hypothétique où les paires couplées sélectionnées par l'ordinateur sont traitées comme étant correctes. En réalité, le niveau d'erreurs ne sera pas nul, si bien que l'intervalle de confiance pour le modèle de mélange de lois n'est pas nécessairement erroné.

Tableau 4
Estimations de l'erreur de couplage par la méthode du modèle de mélange de lois

MEPS de 1996	Pourcentage de faux appariements			
	Taux prévu	Borne inférieure*	Borne supérieure*	Taux observé
Appariement manuel	9,1	6,0	12,2	14,5
Appariement par le système	0,9	0,6	1,2	0,0

* Les bornes inférieure et supérieure sont celles de l'intervalle de confiance à 95 % du taux d'erreur prévu.

Nous avons produit des estimations globales des paramètres en utilisant les échantillons d'apprentissage sélectionnés manuellement ainsi que par le système pour les MEPS de 1996 et de 1997, afin de créer quatre ensembles de données d'entrée pour produire des estimations globales pour modéliser l'erreur de couplage pour la MEPS de 1998. Cela a été possible, parce que les données sont restées semblables et que les paires d'enregistrement ont été sélectionnées en appliquant les mêmes règles d'appariement pour les trois années. La seule différence était que nous n'avons pas procédé à un examen manuel pour la MEPS de 1998 et que nous n'avons pas pu utiliser la procédure de Box–Cox pour l'estimation globale des paramètres pour 1998 (parce qu'il n'existait pas d'indicateur manuel distinct pour les paires vraies et fausses). Pour cette application, nous avons utilisé une méthode bootstrap dans la procédure de calage de Belin–Rubin afin de nous appuyer sur plusieurs ensembles de paramètres de façon à refléter les incertitudes de l'estimation. Cependant, cette application n'a pas convergé après 150 itérations de la procédure d'estimation. Nous avons seulement pu conclure que les estimations globales des paramètres faites d'après des échantillons d'apprentissage antérieurs ne pouvaient être généralisées et fournir des estimations du taux d'erreurs pour des applications de couplage répétées.

8. Conclusion et incidences analytiques

La sélection d'un seuil et l'estimation de l'erreur de couplage représente un processus itératif comprenant des cycles répétés d'observation, d'estimation et de modélisation. Notre étude s'appuie sur des approches de modélisation pour estimer les erreurs de couplage et évaluer le pouvoir prédicteur du système de couplage. Les deux méthodes fournissent des renseignements valables pour déterminer la sélection des couplages et pour évaluer la

qualité des appariements déclarés, comme nous l'avons constaté dans le cas de la MEPS.

La méthode d'estimation fondée sur les courbes de distribution des poids a l'avantage de permettre de choisir un seuil de sélection pour atteindre le niveau acceptable d'erreurs de couplage. Par exemple, la figure 1 montre l'échantillon d'apprentissage et les courbes de distribution des poids simulées au moyen de SimRate d'après les fichiers d'appariement de la MEPS de 1996. Une droite verticale est tracée au poids seuil de sélection $w = 1$; les niveaux d'erreur pour la MEPS de 1996 (présentés au tableau 3) ont alors été estimés par le pourcentage cumulatif au niveau seuil. En déplaçant ce seuil, on peut tenter de réduire au minimum l'erreur totale de couplage en choisissant un seuil au point d'intersection des courbes *M* et *U*. Dans la présente étude de cas, les seuils optimaux suggérés par les deux ensembles de courbes de distribution des poids sont assez cohérents. Nous avons inclus une échelle du rapport de vraisemblance dans la figure pour donner une interprétation grossière de la vraisemblance du poids d'appariement. Par exemple, pour le poids d'appariement $w = 1$, le rapport de vraisemblance est égal à 2. Autrement dit, pour les enregistrements dont le poids d'appariement est égal ou supérieur à $w = 1$, la vraisemblance relative qu'il s'agisse d'un vrai appariement est au moins de 2 à 1.

En ce qui concerne la qualité des paires couplées, la figure 2 montre les distributions des estimations des taux de faux appariements calculés d'après le modèle de mélange de lois. Elle donne le taux de faux appariements estimé d'après le modèle, les bornes supérieure et inférieure de l'intervalle de confiance à 95 % des estimations du taux d'erreurs et les taux réels observés. Le premier panneau montre les estimations lorsqu'on traite les paires couplées sélectionnées par le système informatique comme étant des appariements vrais. Les deuxième et troisième panneaux montrent les estimations produites d'après les échantillons d'apprentissage tirés de la MEPS de 1996 et de la MEPS de 1997. La différence entre les deuxième et troisième panneaux montre le manque d'uniformité de la sélection manuelle par divers examinateurs dans notre application. Dans aucun des trois panneaux l'intervalle de confiance à 95 % des estimations d'après le modèle ne couvre les valeurs réelles observées. Idéalement, on devrait utiliser à la fois la figure 1 et la figure 2 pour orienter le choix des seuils de sélection.

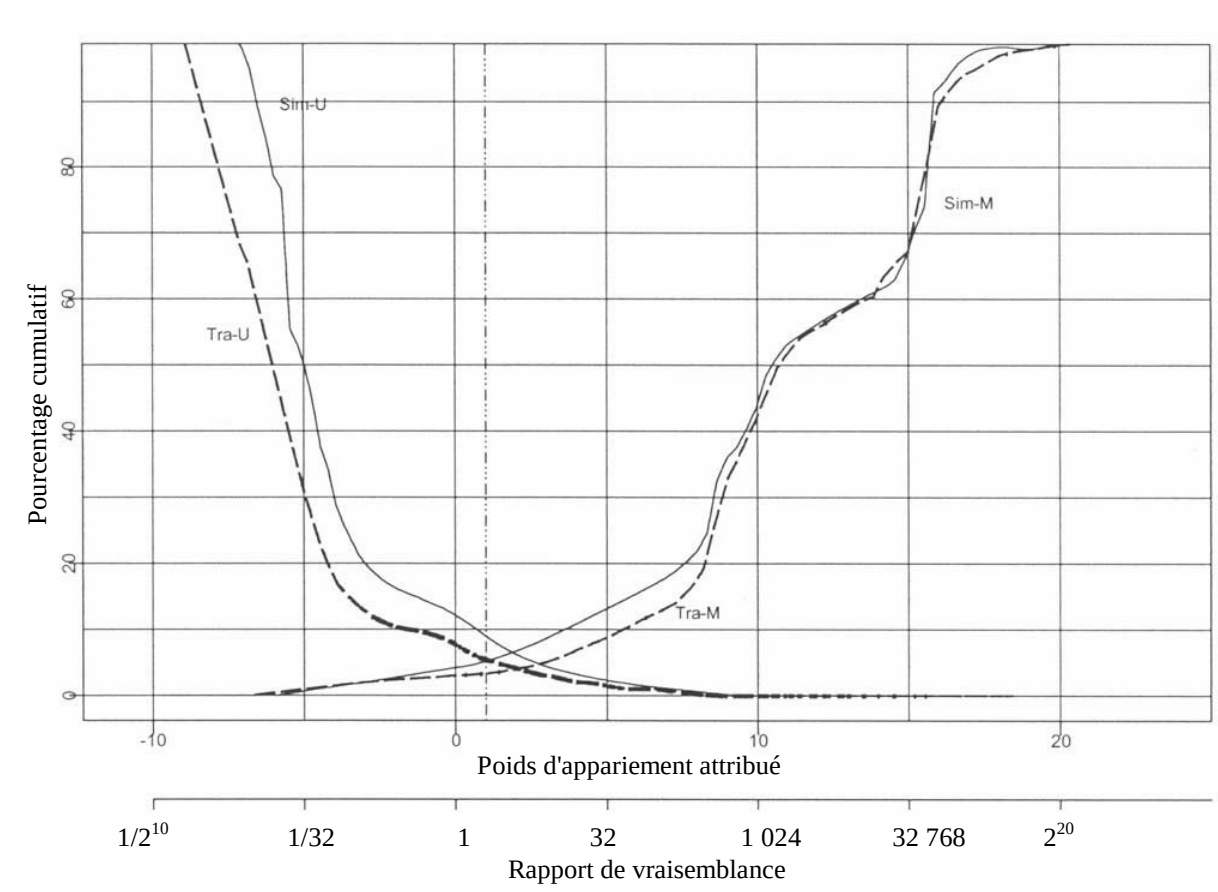

Figure 1. Courbes des poids pour la MEPS de 1996 d'après les méthodes SimRate et d'échantillons d'apprentissage; la droite de référence verticale en pointillé montre la valeur seuil de 1.

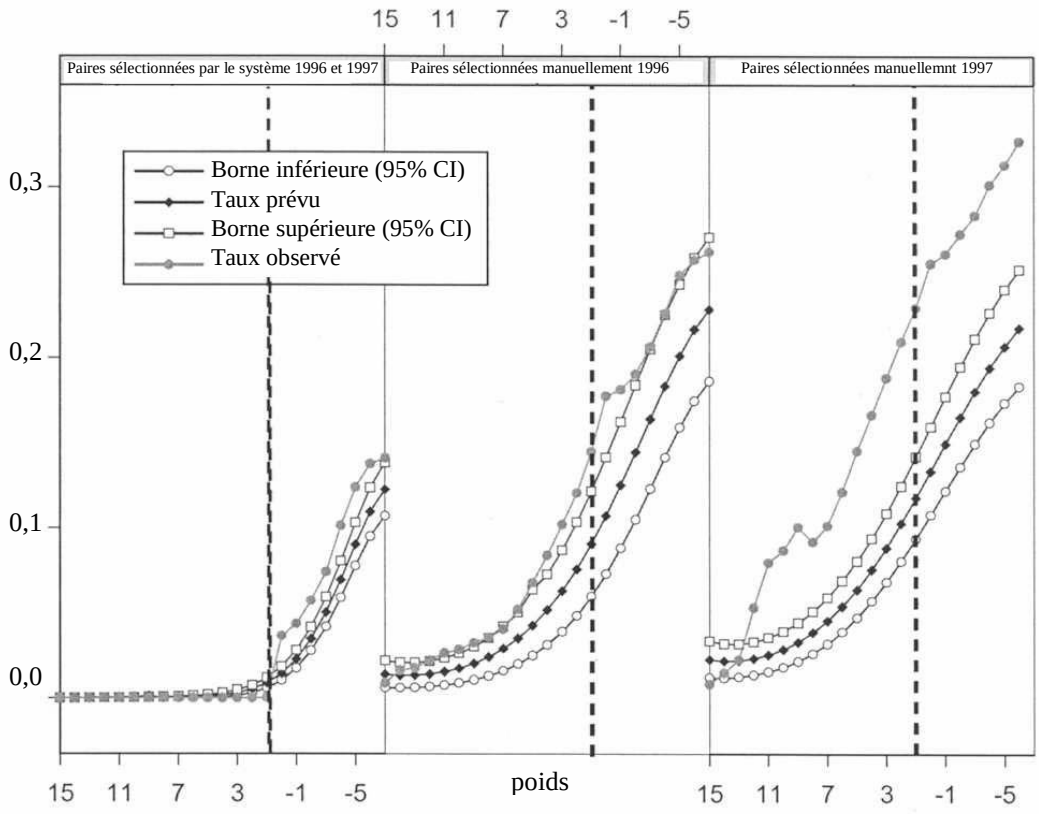

Figure 2. Estimations d'après le modèle de mélange de lois des taux de faux appariements selon le poids, échantillons d'apprentissage tirés de la MEPS de 1996 et de 1997 (une droite verticale est tracée au poids = 1, qui est le seuil).

Dans notre application, SimRate s'est avérée être un outil informatif et souple pour la détermination des seuils de sélection et l'estimation des taux d'erreurs. Étant donné un modèle multinomial ou d'autres modèles pour les variables d'appariement, la méthode SimRate fournit des estimations du taux d'erreurs que l'on obtiendrait par application répétée de l'algorithme d'appariement à un grand nombre de paires d'enregistrement candidates. Elle s'avère également souple en ce qui concerne le choix des ensembles de paires à comparer pour calculer les taux.

Bien que notre application nous ait permis de réaliser nos objectifs d'estimations des taux d'appariement et des taux d'erreurs pour la MEPS, une étude plus approfondie pourrait être réalisée avant l'étape de l'analyse ou durant celle-ci. Faute d'espace, nous ne pouvons élaborer ces travaux dans le contexte de la présente étude de cas, mais nous pourrions mentionner deux approches générales. En premier lieu, il est possible de repondérer les résultats finals et de les corriger pour les faux non-appariements, en traitant ceux-ci d'une façon analogue à la non-réponse unitaire (par exemple, comme dans Oh et Scheuren 1980). Pour traiter les appariements incorrects, les idées proposées dans Scheuren et Winkler (1993 et 1997) et dans Lahiri et Larsen (2002) vaudraient peut-être la peine d'être consultées. La question de savoir si ces étapes supplémentaires sont nécessaires dépend, évidemment, de l'utilisation finale prévue des données couplées.

Remerciements

L'étude fondamentale du couplage d'enregistrements présentée ici a été réalisée en vertu des contrats 290-99-0002 et 290-94-2002 parrainés par l'Agency for Healthcare Research and Quality et le National Center for Health Statistics. Les auteurs remercient Steven B. Cohen, Steven Machlin et Joel Cohen, de l'Agency for Healthcare Research and Quality, de leurs commentaires à diverses étapes de cette étude, et Thomas Belin, pour ses suggestions concernant une version antérieure.

Bibliographie

Agency for Healthcare Research et Quality (2001). MEP – Medical Expenditure Panel Survey. <http://www.ahrq.gov/data/mepsix.htm>.

Armstrong, J.B., et Mayda, J.E. (1993). Estimation modéliste des taux d'erreur liés au couplage d'enregistrements. *Techniques d'enquête*, 19, 147-158.

Bartlett, S., Krewski, D., Wang, Y. et Zielinski, J.M. (1993). Évaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisé. *Techniques d'enquête*, 19, 3-13.

Box, G.E.P., et Cox, D.R. (1964). An analysis of transformations (avec discussion). *Journal of the Royal Statistical Society*, Series B, 26, 206-252.

Belin, T.R. (1993). Évaluation des sources de variation dans le couplage d'enregistrements au moyen d'une expérience factorielle. *Techniques d'enquête*, 19, 15-33.

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

Chambers, J.M., Cleveland, W.S., Kleiner, B. et Tukey, P. (1983). *Graphic Methods for Data Analysis*, Duxbury Press, Boston.

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

Fellegi, I.P. (1997). Record linkage and public policy – A Dynamic Evoluation. *Proceedings of the International Workshop and Exposition, Federal Committee on Statistical Methodology, Office of Management and Budget*, Washington, DC.

Gomatam, S., Carter, R., Ariet, A. et Mitchell, G. (2002). An empirical companion of record linkage procedures. *Statistics in Medicine*, 21, 1485-1496.

Jaro, M.A. (1989). Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

Johnson, N.L., Kotz, S. et Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: John Wiley & Sons, Inc.

Lahiri, P., et Larsen, M.D. (2002). Regression analyses with linked data. (Manuscript d'ébauche).

Larsen, M.D., et Rubin, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32-41.

Matchware Technologies Inc. (1996). *AutoMatch: Generalized Record Linkage System User's Manual*. Silver Spring, MD: Matchware Technologies, Inc.

Newcombe, H.B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. Oxford University Press, New York.

Newcombe, H.B., Kennedy, J.M., Axford, S.J. et James, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

Newcombe, H.B., et Kennedy, J.M. (1962). Record linkage: Making maximum use of the discriminating power of identifying information. *Communications of the Association for Computing Machinery*, 5, 563-567.

Oh, H.L., et Scheuren, F.(1980). Fiddling around with nonmatches and mismatches, *Studies from Interagency Data Linkages Series.* Social Security Administration, Rapport No. 11.

Scheuren, F. (1983). Design and estimation for large federal surveys using administrative records. *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 377-381.

Scheuren, F., et Winkler, W.E. (1993). Analyse de régression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.

Scheuren, F., et Winkler, W.E. (1997). Analyse de régression des fichiers de données appariés par ordinateur - Partie II. *Techniques d'enquête*, 23, 171-180.

S-Plus 2000 (1999). MathSoft, Inc. Data Analysis Products Division, Seattle, Washington.

Tepping, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

Winglee, M., Valliant, R., Brick, J.M. et Machlin, S. (2000). Probability matching of medical events. *Journal of Economic and Social Measurement*, 26, 129-140.

Winkler, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 829-834.

Winkler, W.E. (1994). *Advanced Methods for Record Linkage.* Bureau of the Census Statistical Research Division, Statistical Research Report Series, RR 94/05.

Winkler, W.E. (1995). *Matching and record linkage*. Dans *Business Survey Methods*, (Éds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. College et P.S. Kott) New York: John Wiley & Sons, Inc., 355-384.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 15-24
Statistique Canada, N° 12-001-XIF au catalogue

L'effet des erreurs de couplage d'enregistrements sur les estimations du risque dans les études-cohorte de mortalité

D. Krewski, A. Dewanji, Y. Wang, S. Bartlett, J.M. Zielinski et R. Mallick [1]

Résumé

L'élaboration de la méthodologie de couplage informatisé d'enregistrements a facilité la réalisation d'études-cohorte de mortalité dans lesquelles les données sur l'exposition provenant d'une base de données sont couplées électroniquement à celles sur la mortalité provenant d'une autre base de données. Cependant, cette méthode donne lieu à des erreurs de couplage causées par l'appariement incorrect d'une personne figurant dans l'une des bases de données à une personne différente dans l'autre base de données. Dans le présent article, nous examinons l'effet des erreurs de couplage sur les estimations d'indicateurs épidémiologiques du risque, comme les ratios standardisés de mortalité et les paramètres des modèles de régression du risque relatif. Nous montrons que les effets sur les nombres observé et attendu de décès sont de sens opposé et que, par conséquent, ces indicateurs peuvent présenter un biais et une variabilité supplémentaire en présence d'erreurs de couplage.

Mots clés : Étude de cohorte; couplage informatisé d'enregistrements; erreurs de couplage; poids seuil de couplage; régression de Poisson; régression du risque relatif; ratio standardisé de mortalité.

1. Introduction

Ces dernières années, plusieurs études de cohorte historiques ont été réalisées en épidémiologie environnementale en se servant de bases de données administratives existantes comme sources d'information (Howe et Spasoff 1986; Carpenter et Fair 1990). En termes généraux, cette approche consiste à coupler des enregistrements de données sur l'exposition humaine à des risques environnementaux à des enregistrements de données sur l'état de santé, souvent au moyen de méthodes informatisées d'appariement d'enregistrements individuels provenant de bases de données différentes. Dans le cas d'une étude-cohorte de mortalité, le statut vital de chaque membre de la cohorte est déterminé par couplage aux enregistrements de décès des bases de données sur la mortalité tenues à jour par les organismes gouvernementaux. L'existence d'une surmortalité dans la cohorte comparativement à la population générale pourrait être due aux expositions subies par les membres de la cohorte.

En termes spécifiques, le couplage d'enregistrements est le processus consistant à regrouper deux ou plusieurs éléments d'information enregistrés distincts concernant une même entité (Bartlett, Krewski, Wang et Zielinski 1993). Les procédures de couplage informatisé d'enregistrements (CIE) sont devenues de plus en plus perfectionnées, grâce à l'utilisation d'algorithmes complexes pour évaluer la probabilité que l'appariement de deux enregistrements soit correct (Hill 1988; Newcombe 1988). Statistique Canada a mis au point un système de CIE appelé CANLINK capable de coupler les enregistrements d'un même fichier, ainsi que ceux de deux fichiers distincts (Howe et Lindsay 1981; Smith et Silins 1981). Ce système attribue à chaque paire d'enregistrements un poids reflétant la probabilité qu'il s'agisse d'un appariement. Deux seuils sont fixés : les appariements potentiels dont le poids de couplage est supérieur au seuil supérieur sont considérés comme des couplages, tandis que les appariements potentiels dont le poids de couplage est inférieur au seuil inférieur sont considérés comme des non-couplages. Les cas d'appariement possible dont le poids est compris entre les seuils inférieur et supérieur sont résolus à l'aide de renseignements supplémentaires, lorsqu'ils sont disponibles. Sinon, on choisit un seuil unique pour faire la distinction entre les couplages et les non-couplages.

Lors de toute étude comportant un couplage d'enregistrements, des mesures strictes sont prises pour assurer la non-divulgation des enregistrements protégés aux termes de la *Loi sur la statistique*. Toutes les études qui nécessitent le couplage d'enregistrements faisant partie de bases de données protégées doivent être soumises à un processus d'examen et d'approbation rigoureux avant d'être exécutées

1. D. Krewski, Centre McLaughlin d'évaluation du risque pour la santé des populations, Université d'Ottawa, Ottawa (Ontario), Canada K1N 6N5. School of Mathematics & Statistics, Carleton University, Ottawa (Ontario), Canada, K1S 5B6. La correspondance devrait être adressée à : A. Dewanji, Applied Statistics Unit, Indian Statistical Institute, Kolkata, India; Y. Wang, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; S. Bartlett, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; J. M. Zielinski, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; R. Mallick, Centre McLaughlin d'évaluation du risque pour la santé des populations, Université d'Ottawa, Ottawa (Ontario), Canada K1N 6N5. School of Mathematics & Statistics, Carleton University, Ottawa (Ontario), Canada, K1S 5B6.

conformément à des procédures bien établies en vue d'assurer le respect de la confidentialité des données (Singh, Feder, Dunteman et Yu 2001). Tous les fichiers couplés contenant des renseignements permettant d'identifier des individus restent sous la garde de Statistique Canada (Labossière 1986).

Des méthodes informatisées de couplage d'enregistrements ont été utilisées pour coupler des données sur l'exposition environnementale à celles de la Base canadiennes de données sur la mortalité (BCDM). Par exemple, une étude a été entreprise pour étudier les liens éventuels entre les causes de décès chez plus de 326 000 exploitants agricoles au Canada et diverses variables sociodémographiques et d'exploitation agricole, particulièrement l'utilisation de pesticides (Jordan-Simpson, Fair et Poliquin 1990). Cette étude comportait le couplage des données de la BCDM à celles du Recensement de la population de 1971 et du Recensement de l'agriculture de 1971. Une autre étude permanente de grande portée est fondée sur le Fichier dosimétrique national (FDN) du Canada (Ashmore et Grogan 1985; Ashmore et Davies 1989). Le FDN contient des renseignements remontant jusqu'à 1950 sur les expositions professionnelles aux rayonnement ionisants subies par plus de 400 000 Canadiens. Récemment, les enregistrements du FDN ont été couplés à ceux de la BCDM en vue d'étudier les associations entre la surmortalité due au cancer et l'exposition professionnelle à de faibles niveaux de rayonnements ionisants (Ashmore, Krewski et Zielinski 1997; Ashmore, Krewski, Zielinski, Jiang, Semenciw et Létourneau 1998). Plus récemment, les enregistrements du FDN ont été couplés à ceux de la Base canadienne de données sur l'incidence du cancer (Sont, Zielinski, Ashmore, Jiang, Krewski, Fair, Band et Létourneau 2001). La liste complète des autres études relatives à la santé fondées sur le couplage de données sur l'exposition à celles de la BCDM a été dressée par Fair (1989).

Le succès des études axées sur le couplage d'enregistrements dépend de la qualité des bases de données couplées (Roos, Soodeen et Jebamani 2001). À l'aide de données administratives longitudinales représentatives de la population, Roos et coll. ont examiné les problèmes de qualité de données dans les études sur l'état de santé et les soins de santé. Ardal et Ennis (2001) ont tenu compte des erreurs systématiques présentes dans les bases de données administratives intervenant dans l'analyse secondaire de l'information sur la santé. S'il est vrai que les études fondées sur le couplage d'enregistrements donnent de meilleurs résultats quand les données sont de haute qualité, les contraintes liées à la qualité des données sont compensées dans une certaine mesure par la grande taille des échantillons sur lesquels reposent de nombreuses bases de données administratives.

Les études par couplage d'enregistrements offrent plusieurs avantages par rapport aux études épidémiologiques classiques. L'utilisation des bases de données administratives existantes évite de devoir recueillir de nouvelles données pour les études sur la santé et permet d'obtenir des échantillons de grande taille moyennant assez peu d'efforts. Selon la nature des bases de données utilisées, le couplage d'enregistrements offre un moyen peu coûteux d'explorer de nombreuses associations éventuelles dans le cadre des études épidémiologiques. Le couplage d'enregistrements présente aussi certains inconvénients. Les chercheurs exercent généralement fort peu de contrôle sur l'information recueillie et le nombre de sujets perdus de vue lors des suivis peut être important. Les erreurs de couplage, qui sont le sujet du présent article, sont un autre inconvénient du couplage d'enregistrements. Inévitablement, certains enregistrements concordants ne seront pas couplés et certains enregistrements non concordants seront couplés incorrectement.

Assez peu de travaux ont été accomplis en vue de déterminer l'effet de ces erreurs de couplage sur les inférences statistiques. Neter, Maynes et Ramanathan (1965) ont utilisé un modèle de régression linéaire simple pour analyser l'effet des erreurs introduites durant le processus d'appariement. Selon leurs résultats, les erreurs de couplage font augmenter la variance résiduelle et introduisent un biais dans l'estimation de la pente de la droite de régression. Winkler et Scheuren (1991) établissent une expression du biais dû aux erreurs de couplage dans les estimations des coefficients de régression linéaire. Les progrès concernant l'estimation des taux d'erreurs de couplage réalisés par Belin et Rubin (1991) ont permis à Scheuren et Winkler (1993) de mettre en œuvre une méthode améliorée de correction du biais. L'application des méthodes de régression linéaire à l'analyse des fichiers de données appariées informatiquement est discutée plus en détail par Scheuren et Winkler (1997).

L'objet du présent article est d'étudier l'effet des erreurs de couplage sur les inférences statistiques dans les études-cohorte de la mortalité. À la section 2, nous décrivons les modèles de régression du risque relatif employés pour analyser les données provenant de ce genre d'études et nous élaborons des expressions pour les nombres observé et attendu de décès fondés sur ces modèles. À la section 3, nous discutons de l'effet des erreurs de couplage sur les nombres observé et attendu de décès et de personnes-années à risque. À la section 4, nous analysons l'effet des erreurs de couplage sur les estimations des ratios standardisés de mortalité (RSM) et sur les paramètres de régression du risque relatif. Les deux types d'erreurs peuvent introduire un biais et une variabilité supplémentaire dans les estimations

de ces paramètres. À la section 5, nous présentons nos conclusions.

2. Modèles de régression du risque relatif

Les méthodes statistiques d'analyse des données provenant d'études-cohorte de la mortalité sont bien établies (Breslow et Day 1987). L'objectif principal de ce genre d'analyse est de déterminer si l'exposition à l'agent d'intérêt augmente le taux de mortalité chez les membres de la cohorte. La mortalité est caractérisée par la fonction de risque, qui précise le taux de mortalité sous forme de fonction du temps. Si nous représentons par T le moment du décès, la fonction de risque au temps u se définit formellement comme suit

$$\lambda(u) = \lim_{\Delta u \downarrow 0} \frac{\Pr\{u \le T < u + \Delta u \,|\, T \ge u\}}{\Delta u}. \qquad (1)$$

Soit $\lambda_i(u)$ la fonction de risque pour une cause particulière de décès au temps u pour l'individu $i = 1, ..., N$ dans une cohorte de taille N, et soit $\mathbf{z}_i(u)$ un vecteur correspondant de covariables propres à cet individu. Nous supposons que ces covariables ont pour effet de modifier le risque de base $\lambda^*(u)$ conformément au modèle de régression du risque relatif

$$\lambda_i(u) = \lambda^*(u)\gamma\{\beta'\mathbf{z}_i(u)\}, \qquad (2)$$

où γ est une fonction positive des covariables et β est un vecteur de paramètres de régression.

Deux cas particuliers du modèle général de régression du risque relatif présentant un intérêt sont les modèles multiplicatif et additif de régression du risque. Définissons la fonction γ figurant dans (2) par

$$\log\gamma(z) = \frac{(1+z)^\rho - 1}{\rho}. \qquad (3)$$

Quand $\rho = 1$, le modèle général de régression du risque relatif se réduit au modèle multiplicatif de régression du risque

$$\lambda_i(u) = \lambda^*(u)\exp\{\beta'\mathbf{z}_i(u)\}, \qquad (4)$$

Ce modèle à risques proportionnels, qui a été introduit par Cox (1972), est d'usage très répandu en analyse des données sur la mortalité (Kalbfleish et Prentice 1980). Le modèle additif de régression du risque

$$\lambda_i(u) = \lambda^*(u) + \beta'\mathbf{z}_i(u) \qquad (5)$$

survient en tant que cas limite quand $\rho \to 0$.

Soit t_i^0 et t_i^1 l'âge au moment de l'entrée dans l'étude et l'âge au moment de la perte de vue (due à l'abandon par le sujet, à l'interruption de l'étude ou au décès) du i^{e} sujet de la cohorte, respectivement. Soit $\delta_i = 1$ ou 0, selon que le i^{e} sujet est ou n'est pas décédé au moment de la perte de vue. La fonction de log-vraisemblance fondée sur le modèle du risque relatif (2) peut s'écrire

$$\log L = \sum_{i=1}^{N} \left\{ \begin{array}{l} \delta_i \log(\gamma\{\beta'\mathbf{z}_i(t_i^1)\}) \\ \qquad -\int_{t_i^0}^{t_i^1} \gamma\{\beta'\mathbf{z}_i(u)\}\lambda^*(u)\,du \end{array} \right\}. \quad (6)$$

Lorsqu'il n'existe qu'une covariable $z_i(u) \equiv 1$, l'estimation du maximum de vraisemblance de $\theta = \exp\{\beta\}$ se réduit au ratio standardisé de mortalité RSM = OBS/ATT, où OBS = $\sum_{i=1}^N \delta_i$ et ATT = $\sum_{i=1}^N e_i$ sont les nombres observé et attendu de décès, respectivement, avec $e_i = \int_{t_i^0}^{t_i^1} \lambda^*(u)\,du$.

La maximisation de la fonction de vraisemblance (6) peut donner lieu à des calculs fastidieux dans le cas d'échantillons de grande taille. Breslow, Lubin et Langholz (1983) simplifient cette fonction en supposant que les covariables prennent des valeurs constantes dans les états par lesquels passe un sujet durant le cours de l'étude. Ces états sont définis par des classifications croisées des covariables d'intérêt. Plus précisément, supposons qu'il existe J états de ce genre $\{S_j;\ j = 1, ..., J\}$, tels que $\mathbf{z}_i(u) = \mathbf{z}_j$ chaque fois que le i^{e} sujet se trouve dans l'état S_j au temps u. Ces états sont mutuellement exclusifs et exhaustifs, si bien que, à tout temps u, chaque membre de la cohorte se trouvera dans un état, et uniquement un. La fonction de log-vraisemblance (6) peut s'écrire

$$\log L = \sum_{j=1}^{J} \{d_{jj}\log(\gamma\{\beta'\mathbf{z}_i\}) - \gamma\{\beta'\mathbf{z}_j\}e_j\}, \qquad (7)$$

où

$$e_j = \sum_{i=1}^{N} \int_{[\mathbf{z}_i(u)\in S_j]} \lambda^*(u)\,du \qquad (8)$$

est la contribution au nombre attendu de décès provenant de toutes les personnes-années d'observation dans l'état S_j, et d_{jj} est le nombre total de décès dans cet état. En posant que $\Lambda_j(\beta) = \log(\gamma\{\beta'\mathbf{z_j}\})$, nous obtenons l'estimation du maximum de vraisemblance $\hat\beta$ de β en tant que solution de l'équation de score

$$\frac{\partial \log L}{\partial \beta} = \sum_{j=1}^{J} \frac{\partial \Lambda_j(\hat\beta)}{\partial \beta}\{d_{jj} - \exp\{\Lambda_j(\hat\beta)\}e_j\} = 0. \quad (9)$$

3. L'effet des erreurs de couplage sur les nombres observé et attendu de décès

Deux grands types d'erreurs peuvent se produire lors du couplage de fichiers de données dans le contexte du CIE (Fellegi et Sunter 1969). Un résultat faussement positif a lieu quand un membre de la cohorte encore en vie est

incorrectement désigné comme étant décédé et un résultat faussement négatif survient quand un membre décédé de la cohorte est considéré comme étant en vie. Plus précisément, pour le développement mathématique qui suit, un résultat faussement positif survient dans un état particulier quand un individu qui demeure en vie pendant tout le temps où il se trouve dans cet état est incorrectement étiqueté comme étant décédé dans cet état. Pareillement, un résultat faussement négatif survient dans un état particulier quand un membre de la cohorte qui est décédé avant d'atteindre cet état ou pendant qu'il se trouvait dans cet état est considéré comme étant en vie en étant dans cet état. Dans un état donné, les résultats faussement positifs et faussement négatifs représentent donc des cas particuliers de l'erreur de classification discutée par Anderson (1974, chapitre 6.2.1). À la présente section, nous examinons l'effet de ces deux types d'erreurs de couplage sur les nombres observé et attendu de décès, respectivement. À cet fin, nous commençons par définir des jeux d'indices dans les divers états que nous utiliserons pour représenter les ensembles d'enregistrements correctement appariés et incorrectement appariés.

3.1 Erreurs de couplage

Soit A_j et D_j l'ensemble d'étiquettes pour les membres de la cohorte qui demeurent en vie dans l'état S_j, et pour ceux qui sont décédés dans l'état S_j, respectivement. Soit D_{jj} le sous-ensemble de D_j correspondant aux personnes qui sont décédées dans l'état S_j. Soit A_j^L, D_j^L et D_{jj}^L les ensembles correspondants à la présence d'erreurs de couplage. Définissons en outre D_j^P comme étant l'ensemble d'étiquettes des individus en vie dans l'état S_j (c'est-à-dire dans A_j), mais étiquetés comme étant décédés dans l'état S_j, c'est-à-dire correspondant aux résultats faussement positifs dans S_j. De la même façon, A_j^N est l'ensemble d'individus décédés dans l'état S_j (c'est-à-dire dans D_j), mais étiquetés comme en étant en vie dans l'état S_j, c'est-à-dire correspondant aux résultats faussement positifs dans S_j. Représentons aussi par D_{jj}^P le sous-ensemble de D_j^P correspondant aux individus étiquetés comme étant décédés dans l'état S_j et, pareillement, par A_{jj}^N le sous-ensemble d'individus de A_j^N qui sont décédés dans l'état S_j (c'est-à-dire dans D_{jj}). Ces ensembles satisfont aux relations $A_j^L = (A_j - D_j^P) \cup A_j^N$, $D_j^L = (D_j - A_j^N) \cup D_j^P$ et $D_{jj}^L = (D_{jj} - A_{jj}^N) \cup D_{jj}^P$.

L'effet des erreurs de couplage sur la fonction de vraisemblance donnée par (7) peut être décrit comme suit. Soit t_{ij}^0 le temps auquel le i^{e} individu entre, réellement ou par erreur de couplage, dans le j^{e} état S_j. De même, t_{ij}^1 représente le moment du décès (s'il a lieu, réellement ou par erreur de couplage) du i^{e} individu dans l'état S_j et t_{ij}^2, le moment de la sortie de l'état S_j, réellement ou par erreur de couplage. Notons que, si t_{ij}^1 existe, il est inférieur ou égal à t_{ij}^2. Par souci de simplicité, supposons que t_{ij}^1, s'il existe, est égal à t_{ij}^0; autrement dit, tous les décès qui surviennent dans un état particulier le font au moment correspondant de l'entrée dans cet état. Bien que cette hypothèse produise une sous-estimation du nombre attendu de décès, aux fins de l'étude du biais, elle n'est peut-être pas si contestable. Le fait de supposer que tous les décès surviennent au moment de la sortie des états correspondants offre aussi une simplification comparable. Partant de (8) et de la décomposition de A_j^L, nous pouvons écrire le nombre attendu de décès e_j^L dans S_j en présence d'erreurs de couplage sous la forme

$$\begin{aligned} e_j^L &= \sum_{i \in A_j^L} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du \\ &= \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du + \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du \\ &\qquad - \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du \\ &= e_j - \Delta e_j, \end{aligned} \tag{10}$$

où

$$e_j = \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du, \text{ et } \Delta e_j = e_j^P - e_j^N \tag{11}$$

avec

$$e_j^P = \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du \text{ et } e_j^N = \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du. \tag{12}$$

Pour simplifier la notation, écrivons $T_\lambda(i, j)$ pour $\int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du$ dans la suite. Le terme Δe_j représente le biais introduit par les erreurs de couplage dans le nombre attendu de décès dans le j^{e} état. Il découle de (10) et de (11) que les résultats faussement positifs ont tendance à réduire le nombre attendu de décès et que les résultats faussement négatifs ont tendance à l'augmenter.

En utilisant la décomposition de D_{jj}^L, nous pouvons écrire le nombre observé de décès d_{jj}^L en présence d'erreurs de couplage comme suit

$$d_{jj}^L = d_{jj} + \Delta d_{jj}, \tag{13}$$

où

$$\Delta d_{jj} = d_{jj}^P - a_{jj}^N, \tag{14}$$

avec d_{jj}, d_{jj}^P et a_{jj}^N le nombre d'individus dans les ensembles D_{jj}, D_{jj}^P et A_{jj}^N, respectivement. Le terme Δd_{jj} représente la variation du nombre observé de décès dans le j^{e} état due aux erreurs de couplage. Il découle de (13) et de (14) que les résultats faussement positifs font augmenter le nombre observé de décès et que les résultats faussement négatifs le réduisent.

Le statut vital est souvent déterminé par couplage des données sur la cohorte étudiée à celles de la BCDM, dont l'effectif est généralement beaucoup plus grand que celui de

la cohorte d'intérêt. Lorsque les enregistrements sur l'exposition d'une personne en vie sont associés incorrectement à ceux d'une personne décédée, cette dernière n'appartient habituellement pas à la cohorte. Donc, la contribution de personnes-années à risque de la personne qui demeure en vie cessera prématurément dans l'année du décès présumé; les personnes-années à risque perdues correspondent à la période écoulée de l'année du décès présumé jusqu'à la fin du suivi. Par ailleurs, si les enregistrements sur l'exposition d'un individu décédé sont associés incorrectement à ceux d'une personne en vie, la contribution de personnes-années à risque de cet individu inclura une période supplémentaire s'étendant de l'année réelle du décès jusqu'à la fin du suivi. Par conséquent, les résultats faussement positifs réduiront le nombre de personnes-années à risque dans la cohorte et les résultats faussement négatifs l'augmenteront.

3.2 Espérances et variances des différences dans les nombres observé et attendu de décès

L'effet des erreurs de couplage sur les nombres observé et attendu de décès dépend des taux de résultats faussement positifs et faussement négatifs. Soit p_j^P et p_j^N les taux de résultats faussement positifs et de résultats faussement négatif, respectivement, dans l'état S_j, pour $j = 1, ..., J$, que l'on suppose être constants dans S_j et les mêmes pour tous les individus dans A_j et D_j, respectivement. Cette hypothèse est raisonnable si les individus qui se trouvent dans le même état sont très homogènes, particulièrement en ce qui concerne des attributs tels que la qualité des identificateurs personnels, qui influent sur les taux d'erreurs de couplage. Bien que cette hypothèse idéaliste soit peu susceptible d'être entièrement satisfaite en pratique, elle simplifie considérablement l'évaluation subséquente des effets des erreurs de couplage. Formellement, p_j^P (p_j^N) est la probabilité conditionnelle qu'un individu compris dans $A_j (D_j)$ soit étiqueté comme étant décédé (en vie) dans l'état S_j. Autrement dit, $p_j^P = P[i \in D_j^P \mid i \in A_j]$ et $p_j^N = P[i \in A_j^N \mid i \in D_j]$.

Soit a_j, d_j, a_j^N et d_j^P le nombre d'individus dans A_j, D_j, A_j^N et D_j^P, respectivement. Alors, notons que d_j^P suit une loi binomiale (a_j, p_j^P) et que a_j^N suit une loi binomiale (d_j, p_j^N). En outre, d_{jj}^P suit une loi binomiale (a_j, p_{jj}^P), où p_{jj}^P est la probabilité conditionnelle qu'un individu compris dans A_j soit étiqueté comme étant décédé dans l'état S_j. Autrement dit, $p_{jj}^P = P[i \in D_{jj}^P \mid i \in A_j]$. De toute évidence, $p_{jj}^P \le p_j^P$. De la même façon, a_{jj}^N suit une loi binomiale (d_{jj}, p_{jj}^N), où p_{jj}^N est la probabilité conditionnelle qu'un individu compris dans D_{jj} soit étiqueté comme étant en vie dans l'état S_j. Autrement dit, $p_{jj}^N = P[i \in A_{jj}^N \mid i \in D_{jj}]$. Bien qu'il n'existe pas de relation sans importance entre p_j^N et p_{jj}^N en général, il est raisonnable de supposer que $p_j^N = p_{jj}^N$ dans le contexte des erreurs couplage.

En supposant que les erreurs de couplage associées à divers individus sont indépendantes, l'espérance et la variance de la différence dans le nombre observé de décès dans l'état S_j, donnée par Δd_{jj} dans (14), sont

$$E[\Delta d_{jj}] = E[d_{jj}^P] - E[a_{jj}^N] = a_j\, p_{jj}^P - d_{jj}\, p_j^N \qquad (15)$$

et

$$V[\Delta d_{jj}] = V[d_{jj}^P] + V[a_{jj}^N] = a_j\, p_{jj}^P (1 - p_{jj}^P) + d_{jj}\, p_j^N (1 - p_j^N). \qquad (16)$$

Puisque A_j et D_{jj} sont constitués d'ensembles différents d'individus, d_{jj}^P et a_{jj}^N sont indépendants.

De la même façon, l'espérance et la variance de la différence dans le nombre attendu de décès dans l'état S_j, donnée par Δe_j dans (11), peuvent être calculées comme suit. À cette fin, il est commode d'écrire e_j^P et e_j^N en fonction des variables indicatrices qui suivent. Pour $i \in A_j$, définissons $\xi_{ij} = I\{i \in D_j^P\}$ et $\xi_{ijj} = I\{i \in D_{jj}^P\}$. En outre, pour $i \in D_j$, définissons $\psi_{ij} = I\{i \in A_j^N\}$. Alors, il découle de (12) et des définitions de D_j^P et A_j^N que

$$e_j^P = \sum_{i \in A_j} \xi_{ij}\, T_\lambda(i,\ j) \qquad (17)$$

et

$$e_j^N = \sum_{i \in D_j} \psi_{ij}\, T_\lambda(i,\ j). \qquad (18)$$

En particulier, nous pouvons écrire $d_{jj}^P = \sum_{i \in A_j} \xi_{ijj}$ et $a_{jj}^N = \sum_{i \in D_{jj}} \psi_{ij}$, qui sont utiles pour établir (15) et (16). D'après (17) et (18), nous obtenons

$$E[\Delta e_j] = E[e_j^P] - E[e_j^N] = p_j^P \sum_{i \in A_j} T_\lambda(i,\ j) \ - \ p_j^N \sum_{i \in D_j} T_\lambda(i,\ j), \qquad (19)$$

et

$$V[\Delta e_j] = V[e_j^P] + V[e_j^N] = p_j^P (1 - p_j^P) \sum_{i \in A_j} T_\lambda^2(i,\ j) + p_j^N (1 - p_j^N) \sum_{i \in D_j} T_\lambda^2(i,\ j), \qquad (20)$$

puisque A_j et D_j sont constitués d'ensembles différents d'individus.

Les résultats (15)–(16) et (19)–(20) indiquent que les erreurs de couplage d'enregistrements introduisent un biais et une variation supplémentaire dans les nombres observé et attendu de décès. Minimiser les termes de variance dans (16) et (20) est difficile, puisque les deux taux d'erreurs p_j^P et p_j^N ne sont pas fonctionnellement indépendants. En

général, la diminution de p_j^P donnera lieu à une augmentation de p_j^N, et inversement (voir la section 5 pour une discussion plus approfondie de ce point). Bien que ces taux d'erreurs soient indépendants du modèle de régression du risque relatif sous-jacent γ donné par (2), l'erreur quadratique moyenne obtenue par combinaison des termes d'espérance et de variance ne peut être minimisée sans qu'on spécifie le risque de base $\lambda^*(u)$, qui figure dans T_λ.

4. L'effet des erreurs de couplage sur les estimations des RSM et des coefficients de régression

4.1 Ratios standardisés de mortalité

Pour déterminer l'effet des erreurs de couplage sur les RSM, nous remplaçons les nombres observé et attendu réels de décès d_{jj} et e_j par les nombres observé et attendu de décès en présence d'erreurs de couplage d_{jj}^L et e_j^L dans l'expression $\mathrm{RSM} = \sum d_{jj} / \sum e_j$. En représentant par RSM_L les ratios standardisés de mortalité en présence d'erreurs de couplage, nous obtenons

$$\mathrm{RSM}_L = \mathrm{RSM}\left[1+\frac{\sum \Delta d_{jj}}{\sum d_{jj}}\right] \Big/ \left[1-\frac{\sum \Delta e_j}{\sum e_j}\right]. \qquad (21)$$

Il découle des équations (10) à (14) que les résultats faussement positifs feront augmenter le RSM, tandis que les résultats faussement négatifs le feront diminuer.

En utilisant un développement en série de premier ordre de Taylor comme approximation de RSM_L autour de RSM, la différence $\Delta\mathrm{RSM} = \mathrm{RSM}_L - \mathrm{RSM}$ peut s'exprimer sous la forme

$$\frac{\Delta \mathrm{RSM}}{\mathrm{RSM}} = \frac{\sum_j \Delta d_{jj}}{\sum_j d_{jj}} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \qquad (22)$$

Alors, la moyenne et la variance de la différence relative de RSM peuvent être approximées par

$$E\left[\frac{\Delta \mathrm{RSM}}{\mathrm{RSM}}\right] \simeq \frac{\sum_j E[\Delta d_{jj}]}{\sum_j d_{jj}} + \frac{\sum_j E[\Delta e_j]}{\sum_j e_j} \qquad (23)$$

et

$$\begin{aligned} V\left[\frac{\Delta \mathrm{RSM}}{\mathrm{RSM}}\right] \simeq & \left(\sum_j d_{jj}\right)^{-2} V\left[\sum_j \Delta d_{jj}\right] \\ & + \left(\sum_j e_j\right)^{-2} V\left[\sum_j \Delta e_j\right] \\ & + 2\left(\sum_j d_{jj}\right)^{-1}\left(\sum_j e_j\right)^{-1} \mathrm{Cov}\left[\sum_j \Delta d_{jj}, \sum_j \Delta e_j\right], \end{aligned} \qquad (24)$$

respectivement. Il est facile de calculer le deuxième membre de (23) en utilisant (15) et (19). Pour calculer le deuxième membre de (24), notons que

$$\begin{aligned} V\left[\sum_j \Delta d_{jj}\right] = & \sum_j V[\Delta d_{jj}] \\ & + 2\sum_{j<j'} \mathrm{Cov}[\Delta d_{jj},\ \Delta d_{jj'}], \end{aligned} \qquad (25)$$

$$V\left[\sum_j \Delta e_j\right] = \sum_j V[\Delta e_j] + 2\sum_{j<j'} \mathrm{Cov}[\Delta e_j,\ \Delta e_{j'}], \qquad (26)$$

et

$$\begin{aligned} & \mathrm{Cov}\left[\sum_j \Delta d_{jj},\ \sum_j \Delta e_j\right] \\ & = \sum_j \mathrm{Cov}[\Delta d_{jj},\ \Delta e_j] + \sum_{j\neq j'} \mathrm{Cov}[\Delta d_{jj},\ \Delta e_{j'}]. \end{aligned} \qquad (27)$$

Sans perte de généralité, supposons, pour $j < j'$, que $t_{ij}^0 \le t_{ij'}^0$ pour le même individu i (en vie ou décédé) dans S_j et $S_{j'}$; autrement dit, le moment de l'entrée dans S_j est identique ou antérieur à celui de l'entrée dans $S_{j'}$. Nous avons alors, pour $j < j'$,

$$\begin{aligned} & \mathrm{Cov}[\Delta d_{jj}, \Delta d_{j'j'}] \\ & = -\left(\sum_{i\in A_j\cap A_{j'}} p_{jj}^P\, p_{j'j'}^P + \sum_{i\in A_j\cap D_{j'j'}} p_{jj}^P\, p_{j'}^N\right), \end{aligned} \qquad (28)$$

$$\begin{aligned} & \mathrm{Cov}[\Delta e_j,\ \Delta e_{j'}] \\ & = \sum_{i\in A_j\cap A_{j'}} p_j^P (1-p_{j'}^P)\, T_\lambda(i,\ j)\, T_\lambda(i,\ j') \\ & + \sum_{i\in A_j\cap D_{j'}} p_j^P\ p_{j'}^N\, T_\lambda(i,\ j)\, T_\lambda(i,\ j') \\ & + \sum_{i\in D_j\cap D_{j'}} p_{j'}^N (1-p_j^N)\, T_\lambda(i,\ j)\, T_\lambda(i,\ j'), \end{aligned} \qquad (29)$$

$$\begin{aligned} & \mathrm{Cov}[\Delta d_{jj},\ \Delta e_j] \\ & = \sum_{i\in A_j} p_{jj}^P (1-p_j^P)\, T_\lambda(i,\ j) \\ & + \sum_{i\in D_{jj}} p_j^N (1-p_j^N)\, T_\lambda(i,\ j), \end{aligned} \qquad (30)$$

$$\begin{aligned} & \mathrm{Cov}[\Delta d_{jj},\ \Delta e_{j'}] \\ & = \sum_{i\in A_j\cap A_{j'}} p_{jj}^P (1-p_{j'}^P)\, T_\lambda(i,\ j') \\ & + \sum_{i\in A_j\cap D_{j'}} p_{jj}^P\ p_{j'}^N\, T_\lambda(i,\ j') \\ & + \sum_{i\in D_{jj}\cap D_{j'}} p_{j'}^N (1-p_j^N)\, T_\lambda(i,\ j'), \end{aligned} \qquad \text{et (31)}$$

$$\begin{aligned} & \mathrm{Cov}[\Delta d_{j'j'},\ \Delta e_j] \\ & = -\sum_{i\in A_j\cap A_{j'}} p_j^P\ p_{j'j'}^P\, T_\lambda(i,\ j) \\ & + \sum_{i\in A_j\cap D_{j'j'}} p_j^P\ p_{j'}^N\, T_\lambda(i,\ j). \end{aligned} \qquad (32)$$

En utilisant les équations (25) à (32), nous pouvons approximer la variance de la différence relative $\Delta\text{RSM/RSM}$ au moyen du deuxième membre de (24). Nous pouvons tirer deux conclusions des équations (23) et (24). En premier lieu, les erreurs de couplage peuvent introduire un biais dans l'estimation du RSM. En deuxième lieu, les deux types d'erreurs de couplage introduisent une variation supplémentaire dans les estimations du RSM. Notons que le premier terme de (32) est dominé par le premier terme de (29) pour $p_{j'}^{P} < 0{,}5$, et que le terme de covariance négatif (28) est dominé dans le calcul de la variance dans (25). Par conséquent, la variance supplémentaire (24) est strictement positive, puisque les taux de résultats faussement positifs et de résultats faussement négatifs sont tous deux positifs.

4.2 Paramètres de régression du risque relatif

Pour déterminer l'effet des erreurs de couplage sur les estimations des paramètres de régression, considérons d'abord le modèle général de régression du risque relatif (2). En remplaçant dans la fonction de log-vraisemblance (7) les nombres observé et attendu de décès d_{jj} et e_j par les nombres observé et attendu de décès en présence d'erreurs de couplage d_{jj}^{L} et e_j^{L}, nous obtenons

$$\log L = \sum_{j=1}^{J} \{d_{jj}^{L} \log(\gamma\{\beta' \mathbf{z}_j\}) - \gamma\{\beta' \mathbf{z}_j\} e_j^{L}\}. \qquad (33)$$

Soit $\hat{\beta}$ et $\tilde{\beta}$ les estimations du maximum de vraisemblance de β fondées sur $\{d_{jj}, e_j\}$ et $\{d_{jj}^{L}, e_j^{L}\}$, respectivement. L'équation de score (9) peut s'écrire sous la forme

$$\sum_{j=1}^{J} \frac{\partial \Lambda_j(\tilde{\beta})}{\partial \beta} [d_{jj} + \Delta d_{jj} - \exp\{\Lambda_j(\tilde{\beta})\}(e_j - \Delta e_j)] = 0. \qquad (34)$$

En supposant que $\Delta\beta = \tilde{\beta} - \hat{\beta}$ est faible, un développement en série de premier ordre de $\exp\{\Lambda_j(\tilde{\beta})\}$ autour de $\hat{\beta}$ donne

$$\exp\{\Lambda_j(\tilde{\beta})\} \simeq \exp\{\hat{\Lambda}_j\} + \exp\{\hat{\Lambda}_j\} \frac{\partial \hat{\Lambda}_j}{\partial \beta} \Delta\beta, \qquad (35)$$

où $\hat{\Lambda}_j = \Lambda_j(\hat{\beta})$ et $\partial\hat{\Lambda}_j / \partial\beta$ est $\partial\Lambda_j / \partial\beta$ évalué à $\beta = \hat{\beta}$. En introduisant (35) par substitution dans (34), nous obtenons

$$\sum_{j=1}^{J} \frac{\partial \hat{\Lambda}_j}{\partial \beta} [d_{jj} - \exp\{\hat{\Lambda}_j\} e_j] + \sum_{j=1}^{J} \frac{\partial \hat{\Lambda}_j}{\partial \beta} \begin{bmatrix} \Delta d_{jj} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \\ - \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \Delta\beta \\ + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \Delta\beta \end{bmatrix} \simeq 0. \qquad (36)$$

En utilisant (9), la première somme dans (36) est nulle. Par conséquent, puisque $\Delta e_j \Delta\beta$ est faible, $\Delta\beta$ peut être approximé par

$$\Delta\beta \simeq \left(\sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \{\Delta d_{jj} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j\}. \qquad (37)$$

Il découle de (37) que

$$E[\Delta\beta] \simeq \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \Lambda_j}{\partial \beta} \alpha_j, \qquad (38)$$

où $\alpha_j = E[\Delta d_{jj}] + \gamma\{\hat{\beta}' \mathbf{z}_j\} E[\Delta e_j]$, qui peut être calculé d'après (15) et (19). En outre,

$$V[\Delta\beta] \simeq \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \left(\sum_j \sum_{j'} \frac{\partial \Lambda_j}{\partial \beta} \Theta_{jj'} \frac{\partial \Lambda_{j'}'}{\partial \beta} \right) \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \qquad (39)$$

avec

$$\Theta_{jj'} = \text{Cov}[\Delta d_{jj} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j,\ \Delta d_{j'j'} + \gamma\{\hat{\beta}' \mathbf{z}_{j'}\} \Delta e_{j'}],$$

qui peut aussi être obtenu facilement en utilisant (16), (20) et (28) à (32).

Dans le cas particulier du modèle multiplicatif de risque (4), la différence $\Delta\beta$ due aux erreurs de couplage peut être approximée par

$$\Delta\beta \simeq (X'W\,X)^{-1} X'(\Delta D + \Delta W), \qquad (40)$$

où $X' = (\mathbf{z}_1', \ldots, \mathbf{z}_J')$, $\Delta D' = (\Delta d_{11}, \ldots, \Delta d_{JJ})$, $W = \text{diag}(\exp(\mathbf{z}_1'\hat{\beta})\, e_1, \ldots, \exp(\mathbf{z}_J'\hat{\beta})\, e_J)$, et $\Delta W' = (\exp(\mathbf{z}_1'\hat{\beta})\, \Delta e_1, \ldots, \exp(\mathbf{z}_J'\hat{\beta})\, \Delta e_J)$. Notons que la matrice de poids W est la matrice d'information de Fisher pour $\hat{\beta}$. Il découle de (38) que

$$E[\Delta\beta] \simeq (X'W\,X)^{-1} X'\,\Pi, \qquad (41)$$

où $\Pi' = (\pi_1, \ldots, \pi_J)$ avec π_j identique à α_j, mais $\gamma\{\hat{\beta}' \mathbf{z}_j\}$ remplacé par $\exp(\mathbf{z}_j'\hat{\beta})$.

En outre,

$$V[\Delta\beta] \simeq (X'W\,X)^{-1} X'\,\Psi\, X(X'W\,X)^{-1}, \qquad (42)$$

où Ψ est la matrice des $\Theta_{jj'}$ avec $\gamma\{\hat{\beta}' \mathbf{z}_j\}$ remplacé par $\exp(\mathbf{z}_j'\hat{\beta})$. Notons que les expressions (40) à (42) sont des cas particuliers des expressions (37) à (39), respectivement, écrites en notation matricielle.

Avec une seule covariable $z_i = 1$, $X'W\,X = e^{\hat{\beta}} \sum_j e_j$, $X'\Delta D = \sum_j d_{jj}$ et $X'\Delta W = e^{\hat{\beta}} \sum_j \Delta e_j$. Dans ce cas,

$$\Delta\beta \simeq \frac{\sum_j \Delta d_{jj} + e^{\hat{\beta}} \sum_j \Delta e_j}{e^{\hat{\beta}} \sum_j e_j}. \qquad (43)$$

Puisque le $\mathrm{RSM} = e^{\hat{\beta}} = \sum_j d_{jj} / \sum_j e_j$, avec $\Delta\beta = \Delta\mathrm{RSM}/\mathrm{RSM}$ ici, nous obtenons

$$\Delta\beta \simeq \frac{\sum_j \Delta d_{jj}}{\sum_j d_{jj}} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \qquad (44)$$

Donc, l'expression (44) peut être considérée comme un cas particulier de (22).

Les résultats qui précédent indiquent que les résultats faussement positifs ainsi que faussement négatifs introduisent un biais et une variation supplémentaire dans les estimations des paramètres de régression du risque relatif. La seule contribution négative à cette variance supplémentaire (39) a lieu par la voie de $\mathrm{Cov}[\Delta d_{jj}, \Delta d_{j'j'}]$, donné par (28), et du premier terme de (32) (voir $\Theta_{jj'}$). En utilisant le même argument qu'à la section 4.1, il s'ensuit que cette variance supplémentaire est strictement positive.

5. Conclusion

Le couplage d'enregistrements est maintenant une technique bien établie dans le contexte des études épidémiologiques des risques pour la santé des populations. En couplant l'information sur les expositions des individus provenant d'une base de données à celles sur les résultats en ce qui concerne la santé provenant d'une autre base de données, il est possible de construire de grandes bases de données informatives sur les risques que courent les populations et les sous-groupes de population. Le succès de ce genre d'études dépend, en grande partie, de la qualité des deux bases de données que l'on couple, y compris la quantité d'information sur les identificateurs personnels utilisés pour coupler les individus représentés dans les deux bases de données. Dans la plupart des études, l'exactitude du couplage est examinée en estimant les taux de faux couplages (résultats faussement positifs) et de faux non-couplages (résultats faussement négatifs) associés au processus de couplage. En pratique, on procède habituellement au tirage d'un échantillon d'enregistrements couplés et non couplés, puis on détermine l'exactitude des couplages dans l'échantillon en se servant de données auxiliaires provenant d'autres sources.

Bien que le CIE soit utilisé depuis un certain temps dans les études-cohorte de mortalité, l'effet des erreurs de couplage sur la fiabilité des inférences statistiques faites d'après ce genre d'études n'a pas fait l'objet d'un examen détaillé. Les résultats théoriques présentés dans le présent article visent à combler cette lacune. Ces résultats montrent qu'en plus d'accroître le nombre observé de décès, les résultats faussement positifs ont tendance à réduire le nombre attendu de décès. Inversement, les résultats faussement négatifs accroissent le nombre attendu de décès et réduisent le nombre observé de décès. Nous avons montré que les erreurs de couplage introduisent un biais dans les estimations des RSM. Les estimations des coefficients de régression du risque relatif sont également entachées d'un biais, dont la direction dépend de la nature du coefficient de régression. En plus de ces biais, les erreurs de couplage introduisent une incertitude additionnelle dans les estimations des RSM, ainsi que des coefficients de régression.

Bien que nous émettions l'hypothèse simplificatrice que $t_{ij}^1 = t_{ij}^0$, il est possible d'établir les expressions pertinentes du biais et de la variabilité supplémentaire sans le faire; cependant, les expressions sont trop complexes pour fournir des éclaircissements supplémentaires sur les effets des erreurs de couplage. Il en est également ainsi de l'hypothèse selon laquelle $p_{jj}^N = p_j^N$. La définition de A_j pour le ou les états correspondant à la dernière tranche d'âge, qui est habituellement ouverte jusqu'à l'infini du côté droit, pose un problème technique. Dans ces états, l'hypothèse que $t_{ij}^1 = t_{ij}^0$ est problématique si la probabilité de mourir dans cette dernière tranche d'âge est appréciable. On peut contourner le problème en supposant que la durée de vie humaine a une limite supérieure finie.

Comme nous en discutons à la section 3.1, les résultats faussement positifs surviennent principalement lorsqu'un individu en vie à la fin de la période de suivi est couplé incorrectement à une personne décédée. Cependant, une personne décédée dans l'un des états S_j peut être couplée incorrectement à une autre personne décédée à une période antérieure, ce qui donne un résultat faussement positif qui persiste jusqu'au moment réel du décès; l'analyse de la section 3 tient compte de ce genre d'erreur. De même, une personne décédée peut être couplée incorrectement à une autre personne décédée à une date ultérieure, qui n'est pas en vie à la fin de la période de suivi. Ce cas est traité comme un résultat faussement négatif uniquement jusqu'au moment incorrect du décès. À ce moment-là, il y aura une contribution incorrecte au nombre de décès, erreur qui n'a pas été prise en compte à la section 3. Toutefois, ce genre d'erreur ne serait normalement pas décelé dans les études par couplage d'enregistrements habituelles dans lesquelles on procède à une vérification manuelle simplifiée pour repérer les résultats faussement positifs et faussement négatifs. Puisque ce genre d'erreurs est vraisemblablement rare, nous nous attendons à ce que son effet soit faible.

Afin d'étudier plus en détail l'effet éventuel des erreurs de couplage d'enregistrements, supposons que τ_j est la limite d'âge supérieure pour le j^{e} état S_j. (Notons que certains τ_j peuvent être égaux.) Alors, si nous représentons

par α la probabilité d'une erreur de couplage (de l'un ou l'autre type), nous pouvons écrire les taux de résultats faussement positifs et de résultats faussement négatifs, p_j^P et p_j^N, sous la forme $\alpha P[T \le \tau_j]$ et $\alpha P[T > \tau_j]$, respectivement. En particulier, $p_{jj}^P = \alpha P[\tau_{j-1} < T \le \tau_j]$, où τ_{j-1} est la limite inférieure d'âge pour le j^{e} état, et $p_{jj}^N = p_j^N$. Par conséquent, les taux de résultats faussement positifs peuvent être supérieurs aux taux de résultats faussement négatifs pour les groupes d'âge avancé, l'inverse se produisant pour les groupes d'âge plus jeune. Si l'on suppose que le profil de taille est le même pour les D_j et A_j, certains termes s'annulent dans le calcul de $E[\Delta e_j]$ dans (19) et dans celui de $E[\Delta d_{jj}]$ dans (15). Cet effet d'annulation réduira les biais attendus dans le RSM et dans les paramètres de régression du risque donnés par (23) et (38), respectivement.

Bien que nous ayons considéré uniquement la mortalité toutes causes confondues dans le présent article, la mortalité par cause peut être étudiée en apportant des modifications simples aux définitions de D_{jj}, D_{jj}^L et D_{jj}^P. Ces ensembles devraient alors ne tenir compte que des décès dus à la cause particulière étudiée. Par conséquent, d_{jj} et e_j devraient représenter, respectivement, les nombres observé et attendu de décès du type spécifié dans S_j. Dans (1) et (2), la fonction de risque devrait avoir trait au type spécifique de décès, avec $\lambda^*(u)$ le taux de risque par cause de base correspondant. Enfin, à la section 2, l'indicateur δ_i devrait indiquer le type spécifique de décès.

Les résultats analytiques qui précèdent fournissent d'importants éclaircissements sur les effets des erreurs de couplage dans les études-cohorte de la mortalité, mais il est important d'examiner ce genre d'effets dans des conditions aussi proches que possible de celles rencontrées en pratique. À cette fin, nous avons réalisé une étude en simulation informatisée fondée sur des données réelles provenant du Fichier dosimétrique national du Canada, dans laquelle nous avons introduit des couplages incorrects et des non-couplages incorrects avec probabilités connues pour évaluer plus en profondeur l'effet des erreurs de couplage sur les estimations du risque de cancer (Mallick, Krewski, Dewanji et Zielinski 2002). Les résultats de cette simulation corroborent les résultats théoriques exposés dans l'article.

Alors que les résultats présentés ici permettent de mieux comprendre l'effet des erreurs de couplage sur l'inférence statistique, des méthodes tenant compte de ce genre d'erreurs dans les analyses statistiques n'ont pas encore été élaborées. Ces méthodes pourraient s'inspirer des modèles d'erreur de réponse utilisés dans le domaine du sondage, conjugués aux méthodes statistiques classiques d'analyse des données sur la mortalité des cohortes. Des travaux de recherche dans ce domaine sont en cours.

6. Remerciements

La présente étude a été financée en partie par une bourse du Conseil national de recherches en sciences et en génie du Canada octroyée à D. Krewski, qui est titulaire à l'heure actuelle de la chaire CRSNG-CRHS-McLaughlin d'évaluation du risque pour la santé des populations à l'Université d'Ottawa. Des versions préliminaires du présent article ont été présentées à l'Annual Joint Meeting de l'American Statistical Association qui s'est tenue à San Francisco du 8 au 12 août 1993 et à l'Assemblée annuelle de la Société statistique du Canada qui s'est tenue à Montréal du 10 au 16 juillet 1995. La version finale a été présentée à la session dédiée à J.N.K. Rao du Symposium 2001 de Statistique Canada qui a eu lieu à Ottawa le 18 octobre 2001. L'auteur principal (D. Krewski) est particulièrement reconnaissant d'avoir été invité à prendre la parole à la session en l'honneur de J.N.K. Rao, qui avait été son directeur de thèse de doctorat il y a de nombreuses années. L'étude a été achevée pendant les séjours de A. Dewanji au Centre McLaughlin d'évaluation du risque pour la santé des populations à titre de chercheur invité durant les étés de 2002 et de 2003.

Bibliographie

Anderson, T.W. (1974). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, Inc.

Ardal, S., et Ennis, S. (2001). Enquêtes sur les données : Mise en évidence d'erreurs systématiques dans les bases de données administratives. *Recueil : Symposium 2001, La qualité des données d'un organisme statistique : Une perspective méthodologique*, Statistique Canada, Ottawa.

Ashmore, J.-P., et Grogan, D. (1985). The national dose registry of Canada. *Radiation Protection Dosimetry*, 11, 95-100.

Ashmore, J.-P., et Davies, B.D. (1989). The national dose registry: A centralized record keeping system for radiation workers in Canada. Dans *Applications of Computer Technology to Radiation Protection*, IAEA-SR-136/58, J. Stephan Institute, Ljublyua, 505-520.

Ashmore, J.-P., Krewski, D. and Zielinski, J.M. (1997). Protocol for a cohort mortality study of occupational radiation exposure based on the national dose registry of Canada. *European Journal of Cancer*, 33, S10-S21.

Ashmore, J.-P., Krewski, D., Zielinski, J.M., Jiang, H., Semenciw, R. et Létourneau, E. (1998). First analysis of occupational radiation mortality based on the national dose registry of Canada. *American Journal of Epidemiology*, 148, 564-574.

Bartlett, S., Krewski, D., Wang, Y. et Zielinski, J.M. (1993). Évaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisé. *Techniques d'enquête*, 19, 3-13.

Belin, T.R., et Rubin, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.

Breslow, N.E., Lubin, J.H. et Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, 78, 1-12.

Breslow, N.E., et Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. 2 : *The Design and Analysis of Cohort Studies*. IARC scientific publication No. 82, international agency for research on cancer, Lyon, France.

Carpenter, M., et Fair, M.E. (Eds.) (1990). *Canadian Epidemiology Research Conference – 1989: Proceedings of Record Linkage Sessions & Workshop*. Ottawa Select Printing, Ottawa.

Cox, D.R. (1972). Regression models and life tables (avec discussion). *Journal of Royal Statistical Society*, B, 34, 187-220.

Fair, M.E. (1989). Studies and References Relating to Uses of the Canadian Mortality Data Base. Report from the occupational and environmental health research unit, Division de la Santé, Statistique Canada, Ottawa.

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

Hill, T. (1988). Generalized Iterative Record Linkage System: GIRLS Strategy (Relâcher 2.7). Report from research and general system, informatics services and development division, Statistique Canada, Ottawa.

Howe, G.R., et Lindsay, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research*, 14, 327-340.

Howe, G.R., et Spasoff, R.A. (Eds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research*. University of Toronto Press, Toronto.

Jordan-Simpson, D.A., Fair, M.E. et Poliquin, C. (1990). Étude des exploitants agricoles canadians : Méthodologie. *Rapports sur la santé*, 2, 141-155.

Kalbfleish, J.D., et Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons, Inc.

Labossière, G. (1986). Confidentiality and access to data: The practice at Statistics Canada. *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, University of Toronto Press, Toronto.

Mallick, R., Krewski, D., Dewanji, A. et Zielinski, J.M. (2002). A simulation study of the effect of record linkage errors in cohort mortality data. *Proceedings of International Conference in Recent Advances in Survey Sampling*. Carleton University, Ottawa, à paraître.

Neter, J., Maynes, E.S. et Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.

Newcombe, H.B. (1988). *Handbook of Record Linkage*: *Methods for Health and Statistical Studies, Administration, and Business. Oxford Medical Publications*. Oxford.

Roos, L.L., Soodeen, R. et Jebamani, L. (2001). Un environnement riche en information : La qualité des données des systèmes d'appariement de dossiers au Canada. *Recueil* : *Symposium 2001*, *La qualité des données d'un organisme statistique* : *Une perspective méthodologique*, Statistique Canada, Ottawa.

Scheuren, F., et Winkler, W.E. (1993). Analyse de regression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.

Scheuren, F., et Winkler, W.E. (1997). Analyse de régression des fichiers de données appariés par ordinateur – Partie II. *Techniques d'enquête*, 23, 171-180.

Singh, A.C., Feder, M., Dunteman, G. et Yu, F. (2001). Protection de la confidentialité et maintien de la qualité des microdonnées à grande diffusion. *Recueil* : *Symposium 2001*, *La qualité des données d'un organisme statistique* : *Une perspective méthodologique*, Statistique Canada, Ottawa.

Smith, M.E., et Silins, J. (1981). Generalized iterative record linkage system. *Social Statistics Section*, *Proceedings of the American Statistical Association*, 128-137.

Sont, W.N., Zielinski, J.M., Ashmore, J.P., Jiang, H., Krewski, D., Fair, M.E., Band, P. et Létourneau, E. (2001). First analysis of cancer incidence and occupational radiation exposure based on the national dose registry of Canada. *American Journal of Epidemiology*, 153, 309-318.

Winkler, W.E., et Scheuren, F. (1991). How computer matching error effect regression analysis: Exploratory and confirmatory analysis. Rapport technique, Statistical research division, U.S. Bureau of the Census, Washington, D.C.

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 25-44
Statistique Canada, N° 12-001-XIF au catalogue

Analyse d'expériences intégrées dans des plans de sondage complexes

Jan A. van den Brakel et Robbert H. Renssen [1]

Résumé

Les instituts nationaux de statistique intègrent parfois des expériences dans les enquêtes par sondage courantes afin d'étudier les effets éventuels de diverses techniques d'enquête sur les estimations des paramètres d'une population finie. En vue de tester les hypothèses au sujet des différences entre les estimations par sondage obtenues on applique diverses variantes de l'enquête, nous élaborons une théorie fondée sur le plan de sondage pour analyser des plans en randomisation totale ou des plans en blocs randomisés intégrés dans des plans de sondage complexes généraux. Pour ces deux types de plans d'expérience, nous établissons une statistique de Wald fondée sur le plan de sondage pour l'estimateur d'Horvitz-Thompson et pour l'estimateur par la régression généralisée. Enfin, nous illustrons la théorie au moyen d'une étude en simulation.

Mots clés : Analyse fondée sur le plan de sondage; modèles de l'erreur de mesure; échantillonnage probabiliste; expériences randomisées; superposition.

1. Introduction

Une part de la recherche réalisée dans le domaine de la méthodologie d'enquête consiste à considérer et à évaluer des techniques d'enquête de rechange, en vue d'améliorer la qualité et l'efficacité des processus d'enquête par sondage mis en place par les instituts nationaux de statistique. L'intégration d'expériences à grande échelle sur le terrain dans les enquêtes par sondage courantes convient particulièrement bien pour quantifier l'effet de diverses mises en œuvre d'une enquête sur le comportement de réponse ou sur les estimations des paramètres de population finie. Ainsi, Statistique Pays-Bas a étudié les effets de diverses conceptions de questionnaire, diverses stratégies d'approche ou diverses lettres préalables à l'enquête sur les deux types de paramètres. Consulter à cet égard Van den Brakel et Renssen (1998), Van den Brakel (2001), ainsi que Van den Brakel et Van Berkel (2002). Les instituts nationaux de statistique évitent généralement de modifier les enquêtes par sondage aussi longtemps que possible afin de produire des séries chronologiques ininterrompues d'estimations des paramètres de population. Toutefois, il est immanquable qu'ils doivent rajuster les processus d'enquête de temps en temps. Des expériences intégrées peuvent être utilisées pour déceler et quantifier les ruptures de tendance que peuvent causer dans ces séries chronologiques les changements qu'il faut apporter à une enquête par sondage et pour assurer une transition harmonieuse de l'ancien au nouveau plan de sondage. L'exécution en parallèle de l'ancienne et de la nouvelle enquête au moyen d'une expérience intégrée donne la possibilité de retourner à l'ancienne approche aux fins des publications courantes si la nouvelle s'avère être un échec.

Les applications d'expériences intégrées décrites dans la littérature avaient pour but d'estimer le biais ou les diverses composantes de la variance dans les modèles de l'erreur de mesure totale. Mahalanobis (1946) a probablement été le premier à lancer l'idée d'intégrer des expériences dans les enquêtes par sondage courantes, sous forme de sous-échantillonnage superposé, pour tester les différences entre intervieweurs sous échantillonnage aléatoire simple et randomisation non contrainte des unités d'échantillonnage entre les intervieweurs. Fellegi (1964), ainsi que Hartley et Rao (1978) ont généralisé cette approche pour estimer les variances de réponse sous des plans de sondage plus complexes et la randomisation contrainte des unités d'échantillonnage. Fienberg et Tanur (1987, 1988, 1989) discutent des dissemblances et des parallèles entre la théorie des plans expérimentaux et celle de l'échantillonnage en population finie, et de la façon utile et naturelle dont les méthodes statistiques appliquées dans les deux domaines peuvent être combinées pour concevoir et analyser des expériences intégrées. Leur article de 1988 donne un aperçu exhaustif des applications d'expériences intégrées mentionnées dans la littérature.

La situation type considérée dans le présent article est une expérience sur le terrain conçue pour comparer les effets de K mises en œuvre différentes d'une enquête, c'est-à-dire les traitements, sur les estimations des principaux paramètres de population finie d'une enquête courante. À cette fin, un échantillon probabiliste tiré d'une population finie est subdivisé aléatoirement en K sous-échantillons conformément à un plan d'expérience. Chaque sous-échantillon est assigné à l'un des K traitements. Les plans d'expérience considérés sont les plans en randomisation totale (PRT) et les plans en blocs randomisés (PBR), où les

1. Jan A. van den Brakel et Robbert H. Renssen, Statistics Netherlands, Department of Statistical Methods, P.O. Box 4481, 6401 CZ Heerlen, Pays-Bas.

structures d'échantillonnage, comme les strates, les unités primaires d'échantillonnage (UPE), les grappes ou les intervieweurs sont des variables de bloc éventuelles. En général, on assigne à l'enquête courante un grand sous-échantillon qui est utilisé pour la production des publications officielles et sert simultanément de groupe témoin dans l'expérience. L'objectif des expériences intégrées est d'estimer les paramètres de population finie sous les diverses mises en œuvre de l'enquête et de tester les hypothèses au sujet des écarts entre les diverses estimations ainsi obtenues de ces paramètres.

Au premier abord, on pourrait considérer une approche basée sur un modèle classique pour cette analyse. Cependant, puisque les unités expérimentales sont tirées selon un plan d'échantillonnage complexe sans remise à partir d'une population finie, l'application d'une telle approche risque de produire des estimations des paramètres et des variances biaisées par rapport au plan de sondage. Les résultats de l'analyse ne pourraient alors pas être comparés aux estimations des paramètres et des variances de l'enquête ordinaire, ce qui compliquerait l'interprétation des résultats sous les conditions du plan de l'enquête par sondage. Pour rendre l'analyse plus robuste aux écarts par rapport au modèle hypothétique, il faudrait adopter une approche basée sur le plan de sondage qui tient compte de ce dernier.

Avant de présenter notre approche basée sur le plan de sondage, nous mentionnons deux autres options qui, à première vue, semblent correctes. Toutefois, nous argumentons brièvement du fait que ces deux options produisent généralement des résultats invalides. La première est une analyse par régression linéaire fondée sur le plan de sondage qui tient compte du plan d'échantillonnage pour estimer les effets des K traitements introduits dans le modèle de régression et tester les hypothèses à leur sujet. Cependant, cette approche produit facilement des estimations incorrectes des variances par rapport au plan, puisqu'on ignore la randomisation du plan d'expérience. L'objectif analytique principal des expériences intégrées est de comparer les effets de diverses stratégies d'enquête sur les estimations principales produites d'après l'enquête par sondage courante. Or, une analyse par régression linéaire ne permet pas précisément de réaliser cet objectif, puisque, dans le modèle de régression, les effets de traitement ne sont généralement pas égaux aux écarts entre les estimations sur les sous-échantillons.

La deuxième option consiste à faire à une inférence basée sur le plan de sondage pour comparer les paramètres de domaine, où les K traitements sont considérés comme K domaines. Toutefois, l'objectif d'une expérience intégrée est de comparer les estimations du même paramètre sous diverses stratégies d'enquête, ou traitements, alors que, dans le cas des paramètres de domaine, l'objectif est de comparer les estimations de divers paramètres de population sous essentiellement la même stratégie d'enquête.

L'approche présentée dans le présent article se résume comme suit. En nous fondant sur les K sous-échantillons, nous établissons un estimateur basé sur le plan de sondage du paramètre de population observé sous chacun des K traitements, ainsi qu'un estimateur fondé sur le plan de sondage de la matrice des covariances des $K-1$ contrastes entre ces estimations. Cette méthode d'estimation tient compte de la structure probabiliste du plan d'échantillonnage, de la randomisation des unités d'échantillonnage sur les traitements conformément au plan d'expérience et de la procédure de pondération appliquée dans l'enquête courante pour l'estimation des paramètres cibles. Nous obtenons ainsi une statistique de Wald basée sur le plan de sondage pour tester les hypothèses au sujet des écarts entre les estimations par sondage.

La contribution principale du présent article est d'offrir un cadre général pour la comparaison de K approches d'enquête dans la situation réaliste d'un vrai processus d'enquête par sondage. La sélection aléatoire des unités d'échantillonnage à partir d'une population finie cible selon une méthode d'échantillonnage probabiliste est combinée à la randomisation des unités d'échantillonnage sur les divers traitements conformément à un plan d'expérience. Cette façon de procéder facilite la comparaison des effets de diverses approches d'enquête sur les résultats principaux d'une enquête par sondage, ainsi que la généralisation des résultats observés à des populations plus grandes que l'échantillon inclus dans l'expérience. La méthode d'analyse proposée ici généralise l'analyse des expériences à deux traitements intégrés dans les enquêtes par sondage (Van den Brakel et Renssen (1998) et Van den Brakel et Van Berkel (2002)) en l'étendant aux plans en randomisation totale (PRT) et aux plans en blocs randomisés (PBR) avec $K > 2$ traitements. Un résultat important est que l'estimateur fondé sur le plan de sondage de la matrice des covariances des contrastes entre les estimations sur sous-échantillons possède une structure assez simple, comme si les unités d'échantillonnage étaient tirées avec remise et probabilités de sélection inégales. Par conséquent, la procédure d'estimation de la variance ne nécessite ni les probabilités d'inclusion conjointes ni les covariances fondées sur le plan de sondage entre les estimations sur sous-échantillons, ce qui produit une méthode d'analyse séduisante et assez simple. Un deuxième avantage est que cette méthode permet de tester les hypothèses au sujet des différences entre les estimations par sondage de l'enquête, ce qui facilite l'interprétation des résultats d'analyse dans de nombreuses applications.

À la section 2, nous présentons une théorie fondée sur le plan de sondage de l'analyse des expériences intégrées. À la

section 3, nous expliquons plus en détail pourquoi l'analyse par régression linéaire fondée sur le plan de sondage convient moins bien. À la section 4, nous évaluons la méthode d'analyse fondée sur le plan de sondage proposée au moyen d'une étude en simulation. Enfin, à la section 5, nous résumons nos conclusions.

2. Analyse des expériences intégrées

2.1 Modèles de l'erreur de mesure

Bien que la méthode d'analyse des expériences intégrées proposée à la présente section soit fondée sur le plan de sondage, elle comporte l'utilisation de modèles de l'erreur de mesure. L'évaluation des effets systématiques de diverses techniques d'enquête sur les résultats d'une enquête implique l'existence d'erreurs de mesure. La notion classique selon laquelle les observations obtenues auprès des unités d'échantillonnage sont des valeurs fixes réelles, dépourvues d'erreur, hypothèse qui est généralement faite en théorie de l'échantillonnage fondée sur le plan de sondage, n'est pas défendable dans ce genre de situation. Par conséquent, on spécifie un modèle d'erreur de mesure pour les observations obtenues sous les diverses mises en œuvre de l'enquête, ou traitements de l'expérience. Ce modèle relie les effets de traitement aux écarts systématiques entre les valeurs des paramètres de population finie.

Considérons une population finie U de N individus. Soit y_{ikl} la réponse éventuelle du i^{e} individu $(i = 1, 2, \ldots, N)$ observée au moyen du k^{e} traitement $(k = 1, 2, \ldots, K)$ et du l^{e} intervieweur $(l = 1, 2, \ldots, L)$. Nous supposons que ces observations sont une réalisation du modèle de l'erreur de mesure $y_{ikl} = u_i + \beta_k + \psi_{il} + \varepsilon_{ik}$. Ici, u_i est la valeur réelle, intrinsèque pour le i^{e} individu, β_k est l'effet du k^{e} traitement, ψ_{il} est l'effet du l^{e} intervieweur sur le i^{e} individu et ε_{ik} est une composante de l'erreur du i^{e} individu observé au moyen du k^{e} traitement. L'effet d'intervieweur ψ_{il} tient compte de la mise en grappes et de la corrélation systématique entre les réponses des individus assignés à un même intervieweur à cause des effets d'intervieweur fixes et aléatoires, c'est-à-dire $\psi_{il} = \psi_l + \xi_l$, avec ψ_l l'effet fixe et ξ_l l'effet aléatoire du l^{e} intervieweur. Outre les intervieweurs, des facteurs courants, tels que les codeurs et les superviseurs, pourraient induire une corrélation entre les réponses des individus.

Puisque, pour chaque unité d'échantillonnage, nous définissons une variable de réponse éventuelle pour chacun des K traitements, nous pouvons exprimer le modèle de l'erreur de mesure en notation matricielle sous la forme

$$\mathbf{y}_{il} = \mathbf{j}u_i + \boldsymbol{\beta} + \mathbf{j}\psi_{il} + \boldsymbol{\varepsilon}_i, \tag{1}$$

où $\mathbf{y}_{il} = (y_{i1l}, \ldots, y_{iKl})^t$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^t$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iK})^t$ et $\mathbf{j} = (1, \ldots, 1)^t$. Soit E_m et Cov_m l'espérance et la covariance par rapport au modèle de l'erreur de mesure. Nous formulons au sujet du modèle les hypothèses suivantes :

$$\mathrm{E}_m(\boldsymbol{\varepsilon}_i) = \mathbf{0}, \tag{2}$$

$$\mathrm{Cov}_m(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}^t) = \begin{cases} \boldsymbol{\Sigma}_i : & i = i' \\ \mathbf{O} : & i \neq i' \end{cases}, \tag{3}$$

$$\mathrm{E}_m(\xi_l) = 0, \tag{4}$$

$$\mathrm{Cov}_m(\xi_l, \xi_{l'}) = \begin{cases} \tau_l^2 : & l = l' \\ 0 : & l \neq l' \end{cases}, \tag{5}$$

$$\mathrm{Cov}_m(\varepsilon_{ik}, \xi_l) = 0, \tag{6}$$

où $\mathbf{0}$ est un vecteur de dimension K dont chaque élément est nul et $\mathbf{O}$ est une matrice de dimensions $K \times K$ dont chaque élément est nul. Si $\psi_l = 0$, alors nous obtenons un modèle ne contenant que des effets d'intervieweur aléatoires. Si $\tau_l^2 = 0$, alors nous obtenons un modèle ne contenant que des effets d'intervieweur fixes. Il découle des hypothèses que

$$\mathrm{E}_m(\mathbf{y}_{il}) = \mathbf{j}u_i + \mathbf{j}\psi_l + \boldsymbol{\beta}, \tag{7}$$

et

$$\mathrm{Cov}_m(\mathbf{y}_{il}, \mathbf{y}_{i'l'}^t) = \begin{cases} \boldsymbol{\Sigma}_i + \mathbf{j}\mathbf{j}^t\,\tau_l^2 : & i = i' \quad \text{et} \quad l = l' \\ \mathbf{j}\mathbf{j}^t\,\tau_l^2 : & i \neq i' \quad \text{et} \quad l = l'. \\ \mathbf{O} : & i \neq i' \quad \text{et} \quad l \neq l' \end{cases} \tag{8}$$

Toute corrélation entre les réponses d'individus différents peut être modélisée au moyen d'effets d'intervieweur aléatoires. Tout effet fixe d'intervieweur influence les valeurs de réponse attendues. Dans la suite, afin de simplifier la notation, nous omettrons l'indice inférieur l dans y_{ikl} et $\mathbf{y}_{il}$.

2.2 Vérification des hypothèses

Le modèle de l'erreur de mesure établi pour les observations faites durant l'expérience nous permet de relier les écarts systématiques entre les paramètres de population aux diverses mises en œuvre de l'enquête. Supposons que L intervieweurs soient disponibles pour la collecte des données. Nous pouvons, conceptuellement, diviser la population U de taille N en L groupes U_l de taille $N_l, l = 1, \ldots, L,$ de sorte que tous les individus appartenant à un groupe puissent être interviewés par le même intervieweur. Soit $\overline{\mathbf{Y}} = (\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_K)^t$ le vecteur de dimension K des moyennes de population de $\mathbf{y}_i$, c'est-à-dire

$$\overline{\mathbf{Y}} = \mathbf{j}\frac{1}{N}\sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}\sum_{l=1}^{L} \frac{N_l}{N}\psi_l + \mathbf{j}\sum_{l=1}^{L} \frac{N_l}{N}\xi_l + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\varepsilon}_i. \tag{9}$$

L'objectif de l'expérience est de déterminer s'il existe des différences systématiques entre les K moyennes de population de $\overline{\mathbf{Y}}$ dues aux K stratégies d'enquête différentes, ou traitements. Nous pouvons pour cela formuler des hypothèses au sujet de

$$\mathrm{E}_m(\overline{\mathbf{Y}}) = \mathbf{j}\frac{1}{N}\sum_{i=1}^{N} u_i + \mathbf{j}\sum_{l=1}^{L} \frac{N_l}{N}\psi_l + \boldsymbol{\beta}, \qquad (10)$$

où l'espérance est prise sur le modèle de l'erreur de mesure. Nous obtenons ainsi les hypothèses suivantes :

$$H_0 : \mathbf{C}\mathrm{E}_m\overline{\mathbf{Y}} = 0,$$
$$H_1 : \mathbf{C}\mathrm{E}_m\overline{\mathbf{Y}} \neq 0, \qquad (11)$$

où $\mathbf{C}$ représente une matrice de dimensions $(K-1)\times K$ avec $K-1$ contrastes et $\mathbf{0}$ représente un vecteur de dimension $K-1$ de zéros. Puisque $\mathbf{Cj} = \mathbf{0}$, il s'ensuit que $\mathbf{C}\mathrm{E}_m\overline{\mathbf{Y}} = \mathbf{C}\boldsymbol{\beta}$ et les hypothèses (11) concernent les effets de traitement représentés par $\boldsymbol{\beta}$ dans le modèle d'erreur de mesure (1). Les contrastes entre les paramètres de population correspondent commodément à ces effets de traitement. En ce qui concerne les expériences randomisées considérées dans le présent article, il est vérifié que chaque unité expérimentale affectée à un intervieweur l a une probabilité non nulle d'être affecté à chacun des K traitements. Par conséquent, le biais induit dans les estimations des paramètres par les effets fixes d'intervieweurs est le même sous chacun des K traitements et s'annule dans les $K-1$ contrastes entre les K estimations de paramètre.

Nous vérifierons l'hypothèse (11) en estimant $\mathrm{E}_m\overline{\mathbf{Y}}$ au lieu de $\boldsymbol{\beta}$, en tenant compte du plan d'échantillonnage, du plan d'expérience et de la méthode de pondération appliquée dans l'enquête courante pour estimer les paramètres de population. Pour vérifier (11), nous disposons d'un échantillon probabiliste tiré d'une population finie. Les unités d'échantillonnage (unités expérimentales) sont randomisées sur K sous-échantillons et assignées à l'un des K traitements. À la section 2.3, nous élaborons un estimateur sans biais par rapport au plan de sondage de $\mathrm{E}_m\overline{\mathbf{Y}}$, que nous notons $\hat{\overline{\mathbf{Y}}}$. Par exemple, $\hat{\overline{\mathbf{Y}}}$ pourrait être l'estimateur d'Horvitz-Thompson ou l'estimateur par la régression généralisée. Soit $\mathbf{V}$ la matrice des covariances de $\hat{\overline{\mathbf{Y}}}$. Un estimateur (approximativement) sans biais par rapport au plan de sondage de la matrice des covariances des $K-1$ contrastes de $\hat{\overline{\mathbf{Y}}}$, noté $\mathbf{C}\hat{\mathbf{V}}\mathbf{C}^t$, sera établi à la section 2.4. Maintenant, nous pouvons vérifier l'hypothèse (11) au moyen de la statistique de Wald basée sur le plan de sondage qui suit :

$$W = \hat{\overline{\mathbf{Y}}}^t\mathbf{C}^t(\mathbf{C}\hat{\mathbf{V}}\mathbf{C}^t)^{-1}\mathbf{C}\hat{\overline{\mathbf{Y}}}. \qquad (12)$$

Pour des considérations d'ordre mathématique, nous préférons la matrice de contrastes $\mathbf{C} = (\mathbf{j} : -\mathbf{I})$, où $\mathbf{j}$ est un vecteur de dimension $K-1$ de 1 et $\mathbf{I}$ est la matrice identité de dimensions $(K-1)\times(K-1)$.

2.3 Estimation des effets de traitement

2.3.1 Estimateur d'Horvitz-Thompson

Considérons un échantillon s tiré selon un plan de sondage généralement complexe, qui peut être décrit par les probabilités d'inclusion de premier et de deuxième ordres π_i et $\pi_{ii'}$ de la i^{e} et des i, i'^{e} unité(s) d'échantillonnage, respectivement. Dans le cas d'un PRT, l'échantillon s est subdivisé aléatoirement en K sous-échantillons s_k de taille n_k. Si $n_+ = \sum_{k=1}^{K} n_k$ représente le nombre d'unités d'échantillonnage dans s, alors la probabilité conditionnelle que la i^{e} unité d'échantillonnage soit sélectionnée dans le sous-échantillon s_k, sachant que l'échantillon s est sélectionné, est égale à n_k / n_+. Dans le cas d'un PBR, les unités d'échantillonnage sont, sachant la réalisation de s, subdivisées de façon déterministe en J blocs s_j. Les variables de bloc possibles sont les structures d'échantillonnage telles que les strates, les grappes, les UPE, les intervieweurs et ainsi de suite. Dans chaque bloc, les unités d'échantillonnage sont randomisées sur les K traitements. Soit n_{jk} le nombre d'unités d'échantillonnage dans le bloc j assigné au traitement k. Alors, $n_{j+} = \sum_{k=1}^{K} n_{jk}$ est la taille du bloc j, $n_{+k} = \sum_{j=1}^{J} n_{jk}$ est la taille du sous-échantillon s_k et $n_{++} = \sum_{k=1}^{K}\sum_{j=1}^{J} n_{jk}$ est la taille de l'échantillon s. La probabilité conditionnelle que la i^{e} unité d'échantillonnage soit sélectionnée dans le sous-échantillon s_k, sachant que l'échantillon s est sélectionné et que $i \in s_j$, est égale à n_{jk} / n_{j+}.

Nous pouvons considérer chaque sous-échantillon s_k comme un échantillon à deux phases, où les probabilités d'inclusion de premier ordre de l'échantillon de la première phase sont obtenues d'après le plan de sondage et les probabilités d'inclusion conditionnelles de premier ordre de l'échantillon de la deuxième phase sont obtenues d'après le plan d'expérience. De ce point de vue, les probabilités d'inclusion de premier ordre des éléments de s_k sont égales à $\pi_i^* = (n_k / n_+)\,\pi_i$ pour les PRT et à $\pi_i^* = (n_{jk} / n_{j+})\,\pi_i$ pour les PBR si cette i^{e} unité d'échantillonnage est assignée au j^{e} bloc. Il s'ensuit que l'estimateur d'Horvitz-Thompson de $\overline{Y}_k$, fondé sur les n_{+k} observations obtenues à partir du sous-échantillon s_k peut être défini par :

$$\hat{\overline{Y}}_{k;\mathrm{HT}} = \frac{1}{N}\sum_{i=1}^{n_{+k}} \frac{y_{ik}}{\pi_i^*} \equiv \frac{1}{N}\sum_{i=1}^{n_{++}} \frac{\mathbf{p}_{ik}^t\,\mathbf{y}_i}{\pi_i}, \qquad (13)$$

où les $\mathbf{p}_{ik}$ sont les vecteurs de dimension K qui décrivent le mécanisme de randomisation du plan expérimental. Pour un PRT, il s'ensuit que

$$\mathbf{p}_{ik} \equiv \begin{cases} \dfrac{n_+}{n_k}\mathbf{r}_k & \text{si} \quad i \in s_k \\ \mathbf{0} & \text{si} \quad i \notin s_k \end{cases}, \tag{14}$$

et pour un PBR

$$\mathbf{p}_{ik} \equiv \begin{cases} \dfrac{n_{j+}}{n_{jk}}\mathbf{r}_k & \text{si} \quad i \in s_{jk} \\ \mathbf{0} & \text{si} \quad i \notin s_{jk} \end{cases}, \tag{15}$$

où $\mathbf{r}_k$ est le vecteur unitaire de dimension K avec le k^{e} élément égal à un et les autres éléments égaux à zéro, et $\mathbf{0}$ représente un vecteur de dimension K de zéros. Les propriétés des vecteurs $\mathbf{p}_{ik}$ sont données en annexe.

Maintenant, puisque s_k peut être considéré comme un échantillon à deux phases, il est vérifié que $\mathrm{E}_s\mathrm{E}_e(\hat{\bar{Y}}_{k;\mathrm{HT}}|s,m) = \bar{Y}_k$, où E_s et E_e représentent l'espérance par rapport au plan de sondage et au plan d'expérience, respectivement. Donc, sachant m, nous proposons le vecteur $\hat{\bar{\mathbf{Y}}}_{\mathrm{HT}} = (\hat{\bar{Y}}_{1;\mathrm{HT}}, \ldots, \hat{\bar{Y}}_{K;\mathrm{HT}})^t$ comme estimateur sans biais de $\bar{\mathbf{Y}}$. Mais alors, $\hat{\bar{\mathbf{Y}}}_{\mathrm{HT}}$ est sans biais pour $E_m\,\bar{\mathbf{Y}}$.

2.3.2 Estimateur par la régression généralisée

Dans le cas de l'échantillonnage en population finie, on augmente habituellement la précision de l'estimateur d'Horvitz-Thompson, si l'on dispose de données auxiliaires appropriées, au moyen de l'estimateur par la régression généralisée [consulter, par exemple, Bethlehem et Keller (1987) et Särndal, Swensson et Wretman (1992)]. L'estimateur par la régression généralisée nous permet d'intégrer le schéma de pondération de l'enquête courante dans l'analyse des expériences intégrées. Cela pourrait réduire la variance par rapport au plan de sondage, ainsi que le biais dû à la non-réponse sélective et, par conséquent, accroître la précision de l'expérience. Dans le présent contexte, l'estimateur par la régression généralisée représente donc un analogue fondé sur le plan de sondage de l'analyse de covariance de la méthodologie classique des plans d'expérience.

En plus des valeurs de la variable de réponse $\mathbf{y}_i$, nous associons à chaque unité de la population un vecteur d'information auxiliaire $\mathbf{x}_i$ de dimension H. Nous supposons que la moyenne en population finie de ces variables auxiliaires est connue et nous la représentons par $\bar{\mathbf{X}}$. Nous supposons aussi que les variables auxiliaires sont des valeurs intrinsèques, qui peuvent être observées sans erreur de mesure et ne sont, par conséquent, pas affectées par les traitements. Si nous suivons l'approche assistée par modèle de Särndal et coll. (1992), nous supposons que les valeurs intrinsèques u_i du modèle de l'erreur de mesure de la section 2.1 pour chaque unité de la population sont des réalisations indépendantes du modèle de régression linéaire :

$$u_i = B^t\mathbf{x}_i + e_i, \tag{16}$$

où B est un vecteur de dimension H contenant les coefficients de régression et les e_i sont les résidus. Dans l'approche assistée par modèle de Särndal et coll. (1992), les valeurs intrinsèques u_i sont considérées chacune comme une réalisation d'un modèle de superpopulation sous-jacent défini par (16). Dans ce cas, les résidus e_i sont des variables aléatoires indépendantes de variance ω_i^2. Alors, il est nécessaire de connaître tous les ω_i^2 jusqu'à un facteur d'échelle commun; autrement dit, $\omega_i^2 = v_i\omega^2$ avec v_i connu. D'un point de vue strictement axé sur le plan de sondage, proposé par Bethlehem et Keller (1987), il n'est pas nécessaire d'adopter un modèle de superpopulation. Alors, les résidus sont des valeurs intrinsèques fixes des éléments de la population finie et aucune hypothèse de modélisation ne doit être formulée au sujet des résidus. Ici, nous adoptons l'approche assistée par modèle de Särndal. Cela signifie que les espérances par rapport au modèle de mesure, comme en (7) et (10), sont les espérances conditionnelles sachant la réalisation des valeurs intrinsèques $u_i, i = 1, \ldots, N$, dans la population finie conformément au modèle de superpopulation (16).

Les coefficients de régression du modèle linéaire (16) dans la population finie sont définis par

$$\mathbf{b} = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^t}{\omega_i^2}\right)^{-1} \sum_{i=1}^{N} \frac{\mathbf{x}_i u_i}{\omega_i^2}. \tag{17}$$

Les valeurs intrinsèques u_i ne sont pas observables à cause des erreurs de mesure et des effets de traitement. Par conséquent, nous ne pouvons calculer (17), même si nous dénombrons entièrement la population finie. Dans le cas d'un recensement complet sous le k^{e} traitement

$$\tilde{\mathbf{b}}_k = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^t}{\omega_i^2}\right)^{-1} \sum_{i=1}^{N} \frac{\mathbf{x}_i y_{ik}}{\omega_i^2}, \quad k = 1, 2, \ldots, K, \tag{18}$$

représente les coefficients de régression en population finie du modèle linéaire (16). Sachant la réalisation de $u_i, i = 1, \ldots, N$, l'espérance des coefficients de régression en population finie $\tilde{\mathbf{b}}_k$ par rapport au modèle de l'erreur de mesure est donnée par

$$\mathrm{E}_m\tilde{\mathbf{b}}_k = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^t}{\omega_i^2}\right)^{-1} \sum_{i=1}^{N} \frac{\mathbf{x}_i(u_i + \beta_k + \psi_l)}{\omega_i^2} \equiv \mathbf{b}_k, \quad k = 1, 2, \ldots, K. \tag{19}$$

Nous pouvons estimer les coefficients de régression en population finie $\tilde{\mathbf{b}}_k$ et $\mathbf{b}_k$ en utilisant les données d'échantillon provenant du sous-échantillon s_k avec l'estimateur d'Horvitz-Thompson :

$$\hat{\mathbf{b}}_k = \left(\sum_{i=1}^{n_{+k}} \frac{\mathbf{x}_i\mathbf{x}_i^t}{\omega_i^2\pi_i^*}\right)^{-1} \sum_{i=1}^{n_{+k}} \frac{\mathbf{x}_i y_{ik}}{\omega_i^2\pi_i^*}, \quad k = 1, 2, \ldots, K.$$

Maintenant, l'estimateur par la régression généralisée de $\bar{Y}_k$, fondé sur les n_{+k} observations du sous-échantillon s_k, est défini par

$$\hat{\bar{Y}}_{k;\text{greg}} = \hat{\bar{Y}}_{k;\text{HT}} + \hat{\mathbf{b}}_k^t(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}_{\text{HT}}), \quad k = 1, 2, \ldots, K, \qquad (20)$$

où $\hat{\bar{\mathbf{X}}}_{\text{HT}}$ est l'estimateur d'Horvitz-Thompson pour les moyennes de population des variables auxiliaires $\bar{\mathbf{X}}$ fondées sur les n_{+k} unités d'échantillonnage du sous-échantillon s_k.

En exprimant (20) sous forme d'une fonction de $(\hat{\bar{Y}}_{k;\text{HT}}, \hat{\mathbf{b}}_k, \hat{\bar{\mathbf{X}}}_{\text{HT}})$, nous pouvons approximer l'estimateur par la régression généralisée au moyen d'une linéarisation de Taylor de premier ordre autour de $(\text{E}_m \bar{Y}_k, \mathbf{b}_k, \bar{\mathbf{X}})$, où $\mathbf{b}_k$ est défini dans (19). Ceci nous donne :

$$\hat{\bar{Y}}_{k;\text{greg}} \doteq \hat{\bar{Y}}_{k;\text{HT}} + \mathbf{b}_k^t(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}_{\text{HT}}) = \hat{\bar{E}}_{k;\text{HT}} + \mathbf{b}_k^t \bar{\mathbf{X}}, \; k = 1, 2, \ldots, K,$$

avec

$$\hat{\bar{E}}_{k;\text{HT}} = \hat{\bar{Y}}_{k;\text{HT}} - \mathbf{b}_k^t \hat{\bar{\mathbf{X}}}_{\text{HT}} = \sum_{i \in s} \left(\frac{\mathbf{p}_{ik}^t (\mathbf{y}_i - \mathbf{B}^t \mathbf{x}_i)}{\pi_i N} \right),$$

et où $\mathbf{B}$ est une matrice de dimensions $H \times K$ dont les colonnes sont les vecteurs $\mathbf{b}_k$ de dimension H. Maintenant, nous proposons $\hat{\bar{\mathbf{Y}}}_{\text{GREG}} = (\hat{\bar{Y}}_{1;\text{greg}}, \ldots, \hat{\bar{Y}}_{K;\text{greg}})^t$ comme estimateur approximativement sans biais de $E_m \bar{\mathbf{Y}}$.

2.4 Estimation de la variance des effets des traitements

Soit $\mathbf{V}$ la matrice des covariances de $\hat{\bar{\mathbf{Y}}}_{\text{GREG}}$. Pour estimer les termes de covariance de $\mathbf{V}$, nous avons besoin des vecteurs $\mathbf{y}_i$ contenant les observations obtenues auprès de chaque unité d'échantillonnage pour les K traitements. Puisque, dans les plans d'expérience considérés ici, chaque unité d'échantillonnage est assignée à l'un des K traitements, nous n'observons effectivement qu'une seule des composantes de $\mathbf{y}_i$, pour $i \in s$. Par conséquent, nous ne pouvons établir un estimateur sans biais par rapport au plan de sondage pour $\mathbf{V}$. Van den Brakel et Binder (2000, 2004) ont essayé de surmonter ce problème en imputant les composantes inobservées. Cependant, l'utilité de leurs résultats dépend de l'exactitude du modèle d'imputation. Ici, nous contournons le problème en établissant un estimateur basé sur le plan de sondage pour $\mathbf{CVC}^t$, c'est-à-dire la matrice des covariances des contrastes de $\hat{\bar{\mathbf{Y}}}_{\text{GREG}}$, qui est suffisant pour la statistique de Wald (12).

Nous commençons par établir les expressions pour l'estimateur par la régression généralisée. Les résultats pour l'estimateur d'Horvitz-Thompson sont donnés à titre de cas particulier. La matrice des covariances des contrastes de $\hat{\bar{\mathbf{Y}}}_{\text{GREG}}$ peut être approximée par la matrice des covariances des contrastes de $\hat{\bar{\mathbf{E}}}_{\text{HT}} = (\hat{\bar{E}}_{1;\text{HT}}, \ldots, \hat{\bar{E}}_{K;\text{HT}})^t$. Soit Cov_s et Cov_e les covariances par rapport au plan de sondage et au plan d'expérience, respectivement. Maintenant, considérons la décomposition de la variance suivante :

$$\mathbf{CVC}^t = \text{Cov}_m \text{E}_s \text{E}_e (\mathbf{C}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) + \text{E}_m \text{Cov}_s \text{E}_e (\mathbf{C}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) + \text{E}_m \text{E}_s \text{Cov}_e (\mathbf{C}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s). \qquad (21)$$

Puisque $\text{E}_e(\mathbf{p}_{ik}) = \mathbf{r}_k$ (voir (42) à l'annexe), il s'ensuit que

$$\text{E}_e(\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) = \sum_{i \in s} \left(\frac{(\mathbf{y}_i - \mathbf{B}^t \mathbf{x}_i)}{\pi_i N} \right). \qquad (22)$$

Sous la condition qu'il existe un vecteur de dimension H constant $\mathbf{a}$, tel que $\mathbf{a}^t \mathbf{x}_i = 1$ pour tout $i \in U$, nous prouverons à l'annexe que

$$\mathbf{C}(\mathbf{y}_i - \mathbf{B}^t \mathbf{x}_i) = \mathbf{C}\boldsymbol{\varepsilon}_i. \qquad (23)$$

La condition énoncée suppose implicitement que l'on connaît la taille de la population finie et qu'on l'utilise comme information auxiliaire. Cette condition tient pour les modèles de pondération qui contiennent une ordonnée à l'origine ou bien une ou plusieurs variables nominales qui partitionnent la population en sous-populations. En utilisant les hypothèses de modélisation (2) et (3), il découle de (22) et (23) que

$$\text{Cov}_m \text{E}_s \text{E}_e (\mathbf{C}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) = \text{Cov}_m \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{C}\boldsymbol{\varepsilon}_i \right) = \frac{1}{N^2} \sum_{i=1}^{N} \mathbf{C} \sum\nolimits_i \mathbf{C}^t, \qquad (24)$$

et

$$\text{E}_m \text{Cov}_s \text{E}_e (\mathbf{C}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) = E_m \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} (\pi_{ii'} - \pi_i \pi_{i'}) \times \frac{\mathbf{C}\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_{i'}^t \mathbf{C}^t}{\pi_i \pi_{i'}} = \frac{1}{N^2} \sum_{i=1}^{N} \left(\frac{1}{\pi_i} - 1 \right) \mathbf{C} \sum\nolimits_i \mathbf{C}^t. \qquad (25)$$

Pour le troisième terme de (21), nous prouvons à l'annexe, dans le cas d'un PBR, que

$$\text{E}_m \text{E}_s \text{Cov}_e (\mathbf{C}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) = \text{E}_m \text{E}_e (\mathbf{CDC}^t) - \frac{1}{N^2} \sum_{i=1}^{N} \frac{\mathbf{C} \sum\nolimits_i \mathbf{C}^t}{\pi_i}, \qquad (26)$$

où $\mathbf{D}$ est une matrice diagonale de dimensions $K \times K$ dont les éléments diagonaux sont donnés par

$$d_k = \sum_{j=1}^{J} \frac{1}{n_{jk}} \frac{1}{n_{j+} - 1} \sum_{i=1}^{n_{j+}} \left(\frac{n_{j+}(y_{ik} - \mathbf{b}_k^t \mathbf{x}_i)}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{n_{j+}(y_{i'k} - \mathbf{b}_k^t \mathbf{x}_{i'})}{N \pi_{i'}} \right)^2 \equiv \sum_{j=1}^{J} \frac{S_{E_{jk}}^2}{n_{jk}}. \qquad (27)$$

Si nous insérons les résultats obtenus en (24), (25) et (26) dans (21), il s'ensuit que

$$\mathbf{CVC}^t = \text{E}_m \text{E}_s \mathbf{CDC}^t. \qquad (28)$$

Sachant la réalisation de *m* et de *s*, nous pouvons établir un estimateur approximativement sans biais par rapport au plan de sondage pour **D** dans (28). Par conséquent, nous pouvons, de façon commode, exprimer implicitement $\mathbf{CVC}^t$ en tant qu'espérance sur le modèle de l'erreur de mesure et le plan de sondage. Voir Van den Brakel (2001) pour les expressions explicites de $\mathbf{CVC}^t$. Sachant la réalisation de *m* et de *s*, la répartition des unités d'échantillonnage contenues dans chaque bloc entre les sous-échantillons s_{jk} peut être considérée comme un échantillonnage aléatoire simple sans remise à partir du bloc s_j. Par conséquent, pour un PBR, il s'ensuit qu'un estimateur approximativement sans biais par rapport au plan de sondage de **D** est donné par une matrice diagonale $\hat{\mathbf{D}}$ de dimensions $K \times K$ dont les éléments diagonaux sont

$$\hat{d}_k = \sum_{j=1}^{J} \frac{1}{n_{jk}} \frac{1}{n_{jk}-1} \sum_{i=1}^{n_{jk}} \left(\frac{n_{j+}(y_{ik} - \hat{\mathbf{b}}_k^t \mathbf{x}_i)}{N \pi_i} - \frac{1}{n_{jk}} \sum_{i'=1}^{n_{jk}} \frac{n_{j+}(y_{i'k} - \hat{\mathbf{b}}_k^t \mathbf{x}_{i'})}{N \pi_{i'}} \right)^2 \equiv \sum_{j=1}^{J} \frac{\hat{S}_{E_{jk}}^2}{n_{jk}}. \quad (29)$$

Un estimateur approximativement sans biais par rapport au plan de sondage de la matrice $\mathbf{CVC}^t$ définie en (28) est donné par $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t$. Les résultats pour un PRT découlent directement, à titre de cas particulier, de (27) et (29), où $J = 1, n_{j+} = n_+$ et $n_{jk} = n_k$. Une autre solution consiste à multiplier les résidus $(y_{ik} - \hat{\mathbf{b}}_k^t \mathbf{x}_i)$ dans (29) par les poids de correction (également appelés poids *g*, Särndal et coll. 1992, résultat 6.6.1). Puisque $\mathbf{CVC}^t$ dans (28) est définie implicitement comme étant l'espérance sur le plan de sondage, (29) est approximativement sans biais par rapport au plan de sondage sous des plans d'échantillonnage complexes généraux. Cet estimateur de la variance nécessite uniquement que soient fixées d'avance les fractions d'unités d'échantillonnage assignées aux divers traitements conformément au plan d'expérience. La taille de l'échantillon, ainsi que les blocs peuvent être aléatoires en ce qui concerne le plan de sondage, par exemple, dans le cas d'un PBR où les grappes ou bien les UPE sont la variable de bloc.

L'estimateur de la variance $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t$ a la même structure que si les *K* sous-échantillons avaient été tirés indépendamment l'un de l'autre, où les unités d'échantillonnage sont sélectionnées avec probabilités inégales (π_i / n_+) avec remise dans le cas d'un PRT, ou (π_i / n_{j+}) avec remise dans chaque bloc *j* dans le cas d'un PBR (comparer (29) à l'équation (9A.16) dans Cochran 1977). Il est remarquable que les probabilités d'inclusion de deuxième ordre du plan de sondage aient disparu. Cette disparition a pour causes :

1. l'hypothèse d'additivité des effets de traitement dans le modèle de l'erreur de mesure, c'est-à-dire β_k pour tout $i \in U$ observé sous le traitement *k*;
2. l'hypothèse que les erreurs de mesure entre individus sont indépendantes;
3. un schéma de pondération correctement choisi pour que la condition $\mathbf{a}^t \mathbf{x}_i = 1$ pour tout $i \in U$ soit satisfaite;
4. le fait que les variances sont calculées pour les contrastes entre les moyennes de sous-échantillons.

La variance par rapport au plan de sondage de l'approximation par développement en série de Taylor de premier ordre de l'estimateur par la régression généralisée est constituée des résidus $(y_{ik} - \mathbf{b}_k^t \mathbf{x}_i)$. De la preuve de (23) il découle que, sous un schéma de pondération qui satisfait à la condition $\mathbf{a}^t \mathbf{x}_i = 1$ pour tout $i \in U$, les effets de traitement β_k disparaissent des résidus $(y_{ik} - \mathbf{b}_k^t \mathbf{x}_i)$ dans (23). Trois termes demeurent dans ces résidus, à savoir :

1. le résidus du modèle de régression linéaire de la valeur intrinsèque, c'est-à-dire $e_i = u_i - \mathbf{b}^t \mathbf{x}_i$;
2. un terme concernant le biais dû aux effets d'intervieweur, qui est égal à $\psi_{il} - \mathbf{d}^t \mathbf{x}_i$, où **d** représente les coefficients de régression provenant de la fonction de régression des effets d'intervieweur sur les variables auxiliaires $\mathbf{x}_i$, (voir la preuve de (23) à l'annexe);
3. les erreurs de mesure ε_{ik}.

Les résidus des valeurs intrinsèques e_i et le biais dû aux effets d'intervieweur ne dépendent pas des divers traitements et, par conséquent, s'annulent dans les contrastes des résidus dans (23). Seules les erreurs de mesure $\boldsymbol{\varepsilon}_i$ persistent dans ces contrastes. Par conséquent, les deux termes $\mathrm{Cov}_m \mathrm{E}_s \mathrm{E}_e (\mathbf{C}\hat{\tilde{\mathbf{E}}}_{\mathrm{HT}} | m, s)$ et $\mathrm{E}_m \mathrm{Cov}_s \mathrm{E}_e (\mathbf{C}\hat{\tilde{\mathbf{E}}}_{\mathrm{HT}} | m, s)$ contiennent uniquement les erreurs de mesure ε_{ik}. Étant donné l'hypothèse d'indépendance des erreurs de mesure entre les individus, les produits croisés entre individus, qui contiennent les probabilités d'inclusion de deuxième ordre dans (24) et (25), disparaissent. La structure de covariance du troisième terme de (21) est déterminée principalement par le mécanisme de randomisation du plan d'expérience. Pour un PRT, ce mécanisme se résume à la sélection de *K* sous-échantillons à partir de *s* par échantillonnage aléatoire simple sans remise. Pour un PBR, il se résume à la sélection de *K* sous-échantillon à partir de *s* par échantillonnage aléatoire simple stratifié sans remise, où les strates correspondent aux blocs de l'expérience. Dans la variance des contrastes des moyennes de sous-échantillons, les corrections pour population finie dans la variance par rapport au plan de sondage des moyennes de sous-échantillons sont annulées par les covariances par rapport au plan de sondage entre les moyennes de sous- échantillons. Par conséquent, le terme principal de (26), c'est-à-dire $\mathrm{E}_m \mathrm{E}_s \mathbf{CDC}^t$, a la même structure que si les *K* sous-échantillons étaient

sélectionnés indépendamment l'un de l'autre par échantillonnage aléatoire simple avec remise dans le cas d'un PRT, ou d'un échantillonnage aléatoire simple stratifié avec remise dans le cas d'un PBR. Les probabilités d'inclusion de deuxième ordre apparaissent si, dans (28), l'espérance par rapport au plan de sondage est rendue explicite, voir Van den Brakel (2001).

L'utilisation minimale d'informations auxiliaires correspond à un schéma de pondération où $\mathbf{x}_i = (1)$ et $\omega_i^2 = \omega^2$ pour tout $i \in U$. Sous ce schéma de pondération, il s'ensuit que

$$\hat{\bar{Y}}_{k;\text{greg}} = \left(\sum_{i=1}^{n_{+k}} \frac{1}{\pi_i^*} \right)^{-1} \left(\sum_{i=1}^{n_{+k}} \frac{y_{ik}}{\pi_i^*} \right) \equiv \tilde{y}_k, \qquad (30)$$

qui s'avère être l'estimateur par le ratio d'une moyenne de population, proposé au départ par Hájek (1971). Il s'ensuit aussi que $\hat{\mathbf{b}}_k = (\tilde{y}_k)$ et qu'un estimateur approximativement sans biais par rapport au plan de sondage de la matrice des covariances des effets de traitement est donné par (29) avec $\hat{\mathbf{b}}_k^t \mathbf{x}_i = \tilde{y}_k$.

Si $\sum_{i=1}^{n_{+k}} 1/\pi_i^* \equiv \hat{N} = N$, alors l'estimateur par le ratio (30) correspond à l'estimateur d'Horvitz-Thompson courant. Cette condition est satisfaite dans le cas d'un PRT ou d'un PBR intégré dans un plan d'échantillonnage aléatoire simple, d'un PBR intégré dans un plan d'échantillonnage aléatoire simple stratifié où les strates sont utilisées comme variables de blocs, ou d'un PRT intégré dans un plan d'échantillonnage aléatoire simple stratifié avec répartition proportionnelle. Sous la condition $\hat{N} = N$, les expressions de la variance par rapport au plan de sondage de l'estimateur d'Horvitz-Thompson sont données par (27) et (29), où $y_{ik} - \mathbf{b}_k^t \mathbf{x}_i$ et $y_{ik} - \hat{\mathbf{b}}_k^t \mathbf{x}_i$ sont remplacés par y_{ik}. Les expressions de la variance de l'estimateur d'Horvitz-Thompson sont plus compliquées si $\hat{N} \neq N$; voir Van den Brakel (2001).

2.5 Test de Wald

L'insertion des estimateurs sans biais des moyennes de sous-échantillon et de la matrice des covariances des contrastes entre ces moyennes de sous-échantillons dans l'équation (12) donnent la statistique de Wald basée sur le plan de sondage

$$W = \hat{\bar{\mathbf{Y}}}_{\text{GREG}}^t \, \mathbf{C}^t (\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t)^{-1} \mathbf{C} \, \hat{\bar{\mathbf{Y}}}_{\text{GREG}}. \qquad (31)$$

Nous prouvons à l'annexe que cette expression peut se simplifier en :

$$W = \sum_{k=1}^{K} \frac{\hat{\bar{Y}}_{k;\text{greg}}^2}{\hat{d}_k} - \frac{1}{\sum_{k=1}^{K} \frac{1}{\hat{d}_k}} \left(\sum_{k=1}^{K} \frac{\hat{\bar{Y}}_{k;\text{greg}}}{\hat{d}_k} \right)^2. \qquad (32)$$

Pour les plans de sondage généraux, la loi asymptotique de cette statistique de test sera inconnue. Cependant, si le plan de sondage est un plan aléatoire simple sans remise et que le plan d'expérience est un PRT, Lehmann (1975, annexe 8) donne, en s'inspirant des travaux de Hájek (1960), les conditions suffisantes sous lesquelles $\hat{\bar{\mathbf{E}}}_{\text{HT}}$ est distribué asymptotiquement suivant une loi normale multivariée de moyenne $\mathrm{E}_s \mathrm{E}_e(\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) = \bar{\mathbf{E}}$ et matrice des covariances $\tilde{\mathbf{V}} = \mathrm{Cov}_s \mathrm{E}_s(\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) + \mathrm{E}_s \mathrm{Cov}_e(\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s)$ si $n_{+k} \to \infty$ et $(N - n_{++}) \to \infty$: $(\hat{\bar{\mathbf{E}}}_{\text{HT}} | m) \to N(\bar{\mathbf{E}}, \tilde{\mathbf{V}})$. Donc, $(\mathbf{C}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m) \to N(\mathbf{C}\bar{\mathbf{E}}, \mathbf{C}\tilde{\mathbf{V}}\mathbf{C}^t)$, avec $\mathbf{C}\bar{\mathbf{E}} = (1/N)\sum_{i=1}^{N} \mathbf{C}\boldsymbol{\varepsilon}_i$. Puisque les $\mathbf{C}\boldsymbol{\varepsilon}_i$ sont des variables aléatoires mutuellement indépendantes de moyenne nulle et matrice des covariances $\mathbf{C}\sum_i \mathbf{C}^t$, en vertu du théorème central limite, nous avons $(\mathbf{C}\bar{\mathbf{E}}) \to N(0, (1/N^2)\sum_{i=1}^{N} \mathbf{C}\sum_i \mathbf{C}^t)$. En combinant les deux lois limites, nous obtenons que, inconditionnellement, $\mathbf{C}\hat{\bar{\mathbf{E}}}_{\text{HT}} \to N(0, \mathbf{C}\mathbf{V}\mathbf{C}^t)$ et donc $\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\text{GREG}} \to N(\mathbf{C}\boldsymbol{\beta}, \mathbf{C}\mathbf{V}\mathbf{C}^t)$. Par conséquent, il s'ensuit, sous l'hypothèse nulle, que W suit asymptotiquement une loi du chi-carré à $K-1$ degrés de liberté (Searle 1971, théorème 2, chapitre 2). Pour des plans de sondage plus complexes, il est habituellement conjecturé que $\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\text{GREG}} \to N(\mathbf{C}\boldsymbol{\beta}, \mathbf{C}\mathbf{V}\mathbf{C}^t)$. Alors, W est encore distribué asymptotiquement selon une loi du chi-carré à $K-1$ degrés de liberté. La validité de cette conjecture a été confirmée par des études en simulation [voir la section 4 et Van den Brakel (2001)].

2.6 Estimateurs groupés de la variance

Dans le cas d'un PBR, les n_{++} unités d'échantillonnage de s sont réparties en JK groupes de taille n_{jk}. Il faut estimer une variance de population $\hat{S}_{E_{jk}}^2$ distincte pour chacun de ces JK sous-échantillons. Si le nombre d'unités expérimentales n_{jk} disponibles pour l'estimation de ces variances de population devient trop petit, les estimations risquent de devenir instables. Le cas échéant, on peut obtenir des estimations plus stables en regroupant les estimations des variances de population dans les blocs.

Les résidus de l'estimateur par la régression généralisée, $(y_{ik} - \mathbf{b}_k^t \mathbf{x}_i)$, dépendent uniquement du k^{e} effet de traitement par la voie des erreurs de mesure ε_{ik}. Sous l'hypothèse que $\sum_i = \sigma^2 \mathbf{I}$ dans (3) pour tout $i \in U$, il s'ensuit que les $S_{E_{jk}}^2$ dans chaque bloc sont des paramètres identiques, c'est-à-dire $S_{E_{j1}}^2 = \ldots = S_{E_{jK}}^2 = S_{E_j}^2$, pour $j = 1, 2, \ldots, J$. Sous cette hypothèse, il est efficace d'utiliser un estimateur groupé pour $S_{E_j}^2$;

$$\hat{S}_{E_j;\,P_1}^2 = \frac{1}{(n_{j+} - 1)} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} \left(\frac{n_{j+}(y_{ik} - \hat{\mathbf{b}}_k^t \mathbf{x}_i)}{N \pi_i} - \frac{1}{n_{j+}} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} \frac{n_{j+}(y_{ik} - \hat{\mathbf{b}}_k^t \mathbf{x}_i)}{N \pi_i} \right)^2 \qquad (33)$$

ou, alternativement,

$$\hat{S}^2_{E_j;\,P_2} = \frac{1}{(n_{j+}-K)}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} \left(\frac{n_{j+}(y_{ik}-\hat{\mathbf{b}}_k^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}}\frac{n_{j+}(y_{ik}-\hat{\mathbf{b}}_k^t\mathbf{x}_i)}{N\pi_i}\right)^2. \quad (34)$$

Dans plusieurs cas particuliers, la statistique de Wald basée sur le plan de sondage coïncide avec la statistique F utilisée dans les méthodes d'analyse basées sur un modèle plus conventionnel. Considérons un PBR intégré dans un plan de sondage autopondéré, où les unités d'échantillonnage sont réparties proportionnellement entre les traitements sur les blocs, c'est-à-dire $\pi_i = n_{++}/N$ et $n_{jk}/n_{j+} = n_{+k}/n_{++}$ pour tout $j = 1, \ldots, J$. Alors, il découle des résultats obtenus pour l'estimateur par le ratio (30) que $\bar{Y}_{k;\text{greg}} = 1/n_{+k}\sum_{i=1}^{n_{+k}} y_{ik} \equiv \bar{y}_{+k}$ et $\hat{\mathbf{b}}_k^t\mathbf{x}_i = \bar{y}_{+k}$. Notons $\bar{y}_{j+} = 1/n_{j+}\sum_{i=1}^{n_{j+}} y_{ik}$ et $\bar{y}_{++} = 1/n_{++}\sum_{k=1}^{K}\sum_{i=1}^{n_{+k}} y_{ik}$, alors nous avons

$$\frac{1}{n_{j+}}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} y_{ik} = \bar{y}_{j+}, \qquad \frac{1}{n_{j+}}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} \hat{\mathbf{b}}_k^t\mathbf{x}_i =$$

$$\frac{1}{n_{j+}}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} \bar{y}_{+k} = \sum_{k=1}^{K}\frac{n_{jk}}{n_{j+}}\bar{y}_{+k} = \sum_{k=1}^{K}\frac{n_{+k}}{n_{++}}\bar{y}_{+k} = \bar{y}_{++}.$$

Si $n_{j+} \approx n_{j+} - 1$, alors il s'ensuit sous l'estimateur groupé de la variance (33) que

$$\hat{d}_k = \sum_{j=1}^{J}\frac{n_{j+}}{n_{jk}}\frac{n_{j+}}{n_{j+}-1}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} \left(\frac{y_{ik}-\hat{\mathbf{b}}_k^t\mathbf{x}_i}{N\pi_i} - \frac{1}{n_{j+}}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}\frac{y_{ik}-\hat{\mathbf{b}}_k^t\mathbf{x}_i}{N\pi_i}\right)^2$$

$$\approx \frac{1}{n_{+k}}\frac{1}{n_{++}}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ik}-\bar{y}_{+k}-\bar{y}_{j+}+\bar{y}_{++})^2 \equiv \frac{\hat{d}_{P_1}}{n_{+k}}. \quad (35)$$

Notons $\bar{y}_{jk} = 1/n_{jk}\sum_{i=1}^{n_{jk}} y_{ik}$. Sous l'estimateur groupé de la variance (34), il s'ensuit que

$$\hat{d}_k = \sum_{j=1}^{J}\frac{n_{j+}}{n_{jk}}\frac{n_{j+}}{n_{j+}-K)}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} \left(\frac{y_{ik}-\hat{\mathbf{b}}_k^t\mathbf{x}_i}{N\pi_i} - \frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}}\frac{y_{ik}-\hat{\mathbf{b}}_k^t\mathbf{x}_i}{N\pi_i}\right)^2$$

$$\approx \frac{1}{n_{+k}}\frac{1}{n_{++}}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ik}-\bar{y}_{jk})^2 \equiv \frac{\hat{d}_{P_2}}{n_{+k}}. \quad (36)$$

En introduisant par substitution ces estimateurs groupés de la variance dans la statistique de Wald (32), nous obtenons

$$W = \frac{1}{\hat{d}_{P_a}}\left(\sum_{k=1}^{K} n_{+k}(\bar{y}_{+k})^2 - n_{++}(\bar{y}_{++})^2\right), \quad (37)$$

où $\hat{d}_{P_a}$ est donné par (35) pour $a = 1$ ou par (36) pour $a = 2$. Il est reconnaissable que $W/(K-1)$ dans (37) avec $\hat{d}_{P_1}$ l'estimateur groupé de la variance donné par (35) correspond à la statistique F d'une analyse de variance à deux critères de classification sans interactions. Si l'on insère $\hat{d}_{P_2}$ donné par (36), alors $W/(K-1)$ correspond à la statistique F d'une analyse de variance à deux critères de classification avec interactions (Scheffé 1959, chapitre 4). Un estimateur groupé de la variance pour un PRT découle, en tant que cas particulier, de (35) et (36). Sous les deux estimateurs, il s'ensuit que $W/(K-1)$ correspond à la statistique F d'une analyse de variance à un critère de classification (Scheffé 1959, chapitre 3).

2.7 Avantages des plans en blocs randomisés

Le principal avantage des PBR est l'élimination de la variation entre les blocs dans l'analyse des effets de traitement. Les unités d'échantillonnage provenant d'une même strate, UPE ou grappe sont plus homogènes que celles provenant de strates, UPE ou grappes différentes, ce qui donne à penser qu'on devrait utiliser des structures d'échantillonnage telles que les strates, les UPE ou les grappes comme variables de bloc dans un PBR (Fienberg et Tanur 1987, 1988). Cette approche assure que chaque strate, UPE ou grappe soit suffisamment représentée dans chaque sous-échantillon. Les intervieweurs peuvent aussi être utilisés comme variables de bloc, puisque la variation des observations due aux effets fixes ou aléatoires d'intervieweurs spécifiés dans le modèle de l'erreur de mesure (1) est alors éliminée. Dans les enquêtes où les intervieweurs recueillent les données par IPAO dans des régions géographiques distinctes, la constitution de blocs d'après les intervieweurs élimine aussi cette variation régionale de la variable cible. La puissance d'une expérience est maximisée si l'on répartit les unités d'échantillonnage proportionnellement entre les traitements sur les blocs, c'est-à-dire $n_{jk}/n_{j+} = n_{+k}/n_{++}$ pour tout $j = 1, \ldots, J$ (voir Van den Brakel 2001, chapitre 6). Le meilleur moyen de préserver cette répartition est d'utiliser les intervieweurs comme variables de bloc, puisque le taux de réponse varie considérablement d'un intervieweur à l'autre. La randomisation non contrainte au moyen d'un PRT n'est pas toujours réalisable en pratique. Par exemple, dans le cas des enquêtes par IPAO, où les intervieweurs recueillent les données dans des régions géographiques autour de leur lieu de résidence, il pourrait être nécessaire de restreindre la randomisation des unités d'échantillonnage aux intervieweurs ou aux régions géographiques qui sont des unions de régions d'intervieweur adjacentes pour éviter d'accroître de façon inacceptable la distance que doivent parcourir les intervieweurs. Cette approche mène naturellement au PBR avec les intervieweurs ou les régions comme variables de bloc.

3. Analyse par régression linéaire basée sur le plan de sondage

Nous pourrions envisager une régression linéaire basée sur le plan de sondage pour remplacer l'analyse des expériences intégrées. On suppose, dans ce cas, que les observations sont le résultat d'un modèle de régression linéaire $y_i = B^t\mathbf{x}_i + e_i$, avec $\mathbf{x}_i$ le vecteur contenant Q variables explicatives, B le vecteur contenant les coefficients de régression et e_i un résidu. Ce modèle est déterminé principalement par le plan d'expérience et contient les facteurs de traitement, les facteurs de contrôle locaux (par exemple, blocs) et les covariables comme variables explicatives (voir, par exemple, Montgomery 2001). Des covariables possibles sont les variables auxiliaires du schéma de pondération de l'estimateur par la régression généralisée. Les paramètres d'intérêt sont les coefficients de régression dans la population finie, qui sont définis par $\boldsymbol{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, où $\mathbf{X}$ est la matrice de plan d'expérience de dimensions $N \times Q$ et $\mathbf{y}$ est un vecteur de dimension N contenant les observations obtenues sous les divers traitements, comme si la population finie complète était incluse dans l'expérience. La matrice de plan d'expérience subdivise conceptuellement la population en K sous-populations ou domaines, qui sont observés sous chacun des K traitements de l'expérience. La taille de chaque sous-population est déterminée par la fraction d'unités d'échantillonnage assignées à chaque traitement dans l'expérience. Un estimateur des coefficients de régression fondé sur le plan est donné par $\hat{\boldsymbol{\beta}} = (\mathbf{X}_n^t\boldsymbol{\Pi}^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n^t\boldsymbol{\Pi}^{-1}\mathbf{y}_n$, (Särndal et coll. 1992, section 5.10). Ici, $\mathbf{X}_n$ est la matrice de plan d'expérience de dimensions $n \times Q$, $\mathbf{y}_n$ est un vecteur contenant n observations obtenues sous les divers traitements des n unités incluses dans l'échantillon, et $\boldsymbol{\Pi}$ est une matrice diagonale de dimensions $n \times n$ contenant les probabilités d'inclusion de premier ordre π_i du plan de sondage. La matrice des covariances approximative de $\hat{\boldsymbol{\beta}}$ est donnée par (Särndal et coll. 1992, section 5.10)

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^t\mathbf{X})^{-1}\boldsymbol{\Lambda}(\mathbf{X}^t\mathbf{X})^{-1}, \qquad (38)$$

avec $\boldsymbol{\Lambda} = \mathrm{Var}_s(\mathbf{X}_n^t\boldsymbol{\Pi}^{-1}\mathbf{y}_n - \mathbf{X}_n^t\boldsymbol{\Pi}^{-1}\mathbf{X}_n\boldsymbol{\beta})$. Les éléments de $\boldsymbol{\Lambda}$ sont donnés par

$$\lambda_{qq'} = \sum_{i\in U}\sum_{i'\in U}(\pi_{ii'} - \pi_i\pi_{i'})\frac{x_{iq}e_i}{\pi_i}\frac{x_{i'q'}e_{i'}}{\pi_{i'}}, \quad q,q' = 1,\ldots,Q,$$

avec $e_i = y_i - \boldsymbol{\beta}^t\mathbf{x}_i$. Les hypothèses au sujet du sous-ensemble de coefficients de régression qui reflète les effets de traitement sont soumises à un test de Wald; voir, par exemple, Skinner (1989).

Le principal inconvénient de cette approche est que la méthode d'estimation ne tient pas compte de l'affectation aléatoire des unités d'échantillonnage aux traitements conformément aux plans d'expérience. En procédant comme cela, les estimations sur les sous-échantillons sont traitées incorrectement comme s'il s'agissait d'estimations de domaine, ce qui donne des variances par rapport aux plans de sondage incorrectes. La matrice des covariances des effets de traitement (28), établie à la section 2.4, illustre le fait que la superposition du plan d'expérience au plan d'échantillonnage détermine quelles caractéristiques particulières du plan d'échantillonnage sont annulées ou préservées. Par exemple, l'effet de l'échantillonnage stratifié ou de l'échantillonnage à deux degrés sur la variance des effets de traitement est annulé sous un plan en randomisation totale. Toutefois, l'approche de la régression linéaire ignore cet effet, puisque $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ tient compte uniquement de la variance du plan de sondage. Le fait qu'il n'est pas tenu compte du plan d'expérience dans la méthode d'estimation de la variance devient encore plus évident sous un dénombrement complet de la population finie. En raison du plan d'expérience, la population finie entière est subdivisée aléatoirement en K sous-échantillons et les paramètres sous les divers traitements sont encore estimés avec une variance de plan non nulle. Dans cette situation, il s'ensuit, pour l'approche de la régression linéaire, que $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ et que $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ est nulle parce que la variance de plan induite par le plan d'expérience n'est pas prise en compte. Ceci fait contraste avec (28) qui, sous un recensement complet, reflète encore la variance de plan due au plan d'expérience.

La façon de corriger l'approche de la régression linéaire pour tenir compte de la randomisation due au plan d'échantillonnage ainsi qu'au plan d'expérience n'est pas évidente. Sachant la réalisation de l'échantillon, le plan d'expérience peut être décrit par des probabilités d'inclusion de premier et de deuxième ordres. Soit $\pi_{i|s}^k$ la probabilité d'inclusion de premier ordre que la i^{e} unité d'échantillonnage soit assignée au k^{e} traitement et soit $\pi_{ii'|s}^{kk'}$ la probabilité d'inclusion de deuxième ordre que la i^{e} unité d'échantillonnage soit assignée au k^{e} traitement et que la i'^{e} unité d'échantillonnage soit assignée au k'^{e} traitement. Un estimateur fondé sur le plan de $\boldsymbol{\beta}$, qui tient compte du plan de sondage et du plan d'expérience, est donné par $\hat{\boldsymbol{\beta}} = (\mathbf{X}_n^t\boldsymbol{\Pi}^{*-1}\mathbf{X}_n)^{-1}\mathbf{X}_n^t\boldsymbol{\Pi}^{*-1}\mathbf{y}_n$, où $\boldsymbol{\Pi}^*$ est la matrice diagonale de dimensions $n \times n$ avec les probabilités d'inclusion de premier ordre $\pi_i^* = \pi_i\,\pi_{i|s}^k$. Une approximation de la matrice des covariances de $\hat{\boldsymbol{\beta}}$ est donnée par (38), où $\boldsymbol{\Lambda}$ est obtenue en imposant comme condition la réalisation de l'échantillon, c'est-à-dire

$$\boldsymbol{\Lambda} = \mathrm{Var}_s\mathrm{E}_e(\mathbf{X}_n^t\boldsymbol{\Pi}^{*-1}\mathbf{y}_n - \mathbf{X}_n^t\boldsymbol{\Pi}^{*-1}\mathbf{X}_n\boldsymbol{\beta}) + \mathrm{E}_s\mathrm{Var}_e(\mathbf{X}_n^t\boldsymbol{\Pi}^{*-1}\mathbf{y}_n - \mathbf{X}_n^t\boldsymbol{\Pi}^{*-1}\mathbf{X}_n\boldsymbol{\beta}).$$

Ceci mène pour les éléments de $\boldsymbol{\Lambda}$ à l'expression :

$$\lambda_{qq'} = \sum_{i\in U}\sum_{i'\in U}(\pi_{ii'} - \pi_i\pi_{i'})\frac{x_{iq}e_i}{\pi_i}\frac{x_{i'q'}e_{i'}}{\pi_{i'}}$$

$$+ \sum_{i\in U}\sum_{i'\in U}\pi_{ii'}(\pi^{kk'}_{ii'|s} - \pi^{k}_{i|s}\pi^{k'}_{i'|s})\frac{x_{iq}e_i}{\pi^*_i}\frac{x_{i'q'}e_{i'}}{\pi^*_{i'}},$$

qui a la structure de variance d'un échantillon à deux phases, où la première phase correspond au plan d'échantillonnage et la deuxième, au plan d'expérience. Les unités d'échantillonnage sont, suivant le plan d'expérience, assignées à un seul des K traitements. Il s'ensuit que $\pi^{kk'}_{ii'|s} = 0$ pour $k \neq k'$, et $i = i'$, ce qui rend difficile l'établissement d'un estimateur approximativement sans biais par rapport au plan pour les termes de covariance de $\mathrm{Var}(\hat{\boldsymbol{\beta}})$; voir aussi Van den Brakel et Binder (2000, 2004). Dans la méthode d'analyse proposée à la section 2, ce problème est contourné en établissant un estimateur fondé sur le plan pour la matrice des covariances des contrastes de $\mathbf{C}\hat{\overline{\mathbf{Y}}}_{\mathrm{GREG}}$ au lieu d'un estimateur de la matrice des covariances de $\hat{\overline{\mathbf{Y}}}_{\mathrm{GREG}}$ proprement dit.

4. Étude en simulation

À la sous-section 4.1, nous procédons à une étude en simulation en vue d'évaluer la performance de l'estimateur fondé sur le plan de la matrice des covariances des contrastes entre les estimations sur sous-échantillon $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t$ avec éléments diagonaux donnés par (29), ainsi que la statistique de Wald fondée sur le plan W définie par (32) pour tester les hypothèses au sujet de ces contrastes. Puis, à la sous-section 4.2, nous appliquons ce test de Wald fondé sur le plan, l'approche de la régression linéaire fondée sur le plan et une analyse de variance classique Anova à l'analyse d'un plan en randomisation totale et d'un plan en blocs randomisés.

4.1 Évaluation de l'absence de biais dans $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t$ et de la loi de W

Dans la présente étude en simulation, nous supposons que le modèle de l'erreur de mesure ne contient pas d'effet d'intervieweur, c'est-à-dire que

$$y_{ik} = u_i + \beta_k + \varepsilon_{ik}. \qquad (39)$$

Nous générons une population artificielle constituée de trois strates, 450 UPE et 109 500 USE par tirage aléatoire de valeurs strictement positives pour les valeurs intrinsèques u_i d'un paramètre cible. Les tailles des UPE dans la population sont inégales. Nous générons les valeurs intrinsèques en deux étapes. Premièrement, nous tirons une valeur positive pour chaque UPE dans la population à partir d'une loi uniforme. Puis, nous tirons une valeur positive pour chaque USE, également à partir d'une loi uniforme et nous l'ajoutons à la valeur obtenue pour l'UPE à la première étape. Dans chaque strate nous appliquons des bornes supérieure et inférieure et des largeurs d'intervalle différentes pour ces lois uniformes, de sorte que la population puisse être stratifiée en trois sous-populations relativement homogènes. Les intervalles des lois uniformes qui sont appliquées à la deuxième étape sont plus petits que les intervalles des lois uniformes appliquées à la première étape. Il en résulte une population où les valeurs intrinsèques pour les USE contenues dans chaque UPE sont mises en grappes. La structure de la population est résumée au tableau 1.

Tableau 1
Population

			Valeur intrinsèque du paramètre cible			
Strate	Nombre d'UPE	Nombre d'USE	Moyenne	Écart-type	Valeur min.	Valeur max.
1	70	6 250	22 183	12 001	7 607	50 915
2	130	18 250	6 128	1 866	3 007	10 490
3	250	85 000	1 407	732	512	3 248
Total	450	109 500	3 380	5 803	512	50 915

Nous tirons des échantillons de façon répétée de cette population selon un plan d'échantillonnage stratifié à deux degrés sans remise avec probabilités d'inclusion inégales. Les probabilités d'inclusion sont choisies proportionnellement à la taille du paramètre cible. Les tailles d'échantillon pour les diverses strates sont résumées au tableau 2. Pour chaque échantillon, nous générons une nouvelle erreur de mesure pour chaque élément de population. Ces erreurs de mesure sont tirées d'une loi normale de moyenne nulle et d'écart-type proportionnel à la grandeur des valeurs intrinsèques. La fourchette des écarts-types varie de 1 000 pour les USE avec les valeurs intrinsèques les plus grandes dans la première strate à 10 pour les USE avec les valeurs intrinsèques les plus petites dans la troisième strate.

Tableau 2
Plan d'échantillonnage

Strate	Nombre d'UPE	Nombre d'USE
1	25	900
2	30	1 080
3	50	1 800
Total	105	3 780

Enfin, nous subdivisons aléatoirement les échantillons, conformément à un plan d'expérience, en quatre sous-échantillons contenant chacun 945 USE. Nous appliquons deux plans d'expérience distincts. Dans le premier, les USE sont randomisées sur les quatre traitements différents selon un PRT. Dans le deuxième, elles sont randomisées sur les quatre traitements différents selon un PBR, où les trois strates sont utilisées comme variables de bloc. Dans chaque bloc ou strate, un quart des USE est assigné aléatoirement à chaque traitement. Sous les deux plans d'expérience, quatre ensembles distincts d'effets de traitement sont appliqués, l'un sous l'hypothèse nulle et les trois autres sous des

hypothèses alternatives distinctes. On obtient ainsi huit simulations différentes, dont les spécifications sont données au tableau 3. Chaque simulation est fondée sur $R = 100\,000$ rééchantillonnages. Les observations du paramètre cible sont obtenues en ajoutant une erreur de mesure et un effet de traitement aux valeurs intrinsèques conformément à (39).

Tableau 3
Sommaire des conditions de simulation

Plan d'expérience		Effets de traitement			
		β_1	β_2	β_3	β_4
PRT	PBR	0	0	0	0
PRT	PBR	0	20	40	60
PRT	PBR	0	40	80	120
PRT	PBR	0	80	160	240

Les données obtenues lors de chaque rééchantillonnage sont analysées au moyen de l'estimateur d'Horvitz-Thompson étendu (30). Soit $\tilde{y}_k^r$ l'estimation sur sous-échantillon obtenue sous le k^{e} traitement dans le r^{e} ré-échantillonnage. Le vecteur contenant les quatre estimations sur sous-échantillon obtenues lors du r^{e} rééchantillonnage est donné par $\tilde{\mathbf{Y}}^r = (\tilde{y}_1^r, \tilde{y}_2^r, \tilde{y}_3^r, \tilde{y}_4^r)^t$. Le vecteur avec les trois contrastes pour le r^{e} rééchantillonnage est égal à $\mathbf{C}\tilde{\mathbf{Y}}^r$, avec $\mathbf{C} = (\mathbf{j} \vdots -\mathbf{I})$, $\mathbf{j}$ un vecteur de dimension 3 dont chaque élément est égal à 1 et $\mathbf{I}$ la matrice identité de dimensions 3×3. En outre, $\hat{d}_k^r$ représente les éléments diagonaux de la matrice des covariances estimée, obtenue sous le r^{e} rééchantillonnage. Une expression de $\hat{d}_k^r$ est donnée par (29) avec $\hat{\mathbf{b}}_k^t \mathbf{x}_i = \tilde{y}_k^r$. La matrice estimée des covariances des effets de traitement est égale à $\mathbf{C}\hat{\mathbf{D}}^r\mathbf{C}^t$, avec $\hat{\mathbf{D}}^r = \mathbf{diag}(\hat{d}_1^r, \hat{d}_2^r, \hat{d}_3^r, \hat{d}_4^r)$. Enfin, $W^r = (\mathbf{C}\tilde{\mathbf{Y}}^r)^t(\mathbf{C}\hat{\mathbf{D}}^r\mathbf{C}^t)^{-1}(\mathbf{C}\tilde{\mathbf{Y}}^r)$ représente la statistique de Wald observée lors du r^{e} rééchantillonnage. D'après les $R = 100\,000$ rééchantillonnages pour chaque simulation, les paramètres de population sous les divers traitements peuvent être approximés par

$$\overline{\mathbf{Y}} = \frac{1}{R}\sum_{r=1}^{R} \tilde{\mathbf{Y}}^r,$$

avec $\overline{\mathbf{Y}} = (\overline{Y}_1, \overline{Y}_2, \overline{Y}_3, \overline{Y}_4)^t$. De (10) il découle que les effets de traitement réels dans le modèle de l'erreur de mesure peuvent être approximés par $\mathbf{C}\overline{\mathbf{Y}} \approx \mathbf{C}\boldsymbol{\beta}$. En outre, la moyenne des matrices estimées des covariances de rééchantillonnage peut être calculée selon

$$\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t = \frac{1}{R}\sum_{r=1}^{R} \mathbf{C}\hat{\mathbf{D}}^r\mathbf{C}^t,$$

et la moyenne des statistiques de Wald de rééchantillonnage selon

$$\overline{W} = \frac{1}{R}\sum_{r=1}^{R} W^r. \qquad (40)$$

Une approximation de la matrice réelle des covariances des effets de traitement est donnée par

$$\mathbf{CVC}^t = \frac{1}{R-1}\sum_{r=1}^{R} \mathbf{C}\,(\tilde{\mathbf{Y}}^r - \overline{\mathbf{Y}})\,(\tilde{\mathbf{Y}}^r - \overline{\mathbf{Y}})^t\mathbf{C}^t. \qquad (41)$$

Nous évaluons la performance de la méthode d'estimation de la variance par comparaison de $\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$ à $\mathbf{CVC}^t$. Si l'estimateur de la variance établi $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t$ est approximativement sans biais par rapport au plan, alors la moyenne des matrices des covariances de rééchantillonnage $\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$ doit tendre vers la matrice réelle des covariances $\mathbf{CVC}^t$, pour $R \to \infty$. Le calcul de l'écart-type des éléments de $\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$, noté $\sigma(\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t)$, donne une idée de la précision de l'estimateur de la variance établi. Les éléments diagonaux de $\overline{\mathbf{D}}$ sont notés $\overline{d}_k$.

Si $\mathbf{C}\hat{\overline{\mathbf{Y}}}_{\text{GREG}} \to N(\mathbf{C}\boldsymbol{\beta}, \mathbf{CVC}^t)$, il s'ensuit que $W \to \chi^2_{[K-1][\delta]}$, où $K-1$ est le nombre de degrés de liberté et $\delta = 1/2(\mathbf{C}\boldsymbol{\beta})^t(\mathbf{CVC}^t)^{-1}(\mathbf{C}\boldsymbol{\beta})$ est le paramètre de non-centralité de la loi du chi-carré. Dans l'étude en simulation, nous pouvons calculer le paramètre de non-centralité sous les hypothèses alternatives en insérant (41) dans l'expression de δ. Ensuite, nous pouvons calculer la puissance de la statistique de Wald pour un ensemble particulier d'effets de traitement par $P(W) = P(\chi^2_{[K-1][\delta]} > \chi^2_{[1-\alpha][K-1]})$, où $\chi^2_{[1-\alpha][K-1]}$ est le $(1-\alpha)^{\text{e}}$ percentile de la loi du chi-carré centrée à $K-1$ degrés de liberté. Nous évaluons les propriétés de la statistique de Wald en comparant $P(W)$ à la puissance simulée, qui est définie comme étant la fraction d'exécutions significatives observées dans les R rééchantillonnages, c'est-à-dire

$$P^{\text{sim}}(W) = \frac{1}{R}\sum_{r=1}^{R} I(W^r > \chi^2_{[1-\alpha][K-1]}),$$

où $I(B)$ est la variable indicatrice qui est égale à l'unité si B est vrai et nulle autrement. Les résultats des simulations sont résumés aux tableaux 4.1 à 4.8.

Les moyennes des estimations sur les sous-échantillons $\overline{Y}_k$ sous l'hypothèse nulle présentées aux tableaux 4.1 et 4.5 surestiment légèrement la moyenne de population donnée au tableau 1. Cette différence peut être attribuée au biais de l'estimateur étendu d'Horvitz-Thompson. Par contre, les moyennes des contrastes entre les estimations sur les sous-échantillons $\mathbf{C}\overline{\mathbf{Y}}$ concordent presque parfaitement avec les effets de traitement réels $\mathbf{C}\boldsymbol{\beta}$. Les moyennes des matrices des covariances de rééchantillonnage $\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$ tendent vers les valeurs des matrices des covariances réelles $\mathbf{CVC}^t$, résultat qui montrent que la méthode d'estimation de la variance établie à la section 2.4 est approximativement sans biais par rapport au plan. La précision relative des éléments diagonaux de $\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$ est d'environ 10,5 % sous cette taille d'échantillon particulière. La puissance simulée fondée sur la distribution de rééchantillonnage de la

statistique de Wald approxime raisonnablement bien la puissance réelle. En moyenne, la puissance simulée est légèrement plus élevée. L'espérance de la loi du chi-carré est égale à $E(\chi^2_{[K-1][\delta]}) = (K-1) + 2\delta$ (Searle 1971, section 2.4.h). Si la distribution de rééchantillonnage de la statistique de Wald tend vers une loi $\chi^2_{[K-1][\delta]}$, alors la moyenne de la statistique de Wald de rééchantillonnage $\overline{W}$ (40) doit tendre vers l'espérance de la loi du chi-carré. En effet, il découle des tableaux 4.1 à 4.8 que $\overline{W} \approx (K-1) + 2\delta$. De plus, nous vérifions l'hypothèse que la distribution de rééchantillonnage de la statistique de Wald sous l'hypothèse nulle est égale à la loi du chi-carré centrée au moyen du test de Kolmogorov-Smirnov pour un échantillon. Cette hypothèse n'est rejetée au seuil de signification de 5 % ni pour le PRT ni pour le PBR, et confirme la conjecture selon laquelle la statistique de Wald suit asymptotiquement une loi du chi-carré sous échantillonnage stratifié à deux degrés sans remise, probabilités d'inclusion inégales et fractions d'échantillonnage relativement grandes. La comparaison des simulations sous un PRT et sous un PBR montre que la constitution de blocs sur les strates augmente considérablement la précision des contrastes estimés et la puissance des tests dans cette situation particulière.

Tableau 4.1
Résultats de la simulation PRT $\boldsymbol{\beta} = (0, 0, 0, 0)^t$

Sous-échantillons				Contrastes					Statistique de Wald		
						Éléments diagonaux de					
k	β_k	$\overline{Y}_k$	$\overline{d}_k$	$k-k'$	$\mathbf{C}\overline{\mathbf{Y}}$	$\mathbf{CVC}^t$	$\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$	$\sigma(\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t)$	α	$P(W)$	$P^{\text{sim}}(W)$
1	0	3 392	14 311	1 – 2	0	28 725	28 616	3 019	0,050	0,05000	0,05072
2	0	3 392	14 305	1 – 3	0	28 892	28 616	3 019	0,025	0,02500	0,02506
3	0	3 392	14 306	1 – 4	2	28 787	28 603	3 019	0,010	0,01000	0,01017
4	0	3 390	14 292						$\overline{W}$: 3,01591		δ : 0,0000

Tableau 4.2
Résultats de la simulation PRT $\boldsymbol{\beta} = (0, 20, 40, 60)^t$

Sous-échantillons				Contrastes					Statistique de Wald		
						Éléments diagonaux de					
k	β_k	$\overline{Y}_k$	$\overline{d}_k$	$k-k'$	$\mathbf{C}\overline{\mathbf{Y}}$	$\mathbf{CVC}^t$	$\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$	$\sigma(\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t)$	α	$P(W)$	$P^{\text{sim}}(W)$
1	0	3 392	14 307	1 – 2	– 20	28 635	28 614	3 026	0,050	0,05842	0,05925
2	20	3 412	14 307	1 – 3	– 40	28 918	28 620	3 033	0,025	0,03008	0,03040
3	40	3 432	14 314	1 – 4	– 58	28 624	28 597	3 025	0,010	0,01257	0,01255
4	60	3 450	14 291						$\overline{W}$: 3,14037		δ : 0,0697

Tableau 4.3
Résultats de la simulation PRT $\boldsymbol{\beta} = (0, 40, 80, 120)^t$

Sous-échantillons				Contrastes					Statistique de Wald		
						Éléments diagonaux de					
k	β_k	$\overline{Y}_k$	$\overline{d}_k$	$k-k'$	$\mathbf{C}\overline{\mathbf{Y}}$	$\mathbf{CVC}^t$	$\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$	$\sigma(\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t)$	α	$P(W)$	$P^{\text{sim}}(W)$
1	0	3 392	14 314	1 – 2	– 40	28 597	28 621	3 020	0,050	0,08503	0,08523
2	40	3 432	14 307	1 – 3	– 80	28 947	28 622	3 022	0,025	0,04704	0,04760
3	80	3 472	14 307	1 – 4	– 119	28 713	28 609	3 021	0,010	0,02150	0,02165
4	120	3 511	14 295						$\overline{W}$: 3,55406		δ : 0,2783

Tableau 4.4
Résultats de la simulation PRT $\boldsymbol{\beta} = (0, 80, 160, 240)^t$

Sous-échantillons				Contrastes					Statistique de Wald		
						Éléments diagonaux de					
k	β_k	$\overline{Y}_k$	$\overline{d}_k$	$k-k'$	$\mathbf{C}\overline{\mathbf{Y}}$	$\mathbf{CVC}^t$	$\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$	$\sigma(\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t)$	α	$P(W)$	$P^{\text{sim}}(W)$
1	0	3 392	14 306	1 – 2	– 80	28 748	28 616	3 026	0,050	0,21198	0,2116
2	80	3 472	14 310	1 – 3	– 160	28 784	28 618	3 030	0,025	0,13809	0,13885
3	160	3 552	14 312	1 – 4	– 239	28 538	28 598	3 022	0,010	0,07703	0,07781
4	240	3 631	14 291						$\overline{W}$: 5,22065		δ : 1,1203

Tableau 4.5
Résultats de la simulation PBR $\boldsymbol{\beta} = (0, 0, 0, 0)^t$

Sous-échantillons				Contrastes					Statistique de Wald		
						Éléments diagonaux de					
k	β_k	$\overline{Y}_k$	$\overline{d}_k$	$k-k'$	$\mathbf{C}\overline{\mathbf{Y}}$	$\mathbf{CVC}^t$	$\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t$	$\sigma(\mathbf{C}\overline{\mathbf{D}}\mathbf{C}^t)$	α	$P(W)$	$P^{\text{sim}}(W)$
1	0	3 389	3 088	1 – 2	0	6 175	6 176	647	0,050	0,05000	0,05168
2	0	3 389	3 088	1 – 3	0	6 216	6 176	647	0,025	0,02500	0,02640
3	0	3 389	3 088	1 – 4	0	6 217	6 176	647	0,010	0,01000	0,01060
4	0	3 389	3 088						$\overline{W}$: 3,01483		δ : 0,0000

4.2 Comparaison de trois méthodes d'analyse

De surcroît, nous comparons trois méthodes d'analyse possibles pour les expériences intégrées, c'est-à-dire le test de Wald fondé sur le plan proposé à la section 2, une analyse de variance Anova classique où les observations sont équipondérées et considérées comme étant i.i.d., et l'approche de la régression linéaire fondée sur le plan décrite à la section 3. À cette fin, nous tirons deux échantillons, chacun de taille égale à 3 780 USE, de la population finie spécifiée au tableau 1, selon le plan d'échantillonnage stratifié à deux degrés qui a été utilisé pour la simulation précédente (voir le tableau 2). Nous répartissons les USE de l'un de ces échantillons aléatoirement en quatre sous-échantillons, chacun de taille égale à 945, selon un PRT et les USE de l'autre échantillon aléatoirement en quatre sous-échantillons, chacun de taille égale à 945, selon un PBR où les strates sont utilisées comme variables de bloc. Les deux expériences sont réalisées sous l'hypothèse alternative que les effets de traitement dans la population finie sont égaux à $\boldsymbol{\beta} = (0, 80, 160, 240)^t$. Nous exécutons l'analyse par régression linéaire fondée sur le plan au moyen de la procédure SVYREG de Stata qui tient compte de la stratification, de l'échantillonnage à deux degrés et des probabilités de sélection inégales du plan d'échantillonnage (StataCorp. 2001). Nous exécutons l'analyse de variance avec la procédure ANOVA de Stata (StataCorp. 2001). Les résultats de l'analyse sous un PRT sont résumés au tableau 5.1 pour le test de Wald fondé sur le plan, au tableau 5.2, pour l'approche par la régression linéaire fondée sur le plan et au tableau 5.3 pour l'analyse de variance. De même, les résultats de l'analyse sous un PBR sont résumés aux tableaux 6.1, 6.2 et 6.3.

Tableau 4.6
Résultats de la simulation PBR $\boldsymbol{\beta} = (0, 20, 40, 60)^t$

Sous-échantillons				Contrastes					Statistique de Wald		
						Éléments diagonaux de					
k	β_k	$\bar{Y}_k$	$\bar{d}_k$	$k-k'$	$\mathbf{C}\bar{\mathbf{Y}}$	$\mathbf{CVC}^t$	$\mathbf{C}\bar{\mathbf{D}}\mathbf{C}^t$	$\sigma(\mathbf{C}\bar{\mathbf{D}}\mathbf{C}^t)$	α	$P(W)$	$P^{\text{sim}}(W)$
1	0	3 390	3 090						0,050	0,09099	0,09371
2	20	3 410	3 089	1 – 2	– 20	6 225	6 180	648	0,025	0,05096	0,05238
3	40	3 430	3 090	1 – 3	– 40	6 177	6 181	648	0,010	0,02365	0,02405
4	60	3 450	3 090	1 – 4	– 60	6 184	6 180	649	$\bar{W}$: 3,66771		δ : 0,3226

Tableau 4.7
Résultats de la simulation PBR $\boldsymbol{\beta} = (0, 40, 80, 120)^t$

Sous-échantillons				Contrastes					Statistique de Wald		
						Éléments diagonaux de					
k	β_k	$\bar{Y}_k$	$\bar{d}_k$	$k-k'$	$\mathbf{C}\bar{\mathbf{Y}}$	$\mathbf{CVC}^t$	$\mathbf{C}\bar{\mathbf{D}}\mathbf{C}^t$	$\sigma(\mathbf{C}\bar{\mathbf{D}}\mathbf{C}^t)$	α	$P(W)$	$P^{\text{sim}}(W)$
1	0	3 389	3 088						0,050	0,23999	0,24310
2	40	3 429	3 088	1 – 2	– 40	6 178	6 176	647	0,025	0,15999	0,16302
3	80	3 469	3 088	1 – 3	– 80	6 183	6 176	649	0,010	0,09181	0,09458
4	120	3 509	3 088	1 – 4	– 120	6 189	6 176	649	$\bar{W}$: 5,62182		δ : 1,2905

Tableau 4.8
Résultats de la simulation PBR $\boldsymbol{\beta} = (0, 80, 160, 240)^t$

Sous-échantillons				Contrastes					Statistique de Wald		
						Éléments diagonaux de					
k	β_k	$\bar{Y}_k$	$\bar{d}_k$	$k-k'$	$\mathbf{C}\bar{\mathbf{Y}}$	$\mathbf{CVC}^t$	$\mathbf{C}\bar{\mathbf{D}}\mathbf{C}^t$	$\sigma(\mathbf{C}\bar{\mathbf{D}}\mathbf{C}^t)$	α	$P(W)$	$P^{\text{sim}}(W)$
1	0	3 390	3 091						0,050	0,77340	0,77712
2	80	3 470	3 090	1 – 2	– 80	6 204	6 180	648	0,025	0,68135	0,68789
3	160	3 550	3 090	1 – 3	– 160	6 210	6 181	648	0,010	0,55796	0,56701
4	240	3 630	3 090	1 – 4	– 240	6 214	6 181	648	$\bar{W}$: 13,48594		δ : 5,1331

Tableau 5.1
Statistique de Wald fondée sur le plan, PRT

Sous-échantillons			Contrastes			Statistique de Wald		
k	β_k	$\tilde{y}_k$	$k-k'$	$\tilde{y}_k - \tilde{y}_{k'}$	$\sqrt{\hat{d}_k + \hat{d}_{k'}}$	W	ddl	Valeur p
1	0	3 414	1 – 2	– 124	164,915	2,4740	3	0,480
2	80	3 538	1 – 3	– 182	162,542			
3	160	3 596	1 – 4	– 249	164,782			
4	240	3 663						

Tableau 5.2
Régression fondée sur le plan, PRT

Source	Coefficient	Erreur-type	Statistique de Wald	*ddl*	Valeur p
Traitement			2,907	3	0,4062
Traitement 1	– 182,14	177,60			
Traitement 2	– 58,36	175,56			
Traitement 4	66,79	170,46			
Constante	3 596,47	194,75			

Tableau 5.3
Analyse de variance classique, PRT

k	β_k	$\bar{y}_k$	Contraste		ANOVA				
1	0	8021	$k-k'$	$\bar{y}_k - \bar{y}_{k'}$	Source	*ddl*	*CM*	*F*	Valeur p
2	80	8094	1 – 2	– 73	Entre traitements	3	14 432 816	0,14	0,9376
3	160	7955	1 – 3	66	Résidu	3 776	104 924 668		
4	240	8242	1 – 4	– 221	Total	3 779			

Tableau 6.1
Statistique de Wald fondée sur le plan, PBR

Sous-échantillons			Contrastes			Statistique de Wald		
k	β_k	$\tilde{y}_k$	$k-k'$	$\tilde{y}_k - \tilde{y}_{k'}$	$\sqrt{\hat{d}_k + \hat{d}_{k'}}$	*W*	*ddl*	Valeur p
1	0	3 395	1 – 2	– 25	81,247	9,93011	3	0,0192
2	80	3 420	1 – 3	– 120	80,697			
3	160	3 515	1 – 4	– 231	82,383			
4	240	3 626						

Tableau 6.2
Régression fondée sur le plan, PBR

Source	Coefficient	Erreur-type	Statistique de Wald	*ddl*	Valeur p
Bloc					
Bloc 2	–17 068,28	2 556,46			
Bloc 3	–21 999,39	2 540,98			
Traitement			18,4212	3	0,00036
Traitement 1	–211,51	74,84			
Traitement 2	–246,78	60,05			
Traitement 3	–97,91	73,39			
Constante	23 589,64	2543,25			

Tableau 6.3
Analyse de variance classique, PBR

k	β_k	$\bar{y}_{+k}$	Contraste		ANOVA				
1	0	8 815	$k-k'$	$\bar{y}_{+k} - \bar{y}_{+k'}$	Source	*ddl*	*CM*	*F*	Valeur p
2	80	8 150	1 – 2	665	Entre blocs	2	1,6773 E+11		
3	160	8 566	1 – 3	249	Entre traitements	3	84 377 227	1,99	0,1126
4	240	8 746	1 – 4	69	Résidu	3 774	42 310 035		
					Total	3 779	131 089 505		

Comme nous le soulignons à la section 3, l'approche par la régression linéaire ne tient pas compte de la variance de plan due à la randomisation des unités d'échantillonnage sur les sous-échantillons selon le plan d'expérience. Par conséquent, les erreurs-types des effets de traitement sont plus faibles sous l'approche par la régression linéaire que dans le cas du test de Wald fondé sur le plan, et l'approche de la régression fondée sur le plan produit des valeurs p plus faibles pour le test des effets de traitement.

L'analyse de variance classique est une approche naïve, puisqu'elle ne tient pas compte de la stratification, de la mise en grappes et de la sélection des unités d'échantillonnage en utilisant des probabilités d'inclusion proportionnelles à la valeur du paramètre cible. Omettre de tenir compte de ces aspects du plan d'échantillonnage dans l'analyse a pour résultat net d'exagérer fortement les estimations sur les sous-échantillons, ainsi que les erreurs-types. Comparativement aux deux autres méthodes fondées sur le plan

de sondage, cette approche produit des valeurs p plus grandes pour le test des effets de traitement.

Un autre avantage important du test de Wald fondé sur le plan de sondage comparativement à l'approche par la régression linéaire fondée sur le plan de sondage est qu'il a toujours trait aux différences entre les estimations sur les sous-échantillons, ce qui facilite l'interprétation des résultats. Cette propriété est particulièrement importante pour les expériences intégrées visant à quantifier des ruptures de tendance concernant les paramètres d'une enquête causées par des ajustements du plan de sondage. Dans le cas d'un plan en randomisation totale, le modèle de régression linéaire est constitué d'une ordonnée à l'origine et de trois coefficients pour les effets de traitement. Dans cette situation particulièrement simple, les coefficients des effets de traitement sont exactement égaux aux différences entre les estimations sur les sous-échantillons. Cependant, cette propriété ne tient pas pour les effets de traitement obtenus sous des modèles plus complexes, comme le cas du plan en blocs randomisés.

5. Discussion et conclusions

Nous discutons dans le présent article de la manière dont la méthodologie statistique des expériences randomisées et de l'échantillonnage aléatoire peut appuyer la conception et l'analyse d'expériences intégrées dans les enquêtes par sondage courantes. Le plan de l'enquête par sondage constitue un cadre a priori pour l'application de principes, tirés de la théorie des plans expérimentaux, tels que la randomisation et le contrôle local par constitution de blocs sur les strates, les UPE, les grappes ou les intervieweurs. Pour tester les hypothèses au sujet des estimations des paramètres de population finie obtenues sous divers traitements de l'expérience, nous établissons une statistique de Wald fondée sur le plan de sondage pour l'analyse des plans en randomisation totale et des plans en blocs randomisés intégrés dans des plans d'échantillonnage complexes généraux en utilisant l'estimateur d'Horvitz-Thompson et l'estimateur par la régression généralisée. La combinaison de l'échantillonnage aléatoire d'une population finie et de cette méthode d'analyse fondée sur le plan de sondage nous permet de généraliser les résultats de l'expérience observés dans l'échantillon particulier à l'ensemble de la population d'enquête.

Puisque nous tenons compte de plans de sondage complexes généraux, nous nous attendons à obtenir une expression assez compliquée pour la matrice des covariances des effets de traitement, avec des éléments hors diagonal non nuls. Cependant, l'estimateur établi pour cette matrice des covariances a la même structure que si les unités d'échantillonnage étaient tirées avec remise et avec probabilités de sélection inégales. Ni les probabilités d'inclusion de deuxième ordre ni les covariances par rapport au plan de sondage entre les estimations sur sous-échantillon ne doivent être connues, ce qui simplifie considérablement l'analyse. Ainsi, dans le cas de l'échantillonnage aléatoire simple avec remise, ce résultat signifie que l'on devrait laisser tomber le facteur de correction pour population finie dans l'estimation de la variance des contrastes. Par conséquent, nous obtenons une statistique de Wald, établie dans une perspective axée sur le plan de sondage sous des plans de sondage complexes généraux, qui retient la structure assez simple intéressante des méthodes classiques d'analyse fondées sur un modèle.

Pour les PRT et les PBR intégrés dans un plan d'échantillonnage autopondéré analysés au moyen de l'estimateur étendu d'Horvitz-Thompson et d'un estimateur groupé de la variance, la statistique de Wald coïncide avec la statistique F d'une analyse de variance à un ou à deux critères de classification. Pour l'analyse de l'expérience à deux traitements intégrés, on peut établir une version fondée sur le plan de sondage de la statistique t en tant que cas particulier de la statistique de Wald. Les expressions et des renseignements supplémentaires au sujet de cette statistique t fondée sur le plan de sondage et de sa relation avec la statistique t de Welch et avec la statistique t standard figurent dans Van den Brakel et Renssen (1998), Van den Brakel (2001) ou Van den Brakel et Van Berkel (2002).

La méthode d'analyse proposée dans le présent article est implémentée dans un progiciel appelé X-tool. Cet outil sera disponible en tant que composante du progiciel de traitement des données d'enquête Blaise développé par Statistique Pays-Bas.

Annexe

Propriétés des vecteurs de randomisation $\mathbf{p}_{ik}$

Pour les PRT et les PBR, les vecteurs de randomisation $\mathbf{p}_{ik}$ sont définis par (14) et (15). En raison du mécanisme de randomisation du plan d'expérience, les vecteurs $\mathbf{p}_{ik}$ sont aléatoires et ont les fonctions de masse de probabilité conditionnelle qui suivent. Pour un PRT, nous avons

$$P\left(\mathbf{p}_{ik}=\frac{n_{+}}{n_k}\mathbf{r}_k \mid s\right)=\frac{n_k}{n_{+}} \quad \text{et} \quad P(\mathbf{p}_{ik}=0 \mid s)=1-\frac{n_k}{n_{+}}.$$

Pour un PBR, nous avons

$$P\left(\mathbf{p}_{ik}=\frac{n_{j+}}{n_{jk}}\mathbf{r}_k \mid s_j\right)=\frac{n_{jk}}{n_{j+}} \text{ et } P(\mathbf{p}_{ik}=0 \mid s_j)=1-\frac{n_{jk}}{n_{j+}}.$$

Nous établissons les propriétés de ces vecteurs pour un PBR. Les propriétés pour un PRT en découlent comme cas

particulier, puisqu'un PRT peut être considéré comme un PBR à un bloc. Notons a. pr. pour « avec probabilité ».

$$\mathbf{p}_{ik}\mathbf{p}_{ik}^t = \begin{cases} \left(\dfrac{n_{j+}}{n_{jk}}\right)^2 \mathbf{r}_k\mathbf{r}_k^t & \text{a. pr.: } \dfrac{n_{jk}}{n_{j+}} \quad \text{si } i\in s_j \\ \mathbf{O} & \text{a. pr.: } 1-\dfrac{n_{jk}}{n_{j+}} \end{cases}$$

$$\mathbf{p}_{ik}\mathbf{p}_{ik'}^t = \begin{cases} \dfrac{n_{j+}}{n_{jk}}\dfrac{n_{j+}}{n_{jk'}}\mathbf{r}_k\mathbf{r}_{k'}^t & \text{a. pr.: } 0 \quad \text{si } i\in s_j \\ \mathbf{O} & \text{a. pr.: } 1 \end{cases}$$

$$\mathbf{p}_{ik}\mathbf{p}_{i'k'}^t = \begin{cases} \dfrac{n_{j+}}{n_{jk}}\dfrac{n_{j+}}{n_{jk'}}\mathbf{r}_k\mathbf{r}_{k'}^t & \text{a. pr.: } \dfrac{n_{jk}}{n_{j+}}\dfrac{n_{jk'}}{(n_{j+}-1)}, \\ & \quad \text{si } i\in s_j, i'\in s_j \\ \mathbf{O} & \text{a. pr.: } 1-\dfrac{n_{jk}}{n_{j+}}\dfrac{n_{jk'}}{(n_{j+}-1)}, \\ & \quad \text{si } i\in s_j, i'\in s_j \\ \dfrac{n_{j+}}{n_{jk}}\dfrac{n_{j'+}}{n_{j'k'}}\mathbf{r}_k\mathbf{r}_{k'}^t & \text{a. pr.: } \dfrac{n_{jk}}{n_{j+}}\dfrac{n_{j'k'}}{n_{j'+}}, \\ & \quad \text{si } i\in s_j, i'\in s_{j'} \\ \mathbf{O} & \text{a. pr.: } 1-\dfrac{n_{jk}}{n_{j+}}\dfrac{n_{j'k'}}{n_{j'+}}, \\ & \quad \text{si } i\in s_j,\ i'\in s_{j'} \end{cases}$$

$$\mathbf{p}_{ik}\mathbf{p}_{i'k}^t = \begin{cases} \left(\dfrac{n_{j+}}{n_{jk}}\right)^2 \mathbf{r}_k\mathbf{r}_k^t & \text{a. pr.: } \dfrac{n_{jk}}{n_{j+}}\dfrac{(n_{jk}-1)}{(n_{j+}-1)}, \text{si } i\in s_j, i'\in s_j \\ \mathbf{O} & \text{a. pr.: } 1-\dfrac{n_{jk}}{n_{j+}}\dfrac{(n_{jk}-1)}{(n_{j+}-1)}, \text{si } i\in s_j\ i'\in s_j \\ \dfrac{n_{j+}}{n_{jk}}\dfrac{n_{j'+}}{n_{j'k}}\mathbf{r}_k\mathbf{r}_k^t & \text{a. pr.: } \dfrac{n_{jk}}{n_{j+}}\dfrac{n_{j'k}}{n_{j'+}}, \text{si } i\in s_j\ i'\in s_{j'} \\ \mathbf{O} & \text{a. pr.: } 1-\dfrac{n_{jk}}{n_{j+}}\dfrac{n_{j'k}}{n_{j'+}}, \text{si } i\in s_j\ i'\in s_{j'} \end{cases}$$

L'espérance de $\mathbf{p}_{ik}$ par rapport au plan d'expérience est donnée par :

$$E_e(\mathbf{p}_{ik}) = P\left(\mathbf{p}_{ik} = \frac{n_{j+}}{n_{jk}}\mathbf{r}_k\right)\frac{n_{j+}}{n_{jk}}\mathbf{r}_k + P(\mathbf{p}_{ik}=\mathbf{0})\mathbf{0} = \mathbf{r}_k. \quad (42)$$

Nous pouvons établir les covariances qui suivent par rapport au plan d'expérience :

$$\mathrm{Cov}_e(\mathbf{p}_{ik}\,\mathbf{p}_{ik}^t) = \frac{(n_{j+}-n_{jk})}{n_{jk}}\mathbf{r}_k\,\mathbf{r}_k^t \quad (43)$$

$$\mathrm{Cov}_e(\mathbf{p}_{ik}\,\mathbf{p}_{ik'}^t) = -\mathbf{r}_k\,\mathbf{r}_{k'}^t \quad (44)$$

$$\mathrm{Cov}_e(\mathbf{p}_{ik}\mathbf{p}_{i'k'}^t) = \begin{cases} \dfrac{1}{(n_{j+}-1)}\mathbf{r}_k\,\mathbf{r}_{k'}^t & \text{si } i\in s_j \text{ et } i'\in s_j \\ \mathbf{O} & \text{si } i\in s_j \text{ et } i'\in s_{j'} \end{cases} \quad (45)$$

$$\mathrm{Cov}_e(\mathbf{p}_{ik}\,\mathbf{p}_{i'k}^t) = \begin{cases} -\dfrac{(n_{j+}-n_{jk})}{n_{jk}}\dfrac{1}{(n_{j+}-1)}\mathbf{r}_k\mathbf{r}_k^t & \text{si } i\in s_j \text{ et } i'\in s_j \\ \mathbf{O} & \text{si } i\in s_j \text{ et } i'\in s_{j'} \end{cases} \quad (46)$$

Preuve de la formule (23)

Sous la condition énoncée qu'il existe un vecteur $\mathbf{a}$ de dimension H constant tel que $\mathbf{a}^t\mathbf{x}_i = 1$ pour tout $i\in U$, et sachant la réalisation de $u_i, i=1,\ldots,N$, il découle du modèle de superpopulation (16) que $\tilde{\mathbf{b}}_k$ dans (18) peut être évalué sous la forme

$$\begin{aligned} E_m(\tilde{\mathbf{b}}_k) &= E_m\left(\sum_{i=1}^N \frac{\mathbf{x}_i\mathbf{x}_i^t}{\omega_i^2}\right)^{-1}\sum_{i=1}^N \frac{\mathbf{x}_i y_{ik}}{\omega_i^2} \\ &= \left(\sum_{i=1}^N \frac{\mathbf{x}_i\mathbf{x}_i^t}{\omega_i^2}\right)^{-1}\sum_{i=1}^N \frac{\mathbf{x}_i(u_i+\psi_I)}{\omega_i^2} + \mathbf{a}\,\beta_k \\ &= \mathbf{b}+\mathbf{d}+\mathbf{a}\,\beta_k, \end{aligned} \quad (47)$$

où $\mathbf{b}$ représente les coefficients de régression définis par (17) et $\mathbf{d}$, les coefficients de régression de la fonction de régression des effets d'intervieweur sur les variables auxiliaires $\mathbf{x}_i$. Du résultat (47) il découle que $\mathbf{B}^t\mathbf{x}_i = \mathbf{j}(\mathbf{b}^t\mathbf{x}_i + \mathbf{d}^t\mathbf{x}_i) + \boldsymbol{\beta}$. Puisque $\mathbf{C}\,\mathbf{j} = \mathbf{0}$, et d'après le modèle de l'erreur de mesure (1) et le modèle de régression linéaire (16), il découle que

$$\begin{aligned} \mathbf{C}(\mathbf{y}_i - \mathbf{B}^t\mathbf{x}_i) &= \mathbf{C}(\mathbf{j}u_i + \mathbf{j}\psi_{il} + \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i - \mathbf{j}(\mathbf{b}^t + \mathbf{d}^t)\mathbf{x}_i - \boldsymbol{\beta}) \\ &= \mathbf{C}\boldsymbol{\varepsilon}_i, \quad \textbf{C.Q.D.F.} \end{aligned}$$

Preuve de la formule (26) pour un PBR

Nous commençons par établir une expression pour $\mathrm{Cov}_e(\mathbf{C}\hat{\bar{\mathbf{E}}}_{\mathrm{HT}} \mid m, s)$. Soit $\mathbf{e}_i = (e_{i1},\ldots,e_{iK})^t$ un vecteur de dimension K avec éléments $e_{ik} = y_{ik} - \mathbf{b}_k^t\mathbf{x}_i$. Par conséquent, $\mathbf{e}_i = \mathbf{y}_i - \mathbf{B}^t\mathbf{x}_i$. Notons que $\mathrm{E}_m\mathrm{E}_s\mathrm{Cov}_e(\mathbf{C}\hat{\bar{\mathbf{E}}}_{\mathrm{HT}} \mid m, s) = \mathbf{C}\mathrm{E}_m\mathrm{E}_s\mathrm{Cov}_e(\hat{\bar{\mathbf{E}}}_{\mathrm{HT}} \mid m, s)\mathbf{C}^t$ avec $\hat{\bar{\mathbf{E}}}_{\mathrm{HT}} = (\hat{\bar{E}}_{1;\mathrm{HT}},\ldots,\hat{\bar{E}}_{\mathrm{K};\mathrm{HT}})^t$. En outre, notons que

$$\hat{\bar{E}}_{k;\mathrm{HT}} = \sum_{i=1}^{n_{++}} \left(\frac{\mathbf{p}_{ik}^t (\mathbf{y}_i - \mathbf{B}^t \mathbf{x}_i)}{\pi_i N} \right) = \sum_{i=1}^{n_{++}} \frac{\mathbf{p}_{ik}^t \mathbf{e}_i}{\pi_i N}. \qquad (48)$$

En nous servant de (43) et (46), nous pouvons élaborer les éléments diagonaux de $\mathrm{Cov}_e(\hat{\bar{\mathbf{E}}}_{\mathrm{HT}} \mid m, s)$ comme suit

$$\mathrm{Var}_e(\hat{\bar{E}}_{k;\mathrm{HT}} \mid m, s)$$

$$= \mathrm{Cov}_e \left(\sum_{i=1}^{n_{++}} \frac{\mathbf{p}_{ik}^t \mathbf{e}_i}{\pi_i N}, \sum_{i'=1}^{n_{++}} \frac{\mathbf{p}_{i'k}^t \mathbf{e}_{i'}}{\pi_{i'} N} \mid m, s \right)$$

$$= \sum_{j=1}^{J} \left(\begin{array}{l} \displaystyle\sum_{i=1}^{n_{j+}} \frac{\mathbf{e}_i^t}{\pi_i N} \mathrm{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{ik}^t \mid m, s) \frac{\mathbf{e}_i}{\pi_i N} \\ \displaystyle + \sum_{i=1}^{n_{j+}} \sum_{i' \neq i=1}^{n_{j+}} \frac{\mathbf{e}_i^t}{\pi_i N} \mathrm{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'k}^t \mid m, s) \frac{\mathbf{e}_{i'}}{\pi_{i'} N} \end{array} \right)$$

$$= \sum_{j=1}^{J} \left(\begin{array}{l} \displaystyle\frac{n_{j+}}{(n_{j+}-1)} \frac{n_{j+}}{n_{jk}} \sum_{i=1}^{n_{j+}} \left(\frac{e_{ik}}{\pi_i N} - \frac{1}{n_{j+}} \sum_{i=1}^{n_{j+}} \frac{e_{ik}}{\pi_i N} \right)^2 \\ \displaystyle - \frac{n_{j+}}{(n_{j+}-1)} \sum_{i=1}^{n_{j+}} \left(\frac{e_{ik}}{\pi_i N} - \frac{1}{n_{j+}} \sum_{i=1}^{n_{j+}} \frac{e_{ik}}{\pi_i N} \right)^2 \end{array} \right). \qquad (49)$$

En nous servant de (44) et (45), nous pouvons élaborer les éléments hors diagonale de $\mathrm{Cov}_e(\hat{\bar{\mathbf{E}}}_{\mathrm{HT}} \mid m, s)$ comme suit

$$\mathrm{Cov}_e(\hat{\bar{E}}_{k;\mathrm{HT}}, \hat{\bar{E}}_{k';\mathrm{HT}} \mid m, s)$$

$$= \mathrm{Cov}_e \left(\sum_{i=1}^{n_{++}} \frac{\mathbf{p}_{ik}^t \mathbf{e}_i}{\pi_i N}, \sum_{i'=1}^{n_{++}} \frac{\mathbf{p}_{i'k'}^t \mathbf{e}_{i'}}{\pi_{i'} N} \mid m, s \right)$$

$$= \sum_{j=1}^{J} \left(\begin{array}{l} \displaystyle\sum_{i=1}^{n_{j+}} \frac{\mathbf{e}_i^t}{\pi_i N} \mathrm{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{ik'}^t \mid m, s) \frac{\mathbf{e}_i}{\pi_i N} \\ \displaystyle + \sum_{i=1}^{n_{j+}} \sum_{i' \neq i=1}^{n_{j+}} \frac{\mathbf{e}_i^t}{\pi_i N} \mathrm{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'k'}^t \mid m, s) \frac{\mathbf{e}_{i'}}{\pi_{i'} N} \end{array} \right)$$

$$= \sum_{j=1}^{J} -\frac{n_{j+}}{(n_{j+}-1)} \sum_{i=1}^{n_{j+}} \left(\frac{e_{ik}}{\pi_i N} - \frac{1}{n_{j+}} \sum_{i=1}^{n_{j+}} \frac{e_{ik}}{\pi_i N} \right) \left(\frac{e_{ik'}}{\pi_i N} - \frac{1}{n_{j+}} \sum_{i=1}^{n_{j+}} \frac{e_{ik'}}{\pi_i N} \right). \qquad (50)$$

Les résultats (49) et (50) peuvent s'écrire en notation matricielle;

$$\mathrm{Cov}_e(\hat{\bar{\mathbf{E}}}_{\mathrm{HT}} \mid m, s)$$

$$= \mathbf{D} - \sum_{j=1}^{J} \frac{n_{j+}}{n_{j+}-1} \sum_{i=1}^{n_{j+}} \left(\frac{\mathbf{y}_{ik} - \mathbf{B}_k^t \mathbf{x}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{y}_{i'} - \mathbf{B}\mathbf{x}_i}{N \pi_{i'}} \right) \left(\frac{\mathbf{y}_{ik} - \mathbf{B}_k^t \mathbf{x}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{y}_{i'} - \mathbf{B}\mathbf{x}_{i'}}{N \pi_{i'}} \right)^t$$

où $\mathbf{D}$ représente une matrice diagonale de dimensions $K \times K$ avec éléments

$$d_k = \sum_{j=1}^{J} \frac{n_{j+}}{n_{j+}-1} \frac{n_{j+}}{n_{jk}} \sum_{i=1}^{n_{j+}} \left(\frac{y_{ik} - \mathbf{b}_k^t \mathbf{x}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{y_{i'k} - \mathbf{b}_k^t \mathbf{x}_{i'}}{N \pi_{i'}} \right)^2.$$

Il découle de (23) que

$$\mathrm{Cov}_e(\mathbf{C}\hat{\bar{\mathbf{E}}}_{\mathrm{HT}} \mid m, s)$$

$$= \mathbf{C}\mathbf{D}\mathbf{C}^t - \sum_{j=1}^{J} \frac{n_{j+}}{n_{j+}-1} \sum_{i=1}^{n_{j+}} \left(\frac{\mathbf{C}\boldsymbol{\varepsilon}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{C}\boldsymbol{\varepsilon}_{i'}}{N \pi_{i'}} \right) \left(\frac{\mathbf{C}\boldsymbol{\varepsilon}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{C}\boldsymbol{\varepsilon}_{i'}}{N \pi_{i'}} \right)^t. \qquad (51)$$

La partie finale de la preuve consiste à prendre l'espérance de $\mathrm{Cov}_e(\mathbf{C}\hat{\bar{\mathbf{E}}}_{\mathrm{HT}} \mid m, s)$ par rapport au plan d'échantillonnage et au modèle de l'erreur de mesure. La preuve est donnée pour les PBR où les UPE sont les variables de bloc. Selon un plan d'échantillonnage à deux degrés, nous tirons J blocs ou UPE d'une population finie de J_u blocs avec probabilités d'inclusion de premier ordre π_j^I. Dans chaque UPE, nous tirons n_{j+} USE au deuxième degré avec probabilités d'inclusion de premier et de deuxième ordres $\pi_{i|j}^{II}$ et $\pi_{ii'|j}^{II}$. Les probabilités d'inclusion de premier ordre des individus dans l'échantillon sont $\pi_i = \pi_j^I \pi_{i|j}^{II}$. En outre, soit

$$\bar{\Delta}_j = \sum_{i=1}^{N_j} \frac{\boldsymbol{\varepsilon}_i}{N_j}$$

la moyenne de population des erreurs de mesure des individus du bloc j. Alors

$$\hat{\bar{\Delta}}_j = \sum_{i=1}^{n_{j+}} \frac{\boldsymbol{\varepsilon}_i}{N_j \pi_{i|j}^{II}}$$

est l'estimateur d'Horvitz-Thompson pour $\bar{\Delta}_j$. Maintenant, nous avons

$$\sum_{j=1}^{J}\sum_{i=1}^{n_{j+}}\left(\frac{\boldsymbol{\varepsilon}_i}{N\pi_i}-\frac{1}{n_{j+}}\sum_{i'=1}^{n_{j+}}\frac{\boldsymbol{\varepsilon}_{i'}}{N\pi_{i'}}\right)\left(\frac{\boldsymbol{\varepsilon}_i}{N\pi_i}-\frac{1}{n_{j+}}\sum_{i'=1}^{n_{j+}}\frac{\boldsymbol{\varepsilon}_{i'}}{N\pi_{i'}}\right)^t$$

$$=\sum_{j=1}^{J}\left(\frac{1}{\pi_j^I}\right)^2\left(\frac{N_j}{N}\right)^2$$

$$\left(\begin{array}{l}\dfrac{1}{n_{j+}^2}\displaystyle\sum_{i=1}^{n_{j+}}\left(\frac{n_{j+}\boldsymbol{\varepsilon}_i}{N_j\pi_{i|j}^{II}}-\overline{\Delta}_j\right)\left(\frac{n_{j+}\boldsymbol{\varepsilon}_i}{N_j\pi_{i|j}^{II}}-\overline{\Delta}_j\right)^t\\ -\dfrac{1}{n_{j+}}\left(\hat{\overline{\Delta}}_j-\overline{\Delta}_j\right)\left(\hat{\overline{\Delta}}_j-\overline{\Delta}_j\right)^t\end{array}\right). \qquad (52)$$

Soit E_{s_I} l'espérance par rapport au premier degré du plan d'échantillonnage et $\mathrm{E}_{s_{II}}$ l'espérance par rapport au deuxième degré du plan d'échantillonnage. En prenant l'espérance par rapport au modèle de l'erreur de mesure et le plan d'échantillonnage de la première partie de (52) et en utilisant l'hypothèse de modélisation (3), nous arrivons à

$$\mathrm{E}_m\mathrm{E}_{s_I}\mathrm{E}_{s_{II}}\sum_{j=1}^{J}\left(\frac{1}{\pi_j^I}\right)^2\frac{1}{n_{j+}^2}\sum_{i=1}^{n_{j+}}\left(\frac{n_{j+}\boldsymbol{\varepsilon}_i}{N_j\pi_{i|j}^{II}}-\overline{\Delta}_j\right)\left(\frac{n_{j+}\boldsymbol{\varepsilon}_i}{N_j\pi_{i|j}^{II}}-\overline{\Delta}_j\right)^t$$

$$=\mathrm{E}_m\mathrm{E}_{s_I}\sum_{j=1}^{J}\left(\frac{1}{\pi_j^I}\right)^2\frac{1}{n_{j+}}\left(\sum_{i=1}^{N_j}\frac{n_{j+}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^t}{N_j^2\pi_{i|j}^{II}}-\overline{\Delta}_j\overline{\Delta}_j^t\right)$$

$$=\frac{1}{\pi_j^I n_{j+}N_j^2}\sum_{i=1}^{N_j}\left(\frac{n_{j+}}{\pi_{i|j}^{II}}-1\right)\boldsymbol{\Sigma}_i. \qquad (53)$$

Notons que $\mathrm{E}_{s_{II}}(\hat{\overline{\Delta}}_j-\overline{\Delta}_j)(\hat{\overline{\Delta}}_j-\overline{\Delta}_j)^t$ dans (52) est égale à la variance de plan de sondage de $\hat{\overline{\Delta}}_j$ par rapport au deuxième degré du plan d'échantillonnage dans le bloc *j*. En prenant l'espérance par rapport au modèle de l'erreur de mesure et au plan d'échantillonnage de la deuxième partie de (52) et en utilisant l'hypothèse de modélisation (3), nous arrivons à

$$\mathrm{E}_m\mathrm{E}_{s_I}\mathrm{E}_{s_{II}}\sum_{j=1}^{J}\left(\frac{1}{\pi_j^I}\right)^2\frac{1}{n_{j+}}\left(\hat{\overline{\Delta}}_j-\overline{\Delta}_j\right)\left(\hat{\overline{\Delta}}_j-\overline{\Delta}_j\right)^t$$

$$=\mathrm{E}_m\frac{1}{\pi_j^I N_j^2}\sum_{i=1}^{N_j}\sum_{i'=1}^{N_j}(\pi_{ii'|j}^{II}-\pi_{i|j}^{II}\pi_{i'|j}^{II})\frac{\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_{i'}^t}{\pi_{i|j}^{II}\pi_{i'|j}^{II}}$$

$$=\frac{1}{\pi_j^I N_j^2}\sum_{i=1}^{N_j}\left(\frac{1}{\pi_{i|j}^{II}}-1\right)\boldsymbol{\Sigma}_i. \qquad (54)$$

Au moyen des résultats (52), (53) et (54), nous pouvons élaborer le deuxième terme du deuxième membre de (51) sous la forme

$$\mathrm{E}_m\mathrm{E}_s\sum_{j=1}^{J}\frac{n_{j+}}{n_{j+}-1}\sum_{i=1}^{n_{j+}}\left(\frac{\mathbf{C}\boldsymbol{\varepsilon}_i}{N\pi_i}-\frac{1}{n_{j+}}\sum_{i'=1}^{n_{j+}}\frac{\mathbf{C}\boldsymbol{\varepsilon}_{i'}}{N\pi_{i'}}\right)$$

$$\left(\frac{\mathbf{C}\boldsymbol{\varepsilon}_i}{N\pi_i}-\frac{1}{n_{j+}}\sum_{i'=1}^{n_{j+}}\frac{\mathbf{C}\boldsymbol{\varepsilon}_{i'}}{N\pi_{i'}}\right)^t=\frac{1}{N^2}\sum_{i=1}^{N}\frac{\mathbf{C}\boldsymbol{\Sigma}_i\mathbf{C}^t}{\pi_i}. \qquad (55)$$

Enfin, il découle de (51) et (55) que

$$\mathrm{E}_m\mathrm{E}_s\mathrm{Cov}_e(\mathbf{C}\hat{\overline{\mathbf{E}}}_{\mathrm{HT}}\mid m,s)=$$

$$\mathrm{E}_m\mathrm{E}_s\mathbf{C}\boldsymbol{\Delta}\mathbf{C}^t-\frac{1}{N^2}\sum_{i=1}^{N}\frac{\mathbf{C}\boldsymbol{\Sigma}_i\mathbf{C}^t}{\pi_i},\quad \textbf{C.Q.F.D.}$$

Le calcul pour PBR où les strates sont les variables de bloc s'ensuit directement en tant que cas particulier d'un PBR où les UPE sont les variables de bloc avec $\pi_j^I=1, \pi_{i|j}^{II}=\pi_i,\ \pi_{ii'|j}^{II}=\pi_{ii'}$ et $J=J_u$. La preuve pour un PBR où les grappes sont des variables de bloc s'ensuit directement en tant que cas particulier d'un PBR où les UPE sont les variables de bloc avec $\pi_{i|j}^{II}=1$ et $\pi_{ii'|j}^{II}=1$.

L'espérance de $\mathrm{Cov}_e(\mathbf{C}\hat{\overline{\mathbf{E}}}_{\mathrm{HT}}\mid m,s)$ par rapport au plan d'échantillonnage et au modèle de l'erreur de mesure pour un PBR où les intervieweurs sont les variables de bloc ne découle pas à titre de cas particulier d'un PBR où les UPE sont des variables de bloc. Puisque les variables de bloc ne sont pas liées directement au plan d'échantillonnage, les blocs devraient être considérés comme des domaines où la taille de bloc n_{j+} est aléatoire par rapport au plan d'échantillonnage. Le calcul suit les mêmes étapes que celles de la preuve pour la constitution de blocs sur les UPE et est donné par Van den Brakel (2001).

Preuve de la formule (32)

La matrice $\hat{\mathbf{D}}$ peut être partitionnée comme suit :

$$\hat{\mathbf{D}}=\begin{pmatrix}\hat{d}_1 & \mathbf{0}^t\\ \mathbf{0} & \hat{\mathbf{D}}_*\end{pmatrix}.$$

Il découle de l'identité de Bartlett (Morisson 1990, chapitre 2) que :

$$(\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t)^{-1}=(\hat{d}_1\mathbf{j}\mathbf{j}^t+\hat{\mathbf{D}}_*)^{-1}=\hat{\mathbf{D}}_*^{-1}-\frac{1}{\mathrm{trace}(\hat{\mathbf{D}}^{-1})}\hat{\mathbf{D}}_*^{-1}\mathbf{j}\mathbf{j}^t\hat{\mathbf{D}}_*^{-1}.$$

De ce résultat il découle que

$$\mathbf{C}^t(\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t)^{-1}\mathbf{C}=\mathbf{C}^t\hat{\mathbf{D}}_*^{-1}\mathbf{C}-\frac{1}{\mathrm{trace}(\hat{\mathbf{D}}^{-1})}\mathbf{C}^t\hat{\mathbf{D}}_*^{-1}\mathbf{j}\mathbf{j}^t\hat{\mathbf{D}}_*^{-1}\mathbf{C}$$

$$=\hat{\mathbf{D}}^{-1}-\frac{1}{\mathrm{trace}(\hat{\mathbf{D}}^{-1})}\hat{\mathbf{D}}^{-1}\mathbf{j}\mathbf{j}^t\hat{\mathbf{D}}^{-1}. \qquad (56)$$

L'insertion de (56) dans (31) mène à (32), **C.Q.F.D.**

Remerciements

Les auteurs remercient le rédacteur associé, les examinateurs, Paul Smith et Rachel Vis-Visschers de leurs commentaires constructifs au sujet d'ébauches antérieures du présent article. Jan remercie aussi les professeurs Stephen E. Fienberg et Peter Kooiman de leur soutien à titre de conseillers de thèse de doctorat durant les présents travaux.

Bibliographie

Bethlehem, J.G., et Keller, W.G. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3(2), 141-153.

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

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

Fienberg, S.E., et Tanur, J.M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *Revue Internationale de Statistique*, 55(1), 75-96.

Fienberg, S.E., et Tanur, J.M. (1988). From the inside out and the outside in: Combining experimental and sampling structures. *The Canadian Journal of Statistics*, 16(2), 135-151.

Fienberg, S.E., et Tanur, J.M. (1989). Combining cognitive and statistical approaches to survey design. *Science*, 243, 1017-1022.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Mathematical Institute of the Hungarian Academy of Sciences*, 5, 361-374.

Hájek, J. (1971). Comment on a paper by D. Basu. Dans *Foundations of Statistical Inference*, (Éds. V.P. Godambe et D.A. Sprott). Toronto: Holt, Rinehart et Winston. 236.

Hartley, H.O., et Rao, J.N.K. (1978). Estimation of nonsampling variance components in sample surveys. Dans *Survey Sampling and Measurement*, (Éds. N.K. Namboodiri). New York: Academic Press. 35-43.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. New York: McGraw-Hill.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian statistical institute. *Journal of the Royal Statistical Society*, 109, 325-370.

Montgomery, D.C. (2001). *Design and Analysis of Experiments*. New York: John Wiley & Sons, Inc.

Morisson, D.F. (1990). *Multivariate Statistical Methods*. Singapore: McGraw-Hill.

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

Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons, Inc.

Searle, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys*, (Éds. C.J. Skinner, D. Holt et T.M.F. Smith). Chichester: Wiley & Sons, Inc. 59-87.

Statacorp. (2001). Stata Reference Manual Release 7.0. College Station, Texas.

Van den Brakel, J.A. (2001). Design and Analysis of Experiments Embedded in Complex Sample Surveys. Thèse de doctorat. Rotterdam: Erasmus University of Rotterdam.

Van den Brakel, J.A. et Binder, D. (2000). Variance estimation for experiments embedded in complex sampling schemes. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Indianapolis, August 13-17. 805-810.

Van den Brakel, J.A., et Binder, D. (2004). Variance estimation for experiments embedded in complex sampling designs. Article de recherche non publiée, BPA nr.: H894-04-TMO. Heerlen: Statistics Netherlands.

Van den Brakel, J.A., et Renssen, R.H. (1998). Design and analysis of experiments embedded in sample surveys. *Journal of Official Statistics*, 14(3), 277-295.

Van den Brakel, J.A., et Van Berkel, C.A.M. (2002). A design-based analysis procedure for two-treatment experiments embedded in sample surveys. An application in the Dutch labor force survey. *Journal of Official Statistics*, 18(2), 217-231.

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 45-55
Statistique Canada, N° 12-001-XIF au catalogue

Estimateurs de domaine pour la technique du dénombrement d'items

Takahiro Tsuchiya [1]

Résumé

La technique du dénombrement d'items (*item count technique*) est une méthode de questionnement indirect qui a été conçue pour estimer la proportion de personnes pour lesquelles un item important de nature délicate est vrai. Elle consiste à demander aux répondants d'indiquer, parmi une liste de phrases descriptives, le nombre d'entre elles qu'ils estiment s'appliquer à eux. Une liste comprenant l'item clé est présentée à une moitié de l'échantillon et une liste ne le contenant pas est présentée à l'autre moitié. La différence entre les nombres moyens de phrases sélectionnées est un estimateur de la proportion recherchée. Dans le présent article, nous proposons deux nouvelles méthodes, appelées méthode par croisement et méthode par double croisement, où les proportions dans les sous-groupes ou domaines sont estimées d'après les données obtenues par la technique du dénombrement d'items. Afin d'évaluer la précision des méthodes proposées, nous réalisons des expériences par simulation au moyen de données tirées d'une enquête sur le caractère national japonais. Les résultats montrent que la méthode par double croisement est beaucoup plus précise que la méthode stratifiée traditionnelle et moins susceptible de produire des estimations illogiques.

Mots clés : Techniques de questionnement indirect; technique du dénombrement d'items; estimateurs de domaine; enquête sur le caractère national japonais.

1. Introduction

1.1 Techniques de questionnement indirect

Supposons qu'une population U soit divisée en deux sous-populations $U_{(T)}$ et $U^c_{(T)}$, où $U_{(T)}$ est un ensemble d'éléments ayant un attribut T, et $U^c_{(T)}$ est un complément de $U_{(T)}$. L'un des objectifs des enquêtes sociales est d'estimer $\pi = \overline{Y} = P(Y=1)$, où

$$Y_k = \begin{cases} 1 & \text{si } k \in U_{(T)} \\ 0 & \text{autrement} \end{cases}$$

et $P(\cdot)$ représente la proportion d'unités ayant une valeur particulière de la variable. Par exemple, quand T représente « appuie le cabinet actuel », π indique le taux de soutien pour le cabinet et quand T signifie « consomme une drogue illicite particulière », π représente le taux de prévalence de la consommation de cette drogue.

Dans le cas d'une technique de questionnement direct, les chercheurs demandent aux répondants « Appartenez-vous à $U_{(T)}$? » et obtiennent directement la valeur indicatrice y_i sous forme d'une réponse « Oui » ou « Non » (Cochran 1977, page 50). Si la probabilité de sélection est la même pour tous les répondants, la moyenne d'échantillon $\bar{y}$ sert d'estimateur de π.

Par ailleurs, certaines techniques de questionnement indirect, y compris la technique de la réponse aléatoire (Warner 1965), la technique nominative (Miller 1985), la technique du dénombrement d'items (Droitcour, Caspar, Hubbard, Parsley, Visscher et Ezzati 1991) et la technique des trois cartes (Droitcour, Larson et Scheuren 2001), sont conçues pour contourner le fait que certains répondants essayent d'éviter les questions délicates, comme celles portant sur des sujets très intimes, des comportements socialement inacceptables ou pervers, ou des actes illégaux. La caractéristique essentielle des techniques indirectes est qu'au lieu d'observer directement Y, on observe une autre variable $X = g(Y, \boldsymbol{V})$, qui est une certaine fonction de Y et, au besoin, d'autres variables aléatoires $\boldsymbol{V}$, de sorte que les répondants aient l'impression que leur réponse réelle pour Y n'est pas révélée. Bien que cette caractéristique permette, en principe, de dériver une réponse correcte pour des répondants évasifs, les procédures tant de questionnement que d'estimation sont assez compliquées comparativement à la technique de questionnement direct, en partie parce que la fonction $g(\cdot)$ comprend parfois des processus de randomisation. Nous décrirons deux techniques indirectes dans les grandes lignes plus loin.

La technique de la réponse aléatoire est la plus populaire des techniques indirectes et diverses versions ont été proposées (Abul-Ela, Greenberg et Horvitz 1967; Warner 1971; Chaudhuri et Mukerjee 1988; Greenberg, Abul-Ela, Simmons et Horvitz 1969; Takahasi et Sakasegawa 1977). Bien que cette technique ne soit pas le sujet du présent article, nous décrivons brièvement la procédure originale de Warner à titre de référence, car nous en décrirons la simulation à la dernière section.

1. Préparer deux types de questionnaires. Le questionnaire A comprend la question « Appartenez-vous à

1. Takahiro Tsuchiya, The Institute of Statistical Mathematics, 4-6-7, Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japon. Courriel : taka@isma.ac.jp.

$U_{(T)}$? » et le questionnaire *B*, la question « Appartenez-vous à $U^{c}_{(T)}$? ».

2. Soit $p\,(\neq 0{,}5)$ la probabilité prédéterminée. Chaque répondant choisit le questionnaire *A* ou *B* avec la probabilité p ou $1-p$ respectivement, mais personne d'autre que le répondant ne sait quel questionnaire il a sélectionné.

3. Supposer que *X* est une variable indicateur dont la valeur est 1 si la réponse est « oui » ou 0 si la réponse est « non ». L'estimateur de π est donné par

$$\hat{\pi} = \frac{p-1+\bar{x}}{2p-1}, \qquad (1)$$

où $\bar{x}$ est une moyenne d'échantillon de *X*.

Puisque les chercheurs n'ont aucune information quant au type de questionnaire choisi par chaque répondant, un plus grand nombre de répondants devraient, en principe, répondre honnêtement que si on leur posait directement les questions.

La technique du dénombrement d'items, qui est le sujet du présent article, n'est pas aussi répandue malgré sa simplicité. Cette technique s'avère, elle aussi, efficace lorsqu'on doit poser des questions délicates, parce qu'on demande aux répondants non pas de répondre directement à ces questions, mais plutôt de déclarer le nombre d'items qui, parmi une liste, sont vrais dans leur cas. Les procédures de la technique du dénombrement d'items sont les suivantes :

1. Préparer l'item clé *T*, qui est le point de concentration de l'étude, et *G* autres items non clés $E_1, \ldots, E_G$. Par exemple, *T* pourrait être « consommer une certaine drogue illicite », comme mentionné plus haut, et E_g, une description non délicate quelconque, telle que « posséder une bicyclette ».

2. Préparer deux types de questionnaires, *A* et *B*. Dans le questionnaire *A*, on demande aux répondants d'indiquer le nombre C^A d'items qui sont vrais en ce qui les concerne parmi les *G* items non clés. Dans le questionnaire *B*, on demande aux répondants d'indiquer le nombre C^B d'items qui sont vrais en ce qui les concerne parmi les $G+1$ items, y compris l'item clé *T*.

   Le tableau 1 donne des exemples de listes d'items. Notre but est d'estimer la proportion de personnes qui utilisent une drogue illicite particulière. L'item clé est « avoir consommé une drogue illicite particulière » dans le questionnaire *B* et les quatre autres items ne sont pas des items clés. Sauf si une réponse au questionnaire *B* est $C^B = 0$ ou $C^B = 5$, les chercheurs ne peuvent déterminer quels items sont vrais pour le répondant. Ainsi, un répondant pourrait indiquer que quatre items du questionnaire *B* sont vrais, mais nous ne pouvons être certains qu'il consomme de la drogue. Par conséquent, on devrait s'attendre à ce qu'un plus grand nombre de répondants consommant une drogue illicite donnent une réponse honnête dans une telle situation que si on leur posait une question directe.

3. Diviser l'échantillon en deux sous-groupes, *A* et *B*, de taille n^A et n^B au hasard, de sorte que chaque questionnaire soit attribué à un sous-groupe particulier.

Tableau 1
Exemples de listes d'items

Questionnaire *A*	Questionnaire *B*
Combien parmi les items suivants sont vrais pour vous? – posséder une bicyclette – avoir voyagé à l'étranger – avoir appelé une ambulance – posséder une villa d'été	Combien parmi les items suivants sont vrais pour vous? – posséder une bicyclette – avoir voyagé à l'étranger – avoir appelé une ambulance – avoir consommé une drogue illicite particulière – posséder une villa d'été

4. L'estimateur de π est donné par

$$\hat{\pi} = \hat{\bar{C}}^B - \hat{\bar{C}}^A, \qquad (2)$$

où $\hat{\bar{C}}^A$ et $\hat{\bar{C}}^B$ sont les moyennes estimées de C^A et C^B, respectivement. La justification de (2) est donnée à la section 2.1. Si la probabilité de sélection est la même pour toutes les unités de l'échantillon, $\hat{\pi}$ peut s'écrire

$$\hat{\pi} = \sum_{c=0}^{G+1} c\,\frac{n_c^B}{n^B} - \sum_{c=0}^{G} c\,\frac{n_c^A}{n^A}, \qquad (3)$$

où n_c^A et n_c^B sont les nombres de répondants dont les réponses sont $C^A = c$ et $C^B = c$, respectivement. De surcroît, si l'on dispose d'une variable auxiliaire *Z* et que l'on connaît sa distribution $P(Z=z) = m_z$ dans la population, par exemple d'après un recensement, on recourt souvent à la poststratification pour corriger la distribution d'échantillon de *Z* d'après la distribution de population. Autrement dit, l'estimateur poststratifié de π est donné par

$$\begin{aligned}\hat{\pi}_{\mathrm{PS}} &= \sum_{c=0}^{G+1} c\,\frac{\sum_z v_z^B n_{cz}^B}{n^B} - \sum_{c=0}^{G} c\,\frac{\sum_z v_z^A n_{cz}^A}{n^A} \\ &= \sum_{c=0}^{G+1} c \sum_z \frac{m_z}{n_{\cdot z}^B} n_{cz}^B - \sum_{c=0}^{G} c \sum_z \frac{m_z}{n_{\cdot z}^A} n_{cz}^A, \end{aligned} \qquad (4)$$

où n_{cz}^A est le nombre de répondants pour chaque $C^A = c$ et $Z = z$,

$$n_{\cdot z}^A = \sum_{c=0}^{G} n_{cz}^A,\ n^A = \sum_z n_{\cdot z}^A,\ v_z^A = \frac{m_z n^A}{n_{\cdot z}^A}$$

et $n_{cz}^B, n_{\cdot z}^B, n^B$ et v_z^B sont définis de façon analogue.

L'un des avantages de la technique du dénombrement d'items est qu'elle ne nécessite aucun des mécanismes de randomisation utilisé dans la technique de la réponse aléatoire. Ce n'est pas le répondant, mais un chercheur, qui choisit le questionnaire auquel il faut répondre. Donc, la technique du dénombrement d'items est facile à appliquer au moyen d'une enquête avec questionnaire à remplir soi-même ou d'une enquête téléphonique. Une comparaison plus approfondie de la technique de la réponse aléatoire et de la technique du dénombrement d'items figure dans Hubbard, Casper et Lessler (1989).

Le questionnaire A est introduit pour obtenir la distribution du nombre d'items non clés. Autrement dit, les personnes qui répondent au questionnaire A ne répondent pas à la question délicate. Par conséquent, il est possible d'accroître la précision de l'estimateur en utilisant la version à liste double de la technique du dénombrement d'items (Droitcour et coll., 1991), en vertu de laquelle il y a échange de rôle entre les deux sous-groupes. Cependant, ici, nous limitons notre argument à une version à liste unique, parce que l'extension des estimateurs à la version à liste double est simple.

1.2 But du présent article

Jusqu'à présent, nous nous sommes concentrés sur le paramètre $\pi = \bar{Y} = P(Y = 1)$ de l'ensemble d'une population. Cependant, il est souvent nécessaire d'obtenir des estimateurs pour des sous-populations ou domaines (Särndal, Swesson et Wretman 1992, page 5), c'est-à-dire d'estimer une proportion conditionnelle $P(Y = 1 | Z = z)$ ou une proportion conjointe $P(Y = 1, Z = z)$, où la population est subdivisée en plusieurs domaines par la valeur Z. Ici, nous donnons à la variable Z le nom de variable de domaine. Les variables de domaine souvent utilisées sont les caractéristiques démographiques, comme le sexe ou l'âge. Par exemple, certains organismes gouvernementaux voudraient connaître, pour chaque groupe d'âge, la proportion de personnes qui consomment une drogue illicite particulière. Même si, dans l'équation (4), l'estimateur poststratifié $\hat{\pi}_{PS}$ utilise la variable de domaine Z, le but est d'estimer $P(Y = 1)$ pour l'ensemble de la population. Notre objectif, dans le présent article, est d'obtenir des estimations distinctes de $P(Y = 1 | Z = z)$ pour chaque domaine.

Voici une méthode simple d'estimation :

1. Poststratifier l'échantillon en strates ou domaines en se basant sur la valeur Z.
2. Dans chaque strate ou domaine, déterminer séparément $p(Y = 1 | Z = z)$ en se servant de (1) ou (2), où $p(\cdot)$ est une estimation échantillonnale de $P(\cdot)$.
3. Au besoin, estimer $p(Y = 1, Z = z)$ en multipliant une proportion de domaine connue, $P(Z = z)$, ou une proportion de domaine estimée, $p(Z = z)$.

Dans tout l'article, nous appelons la méthode susmentionnée méthode stratifiée, parce que les estimations sont obtenues séparément pour chaque strate ou domaine. Bien que Rao (2003) donne à cette méthode le nom d'estimation directe, nous avons choisi de ne pas utiliser le terme « directe » afin d'éviter toute confusion avec l'expression « technique de questionnement direct ».

L'un des avantages de la méthode stratifiée est qu'elle s'applique à toute technique de questionnement indirect, y compris celles de la réponse aléatoire et du dénombrement d'items. Aux États-Unis, le General Accounting Office (1999) a adopté la méthode stratifiée pour produire des estimations de domaine par la technique des trois cartes. Cependant, l'un des inconvénients sérieux de la méthode stratifiée est qu'elle produit souvent des estimations illogiques, surtout des estimations négatives, dans le cas de la réponse aléatoire et du dénombrement d'items, comme nous l'expliquons plus loin. Ce problème tient principalement au fait que la réduction de la taille d'échantillon dans chaque strate accroît l'erreur-type des estimateurs (Lessler et O'Reilly 1997). Ainsi, Droitcour et coll. (1991, page 206) ont calculé des estimations distinctes pour les trois strates de risque et obtenu des estimations négatives du taux de prévalence de la consommation de drogue.

Dans le cas de la technique de la réponse aléatoire, les possibilités de produire des estimations de domaine autrement que par la méthode stratifiée sont peu nombreuses, parce que l'information sur le type de questionnaire choisi par les répondants n'est pas disponible. Par contre, dans le cas du dénombrement d'items, on sait à quel questionnaire a répondu chaque personne. Par conséquent, la précision des estimateurs de domaine devrait, en principe, augmenter si l'on utilise des données auxiliaires, plus précisément des tableaux de contingence entre Z et C^A ou C^B.

Dans le présent article, nous proposons pour la technique du dénombrement d'items, de nouveaux estimateurs de domaine que nous appelons méthode par croisement et méthode par double croisement, respectivement. En outre, nous montrons que les nouveaux estimateurs sont plus efficaces que la méthode stratifiée classique en simulant la technique du dénombrement d'items au moyen de données tirées de l'enquête sur le caractère national japonais visant à déterminer les attributs significatifs du caractère japonais.

2. Estimateurs de domaine pour la technique du dénombrement d'items

2.1 Méthode stratifiée

Ici, nous reformulons la méthode stratifiée. Supposons que les équations qui suivent soient vérifiées pour chaque valeur de *c* et de *z*.

Hypothèse 1.

$$P(C^B = c|Z = z) = P(C^A = c, Y = 0|Z = z) + P(C^A = c-1, Y = 1|Z = z),$$
$$P(C^A = G+1, Y = 0|Z = z) = 0.$$

Ces hypothèses sous-entendent que la différence entre les distributions de C^A et C^B dépend uniquement de *Y*. Les effets de question, dont les effets d'ordre et les effets de contexte (Schuman et Presser 1981), ne sont pas pris en considération.

En nous fondant sur ces hypothèses, nous obtenons le résultat suivant.

Méthode stratifiée.

$$P(Y=1|Z=z) = \sum_{c=0}^{G+1} c\,P(C^B = c|Z=z) - \sum_{c=0}^{G} c\,P(C^A = c|Z=z) \quad (5)$$
$$= \overline{C}_z^B - \overline{C}_z^A, \quad (6)$$

où $\overline{C}_z^A$ et $\overline{C}_z^B$ sont les moyennes de domaine de C^A et C^B.

Calculs.

$$\sum_{c=0}^{G+1} cP(C^B = c|Z=z)$$
$$= \sum_{c=0}^{G+1} cP(C^A = c, Y=0|Z=z) + \sum_{c=0}^{G+1} cP(C^A = c-1, Y=1|Z=z)$$
$$= \sum_{c=0}^{G} cP(C^A = c, Y=0|Z=z) + \sum_{c=0}^{G} (c+1)P(C^A = c, Y=1|Z=z)$$
$$= \sum_{c=0}^{G} c\{P(C^A = c, Y=0|Z=z) + P(C^A = c, Y=1|Z=z)\} + \sum_{c=0}^{G} P(C^A = c, Y=1|Z=z)$$
$$= \sum_{c=0}^{G} cP(C^A = c|Z=z) + P(Y=1|Z=z).$$

Le transfert du premier terme dans le premier membre de l'équation donne la méthode stratifiée (5).

L'estimateur $p(Y=1|Z=z)$ s'obtient en remplaçant les moyennes de domaine $\overline{C}_z^A$ et $\overline{C}_z^B$ par leurs estimateurs, $\hat{\overline{C}}_z^A$ et $\hat{\overline{C}}_z^B$.

$$p(Y=1|Z=z) = \hat{\overline{C}}_z^B - \hat{\overline{C}}_z^A. \quad (7)$$

Quand les probabilités de sélection sont égales pour toutes les unités de l'échantillon, l'estimateur $P(Y=1|Z=z)$ s'écrit

$$p(Y=1|Z=z) = \sum_{c=0}^{G+1} c\,\frac{n_{cz}^B}{n_{\cdot z}^B} - \sum_{c=0}^{G} c\,\frac{n_{cz}^A}{n_{\cdot z}^A}, \quad (8)$$

où $n_{cz}^A, n_{cz}^B, n_{\cdot z}^A$ et $n_{\cdot z}^B$ sont définis à la section 1.1. Les équations (2) et (3) pour l'ensemble de la population sont des cas particuliers de (7) et (8).

L'un des avantages de la méthode stratifiée est que l'estimateur de la variance de $p(Y=1|Z=z)$ s'obtient facilement par

$$\hat{\mathrm{V}}\mathrm{ar}\,(p(Y=1|Z=z)) = \hat{\mathrm{V}}\mathrm{ar}\,(\hat{\overline{C}}_z^B) + \hat{\mathrm{V}}\mathrm{ar}\,(\hat{\overline{C}}_z^A). \quad (9)$$

Par ailleurs, comme nous l'avons souligné à la section précédente, la réduction de la taille d'échantillon dans chaque strate augmente les variances estimées dans (9). De surcroît, l'estimateur marginal $p(Y=1)$ obtenu en utilisant (8) ne correspond pas à celui obtenu directement au moyen de (3), à moins que $n_{\cdot z}^A = n_{\cdot z}^B$ pour tout *z*. Autrement dit, si $p(Z=z)$ n'est pas connue, son estimateur est donné par

$$p(Z=z) = (n_{\cdot z}^A + n_{\cdot z}^B)/(n^A + n^B)$$

et

$$\sum_z p(Y=1|Z=z)p(Z=z) = \sum_z \frac{n_{\cdot z}^A + n_{\cdot z}^B}{n^A + n^B}\left\{\sum_{c=0}^{G+1} c\,\frac{n_{cz}^B}{n_{\cdot z}^B} - \sum_{c=0}^{G} c\,\frac{n_{cz}^A}{n_{\cdot z}^A}\right\}$$
$$\neq \sum_{c=0}^{G+1} c\,\frac{n_{c\cdot}^B}{n^B} - \sum_{c=0}^{G} c\,\frac{n_{c\cdot}^A}{n^A} = \hat{\pi}. \quad (10)$$

Si l'on connaît la proportion de domaine $p(Z=z) = m_z$, l'estimateur marginal correspond à l'estimateur post-stratifié (4).

$$\sum_z p(Y=1|Z=z)P(Z=z) = \sum_z m_z\left\{\sum_{c=0}^{G+1} c\,\frac{n_{cz}^B}{n_{\cdot z}^B} - \sum_{c=0}^{G} c\,\frac{n_{cz}^A}{n_{\cdot z}^A}\right\}$$
$$= \hat{\pi}_{\mathrm{PS}}.$$

Ces résultats indiquent que nous devrions utiliser un estimateur poststratifié $\hat{\pi}_{\mathrm{PS}}$ avec les estimateurs de domaine si nous utilisons la méthode stratifiée.

2.2 Méthode par croisement

Dans la méthode stratifiée, nous subdivisons un échantillon de l'ensemble de la population en strates afin de procéder à l'estimation directe de $P(Y=1|Z=z)$, ce qui cause une réduction de la taille d'échantillon. Par conséquent, dans la méthode par croisement proposée ici,

nous commençons par estimer la proportion conjointe $P(Y=1, Z=z)$ afin d'utiliser l'échantillon complet, puis nous obtenons la proportion conditionnelle par

$$p(Y=1|Z=z)=\frac{p(Y=1, Z=z)}{p(Z=z)}$$

$$\text{ou} \quad p(Y=1|Z=z)=\frac{p(Y=1, Z=z)}{P(Z=z)}$$

Nous utilisons la dénomination « méthode par croisement », parce que cette méthode s'appuie sur des totalisations croisées $P(Z=z|C^B=c)$, comme le montre (19).

Dans la méthode par croisement, nous supposons que les équations qui suivent tiennent pour chaque valeur de *c*.

Hypothèse 2.

$$P(C^B=c+1, Y=1)=P(C^A=c, Y=1), \qquad (11)$$

$$P(C^B=0, Y=1)=P(C^A=-1, Y=1)=0, \qquad (12)$$

$$P(C^B=c, Y=0)=P(C^A=c, Y=0). \qquad (13)$$

Ces hypothèses impliquent aussi que la différence entre les distributions de C^A et C^B dépend uniquement de *Y*.

En nous fondant sur ces hypothèses, nous obtenons le résultat suivant.

Méthode par croisement.

$$P(Y=1, Z=z)=\sum_{c=1}^{G+1} P(Z=z|C^B=c)Q_{c-1}, \qquad (14)$$

où

$$Q_c=\sum_{d=0}^{c}\{P(C^A=d)-P(C^B=d)\}.$$

De plus, on suppose que $P(Z=z|C^B=c,\ Y=1) = P(Z=z|C^B=c)$ pour tout $c > 0$. Cette hypothèse serait valide jusqu'à un certain point, quand les items clés et non clés décrivent tous les deux le même type de comportement stigmatisant.

Calculs.

D'après les hypothèses, nous avons

$$P(C^B=c)=P(C^B=c, Y=1)+P(C^B=c, Y=0)$$

$$=P(C^A=c-1, Y=1)+P(C^A=c, Y=0). \qquad (15)$$

L'équation qui suit est vérifiée pour toute valeur de *c*.

$$P(C^A=c, Y=0)=P(C^A=c)-P(C^A=c, Y=1). \qquad (16)$$

Donc, la substitution de (16) dans (15) donne

$$P(C^B=c)=P(C^A=c-1,\ Y=1)$$

$$+\{P(C^A=c)-P(C^A=c, Y=1)\}. \qquad (17)$$

Par sommation de (17) sur *c* jusqu'à une certaine valeur *g*, nous obtenons

$$\sum_{c=0}^{g} P(C^B=c)=\sum_{c=0}^{g} P(C^A=c-1,\ Y=1)$$

$$+\sum_{c=0}^{g}\{P(C^A=c)-P(C^A=c, Y=1)\}$$

$$=\sum_{c=0}^{g} P(C^A=c)-P(C^A=g, Y=1).$$

Par transposition des termes, nous définissons Q_c.

$$Q_c=\sum_{d=0}^{c}\{P(C^A=d)-P(C^B=d)\}$$

$$=P(C^A=c, Y=1)$$

$$=P(C^B=c+1, Y=1). \qquad (18)$$

Ici, la proportion conjointe $P(Y=1, Z=z)$ se décompose comme suit

$$P(Y=1, Z=z)=\sum_{c=0}^{G+1} P(Z=z|C^B=c)P(C^B=c, Y=1). \qquad (19)$$

La substitution de l'équation (18) et de l'hypothèse (12) dans l'équation (19) donne la méthode par croisement.

L'estimateur conjoint $P(Y=1, Z=z)$ s'obtient par remplacement de chaque terme de (14) par son estimateur. Si l'échantillon est autopondéré, l'estimateur s'écrit

$$P(Y=1, Z=z)=\sum_{c=1}^{G+1}\frac{n_{cz}^B}{n_{c\cdot}^B}\sum_{d=0}^{c-1}\left(\frac{n_{d\cdot}^A}{n^A}-\frac{n_{d\cdot}^B}{n^B}\right), \qquad (20)$$

où

$$n_{c\cdot}^A=\sum_z n_{cz}^A \quad \text{et} \quad n_{c\cdot}^B=\sum_z n_{cz}^B.$$

Nous obtenons l'estimateur conditionnel $p(Y=1|Z=z)$ en divisant $p(Y=1|Z=z)$ par les proportions de domaine $P(Z=z)$ ou leur estimateur $p(Z=z)$.

Comme nous l'avons mentionné plus haut, la caractéristique principale de la méthode par croisement est qu'on commence par estimer $p(Y=1, Z=z)$ pour l'échantillon complet. Par conséquent, la variance de $p(Y=1|Z=z)$ devrait être plus faible dans le cas de la méthode par croisement que dans celui de la méthode stratifiée. En outre, la méthode par croisement produit rarement des valeurs négatives, tandis que la méthode stratifiée en produit fréquemment. De surcroît, l'estimateur marginal $p(Y=1)$ obtenu par sommation de (20) est égal à l'estimateur (3), à moins que $n_{c\cdot}^B=0$ pour certaines valeurs de *c* :

$$\sum_z p(Y=1, Z=z) = \sum_z \sum_{c=1}^{G+1} \frac{n_{cz}^B}{n_{c\cdot}^B} \sum_{d=0}^{c-1} \left(\frac{n_{d\cdot}^A}{n^A} - \frac{n_{d\cdot}^B}{n^B} \right)$$
$$= \sum_{c=1}^{G+1} \sum_{d=0}^{c-1} \left(\frac{n_{d\cdot}^A}{n^A} - \frac{n_{d\cdot}^B}{n^B} \right)$$
$$= \sum_{c=1}^{G+1} \left\{ \left(1 - \sum_{d=c}^{G} \frac{n_{d\cdot}^A}{n^A} \right) - \left(1 - \sum_{d=c}^{G+1} \frac{n_{d\cdot}^B}{n^B} \right) \right\}$$
$$= \sum_{c=0}^{G+1} c \frac{n_{c\cdot}^B}{n^B} - \sum_{c=0}^{G} c \frac{n_{c\cdot}^A}{n^A} = \hat{\pi}. \quad (21)$$

Naturellement, si nous connaissons les proportions de domaine $P(Z=z) = m_z$, nous pouvons les utiliser pour obtenir un estimateur poststratifié $p(C^A = d)$ de $P(C^A = d)$ dans Q_{c-1} de (14),

$$p(C^A = d) = \sum_z \frac{m_z}{n_{\cdot z}^B} n_{dz}^B.$$

Dans ce cas, $\sum_z p(Y=1, Z=z)$ coïncide avec l'estimateur poststratifié $\hat{\pi}_{PS}$.

Un inconvénient de la méthode par croisement est que la variance de $p(Y=1|Z=z)$ est presque impossible à estimer algébriquement. Par conséquent, il faut utiliser une méthode par rééchantillonnage, telle que le jacknife ou le bootstrap. De plus, puisqu'il est impossible de déterminer laquelle, de la méthode stratifiée et de la méthode par croisement, est la plus efficace, nous décrivons plus loin la réalisation d'une étude en simulation.

2.3 Méthode par double croisement

Avant de passer à l'étude en simulation, nous proposons une version modifiée de la méthode par croisement. Dans l'équation (19) de la méthode par croisement, nous utilisons $P(Z=z|C^B=c)$. De la même façon, l'utilisation de $P(Z=z|C^A=c)$ donne

$$P(Y=1, Z=z) = \sum_{c=0}^{G} P(Z=z|C^A=c) P(C^A=c, Y=1)$$
$$= \sum_{c=0}^{G} P(Z=z|C^A=c) Q_c. \quad (22)$$

Par conséquent, nous obtenons une méthode par double croisement en combinant (14) et (22) comme suit :

$$P(Y=1, Z=z) = \sum_{c=0}^{G} \left\{ \begin{matrix} w^A P(Z=z|C^A=c) \\ + w^B P(Z=z|C^B=c+1) \end{matrix} \right\} Q_c, \quad (23)$$

où w^A et w^B sont les poids non négatifs de chaque sous-groupe, dont la somme est égale à 1.

L'équation qui suit est également vraie pour la méthode par double croisement de tout poids w^A et w^B, à moins que $n_{c\cdot}^A = 0$ ou $n_{c\cdot}^B = 0$ pour certaines valeurs de c.

$$\sum_z p(Y=1, Z=z) = \hat{\pi}. \quad (24)$$

3. Expériences numériques

3.1 Ensemble de données

Afin de comparer la précision des estimateurs, nous avons réalisé des expériences en simulation en nous servant de données tirées de l'enquête sur le caractère national japonais (Sakamoto, Tsuchiya, Nakamura, Maeda et Fouse, 2000). Bien que les répondants aient été sélectionnés par échantillonnage stratifié à deux degrés parmi la population du Japon de 20 ans et plus, nous n'avons pas tenu compte du plan d'échantillonnage, parce que nous avons traité l'échantillon recueilli de $N = 1\,339$ comme étant la population « réelle » dans cette expérience. Le tableau 2 donne les résultats pour une question au sujet des attributs significatifs du caractère japonais. Lors d'une interview sur place, on a demandé aux répondants de choisir, parmi une liste de dix adjectifs, tous ceux qui, selon eux, décrivaient le caractère japonais.

Tableau 2
Attributs significatifs du caractère japonais

					N = 1 339
(Montrer la carte) Selon vous, lesquels, parmi les adjectifs suivants décrivent le caractère du peuple japonais? Choisissez autant d'adjectifs que vous souhaitez.					
1	Rationnel	18 %	6	Gentil	42 %
2	Diligent	71 %	7	Original	7 %
3	Libre	13 %	8	Poli	50 %
4	Ouvert, franc	14 %	9	Joyeux	8 %
5	Persistent	51 %	10	Idéaliste	23 %

La forme de cette question diffère de celle de la technique du dénombrement d'items, qui consiste à demander aux répondants d'« indiquer le nombre d'adjectifs ». Dans l'enquête décrite ici, on demande aux répondants d'« encercler autant d'adjectifs qu'ils jugent appropriés ». En outre, les dix items ne sont pas de nature très délicate, si bien que les répondants ne devraient pas hésiter durant la sélection. Cependant, puisque nous obtenons les tableaux de contingence réels entre chacun des dix éléments et une autre variable Z, nous pouvons évaluer les propriétés des estimateurs grâce à une pseudo procédure de dénombrement d'items.

Nous avons choisi chacun des trois items suivants comme item clé Y, où $Y=1$ signifie que l'item a été sélectionné.

- 7 Original (π la plus faible parmi les dix items)
- 8 Poli (π exactement égale à 50 %)
- 2 Diligent (π la plus grande parmi les dix items)

Nous utilisons trois combinaisons d'items non clés, telles qu'énumérées au tableau 3. La combinaison 1 comprend deux items pour lesquels la proportion est faible, tandis que la combinaison 2 comprend deux items dont la proportion est élevée. La combinaison 3 est le cas où le nombre d'items non clés est maximal.

Tableau 3
Trois combinaisons d'items non clés

	Items non clés	
Combinaison 1 (G = 2):	9 Joyeux	(8 %)
	3 Libre	(13 %)
Combinaison 2 (G = 2):	5 Persistent	(51 %)
	6 Gentil	(42 %)
Combinaison 3 (G = 2):	Neuf items autres que l'item clé	

Nous utilisons soit le sexe, soit l'âge comme variable de domaine *Z*. Le sexe est masculin ou féminin, et les groupes d'âges sont « 20 à 29 ans », « 30 à 39 ans », « 40 à 49 ans », « 50 à 59 ans », « 60 à 69 ans » et « 70 ans et plus ».

3.2 Questionnement direct contre technique du dénombrement d'items

3.2.1 Méthodes de simulation

Premièrement, nous comparons les erreurs-types des techniques de questionnement direct et de dénombrement d'items. Dans cette expérience, nous avons testé la combinaison de « 7 Original » (item clé), la combinaison 3 (items non clés) et le sexe (variable de domaine). Le tableau de contingence fondée sur l'échantillon complet de $N = 1\,339$ figure au tableau 4.

Tableau 4
Tableau de contingence entre « 7 Original » et le sexe

	7 Original		
	$Y = 1$	$Y = 0$	Total
Hommes	46 (7,5)	569 (92,5)	615 (100,0)
Femmes	51 (7,0)	673 (93,0)	724 (100,0)
Total	97 (7,2)	1 242 (92,8)	1 339 (100,0)

La simulation a été réalisée selon la procédure suivante :

Étape 1. Poser que l'échantillon total de $N = 1\,339$ est une population.

Étape 2. Tirer un sous-échantillon *S* de taille $N\,f$, où *f* est la fraction d'échantillonnage sous échantillonnage aléatoire simple sans remise.

Étape 3. À titre de résultat simulé de la méthode de questionnement direct, calculer directement les proportions $p(Y = 1|Z = \text{hommes})$ et $p(Y = 1|Z = \text{femmes})$.

Étape 4. Diviser le sous-échantillon *S* en deux groupes S^A et S^B de tailles n^A et n^B qui ne sont pas nécessairement égales. Compter le nombre C^A d'éléments non clés sélectionnés pour chaque répondant dans S^A. En outre, compter le nombre C^B d'éléments sélectionnés, y compris l'élément clé et les éléments non clés, dans S^B.

Étape 5. À titre de résultat simulé de la technique de dénombrement d'items, calculer $p(Y=1|Z=$ hommes) et $p(Y=1|Z=$femmes) par les trois méthodes, à savoir la méthode stratifiée, la méthode par croisement et la méthode par double croisement. Dans la méthode par double croisement, poser que $w^A = n^A/(n^A + n^B)$ et $w^B = n^B/(n^A + n^B)$.

Étape 6. Poser que $f = 0{,}1$ à l'étape 2 et exécuter les étapes 2 à 5 pour 2 000 itérations. Calculer les moyennes E_D, E_S, E_C et E_W et les écarts-types SE_D, SE_S, SE_C et SE_W pour chaque méthode d'estimation afin d'obtenir une approximation des espérances et des erreurs-types des estimateurs, où les indices *D*, *S*, *C* et *W* indiquent la méthode de questionnement direct, la méthode stratifiée, la méthode par croisement et la méthode par double croisement, respectivement. De la même façon, poser que $f = 0{,}2$ et exécuter les étapes 2 à 5 pour 2 000 itérations, et ainsi de suite jusqu'à $f = 0{,}9$ inclusivement.

3.2.2 Résultats des simulations

La figure 1 donne les espérances et les erreurs-types approximatives des estimateurs. Les axes horizontaux donnent la fraction d'échantillonnage *f*. Dans les deux cas, c'est-à-dire les hommes et les femmes, l'espérance approximative de E_D est stable pour les diverses valeurs de *f*, tandis que les espérances de E_S, E_C et E_W pour la technique du dénombrement d'items fluctuent irrégulièrement. Ces fluctuations sont dues au fait que la méthode du dénombrement d'items comprend deux randomisations, à la phase d'échantillonnage et à la phase de la subdivision de l'échantillon, tandis que le scénario de questionnement direct ne comprend qu'une seule randomisation à la phase d'échantillonnage. Même si $f = 1$, l'estimateur sous la technique du dénombrement direct présente une certaine variance due à la randomisation à la phase de la subdivision de l'échantillon. Comme l'étendue des fluctuations est négligeable comparativement à la grandeur des erreurs-types, illustrées plus bas, nous concluons que le nombre de répétitions était suffisant.

Les erreurs-types, SE_D, pour la méthode de questionnement direct sont nettement plus faibles que celles observées pour la technique du dénombrement d'items, pour

laquelle les erreurs-types ne convergent pas vers zéro même si $f = 1$. Comme nous l'avons mentionné plus haut, cette situation est due à la randomisation introduite à la phase de subdivision de l'échantillon. Dans le cas de la méthode stratifiée, les erreurs-types sont manifestement plus grandes que pour les deux méthodes par croisement. Les courbes produites pour les méthodes par croisement et par double croisement se superposent pour ainsi dire et ne présentent aucune différence marquante.

Afin d'évaluer le degré de variance ou l'erreur-type des estimateurs, considérons le critère suivant, qui est analogue à l'effet du plan (Kish 1965),

$$\mathrm{Def}_{M_1, M_2} = \frac{SE^2_{M_1}}{SE^2_{M_2}},$$

où M_1 et M_2 indiquent l'une des quatre méthodes *D*, *S*, *C* et *W*. Nous ne présentons pas les résultats détaillés, mais brièvement, $\mathrm{Def}_{C,D}$ varie de 50 (quand $f = 0{,}1$) à 700 (quand $f = 0{,}9$). Autrement dit, même si nous utilisons la méthode par croisement, l'erreur-type de la technique du dénombrement d'items est près de 7 à 26 fois plus élevée que pour la méthode de questionnement direct. Cependant, la réduction de la variance due à l'utilisation de la méthode par double croisement au lieu de la méthode stratifiée varie de $\mathrm{Def}_{W,S} = 0{,}39$ (hommes) à 0,55 (femmes). Donc, l'erreur-type de la méthode par double croisement correspond à environ 62 % de celle de la méthode stratifiée, pour la valeur minimale et à 74 %, pour la valeur maximale.

3.3 Méthode stratifiée contre méthode par croisement

3.3.1 Méthodes de simulation

À la section précédente, nous avons montré que la précision des méthodes par croisement et par double croisement semble être plus grande que celle de la méthode stratifiée. Nous allons maintenant vérifier la précision de ces méthodes pour d'autres combinaisons de l'item clé, de la combinaison d'items non clés et de la variable de domaine Z grâce à des expériences en simulation.

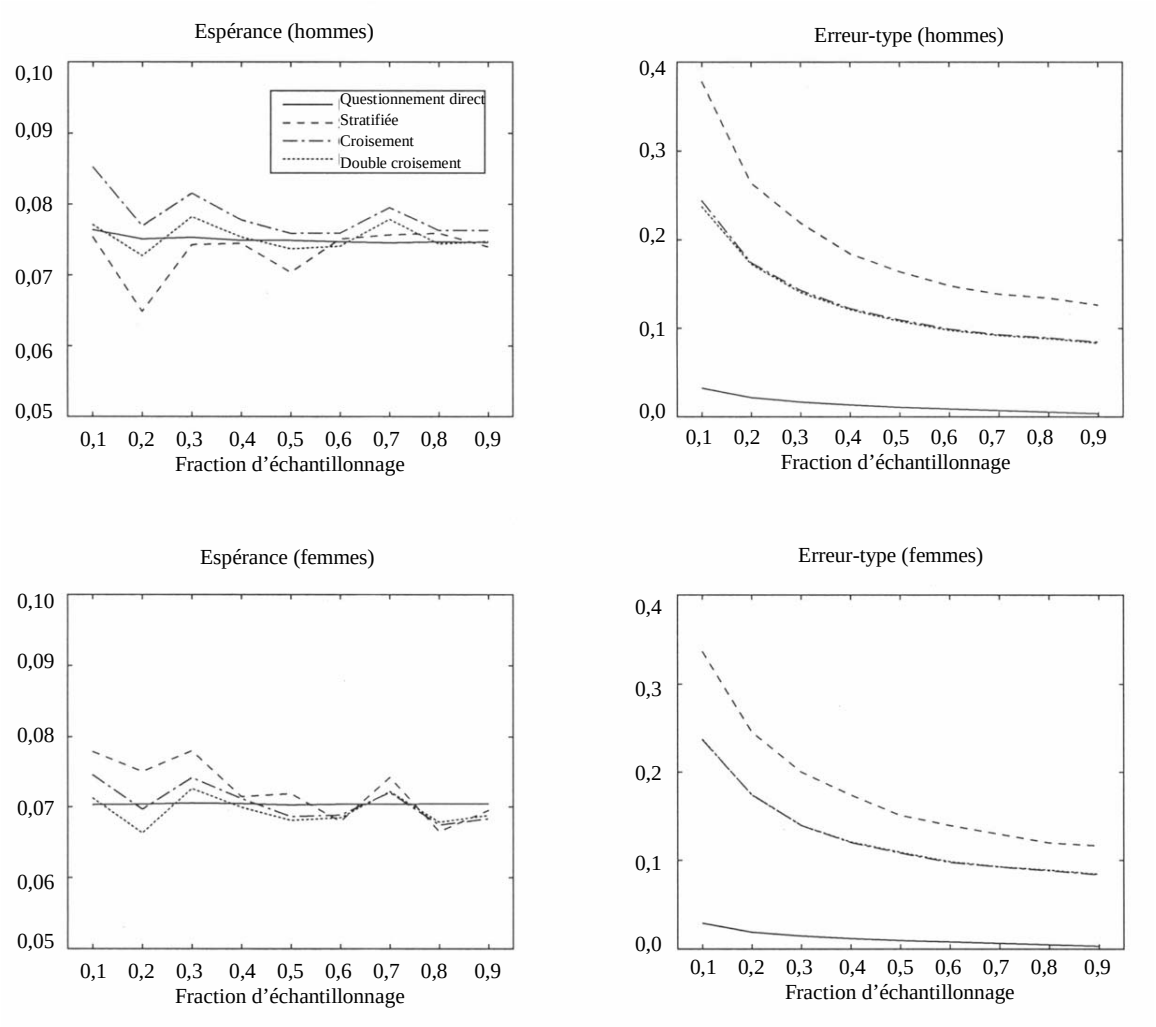

Figure 1. Espérance et erreur-type approximative des estimateurs.

Nous utiliserons tous les échantillons comme suit :

Étape 1. Calculer $P(Y=1|Z=z)$ pour chaque z d'après toutes les données pour la taille $N=1\,339$.

Étape 2. Diviser l'échantillon total ($N=1\,339$) aléatoirement en un groupe *A* et un groupe *B* de taille n^A et n^B où $N=n^A+n^B$.

Étape 3. Compter le nombre C^A d'items non clés sélectionnés pour chaque répondant du groupe *A* et compter le nombre C^B d'items sélectionnés, y compris l'item clé et les items non clés, dans le groupe *B*.

Étape 4. Estimer $p(Y=1|Z=z)$ par la méthode stratifiée, la méthode par croisement et la méthode par double croisement, respectivement.

Étape 5. Calculer la distance du chi-carré e^2 entre $P(Y=1|Z=z)$ et $p(Y=1|Z=z)$ pour chaque méthode.

$$e^2=\sum_z \frac{\{p(Y=1|Z=z)-P(Y=1|Z=z)\}^2}{P(Y=1|Z=z)}$$

Étape 6. Répéter la procédure susmentionnée de l'étape 2 à l'étape 5 pour 1 000 itérations. Calculer la moyenne et l'écart-type de e^2 pour chaque méthode.

En outre, à titre de référence, nous avons simulé la méthode stratifiée sous réponse aléatoire par la procédure suivante :

Étape 1. Soit p une proportion telle que décrite plus loin. Diviser l'échantillon total ($N=1\,339$) aléatoirement en deux groupes. Le groupe *A* est composé de Np répondants et le groupe *B*, de $N(1-p)$ répondants.

Étape 2. Soit n_z^A le nombre de répondants qui ont choisi l'item clé et $Z=z$ dans le groupe *A*. Soit n_z^B le nombre de répondants qui n'ont pas choisi l'item clé et $Z=z$ dans le groupe *B*. Soit n_z le nombre de répondants avec $Z=z$. Calculer

$$p(Y=1|Z=z)=\frac{n_z}{1\,339}\left(\frac{p-1+(n_z^A+n_z^B)/n_z}{2p-1}\right).$$

Étape 3. Calculer e^2 en se servant de la même équation que celle utilisée pour la technique du dénombrement d'items.

Étape 4. Répéter la procédure susmentionnée de l'étape 1 à l'étape 3 pour 1 000 itérations. Calculer la moyenne et l'écart-type de e^2 pour chaque méthode.

Nous avons utilisé trois valeurs de *p*, à savoir $p=0{,}2$, $p=0{,}3$ et $p=0{,}4$.

3.3.2 Résultats des simulations

Les tableaux 5 et 6 donnent la moyenne et l'écart-type pour 1 000 e^2 pour la variable de domaine *Z* de sexe et d'âge, respectivement. Les estimateurs de domaine sont d'autant plus précis que la moyenne de la « valeur e^2 » est faible. Pour certaines répétitions, nous avons obtenu des estimations illogiques de $p(Y=1|Z=z)$, qui s'écartent de l'intervalle [0, 1]. Les colonnes des tableaux intitulées « Inférieure » indiquent le nombre de répétitions pour lesquelles au moins une des estimations de $p(Y=1|Z=z)$ était inférieure à 0 et les colonnes intitulées « Supérieure » indiquent le nombre d'estimations qui étaient supérieures à 1. Idéalement, les chiffres figurant dans les colonnes « *p* illogique » devraient être nuls.

Pour toute combinaison de l'item clé, des items non clés et de la variable de domaine *Z*, les moyennes de e^2 calculées pour la méthode par double croisement sont les plus faibles et celles pour la méthode par croissement viennent au deuxième rang, l'écart étant très faible. Quand la proportion π de l'item clé est faible (« 7 Original »), que le nombre d'items non clés est grand (combinaison 3) et que le nombre d'options de la variable de domaine Z est grand (âge), la précision de la méthode stratifiée diminue considérablement comparativement aux autres combinaisons.

En outre, quand la proportion π de l'item clé est faible, l'utilisation de la méthode stratifiée produit souvent des estimations négatives. Par exemple, si l'on combine « 7 Original », la combinaison 3 et l'âge, la fréquence observée des estimations négatives est de 926 pour 1 000 itérations. Si l'on utilise la méthode par double croisement, les estimations négatives sont moins fréquentes.

Pour la méthode de la réponse aléatoire, si le nombre d'options de la variable de domaine *Z* est faible (sexe), les estimations semblent avoir la même précision que celles obtenues par les méthodes par croisement et par double croisement. Cependant, la moyenne de e^2 est un peu plus grande que celle observée pour la méthode par croisement quand le nombre d'options pour la variable de domaine *Z* est grand (âge). La méthode de la réponse aléatoire, pour laquelle seule la méthode stratifiée est disponible, produit aussi des estimations négatives, surtout quand π est faible (« 7 Original »).

Tableau 5
Moyenne et écart-type de e^2 et nombre d'estimations illogiques obtenues (la variable de domaine *Z* est le sexe)

	7 Original (7 %)				8 Poli (50 %)				2 Diligent (71 %)			
	Valeur e^2		*p* illogique		Valeur e^2		*p* illogique		Valeur e^2		*p* illogique	
	moyenne	(é.-t.)	inférieure	supérieure	moyenne	(é.-t.)	inférieure	supérieure	moyenne	(é.-t.)	inférieure	supérieure
Méthode stratifiée												
Combination 1	38	(36)	39	0	6	(6)	0	0	4	(4)	0	0
Combination 2	89	(92)	179	0	16	(17)	0	0	10	(11)	0	0
Combination 3	341	(330)	457	0	44	(43)	0	0	33	(32)	0	7
Méthode par croisement												
Combination 1	18	(24)	1	0	4	(5)	0	0	3	(3)	0	0
Combination 2	45	(65)	41	0	10	(12)	0	0	7	(8)	0	0
Combination 2	163	(239)	186	0	22	(31)	0	0	17	(23)	0	1
Méthode par double croisement												
Combination 1	18	(24)	1	0	3	(4)	0	0	2	(3)	0	0
Combination 2	45	(65)	31	0	9	(12)	0	0	6	(8)	0	0
Combination 3	163	(240)	177	0	21	(31)	0	0	16	(23)	0	0
Réponse aléatoire												
$p = 0,2$	12	(14)	0	0	3	(3)	0	0	2	(2)	0	0
$p = 0,3$	35	(43)	41	0	8	(7)	0	0	5	(5)	0	0
$p = 0,4$	158	(181)	305	0	35	(34)	0	0	23	(23)	0	3

Nota : La valeur e^2 est multipliée par 10^3.

Tableau 6
Moyenne et écart-type de e^2 et nombre d'estimations illogiques obtenues (la variable de domaine *Z* est l'âge)

	7 Original (7 %)				8 Poli (50 %)				2 Diligent (71 %)			
	Valeur e^2		*p* illogique		Valeur e^2		*p* illogique		Valeur e^2		*p* illogique	
	moyenne	(é.-t.)	inférieure	supérieure	moyenne	(é.-t.)	inférieure	supérieure	moyenne	(é.-t.)	inférieure	supérieure
Méthode stratifiée												
Combination 1	375	(226)	609	0	60	(39)	0	0	39	(26)	0	0
Combination 2	859	(507)	799	0	152	(91)	0	0	97	(58)	0	18
Combination 3	3 410	(2,108)	926	1	446	(290)	48	41	333	(217)	9	353
Méthode par croisement												
Combination 1	93	(82)	8	0	32	(20)	0	0	28	(16)	0	0
Combination 2	175	(195)	138	0	80	(42)	0	0	59	(33)	0	0
Combination 3	536	(733)	273	0	89	(95)	0	0	70	(71)	0	10
Méthode par double croisement												
Combination 1	70	(75)	8	0	13	(13)	0	0	9	(8)	0	0
Combination 2	153	(202)	93	0	45	(35)	0	0	31	(23)	0	0
Combination 3	526	(745)	246	0	72	(94)	0	0	52	(70)	0	1
Réponse aléatoire												
$p = 0,2$	158	(101)	284	0	25	(14)	0	0	17	(11)	0	0
$p = 0,3$	476	(294)	720	0	74	(42)	0	0	51	(31)	0	2
$p = 0,4$	2 181	(1 348)	945	0	335	(193)	9	9	232	(136)	0	217

Nota : La valeur e^2 est multipliée par 10^3.

4. Conclusion

Nos expériences en simulation ont produit les résultats suivants :

– La méthode par croisement ou la méthode par double croisement proposée dans le présent article devrait être utilisée pour estimer les paramètres de domaine lorsque les données sont obtenues par la technique du dénombrement d'items. Dans la première simulation, la variance des estimateurs par croisement était réduite à 39 % de la variance de l'estimation stratifiée dans le cas du minimum et à 55 %, dans le cas du maximum. Dans les études en simulation, la méthode par double croisement n'a pas produit d'amélioration importante de la précision comparativement à la méthode par croisement.

– Même si l'on utilise la méthode par double croisement, l'erreur-type des estimateurs de domaine est beaucoup plus grande que celle produite par la technique de questionnement direct.

Pour une question à laquelle les répondants évitent de donner une réponse honnête, la valeur réelle de $\pi = \bar{Y} = P(Y = 1)$ serait souvent faible. En outre, la méthode de questionnement utilisée est indirecte afin d'assurer la protection des renseignements personnels. Les répondants ont le sentiment que leur vie privée est protégée si un grand nombre d'items non clés sont inclus (Hubbard et coll. 1989). Les études en simulation montrent que, dans de telles situations, la méthode par croisement ou par double croisement est plus efficace que la méthode stratifiée classique.

Les estimateurs de domaine obtenus par la méthode stratifiée classique ne convergent généralement pas vers l'estimateur $\hat{\pi}$ comme le montre l'équation (10). L'utilisation de l'estimateur $\hat{\pi}_{PS}$ poststratifié par la variable de domaine étudiée est essentielle pour assurer la convergence. Autrement, il faut diviser l'échantillon total en deux sous-groupes de façon telle que les distributions de leur variable de domaine concordent a priori. Par contre, les estimateurs de domaine obtenus par les méthodes par croisement et par double croisement convergent vers $\hat{\pi}$ tel que le montre l'équation (21). Cependant, cela ne signifie pas que la méthode par croisement donne automatiquement l'ajustement des deux sous-groupes de sorte que les distributions d'échantillon de la variable de domaine dans les deux sous-groupes concordent. Pour la méthode par croisement, la poststratification par les variables de domaine ou d'autres variables démographiques est acceptable, mais non indispensable.

Même si l'on utilise la méthode par double croisement, on observe parfois des estimations de domaine négatives. Il est possible d'éviter ces estimations négatives en permettant à une estimation négative q_c de Q_c dans (23) d'être nulle. Cependant, ce genre de correction produit un biais positif dans $p(Y = 1 | Z = z)$.

Les données de l'enquête sur le caractère national japonais, qui ont été utilisées pour les expériences en simulation, ne sont pas délicates et n'ont pas été obtenues par la technique du dénombrement d'items. Dans l'avenir, il faudrait évaluer les propriétés de la méthode proposée en l'appliquant à des données obtenues par cette technique.

Remerciements

L'auteur remercie deux examinateurs anonymes et le rédacteur adjoint de leurs commentaires constructifs au sujet d'une version antérieure du présent article.

Bibliographie

Abul-Ela, Abdel-Latif, A., Greenberg, B.G. et Horvitz, D.G. (1967). A multiproportions RR model. *Journal of the American Statistical Association*, 62, 990-1008.

Chaudhuri, A., et Mukerjee, R.M. (1988). *Randomized Response: Theory and Techniques.* New York: Marcel Dekker.

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

Droitcour, J., Caspar, R.A., Hubbard, M.L., Parsley, T.L., Visscher, W. et Ezzati, T.M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. Dans *Measurement Errors in Surveys* (Éds. P.P. Biemer et coll.), New York: John Wiley & Sons, Inc.

Droitcour, J.A., Larson, E.M. et Scheuren, F.J. (2001). The three card method: Estimating sensitive survey items-with permanent anonymity of response. *Proceedings of the Social Statistics Section of the American Statistical Association.* Alexandria, V.A.: American Statistical Association.

Greenberg, B.G., Abul-Ela, Abdel-Latif, A., Simmons, W.R. et Horvitz, D.G. (1969). The unrelated question RR model: Theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.

Hubbard, M.L., Casper, R.A. et Lessler, J.T. (1989). Respondent reactions to item count lists and randomized response. *Proceedings of the Survey Research Section of the American Statistical Association.* Washington, D.C.: American Statistical Association. 544-548.

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

Lessler, J.T., et O'Reilly J.M. (1997). Mode of interview and reporting sensitive issues: Design and implementation of audio computer-assisted self-interviewing. *NIDA Research Monograph*, 167, 366-382.

Miller, J.D. (1985). The nominative technique: A new method of estimating heroin prevalence. *NIDA Research Monograph*, 57, 104-124.

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

Sakamoto, Y., Tsuchiya, T., Nakamura, T., Maeda, T. et Fouse, D.B. (2000). *A Study of the Japanese National Character: The Tenth Nationwide Survey (1998).* Tokyo: The Institute of Statistical Mathematics Research Report General Series 85.

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

Schuman, H., et Presser, S. (1981). *Questions & Answers in Attitude Surveys.* New York: Academic Press.

Takahasi, K., et Sakasegawa, H. (1977). A randomized response technique without making use of any randomizing device. *Annals of the Institute of Statistical Mathematics*, 29, 1-8.

U.S. General Accounting Office (1999). *Survey Methodology. An Innovative Technique for Estimating Sensitive Items.* Washington D.C.: General Accounting Office.

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

Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 57-68
Statistique Canada, N° 12-001-XIF au catalogue

Vérification des erreurs systématiques d'unité de mesure au moyen de la modélisation par mélanges

Marco Di Zio, Ugo Guarnera et Orietta Luzi [1]

Résumé

Dans le domaine de la statistique officielle, le processus de vérification des données joue un rôle important dans la rapidité de production, l'exactitude des données et les coûts d'enquête. Les techniques adoptées pour déceler et éliminer les erreurs que contiennent les données doivent essentiellement tenir compte simultanément de tous ces aspects. L'une des erreurs systématiques que l'on observe fréquemment dans les enquêtes visant à recueillir des données numériques est celle de l'unité de mesure. Cette erreur a une forte incidence sur la rapidité de production, l'exactitude des données et le coût de la phase de vérification et d'imputation. Dans le présent article, nous proposons une formalisation probabiliste du problème basée sur des modèles de mélanges finis. Ce cadre nous permet de traiter le problème dans un contexte multivarié et fournit en outre plusieurs diagnostics utiles pour établir la priorité des cas qui doivent être examinés plus en profondeur par examen manuel. Le classement des unités par ordre de priorité est important si l'on veut accroître l'exactitude des données, tout en évitant de perdre du temps en faisant le suivi d'unités qui ne sont pas vraiment critiques.

Mots clés : Vérification; erreur aléatoire; erreur systématique; vérification sélective; classification fondée sur un modèle.

1. Introduction

Les éléments qui déterminent la qualité d'un processus de vérification et d'imputation (V et I) sont multiples et ont été décrits en détail dans la littérature (Granquist 1995). Nous nous intéressons à une erreur non due à l'échantillonnage particulière qui a une forte incidence sur deux dimensions concurrentes importantes de la qualité, à savoir la rapidité de production et l'exactitude des données. En ce qui concerne l'exactitude, nous adoptons la définition proposée dans l'Encyclopedia of Statistical Sciences (1999) : [traduction] « L'exactitude s'entend de la concordance entre les statistiques et les caractéristiques cibles ». Un certain nombre de facteurs peuvent causer des inexactitudes tout au long du processus d'enquête statistique. L'inexactitude peut être réduite durant la phase de vérification et d'imputation, qu'on peut considérer comme un « outil d'amélioration de l'exactitude grâce auquel les données erronées ou très suspectes sont découvertes et, au besoin, corrigées (imputées) » (Federal Committee on Statistical Methodology 1990).

Étant donné la complexité des phénomènes étudiés et l'existence de plusieurs types d'erreur non due à l'échantillonnage, la phase de vérification et d'imputation peut être très longue et complexe (Granquist 1996). Dans la littérature spécialisée, une classification courante des erreurs repose sur la définition de deux catégories d'erreur, à savoir l'*erreur systématique* et l'*erreur aléatoire*. La première catégorie comprend les erreurs qui sont toutes de même signe et produisent un biais en statistique, tandis que la seconde englobe les erreurs qui sont réparties aléatoirement autour de zéro et ont une incidence sur la variance des estimations (*Encyclopedia of Statistical Sciences* 1999). Comprendre la nature des erreurs aide non seulement à en déterminer la source et en évaluer les effets sur les estimations, mais aussi à adopter la méthode convenant le mieux pour les corriger (Di Zio et Luzi 2002). Alors que l'approche de Fellegi–Holt (Fellegi et Holt 1976) est un modèle reconnu pour s'occuper des erreurs aléatoires, des solutions ponctuelles sont généralement adoptées pour traiter les erreurs systématiques (voir, par exemple, Euredit 2003, vol. 1, chapitre 5). Habituellement, les erreurs systématiques sont traitées avant les erreurs aléatoires, particulièrement si ces dernières le sont au moyen d'un logiciel automatisé, comme le Système généralisé de vérification et d'imputation (SGVI) (Kovar, Mac Millan et Whitridge 1988) et, plus récemment, De Waal (2003).

Dans la famille des erreurs systématiques, l'une dont l'incidence sur les estimations finales est importante et qui affecte fréquemment les données des enquêtes statistiques visant à évaluer des caractéristiques quantitatives (par exemple, enquêtes auprès des entreprises) est l'*erreur d'unité de mesure multipliée par une constante* (par exemple, 100 ou 1 000). Cette erreur est due au fait que certains répondants choisissent incorrectement l'unité de mesure lors de la déclaration de la quantité de certains items du questionnaire.

À tire d'exemples d'enquêtes souffrant de ce genre d'erreur, nous avons choisi deux enquêtes réalisées par ISTAT, c'est-à-dire l'Enquête italienne sur le coût de la

1. Marco Di Zio, Ugo Guarnera et Orietta Luzi, Italian National Statistical Institute, Via Cesare Balbo 16, 00184 Roma, Italie.

main-d'œuvre (CMO) de 1997 et le Système d'enquêtes sur l'eau (SEE) de l'Italie de 1999.

La CMO est une enquête par sondage périodique conçue pour recueillir des renseignements sur l'emploi, le nombre d'heures travaillées, les traitements et salaires, et le coût de la main-d'œuvre auprès d'environ 12 000 entreprises comptant plus de dix employés. La figure 1 représente le logarithme du coût de la main-d'œuvre (LCOST), du nombre d'employés (LEMPLOY) et du nombre d'heures travaillées (LWORKEDH) dans une matrice de diagrammes de dispersion. Notons qu'à l'étape de la vérification, la variable d'emploi est sans erreur, en raison d'une vérification préliminaire de l'information provenant des registres des entreprises (Cirianni, Di Zio, Luzi et Seeber 2000). L'analyse de la figure 1 montre que le coût de la main-d'œuvre est affecté par deux types d'erreur d'unité de mesure (c'est-à-dire facteurs de 1 million et de 1 000), tandis que le nombre d'heures travaillées ne présente que l'erreur du facteur 1 000. Ces erreurs donnent lieu à la formation de diverses grappes dans la figure 1. Il convient de souligner, que, dans chaque diagramme de dispersion, les grappes qui se trouvent dans le coin inférieur gauche représentent les données non erronées.

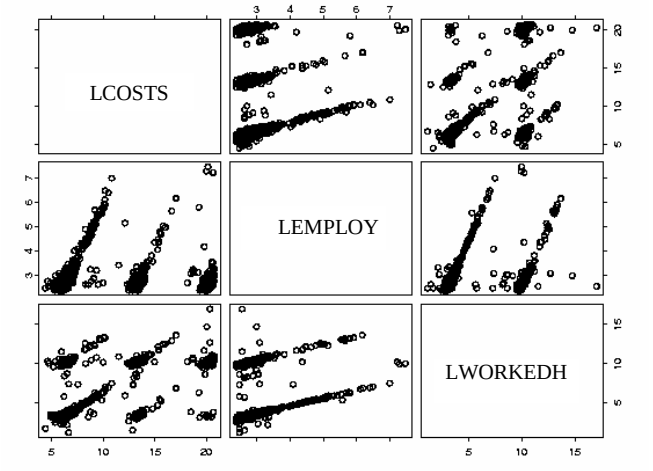

Figure 1. Diagramme de dispersion multiple du coût total de la main-d'œuvre, du nombre d'employés et du nombre d'heures travaillées (échelle logarithmique).

L'exemple du SEE sera décrit en détail à la sous-section 4.2, où nous présenterons une application de la méthode proposée dans le présent article en vue de repérer et de traiter les erreurs d'unité de mesure.

Dans le cas de l'erreur d'unité de mesure, l'élément essentiel est la localisation des items erronés plutôt que leur traitement. En fait, une fois qu'un item est classé comme étant erroné, le traitement optimal est déterminé de façon unique et consiste à prendre une mesure déterministe de correction de la valeur originale par une opération inverse (par exemple, division par 1 000) qui neutralise l'effet de l'erreur.

En général, on s'attaque à l'erreur d'unité de mesure par des procédures ponctuelles s'appuyant essentiellement sur la représentation graphique de distributions marginales ou bivariées, et sur des *vérifications de rapport*. Une vérification de rapport est une règle énonçant que la valeur d'un rapport donné entre deux variables doit être comprise dans un intervalle prédéfini. Les bornes de l'intervalle sont généralement déterminées d'après des renseignements a priori ou grâce à une analyse exploratoire des données, en utilisant éventuellement des données auxiliaires fiables. Pour le genre d'erreur susmentionné, les vérifications de rapport sont efficaces si les données sur l'une des deux variables ne contiennent pas d'erreur. En outre, elles ne permettent de tenir compte que des relations bivariées entre variables et, même si l'on recourt à l'inspection graphique interactive (par exemple, matrice de diagrammes de dispersion), on doit se limiter à une analyse par paire, et ne pas tenir compte des interactions plus complexes entre les variables. Enfin, il faut souligner que le recours à l'analyse par paire implique que les variables doivent être traitées selon une hiérarchie prédéterminée, ce qui accroît la complexité de la méthode de localisation de l'erreur.

Si l'on s'en tient aux approches classiques, le problème de localisation de l'erreur est non seulement complexe, mais également long et coûteux. La durée et le coût dépendent principalement : 1) de la complexité de la conception et de la mise en œuvre de procédures déterministes automatisées *ponctuelles* et 2) des ressources consacrées à la vérification manuelle des observations dont la probabilité d'être erronées est faible et(ou) dont l'effet sur les estimations cibles est faible (*survérification*).

Dans le présent article, nous proposons une formalisation probabiliste du problème au moyen de modèles de mélanges finis (McLachlan et Basford 1988; McLachlan et Peel 2000).

Cette modélisation peut offrir une approche statistique disciplinée, permettant d'estimer la probabilité conditionnelle qu'une observation soit affectée par une erreur d'unité de mesure. L'avantage de l'approche proposée est qu'elle représente une méthode générale permettant de faire une analyse multivariée des données et fournissant des éléments qui peuvent être utilisés pour optimiser l'équilibre entre les composantes automatiques et interactives de la procédure de vérification, c'est-à-dire entre la durée et l'exactitude (Granquist et Kovar 1997).

La présentation de l'article est la suivante. À la section 2, nous décrivons le modèle proposé, ainsi que l'algorithme EM utilisé pour estimer les paramètres. À la section 3, nous décrivons les diagnostics pour la vérification sélective. À la section 4, nous illustrons les résultats de l'application de la méthode proposée à des données simulées, ainsi que des

données réelles. Enfin, à la section 5, nous présentons nos conclusions et les futures travaux de recherche.

2. Le modèle

Il est difficile de donner une formalisation complète des erreurs aléatoires et systématiques. Dans le présent contexte, nous fournissons une définition qui, bien qu'elle ne soit pas exhaustive, inclut un grand nombre de situations courantes. Soit $\boldsymbol{X}^*$ le vecteur de variables cibles de l'enquête et $(\boldsymbol{\mu}, \Sigma^*)$ le vecteur de moyennes correspondant et la matrice des covariances. Supposons que le processus de mesure soit affecté par un mécanisme d'erreur aléatoire *R* ayant un effet sur la structure de covariance de $\boldsymbol{X}^*$, mais laissant le vecteur de moyennes inchangé et, conséquemment, représentons par $\boldsymbol{X}$ la variable « contaminée » correspondante, avec $E(\boldsymbol{X}) = E(\boldsymbol{X}^*) = \boldsymbol{\mu}$, $\mathrm{Var}(\boldsymbol{X}) = \Sigma$. En outre, nous supposons que $\boldsymbol{X}$ peut, à son tour, être affectée par un mécanisme d'erreur systématique *S* agissant uniquement sur son espérance mathématique $\mu \xrightarrow{S} \varphi(\boldsymbol{\mu})$ pour une certaine fonction φ (par exemple, si l'on suppose que le mécanisme d'erreur est additif, $\varphi(\boldsymbol{\mu}) = \boldsymbol{\mu} + \textit{constante}$). En raison des deux mécanismes d'erreur, que nous supposons être indépendants l'un de l'autre, nous pouvons décrire les données observées au moyen d'un vecteur aléatoire $\boldsymbol{Y}$ dont la loi de probabilité, sachant $\boldsymbol{X}$, dépend uniquement du mécanisme d'erreur systématique. Notre façon d'aborder le traitement des erreurs systématiques consiste à construire pour $\boldsymbol{Y}$ un modèle axé uniquement sur la détection des erreurs systématiques, donc visant à récupérer les données aléatoirement contaminées représentées par le vecteur aléatoire $\boldsymbol{X}$. Cette approche est celle généralement adoptée dans les procédures de vérification où les erreurs systématiques et les erreurs aléatoires sont traitées séparément et hiérarchiquement.

La définition donnée plus haut de l'erreur systématique comprend l'erreur d'unité de mesure, une fois que les données ont été transformées en échelle logarithmique. En fait, l'erreur d'unité de mesure a habituellement l'effet de multiplier les variables par un facteur constant. Donc, sur une échelle logarithmique, les données erronées apparaissent comme la translation d'un vecteur de constantes qui dépend des items qui sont erronés (« patron d'erreur »), tandis que la structure de covariance est la même pour chaque patron d'erreur. Qui plus est, en fait, les variables des enquêtes-entreprises sont fréquemment considérées comme suivant une loi log-normale. Donc, en échelle logarithmique, nous pouvons adopter les paramètres gaussiens.

Partant de la formalisation exposée jusqu'à présent, notre but est maintenant d'affecter chaque observation à un « patron d'erreur » particulier, ce qui revient à localiser les items entachés d'une erreur. Si nous interprétons chaque patron d'erreur comme étant une « grappe », le problème de localisation de l'erreur devient un problème de classification (*cluster analysis*) et nous pouvons profiter des enseignements de la théorie de la classification basée sur un modèle (Fraley et Raftery 2002).

Plus précisément, supposons que nous ayons n observations indépendantes $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iq})$, $i = 1, ..., n$, correspondant aux vecteurs de dimension q représentés par $\boldsymbol{X}_i = (X_{i1}, ..., X_{iq})$ avec la f.d.p. $f(x_1, ..., x_q; \boldsymbol{\theta})$, telle que $E(X_1, ..., X_q) = (\mu_1, ..., \mu_q) = \boldsymbol{\mu}$, et $\mathrm{Var}(X_1, ..., X_q) = \Sigma$.

Si nous supposons que le seul effet des erreurs systématiques sur le vecteur aléatoire $\boldsymbol{X}$ est la transformation de son espérance $\boldsymbol{\mu}$ en $\varphi_g(\boldsymbol{\mu})$, où $\varphi_g(\cdot): \mathrm{R}^q \to \mathrm{R}^q$, pour $g = 1, ..., h$, est un ensemble de fonctions connues, les fonctions φ_g caractérisent de façon univoque h grappes (patrons d'erreur) distinctes, qui ne diffèrent l'une de l'autre que par le paramètre d'emplacement. Par exemple, si l'erreur systématique affectait toutes les variables X_s, pour $s = 1, ..., q$, de la même façon en transformant leur espérance μ_s conformément à $\mu_s \to \mu_s + C$, où C est une constante connue, le nombre de grappes serait $h = 2^q$, c'est-à-dire le nombre de combinaisons différentes de l'occurrence de l'erreur sur les q variables (y compris le cas où il n'y a pas d'erreur). Dans ce cas, chaque fonction φ_g et chaque grappe correspondante sera associée à l'un des 2^q sous-ensembles possibles de variables affectées par l'erreur; par exemple, le groupe G caractérisé par le vecteur de moyennes $\boldsymbol{\mu}_G = (\mu_1, \mu_2 + C, \mu_3, \mu_4, ..., \mu_q)$ est une grappe d'unités présentant une erreur n'affectant que la variable X_2. Soulignons que nous supposons qu'il existe un matrice des covariances commune, parce que nous émettons l'hypothèse que l'erreur aléatoire possible agit de la même façon sur toutes les données.

Aux fins de la localisation de l'erreur, nous suivons une approche par modèle basée sur des modèles de mélanges finis de lois où chaque composante du mélange G_g, $g = 1, ..., h$, représente un patron d'erreur particulier. Formellement, nous supposons que les $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iq})$, pour $i = 1, ..., n$, sont i.i.d. par rapport à $\sum_{t=1}^{h} \pi_t f_t(\cdot; \boldsymbol{\theta}_t)$, où $\sum_t \pi_t = 1$ et $\pi_t \geq 0$. Les paramètres de mélange π_t représentent la probabilité qu'une observation appartienne à la t^{e} composante du mélange.

Afin de classer une observation $\boldsymbol{y}_i$ dans l'un des h groupes, nous calculons la probabilité a posteriori $\tau_g(\boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi}) = \mathrm{pr}(i^{\mathrm{e}} \text{ observation} \in G_g \mid \boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi})$, c'est-à-dire

$$\tau_g(\boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi}) = \pi_g f_g(\boldsymbol{y}_i; \boldsymbol{\theta}_g) \Big/ \sum_{t=1}^{h} \pi_t f_t(\boldsymbol{y}_i; \boldsymbol{\theta}_t) \qquad g = 1, ..., h. \quad (1)$$

La i^{e} observation est assignée à la grappe G_t si

$$\tau_t(\boldsymbol{y}_i;\ \boldsymbol{\theta},\ \boldsymbol{\pi}) > \tau_g(\boldsymbol{y}_i;\ \boldsymbol{\theta},\ \pi)\ \ g = 1,\ ...,\ h;\ g \neq t.$$

La règle d'affectation qui précède est la solution optimale du problème de classification, en ce sens qu'elle minimise le taux global d'erreur (Anderson 1984, chapitre 6).

Puisque, au lieu des paramètres $(\boldsymbol{\theta},\ \pi)$, généralement inconnus, nous utilisons les estimations du maximum de vraisemblance $(\hat{\boldsymbol{\theta}},\ \hat{\pi})$, la règle de classification devient :

$$\tau_t(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}},\ \hat{\pi}) > \tau_g(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}},\ \hat{\pi})\ \ g = 1,\ ...,\ h;\ g \neq t. \qquad (2)$$

Nous supposons que la fonction $f_t(\boldsymbol{y};\boldsymbol{\theta}_t)$ est une fonction de densité multivariée $MN(\boldsymbol{\mu}_t,\ \boldsymbol{\Sigma})$ et que chaque fonction $\varphi_g(\cdot)$ agit sur le vecteur de moyennes $\boldsymbol{\mu}$ comme une translation : $\varphi_g(\boldsymbol{\mu}) = \boldsymbol{\mu} + \boldsymbol{C}_g$, où $\boldsymbol{C}_g$ représente le vecteur de translation pour la moyenne de la g^{e} grappe que nous supposons être connue. Ce cadre, comme nous l'avons déjà souligné, convient pour le traitement de l'erreur d'unité de mesure. Afin de calculer les estimations de vraisemblance, nous utilisons l'algorithme EM, tel que proposé dans McLachlan et Basford (1988). Néanmoins, un effort supplémentaire est nécessaire pour adapter l'algorithme à notre situation particulière, où les vecteurs moyens des composantes du mélange sont liés par une relation fonctionnelle connue. Donc, alors que dans le cas sans contrainte (McLachlan et Basford 1988), un vecteur de moyennes distinct doit être estimé pour chaque composante du mélange, dans notre situation contrainte, il suffit d'en estimer un seul. L'algorithme EM modifié résultant consiste à définir une valeur initiale estimée au jugé pour les paramètres à estimer $\hat{\pi}_g^{(0)}$ pour $g = 1,\ ...,\ h$, $(\hat{\boldsymbol{\mu}}^{(0)},\ \hat{\boldsymbol{\Sigma}}^{(0)})$ et à appliquer jusqu'à la convergence le schéma récursif suivant :

i) calculer les probabilités a posteriori $\tau_{gi}^{(k)} = \tau_g^{(k)}(\boldsymbol{y}_i;\boldsymbol{\theta},\ \boldsymbol{\pi})$ sous les estimations courantes $\hat{\pi}^{(k)}$, $\hat{\boldsymbol{\mu}}^{(k)}$, $\hat{\boldsymbol{\Sigma}}^{(k)}$ (k est l'indice supérieur désignant le k^{e} cycle)

$$\hat{\tau}_{gi}^{(k)} = \frac{\hat{\pi}_g^{(k)} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_g^{(k)}\right)'\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_g^{(k)}\right)\right\}}{\sum_{t=1}^{h}\hat{\pi}_t^{(k)} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_t^{(k)}\right)'\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_t^{(k)}\right)\right\}}$$

ii) calculer les nouvelles estimations au moyen des équations récursives suivantes :

$$\hat{\pi}_g^{(k+1)} = \sum_{i=1}^{n}\hat{\tau}_{gi}^{(k)}/n$$

$$\hat{\boldsymbol{\mu}}^{(k+1)} = \sum_{g=1}^{h}\sum_{i=1}^{n}\hat{\tau}_{gi}^{(k)}\boldsymbol{y}_i/n - \sum_{g=1}^{h}\boldsymbol{C}_g\hat{\pi}_g^{(k+1)}$$

$$\hat{\boldsymbol{\Sigma}}^{(k+1)} = \sum_{g=1}^{h}\sum_{i=1}^{n}\hat{\tau}_{gi}^{(k)}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_g^{(k+1)}\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}_g^{(k+1)}\right)'/n\hat{\pi}_g^{(k+1)}.$$

Soulignons que $\hat{\boldsymbol{\mu}}_g^{(k)}$ représente $\hat{\boldsymbol{\mu}}^{(k)} + \boldsymbol{C}_g$.

Dans les applications pratiques, le choix des points de départ s'avère essentiel, comme d'habitude dans les algorithmes EM (voir Biernacki, Celeux et Govaert 2003). Pour surmonter ce problème, nous utilisons une stratégie d'initialisation, inspirée de Biernacki et coll. (2003), qui consiste en plusieurs exécutions brèves, en ce qui concerne le nombre d'itérations, de l'algorithme provenant des initialisations aléatoires, suivies par une longue exécution de l'algorithme EM provenant de la solution qui maximise la log-vraisemblance observée.

Il mérite d'être mentionné qu'à cause des contraintes d'emplacement, les paramètres qui doivent être estimés sont sensiblement moins nombreux que ceux d'un problème de mélange de lois habituel. En fait, cette différence est d'autant plus importante que le nombre de variables à analyser est grand; par exemple, dans le cas de 3 variables et de 8 grappes, nous devons estimer 16 paramètres au lieu de 37. Cet aspect est particulièrement important quand nous avons affaire à de petits échantillons. De surcroît, les contraintes sur les emplacements des grappes permettent de repérer plus facilement les « grappes rares ». En fait, les distances relatives entre les vecteurs de moyennes étant fixes, le problème d'estimation se réduit à estimer l'emplacement du polyèdre convexe dont les sommets sont les centroïdes des grappes. Autrement dit, puisque l'emplacement d'un centroïde détermine sans ambiguïté les positions de tous les autres, les paramètres des petites grappes sont estimés plus facilement que s'il n'existait pas de contrainte.

Puisque la modélisation présentée est fondée sur l'hypothèse que les observations suivent une loi normale, la validation du modèle est une question dont il faut tenir compte. Le problème de l'évaluation de la normalité dans les modèles de mélanges de lois est bien décrit dans McLachlan et Basford (1988). Il est essentiellement fondé sur les quantités $\hat{a}_{gi}$ décrites plus bas. Soit $\boldsymbol{y}_{gi}$ pour $i = 1,\ ...,\ \hat{m}_g$ les observations assignées à la g^{e} grappe pour $g = 1,\ ...,\ h$, conformément au modèle estimé. Soit $\hat{p}_{gi}$ la valeur calculée en se servant des paramètres estimés, au moyen de la formule :

$$\hat{p}_{gi} = \frac{(\nu\hat{m}_g/q)\,D\left(\boldsymbol{y}_{gi},\hat{\boldsymbol{\mu}}_g;\hat{\boldsymbol{\Sigma}}\right)}{(\nu+q)(\hat{m}_g - 1) - \hat{m}_g D\left(\boldsymbol{y}_{gi},\hat{\boldsymbol{\mu}}_g;\hat{\boldsymbol{\Sigma}}\right)}, \qquad (3)$$

où $\mathrm{D}(\cdot,\cdot;M)$ est le carré de la distance de Mahalanobis basée sur la mesure M, et $\nu = n - h - q$. Nous définissons

$\hat{a}_{gi}$ comme étant l'aire à la droite de la valeur $\hat{p}_{gi}$ sous la distribution $F_{q,\nu}$ (pour des précisions, consulter McLachlan et Basford 1988, chapitre 2).

Sous l'hypothèse de normalité, $\hat{a}_{gi}$ pour $i = 1, ..., \hat{m}_g$ est approximativement uniformément distribué sur l'intervalle (0,1). Hawkins (1981) propose d'utiliser la statistique d'Anderson-Darling pour évaluer la distribution uniforme de $\hat{a}_{gi}$. Les $\hat{a}_{gi}$ sont également utiles pour déceler les valeurs extrêmes, c'est-à-dire les observations aberrantes par rapport au modèle. Dans McLachlan et Basford (1988), la probabilité que $\boldsymbol{y}_{gi}$ soit atypique est d'autant plus élevée que $\hat{a}_{gi}$ est faible, si bien que toutes les observations avec $\hat{a}_{gi} < \alpha$, où α est un seuil spécifié, peuvent être considérées comme étant atypiques. Les valeurs proposées du seuil varient de $\alpha = 0{,}05$ à $\alpha = 0{,}005$, selon les observations aberrantes (valeurs plus ou moins extrêmes) qu'il faut sélectionner.

3. Diagnostics pour la vérification sélective

Une fois que les paramètres du mélange sont estimés, nous pouvons classer les données dans les diverses grappes; autrement dit, pour chaque observation, nous pouvons déterminer s'il s'agit d'une erreur ou non, et sur quelle variable l'erreur porte. Cependant, divers types d'observations critiques peuvent être définis après la phase de modélisation, à savoir les unités classées dans une grappe, mais ayant une probabilité non négligeable d'appartenir à une autre grappe, et les observations qui sont des valeurs aberrantes par rapport au modèle.

Afin d'accroître l'exactitude des données, il serait utile de procéder à une double vérification des observations critiques (par examen manuel, ou, dans les cas les plus difficiles, par un suivi). Par ailleurs, afin de réduire la survérification éventuelle et les coûts de vérification, il convient de concentrer l'examen manuel et(ou) le suivi sur les observations les plus critiques. Le modèle de mélanges proposé fournit des diagnostics directs que l'on peut utiliser à cette fin.

Un premier type d'unités critiques est représenté par les observations éventuellement classées incorrectement. Afin de déterminer le degré de confiance dans la classe attribuée à une observation $\boldsymbol{y}_i$, nous pouvons considérer la probabilité correspondante résultant de (2). Les observations pour lesquelles cette probabilité n'est pas très proche de l'unité ont une probabilité non négligeable d'appartenir à une autre grappe. Ces observations sont celles situées dans la région où les composantes du mélange se superposent.

En plus du type susmentionné d'unités critiques, il existe d'autres observations qui sont éloignées de toutes les grappes (toutes les composantes du mélange), c'est-à-dire les valeurs aberrantes par rapport au modèle. Ces observations représentent aussi des situations critiques. Afin de repérer ce genre de valeur aberrante, nous nous servons des quantités $\hat{a}_{ij}$ décrites à la section précédente.

La probabilité de classification et l'indice d'atypicalité $\hat{a}_{gi}$ devraient être utilisés, conformément à une approche de vérification sélective/selon l'importance (Latouche et Berthelot 1992; Lawrence et McKenzie 2000), pour construire des fonctions de score appropriées afin de déterminer l'ordre de priorité des unités critiques. Nous donnons un exemple d'utilisation de ces diagnostics dans ce but à la sous-section 4.2.

4. Exemples

Nous décrivons ici certaines expériences réalisées en vue d'étudier les particularités de la méthode proposée. En premier lieu, grâce à une étude en simulation, nous analysons les propriétés du modèle proposé lorsqu'il est appliqué à des données qui s'écartent de la normalité. Deuxièmement, au moyen de données réelles, nous décrivons comment l'approche peut être appliquée dans le domaine de la statistique officielle.

Toutes les expériences sont réalisées dans l'environnement R pour calcul statistique (http://www.r-project.org/).

4.1 Exemple simulé : Déviation par rapport à la loi normale

Dans cette expérience, nous décrivons les résultats obtenus en appliquant la méthode du mélange de lois aux trois populations distinctes illustrées à la première ligne de la figure 2. La première distribution est une loi normale bivariée (MN), qui représente donc le cas où le modèle est spécifié correctement. La deuxième correspond à une loi *t* bivariés (MT), c'est-à-dire qu'elle mime la situation où la déviation par rapport à la loi normale se résume essentiellement à des queues plus lourdes. La dernière est une loi *t* asymétrique bivariée (ST) (Azzalini et Capitanio 2003; Azzalini, Dal Cappello et Kotz 2003), qui représente une population distribuée conformément à une loi asymétrique à queues lourdes.

Pour ces distributions, nous construisons un modèle de mélanges à quatre composantes en ajoutant à chaque unité l'un des quatre vecteurs de translation $\mathbf{C}_1 = (0, 0)$, $\mathbf{C}_2 = (0, \log(1\,000))$, $\mathbf{C}_3 = (\log(1\,000), 0)$ ou $\mathbf{C}_4 = (\log(1\,000), \log(1\,000))$, avec les probabilités $\pi_1 = 0{,}5$, $\pi_2 = 0{,}1$, $\pi_3 = 0{,}1$ et $\pi_4 = 0{,}3$, respectivement. Ces paramètres représentent les proportions de mélange du modèle et ont trait, respectivement, aux probabilités de l'absence de translation dans les variables, d'une translation dans une des deux variables seulement et d'une translation dans les deux variables, respectivement. À partir de chaque mélange, nous tirons 100 échantillons de 1 000 observations. À la deuxième ligne de la figure 2 nous présentons l'un de ces échantillons (MN–Mixt, MT–Mixt et

ST–Mixt), correspondant aux trois populations distinctes MN, MT et ST, respectivement.

Pour chaque échantillon, nous calculons le nombre de classifications correctes obtenues en utilisant le modèle de mélanges décrit à la section 2. Le nombre moyen de classifications correctes sur les 100 échantillons est présenté au tableau 1.

L'examen du tableau 1 montre que la fréquence des classifications correctes diminue lorsque la déviation par rapport à la loi normale augmente. Cependant, elle semble acceptable même dans le cas critique ST où la population est caractérisée par une distribution à la fois asymétrique et à queues lourdes.

Tableau 1
Fréquence des classifications correctes

	MN	MT	ST
% correctement classée	98,5	97,5	95,6

Comme nous en avons discuté à la section 3, l'approche du modèle de mélanges fournit des éléments (tels que le degré d'atypicalité et la probabilité de classification) qui peuvent être utilisés pour déterminer l'ordre de priorité des unités pour l'examen manuel. Par conséquent, une évaluation globale de la procédure devrait aussi tenir compte des résultats d'une approche de vérification sélective fondée sur ces diagnostics du modèle.

Afin d'analyser les caractéristiques de l'indice d'atypicalité et de la probabilité de classification, nous examinons un seul échantillon de 1 000 observations tiré à partir des trois populations présentées jusqu'à présent. La figure 3 illustre les trois échantillons MN–Mixt(a), MT–Mixt(a) et ST–Mixt(a), où les unités classées incorrectement sont représentées par une croix sur le même graphique. Le nombre d'unités classées incorrectement est 19 pour MN–Mixt, 20 pour MT–Mixt et 36 pour ST–Mixt.

Pour cet échantillon, nous nous concentrons sur l'effet de divers seuils pour l'atypicalité (α) et la probabilité de classification (β). Pour chaque seuil, nous présentons aux tableaux 2 et 3 le nombre d'unités situées sous le seuil, c'est-à-dire le nombre d'observations critiques (*Atyp. – OC*, *Pr. class. – OC*), et parmi ces observations, le nombre d'unités mal classées (*Atyp. – MC*, *Pr. class. – MC*).

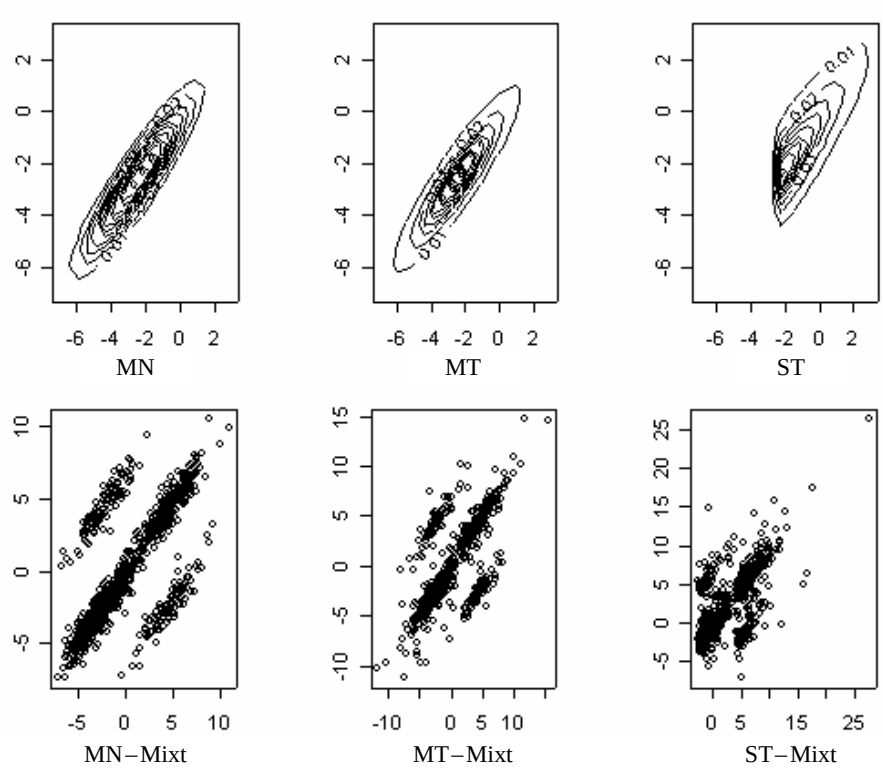

Figure 2. Tracés de contours des trois distributions bivariées : multinormale (MN), *t* de Student (MT), *t* asymétrique (ST), et diagrammes de dispersion des mélanges correspondant MN–Mixt, MT–Mixt et ST–Mixt.

En ce qui concerne l'atypicalité, nous constatons que, si le modèle est spécifié correctement, le rôle de l'indice d'atypicalité dans la correction des unités mal classées est négligeable, tandis que les probabilités de classification donnent de meilleurs résultats. Par ailleurs, l'indice d'atypicalité est important si le modèle dévie de la normalité. Il est clair que le nombre d'observations sélectionnées pour une combinaison de seuils α et β n'est pas égal à la somme des fréquences obtenues dans les tableaux 2 et 3. Donc, afin d'évaluer l'effet collectif des deux indices, nous choisissons les deux seuils suivants $\alpha = 0{,}005$ et $\beta = 0{,}975$. Nous présentons à la figure 3 (deuxième ligne) les unités sélectionnées pour la valeur d'atypicalité seulement (carrés), pour la probabilité de classification seulement (triangles) et pour les deux seuils conjugués (croix). L'examen de ces figures montre que l'atypicalité a surtout une incidence sur la détection des valeurs aberrantes, tandis que la probabilité de classification joue sur les régions chevauchantes. Le tableau 4 donne le nombre d'unités critiques sélectionnées et, parmi celles-ci, le nombre d'unités mal classées.

Nous notons que, pour la population MN–Mixt, à part une observation, toutes les unités mal classées sont sélectionnées. Pour MT–Mixt, nous arrivons à sélectionner 14 des 20 unités mal classées et, dans l'échantillon le plus critique ST–Mixt, nous sélectionnons 24 des 36 unités mal classées.

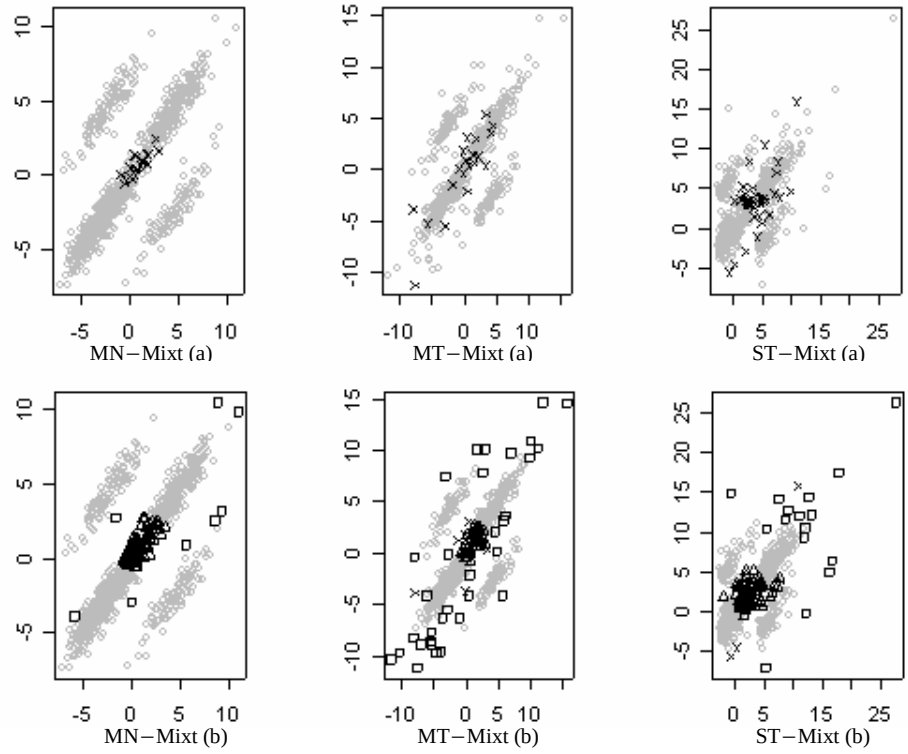

Figure 3. Unités mal classées (croix) dans MN–Mixt(a), MT–Mixt(a) et ST–Mixt(a). Unités critiques pour l'atypicalité (carrés), pour la probabilité de classification (triangles) et pour les deux (croix) dans MN–Mixt(b), MT–Mixt(b) et ST–Mixt(b).

Tableau 2
Nombres d'observations critiques (OC) et d'unités mal classées (MC) pour trois seuils distincts d'atypicalité

	MN–Mixt		**MT–Mixt**		**ST–Mixt**	
α	*Atyp – OC*	*Atyp – MC*	*Atyp – OC*	*Atyp – MC*	*Atyp – OC*	*Atyp – MC*
0,05	50	1	84	9	68	14
0,01	15	0	50	7	33	8
0,005	8	0	39	7	20	5
0,001	4	0	25	4	14	2

Tableau 3
Nombres d'observations critiques (OC) et d'unités mal classées (MC) pour trois seuils distincts de probabilité de classification

	MN–Mixt		MT–Mixt		ST–Mixt	
β	*Pr. Class – OC*	*Pr. Class – MC*	*Pr. Class – OC*	*Pr. Class – MC*	*Pr. Class – OC*	*Pr. Class – MC*
0,99	119	19	63	12	182	26
0,975	76	18	46	11	82	26
0,95	55	14	35	9	66	21

Tableau 4
Nombres d'observations critiques (OC) et d'unités mal classées (MC) pour l'atypicalité et la probabilité de classification

	MN–Mixt		MT–Mixt		ST–Mixt	
Seuils	*OC*	*MC*	*OC*	*MC*	*OC*	*MC*
$\alpha = 0{,}005,\ \beta = 0{,}975$	84	18	79	14	98	24

4.2 Appplication à des données réelles : Le Système d'enquêtes sur l'eau de Italie de 1999

À la présente section, nous décrivons une application de l'approche du modèle de mélanges à des données d'enquête réelles. Ces données sont tirées du Système d'enquêtes sur l'eau (SEE) de l'Italie de 1999. Le SEE est un recensement visant à recueillir des renseignements sur le prélèvement, la fourniture et la consommation d'eau dans les 8 100 municipalités italiennes. Nous limitons l'analyse aux municipalités appartenant à l'un des domaines de données définis par altimétrie (2 041 observations) et aux variables principales *Total de l'eau facturée* (TI) et *Total de l'eau fournie* (TS). Ces variables ont trait toutes deux à des volumes d'eau et il est demandé aux répondants de déclarer ces volumes en milliers de mètres cubes. Le diagramme de dispersion en échelle logarithmique de la quantité d'eau facturée par habitant (*WI*) en fonction de la quantité d'eau fournie par habitant (*WS*) (figure 4) montre l'existence de quatre grappes correspondant à une erreur d'unité de mesure dans l'une, les deux ou ni l'une ni l'autre des deux variables cibles. Ces erreurs sont probablement dues à un malentendu chez certains répondants qui ont exprimé les volume en litres ou en mètres cubes plutôt qu'en milliers de mètres cubes, comme il l'était demandé. Comme il fallait s'y attendre, les deux grappes dont la population est la plus nombreuse sont celles correspondant aux unités non erronées et aux unités où les deux variables sont erronées. Néanmoins, nous observons la présence de deux grappes rares correspondant aux observations où l'erreur d'unité de mesure a trait uniquement à *TI* ou à *TS*, respectivement.

Dans le tableau 5, une étiquette est attribuée à chaque grappe associée à un patron d'erreur particulier. Par souci de simplicité, nous introduisons deux drapeaux E_{TS} et E_{TI} dont la valeur est égale à 1 ou 0, selon que la variable correspondante est affectée ou non par l'erreur d'unité de mesure ou non.

Afin de dépister et de corriger les erreurs d'unité de mesure, nous appliquons la procédure décrite aux sections 2 et 3. Nous classons chaque observation en fonction d'un patron d'erreur particulier, autrement dit nous affectons chaque unité à l'une des grappes G_t, pour $t = 1, ..., 4$. Les résultats sont présentés au tableau 6.

Pour chaque unité, nous calculons aussi l'indice d'atypicalité et nous choisissons le seuil $\alpha = 0{,}005$ afin de repérer les unités atypiques. Pour ce seuil, 71 observations sont sélectionnées comme étant atypiques, et marquées par des « croix » dans la figure 7. Après avoir calculé les valeurs $\hat{a}_{gi}$ conformément à la formule (3), nous pouvons faire un test pour évaluer l'hypothèse de normalité. En fait, à l'instar de McLachlan et Basford (1988, chapitre 2), nous appliquons le test d'Anderson-Darling de l'uniformité de $\hat{a}_{gi}$ à chaque grappe estimée individuelle. La valeur *p* est inférieure à 0,001 pour les deux plus grandes grappes. Puisque le test est fondé sur des approximation asymptotiques, nous ne tenons pas compte des résultats obtenus pour les deux autres populations rares. À la figure 5, nous donnons les quantiles empiriques d'échantillon en fonction des quantiles normaux des variables log(WI) et log(WS), en nous concentrant uniquement sur le sous-ensemble de données classées comme étant non erronées. Nous constatons que l'écart par rapport à la loi normale est dû principalement à des queues lourdes. Compte tenu des résultats obtenus à la section 4.1, où la méthode a donné des résultats satisfaisants également dans des conditions non gaussiennes, nous nous attendons à ce que l'approche du mélange de lois donne de bons résultats pour les données d'enquête. Les résultats de l'application illustrés ci-après confirment ce comportement.

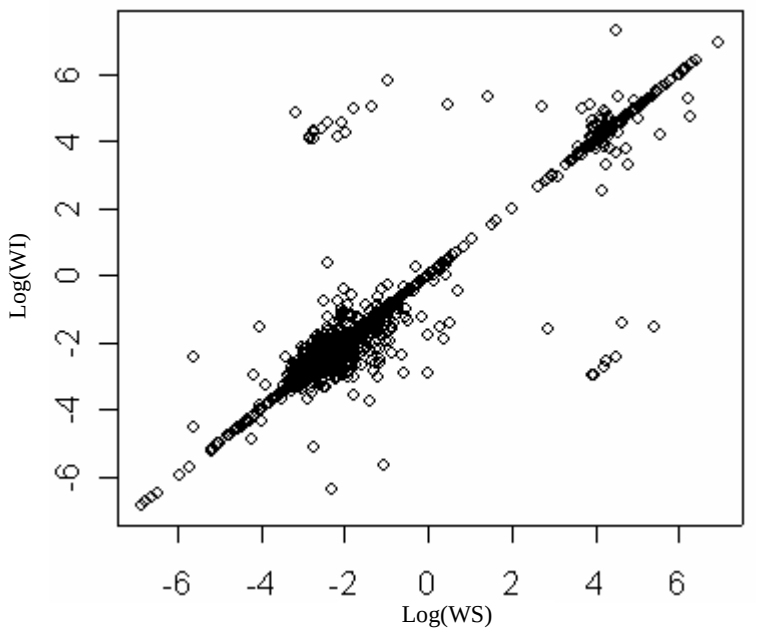

Figure 4. Diagramme de dispersion de log(WI) en fonction de log(WS).

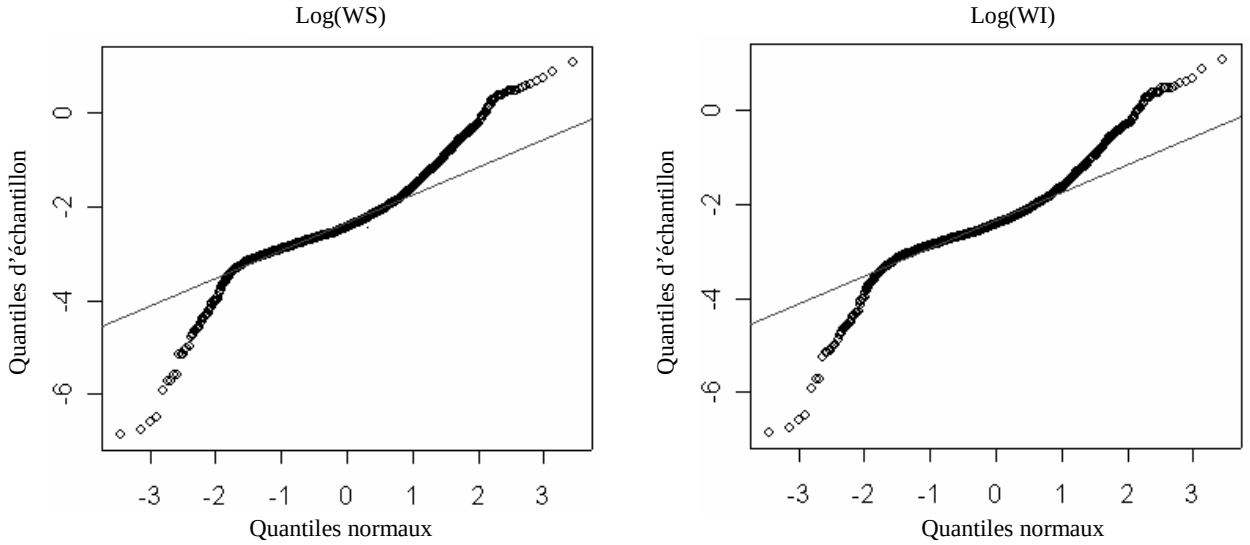

Figure 5. Graphique QQ normal de log(WS) et log(WI).

Tableau 5
Patrons d'erreur et étiquettes de grappe

Patron d'erreur	$E_{TS} = 0$ $E_{TI} = 0$	$E_{TS} = 0$ $E_{TI} = 1$	$E_{TS} = 1$ $E_{TI} = 0$	$E_{TS} = 1$ $E_{TI} = 1$
Étiquette de grappe	*G1*	*G2*	*G3*	*G4*

Tableau 6
Nombre d'unités assignées à chaque grappe

Étiquette de grappe	**G1**	**G2**	**G3**	**G4**
N. d'unités	1 800	16	10	215
%	88,2	0,8	0,5	10,5

Dans la dernière partie de cette section, nous montrons comment utiliser les probabilités a posteriori afin d'accorder la priorité, pour l'examen, aux unités qui, en principe, offriront le plus grand avantage de vérification, compte tenu de l'effet éventuel de la vérification manuelle sur les estimations. À cette fin, notons que la classification erronée d'une observation a pour conséquence que les valeurs finales d'au moins une variable diffèrent par un facteur de multiplication des valeurs réelles correspondantes. Ces écarts peuvent affecter sérieusement l'exactitude des

estimations et causer un biais important. Afin de sélectionner les unités éventuellement erronées qui sont les plus susceptibles d'avoir un effet important sur le estimations cibles, nous adoptons l'*approche de la vérification sélective*. Soit X_1, X_2 les variables *TS* et *TI*, respectivement. Pour chaque unité u_i, $i = 1, ..., n$, et pour chaque variable X_j, $j = 1, 2$, définissons :

X_{ij} : données dépourvues d'erreur systématique;

Y_{ij} : données observées;

$\tilde{X}_{ij}$: données après le traitement de l'erreur systématique d'après la classification au moyen d'un modèle de mélanges (c'est-à-dire $\tilde{X}_{ij} = Y_{ij}$ ou $\tilde{X}_{ij} = Y_{ij}/1\,000$ selon la grappe à laquelle l'unité u_i est assignée).

Supposons que les estimations cibles soient les totaux de population $T(X_j) = \sum_i X_{ij}$. En outre, représentons par $E_\xi(\cdot)$ l'espérance sur la distribution de la variable aléatoire X_j, sachant les données observées Y_{ij} et les données après correction $\tilde{X}_{jj}$. Alors, il découle de l'inégalité $|\sum_i E_\xi(X_{ij} - \tilde{X}_{ij})| \leq \sum_i E_\xi |X_{ij} - \tilde{X}_{ij}|$ que la quantité dans le deuxième membre peut être considérée comme une borne supérieure du biais probable de l'estimation du total pour la variable X_j fondée sur les valeurs corrigées $\tilde{X}_{ij}$. La dernière considération donne à penser à une méthode pour sélectionner les unités les plus « influentes » en ce qui concerne l'estimation $T(X_j)$: afin de garantir le niveau requis d'exactitude et de réduire au minimum les coûts de la vérification manuelle, nous définissons une fonction de score local $S_{ij} = (E_\xi |X_{ij} - \tilde{X}_{ij}|)/\hat{T}(X_j)$, où $\hat{T}(X_j)$ est une estimation de référence pour $T(X_j)$, par exemple l'estimation provenant d'une enquête antérieure, ou une estimation robuste. Dans notre cas, afin de rendre robuste l'estimation préliminaire, nous commençons par exclure des données les observations atypiques, puis nous calculons la valeur moyenne sur ce sous-ensemble, et nous la multiplions par le nombre total d'unités.

Le score local S_{ij} mesure l'effet de l'erreur d'unité de mesure éventuellement associée à l'unité u_i sur l'estimation cible $T(X_j)$. Alors, nous pouvons trier les unités en fonction de leur score S_{ij} et, en commençant par les valeurs les plus élevées, sélectionner les premières unités jusqu'à ce que la somme des valeurs S_{ij} restantes soient inférieures à un seuil préétabli.

Si nous considérons simultanément les deux variables TS et TI, nous pouvons obtenir un score global S_i, pour $i = 1, ..., n$, en combinant comme il convient les fonctions de score local S_{ij}, $j = 1, 2$. Les choix possibles sont $S_i = (S_{i1} + S_{i2})/2$, ou $S_i = \max_{j=1,2} S_{ij}$. Par exemple, la dernière fonction assure que l'effet de l'erreur d'unité de mesure éventuellement associée à u_i sur chaque estimation ne soit pas supérieur à S_i.

Afin de calculer les scores S_{ij}, nous devons estimer l'espérance conditionnelle $E_\xi |X_{ij} - \tilde{X}_{ij}|$ pour chaque unité u_i, $i = 1, ..., n$, et pour chaque variable X_j pour $j = 1, 2$, ce qui peut se faire facilement au moyen des probabilités a posteriori. Par exemple, supposons que l'unité u_i ait été assignée à la grappe G_2. Cela signifie que, pour cette unité, la valeur observée de TS (Y_{i1}) a été considérée correcte, tandis que la valeur observée de TI (Y_{i2}) a été considérée comme étant affectée d'une erreur d'unité de mesure (c'est-à-dire multipliée par 1 000). La correction consiste à diviser par 1 000 la valeur observée de TI, c'est-à-dire ($\tilde{X}_{i1} = Y_{i1}$, $\tilde{X}_{i2} = Y_{i2}/1\,000$). L'espérance conditionnelle $E_\xi |X_{ij} - \tilde{X}_{ij}|$ peut être calculée comme suit :

$$
\begin{aligned}
E_\xi \left| X_{i1} - \tilde{X}_{i1} \right| &= \left| Y_{i1} - Y_{i1} \right| \Pr(u_i \in G_1 \cup G_2) \\
&\quad + \left| \frac{Y_{i1}}{1\,000} - Y_{i1} \right| \Pr(u_i \in G_3 \cup G_4) \\
&= \frac{999}{1\,000} Y_{i1} (\hat{\tau}_{3i} + \hat{\tau}_{4i})
\end{aligned}
$$

$$
\begin{aligned}
E_\xi \left| X_{i2} - \tilde{X}_{i2} \right| &= \left| \frac{Y_{i2}}{1\,000} - \frac{Y_{i2}}{1\,000} \right| \Pr(u_i \in G_2 \cup G_4) \\
&\quad + \left| Y_{i2} - \frac{Y_{i2}}{1\,000} \right| \Pr(u_i \in G_1 \cup G_3) \\
&= \frac{999}{1\,000} Y_{i2} (\hat{\tau}_{1i} + \hat{\tau}_{3i}),
\end{aligned}
$$

où $\hat{\tau}_{gi}$ est la probabilité estimée que l'unité u_i appartienne à la grappe G_g. De façon semblable, nous pouvons calculer les fonctions de score pour toutes les unités.

En pratique, dans notre application, nous trions les unités en fonction de leur score global S_i, $\max_{j=1,2} S_{ij}$ (ordre ascendant). Puis, nous excluons de l'examen manuel toutes les premières observations, de sorte que la somme cumulative de leurs S_i soit inférieure à δ, où δ est un seuil de tolérance spécifié pour l'effet sur les estimations des erreurs encore présentes dans les données. À la figure 6, nous présentons le comportement de la somme cumulative de S_i, $S_{(i)} = \sum_{k \leq i} S_k$, pour les 10 premières observations les plus critiques. Il convient de souligner que, par souci de clarté, nous n'avons pas présenté toutes les observations, parce, pour la plupart d'entre elles, $S_{(i)}$ est proche de zéro, ce qui produit une image illisible en ce qui concerne les différences de grandeur. Notons que nous prévoyons une erreur relative résiduelle inférieure à $\delta = 0{,}001$ en sélectionnant uniquement les deux premières unités (représentées par des croix).

À la figure 7, nous représentons toutes les unités sélectionnées parce qu'elles étaient atypiques (71) et(ou) à cause de l'effet relatif de leur erreur potentielle sur les estimations (2) : les croix correspondent aux observations qui sont critiques du point de vue de l'atypicalité, tandis que les carrés indiquent les deux autres types d'unités critiques.

Nous avons comparé ces résultats à ceux obtenus par la procédure officielle. Sur les 1 968 unités non sélectionnées pour un examen manuel, 1 911 observations sont sans erreur ou entachées d'une erreur d'unité de mesure uniquement. Pour toutes, la classification donnée par le modèle de mélanges est correcte. Parmi les 57 unités caractérisées par d'autres typologies d'erreur, 45 sont classées dans la catégorie des unités non affectées par l'erreur d'unité de mesure, tandis que 12 sont classées dans la catégorie de l'erreur par un facteur 1 000 dans les deux variables. Cette dernière erreur de classification peut être expliquée par l'existence d'autres erreurs systématiques (facteurs 100 et 10 000) qui ne sont pas prises en compte dans le modèle utilisé pour notre exemple.

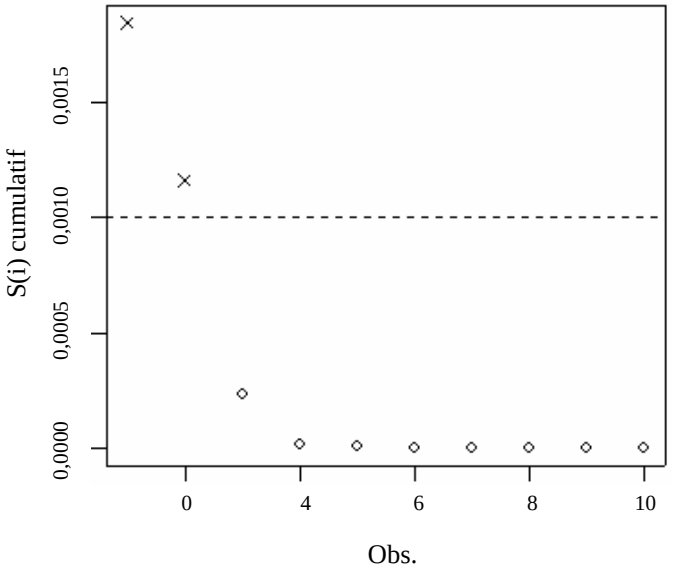

Figure 6. Tracé du score cumulatif $S_{(i)}$ pour les 10 premières observations les plus critiques.

Une autre comparaison a trait à l'estimation des totaux. Sous l'hypothèse que les valeurs sélectionnées pour un examen manuel des valeurs critiques sont groupées convenablement, les écarts relatifs entre la valeur réelle du total d'après la procédure officielle $T(X_j)$ et l'estimation du modèle $\hat{T}(X_j)$ sous la forme $B(X_j) = (|\hat{T}(X_j) - T(X_j)|)/T(X_j)$, pour $j = 1, 2$, sont $B(X_1) = 0{,}005$ et $B(X_2) = 0{,}002$. Ces valeurs ne sont pas directement comparables au seuil de tolérance de $\delta = 0{,}001$; en fait, ce seuil a trait uniquement à l'effet des erreurs d'unité de mesure encore présentes, tandis que $B(X_j)$ est également affecté par d'autres formes d'erreurs. Donc, pour une comparaison plus directe, nous remplaçons pour ces unités les valeurs incorrectes par les valeurs « vraies » et obtenons $B(X_1) = B(X_2) = 0$. Ce degré de performance particulièrement élevé du modèle s'explique par le faible degré de superposition des grappes, comme le montre clairement la figure 7.

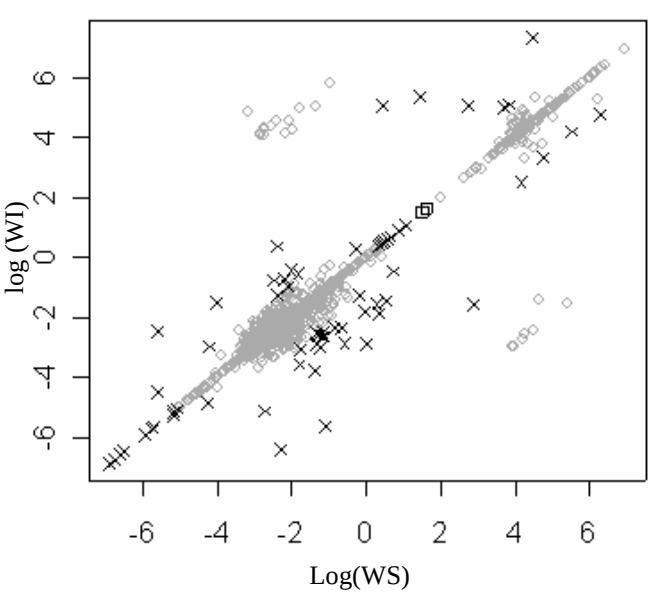

Figure 7. Diagramme de dispersion de log(WI) par rapport à log(WS). Les croix indiquent les unités critiques pour l'atypicalité et les carrés, les unités critiques pour l'effet de leur erreur éventuelle.

5. Mot de la fin et futurs travaux

Dans le présent article, nous proposons un modèle de mélanges finis pour traiter un type particulier d'erreur systématique qui entache fréquemment les données d'enquête numériques continues. L'approche proposée a les avantages, comparativement aux approches classiques, d'énoncer formellement le problème dans un contexte multivarié, d'être facilement implantée dans un logiciel généralisé et de fournir naturellement des diagnostics utiles pour établir l'ordre de priorité des unités douteuses contenant éventuellement des erreurs influentes. Cette dernière caractéristique est particulièrement importante quand la situation est critique, c'est-à-dire quand différents patrons d'erreur se superposent ou, autrement dit, quand les erreurs d'unité de mesure se situent parmi les observations plausibles. Dans ces circonstances, un examen manuel est nécessaire. Par conséquent, il est important d'optimiser la sélection des observations critiques afin de gagner du temps et d'économiser de l'argent. Tous ces avantages sont dus à l'adoption d'une approche basée sur un modèle. Par ailleurs, il est évident qu'une telle approche sous-entend des problèmes associés aux hypothèses sous-jacentes. Cependant, si l'on s'en tient aux expériences décrites dans l'article, il semble que la

technique proposée donne également des résultats satisfaisants dans les cas de déviation par rapport à l'hypothèse de normalité. Néanmoins, il faut mentionner qu'en cas d'écart extrême par rapport à la loi normale, c'est-à-dire quand la distribution n'est pas unimodale, la méthode échouera vraisemblablement. Cela peut se produire dans des situations réelles, lorsque les données contiennent diverses grappes; par exemple, les différences entre le revenu des hommes et des femmes pourraient donner lieu à une distribution bimodale du revenu proprement dit. Dans certains cas, le problème peut être surmonté en stratifiant les données d'après certaines variables explicatives, comme le sexe dans l'exemple précédent. Un autre moyen d'aborder ce genre de problème consisterait à modéliser à leur tour chaque grappe par un mélange de gaussiennes, donc obtenir un « mélange de modèles de mélanges » (McLachlan et Peel 2000; Di Zio, Guarnera et Rocci 2004).

Enfin, une dernière préoccupation a trait au nombre de variables qui peuvent être traitées simultanément. En réalité, le nombre de grappes et, donc, le nombre de paramètres de mélange π_t peuvent croître exponentiellement relativement au nombre de variables, ce qui rend l'estimation des paramètres ardue. Cependant, nous mentionnerons que le nombre de paramètres reliés au vecteur de moyennes et à la matrice des covariances augmente nettement plus lentement, à cause des contraintes caractéristiques de notre modèle.

Remerciements

Nous remercions les examinateurs et le rédacteur associé de leurs commentaires constructifs.

Bibliographie

Anderson, T.W. (1984). *An introduction to Multivariate Statistical Analysis*. Deuxième édition. New York: John Wiley & Sons, Inc.

Azzalini, A., et Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew–t distribution. *Journal of the Royal Statistical Society* (*B*), 65, 367-389.

Azzalini, A., Dal Cappello, T. et Kotz, S. (2003). Log-skew-normal and log-skew–t distributions as models for family income data. *Journal of Income Distribution*, 11, 13-21.

Biernacki, C., Celeux, G. et Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41, 561-575.

Cirianni, A., Di Zio M., Luzi O. et Seeber, A.C. (2000). The new integrated data editing procedure for the Italian Labour Cost survey: Measuring the effects on data of combined techniques. *Proceedings of the International Conference on Establishment Surveys II*, Buffalo, 7-21.

De Waal, T. (2003). Résolution du problème de localisation des erreurs par la génération de sommets. *Techniques d'enquête*, 29, 1, 81-90.

Di Zio, M., Guarnera, U. et Rocci, R. (2004). A mixture of mixture models to detect unity measure error. *Proceedings in Computational Statistics*, (Éd. Antoch Jaromir), 919-927, Physica Verlag, Prague, August 23-28.

Di Zio, M., et Luzi, O. (2002). Combining methodologies in a data editing procedure: an experiment on the survey of Balance Sheets of Agricultural Firms. *Italian Journal of Applied Statistics*, 14, 1, 59-80.

Encyclopedia of Statistical Sciences (1999). New York: John Wiley & Sons, Inc. Mise à jour. 3, 621-629.

Euredit (2003). *Towards Effective Statistical Editing and Imputation Strategies – Findings of the Euredit project*, 1, 2. À paraître. Maintenant disponible à http://www.cs.york.ac.uk/euredit/.

Federal Committee on Statistical Methodology (1990). *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18.

Fellegi, I.P., et Holt, D. (1976). A systematic approach to edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

Fraley, C., et Raftery, A. (2002). Model-Based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611-631.

Granquist, L. (1995). Improving the traditional editing process. Dans *Business Survey Methods*, (Éds. B.G. Cox et D.A. Binder).

Granquist, L. (1996). The new view on editing. *Revue Internationale de Statistique*, 65, 3, 381-387.

Granquist, L., et Kovar, J. (1997). Editing of survey data: How much is enough? 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., 415-435.

Hawkins, D.M. (1981). A new test for multivariate normality and homoscedasticity. *Technometrics*, 23, 105-110.

Kovar, J.G., Mac Millian, I.H. et Whitridge, P. (1988). Overview and strategy for the generalized edit and imputation system, (mis à jour février 1991). Statistique Canada, document de travail, direction de la méthodologie, BSMD-88-007E/F.

Latouche, M., et Berthelot, J.M. (1992). Use of a score function to prioritise and limit recontacts in business surveys. *Journal of Official Statistics*, 8, 389-400.

Lawrence, D., et McKenzie, R. (2000). The general application of significance editing. *Journal of Official Statistics*, 16, 243-253.

McLachlan, G.J., et Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker.

McLachlan G.J., et Peel D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 69-78
Statistique Canada, N° 12-001-XIF au catalogue

Utilisation de substituts appariés pour améliorer les imputations dans les bases de données couplées géographiquement

Wai Fung Chiu, Recai M. Yucel, Elaine Zanutto et Alan M. Zaslavsky [1]

Résumé

Lorsqu'on couple géographiquement les enregistrements d'une base de données administratives à des groupes d'îlots de recensement, les caractéristiques locales tirées du recensement peuvent être utilisées comme variables contextuelles susceptibles de compléter utilement les variables qui ne peuvent être observées directement à partir des dossiers administratifs. Les bases de données contiennent souvent des enregistrements dont les renseignements sur l'adresse ne suffisent pas pour le couplage géographique avec des groupes d'îlots de recensement; par conséquent, les variables contextuelles pour ces enregistrements ne sont pas observées. Nous proposons une nouvelle méthode qui consiste à utiliser l'information provenant des « cas appariés » et des modèles de régression multivariée pour créer des imputations multiples pour les variables non observées. Notre méthode donne de meilleurs résultats que d'autres dans les études par simulation au moyen de données du recensement et a été appliquée à un ensemble de données choisi pour étudier les profils de traitement des personnes atteintes d'un cancer du côlon et du rectum.

Mots clés : Non-réponse totale; imputation multiple; variables contextuelles; substituts appariés; dossiers administratifs.

1. Introduction

Afin d'étudier les profils de traitement des personnes atteintes d'un cancer du côlon et du rectum, le revenu et le niveau de scolarité sont des variables qu'il est souhaitable d'utiliser pour construire des modèles statistiques pertinents du point de vue scientifique. Malheureusement, les données individuelles sur ces variables ne peuvent être extraites directement des bases de données des registres du cancer créées d'après les dossiers hospitaliers, qui, comme de nombreuses bases de données administratives, contiennent principalement des renseignements requis à des fins administratives. Par conséquent, on utilise les valeurs moyennes de ces variables pour de petites régions géographiques (groupe d'îlots ou secteur de recensement) comprenant le lieu de résidence du sujet comme variable indépendante afin d'estimer les effets du revenu et du niveau de scolarité. Les analyses fondées sur ce genre de « variables contextuelles » sont fréquentes en épidémiologie et en recherche sur les services de santé (Krieger, Williams et Andmoss 1997), et produisent souvent des résultats semblables, de façon générale, à ceux obtenus en se fondant sur des variables individuelles. Si l'on disposait à la fois de variables individuelles et de variables contextuelles, il serait possible de faire la distinction entre les effets des caractéristiques des individus et du contexte; dans une analyse purement contextuelle, ces effets sont confondus. Néanmoins, l'observation d'associations entre les caractéristiques socioéconomiques contextuelles et la qualité des soins donneraient à penser qu'il existe un problème d'équité, que ces associations reflètent principalement des relations de niveau individuel ou des relations de niveau communautaire.

Dans l'étude du traitement du cancer du côlon et du rectum, on suppose que chaque variable contextuelle pour un enregistrement patient donné est la valeur moyenne de la variable pour le groupe d'îlots (ou secteur) de recensement obtenus par couplage géographique de l'adresse figurant dans l'enregistrement à un groupe d'îlots (ou secteur) de recensement. Un pourcentage faible, mais important, d'enregistrements patient (environ 3,3 %, soit 1 696 enregistrements) ne contiennent pas suffisamment d'information sur l'adresse pour permettre les couplages aux groupes d'îlots de recensement, ce qui rend les variables contextuelles correspondantes inobservables. Nous dirons que de tels enregistrements sont *non géocodables*, tandis que les enregistrements qui peuvent être couplés à des groupes d'îlots de recensement sont *géocodables*. Pour générer des imputations multiples pour les variables contextuelles non observées, nous proposons une stratégie qui consiste à utiliser l'information provenant de plus d'un « cas apparié » pour faciliter la création de modèles paramétriques/non paramétriques d'imputation. plus précisément, l'information provenant des cas appariés rend compte des effets de petite région dans le modèle d'imputation, si bien qu'il n'est pas nécessaire de les modéliser explicitement.

1. Wai Fung Chiu, Department of Statistics, Harvard University, One Oxford Street, Cambridge MA 02138. Courriel : wfchiu@post.harvard.edu; Recai M. Yucel, Department of Biostatistics and Epidemiology, 408 Arnold House, School of Public Health and Health Sciences, University of Massachusetts, 715 North Pleasant Street, Amherst, MA 01003-9304. Courriel : yucel@schoolph.umass.edu; Elaine Zanutto, The Wharton School, University of Pennsylvania, 466 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia PA 19104. Courriel : zanutto@wharton.upenn.edu; Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston MA 02115. Courriel : zaslavsky@hcp.med.harvard.edu.

Rubin et Zanutto (2001) utilisent l'expression « substitut apparié » au lieu de « cas apparié » et proposent un modèle d'imputation paramétrique utilisant un seul substitut apparié par enregistrement. Les résultats des analyses réalisées au moyen de leur modèle ont été comparés à ceux obtenus par d'autres méthodes analytiques dans le cadre d'une grande étude par simulation, mais la méthode n'a pas été appliquée à des données réelles. Nous étendons la méthode de Rubin et Zanutto 1) en permettant l'utilisation d'information provenant de plus d'un cas apparié par enregistrement et 2) en utilisant une loi empirique plutôt que paramétrique pour les résidus.

La présente étude a été motivée par la nécessité de procéder à des imputations multiples pour les variables partiellement observées dans l'étude des profils de traitement chez les personnes atteintes d'un cancer du côlon et du rectum. Ayanian, Zaslavsky, Fuchs, Guadagnoli, Creech, Cress, O'Connor, West, Allen, Wolf et Wright (2003) ont analysé un ensemble de données comprenant des imputations générées par notre méthode, en faisant référence à Rubin et Zanutto (2001) et à une version provisoire du présent article qui a paru dans un recueil de comptes rendus (Chiu, Yucel, Zanutto et Zaslavsky 2001). Le présent article est la première publication complète de notre méthode et le premier rapport jamais publié décrivant une application de la méthode de Rubin et Zanutto à des données réelles.

La présentation de la suite de l'article est la suivante. À la section 2, nous résumons la méthode de Rubin et Zanutto, puis donnons une description générale de la nôtre. À la section 3, nous décrivons dans les grandes lignes l'application de notre méthode à l'étude du traitement du cancer du côlon et du rectum. À la section 4, nous illustrons, par une étude par simulation, les résultats de notre méthode comparativement à trois méthodes de correction pour la non-réponse utilisées fréquemment.

2. Méthode d'imputation

Nous commençons par résumer la méthode de Rubin et Zanutto, puis nous donnons une description générale de notre méthode comprenant une discussion de l'appariement hors échantillon par opposition à l'appariement intra-échantillon, les détails de la modélisation et des tâches d'imputation multiple, ainsi qu'une analyse de l'efficacité en fonction du nombre de cas appariés utilisés.

2.1 Appariement, modélisation et imputation multiple

Rubin et Zanutto (2001) ont proposé une méthode appelée « appariement, modélisation et imputation multiple » (AMI) qui consiste à utiliser des substituts appariés pour produire des imputations multiples pour les non-répondants aux enquêtes par sondage, sans exiger que les substituts soient des remplacements parfaits des non-répondants. Les substituts appariés sont des unités qui répondent à l'enquête choisies de façon à ce qu'elles coïncident avec les non-répondants sur une ou plusieurs « covariables d'appariement », c'est-à-dire des variables pour lesquelles les données sont disponibles avant le sondage et qui sont commodes pour l'appariement, mais pas forcément pour la modélisation. À cause de l'appariement, les non-répondants et leurs substituts peuvent posséder des valeurs en commun pour les « covariables de terrain », c'est-à-dire des variables qui ne sont observées qu'implicitement et ne sont, par conséquent, pas disponibles pour l'analyse des données. Les « covariables de modélisation » sont des variables qui peuvent être incluses dans les modèles statistiques pour tenir compte des différences observées entre les non-répondants et leurs substituts, mais qui pourraient ne pas être disponibles ou utilisées pour l'appariement. L'essence de la méthode AMI est que l'on utilise à la fois les covariables d'appariement et de modélisation, dans le contexte de l'imputation multiple appropriée (Little et Rubin 1987, pages 258–259 et les références qui y sont mentionnées).

Considérons un exemple simple où les données sont disponibles pour les covariables d'âge et d'adresse pour toutes les unités d'une population avant l'échantillonnage. L'obtention de substituts coïncidant avec les non-répondants en ce qui concerne l'âge ainsi que l'adresse pourrait être difficile. Notre solution consiste à fonder l'appariement uniquement sur l'adresse (par exemple, choisir un voisin comme substitut) et à corriger les différences systématiques d'âge entre les non-répondants et les substituts appariés par modélisation statistique. Si des ménages voisins étaient choisis comme substituts appariés des ménages non-répondants, il se pourrait que les substituts et les non-répondants vivent dans le même contexte socioéconomique (par exemple, taux de criminalité, accès aux moyens de transport en commun, *etc.*) même si l'on n'a pas enregistré ces caractéristiques. Dans le présent exemple, l'adresse est une covariable d'appariement, l'âge est une covariable de modélisation et les caractéristiques socioéconomiques contextuelles sont les covariables de terrain.

Brièvement, la méthode AMI consiste à i) choisir des substituts appariés pour les non-répondants et certains répondants d'après des covariables d'appariement, ii) utiliser des covariables de modélisation pour ajuster un modèle d'estimation des différences systématiques de réponse entre les paires répondant-substitut, iii) procéder à l'imputation multiple des valeurs non observées à l'aide du modèle obtenu en (iii) sous l'hypothèse que la même relation est vérifiée entre les paires non-répondant-substitut et iv) éliminer tous les substituts appariés après l'imputation.

2.2 Appariement hors échantillon et intra-échantillon

Les cas appariés peuvent être obtenus à partir de données hors échantillon ou de données intra-échantillon. Dans l'approche de Rubin et Zanutto, les substituts appariés sont obtenus à partir de données hors échantillon, *après* avoir décelé l'absence de données. La description de ces auteurs met l'accent sur le fait que les substituts appariés doivent être éliminés après l'imputation, puisque le fait d'inclure ces cas supplémentaires dans l'inférence modifierait le plan de sondage par ajout de cas supplémentaires dans les « îlots » qui contiennent des données non observées. Les cas appariés sont considérés comme provenant de données intra-échantillon s'il sont obtenus à partir de la base de données disponible *avant* de faire l'imputation ou même avant de déterminer quels enregistrements de la base de données contiennent des variables non observées. En ce qui concerne les objectifs globaux d'inférence, au lieu d'être des cas supplémentaires, ces cas appariés font partie de la série originale de données et, par conséquent, seront inclus dans les analyses scientifiques.

En supposant que l'appariement est intra-échantillon, nous traitons les enregistrements non géocodables comme des non-répondants et les enregistrement géocodables, comme des répondants. Pour chaque enregistrement non géocodable, nous choisissons aléatoirement un nombre donné de cas appariés à partir d'un groupe d'enregistrements géocodables dans la même petite région géographique (par exemple, code postal, c'est-à-dire un code de livraison postale qui, aux États-Unis, représente habituellement une région desservie par un bureau de poste principal unique). De façon semblable, nous choisissons le même nombre de cas appariés pour chaque enregistrement géocodable échantillonné aléatoirement [voir Rubin et Zanutto (2001) pour des recommandations sur la taille d'un échantillon de ce genre comparativement au nombre total d'enregistrements non géocodables dans un ensemble de données particulier]. Si un nombre de cas appariés plus grand que celui disponible dans la même petite région était nécessaire, le groupe de sélection serait étendu aux régions géographiques « les plus proches » jusqu'à l'obtention du nombre requis.

Dans l'étude du traitement du cancer du côlon et du rectum, tous les cas appariés provenaient de la même base de données sur le cancer. En général, les cas appariés ne doivent pas nécessairement être tirés de la même population que celle dont proviennent les non-répondants et les répondants. Par exemple, pour les enregistrements de cas de cancer du côlon et du rectum, on peut obtenir les cas appariés à partir d'une population générale de cancéreux, puis ajuster un modèle pour tenir compte des différences systématiques. Notons que, si les cas appariés proviennent d'une population fort semblable, il est possible de construire des modèles plus robustes contenant un plus grand nombre de covariables. Dans notre exemple, puisque nous utilisons d'autres patients atteints de la même forme de cancer, les relations en ce qui concerne les variables de processus thérapeutique et les variables de résultat sont vraisemblablement cohérentes.

2.3 Modélisation et imputation multiple

Nous donnons ici un exemple simple de notre méthode afin de communiquer l'idée de base; en pratique, il faut souvent utiliser des modèles plus complexes. Supposons qu'est vérifiée dans la population la relation suivante

$$y_{ik} = \boldsymbol{x}_{ik}^T \boldsymbol{\beta} + \delta_i + \varepsilon_{ik}, \qquad (1)$$

où l'indice i désigne la petite région géographique et l'indice k, l'unité dans la région, et y_{ik} et $\boldsymbol{x}_{ik}$ sont, respectivement, la réponse et les caractéristiques de la k^{e} unité dans la région géographique i. Ce modèle comprend une prévision par régression $\boldsymbol{x}_{ik}^T \boldsymbol{\beta}$, un effet de petite région δ_i, et un résidus particulier à l'unité ε_{ik}. Nous supposons que ε_{ik} suit une loi F_ε de moyenne nulle et de variance σ^2. Notons que ce développement se généralise directement à une réponse $\boldsymbol{y}_{ik}$ multivariée.

Nous étendons la méthode de Rubin et Zanutto de façon à permettre plus d'un appariement dans la même petite région, parce que l'obtention de plusieurs appariements dans les petites régions est possible (souvent commode et peu coûteuse) dans le cas des données de recensement ou de grands ensembles de données administratives. L'hypothèse d'un seul appariement émise par Rubin et Zanutto est appropriée dans le cas de la collecte de données d'enquête nécessitant du travail supplémentaire sur le terrain pour chaque appariement.

Nous estimons les coefficients de régression de l'équation (1) à l'aide d'une série d'observations associées à deux enregistrements ou plus par petite région en vue d'ajuster le modèle de régression dans lequel les δ_i sont traités comme des effets fixes. S'il n'existe que deux cas par région, on peut estimer $\boldsymbol{\beta}$ à partir de la régression dans la région

$$(y_{i1} - y_{i2}) = (\boldsymbol{x}_{i1}^T - \boldsymbol{x}_{i2}^T)\boldsymbol{\beta} + (\varepsilon_{i1} - \varepsilon_{i2}), \qquad (2)$$

où l'effet de petite région s'élimine. Les résidus de cette régression suivent une loi symétrique de variance $2\sigma^2$.

En supposant pour le moment que nous avons procédé à un tirage à partir de la loi a posteriori de $\boldsymbol{\beta}$, nous exécutons le reste de l'analyse sachant ce tirage. Supposons maintenant que nous voulions faire une imputation pour une nouvelle unité (dont l'indice est $k = 0$) dans la région i, et que nous ayons obtenu $K_i \geq 1$ cas appariés pour cette unité.

Dénotons les résultats pour ces cas appariés par le vecteur $\boldsymbol{y}_i = (y_{i1}, ..., y_{iK_i})^T$ et les caractéristiques correspondantes par la matrice $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iK_i})^T$. Avec une loi a priori uniforme pour δ_i, la loi a posteriori de $\delta_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\beta}$ est de moyenne

$$\bar{y}_i - \bar{\boldsymbol{x}}_i^T \boldsymbol{\beta} \tag{3}$$

et de variance σ^2 / K_i, où $\bar{y}_i = \sum_{k=1}^{K_i} y_{ik} / K_i$ et $\bar{\boldsymbol{x}}_i = \sum_{k=1}^{K_i} \boldsymbol{x}_i / K_i$. Donc, la loi prédictive pour $y_{i0} \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{x}_{i0}, \boldsymbol{\beta}$ est de moyenne

$$\bar{y}_i + (\boldsymbol{x}_{i0}^T - \bar{\boldsymbol{x}}_i^T)\boldsymbol{\beta} \tag{4}$$

et de variance $(1 + 1/K_i)\sigma^2$, ce qui est égal à la somme de la variance prédictive sous le modèle, sachant tous les paramètres et la variance a posteriori de δ_i. Ces énoncés supposent que la moyenne des résidus est une statistique suffisante pour δ_i. Cette hypothèse est vraie pour la loi normale (ou les observations naturelles de toute loi de la famille exponentielle); nous supposons qu'elle est, au moins, approximativement vraie pour F_ε, afin que nous puissions fonder nos inférences sur cette moyenne. Notons que l'utilisation d'une loi à priori uniforme donne lieu à des tirages surdispersés comparativement à ceux que l'on obtiendrait avec une loi a priori appropriée à partir d'un modèle hiérarchique, mais qu'il s'agit d'une approche beaucoup plus simple (particulièrement dans les analyses avec des résultats multivariés).

Nous pouvons générer une imputation pour y_{i0} en tirant d'abord σ^2 à partir de sa loi a posteriori, en tirant ensuite $\boldsymbol{\beta}$ sachant le tirage de σ^2, puis en calculant la moyenne prédictive donnée par l'équation (4) à partir du tirage de $\boldsymbol{\beta}$ et en ajoutant enfin un résidu de variance $(1 + 1/K_i)\sigma^2$ à la moyenne prédictive. Dans le cas des sondage simple avec $\boldsymbol{\beta}$ estimé par l'équation (2), la loi a posteriori de $\boldsymbol{\beta}$ (sachant σ^2 et les données) sous une loi a priori uniforme est approximativement $\mathrm{N}(\hat{\boldsymbol{\beta}}, (\boldsymbol{X}^T \boldsymbol{X})^{-1}\sigma^2)$, où la i^{e} ligne de $\boldsymbol{X}$ est $(\boldsymbol{x}_{i1}^T - \boldsymbol{x}_{i2}^T)$. Dans le cas de plans de sondage plus complexes, la loi a posteriori de $\boldsymbol{\beta}$ peut être approximée à l'aide de l'estimation ponctuelle et de la variance d'échantillonnage calculée sous le plan de sondage associé.

Nous pouvons obtenir le résidu par modélisation *ou* par échantillonnage. La modélisation comprend l'estimation de σ^2 en utilisant la variance résiduelle de l'équation (1) et en tirant le résidu sous une loi normale univariée [voir Rubin et Zanutto (2001) pour le cas particulier où un seul cas apparié a été obtenu pour chaque enregistrement] ou sous une autre loi paramétrique. Nous donnons à cette approche le nom de **AMI paramétrique** (AMIP). Une autre option consiste à échantillonner aléatoirement un résidu de régression à partir de *toute* région *j* dont les résidus pourraient être considérés comme permutables avec ceux provenant de la région *i* (Rubin 1987, pages 166–168). Consulter aussi Lessler et Kalsbeek (1992, section 8.2.2.4), Kalton et Kasprzyk (1986), ainsi que Kalton (1983). Puisque la variance d'un tel résidu est $[(K_j - 1)/K_j]\sigma^2$, nous multiplions le résidu échantillonné aléatoirement par $\sqrt{[(K_i + 1)/K_i][K_j/(K_j - 1)]}$ pour obtenir la variance prédictive correcte. Nous appelons cette approche **AMI non paramétrique** (AMINp).

Brièvement, notre méthode comprend trois étapes fondamentales :

1. tirer des cas appariés pour les enregistrements non géocodables et pour certains enregistrements géocodables échantillonnés aléatoirement;
2. utiliser les enregistrements géocodables échantillonnés et leurs cas appariés pour ajuster l'équation (1), où les δ_i sont traités comme des effets fixes et sauvegarder les résidus;
3. répéter m (habituellement de 5 à 10) fois les étapes suivantes :
   a) tirer σ^2 à partir de sa loi a posteriori, puis $\boldsymbol{\beta}$ sachant le tirage de σ^2 ;
   b) pour chaque enregistrement non géocodable, traiter la somme du vecteur des moyennes prédictives obtenu à partir de l'équation (4) et d'un vecteur de résidu tiré en utilisant soit AMIP soit AMINp en tant que réalisation du vecteur non observé de variables contextuelles.

2.4 Efficacité

L'efficacité d'une imputation dépend du nombre de cas appariés utilisés. Soit V_K la variance prédictive d'un modèle d'imputation, où K cas appariés par enregistrement sont utilisés. Pour le modèle de la section 2.3, $V_K = (1 + 1/K)\sigma^2$. Définissons l'efficacité comme étant

$$E_K = \frac{V_\infty}{V_k} = \frac{\sigma^2}{(1 + 1/K)\sigma^2} = \frac{K}{K + 1}, \tag{5}$$

pour tout nombre entier positif K. L'efficacité augmente parallèlement au nombre de cas appariés par enregistrement; par exemple, $E_2 \approx 0{,}67, E_4 = 0{,}8, E_{10} \approx 0{,}91$ et $E_{20} \approx 0{,}95$.

En théorie, il peut exister autant de cas appariés par enregistrement que le permettent les ressources disponibles. En pratique, le nombre de cas appariés utilisés dépend souvent du coût de l'obtention de ces cas et de celui du traitement informatique requis pour ajuster le modèle. Dans le cas de notre méthode, le coût du traitement informatique nécessaire pour chaque cas apparié supplémentaire par enregistrement est négligeable. Dans l'étude du traitement du cancer du côlon et du rectum, l'obtention des cas appariés était gratuite, mais il était essentiel de pouvoir procéder à l'imputation en s'appuyant sur un nombre limité de cas appariés, parce que des contraintes de protection des

renseignements personnels empêchaient les chercheurs d'utiliser l'ensemble de données complet pour la modélisation avec les codes postaux annexés (même sous forme cryptée). À titre d'illustration, nous utiliserons deux cas appariés par enregistrement dans les analyses qui suivent.

3. Application : Étude sur le cancer du côlon et du rectum

La base de données sur le cancer du côlon et du rectum contient, en tout, 50 740 enregistrements patient, dont environ 3,3 % ne sont pas géocodables. Parmi ceux-ci, environ la moitié contiennent une adresse de case postale (souvent dans une région rurale) et les autres une adresse tapée incorrectement ou une adresse appartenant à une région nouvellement développée qui ne figure pas encore dans la base de données sur les adresses. Dans le cadre d'une étude des prédicteurs de l'administration d'une chimiothérapie aux patients atteints d'un cancer du côlon et du rectum, les chercheurs ont estimé que les trois moyennes de groupe d'îlots de recensement qui suivent seraient des variables contextuelles utiles :

Y_1 = revenu médian du ménage;

Y_2 = pourcentage ne possédant pas de diplôme d'études secondaires;

Y_3 = pourcentage sous le seuil de pauvreté.

Ces variables ont été observées dans les enregistrements géocodables, mais non observées dans les enregistrements non géocodables. La tâche consistait à produire des imputations multiples pour les variables de recensement non observées à l'aide des méthodes décrites à la section 2.

Chacune des moyennes de groupe d'îlots était publiée dans les données du recensement pour six groupes ethniques et les analyses scientifiques ont porté uniquement sur l'ensemble des moyennes de groupe d'îlots correspondant à la race/ethnicité de chaque patient. Pour les imputations utilisées dans Ayanian et coll. (2003), nous avons par conséquent ajusté six modèles distincts pour imputer les $18 (6 \times 3)$ valeurs pour chaque cas non géocodables, puis nous avons sélectionné les trois variables pertinentes pour chaque patient; les lois conjointes pour divers groupes raciaux/ethniques n'étaient pas importantes, parce que chaque imputation ne comportait de valeurs que pour un seul groupe. Une autre solution aurait été d'utiliser la race comme variable d'appariement, mais cela nous aurait obligé à rechercher des appariements à une distance géographique nettement plus grande, ce qui aurait affaibli la valeur prédictive de l'appariement géographique.

Aux fins de l'exposé, nous supposons donc que nous disposons uniquement de la moyenne de groupe d'îlots correspondant à la race pour chaque répondant, et que nous n'avons pas accès aux moyennes correspondant aux cinq autres races qui sont disponibles simultanément dans les données du recensement. Cette situation est plus typique des données qui seraient recueillies directement auprès du répondant, où la variable de race proprement dite (en tant que variable de modélisation) est relativement prédictive, parce que les données sur le revenu de personnes de races différentes reflètent les différences de revenu associées à la race.

3.1 Appariement et l'ensemble de données

L'adresse de plus de 90 % des enregistrements non géocodables contient le code postal. Par conséquent, nous avons choisi ce dernier comme covariable d'appariement. Un diagnostic simple de son utilité figure à la section 3.2. La série numérique de codes postaux ne correspond pas toujours aux relations de distance entre les quartiers. Par exemple, Cambridge, Massachusetts, possède un bureau de poste 02138 qui utilise aussi le code postal 02238 pour les boîtes aux lettres et à Boston, située tout près de là, il existe un code postal 02215 qui a été pris à la région 02115. Au lieu d'utiliser la série numérique de codes postaux, nous avons calculé les distances entre les codes postaux d'après les latitude et longitude du bureau de poste principal correspondant, sous l'hypothèse que deux codes postaux étaient les plus proches l'un de l'autre si leurs bureaux de poste principaux étaient les plus proches l'un de l'autre.

La base de données sur le cancer du côlon et du rectum contient 1 696 enregistrements non géocodables. Nous avons sélectionné le même nombre $(n^* = 1696)$ d'enregistrements géocodables aléatoirement à partir de la même base de données. Pour chacun de ces 3 392 enregistrements, nous avons sélectionné aléatoirement deux cas géocodables appariés à partir du code postal de l'enregistrement en question ou (au besoin) de codes postaux voisins. Nous avons obtenu ainsi un ensemble de données contenant $3\,392 \times 3 = 10\,176$ enregistrements. Notons qu'il était commode de choisir n^*, parce que les données étaient gratuites. En général, le choix de n^* pourrait avoir une incidence sur le coût total ainsi que sur la précision des estimations. Tant les enregistrements géocodables que les cas appariés sélectionnés aléatoirement correspondaient à des données intra-échantillon et ont donc été retenus dans les analyses menées par Ayanian et coll. (2003). Nous avons demandé au Registre du cancer de nous fournir uniquement ces cas, car, pour des raisons de confidentialité, nous ne pouvions procéder nous-mêmes à l'appariement aux données que nous possédions (pour les mêmes cas).

Les covariables de modélisation que nous avons utilisées dans le modèle d'imputation étaient les huit variables d'enregistrement administratif, à savoir l'âge, le sexe, la

race, l'état matrimonial, le stade du cancer, le traitement par chimiothérapie, le type de cancer et le traitement par radiothérapie, et la catégorie d'agrément accordée par l'American College of Surgeons à l'hôpital prodiguant le traitement en 1999 (ACOS99). Ces variables sont observées pour les 10 176 enregistrements inclus dans le modèle d'imputation. (Certaines de ces variables sont des prédicteurs et certaines sont des résultats dans les modèles scientifiques des analyses principales, mais la distinction est sans importance pour l'imputation.) Les valeurs moyennes au recensement Y_1, Y_2 et Y_3 sont observées dans les enregistrements géocodables, mais non dans les enregistrements non géocodables. Ces variables ont été traitées comme des variables de résultat du modèle d'imputation à la section 2.3. La structure des données est représentée au tableau 1.

Tableau 1
Structure des données utilisées pour l'imputation dans l'étude sur le cancer du côlon et du rectum

Données*	Huit covariables de modélisation				Variables de recensement		
	Âge	Sexe	…	ACOS99	Y_1	Y_2	Y_3
Non géocodable	√	√	…	√	?	?	?
Premier appariement	√	√	…	√	√	√	√
Deuxième appariement	√	√	…	√	√	√	√
Géocodable	√	√	…	√	√	√	√
Premier appariement	√	√	…	√	√	√	√
Deuxième appariement	√	√	…	√	√	√	√

* Il existait 1 696 enregistrements pour chacun des six type de données.
√ = observée ? = non observée

Avant d'ajuster le modèle, nous avons transformé les variables de résultat en pourcentage y_2 et y_3 à l'aide de la fonction logit mise à l'échelle :

$$\log\left(\frac{(y-a)/(b-a)}{1-(y-a)/(b-a)}\right), \qquad (6)$$

avec $a=-0{,}5$ et $b=100{,}5$ de sorte qu'après les imputations, la transformation inverse avec arrondissement au nombre entier le plus proche produise des valeurs imputées comprises entre 0 et 100 inclusivement (Schafer 1999). De même, nous avons appliqué une transformation logarithmique à la variable de revenu y_1 de sorte que les valeurs imputées de revenu ne soient pas négatives. Notons que les lois suivies par les variables transformées sont plus proches de la loi normale qu'elles ne le sont sur l'échelle originale (Schafer 1997). Pour simplifier la notation, nous redéfinissons y_1, y_2 et y_3 comme étant les versions transformées.

3.2 Diagnostics préliminaires

Un test diagnostique simple de l'utilité des covariables d'appariement consiste à comparer les valeurs de R^2 corrigé pour les modèles de régression prédisant les trois variables de recensement contenant uniquement les covariables de modélisation, les modèles contenant uniquement les covariables d'appariement et les modèles contenant les deux types de variables. Dans la présente application, le code postal était la seule covariable d'appariement. Il existait 1 133 codes postaux distincts (donc 1 132 variables nominales) dans les 8 480 enregistrements observés entièrement (les enregistrements géocodables et tous les premiers et deuxièmes appariements). Le tableau 2 donne le R^2 corrigé pour les modèles contenant uniquement les huit covariables de modélisation, les modèles contenant uniquement le code postal et les modèles contenant les covariables de modélisation ainsi que le code postal. La valeur de R^2 corrigé est plus élevée pour les modèles contenant à la fois les covariables de modélisation et le code postal que pour les modèles correspondants ne contenant qu'un des deux types de covariable. Notre méthode d'imputation utilise l'information provenant des covariables d'appariement et de modélisation et, donc, devrait en principe donner de meilleurs résultats que les méthodes utilisant uniquement les covariables d'appariement ou de modélisation (comme le montre l'étude par simulation décrite à la section 4). Bien que la contribution des variables de modélisation à R^2 soit assez modeste, il est important de les inclure dans le modèle afin d'éliminer les biais systématiques et de représenter correctement les relations qui pourraient être importantes dans les modèles scientifiques.

Tableau 2
R^2 corrigé pour divers modèles de régression

	Covariables de modélisation uniquement	Covariable d'appariement uniquement (code postal)	Covariables de modélisation et d'appariement
Revenu médian du ménage (INC)	0,091	0,453	0,496
Pourcentage sans diplôme d'études secondaires (EDU)	0,115	0,452	0,503
Pourcentage sous le seuil de pauvreté (POV)	0,047	0,327	0,343
Nombre de degrés de liberté du modèle[a]	26[b]	1 133	1 158
Taille de l'échantillon	8 480	8 480	8 480
Nombre de degrés de liberté du résidu	8 454	7 347	7 322

(a) Avec l'ordonnée à l'origine.
(b) Les covariables de modélisation sont l'âge, le sexe (2 niveaux), la race (6 niveaux), l'état matrimonial (6 niveaux), le stade du cancer (6 niveaux), la chimiothérapie (2 niveaux), le type de cancer et la radiothérapie (3 niveaux), et la catégorie d'agrément octroyée par l'American College of Surgeons en 1999 à l'hôpital prodiguant le traitement (6 niveaux).

Pour déterminer s'il fallait utiliser un modèle multivarié, nous avons ajusté un modèle de régression à résultats

multivariés avec les covariables de modélisation et le code postal. Les corrélations estimées entre les résidus étaient : $r_{12} \approx -0{,}194$, $r_{13} \approx -0{,}297$ et $r_{23} \approx 0{,}357$, où la « variable 1 » est le revenu médian du ménage, la « variable 2 » est le pourcentage sans diplôme d'études secondaires et la « variable 3 » est le pourcentage sous le seuil de pauvreté. Ces estimations différaient significativement de zéro, indiquant qu'il fallait utiliser les versions multivariées des méthodes décrites à la section 2.3 pour produire les imputations.

3.3 Résultats de l'imputation multiple et comparaisons

Ayanian et coll. (2003) ont utilisé l'imputation par la méthode AMINp dans l'étude des prédicteurs de l'administration d'une chimiothérapie aux patients atteints d'un cancer du côlon et du rectum. Leur modèle comprenait trois variables indicatrices pour des fourchettes de revenu contextuel, ainsi que 21 autres variables représentant les caractéristiques du patient et de l'hôpital. L'analyse de l'imputation multiple montre que l'information perdue à cause de données manquantes était systématiquement inférieure à 0,1 %, proportion nettement plus faible que la fraction d'enregistrements non géocodables (3,3 %). Comme il faillait s'y attendre, les fractions les plus importantes d'information manquante ont été relevées pour les variables de revenu. Les résultats scientifiques exposés dans Ayanian et coll. (2003) n'auraient pas variés énormément si les cas pour lesquels les données étaient incomplètes avaient été éliminés. Néanmoins, dans ce genre d'étude, chaque cas est précieux et coûteux, et sauvegarder les 3,3 % pour lesquels des données manquaient représentait une contribution à l'étude.

Aux fins de comparaison, les variances des paramètres dans l'analyse portant sur les cas complets étaient, en moyenne, supérieures de 4,0 % à celles observées dans l'analyse sous imputation multiple. Cet écart en pourcentage est proche de la fraction de cas incomplets supprimés de l'analyse. Après l'introduction des imputations générées par notre méthode dans l'analyse scientifique, la précision de l'estimation de l'effet de « région rurale » a augmenté considérablement (l'utilisation des cas pour lesquels les données étaient complètes uniquement a fait augmenter la variance de 41,6 %), à cause de la concentration des enregistrements non géocodables dans les régions rurales (21,6 % d'enregistrements ruraux, mais seulement 3,1 % d'enregistrements non ruraux sont non géocodables).

4. Une étude par simulation

Cette étude par simulation a pour but de comparer la performance de notre nouvelle méthode à celle de trois autres méthodes d'ajustement pour la non-réponse utilisées fréquemment. La population visée par l'étude est celle des 1 696 triplets entièrement observés, c'est-à-dire les 1 696 enregistrements géocodables et les premiers et deuxièmes appariements correspondants (une rangée de chacun des trois derniers blocs horizontaux au tableau 1), soit 5 088 observations. Par souci de simplicité, nous avons supposé que les triplets provenaient de codes postaux (grappes) distincts, donc, que $i = 1, 2, ..., I = 1\,696$. Chaque grappe i contenait trois unités $(u = 1, 2, 3)$, et l'enregistrement de chaque unité était constitué de $\boldsymbol{x}_{iu}$ (les covariables) et de $\boldsymbol{y}_{iu}$ (les variables de recensement).

4.1 Données simulées et mécanisme de réponse

En supposant un échantillonnage en grappes avec taille d'échantillon de 800, nous avons tiré des échantillons aléatoires contenant 800 grappes. Pour chaque échantillon aléatoire, nous avons sélectionné aléatoirement environ la moitié des 800 grappes de façon à ce qu'elles contiennent un enregistrement non géocodable dans lequel les variables de recensement étaient non observées, la probabilité que les données manquent étant fonction de la race de la personne et du revenu moyen de la grappe (code postal). Nous avons simulé l'absence des données sous un modèle logit multinomial où les résultats étaient : aucune variable non observée $(w_{i0} = 1)$, $\boldsymbol{y}_{i1}$ non observée $(w_{i1} = 1)$, $\boldsymbol{y}_{i2}$ non observée $(w_{i2} = 1)$ et $\boldsymbol{y}_{i3}$ non observée $(w_{i3} = 1)$. Plus précisément, pour chaque $i = 1, 2, ..., I$, soit $z_{i0} = 0$ et

$$z_{iu} = a + b \times I(\text{unité } iu \text{ est race blanche}) + c \times (\text{revenu moyen dans le code postal } i) \quad (7)$$

où $u = 1, 2, 3$. Alors,

$$\Pr(w_{iu} = 1) = \exp(z_{iu}) \Big/ \sum_{u=0}^{3} \exp(z_{iu}) \quad \text{pour } u = 0, 1, 2, 3. \quad (8)$$

Les résultats de cette étude par simulation ont été fondés sur des ensembles de données générés par le mécanisme susmentionné avec $a = -1$, $b = 11$ et $c = 0{,}0003$, afin de rendre non géocodables environ 17 % des unités dans un échantillon aléatoire, avec probabilité de géocodage associée positivement à la race blanche et à un revenu plus élevé au niveau de l'îlot. La tâche consistait à utiliser l'échantillon aléatoire pour estimer $\bar{y}$, c'est-à-dire les valeurs moyennes de population (1 696 grappes).

Nous avons établi les conditions de simulation décrites plus haut pour produire un test rigoureux de la méthode et des autres options en exagérant l'effet des données non observées et en faisant en sorte que l'absence de données soit fortement liée aux caractéristiques de l'individu et de la région. Nous n'essayions pas de simuler les conditions

exactes de l'application décrite à la section 3, mais plutôt d'utiliser une population artificielle caractérisée par des lois semblables à celles de la population réelle pour illustrer le fonctionnement de notre méthode et de ses concurrentes.

4.2 Méthodes d'inférence et mesures de performance

Les résultats préliminaires indiquaient que la performance des méthodes AMIP et AMINp était semblable; cependant, la méthode AMINp est plus simple (surtout dans les analyses avec résultats multivariés), parce qu'elle ne nécessite pas la modélisation paramétrique explicite de la variance résiduelle. Nos simulations avaient pour but de comparer la performance de la méthode AMINp (en utilisant deux cas appariés par enregistrement) à trois autres méthodes d'ajustement pour la non-réponse utilisées fréquemment.

1. **Méthode des cas complets (MCC)**
   Les moyennes de population sont estimées d'après l'ensemble des unités géocodables d'un échantillon aléatoire.
2. **Imputation simple par substitut (ISS)**
   Il s'agit de l'utilisation habituelle de substituts. Les variables de recensement non observées pour chaque unité non géocodable sont remplacées par les valeurs des variables de recensement d'une unité sélectionnée aléatoirement dans la même grappe. L'échantillon résultant est traité comme s'il ne contenait aucune unité non géocodable; les 800 grappes comprises dans un tel échantillon sont toutes utilisées pour estimer les moyennes de population.
3. **Imputation multiple normale multivariée (IMNM)**
   Cette méthode consiste à utiliser uniquement une unité tirée aléatoirement de chacune des grappes entièrement observées dans un échantillon aléatoire pour ajuster la régression linéaire normale multivariée

   $$\boldsymbol{y}_i^T \sim \mathrm{N}(\boldsymbol{\beta}_0^T + \boldsymbol{x}_i^T\, \mathbf{B},\ \boldsymbol{\Sigma}),$$

   avec une loi a priori non informative sur les paramètres. Le modèle est alors utilisé pour créer m ensembles d'imputations multiples pour les variables de recensement non observées en utilisant une généralisation multivariée directe de l'algorithme donné par Rubin (1987, page 167).

Notons que la MCC ne comprend *ni* des covariables d'appariement *ni* des covariables de modélisation, que l'ISS comprend *uniquement la covariable d'appariement* (code postal), que la méthode AMIN utilise *uniquement les covariables de modélisation* et que la méthode AMINp utilise *à la fois* la covariable d'appariement et les covariables de modélisation.

Les données MCC et ISS sont analysées par la méthode des données complètes habituelle qui consiste à estimer la moyenne de population à partir des données à l'aide de l'estimateur approprié pour l'échantillonnage en grappes à partir d'une population finie, y compris la correction pour population finie (Cochran 1977, chapitres 9–10). Les méthodes AMIN et AMINp produisent toutes deux m ensembles de données complètes, qui sont analysés chacun par la même méthode des données complètes utilisée pour les données MCC et ISS; les m ensembles d'estimations ponctuelles et d'estimations de la variance sont alors combinés en appliquant la règle de combinaison de l'imputation multiple (Rubin 1987; Schafer 1997, pages 108–110).

Pour chaque simulation $t \in \{1, 2, ..., T\}$, nous dénotons les estimations ponctuelles obtenues à partir des quatre modèles par $\bar{y}_{\mathrm{CC}}(t)$, $\bar{y}_{\mathrm{SS}}(t)$, $\bar{y}_{\mathrm{MN}}(t)$ et $\bar{y}_{\mathrm{Np}}(t)$, et les moyennes de ces quantités sur l'ensemble des simulations par $\bar{y}_{\mathrm{CC}}$, $\bar{y}_{\mathrm{SS}}$, $\bar{y}_{\mathrm{MN}}$ et $\bar{y}_{\mathrm{Np}}$. L'évaluation de la performance des quatre méthodes de correction pour la non-réponse sera fondée sur trois mesures :

1. **Réduction en pourcentage du biais moyen d'un estimateur relativement au biais moyen de l'estimateur MCC.** Représentons le biais moyen d'un estimateur par $\bar{\boldsymbol{b}}_{\mathrm{E}}$. Alors

   $$\bar{\boldsymbol{b}}_{\mathrm{E}} = \bar{y}_E - \bar{y},$$

   où $E \in \{\mathrm{CC}, \mathrm{SS}, \mathrm{MN}, \mathrm{Np}\}$. Nous définissons la réduction en pourcentage du biais moyen d'un estimateur comparativement au biais moyen de l'estimateur MCC comme étant

   $$R(\bar{b}_{\mathrm{E}},\ \bar{b}_{\mathrm{CC}}) = \frac{|\,\bar{b}_{\mathrm{CC}}\,| - |\,\bar{b}_{\mathrm{E}}\,|}{|\,\bar{b}_{\mathrm{CC}}\,|},$$

   où $\bar{b}_{\mathrm{E}}$ est un élément de $\bar{\boldsymbol{b}}_{\mathrm{E}}$ et $\bar{b}_{\mathrm{CC}}$ est l'élément correspondant dans $\bar{\boldsymbol{b}}_{\mathrm{CC}}$. Par définition, $R(\bar{b}_{\mathrm{CC}},\ \bar{b}_{\mathrm{CC}})$ est nul.
2. **Couverture estimée des intervalles de confiance à 95 % nominaux pour $\bar{y}$.** Les intervalles produits par les estimations MCC ou ISS ont été construits sous les lois t appropriées. Pour les intervalles associés aux estimations AMIN ou AMINp, nous avons suivi la procédure décrite dans Schafer (1997, pages 109–110) et remplacé le nombre de degrés de liberté v par la version mise à jour de Barnard et Rubin (1999).
3. **Fraction estimée d'information manquante au sujet de $\bar{y}$.** Pour la méthode AMIN ainsi que pour la méthode AMINp, nous avons calculé $\hat{\lambda}$, une estimation de la fraction d'information manquante au sujet de $\bar{y}$ (voir Barnard et Rubin (1999) pour l'expression la plus récente).

4.3 Résultats

Nous avons exécuté la procédure de simulation 2 000 fois et utilisé $m = 10$ pour les méthodes AMIN et AMINp. Les valeurs moyennes des variables de recensement dans la population étaient $\bar{\boldsymbol{y}} = (40\,642,\ 21{,}65,\ 9{,}55)^T$. Le biais moyen de l'estimateur MCC était $\bar{\boldsymbol{b}}_{\mathrm{CCM}} = (-5\,405, -3{,}97, -1{,}79)^T$. Les autres résultats sont résumés au tableau 3. La méthode AMINp a produit d'importantes réductions en pourcentage du biais moyen relatif (de 95,0 % à 99,5 %). La méthode ISS a réduit plus fortement les biais que la méthode AMIN, parce que la covariable d'appariement (code postal) était nettement plus informative que l'ensemble des covariables de modélisation (section 3.2). Puisque le mécanisme de réponse était *non ignorable* (les probabilités de réponse dépendaient partiellement du revenu), les mauvais résultats de la méthode AMIN, qui ne s'appuyait pas sur l'utilisation de l'information géographique pour prédire le revenu, étaient prévisibles. Notons que la méthode AMIN est biaisée et que le biais est suffisamment important pour que, avec la taille d'échantillon considérée dans le présent article, les intervalles de confiance ne couvrent jamais les valeurs hypothétiques de population.

Dans le cas des méthodes AMIN et AMINp, le pourcentage d'information manquante était nettement plus faible que le pourcentage moyen de données non observées. Le pourcentage d'information manquante était plus faible pour la méthode AMINp que pour la méthode AMIN. Seule la méthode AMINp a produit des intervalles bien étalonnés avec couverture correcte. Bref, la méthode AMINp combine les meilleures caractéristiques des deux autres méthodes, à savoir une couverture proche de la couverture nominale et moins d'information manquante.

Tableau 3
Résultats des simulations[(a)] : réduction du biais, couverture et fraction d'information manquante

Mesure	Moyenne	Méthode AMINp	Méthode AMIN	Méthode ISS
Réduction du biais en pourcentage $100R(\bar{b}_E,\ \bar{b}_{\mathrm{CCM}})$[(b)]	INC	99,5	44,6	95,2
	EDU	95,0	40,6	83,7
	POV	96,8	32,6	80,3
Couverture estimée des IC à 95 %[(c)]	INC	95,1	0,00	89,8
	EDU	94,8	0,00	65,7
	POV	95,2	0,00	66,0
$100\times$ fraction estimée d'information $\hat{\lambda}$[(d)]	INC	1,00	9,92	
	EDU	0,05	0,07	
	POV	0,07	0,08	

(a) Fondés sur 2 000 répétitions et $m = 10$.

(b) Par définition, $100R(\bar{b}_{\mathrm{CCM}},\ \bar{b}_{\mathrm{CCM}}) = 0$.

(c) Le résultats pour les estimations MCC étaient tous nuls.

(d) Le pourcentage moyen de données non observées était environ 17%.

5. Conclusion

Les présents travaux prolongent ceux de Rubin et Zanutto (2001) à deux égards. Premièrement, notre méthode permet d'utiliser plus d'un cas apparié par enregistrement. Nous montrons théoriquement que l'efficacité de l'imputation augmente à mesure que croît le nombre de cas appariés par enregistrement. Quand le coût des cas appariés est assez faible, notre méthode offre l'option d'utiliser l'information provenant de plus d'un cas apparié par enregistrement pour faciliter l'ajustement des modèles d'imputation, moyennant une dépense de traitement informatique négligeable. Deuxièmement, la méthode AMINp ne nécessite pas de modélisation paramétrique explicite de la ou des variances résiduelles, ce qui simplifie la tâche de modélisation (particulièrement dans le cas d'analyses avec résultats multivariés). Cette approche non paramétrique permet d'appliquer notre méthode à des ensembles de données présentant des structures de modèle complexes. Dans une étude par simulation, la méthode AMINp a produit des estimations dont le biais était réduit considérablement et des intervalles de confiance dont la couverture était correcte.

Bien que nous nous soyons concentrés sur l'appariement basé sur la géographie pour compléter les données sur les variables non observées couplées géographiquement, les procédures décrites dans l'article peuvent être généralisées à d'autres variables d'appariement. Par exemple, pour imputer des variables cliniques, il serait peut-être plus approprié de procéder à l'appariement à un autre patient dans le même hôpital, s'il est probable que les caractéristiques cliniques et les traitements soient plus fortement associés à l'hôpital qu'à l'emplacement géographique de la résidence du patient.

Remerciements

La présente étude a été financée en partie par le Bureau of the Census aux termes d'un contrat avec le National Opinion Research Center and Datametrics, Inc., et par une bourse de l'Agency for Healthcare Research and Quality (AHRQ) et du National Cancer Institute (HS09869). Les auteurs remercient John Z. Ayanian d'avoir dirigé le projet de recherche Quality of Cancer Care, Mark Allen et Robert Wolf d'avoir préparé les données, Bill Wright d'avoir appuyé la présente étude, ainsi que le rédacteur adjoint et deux examinateurs anonymes de leurs commentaires constructifs.

Bibliographie

Ayanian, J.Z., Zaslavsky, A.M., Fuchs, C.S., Guadagnoli, E., Creech, C.M., Cress, R.D., O'connor, L.C., West, D.W., Allen, M.E., Wolf, R.E. et Wright, W.E. (2003). Use of adjuvant chemotherapy and radiation therapy for colorectal cancer in a population-based cohort. *Journal of Clinical Oncology*, 21, 1293-1300.

Barnard, J., et Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948-955.

Chiu, W.F., Yucel, R.M., Zanutto, E. et Zaslavsky, A.M. (2001). Using matched substitutes to improve imputations for geographically linked databases. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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

Kalton, G. (1983). *Compensating for Missing Survey Data*. Research Report Series, Ann Arbor, MI : Institute for Social Research.

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

Krieger, N., Williams, D. et Andmoss, N. (1997). Measuring social class in U.S. public health research: Concepts, methodologies, and guidelines. *Annual Review of Public Health*, 18, 341-378.

Lessler, J.T., et Kalsbeek, W.D. (1992). *Nonsampling Errors in Surveys*. New York : John Wiley & Sons, Inc.

Little, R.J.A., et Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley & Sons, Inc.

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

Rubin, D.B., et Zanutto, E. (2001). Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations. Dans *Survey Nonresponse*, (Éds. R. Groves, R. Little et J. Eltinge), New York : John Wiley & Sons, Inc., 389-402.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London : Chapman & Hall.

Schafer, J.L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Logiciel pour Windows 95/98/NT disponible à http://www.stat.psu.edu/~jls/misoftwa.html.

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 79-92
Statistique Canada, N° 12-001-XIF au catalogue

Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines : Une application aux données de la NHANES

Balgobin Nandram et Jai Won Choi [1]

Résumé

Nous utilisons des modèles hiérarchiques bayésiens pour analyser les données sur l'indice de masse corporelle (IMC) des enfants et des adolescents en présence de non-réponse non-ignorable, c'est-à-dire informative, tirées de la troisième National Health and Nutrition Examination Survey (NHANES III). Notre objectif est de prédire l'IMC moyen en population finie et la proportion de répondants pour les domaines formés par l'âge, la race et le sexe (covariables dans les modèles de régression) pour chacun des 35 grands comtés, en tenant compte des non-répondants. Nous utilisons des méthodes de Monte Carlo par chaîne de Markov pour ajuster les modèles (deux modèles de sélection et deux modèles de mélange de schémas d'observation) aux données sur l'IMC provenant de la NHANES III. Au moyen d'une mesure de déviance et d'une étude de validation croisée, nous montrons que le modèle de sélection sous non-réponse non-ignorable est le meilleur des quatre modèles. Nous montrons aussi que l'inférence au sujet de l'IMC n'est pas trop sensible au choix du modèle. Nous obtenons une amélioration en incluant une régression spline dans le modèle de sélection pour tenir compte de l'évolution de la relation entre l'IMC et l'âge.

Mots clés : Validation croisée; déviance; échantillonneur de Metropolis-Hastings; modèle de régression logistique-normale; modèle de régression spline.

1. Introduction

La National Health and Nutrition Examination Survey (NHANES III) est l'une des enquêtes utilisées par le National Center for Health Statistics (NCHS) pour évaluer la santé de la population américaine. L'une des variables de cette enquête est l'indice de masse corporelle (IMC), qui est utilisé par l'Organisation mondiale de la santé pour définir l'embonpoint et l'obésité. Sous des conditions d'ignorabilité de la non-réponse, les estimateurs obtenus d'après les données de la NHANES III sont biaisés, parce que le nombre de non-répondants est élevé. Par conséquent, la question qui nous préoccupe principalement ici est qu'il faut tenir compte de la non-réponse, parce que les répondants et les non-répondants pourraient avoir des caractéristiques différentes. L'objectif de l'étude est de prédire l'IMC moyen en population finie des enfants et des adolescents, poststratifiés selon le comté pour chaque domaine formé par l'âge, la race et le sexe, et de déterminer quels ajustements sont nécessaires pour tenir compte de la non-réponse non-ignorable. Notre approche consiste à ajuster plusieurs modèles hiérarchiques bayésiens de façon à refléter le mécanisme de non-réponse.

Récemment, plusieurs articles traitant de l'embonpoint et de l'obésité ont été publiés. Dans son survol du premier plan national de lutte contre l'embonpoint et l'obésité, le Directeur du Service de santé publique des États-Unis a indiqué qu'il était nécessaire de procéder à des changements radicaux dans les écoles, les restaurants, les lieux de travail et les collectivités afin de combattre l'épidémie croissante d'embonpoint et d'obésité chez les Américains. Il a déclaré dans le rapport sur l'obésité qu'il ne s'agissait ni d'esthétique ni d'apparence, mais bien d'une question de santé. Comme l'a souligné Squires (2001), le coût total des soins de santé liés à l'embonpoint et à l'obésité sont de l'ordre de 117 milliards de dollars annuellement. Les enfants qui ont un surpoids font souvent de l'embonpoint à l'âge adulte, et chez l'adulte, l'embonpoint pause un risque pour la santé (Wright, Parker, Lamont et Craft 2001). Dans un article fort intéressant fondé sur les données de la NHANES, Ogden, Flegal, Carroll et Johnson (2002) décrivent les estimations nationales les plus récentes de la prévalence et de la tendance de l'embonpoint chez les enfants et les adolescents américains. Partant d'une analyse limitée, ils concluent qu'aux États-Unis, la prévalence de l'embonpoint chez les enfants continue de croître, particulièrement chez les adolescents américano-mexicains et noirs d'origine non hispanique. Plusieurs problèmes de santé ont été associés à l'embonpoint durant l'enfance. Un accroissement éventuel de la prévalence du diabète de type 2 est relié à la croissance de la prévalence de l'embonpoint chez les enfants (Fagot-Campagna 2000); il en est de même des facteurs de risque de maladies cardiovasculaires, des taux élevés de cholestérol et des taux

1. Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280. Courriel : balnan@wpi.edu; Jai Won Choi, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782. Courriel : jwc7@cdc.gov.

anormaux de glucose (Dietz 1998). Donc, il serait utile d'étudier l'IMC chez les enfants et chez les adolescents en appliquant des méthodes capables de fournir une correction appropriée pour la non-réponse et une meilleure mesure de précision.

En représentant par x les covariables et par y la variable de réponse, Rubin (1987) et Little et Rubin (1987) décrivent trois catégories de mécanisme produisant des données manquantes. Ces catégories diffèrent selon que la probabilité de réponse a) est indépendante de x et de y, b) dépend de x, mais non de y et c) dépend de y et, éventuellement, de x. Dans le cas a), les données manquent entièrement au hasard (MCAR, *missing completely at random*), dans le cas b), les données manquent au hasard (MAR, *missing at random*) et dans le cas c) les données ne manquent pas au hasard (MNAR, *missing not at random*). Les modèles construits pour les mécanismes MCAR et MAR sont appelés modèles de non-réponse ignorable, si les paramètres de la variable dépendante et de la réponse sont distincts (Rubin 1976). Les modèles pour les mécanismes MNAR de données manquantes sont appelés modèles de non-réponse non-ignorable, ou non-réponse informative.

Les modèles de non-réponse peuvent être classés de façon très générale en une catégorie de modèles de sélection et une catégorie de modèles de mélange de schémas d'observation (par exemple, voir Little et Rubin 1987). Soit $[y]$ et $[r]$ la densité de probabilité de la variable de réponse y et l'indicateur de réponse r, respectivement, avec des notations évidentes pour les lois conjointes et conditionnelles. Alors, le modèle de sélection spécifie que $[y,r]=[r\mid y][y]$ et le modèle de mélange de schémas d'observation spécifie que $[y,r]=[y\mid r][r]$. L'approche par sélection a été élaborée pour étudier les problèmes de sélection d'échantillon (par exemple, Heckman 1976 et Olson 1980). Bien que les deux modèles aient la même loi conjointe, en pratique, on spécifie les composantes $[r\mid y]$ et $[y]$ pour le modèle de sélection et $[y\mid r]$ et $[r]$ pour le modèle de schémas d'observation. Donc, ces modèles peuvent être différents.

Par conséquent, nous utilisons deux modèles de non-réponse non-ignorable, c'est-à-dire un modèle de sélection et un modèle de mélange de schémas d'observation, pour analyser les données de la NHANES III. Nous utilisons chaque modèle dans le cadre hiérarchique bayésien pour résoudre notre problème de non-réponse non-ignorable et nous comparons les résultats afin d'évaluer leur sensibilité au choix du modèle. Dans le modèle de sélection, la propension à répondre est reliée à l'IMC uniquement, si bien que le modèle de l'IMC est linéaire en fonction de l'âge, de la race, du sexe et de l'interaction de la race et du sexe. Dans le modèle de mélange de schémas d'observation, la propension à répondre est reliée à l'âge, à la race et au sexe (mais non à l'IMC), et le modèle de l'IMC présente deux formes linéaires étroitement liées en fonction de l'âge, de la race et du sexe et de l'interaction de la race et du sexe. Ces deux modèles tiennent pour l'ensemble de la population. Les valeurs de l'IMC des non-répondants et des personnes non échantillonnées sont prédites à partir de chaque modèle. Nous préférons le modèle de sélection, parce que nous pouvons intégrer la structure dans les données de la NHANES III et que, d'après des arguments statistiques, cela s'avère vérifié.

Greenlees, Reece et Zieschang (1982) ont élaboré un modèle de régression logistique-normale pour imputer des valeurs manquantes lorsque la probabilité de réponse dépend de la variable imputée. Ils ont appliqué le modèle à des données sur les traitements et salaires provenant de la Current Population Survey (CPS). David, Little, Samuel et Triest (1986) ont comparé la méthode hot deck appliquée aux données de la CPS et le modèle de régression logistique-normale appliqué aux données sur les traitements et salaires provenant d'un même ensemble de données et ont constaté que les résultats des deux méthodes différaient fort peu. Nous notons que le modèle de régression logistique-normale est un modèle de sélection de la non-réponse non-ignorable, mais qu'il ne rend pas compte de la mise en grappes. Dans le cas des données de la NHANES III, pour tenir compte de la mise en grappes dans les comtés, il est naturel de commencer par le modèle logistique-normale.

Notre modèle de sélection hiérarchique bayésien possède une structure spéciale. Dans la NHANES III, la propension à répondre augmente avec l'âge (la race et le sexe jouent un rôle mineur) et les médecins pensent que les personnes obèses ont tendance à ne pas se présenter au rendez-vous pour l'examen physique. Donc, étant donné les valeurs de l'IMC, comme dans Greenlees et coll. (1982), les indicateurs de réponse suivent un modèle de régression logistique où le logarithme de la valeur de l'IMC est la covariable. À leur tour, les logarithme des valeurs de l'IMC sont distribués selon un modèle linéaire dans lequel les covariables sont l'âge, la race et le sexe. Il s'agit de l'information la plus importante que nous intégrons dans le modèle de sélection. En outre, contrairement à Greenlees et coll. (1982), notre modèle inclut des effets de mise en grappes pour tenir compte de l'hétérogénéité entre les comtés au moyen des indicateurs de réponse et des valeurs de l'IMC. Ici, chaque comté possède son propre jeu de paramètres et il existe une distribution commune sur l'ensemble de ces jeux de paramètres. Il s'agit également d'une information a priori importante que nous devons intégrer dans le modèle, ce qui est l'une des caractéristiques intéressantes de la méthode hiérarchique bayésienne.

Dans l'approche bayésienne, la principale difficulté consiste à formuler la relation entre les répondants et les

non-répondants. Dans le cas du modèle de sélection, cette question peut être traitée au moyen de la structure logistique-normale. Nous considérons aussi un modèle hiérarchique bayésien dans le cas de l'approche du mélange de schémas d'observations. Le modèle de mélange de schémas d'observation est une alternative utile pour étudier la sensibilité à l'hypothèse faite dans le modèle de sélection. Pour évaluer l'hypothèse de non-réponse non-ignorable, nous considérons aussi des cas particuliers des modèles de sélection et de mélange de schémas d'observation afin d'obtenir deux modèles de non-réponse ignorable. Nous constatons qu'un cinquième modèle est nécessaire, dans lequel nous étendons notre modèle de sélection à un modèle de régression spline pour tenir compte de la relation dynamique entre l'IMC et l'âge.

Nandram, Han et Choi (2002) ont mis au point une méthode pour analyser les données sur l'IMC selon l'âge, la race et le sexe quand l'IMC est classé en trois intervalles. Cette méthode représente une extension multinomiale de l'analyse de données binaires sous non-réponse non-ignorable de Stasny (1991). Cette méthode s'applique généralement à n'importe quel nombre de cellules dans plusieurs régions (les comtés dans notre application). Nandram et Choi (2002 a, b) considèrent d'autres extensions des travaux de Stasny portant sur les données binaires (c'est-à-dire. les données provenant de la National Health Interview Survey et de la National Crime Survey). Ici, nous ne catégorisons pas les valeurs de l'IMC, mais nous les traitons plutôt, comme il se doit, comme des valeurs continues. Les quantités d'intérêt sont l'IMC moyen en population finie et la proportion de personnes qui répondent dans chaque domaine formé par l'âge, la race, le sexe et le comté.

Le reste de l'article est présenté comme suit. À la section 2, nous décrivons brièvement les données de la NHANES III. À la section 3, nous discutons des modèles hiérarchiques bayésiens pour la non-réponse ignorable et non-ignorable. Nous décrivons aussi l'ajustement du modèle, la sélection du modèle et l'évaluation basée sur la mesure de la déviance et la vérification croisée prédictive. À la section 4, nous décrivons l'analyse des données sur l'IMC provenant de la NHANES III. À la section 5, nous décrivons un modèle de régression spline et nous comparons les résultats. Enfin, à la section 6, nous présentons nos conclusions.

2. Données de la NHANES III

Le plan de sondage est un plan stratifié probabiliste à plusieurs degrés qui est représentatif de l'ensemble de la population civile non placée en établissement, âgée de deux mois ou plus, des États-Unis. Le nombre de personnes échantillonnées dans chaque groupe âge-race-sexe est connu pour chaque comté. La taille de l'échantillon par comté, âge, race et sexe est assez faible. Pour d'autres renseignements sur le plan de sondage de la NHANES III, consulter National Center for Health Statistics (1992, 1994).

La collecte des données de la NHANES III comprend deux volets : le premier est la sélection de l'échantillon et l'interview des membres des ménages échantillonnés en vue de recueillir les renseignements personnels à leur sujet et le second volet est l'examen physique des personnes interviewées dans un centre d'examen mobile (CEM). L'évaluation de la santé comporte des renseignements provenant de l'examen physique, des tests et des mesures faites par des techniciens, ainsi que le prélèvement d'échantillons pour analyse.

L'échantillon a été sélectionné auprès des ménages de 81 comtés des États-Unis continentaux d'octobre 1988 à septembre 1994. Toutefois, pour des raisons de confidentialité, les donnée finales retenues pour l'étude provenaient des 35 plus grands comtés (de 14 États) dont la population est supérieure à 500 000 habitants, pour certains groupes d'âge selon le sexe et la race. Dans le présent article, nous analysons les données à grande diffusion provenant de ces 35 comtés; les variables démographiques sont l'âge, la race et le sexe, et l'indicateur de l'état de santé d'intérêt est l'indice de masse corporelle (IMC), qui est égal au poids en kilogrammes divisé par le carré de la taille en mètres (Kuczmarski, Carrol, Flegal et Troiano 1997). Selon l'Organisation mondiale de la santé (Consultation de l'OMS sur l'obésité 2000), un adulte dont l'IMC est égal ou supérieur à 30 est obèse; la surcharge pondérale, ou embonpoint, s'entend des adultes dont l'IMC est compris dans l'intervalle [25, 30]. Pour les enfants de 1 à 6 ans et les adolescents de 7 à 19 ans, la définition de l'embonpoint et de l'obésité varie selon l'âge.

La non-réponse peut avoir lieu dans les volets interview et examen physique de l'enquête. La non-réponse à l'interview se produit lorsque les personnes échantillonnées ne participent pas à l'interview. Certaines des personnes interviewées et incluses dans le sous-échantillon pour l'évaluation de la santé n'ont pas subi l'examen physique à la maison ou au centre d'examen mobile, et ont donc manqué la totalité ou une partie des examens physiques. Ici, nous ne considérons pas le petit nombre de personnes pour lesquelles les valeurs de IMC et des covariables (âge, race et sexe) manquent (c'est-à-dire les cas de non-réponse totale). Par souci de simplicité et à toutes fins pratiques, il est raisonnable d'inclure toutes les personnes avec les covariables qu'elles ont déclarées (c'est-à-dire données complètes et non-réponses partielles) dans notre analyse. Cohen et Duffy (2002) font remarquer que les enquêtes sur la santé sont un bon exemple de situation où il semble

plausible que la propension à répondre soit liée à l'état de santé. Nous notons aussi que, pour les enfants et les adolescents, le taux observé de non-réponse est d'environ 24 %. L'une des raisons de la non-réponse chez les jeunes enfants est que les parents ou les mères plus âgés se sont montrés extrêmement protecteurs et n'ont pas permis que leur enfant quitte le domicile pour un examen physique.

Nous étudions les données sur l'IMC pour quatre groupes d'âge (2 à 4 ans, 5 à 9 ans, 10 à 14 ans et 15 à 19 ans). Si l'on se souvient qu'il existe 560 ($35 \times 4 \times 2 \times 2$) domaines, la taille d'échantillon par domaine est, en moyenne, très faible (par exemple $2\,647/560 \approx 4$). Donc, il est nécessaire d'« emprunter de l'information » aux autres domaines. En outre, la taille de l'échantillon est petite comparativement à la taille de la population finie (par exemple, $100 \times (2\,647/6\,653\,738) = 0{,}04$ %). Le problème de prédiction demande énormément de calcul. Les données observées indiquent une tendance à la hausse de l'IMC avec l'âge, avec un léger accroissement de la variabilité.

Les données de la NHANES III sont rajustées par étapes multiples de pondération par le quotient afin de les rendre représentatives de la population; voir Mohadjar, Bell et Waksberg (1994). Selon cette méthode d'ajustement par le quotient, la correction pour la non-réponse partielle se fait par estimation par le quotient dans la même classe d'ajustement en supposant que les distributions des répondants et des non-répondants sont identiques. Il est toutefois nécessaire de considérer d'autres méthodes d'ajustement que celle par le quotient pour traiter la non-réponse non-ignorable. Ici, nous présentons une méthode bayésienne comme option possible pour l'étude de la non-réponse dans le cas de la NHANES III.

Schafer, Ezzati-Rice, Johnson, Khare, Little et Rubin (1996) ont entrepris de procéder à une imputation multiple complète des données de la NHANES III pour de nombreuses variables. Le but du projet était d'imputer des données pour tenir compte de la non-réponse en vue de produire plusieurs ensembles de données à grande diffusion. L'une des contraintes imposées était que la procédure utilisée pour créer les données manquantes corresponde à un mécanisme purement ignorable et que la simulation ne fournisse aucune information sur l'effet des écarts possibles par rapport au mécanisme de non-réponse ignorable. Une autre contrainte était que la procédure ne comporte pas de mise en grappes géographique. L'objectif de la présente étude est différent; nous n'avons pas l'intention de fournir des données à grande diffusion imputées. Contrairement à Schafer et coll. (1996), nous incluons la mise en grappes au niveau du comté, bien qu'il puisse être nécessaire d'inclure la mise en grappes au niveau du ménage. Pour les données complètes, il existe 6 440 ménages. De ceux-ci, 52,1 % ont contribué une personne à l'échantillon, 22,5 %, deux personnes et 21,4 %, au moins trois personnes. Nous avons calculé le coefficient de corrélation pour les valeurs de l'IMC par appariement des membres dans les ménages (voir Rao 1973, page 199). La valeur de 0,19 obtenue indique qu'en première approximation, nous pouvons ignorer la mise en grappes dans les ménages.

Pour les besoins de notre application, nous devons faire une inférence pour chaque domaine âge-race-sexe dans un comté. L'une des méthodes standard d'estimation sur petits domaines consiste à identifier chaque petit domaine au moyen d'un paramètre, puis à supposer qu'il existe un processus stochastique commun sur les 560 paramètres. Toutefois, à cause de la rareté des données, l'application de cette méthode n'est pas souhaitable. Donc, nous construisons nos modèles au niveau du comté et nous représentons l'âge, la race et le sexe comme des covariables. Nous procédons à l'inférence pour chaque domaine formé par le recoupement de l'âge, de la race et du sexe dans le comté au moyen de nos modèles de régression, ce qui est un élément essentiel de notre analyse.

3. Méthode hiérarchique bayésienne

À la présente section, nous décrivons deux modèles bayésiens pour la non-réponse non-ignorable et nous déduisons deux modèles supplémentaires pour la non-réponse ignorable à titre de cas particuliers. Nous décrivons la sélection et l'évaluation du modèle pour le modèle choisi (c'est-à-dire le modèle de sélection).

Nous disposons de données provenant de $\ell = 35$ comtés et chaque comté comprend N_i personnes (connues). Nous supposons qu'un échantillon probabiliste de n_i personnes est tiré dans le i^{e} comté. Soit *s* l'ensemble d'unités échantillonnées et *ns* l'ensemble d'unités non échantillonnées. Soit r_{ij} pour $i = 1, 2, \ldots, \ell$ et $j = 1, 2, \ldots, N_i$ l'indicateur de réponse ($r_{ij} = 1$ pour les répondants et $r_{ij} = 0$ pour les non-répondants) pour la j^{e} personne dans le i^{e} comté dans la population. En outre, soit x_{ij} le logarithme de la valeur de l'IMC. Nous avons constaté que la transformation logarithmique donnait une meilleure représentation et nous l'utilisons donc dans tout l'exposé. Il convient de souligner que les valeurs de r_{ij} et x_{ij} sont toutes observées dans l'échantillon *s*, mais qu'elles sont inconnues dans *ns*. Soit $r_i = \sum_{j=1}^{n_i} r_{ij}$ (autrement dit, r_i est le nombre de personnes échantillonnées qui ont répondu dans le i^{e} comté).

Par souci de commodité, nous exprimons le logarithme de l'IMC x_{ij} sous la forme $x_{i1}, x_{i2}, \ldots, x_{ir_i}, x_{ir_i+1}, \ldots, x_{in_i}$ dans *s* et $x_{in_i+1}, \ldots, x_{iN_i}$ dans *ns* pour le comté *i*. Un point important que nous tenons à souligner pour la suite est que les r_i personnes ne sont pas nécessairement des répondants aléatoires provenant des n_i personnes échantillonnées

aléatoirement. Il s'agit là du biais de non-réponse dont nous devons tenir compte. Il est évident que nous devons prédire la valeur de l'IMC, x_{ij}, pour a) les non-répondants dans *s* et b) les personnes dans *ns*. Donc, pour la population finie de N_i personnes, nous avons besoin d'une inférence prédictive bayésienne pour

$$\overline{X}_i = \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i} \quad \text{et} \quad P_i = \frac{\sum_{j=1}^{N_i} r_{ij}}{N_i},$$

pour $i = 1, \ldots, \ell$.

En posant $\overline{x}_i^{(s,r)} = \sum_{j=1}^{r_i} x_{ij}/r_i$, $\overline{x}_i^{(s,nr)} = \sum_{j=r_i+1}^{n_i} x_{ij}/(n_i - r_i)$ et $\overline{x}_i^{(ns)} = \sum_{j=n_i+1}^{N_i} x_{ij}/N_i - n_i)$, nous notons que

$$\overline{X}_i = f_i\left\{g_i^{(s)}\,\overline{x}_i^{(s,r)} + (1-g_i^{(s)})\,\overline{x}_i^{(s,nr)}\right\} + (1-f_i)\,\overline{x}_i^{(ns)} \quad (1)$$

où $f_i = n_i/N_i$ et $g_i^{(s)} = r_i/n_i$. Soulignons que, alors que les f_i sont fixes en vertu du plan de sondage, les g_i et $\overline{x}_i^{(s,r)}$ sont observés. En outre, en posant $\hat{p}_i^{(s)} = r_i/N_i$ et $\hat{p}_i^{(ns)} = (\sum_{j=n_i+1}^{N_i} r_{ij})/(N_i - n_i)$,

$$P_i = f_i\,\hat{p}_i^{(s)} + (1-f_i)\,\hat{p}_i^{(ns)}, \quad (2)$$

$i = 1, \ldots, \ell$. Nous établissons nos modèles hiérarchiques bayésiens de façon à obtenir une inférence prédictive pour des quantités comme (1) et (2) suivant le domaine.

3.1 Modèles concurrents

Nos modèles comprennent deux parties, l'une pour le mécanisme de réponse et l'autre pour la distribution de l'IMC. Ces deux parties sont reliées pour former un modèle unique sous l'hypothèse de non-réponse non-ignorable ou de non-réponse ignorable.

Premièrement, nous décrivons le modèle de sélection. Pour la partie 1 de ce modèle, la réponse dépend de l'IMC comme suit

$$r_{ij} \mid x_{ij}, \boldsymbol{\beta}_i \sim \text{Bernoulli}\left\{\frac{e^{\beta_{0i}+\beta_{1i}\,x_{ij}}}{1+e^{\beta_{0i}+\beta_{1i}\,x_{ij}}}\right\}, \quad (3)$$

$$(\beta_{0i}, \beta_{1i}) \mid \theta_0, \theta_1, \sigma_1^2, \sigma_1^2, \rho_1 \overset{\text{iid}}{\sim} \text{BVNormale}(\theta_0, \theta_1; \sigma_1^2, \sigma_2^2, \rho_1), \quad (4)$$

$$\boldsymbol{\theta} \sim N(\boldsymbol{\theta}^{(0)}, \Delta^{(0)}),\ \sigma_1^{-2}, \sigma_2^{-2} \sim \text{Gamma}(a/2, a/2) \quad \text{et} \quad \rho_1 \sim \text{Uniforme}(-1, 1), \quad (5)$$

où $a, \boldsymbol{\theta}^{(0)}$ et $\Delta^{(0)}$ doivent être spécifiés. Notons que, dans (5), les lois a priori sont conjointement indépendantes. L'hypothèse (3) est importante, car elle établit le lien entre la propension à répondre et les valeurs de l'IMC; les médecins pensent que les personnes qui font de l'embonpoint ou qui sont obèses ont tendance à ne pas se présenter au centre d'examen mobile pour les examens demandés. L'hypothèse (4) tient compte de la mise en grappes dans les comtés et est celle qui permet le « renforcement par emprunt d'information » entre les comtés.

La deuxième partie du modèle a trait à l'IMC. Le prédicteur de loin le plus important de l'IMC est l'âge, le rôle de la race et du sexe étant relativement mineur. Une option consiste à poser que les valeurs de l'IMC sont

$$x_{ij} = \mu_{ij} + \in_{ij}, \quad \mu_{ij} = \alpha_{0ij} + \alpha_{1ij}\,a_{ij}$$

où a_{ij} dénote l'âge et $\in_{ij} \mid \sigma_3^2 \overset{\text{iid}}{\sim} \text{Normale}(0, \sigma_3^2)$ pour $i = 1, \ldots, \ell$ et $j = 1, \ldots, N_i$. En outre, il est nécessaire de comprendre la relation entre l'IMC et l'âge, la race et le sexe. Soit $z_{ij0} = 1$ pour une coordonnée à l'origine, $z_{ij1} = 1$ pour non noir et $z_{ij1} = 0$ pour noir, $z_{ij2} = 1$ pour masculin et $z_{ij2} = 0$ pour féminin, $z_{ij3} = z_{ij1}\,z_{ij2}$ pour l'interaction entre la race et le sexe, et soit $\mathbf{z}_{ij}' = (z_{ij0}, z_{ij1}, z_{ij2}, z_{ij3})$. Alors, pour la régression de l'IMC sur l'âge en corrigeant pour la race et le sexe, en posant $\boldsymbol{\alpha}_1' = (\alpha_{01}, \alpha_{02}, \alpha_{03}, \alpha_{04})$ et $\boldsymbol{\alpha}_2' = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14})$, nous prenons $\alpha_{0ij} = \mathbf{z}_{ij}'\,\boldsymbol{\alpha}_1 + v_{0i}$ et $\alpha_{1ij} = \mathbf{z}_{ij}'\,\boldsymbol{\alpha}_2 + v_{1i}$ pour obtenir

$$\mu_{ij} = (\mathbf{z}_{ij}'\,\boldsymbol{\alpha}_1 + v_{0i}) + (\mathbf{z}_{ij}'\,\boldsymbol{\alpha}_2 + v_{1i})a_{ij}$$

où v_{0i} et v_{1i} sont les effets aléatoires centrés à l'origine avec une loi normale bivariée donnée plus bas pour chaque modèle.

Donc, dans la deuxième partie du modèle de sélection, nous supposons que

$$x_{ij} = (\mathbf{z}_{ij}'\,\boldsymbol{\alpha}_1 + v_{0i}) + (\mathbf{z}_{ij}'\,\boldsymbol{\alpha}_2 + v_{1i})\,a_{ij} + e_{ij} \quad \text{et} \quad e_{ij} \mid \sigma_3^2 \overset{\text{iid}}{\sim} \text{Normale}(0, \sigma_3^2), \quad (6)$$

$$(v_{0i}, v_{1i}) \mid \sigma_4^2, \sigma_5^2, \rho_2 \overset{\text{iid}}{\sim} \text{BVNormale}(0, 0; \sigma_4^2, \sigma_5^2, \rho_2). \quad (7)$$

De nouveau, nous tenons compte de la mise en grappes dans les comtés au moyen de l'hypothèse (7), qui est celle qui permet le « renforcement par emprunt d'information » entre les comtés. Pour cette partie du modèle, nous utilisons les lois a priori

$$\boldsymbol{\alpha}_1 \sim \text{Normale}(\boldsymbol{\alpha}_2^{(0)}, \Delta_2^{(0)}) \text{ et } \boldsymbol{\alpha}_2 \sim \text{Normale}(\boldsymbol{\alpha}_3^{(0)}, \Delta_3^{(0)}),\ \sigma_3^{-2}, \sigma_4^{-2}, \sigma_5^{-2}, \overset{\text{iid}}{\sim} \text{Gamma}(a/2, a/2) \quad \text{et} \quad \rho_2 \overset{\text{iid}}{\sim} \text{Uniforme}(-1, 1) \quad (8)$$

où $a, \boldsymbol{\alpha}_k^{(0)}$ et $\Delta_k^{(0)}, k = 1, 2$ doivent être spécifiés. Notons que, dans (8), les lois a priori sont toutes conjointement indépendantes.

Nous présentons le modèle de mélange de schémas d'observation sous non-réponse non-ignorable à l'annexe A. Nous avons inclus la race, le sexe et leur interaction dans la partie réponse du modèle, quoique cela s'avère non nécessaire. La différence entre les répondants et les non-répondants dans le modèle de mélange de schémas d'observation est que, dans la régression, l'ordonnée à l'origine varie selon le comté pour les répondants, mais non pour les non-répondants; les autres paramètres sont les mêmes. De cette façon, nous pouvons « centrer » le modèle de non-réponse non-ignorable sur le modèle de non-réponse ignorable avec une certaine variation; consulter Nandram et Choi (2002 a) pour une idée comparable. Cette étape est nécessaire parce que les paramètres deviennent non identifiables si l'on suppose sans preuve scientifique qu'il existe une différence importante entre les répondants et les non-répondants dans le modèle de non-réponse non-ignorable. Bien que nous ayons utilisé des effets aléatoires pour faire la distinction entre les répondants et les non-répondants, les paramètres fournissant une différence systématique entre les répondants et les non-répondants dans le modèle de Rubin (1977) ne sont pas identifiables. Il convient de souligner que, dans le modèle de mélange de schémas d'observation donné en (A.4), il existe deux spécifications/schémas pour x_{ij} (*i.e.*, $r_{ij} = 0$ et $r_{ij} = 1$), mais que dans le modèle de sélection, il n'en existe qu'une seule.

Nous montrons comment spécifier les paramètres tels que $\boldsymbol{\theta}^{(0)}, \Delta^{(0)}, \boldsymbol{\alpha}_k^{(0)}, \Delta_k^{(0)}, k = 1, 2$ à l'annexe C. Pour obtenir une loi a priori diffuse appropriée, nous choisissons pour *a* une valeur telle que 0,002. Il est également possible d'utiliser une loi a priori de rétrécissement sur σ_1^{-2} et σ_2^{-2} (voir Natarajan et Kass 2000 et Daniels 1999). Néanmoins, cela n'est pas nécessaire dans le modèle hiérarchique.

L'une des propriétés intéressantes du modèle hiérarchique bayésien est qu'il introduit une corrélation entre les variables. Par exemple, dans le modèle de sélection, (4) et (7) introduisent une corrélation entre les r_{ij} et entre les x_{ij}, respectivement. Il s'agit de l'effet de mise en grappes dans les domaines. Il est possible d'obtenir ce genre d'effet directement, mais la démarche n'est pas aussi simple que dans un modèle hiérarchique. Un autre avantage du modèle hiérarchique est qu'il tient compte des variations extrinsèques entre les domaines, ce qui est intimement relié à l'effet de mise en grappes. Encore un autre avantage est que les spécifications du modèle sont robustes à des niveaux plus profonds que le processus d'échantillonnage (par exemple, l'inférence avec (5) et (8) est assez robuste à des perturbations modérées des spécifications des hyperparamètres). Nous avons observé cette robustesse empiriquement ici et dans d'autres applications.

Nous obtenons un modèle de sélection sous non-réponse ignorable en posant que $\beta_{1i} = 0$ pour tous les comtés avec un ajustement approprié du modèle de sélection. Pour un modèle de mélange de schémas d'observation sous non-réponse ignorable, nous posons que $x_{ij} = (\boldsymbol{z}_{ij}' \boldsymbol{\alpha}_1 + v_{0i}) + (\boldsymbol{z}_{ij}' \boldsymbol{\alpha}_2 + v_{1i})\ a_{ij} + \in_{ij}$ pour les deux valeurs de r_{ij}.

3.2 Ajustement du modèle

À la présente section, nous décrivons comment utiliser l'échantillonneur de Metropolis-Hastings pour ajuster les modèles. Nous utilisons aussi une mesure de déviance pour choisir le meilleur de nos quatre modèles. Puis, nous utilisons une analyse de validation croisée pour évaluer la qualité de l'ajustement du modèle sélectionné et, puisque les mêmes principes généraux s'appliquent aux quatre modèles, nous décrivons l'ajustement du modèle pour le modèle de sélection uniquement.

Donc, nous combinons maintenant le modèle du mécanisme de réponse et le modèle des valeurs de l'IMC pour obtenir la loi conjointe a posteriori de tous les paramètres. Les x_{ij} pour $j = r_i + 1, \ldots, n_i$, $i = 1, \ldots, \ell$ sont inconnus; autrement dit, ce sont des variables latentes. Nous représentons ces variables latentes par $\boldsymbol{x}^{(s,nr)}$ et les données observées par $\boldsymbol{x}^{\text{obs}}$. En nous servant du théorème de Bayes pour combiner la fonction de vraisemblance et la loi conjointe a priori, nous obtenons la loi conjointe a posteriori qui, outre la constante de normalisation, est $p(\boldsymbol{x}^{(s,nr)}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\theta}, \rho_1, \rho_2 \mid \boldsymbol{x}^{(s,r)})$ et est donnée par (B.1) à l'annexe B.

La loi a posteriori (B.1) est complexe, si bien que nous utilisons des méthodes de Monte Carlo par chaîne de Markov (MCMC) pour tirer des échantillons à partir de celle-ci. Plus précisément, nous utilisons l'échantillonneur de Metropolis-Hastings (voir Chib et Greenberg 1995, pour une discussion pédagogique). Nous utilisons aussi les tracés de courbes et les diagnostics d'autocorrélation passés en revue par Cowles et Carlin (1996) pour étudier la convergence et nous suivons la proposition de Gelman, Roberts et Gilks (1996) consistant à surveiller la probabilité de saut à chaque pas de Metropolis dans notre algorithme. Durant l'exécution des calculs, le centrage des valeurs de l'IMC facilite la réalisation de la convergence (voir Gelfand, Sahu et Carlin 1995). Cependant, il ne s'agit pas d'une tâche simple, car dans la régression logistique, le centrage a aussi une incidence sur la partie du modèle ayant trait à l'IMC.

Nous avons obtenu un échantillon de 1 000 itérations que nous avons utilisé pour l'inférence et la vérification du modèle. En utilisant les tracés de courbes, nous avons procédé à 1 000 itérations « d'apprentissage » et, pour annuler l'effet des autocorrélations, nous avons sélectionné ensuite une itération sur dix. Nous avons obtenu cette règle par tâtonnement, durant le réglage fin des pas de Metropolis. Nous avons maintenu les probabilités de saut dans l'intervalle (0,25, 0,50); voir Gelman et coll. (1996).

3.3 Sélection du modèle et évaluation du modèle

Nous utilisons l'approche de la perte prédictive a posteriori minimale (Gelfand et Ghosh 1998) pour sélectionner le meilleur des quatre premiers modèles.

Sous l'erreur quadratique comme fonction de perte, la perte prédictive minimale a posteriori est

$$D_k = P + \frac{k}{k+1} G$$

$$P = \sum_{ij} \mathrm{Var}(x_{ij}^{\mathrm{pre}} \mid \boldsymbol{x}^{\mathrm{obs}}), \quad G = \sum_{ij} \left\{ E(x_{ij}^{\mathrm{pre}} \mid \boldsymbol{x}^{\mathrm{obs}}) - x_{ij}^{\mathrm{obs}} \right\}^2$$

où $f(x_{ij}^{\mathrm{pre}} \mid \boldsymbol{x}^{\mathrm{obs}}) = \int f(x_{ij}^{\mathrm{pre}} \mid \Omega)\pi(\Omega \mid \boldsymbol{x}^{\mathrm{obs}}) d\Omega$ et x_{ij}^{pre} sont les valeurs prédites et Ω est l'ensemble de tous les paramètres. Cette mesure étend celle obtenue antérieurement (Laud et Ibrahim 1995) et nous prenons $k = 100$ pour établir la concordance avec cette version antérieure. Notons que, pour l'application en présence de non-réponse, nous calculons ces mesures uniquement d'après les données complètes sur l'IMC après avoir ajusté nos modèles de non-réponse.

Dans le tableau 1, nous présentons la mesure de la déviance (D_{100}) et ses composantes connexes, la qualité d'ajustement (G) et la pénalité (P), pour les quatre modèles. Si l'on se fonde sur la mesure de déviance, le modèle de sélection est nettement meilleur que les autres. Tandis que la valeur de P est à peu près la même que pour les autres modèles, celle de G est beaucoup plus petite, ce qui rend D_{100} plus petite pour le modèle de sélection. La différence entre les deux modèles de mélange de schémas d'observation est plus importante que celle entre les deux modèles de sélection. Cependant, comme nous ne disposons pas des erreurs-types, il est difficile de dire quel est le degré de signification de la différence.

Tableau 1
Comparaison des modèles de sélection et des modèles de mélange de schémas d'observation sous non-réponse ignorable et non-ignorable au moyen de la mesure de déviance

Modèle	G	P	D_{100}
SEI	135	135	270
SE	118	135	253
MSI	268	135	403
MS	204	135	339

Nota : $D_{100} = G + (100/(100+1))P$ où G est une mesure de la qualité de l'ajustement, P est une pénalité et D est la déviance; le modèle de mélange de schémas d'observation (MS) et le modèle de sélection (SE) sont tous deux des modèles à mécanisme de réponse non-ignorable. SEI est la version à non-réponse ignorable du modèle de sélection et MSI est la version à non-réponse ignorable du modèle de mélange de schémas d'observation.

Ensuite, nous examinons les déficiences du modèle de sélection. Nous utilisons une analyse de validation croisée bayésienne pour évaluer la qualité de l'ajustement du modèle choisi (c'est-à-dire le modèle de sélection). Pour cela, nous utilisons les résidus supprimés sur les valeurs de l'IMC des répondants.

Soit $(\boldsymbol{x}_{(ij)}, \boldsymbol{r}_{(ij)})$ le vecteur de l'ensemble des observations, sauf la $(ij)^{\mathrm{e}}$ observation (x_{ij}, r_{ij}). Alors, le $(ij)^{\mathrm{e}}$ résidu supprimé est donné par

$$\mathrm{DRES}_{ij} = \{x_{ij} - \mathrm{E}(x_{ij} \mid \boldsymbol{x}_{(ij)}, \boldsymbol{r}_{(ij)})\} \,/\, \mathrm{STD}(x_{ij} \mid \boldsymbol{x}_{(ij)}, \boldsymbol{r}_{(ij)}).$$

Ces valeurs sont obtenues en réalisant un échantillonnage préférentiel (pondéré) sur les données de sortie de l'algorithme de Metropolis-Hastings. Nous obtenons les moments a posteriori a partir de

$$f(x_{ij} \mid \boldsymbol{x}_{(ij)}, \boldsymbol{r}_{(ij)}) = \int f(x_{ij} \mid \Omega)\pi(\Omega \mid \boldsymbol{x}_{(ij)}, \boldsymbol{r}_{(ij)}) d\Omega.$$

Pour le modèle de mélange de schémas d'observation,

$$f(x_{ij} \mid \Omega) = f(x_{ij} \mid r_{ij} = 0, \Omega)\, p(r_{ij} = 0 \mid \Omega) + f(x_{ij} \mid r_{ij} = 1, \Omega)\, p(r_{ij} = 1 \mid \Omega)$$

et pour le modèle de sélection

$$f(x_{ij} \mid \Omega) \sim \mathrm{Normale}\{(\boldsymbol{z}_{ij}' \boldsymbol{\alpha}_1 + v_{0i}) + (\boldsymbol{z}_{ij}' \boldsymbol{\alpha}_2 + v_{1i}) a_{ij}, \sigma_3^2\}.$$

Nous avons également considéré l'utilisation de l'ordonnée conditionnelle a posteriori (OCP), qui est $f(x_{ij} \mid \boldsymbol{x}_{(ij)}, \boldsymbol{r}_{(ij)})$ évaluée au x_{ij} observé. Cependant, ces OCP ont mené à des résultats semblables pour le repérage des valeurs extrêmes.

Nous avons tracé des boîtes à moustache (non présentées) pour DRES en fonction des quatre niveaux race-sexe et des 35 comtés, ce qui nous a permis de constater que le modèle de sélection était bien ajusté. Nous avons également tracé les boîtes à moustache pour DRES en fonction de l'âge et, fait intéressant, nous avons observé une tendance. Pour le groupe des 2 à 4 ans, le modèle semble bien ajusté, tandis que pour le groupe des 5 à 9 ans, les valeurs prévues de l'IMC sont un peu élevées et pour les groupes des 10 à 14 ans et des 15 à 19 ans, la variabilité est plus importante. Nous avons examiné les boîtes à moustache pour DRES en fonction de l'âge de façon plus approfondie en traçant les boîtes à moustache pour 18 âges individuels (c'est-à-dire ceux compris entre 2 et 19 ans) (voir la figure 1). Pour les âges 11 à 19, le modèle est bien ajusté, mais pour les âges 2 à 10, il y a un problème (c'est-à-dire une courbure vers le bas dans les médianes). La même tendance s'observe pour les trois autres modèles. Un perfectionnement supplémentaire du modèle de sélection décrit à la section 5 permet de résoudre ce problème.

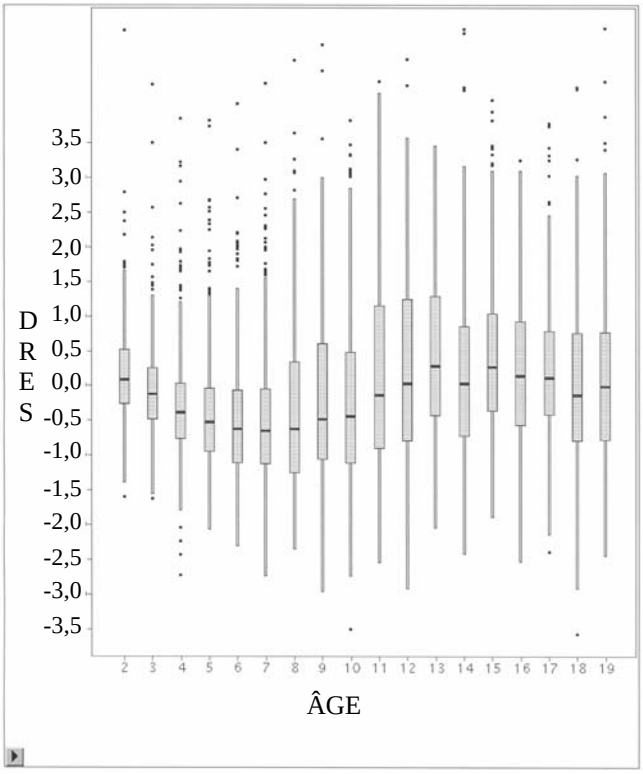

Figure 1. Boîtes à moustache pour les résidus de la vérification croisée (DRES) en fonction de l'âge pour le modèle de sélection

4. Estimation et prédiction

À la présente section, nous analysons les données de la NHANES III sur l'IMC des enfants et des adolescents (c'est-à-dire les jeunes de 2 à 19 ans). Nous utilisons le modèle de sélection, puis, pour étudier sa sensibilité, nous comparons la prédiction sous le modèle de sélection sous non-réponse non-ignorable à celle donnée par les trois autres modèles.

4.1 Estimation

Nous étudions la relation entre l'IMC et l'âge en utilisant les intervalles de confiance à 95 % pour les paramètres du modèle de sélection. En premier lieu, nous notons que l'interaction de la race et du sexe n'est pas importante, mais que, comme il faut s'y attendre, il existe une relation importante entre l'IMC et l'âge. L'IMC augmente considérablement avec l'âge [l'intervalle de confiance à 95 % pour α_{21} est (11,89, 13,67)]. Le taux de croissance est plus faible pour les garçons de race blanche [intervalle de confiance à 95 % pour α_{22} de (–2,30, – 0,19) et intervalle de confiance à 95 % pour α_{23} de (–3,03, – 0,64)]. Donc, bien que l'IMC augmente avec l'âge, l'accroissement est relativement plus faible pour les garçons de race blanche. À part le paramètre θ_1, qui indique un caractère informatif (non-ignorable) important, les autres paramètres sont essentiellement sans importance. Par exemple, les intervalles de confiance à 95 % pour ρ_1 et ρ_2 sont (– 0,53, 0,39) et (– 0,45, 0,45), respectivement, ce qui indique qu'on pourrait utiliser un modèle plus simple (c'est-à-dire $\rho_1 = \rho_2 = 0$).

Pour examiner plus en profondeur la question de l'ignorabilité, nous traçons les boîtes à moustache (non présentées) des lois a posteriori des β_{1i}, obtenues d'après les itérations de l'algorithme de Metropolis-Hastings, selon le comté. Toutes les boîtes à moustache sont situées au-dessus de zéro, ce qui donne à penser que, pour chaque comté, le mécanisme de non-réponse est non-ignorable. En outre, il existe divers degrés de non-ignorabilité. Par exemple, pour plusieurs comtés, la médiane de la boîte à moustache est proche de 1,5, tandis que pour d'autres, elle est proche de 2.

4.2 Prédiction

Il faut prédire la valeur moyenne de l'IMC, ainsi que la proportion de répondants dans la population finie. Les valeurs de l'IMC pour les non-répondants échantillonnés sont obtenues au moyen de leur loi conditionnelle

a posteriori incluse dans l'échantillonneur de Metropolis-Hastings. Les valeurs de l'IMC pour les personnes non échantillonnées doivent être prédites.

Il faut souligner que nous appliquons nos modèles au logarithme de l'IMC en retenant les covariables propres à chaque individu, si bien que le logarithme de chaque valeur non échantillonnée doit être prédit, puis transformé pour le ramener à l'échelle originale. Cependant, le fait que l'on ne connaît pas l'âge, la race et le sexe pour chaque personne non échantillonnée, mais que l'on connaît le nombre de personnes dans chaque domaine âge-race-sexe pour la population américaine selon le comté réduit considérablement les calculs.

Les distributions des personnes non échantillonnées sont

$$f(x_{ij}, r_{ij}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{r}^{\text{obs}}) = \int f(x_{ij}, r_{ij}|\Omega)\pi(\Omega|\boldsymbol{x}^{\text{obs}}, \boldsymbol{r}^{\text{obs}})d\Omega,$$

$i = 1, \ldots, \ell$, $j = n_i + 1, \ldots, N_i$. Pour le modèle de mélange de schémas d'observation, nous avons

$$f(x_{ij}, r_{ij}|\Omega) = f(x_{ij}|r_{ij}, \Omega)p(r_{ij}|\Omega)$$

et pour le modèle de sélection, nous avons

$$f(x_{ij}, r_{ij}|\Omega) = p(r_{ij}|x_{ij}, \Omega)f(x_{ij}|\Omega),$$

où Ω représente l'ensemble complet de paramètres.

Par conséquent, si nous tirons un échantillon de taille M à partir de la loi a posteriori, $\{\Omega^{(h)} : h = 1, \ldots, M\}$, un estimateur de $f(x_{ij}, r_{ij} | \boldsymbol{x}^{\text{obs}})$ est

$$f(x_{ij}, \hat{r}_{ij}|\boldsymbol{x}^{\text{obs}}) = M^{-1}\sum_{h=1}^{M} f(x_{ij}, r_{ij}|\Omega^{(h)}).$$

Donc, nous pouvons introduire les x_{ij} et les r_{ij} pour chaque $\Omega^{(h)}$ obtenu d'après l'algorithme MCMC à partir duquel nous obtenons M réalisations $\bar{X}_i^{(h)}, P_i^{(h)}, h = 1, \ldots, M$. Nous pouvons maintenant faire une inférence au sujet de $\bar{X}_i$ dans (1) et de P_i dans (2).

Nous présentons les intervalles de confiance à 95 % pour la valeur moyenne en population finie (MPF) de l'IMC et la proportion en population finie (PPF) de répondants afin d'évaluer la sensibilité aux quatre modèles. Notons que nous donnons ces intervalles pour chaque domaine race selon le sexe pour chaque groupe d'âge selon le comté et que, comme ils sont fort semblables d'un domaine à l'autre, nous présentons au tableau 2 la moyenne des bornes des intervalles de confiance sur l'ensemble des comtés pour les femmes noires. Pour la MPF, les intervalles sont fort semblables d'un modèle à l'autre. Cependant, pour la PPF, ils sont fort différents. Les intervalles pour le modèle de mélange de schémas d'observation et sa version à non-réponse ignorable sont semblables, sauf pour le groupe des 2 à 4 ans, ce à quoi il faut s'attendre, puisque ces modèles expriment une régression linéaire du logarithme de la cote exprimant la possibilité de répondre en fonction de l'âge. Pour la PPM sous les deux modèles de mélange de schémas d'observation, les intervalles sont essentiellement les mêmes, parce que la relation de l'IMC avec l'âge, la race, le sexe et leur interaction est la même. Pour la version à non-réponse ignorable du modèle de sélection, les intervalles sont tous les mêmes sur l'âge, parce que dans la partie de ce modèle ayant trait à la réponse, l'âge et l'IMC sont tous deux ignorés. Nous notons que, pour le modèle de sélection, les intervalles ont une forme semblable à ceux obtenus pour le modèle de mélange de schémas d'observation et sa version à non-réponse ignorable. Comme l'indiquent les intervalles, la MPF et la PPF augmentent avec l'âge.

Tableau 2
Comparaison des quatre modèles fondée sur la moyenne des bornes des intervalles de confiance à 95 % sur l'ensemble des comtés pour la moyenne en population finie (MPF) de l'IMC et la proportion en population finie (PPF) de répondants pour les femmes noires

		âge			
Modèle		2 à 4 ans	5 à 9 ans	10 à 14 ans	15 à 19 ans
SEI	MPF	(14,80, 16,07)	(17,09, 18,58)	(19,63, 21,61)	(22,40, 25,19)
	PPF	(0,73, 0,79)	(0,73, 0,79)	(0,73, 0,79)	(0,73, 0,79)
SE	MPF	(15,55, 16,21)	(17,49, 18,36)	(19,52, 20,92)	(21,74, 23,91)
	PPF	(0,66, 0,78)	(0,71, 0,81)	(0,75, 0,84)	(0,78, 0,87)
MSI	MPF	(14,75, 16,10)	(17,04, 18,59)	(19,59, 21,55)	(22,42, 25,09)
	PPF	(0,49, 0,70)	(0,72, 0,84)	(0,84, 0,94)	(0,90, 0,98)
MS	MPF	(14,96, 15,79)	(17,16, 18,38)	(19,61, 21,45)	(22,37, 25,07)
	PPF	(0,49, 0,70)	(0,73, 0,84)	(0,84, 0,94)	(0,90, 0,98)

Nota : SEI est la version à non-réponse ignorable du modèle de sélection, MSI est la version à non-réponse ignorable du modèle de mélange de schémas d'observation, MS est le modèle de mélange de schémas d'observations et SE est le modèle de sélection.

5. Un modèle de régression spline

Nous abordons maintenant la question que soulèvent les boîtes à moustache de la figure 1. Examinons plus en profondeur les données observées. Les boîtes à moustache des valeurs observées de l'IMC en fonction de l'âge montrent que l'IMC est pour ainsi dire constant de 2 à 8 ans, puis augmente à peu près linéairement de 8 à 13 ans et enfin, augmente très lentement de 14 à 19 ans. Cette caractéristique apparemment importante n'est pas incluse dans les quatre modèles. Par conséquent, à la présente section, nous essayons d'en tirer parti à l'aide d'un modèle de régression spline.

Nous utilisons la partie 1 du modèle de sélection et pour la partie 2, nous utilisons un modèle de régression join-point. De façon générique, en posant que $c^+ = 0$ si $c \le 0$ et $c^+ = c$ si $c > 0$, nous prenons

$$x_{ij} = \varphi_{0ij} + \varphi_{1ij}(a_{ij} - 8)^+ + \varphi_{2ij} a_{ij} - 13^+ + e_{ij} \qquad (9)$$

où, dans l'esprit de nos quatre modèles,

$$\varphi_{kij} = \boldsymbol{z}_{ij}\, \boldsymbol{\alpha}_k + v_{ki}, \quad k = 0, 1, 2.$$

Dans (9), nous avons posé que

$$e_{ij} \mid \sigma_3^2 \overset{\text{idd}}{\sim} \text{Normale}(0, \sigma_3^2)$$

et, motivé par notre résultat antérieur (les v_{ki} ne sont pas corrélés), au lieu d'une loi normale trivariée sur $\boldsymbol{v}_i = (v_{1i}, v_{2i}, v_{3i})'$, nous posons que

$$v_{ki} \mid \sigma_k^2 \overset{\text{idd}}{\sim} \text{Normale}(0, \sigma_k^2),\ k = 0, 1, 2.$$

Les hypothèses quant à la distribution des hyperparamètres ne changent pas.

Nous calculons la mesure de la déviance pour le modèle spline; voir le tableau 1 pour les quatre autres modèles. Pour ce modèle, $G \approx 129$ et $P \approx 107$, comparativement à $G \approx 118$ et $P \approx 135$ pour le modèle de sélection. Autrement dit, $D_{100} \approx 236$ pour le modèle de régression spline et $D_{100} \approx 253$ pour le modèle de sélection. Donc, le modèle de régression spline représente une amélioration par rapport au modèle de sélection original.

À la figure 2, nous présentons les boîtes à moustache pour DRES en fonction de l'âge. Ce diagramme est nettement meilleur que celui obtenu pour le modèle de sélection (voir la figure 1). Observons que les médianes fluctuent autour de 0 et que les variations sont faibles. Les boîtes à moustache obtenues pour 2, 3, 4, 5, 6 et 7 ans sont un peu moins variables que pour les autres âges. Nous ajustons aussi le modèle quadratique join-point dans lequel nous remplaçons (9) par

$$x_{ij} = \varphi_{0ij} + \varphi_{1ij}(a_{ij} - 8)^+ + \varphi_{2ij}\{(a_{ij} - 13)^+\}^2 + e_{ij},$$

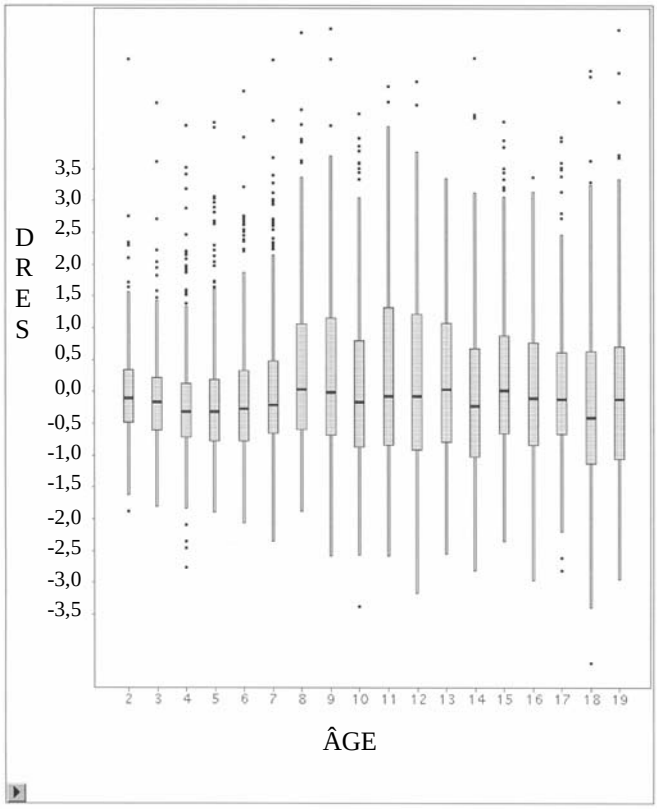

Figure 2. Boîtes à moustache des résidus de validation croisée (DRES) selon l'âge pour le modèle de régression spline

Tableau 3
Comparaison des deux modèles de sélection (régression avec spline et régression sans spline) fondée sur la moyenne des bornes des intervalles de confiance à 95 % sur l'ensemble des comtés pour la moyenne en population finie de l'IMC selon l'âge, la race et le sexe

		âge			
R–S		2 à 4 ans	5 à 9 ans	10 à 14 ans	15 à 19 ans
FN	Pas de spline	(16,26, 16,92)	(16,44, 17,10)	(19,62, 21,41)	(21,35, 25,62)
	Spline	(15,65, 16,31)	(17,62, 18,41)	(19,70, 20,91)	(21,95, 23,82)
MN	Pas de spline	(16,10, 16,76)	(16,26, 16,92)	(18,83, 20,55)	(20,45, 24,53)
	Spline	(15,68, 16,32)	(17,32, 18,11)	(19,03, 20,21)	(20,84, 22,61)
AF	Pas de spline	(16,39, 17,00)	(16,56, 17,17)	(19,48, 21,19)	(21,16, 25,39)
	Spline	(16,01, 16,60)	(17,77, 18,54)	(19,62, 20,79)	(21,61, 23,38)
AM	Pas de spline	(16,53, 17,14)	(16,67, 17,29)	(19,22, 20,95)	(20,83, 24,98)
	Spline	(16,16, 16,74)	(17,74, 18,51)	(19,38, 20,55)	(21,13, 22,87)

Nota : R–S = race-sexe, FN = femme noire, MN = homme noir, AF = autre femme, non noire, et AM = autre homme, non noir.

toutes les autres hypothèses demeurant par ailleurs inchangées. Ce modèle ne présente aucune amélioration appréciable par rapport au modèle spécifié par (9), que nous retenons sans autre perfectionnement.

Au tableau 3, nous comparons la MPF pour les modèles de sélection (régression sans spline et régression avec spline). De nouveau, nous calculons la moyenne des bornes des intervalles de confiance à 95 % sur l'ensemble des comtés. Les intervalles se chevauchent, ce qui porte à croire qu'il existe une similarité entre les modèles avec et sans spline. Cependant, nous notons certaines exceptions. L'écart le plus important entre les intervalles a lieu pour les jeunes de 15 à 19 ans. En général, le modèle spline donne une plus grande précision. Par exemple, pour le groupe des 10 à 19 ans, les intervalles pour le modèle spline sont contenus dans ceux obtenus pour le modèle sans les spline.

6. Conclusions

Pour analyser les données sur l'IMC provenant de la NHANES III selon l'âge, la race et le sexe dans chaque comté, a) nous avons étendu le modèle de régression logistique-normale à deux modèles de sélection hiérarchique bayésien et b) construit un modèle de mélange de schémas d'observation et deux modèles à non-réponse ignorable pour évaluer la sensibilité à l'inférence. Une mesure de déviance montre que, des quatre modèles, le modèle de sélection est le meilleur et une analyse de vérification croisée montre que l'ajustement des modèles est à peu près équivalent.

Une autre contribution de l'étude est le dépistage d'une déficience commune au modèle de sélection, au modèle de mélange de schémas d'observation et aux deux modèles à non-réponse ignorable. D'après les données observées, nous avons constaté qu'il existe une relation dynamique entre l'IMC et l'âge. Par conséquent, nous avons étendu plus loin le modèle de sélection afin d'inclure trois splines linéaires. L'analyse de validation croisée montre que cette approche offre une amélioration comparativement au modèle de sélection et, en fait, la mesure de déviance montre que le modèle de régression spline linéaire est le meilleur des cinq modèles.

Notre étude sur l'obésité est l'une des contributions importantes des travaux décrits ici. La régression spline linéaire de l'IMC sur l'âge, en corrigeant pour la race et le sexe, produit un meilleur ajustement et une plus grande précision que le modèle de sélection sans les splines. Il n'est pas facile de construire un modèle qui satisfasse à tous les aspects des données de la NHANES III simultanément. Nous avons réussi à le faire pour les enfants et les adolescents. L'IMC augmente considérablement avec l'âge; la race et le sexe contribuent négativement à cet accroissement; l'accroissement est relativement plus faible pour les garçons de race blanche. Certaines variations existent entre les 35 comtés.

Annexe A
Le modèle de mélange de schémas d'observation

Pour la partie 1 du modèle de mélange de schémas d'observation, la réponse dépend de l'âge, de la race et du sexe, ainsi que de l'interaction de la race et du sexe par la voie de la régression logistique.

$$r_{ij} \mid \beta_i \overset{\text{ind}}{\sim} \text{Bernoulli}\left\{ e^{\beta_{0i}+\beta_{1i}a_{ij}+\beta_{2i}z_{ij1}+\beta_{3i}z_{ij2}+\beta_{4i}z_{ij3}} / (1+e^{\beta_{0i}+\beta_{1i}a_{ij}+\beta_{2i}z_{ij1}+\beta_{3i}z_{ij2}+\beta_{4i}z_{ij3}}) \right\} \quad \text{(A.1)}$$

$i = 1, \ldots, l,\ j = 1, \ldots, N_i$. Maintenant, posons que $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i}, \beta_{4i})'$, et notons que, alors que le vecteur $\boldsymbol{\beta}_i$ possède $p = 5$ composantes, le vecteur correspondant dans (4) en possède deux. De façon analogue à (4) nous posons

$$\boldsymbol{\beta}_i \mid \boldsymbol{\theta}, \Delta \overset{\text{iid}}{\sim} \text{Normale}(\boldsymbol{\theta}, \Delta), \quad \text{(A.2)}$$

et, pour la loi a priori,

$$\boldsymbol{\theta} \sim \text{Normale}(\boldsymbol{\theta}^{(0)}, \Delta^{(0)})$$
$$\text{et} \quad \Delta^{-1} \sim \text{Wishart}\{(\nu^{(0)}\Lambda^{(0)})^{-1}, \nu^{(0)}\}, \nu^{(0)} > p, \qquad \text{(A.3)}$$

où $\boldsymbol{\theta}^{(0)}, \Delta^{(0)}, \Lambda^{(0)}$ et $\nu^{(0)}$ doivent être spécifiés. La partie 2 de ce modèle pour l'IMC intègre une dépendance à l'égard des indicateurs de réponse, en posant $w_{ij0} = 1, w_{ij1} = a_{ij}$,

$$x_{ij} = \sum_{t=0}^{1} (\boldsymbol{z}'_{ij}\boldsymbol{\alpha}_t + r_{ij}v_{ti})w_{ijt} + e_{ij}, r_{ij} = 0, 1,$$
$$e_{ij}|\sigma_3^2 \overset{\text{iid}}{\sim} \text{Normale}(0, \sigma_3^2). \qquad \text{(A.4)}$$

Les distributions sur les (v_{0i}, v_{1i}) sont les mêmes qu'en (7). Les lois a priori sont exactement celles de la partie 2 du modèle de sélection (c'est-à-dire voir (6) et (7)).

Nous prenons $\nu^{(0)} = 2p$, valeur qui indique une quasi-imprécision, maintient la justesse et permet la stabilité dans les calculs. Nous montrons comment spécifier les paramètres tels que $\boldsymbol{\theta}^{(0)}, \Delta^{(0)}, \boldsymbol{\alpha}_t^{(0)}, \Delta_t^{(0)}, t = 1, 2, 3, \Lambda^{(0)}$ à l'annexe C.

Annexe B
Algorithme de Metropolis-Hastings pour l'ajustement du modèle de sélection

Pour le modèle de sélection en présence de non-réponse non-ignorable, la loi conjointe a posteriori est

$$p(\boldsymbol{x}^{(s,nr)}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\theta}, \rho_1, \rho_2 | \boldsymbol{x}^{(s,r)}) \propto$$
$$\prod_{i=1}^{l}\left\{\prod_{j=1}^{r_i}\frac{1}{\sigma_3}e^{-\frac{1}{2\sigma_3^2}(x_{ij}-\{\boldsymbol{z}'_{ij}(\boldsymbol{\alpha}_1+a_{ij}\boldsymbol{\alpha}_2)+v_{0i}+v_{1i}a_{ij}\})^2}\frac{e^{\beta_{0i}+\beta_{1i}x_{ij}}}{1+e^{\beta_{0i}+\beta_{1i}x_{ij}}}\right\}$$
$$\times\prod_{i=1}^{l}\left\{\prod_{j=r_i+1}^{n_i}\frac{1}{\sigma_3}e^{-\frac{1}{2\sigma_3^2}(x_{ij}-\{\boldsymbol{z}'_{ij}(\boldsymbol{\alpha}_1+a_{ij}\boldsymbol{\alpha}_2)+v_{0i}+v_{1i}a_{ij}\})^2}\frac{1}{1+e^{\beta_{0i}+\beta_{1i}x_{ij}}}\right\}$$
$$\times\left\{\prod_{i=1}^{l}\frac{1}{\sigma_1\sigma_2\sqrt{1-\rho_1^2}}e^{-\frac{1}{2(1-\rho_1^2)}\left[\left(\frac{\beta_{0i}-\theta_0}{\sigma_1}\right)^2-2\rho_1\left(\frac{\beta_{0i}-\theta_0}{\sigma_1}\right)\left(\frac{\beta_{1i}-\theta_1}{\sigma_2}\right)+\left(\frac{\beta_{1i}-\theta_1}{\sigma_1}\right)^2\right]}\right\}$$
$$\times\left\{\prod_{i=1}^{l}\frac{1}{\sigma_4\sigma_5\sqrt{1-\rho_2^2}}e^{-\frac{1}{2(1-\rho_2^2)}\left[\left(\frac{v_{0i}}{\sigma_4}\right)^2-2\rho_2\left(\frac{v_{0i}}{\sigma_4}\right)\left(\frac{v_{1i}}{\sigma_5}\right)+\left(\frac{v_{1i}}{\sigma_5}\right)^2\right]}\right\}$$
$$\times\left\{\prod_{k=1}^{5}\left(\frac{1}{\sigma_k^2}\right)^{\frac{a}{2}+1}e^{-\frac{a}{2\sigma_k^2}}\right\}\left\{e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(0)})'\Delta^{(0)-1}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(0)})}\right\}$$
$$\times\left\{\prod_{k=1}^{2}e^{-\frac{1}{2}(\boldsymbol{\alpha}_k-\boldsymbol{\alpha}_k^{(0)})'\Delta_k^{(0)^{-1}}(\boldsymbol{\alpha}_k-\boldsymbol{\alpha}_k^{(0)})}\right\}. \qquad \text{(B.1)}$$

Soit Ω l'ensemble de paramètres $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\nu}$, $\boldsymbol{\alpha}$, σ_3^2, $\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$ et $\boldsymbol{x}^{(s,nr)}$, où $\boldsymbol{\psi}_1 = (\sigma_1^2, \sigma_2^2, \rho_1)'$ et $\boldsymbol{\psi}_2 = (\sigma_4^2, \sigma_5^2, \rho_2)'$. De façon générique, soit $\Omega_{\boldsymbol{a}}$ l'ensemble des paramètres dans Ω sauf $\boldsymbol{a}$; par exemple, $\Omega_{\boldsymbol{\beta}} = (\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\alpha}, \sigma_3^2, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \boldsymbol{x}^{(s,nr)})$, si bien que la loi conditionnelle a posteriori (LCP) de $\boldsymbol{\beta}$ est représentée par $p(\boldsymbol{\beta}|\Omega_{\boldsymbol{\beta}}, \boldsymbol{x}^{(s,r)})$. Pour exécuter l'algorithme de Metropolis-Hastings, nous avons besoin de la LCP de chaque paramètre sachant les autres et $\boldsymbol{x}^{(s,r)}$. Nous présentons ici brièvement l'algorithme.

Il est facile d'écrire la LCP pour chacun des paramètres $\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\alpha}$ et σ_3^2. Mais nous avons besoin des pas de Metropolis pour les LCP de $\boldsymbol{\beta}, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2$, et $\boldsymbol{x}^{(s,nr)}$.

En conditionnant sur $\Omega_{\boldsymbol{\beta}}$, les paramètres $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_l$, sont indépendants avec

$$p(\boldsymbol{\beta}_i|\boldsymbol{x}^{(s,r)}) \propto \prod_{j=1}^{n_i}\left\{\frac{e^{(\beta_{0i}+\beta_{1i}x_{ij})r_{ij}}}{1+e^{(\beta_{0i}+\beta_{1i}x_{ij})}}\right\}\times e^{-\frac{1}{2}(\boldsymbol{\beta}_i-\boldsymbol{\theta})'\Delta_1^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\theta})},$$

où

$$\Delta_1 = \begin{pmatrix}\sigma_1^2 & \rho_1\sigma_1\sigma_2\\ \rho_1\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$$

et $x_{ij}, i = 1, \ldots, l$ et $j = r_i + 1, \ldots, n_i$ doivent être prédits; voir plus loin. Nous utilisons une technique basée sur la régression logistique pour obtenir une loi instrumentale t de Student multivariée que nous mettons au point en faisant varier le nombre de degrés de liberté.

La méthode pour faire le tirage à partir de la LCP de $\boldsymbol{\psi}_1 = (\sigma_1^2, \sigma_2^2, \rho_1)$ et $\boldsymbol{\psi}_2 = (\sigma_3^2, \sigma_4^2, \rho_2)$ est la même. La LCP de ψ_2 est

$$p(\boldsymbol{\psi}_2|\Omega_{\boldsymbol{\psi}_2}, \boldsymbol{x}^{(s,r)}) \propto \left(\frac{1}{\sigma_4^2\sigma_5^2}\right)^{\frac{a+l}{2}+1}e^{-\frac{b}{2}\left(\frac{1}{\sigma_4^2}+\frac{1}{\sigma_5^2}\right)}$$
$$\times\frac{1}{(1-\rho_2^2)^{l/2}}e^{-\frac{1}{2(1-\rho_2^2)}\left\{\frac{1}{\sigma_4^2}\sum_{i=1}^{l}v_{0i}^2-\frac{2\rho_2}{\sigma_4\sigma_5}\sum_{i=1}^{l}v_{0i}v_{1i}+\frac{1}{\sigma_5^2}\sum_{i=1}^{l}v_{1i}^2\right\}}.$$

Nous avons utilisé la transformation z de Fisher (voir Ruben 1966) pour obtenir une loi instrumentale associée à la loi normale pour $\log\{\rho_2/(1-\rho_2)\}$ et à des lois gamma pour σ_4^2 et σ_5^2.

Enfin, nous considérons le pas de Metropolis pour tirer $\boldsymbol{x}^{(s,nr)}|\Omega_{\boldsymbol{x}^{(s,nr)}}, \boldsymbol{x}^{(s,r)}$. Nous notons que, dans cette LCP, les $x_{ij}, i = 1, \ldots, l, j = r_i + 1, \ldots, n_i,$ sont indépendants avec

$$p(x_{ij}|\Omega_{ij}, \boldsymbol{x}^{(s,r)}) \propto e^{-\frac{1}{2\sigma_3^2}[x_{ij}-\{\boldsymbol{z}_{ij}(\boldsymbol{\alpha}_1+a_{ij}\boldsymbol{\alpha}_2)+v_{0i}+v_{1i}a_{ij}\}]^2}$$
$$\left\{1+e^{\beta_{0i}+\beta_{1i}x_{ij}}\right\}^{-1}.$$

Nous avons construit une loi instrumentale en utilisant les techniques des moindres carrés. Nous notons que la loi

instrumentale Normale $(\boldsymbol{z}_{ij}(\boldsymbol{\alpha}_1 + a_{ij}\boldsymbol{\alpha}_2) + v_{0i} + v_{1i}a_{ij}, \sigma_3^2)$ n'a pas donné de bons résultats (voir Chib et Greenberg 1995).

Annexe C
Spécification des hyperparamètres

Nous discutons de la façon de spécifier les hyperparamètres $(\boldsymbol{\theta}^{(0)}, \Delta^{(0)})$ et $(\boldsymbol{\alpha}_k^{(0)}, \Gamma_k^{(0)})$, $k = 1, 2$, associés à $\boldsymbol{\theta}$ et $\boldsymbol{\alpha}_k$, $k = 1, 2$ dans le modèle de sélection.

Premièrement, considérons $(\boldsymbol{\theta}^{(0)}, \Delta^{(0)})$. Pour $i = 1, \ldots, l$, $j = 1, \ldots, n_i$, ajustons le modèle de régression logistique $r_{ij} \overset{\text{iid}}{\sim}$ Bernoulli $\{e^{\beta_{0i}+\beta_{1i}x_{ij}}(1+e^{\beta_{0i}+\beta_{1i}x_{ij}})^{-1}\}$, où les x_{ij} sont obtenus par prédiction (voir l'annexe A). Posons que $\hat{\boldsymbol{\beta}}_i$, $i = 1, \ldots, l$ représente les estimateurs par les moindres carrés et supposons que $\hat{\boldsymbol{\beta}}_i \overset{\text{iid}}{\sim}$ Normale $(\boldsymbol{\theta}^{(0)}, \tilde{\Delta}^{(0)})$ pour obtenir $\boldsymbol{\theta}^{(0)} = 1/l \sum_{i=1}^{l} \hat{\boldsymbol{\beta}}_i$ et

$$\hat{\tilde{\Delta}}^{(0)} = \frac{1}{l-1}\sum_{i=1}^{l} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\theta}_{(0)})(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\theta}_{(0)}) \qquad \text{(C.1)}$$

et fixons $\Delta^{(0)} = \kappa_1 \hat{\tilde{\Delta}}^{(0)}$, où κ_1 doit être sélectionné.

Puis, nous considérons la façon de spécifier $(\boldsymbol{\alpha}_k^{(0)}, \Gamma_k^{(0)})$, $k = 1, 2$. Nous ajustons $x_{ij} = \boldsymbol{z}_{ij}'(\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 a_{ij}) + e_{ij}$, où a_{ij} est l'âge de la j^{e} personne dans le i^{e} comté, $i = 1, \ldots, l$, $j = 1, \ldots, n_i$ pour obtenir les estimateurs par les moindres carrés, $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2)$ et leur matrice de covariance $\hat{\Gamma}^{(0)}$. Nous fixons $\boldsymbol{\alpha}_k^{(0)} = \hat{\boldsymbol{\alpha}}_k$ et $\Gamma_k^{(0)} = \kappa_2 \hat{\Gamma}_k^{(0)}$, où $\hat{\Gamma}_k^{(0)}$, $k = 1, 2$ est la matrice par blocs correspondante de $\hat{\Gamma}^{(0)}$, $k = 1, 2$ et κ_2 doit être spécifié.

Nous avons expérimenté avec κ_1 dans (C.1). Nous avons utilisé $\kappa_1 = 100$ pour fournir une loi a priori diffuse appropriée; l'utilisation de la valeur $\kappa_1 = 1\,000$ n'a pas modifié nos prédictions. Pareillement, nous avons utilisé $\kappa_2 = 100$.

Bibliographie

Chib, S., et Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49, 327-335.

Cohen, G., et Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics*, 18, 13-23.

Cowles, M., et Carlin, B. (1996). Markov chain Monte Carlo diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.

Daniels, M.J. (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, 27, 569-580.

David, M., Little, R.J.A., Samuel, M.E. et Triest, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.

Dietz, W.H. (1998). Health consequences of obesity in youth: Childhood predictors of adult disease. *Pediatrics*, 101, 518-525.

Fagot-Campagna, A. (2000). Emergence of type 2 diabetes mellitus in children: Epidemiological evidence. *Journal of Pediatric Endocrinology Metabolism*, 13, 1395-1405.

Gelfand, A., et Ghosh, S. (1998). Model choice: A minimum posterior predictive approach. *Biometrika*, 85, 1-11.

Gelfand, A., Sahu, S. et Carlin, B. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82, 479-488.

Gelman, A., Roberts, G.O. et Gilks, W.R. (1996). Efficient Metropolis jumping rules. Dans *Bayesian Statistics* (Éds. J.M. Bernardo, J.O. Berger, A.P. Dawid et A.F.M. Smith, Oxford, U.K.: Oxford University Press, 599-607.

Greenlees, J.S., Reece, W.S. et Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475-492.

Kuczmarski, R.J., Carrol, M.D., Flegal, K.M. et Troiano, R.P. (1997). Varying body mass index cutoff points to describe overweight prevalence among U.S. adults: NHANES III (1988 à 1994). *Obesity Research*, 5, 542-548.

Laud, P., et Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series B*, 57, 247-262.

Little, R.J.A., et Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.

Mohadjar, L., Bell, B. et Waksberg, J. (1994). National health and Nutrition Examination Survey III-Accounting for item nonresponse bias. Rapport interne, National Center for Health Statistics.

Nandram, B,. et Choi, J.W. (2002 a). A hierarchical Bayesian nonresponse model for binary data with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., et Choi, J.W. (2002 b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., Han, G. et Choi, J.W. (2002). Un modèle bayesien hiérarchique de non-réponse non-ignorable pour les données multinomiales des petites régions. *Techniques d'enquête*, 28, 157-170.

Natarajan, R., et Kass, R.E. (2000). Reference Bayesian methods for generalized linear models. *Journal of the American Statistical Association*, 95, 227-237.

National Center for Health Statistics (1992). Third national health and nutrition examination survey. *Vital and Health Statistics Series* 2, 113.

National Center for Health Statistics (1994). Plan and operation of the third national health and nutrition examination survey. *Vital and Health Statistics Series* 1, 32.

Ogden, C.L., Flegal, K.M., Carroll, M.D. et Johnson, C.L. (2002). Prevalence and trends in overweight among us children and adolescents, 1999-2000. *Journal of the American Medical Association*, 288, 1728-1732.

Olson, R.L. (1980). A least square correction for selectivity bias. *Econometrica*, 48, 1815-1820.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons, Inc.

Ruben, H. (1966). Some new results on the distribution of the sample correlation coefficient. *Journal of the Royal Statistical society, Series B*, 28, 513-525.

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

Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 538-543.

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

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M. Little, R.J. A. et Rubin, D.B. (1996). The NHANES III multiple imputation project. *Survey research methods, Proceedings of the American Statistical Association*, 28-37.

Squires, S. (2001). National plan urges to combat obesity: Weight-related illnesses kill 300,000 Americans annually, Surgeon General says. *The Washington Post*, 14 décembre, 2001.

Stasny, E.A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the National Crime Survey. *Journal of the American Statistical Association*, 86, 296-303.

Who Consultation on Obesity (2000). Obesity: Preventing and managing the global epidemic. *WHO Technical Report Series 894*, Geneva, Switzerland: World Health Organization.

Wright, C.M., Parker, L., Lamont, D. et Craft, A.W. (2001). Implications of childhood obesity for adult health: Findings from thousand families cohort study. *British Medical Journal*, 323, 1280-1284.

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 93-101
Statistique Canada, N° 12-001-XIF au catalogue

Vers des poids de régression non négatifs pour les échantillons d'enquête

Mingue Park et Wayne A. Fuller [1]

Résumé

Diverses procédures en vue de construire des vecteurs de poids de régression non négatifs sont considérées. Un vecteur de poids de régression dans lequel les poids initiaux sont les inverses des probabilités de sélection conditionnelles approximatives est présenté. Une étude par simulation permet de comparer les poids obtenus par la régression pondérée, la programmation quadratique, la méthode itérative du quotient, une procédure logit et la méthode du maximum de vraisemblance.

Mots clés : Méthode itérative du quotient; maximum de vraisemblance; programmation quadratique; estimateur simple conditionnellement pondéré.

1. Introduction

Dans le domaine du sondage, on dispose souvent d'information au sujet de la population à l'étape de l'analyse. L'estimation par la régression est l'une des méthodes choisies pour utiliser cette information. La construction d'un estimateur par la régression d'une moyenne ou d'un total de population peut se faire de plusieurs façons. L'un des estimateurs de la moyenne par la régression est

$$\bar{y}_{\text{reg}} = \sum_{i=1}^{n} w_i \, y_i = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\,\tilde{\boldsymbol{\beta}}, \qquad (1)$$

où

$$w_i = \alpha_i + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\left(\sum_{j=1}^{n} \bar{\mathbf{x}}_j' \, \phi_{jj}^{-1} \, \mathbf{x}_j\right)^{-1} \mathbf{x}_i \, \phi_{ii}^{-1}, \quad (2)$$

$$(\bar{y}_\pi \,,\, \bar{\mathbf{x}}_\pi) = \left(\sum_{i=1}^{n} \pi_i^{-1}\right)^{-1} \sum_{i=1}^{n} \pi_i^{-1} (y_i \,,\, \mathbf{x}_i) =: \sum_{i=1}^{n} \alpha_i \, (y_i \,,\, \mathbf{x}_i),$$

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{x}_i' \, \phi_{ii}^{-1} \, \mathbf{x}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i' \, \phi_{ii}^{-1} \, y_i \,,$$

$$\alpha_i = \left(\sum_{j=1}^{n} \pi_j^{-1}\right)^{-1} \pi_i^{-1},$$

$\boldsymbol{\Phi} = \text{diag}(\phi_{11}, \ldots, \phi_{nn})$ est une matrice diagonale non singulière, les π_i sont les probabilités de sélection et $\bar{\mathbf{x}}_N$ est la moyenne de population de $\mathbf{x}$. Un choix possible pour ϕ_{ii}^{-1} est α_i. Une revue de l'utilisation de ce genre d'information dans l'estimation par la régression dans le cas des enquêtes par sondage est donnée par Fuller (2002).

Il est bien connu que les poids de régression utilisés pour définir un estimateur par la régression tel que (2) peuvent être très grands ou (et) négatifs. Si ces poids doivent être utilisés pour estimer un total de population finie dans le cas d'une enquête générale, il semble raisonnable de poser qu'aucun poids individuel ne soit inférieur à 1. En outre, il semble raisonnable, du point de vue de la robustesse, d'éviter les poids dont la valeur est très grande.

Il existe plusieurs moyens de construire des poids de régression dont la fourchette de valeur est réduite. Huang et Fuller (1978) définissent une procédure pour modifier les w_i de sorte qu'il n'y ait aucun poids négatif et aucun poids de valeur élevée. Husain (1969) propose de recourir à la programmation quadratique pour imposer des bornes aux poids. La programmation quadratique et plusieurs autres procédures reposent sur le fait qu'on peut définir les poids comme des valeurs qui optimisent une certaine fonction. Deville et Särndal (1992) envisagent sept fonctions objectives susceptibles d'être utilisées pour construire les poids. Ils proposent des fonctions objectives utilisables pour produire des poids qui tombent dans une fourchette donnée. Deville, Särndal et Sautory (1993) présentent le programme CALMAR, rédigé sous forme de macro SAS, qui peut être utilisé pour calculer les poids correspondant à quatre fonctions objectives distinctes, quand l'information auxiliaire dans l'enquête correspond à des dénombrements de marge connus dans une table de fréquences.

Une autre modification des poids de régression consiste à assouplir certaines contraintes appliquées pour construire l'estimateur. Husain (1969) envisage de modifier les poids d'un échantillon aléatoire simple issu d'une loi normale. Il calcule les poids qui minimisent l'erreur quadratique moyenne (EQM) de l'estimateur résultant. Bardsley et Chambers (1984) considèrent un estimateur fondé sur une fonction objective et sur la division de la variable auxiliaire en deux composantes. Ils étudient le comportement de l'estimateur dans la perspective d'un modèle. Rao et Singh (1997) étudient un estimateur dans lequel des tolérances

1. Mingue Park, University of Nebraska, 103 Miller Hall, Lincoln, NE, 68588-0712, États-Unis; Wayne A. Fuller, Iowa State University, 221 Snedecor Hall, Ames, IA 50011-1210, États-Unis.

sont données pour la différence entre l'estimateur final d'une partie du vecteur de variables auxiliaires et les éléments correspondants du vecteur de population.

Dans le présent article, nous considérons divers types de poids de régression, y compris une procédure fondée sur les probabilités de sélection conditionnelles de Tillé (1998). Nous utilisons les probabilités de sélection conditionnelles approximatives pour calculer des poids de régression qui sont positifs pour la plupart des échantillons. Nous comparons ces poids à ceux obtenus par la MIQ, la programmation quadratique, une procédure logit et l'estimation du maximum de vraisemblance.

2. Maximum de vraisemblance et MIQ

Considérons un tableau à double entrée contenant r lignes et c colonnes. La cellule de population U_{ij} contient N_{ij} éléments; $i = 1, ..., r$, $j = 1, ..., c$. Supposons que les dénombrements de marge $N_{i\cdot}$, $N_{\cdot j}$ soient connus. Les caractéristiques de la population d'intérêt sont les N_{ij} ou, de façon équivalente, les $p_{ij} = N^{-1} N_{ij}$. Pour un échantillon aléatoire simple sans remise de taille n, Deming et Stephan (1940) ont proposé une méthode d'ajustement proportionnel itératif appelée méthode itérative du quotient (MIQ) pour obtenir la solution pour les fréquences de cellule. Voir aussi Stephan (1942). Si nous supposons que l'échantillon est un échantillon aléatoire issu d'une loi multinomiale définie par les valeurs de population dans un tableau à double entrée, nous pouvons construire un estimateur par la méthode du maximum de vraisemblance.

Deville et Särndal (1992) ont défini une classe d'estimateurs par calage, $\bar{y}_{\text{cal}}$, de la moyenne de population de y de la forme

$$\bar{y}_{\text{cal}} = \sum_{i=1}^{n} w_i \, y_i \,, \tag{3}$$

où les w_i minimisent la fonction objective $\sum_{i=1}^{n} G(w_i, \alpha_i)$ sous les contraintes

$$\sum_{i=1}^{n} w_i = 1 \quad \text{et} \quad \sum_{i=1}^{n} w_i \, \mathbf{x}_i = \bar{\mathbf{x}}_N \,, \tag{4}$$

et $G(w_i, \alpha_i)$ est une mesure de la distance entre le poids initial α_i et le poids final w_i. Les estimateurs par la MIQ et par le maximum de vraisemblance de la fraction de population de la cellule, p_{ij}, appartiennent à la classe des estimateurs par calage.

Pour un échantillon aléatoire simple, nous pouvons obtenir les poids par la MIQ pour la fraction de population de la cellule en minimisant

$$\sum_{k=1}^{n} w_k \log\left(\frac{w_k}{n^{-1}}\right) - w_k + n^{-1}, \tag{5}$$

sous les contraintes (4) avec

$$\mathbf{x}_k = (\delta_{1\cdot}, ..., \delta_{r\cdot}, \delta_{\cdot 1}, ..., \delta_{\cdot c}), \tag{6}$$

où $\delta_{i\cdot} = 1$ si le k^{e} élément appartient à la i^{e} ligne et $\delta_{i\cdot} = 0$ autrement, et $\delta_{\cdot j} = 1$ si le k^{e} élément appartient à la j^{e} colonne et $\delta_{\cdot j} = 0$ autrement. L'estimateur par la MIQ pour la fraction de cellule de population p_{ij} est l'estimateur (3) où $y_k = 1$ si le k^{e} élément appartient à la cellule ij et $y_k = 0$ autrement.

Pour l'estimateur par le maximum de vraisemblance de la fraction de population, avec un échantillon aléatoire simple, Deville et Särndal (1992) proposent de minimiser

$$\sum_{k=1}^{n} -n^{-1} \log\left(\frac{w_k}{n^{-1}}\right) + w_k - n^{-1} \tag{7}$$

sous les contraintes (4) avec $\mathbf{x}$ défini dans (6).

Chen et Sitter (1999) proposent *un estimateur de la pseudo-vraisemblance empirique*. Ils définissent la vraisemblance de population de y_i comme suit

$$\sum_{i=1}^{N} \log w_{i,U} \,, \tag{8}$$

où $w_{i,U}$ est la densité à l'observation y_i. Pour un échantillon de taille n, ils proposent l'estimateur de la pseudo-vraisemblance empirique de la forme

$$\bar{y}_{\text{EL}} = \sum_{i=1}^{n} w_i \, y_i \,, \tag{9}$$

où les w_i sont obtenus en minimisant la fonction

$$-\sum_{i=1}^{n} \pi_i^{-1} \log w_i \,, \tag{10}$$

sous les contraintes (4). Les w_i résultants sont égaux à ceux obtenus en minimisant l'expression (7) avec $= N \pi_i$ sous les contraintes (4).

Deville et Särndal (1992) montrent que les estimateurs par la MIQ et par le maximum de vraisemblance équivalent approximativement à un estimateur par la régression de la forme (1) et, par conséquent, ont la même loi limite que l'estimateur par la régression. Les poids pour les estimateurs par la MIQ et par le maximum de vraisemblance sont non négatifs si les solutions pour les poids existent.

3. Régression pondérée en utilisant des probabilités conditionnelles

Tillé (1998) propose d'utiliser des probabilités de sélection conditionnelle approximatives, sachant les estimateurs d'Horvitz-Thompson des variables auxiliaires, pour calculer un estimateur de la moyenne de population de la variable étudiée. Son approximation peut être étendue à la

production de poids de régression qui sont non négatifs avec probabilité élevée.

Supposons que le vecteur des moyennes de population des variables auxiliaires, $\bar{\mathbf{x}}_N$, soit connu. Considérons l'estimateur d'Horvitz-Thompson de $\bar{\mathbf{x}}_N$ donné par

$$\bar{\mathbf{x}}_{\mathrm{HT}} = \frac{1}{N}\sum_{i=1}^{n}\frac{\mathbf{x}_i}{\pi_i}, \qquad (11)$$

où $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ et π_i est la probabilité de sélection inconditionnelle. Tillé (1998) propose l'estimateur simple conditionnellement pondéré (SCP),

$$\bar{y}_{p\pi} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_{i|\bar{x}_{\mathrm{HT}}}}, \qquad (12)$$

où $\pi_{i|\bar{x}_{\mathrm{HT}}}$ est la probabilité de sélection conditionnelle du i^{e} élément, sachant $\bar{\mathbf{x}}_{\mathrm{HT}}$. Pour construire l'estimateur SCP de $\bar{y}_N$, il faut connaître la probabilité de sélection conditionnelle $\pi_{i|\bar{x}_{\mathrm{HT}}}$. Si $\bar{\mathbf{x}}_{\mathrm{HT}}$ prend la valeur $\mathbf{t}$, nous avons

$$\pi_{i|\bar{x}_{\mathrm{HT}}} = \pi_i \frac{P\{\bar{\mathbf{x}}_{\mathrm{HT}} = \mathbf{t} \mid i \in A\}}{P\{\bar{\mathbf{x}}_{\mathrm{HT}} = \mathbf{t}\}}, \qquad (13)$$

où A est l'ensemble d'indices pour les éléments de l'échantillon.

Afin de calculer les probabilités de sélection conditionnelles, il faut connaître la loi de probabilité de $\bar{\mathbf{x}}_{\mathrm{HT}}$ inconditionnelle et conditionnelle à la présence de chaque unité dans l'échantillon. À part certains cas particuliers, cette loi de probabilité est fort complexe. Par conséquent, nous considérons l'approximation de la probabilité de sélection conditionnelle.

Sous l'hypothèse que $\bar{\mathbf{x}}_{\mathrm{HT}}$ suit une loi approximativement normale inconditionnellement et conditionnellement à la présence de chaque unité dans l'échantillon, la probabilité de sélection inconditionnelle (13) peut être approximée par

$$\hat{\pi}_{i|\bar{x}_{\mathrm{HT}}} = \pi_i \left|\sum_{\bar{x}\bar{x}}\right|^{1/2}\left|\sum_{\bar{x}\bar{x},(i)}\right|^{-1/2} \exp\{0{,}5\,(\mathbf{G}_{\bar{x}\bar{x}} - \mathbf{G}_{\bar{x}\bar{x},(i)})\}, \qquad (14)$$

où $\sum_{\bar{x}\bar{x}} = \mathrm{Var}\{\bar{x}_{\mathrm{HT}} \mid F\}$, $\sum_{\bar{x}\bar{x},(i)} = \mathrm{Var}\{\bar{\mathbf{x}}_{\mathrm{HT}} \mid F,\ i \in A\}$,

$$\mathbf{G}_{\bar{x}\bar{x}} = (\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)\sum_{\bar{x}\bar{x}}^{-1}(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)',$$

$$\mathbf{G}_{\bar{x}\bar{x},(i)} = (\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_{N,(i)})\sum_{\bar{x}\bar{x},(i)}^{-1}(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_{N,(i)})',$$

$$\bar{\mathbf{x}}_{N,(i)} = E\{\bar{\mathbf{x}}_{\mathrm{HT}} \mid F,\ i \in A\} = (N\pi_i)^{-1}\mathbf{x}_i + N^{-1}\sum_{\substack{j=1\\ j\neq i}}^{N}(\pi_i\,\pi_j)^{-1}\pi_{ij}\,\mathbf{x}_j,$$

A est l'ensemble d'indices qui apparaissent dans l'échantillon et $F = \{y_1, ..., y_N\}$ est la population finie. Tillé (1998) donne une expression pour $\sum_{\bar{x}\bar{x},(i)}$ pour le cas général.

Supposons que les matrices des covariances de plan de sondage $\sum_{\bar{x}\bar{x}}$ et $\sum_{\bar{x}\bar{x},(i)}$ sont définies positives et que le vecteur de variables auxiliaires suit une loi normale. Tillé (1999) montre que l'estimateur SCP définit en (12) avec les probabilités de sélection conditionnelles approximatives données par (14) satisfait

$$\bar{y}_{p\pi} = \bar{y}_{\mathrm{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\,\boldsymbol{\beta}_N + O_p(n^{-1}) \qquad (15)$$

$$= \bar{y}_{\mathrm{reg}} + O_p(n^{-1}), \qquad (16)$$

où

$$\boldsymbol{\beta}_N = \sum_{\bar{x}\bar{x}}^{-1}\sum_{\bar{x}\bar{y}},$$

$$\bar{y}_{\mathrm{reg}} = \bar{y}_{\mathrm{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\,\hat{\boldsymbol{\beta}},$$

$$\hat{\boldsymbol{\beta}} = \hat{\sum}_{\bar{x}\bar{x}}^{-1}\hat{\sum}_{\bar{x}\bar{y}} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y},$$

$\mathbf{X} = (\mathbf{x}_1', ..., \mathbf{x}_n')'$, $\mathbf{y} = (y_1, ..., y_n)'$, le ij^{e} élément de $\boldsymbol{\Phi}^{-1}$ est $N^{-2}(\pi_{ij}\,\pi_i\,\pi_j)^{-1}(\pi_{ij} - \pi_i\,\pi_j)$, $\sum_{\bar{x}\bar{x}}$ est la variance par rapport au plan de sondage de $\bar{\mathbf{x}}_{\mathrm{HT}}$, $\sum_{\bar{x}\bar{y}}$ est la covariance par rapport au plan de sondage de $\bar{\mathbf{x}}_{\mathrm{HT}}$ et $\bar{y}_{\mathrm{HT}}$, $\hat{\sum}_{\bar{x}\bar{x}}$ est l'estimateur d'Horvitz-Thompson de la variance de $\bar{\mathbf{x}}_{\mathrm{HT}}$, et $\hat{\sum}_{\bar{x}\bar{y}}$ est l'estimateur d'Horvitz-Thompson de la covariance de $\bar{\mathbf{x}}_{\mathrm{HT}}$ et $\bar{y}_{\mathrm{HT}}$.

Dans le cas d'un plan de sondage complexe, un certain nombre de quantités figurant dans (14) sont difficiles à calculer. Cependant, les approximations des estimateurs donnant les mêmes propriétés en grand échantillon sont assez faciles à calculer. Nous remplaçons $\sum_{\bar{x}\bar{x}}$ et $\sum_{\bar{x}\bar{x},(i)}$ par les estimateurs, nous remplaçons $\bar{\mathbf{x}}_{N,(i)}$ par $\bar{\mathbf{x}}_N + \mathbf{d}_{x_i}$, nous définissons

$$\hat{\mathbf{M}}_{\bar{x}\bar{y}} = \sum_{i\in A}(N\pi_i)^{-1} + \mathbf{d}_{x_i}'\,y_i, \qquad (17)$$

et nous supposons que

$$\mathrm{Var}\{n(\hat{\mathbf{M}}_{\bar{x}\bar{y}} - \mathbf{M}_{\bar{x}\bar{y}})\} = O(n^{-1}), \qquad (18)$$

$$\mathbf{d}_{x_i} = O_p(n^{-1}), \qquad (19)$$

où $\mathbf{d}_{x_i}$ est une fonction de l'échantillon et $\mathbf{M}_{\bar{x}\bar{y}}$ est une quantité de population. Souvent, $\mathbf{M}_{\bar{x}\bar{y}}$ est la matrice des covariances de population $\sum_{\bar{x}\bar{y}}$, mais cette égalité n'est pas nécessaire pour que l'estimateur soit bien défini. Dans de nombreux cas, on peut calculer $\mathbf{d}_{x_i}$ sous forme d'un multiple de l'écart jackknife. En outre, dans de nombreuses situations, une valeur adéquate de l'estimateur, $\tilde{\sum}_{\bar{x}\bar{x},(i)}$, de $\sum_{\bar{x}\bar{x},(i)}$ est $n^{-1}(n-1)\hat{\sum}_{\bar{x}\bar{x}}$. Nous écrivons notre généralisation de (14) comme suit

$$\tilde{\pi}_{i|\bar{x}_{\mathrm{HT}}} = \pi_i \left|\hat{\sum}_{\bar{x}\bar{x}}\right|^{1/2}\left|\tilde{\sum}_{\bar{x}\bar{x},(i)}\right|^{-1/2} \exp\{0{,}5\,(\hat{\mathbf{G}}_{\bar{x}\bar{x}} - \tilde{\mathbf{G}}_{\bar{x}\bar{x},(i)})\}, \qquad (20)$$

où

$$\hat{\mathbf{G}}_{\bar{x}\bar{x}} = (\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)\,\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}}^{-1}\,(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)',$$

$$\tilde{\mathbf{G}}_{\bar{x}\bar{x},(i)} = (\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N - \mathbf{d}_{x_i})\,\tilde{\boldsymbol{\Sigma}}_{\bar{x}\bar{x},(i)}^{-1}\,(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N - \mathbf{d}_{x_i})'.$$

Posons que l'estimateur (12) construit avec les $\tilde{\pi}_{i\,|\,\bar{x}_{\mathrm{HT}}}$ de (20) est

$$\bar{y}_{p\tilde{\pi}} = N^{-1} \sum_{i=1}^{n} \tilde{\pi}_{i\,|\,\bar{x}_{\mathrm{HT}}}^{-1}\, y_i. \tag{21}$$

La probabilité de sélection conditionnelle approximative pour un échantillonnage aléatoire simple et une seule variable auxiliaire est donnée par

$$\tilde{\pi}_{i\,|\,\bar{x}_n} = \frac{n}{N}\left[\frac{\hat{\sigma}_{\bar{x}}}{\tilde{\sigma}_{\bar{x},(i)}}\right] \exp\left\{\frac{1}{2}\left[\frac{(\bar{x}_n - \bar{x}_N)^2}{\hat{\sigma}_{\bar{x}}^2} - \frac{(\bar{x}_n - \bar{x}_N - d_{x_i})^2}{\tilde{\sigma}_{\bar{x},(i)}^2}\right]\right\},$$

où

$$d_{x_i} = [\,n(N-1)\,]^{-1}(N-n)\,(x_i - \bar{x}_N),$$

$$\tilde{\sigma}_{\bar{x},(i)}^2 = \frac{(N-n)\,(n-1)}{n^2\,(N-2)}\left[s_x^2 - \frac{N\,(x_i - \bar{x}_n)^2}{(N-1)^2}\right] \approx \frac{n-1}{n}\,\hat{\sigma}_{\bar{x}}^2,$$

$$\hat{\sigma}_{\bar{x}}^2 = (n^{-1} - N^{-1})\,s_x^2,$$

et

$$s_x^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2.$$

Dans ce cas, $d_{x_i} = \bar{x}_{N,(i)} - \bar{x}_N$ et $M_{\bar{x}\bar{y}} = \mathrm{Cov}\,(\bar{x}_{\mathrm{HT}},\, \bar{y}_{\mathrm{HT}})$.

L'estimateur SCP (21) obtenu avec les probabilités de sélection conditionnelles approximatives n'est pas calé; autrement dit l'estimateur (21) de la moyenne du vecteur de variables auxiliaires n'est pas le vecteur de moyennes de population. Il est assez facile de normaliser les probabilités de sorte que leur somme soit égale à l'unité ou à la fraction de strate en cas d'échantillonnage stratifié. Pour construire un estimateur calé pour le cas général, nous proposons de calculer l'estimateur par la régression avec $[\sum_{j=1}^{n} \tilde{\pi}_{j\,|\,\bar{x}_{\mathrm{HT}}}^{-1}]^{-1}\,\tilde{\pi}_{i\,|\,\bar{x}_{\mathrm{HT}}}^{-1}$ comme poids initiaux. L'estimateur proposé est

$$\bar{y}_{w\mathrm{reg}} = \bar{y}_c + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_c)\,\hat{\boldsymbol{\beta}}_{c,1} = \sum_{i=1}^{n} w_i\, y_i, \tag{22}$$

où

$$(\bar{y}_c,\, \bar{\mathbf{x}}_c) = \sum_{i=1}^{n} \alpha_i (y_i,\, \mathbf{x}_i),$$

$$\left(\hat{\beta}_{c,0},\, \hat{\boldsymbol{\beta}}_{c,1}\right)' = \left[\sum_{i=1}^{n} \alpha_i\, \mathbf{z}_i'\, \mathbf{z}_i\right]^{-1}\left[\sum_{i=1}^{n} \alpha_i\, \mathbf{z}_i'\, y_i\right],$$

$$\mathbf{z}_i = (1,\, \mathbf{x}_i - \bar{\mathbf{x}}_c),$$

$$\alpha_i = \left[\sum_{j=1}^{n} \tilde{\pi}_{j\,|\,\bar{x}_{\mathrm{HT}}}^{-1}\right]^{-1} \tilde{\pi}_{i\,|\,\bar{x}_{\mathrm{HT}}}^{-1},$$

$$w_i = \alpha_i + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_c)\left[\sum_{j=1}^{n} \alpha_j\,(\mathbf{x}_j - \bar{\mathbf{x}}_c)'(\mathbf{x}_j - \bar{\mathbf{x}}_c)\right]^{-1} \alpha_i\,(\mathbf{x}_i - \bar{\mathbf{x}}_c)',$$

et $\tilde{\pi}_{i\,|\,\bar{x}_{\mathrm{HT}}}$ est la probabilité de sélection conditionnelle approximative donnée par (20). Nous supposons que le vecteur de variables auxiliaires contient la valeur un si bien que l'estimateur ne varie pas en fonction de la localisation.

L'estimateur (21) est approximativement égal à un estimateur par la régression et l'estimateur (22) est, lui aussi approximativement égal au même estimateur par la régression.

Théorème : Soit une série de populations et d'échantillons, $\{F_N,\, A_N\}$, satisfaisant

$$(\bar{y}_{\mathrm{HT}},\, \bar{\mathbf{x}}_{\mathrm{HT}}) - (\bar{y}_N,\, \bar{\mathbf{x}}_N) = O_p\,(n^{-1/2}). \tag{23}$$

Supposons que les séries de matrices des covariances estimées, $\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}}$ et $\tilde{\boldsymbol{\Sigma}}_{\bar{x}\bar{x},(i)}$, satisfont

$$\left[\mathbf{D}^{-1/2}\,\tilde{\boldsymbol{\Sigma}}_{\bar{x}\bar{x},(i)}\,\mathbf{D}^{-1/2}\right]^{-1} - \left[\mathbf{D}^{-1/2}\,\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}}\,\mathbf{D}^{-1/2}\right]^{-1} = O_p\,(n^{-1}), \tag{24}$$

où $\mathbf{D}$ représente une matrice diagonale ayant sur sa diagonale les éléments de la diagonale de $\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}}$. Soit $\mathbf{d}_{x_i}$ une fonction de l'échantillon satisfaisant (19) et supposons que (18) est vérifiée. Supposons que la série d'estimateurs de la variance d'Horvitz-Thompson satisfait

$$\mathrm{Var}\left\{n\left[\mathrm{Vech}\left(\hat{\boldsymbol{\Sigma}}_{\bar{z}\bar{z},\mathrm{HT}} - \boldsymbol{\Sigma}_{\bar{z}\bar{z}}\right)\right]\right\} = O(n^{-1}), \tag{25}$$

où $\mathbf{z}_i = (\mathbf{x}_i,\, y_i)$ et $\boldsymbol{\Sigma}_{\bar{z}\bar{z}}$ est définie positive. Supposons que $E\{\tilde{\pi}_{i\,|\,\bar{x}_{\mathrm{HT}}}^{-2}\}$ est bornée, où $\tilde{\pi}_{i\,|\,\bar{x}_{\mathrm{HT}}}$ est définie dans (20). Alors, l'estimateur SCP $\bar{y}_{p\tilde{\pi}}$ de (21) satisfait

$$\bar{y}_{p\tilde{\pi}} = \bar{y}_{\mathrm{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\,\boldsymbol{\theta}_N + O_p(n^{-1}) = \bar{y}_{\mathrm{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\,\hat{\boldsymbol{\theta}} + O_p(n^{-1}),$$

où $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}}^{-1}\,\hat{\mathbf{M}}_{\bar{x}\bar{y}}$ et $\boldsymbol{\theta}_N = \boldsymbol{\Sigma}_{\bar{x}\bar{x}}^{-1}\,\mathbf{M}_{\bar{x}\bar{y}}$.

Si $\mathrm{Var}\{\sum_{i=1}^{n} \pi_i^{-1}\} > 0$, supposons que $\mathbf{x}_i$ contient la valeur un comme élément. Supposons que $\mathbf{M}_{\bar{x}\bar{y}} = \boldsymbol{\Sigma}_{\bar{x}\bar{y}}$. Alors, l'estimateur par la régression pondérée de (22) satisfait

$$\bar{y}_{w\mathrm{reg}} = \bar{y}_{\mathrm{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\,\hat{\boldsymbol{\theta}} + O_p(n^{-1}).$$

Pour la preuve, voir l'annexe.

Afin d'illustrer la nature des divers types de poids de régression, nous avons tiré un échantillon aléatoire simple de taille 40 à partir d'une population normale de moyenne nulle et de variance égale à un. La moyenne d'échantillon est – 0,614 et la moyenne de population est nulle. Les poids de l'estimateur par la régression sont donnés par (2) avec

$\alpha_i = \phi_{ii}^{-1} = n^{-1}$. Les poids de l'estimateur par la MIQ et de l'EMV sont obtenus en minimisant les fonctions objectives (5) et (7), respectivement, sous la contrainte (4). Les poids pour l'estimateur par la régression pondérée SCP sont donnés par (22). Les poids sont représentés graphiquement en fonction des valeurs x d'échantillon à la figure 1. Cinq des poids de régression simple sont négatifs, à cause de l'écart important entre les moyennes d'échantillon et de population. Pour l'estimateur par la régression pondérée SCP, l'EMV et la MIQ, tous les poids sont non négatifs. La figure 1 montre que les comportements des poids obtenus par la MIQ et par la régression pondérée SCP sont semblables et que l'EMV produit un poids extrêmement grand dans cet échantillon.

Le tableau 1 contient certains poids pour les valeurs de x les plus petites, les valeurs de x proches de la moyenne d'échantillon, les valeurs de x proches de la moyenne de population et les valeurs de x les plus grandes. Pour les valeurs de x les plus éloignées de la moyenne de population, l'EMV donne les poids les plus grands. Pour les valeurs de x proches de la moyenne d'échantillon, les poids de la régression par les moindres carrés ordinaires sont proches de n^{-1}, tandis que les autres poids sont inférieurs à n^{-1}. Pour les valeurs de x proches de la moyenne de population, les poids de l'EMV sont proches de n^{-1}, tandis que les autres poids sont grands.

Étude par simulation

Afin de comparer les diverses méthodes de construction des poids de régression, nous avons réalisé un étude par simulation. En tout, nous avons sélectionné 30 000 échantillons aléatoires simples de taille 32 à partir d'une loi χ^2 à deux degrés de liberté. Les paramètres estimés sont ceux du mécanisme de génération de données infini. Soit x_i la valeur du i^{e} élément échantillonné. Nous avons considéré six méthodes d'estimation.

1. Régression par les moindres carrés ordinaires (MCO)
2. Programmation quadratique avec bornes supérieure et inférieure (PQ)
3. Régression pondérée avec poids SCP (Rég. SCP)
4. Fonction objective pour le maximum de vraisemblance (EMV)
5. Fonction objective pour la MIQ (Rég. MIQ))
6. Procédure logit avec bornes supérieure et inférieure (Logit)

Tableau 1
Certains poids de régression pour l'exemple illustré

x	Poids multipliés par $n = 40$			
	Rég.	Rég. pond.	MIQ	EMV
– 2,103	– 0,56	0,12	0,16	0,40
– 1,941	– 0,40	0,12	0,20	0,40
– 1,727	– 0,16	0,20	0,24	0,44
– 0,710	0,88	0,68	0,68	0,68
– 0,670	0,96	0,72	0,68	0,68
– 0,468	1,16	0,88	0,84	0,76
– 0,103	1,52	1,28	1,24	0,92
0,021	1,68	1,44	1,40	1,00
0,097	1,76	1,56	1,52	1,08
0,628	2,32	2,60	2,60	1,84
0,662	2,36	2,68	2,72	1,92
1,237	2,96	4,60	4,88	9,12

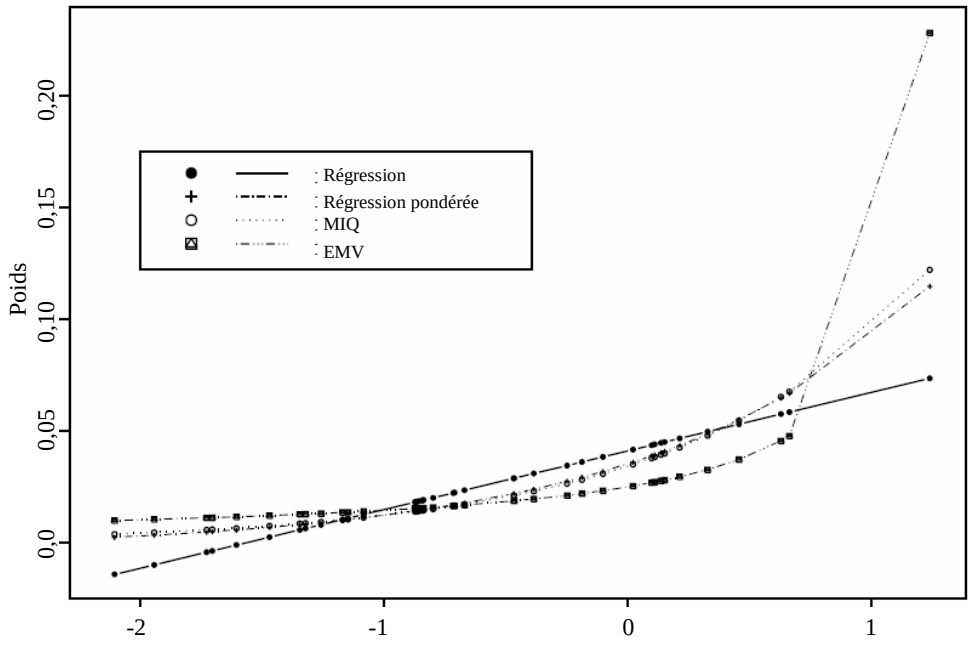

Figure 1. Comparaison de quatre ensembles de poids.

Nous avons calculé les poids pour l'estimateur par les MCO au moyen de (2) avec $\alpha_i = n^{-1}$. Les poids pour la programmation quadratique minimise $\sum_{i=1}^{n} w_i^2$ sous la contrainte $0 \le w_i \le 0{,}065$ pour tout i et sous les contraintes (4). La procédure de programmation quadratique est équivalente à la méthode linéaire tronquée du cas 7 de Deville et Särndal (1992). Les poids pour la régression pondérée SCP ont été calculés en minimisant $\sum_{i=1}^{n} \alpha_i^{-1} w_i^2$ sous les contraintes (4), où α_i est défini dans (22). Pour la MIQ et le maximum de vraisemblance, les poids ont été obtenus en minimisant les fonctions objectives (5) et (7), respectivement, sous les contraintes (4). Les poids calculés par la procédure logit minimisent la fonction $\sum_{i=1}^{n} G(nw_i)$ sous les contraintes (4), où

$$G(nw_i) = a^{-1}\left[(nw_i)\ln(nw_i) + (u - nw_i)\ln\left(\frac{u - nw_i}{u-1}\right)\right],$$

si $0 < nw_i < u$ et ∞ ailleurs, $a = u(u-1)^{-1}$, et $u = 2{,}08$. Notons que la solution pour la procédure logit, si elle existe, satisfait les contraintes de bornes $0 \le w_i \le 0{,}065$ pour tout i. La procédure logit a été introduite en tant que cas 6 dans Deville et Särndal (1992). Nous avons utilisé 0,065 comme borne supérieure du poids, de sorte que 3 026 échantillons (environ 10 %) aient au moins un poids de régression par la MIQ supérieur à 0,065. Pour 99 des 30 000 échantillons, aucune solution n'est possible pour la programmation quadratique et la procédure logit, parce qu'aucun point faisable ne satisfait (4) et la contrainte de borne. Pour ces 99 échantillons, nous avons utilisé le maximum des poids de régression pour les MCO comme borne supérieure pour la programmation quadratique et la procédure logit.

Le tableau 2 montre la moyenne de la somme des carrés pour les six types de poids. Le poids moyen est 1/32 = 0,03125 pour chaque estimateur. Les procédures par les moindres carrés sont celles dont la somme des carrés des poids est la plus faible, parce qu'il s'agit de la fonction objective pour ces procédures. Les procédures par les moindres carrés ont aussi une fourchette de somme des carrés un peu plus petite. Un pourcent des échantillons traités par les moindres carrés ont une moyenne normalisée des carrés supérieure à 1,401, tandis que 1 % des échantillons traités par la MIQ ont une moyenne des carrés supérieure à 1,441.

Tableau 2
Moyenne de Monte Carlo de la somme des carrés des poids

	MCO	PQ	Rég. SCP	EMV	Rég. MIQ	Logit
Moyenne de $\mathbf{w}'\mathbf{w}$ (×32)	1,043	1,044	1,045	1,053	1,045	1,045

Le tableau 3 donne les propriétés du minimum des poids. La méthode du maximum de vraisemblance est celle pour laquelle le poids minimum moyen est le plus grand, tandis que les procédures par les moindres carrés ont un poids minimum moyen plus faible. La variance du poids minimum la plus grande est celle calculée pour les procédures par les moindres carrés ordinaires. Notons que la programmation quadratique permet que les poids soient égaux à la borne inférieure nulle.

Tableau 3
Moyenne, variance et quantiles de Monte Carlo du poids minimum

			Quantiles (×32)				
Procédure	Moyenne ($\times 10^2$)	Variance ($\times 10^5$)	0,01	0,10	0,50	0,90	0,99
MCO	2,22	6,46	– 0,10	0,34	0,79	0,96	1,00
PQ	2,21	6,32	0,00	0,32	0,79	0,96	1,00
Rég. SCP	2,44	3,58	0,22	0,49	0,84	0,97	0,99
EMV	2,45	2,79	0,33	0,52	0,83	0,97	1,00
Rég. MIQ	2,36	3,81	0,20	0,45	0,81	0,97	1,00
Logit	2,25	5,23	0,09	0,36	0,78	0,96	1,00

Parmi les procédures sans contraintes de borne sur les poids, celle des moindres carrés ordinaires produit le poids maximum le plus petit en moyenne et une variance nettement plus faible pour le maximum. Voir le tableau 4. La régression pondérée SCP donne une plus petite fraction de poids très grands que l'EMV ou la MIQ, mais une fraction plus élevée de grands poids que la méthode des moindres carrés ordinaires. La PQ bornée et les procédures logit donnent une moyenne et une variance plus faibles du poids maximum que les procédures sans contrainte sur la borne supérieure.

Tableau 4
Moyenne, variance et quantiles de Monte Carlo du poids maximum

			Quantiles (×32)				
Procédure	Moyenne ($\times 10^2$)	Variance ($\times 10^5$)	0,01	0,10	0,50	0,90	0,99
MCO	4,25	17,35	1,00	1,03	1,20	1,92	2,93
PQ	4,17	11,91	1,00	1,03	1,20	1,92	2,08
Rég. SCP	4,56	26,42	1,03	1,07	1,27	2,12	3,47
EMV	4,75	56,13	1,00	1,04	1,25	2,31	4,72
Rég. MIQ	4,46	30,25	1,00	1,03	1,23	2,09	3,63
Logit	4,13	10,23	1,00	1,03	1,21	1,82	2,08

Pour évaluer les propriétés des procédures quand le modèle linéaire ne tient pas, nous considérons l'estimation des percentiles de la fonction de distribution de x. Le tableau 5 donne le biais de Monte Carlo des estimateurs des percentiles où les entrées du tableau sont

$$[\min\{P, (1-P)\}]^{-1}[\hat{E}\{\hat{P}\} - P] \times 100,$$

et P est le percentile. Par exemple, le biais relatif estimé de Monte Carlo dans l'estimateur par les moindres carrés ordinaires du percentile 0,01 est – 7,75 %. Parmi les procédures sans contrainte sur les bornes, l'estimateur par

les moindres carrés ordinaires est celui qui produit le biais le plus important dans l'estimation des percentiles de population. L'EMV donne le biais le plus faible pour tous les percentiles, sauf les 75^{e}, 95^{e} et 99^{e}, pour lesquels l'estimateur par la régression pondérée SCP produit le biais le plus faible. Un grand nombre d'échantillons de taille 32 ne contiennent aucune observation supérieure au 99^{e} percentile. Les procédures PQ et logit produisent un biais plus important que les autres, sauf pour le 75^{e} percentile. Le biais de ces procédures est assez important pour les percentiles inférieurs.

Le tableau 6 donne l'EQM relative des estimateurs des percentiles où les entrées du tableau sont

$$[\min\{P,\ (1-P)\}]^{-2}[\hat{E}\{\hat{P}-P\}^2]\times 100.$$

Donc, l'erreur quadratique moyenne relative de l'estimateur par les MCO du percentile 0,01 est de 283,27 %. Alors que cet estimateur du percentile 0,01 possède le biais le plus important, il donne l'erreur quadratique moyenne la plus faible parmi les procédures sans contrainte de borne. Les procédures PQ, MCO et logit donnent de meilleurs résultats pour les percentiles extrêmes, tandis que les autres sont meilleures pour les percentiles du milieu.

Tableau 5
Biais normalisé de Monte Carlo dans les estimateurs des percentiles

Percentile	Procédure					
	MCO	PQ	Rég. SCP	EMV	Rég. MIQ	Logit
0,01	– 7,75	– 8,43	– 2,88	– 2,13	– 4,70	– 8,30
0,05	– 7,27	– 7,95	– 2,58	– 1,82	– 4,30	– 7,85
0,10	– 6,66	– 7,31	– 2,27	– 1,57	– 3,91	– 7,26
0,25	– 5,25	– 5,82	– 1,79	– 1,25	– 3,13	– 5,89
0,50	– 3,21	– 3,46	– 1,37	– 1,16	– 2,18	– 3,53
0,75	– 2,30	– 2,07	– 1,60	– 2,21	– 2,25	– 1,78
0,90	4,60	5,31	1,27	0,22	2,62	5,68
0,95	12,75	13,33	6,01	6,41	9,52	13,15
0,99	32,94	32,36	19,03	22,66	26,65	30,03

Tableau 6
EQM relative de Monte Carlo des estimateurs des percentiles

Percentile	Procédure					
	MCO	PQ	Rég. SCP	EMV	Rég. MIQ	Logit
0,01	283,27	282,50	30,.23	311,58	296,37	282,76
0,05	53,91	54,23	57,41	57,07	54,97	54,06
0,10	25,50	25,97	26,40	25,79	25,26	25,80
0,25	8,00	8,41	7,77	7,23	7,42	8,41
0,50	1,99	2,07	1,88	1,71	1,83	2,12
0,75	3,65	3,68	3,62	3,66	3,63	3,67
0,90	14,50	14,60	14,25	14,57	14,36	14,56
0,95	39,40	38,65	40,99	41,66	39,93	37,94
0,99	200,17	196,24	235,71	216,22	205,85	194,33

Dans 562 des 30 000 échantillons, au moins un des poids de régression par les MCO est négatif. Dans 17 échantillons, au moins un des poids de régression SCP originaux était négatif. L'utilisation de la programmation quadratique avec la fonction objective pour les MCO (PQ) en vue de produire des poids égaux ou supérieurs à 0 et inférieur à 0,065 accroît la somme moyenne des carrés de moins de 1 %. Voir le tableau 7. L'utilisation de la programmation quadratique pour imposer aux poids de régression SCP (SCP (PQL)) la borne zéro augmente très peu la somme moyenne des carrés, car seul un très petit nombre de poids sont modifiés.

Tableau 7
Moyenne de Monte Carlo de la somme des carrés des poids pour les échantillons ayant au moins un poids MCO négatif

	MCO	PQ	Rég. SCP SPP	SCP (PQL)	EMV	Rég. MIQ
Moyenne de $\mathbf{w}'\mathbf{w}$ (×32)	1,208	1,217	1,226	1,227	1,342	1,242

Le tableau 8 donne l'EQM de Monte Carlo pour les 562 échantillons ayant des poids par les moindres carrés ordinaires négatifs. La programmation quadratique est supérieure aux autres procédures donnant des poids non négatifs pour le percentile 0,01 et inférieure pour le percentile 0,99. Parmi les 562 échantillons, 497 avaient une moyenne d'échantillon supérieure à la moyenne de population. Rappelons que la population étudiée suit une loi exponentielle. Comme le poids de l'observation la plus grande est nul dans les 497 échantillons, l'erreur de l'estimateur par programmation quadratique du percentile 0,99 est de 100 % pour la plupart des 497 échantillons ayant une moyenne d'échantillon supérieure à la moyenne de population. Si l'on échantillonnait une population finie, la borne sur les poids serait égale ou supérieure à N^{-1} et l'EQM de la programmation quadratique pour le percentile 0,99 serait réduite.

Tableau 8
EQM relative de Monte Carlo des estimateurs des percentiles pour les échantillons ayant au moins un poids MCO négatif

Percentile	Procédure				
	MCO	PQ	SCP (PQL)	EMV	Rég. MIQ
0,01	287,52	291,11	350,58	461,80	344,06
0,05	76,04	70,58	75,80	88,71	72,50
0,10	44,80	40,74	39,31	38,84	36,05
0,25	20,24	19,14	14,72	9,91	12,56
0,50	5,03	5,31	3,65	2,26	3,35
0,75	5,02	4,53	3,36	4,24	3,45
0,90	23,77	23,69	20,04	18,80	20,49
0,95	51,54	46,04	30,79	28,28	32,54
0,99	206,33	90,08	39,40	57,54	43,49

La programmation quadratique est supérieure aux autres procédures calées pour le percentile 0,01 dans les échantillons avec poids par les MCO négatifs. La régression par la MIQ et la régression pondérée SCP sont supérieures à l'EMV pour les percentiles 0,01 et 0,05. Il en est ainsi parce que l'EMV produit souvent le poids maximal le plus grand.

Pour 3 026 des 30 000 échantillons, au moins un des poids de régression par la MIQ est supérieur à 0,065. Pour

2 152 échantillons, au moins un des poids de régression par les MCO est supérieur à 0,065, et pour 3 209 échantillons, au moins un des poids de régression SCP est supérieur à 0,065. L'utilisation de la programmation quadratique avec la fonction objective des MCO pour produire des poids dans l'intervalle (0,000, 0,065) fait augmenter la somme moyenne des carrés de 1,5 %. L'utilisation de la programmation quadratique pour donner aux poids de régression SPC les bornes 0,000 et 0,065 fait augmenter la somme moyenne des carrés de 0,8 %. Voir la colonne SCP (PQ) du tableau 9.

Tableau 9
Moyenne de Monte Carlo de la somme des carrés des poids pour les échantillons dont au moins un poids de régression par la MIQ est supérieur à 0,065

	MCP	PQ	Rég. SCP.	SCP (PQ)	MIQ	Rég. Logit	EMV
Moyenne de $\mathbf{w}'\mathbf{w}$ ($\times 32$)	1,210	1,228	1,221	1,231	1,228	1,232	1,290

Le tableau 10 donne l'EQM relative de Monte Carlo pour les 3 026 échantillons contenant des poids de régression par la MIQ supérieurs à 0,065. La programmation quadratique donne de meilleurs résultats que les procédures SCP (PQ) et Logit pour les percentiles 0,01, 0,95 et 0,99, et la procédure logit est supérieure à la programmation quadratique pour les autres percentiles.

Tableau 10
EQM relative de Monte Carlo des estimateurs des percentiles pour les échantillons dont au moins un poids de régression par la MIQ est supérieur à 0,065

Percentile	Procédure						
	MCO	PQ	Rég. SCP	SCP (PQ)	Rég. MIQ	Logit	EMV
0,01	139,96	130,53	173,86	146,40	124,02	173,65	206,65
0,05	39,83	42,88	39,35	41,69	39,87	37,14	40,83
0,10	26,31	30,92	22,40	28,10	28,88	20,21	19,98
0,25	13,56	17,72	10,13	15,69	17,71	8,65	7,01
0,50	3,95	4,87	3,32	4,75	5,37	3,03	2,28
0,75	4,84	5,35	4,89	5,58	5,37	5,05	5,48
0,90	27,98	29,04	28,70	29,34	29,32	28,79	32,07
0,95	74,15	67,54	85,02	68,12	65,98	83,13	95,99
0,99	198,77	179,58	219,16	181,17	172,45	212,38	226,73

Discussion

Nous avons entrepris l'étude en conjecturant que démarrer avec les poids SCP dans une estimateur par la régression produirait des poids finaux presque toujours positifs et ayant des propriétés désirables, telles que mesurées par la capacité d'estimer la fonction de distribution de x. Nos résultats appuient dans une certaine mesure cette conjecture. Les poids minimaux de la régression SCP sont plus grands que ceux de la régression par les MCO et comparables à ceux de la régression par raking. La programmation quadratique peut être utilisée pour éliminer les poids négatifs dans les quelques échantillons concernés. Si l'on n'impose pas de borne supérieure, les poids maximaux pour la régression pondérée SPC se situent entre ceux obtenus pour les moindres carrés et la MIQ.

Il est connu que toutes les procédures utilisées dans notre étude par simulation ont les mêmes propriétés d'ordre $n^{-1/2}$. Notre simulation et l'étude des procédures de la MIQ généralisé réalisée par Deville et coll. (1993) indiquent qu'il existe aussi des différences modestes dans le cas des petits échantillons. Aucune procédure n'est meilleure que les autres en ce qui concerne l'ensemble des critères. À cause des résultats médiocres pour les percentiles extrêmes, nous déconseillons d'utiliser la fonction objective de l'EMV. La programmation quadratique et la procédure logit produisent, pour les percentiles extrêmes, une somme des carrés des poids, un poids maximal et une EQM marginalement plus faibles que la régression par la MIQ. L'EMV, la régression SCP et les procédures de la MIQ produisent des poids minimaux marginalement plus grands et une EQM marginalement plus faible pour les percentiles du milieu de la distribution de x que la programmation quadratique et la procédure logit. Les propriétés de la programmation quadratique et de la procédure logit sont comparables en ce qui concerne l'estimation de la fonction de distribution de x.

Annexe

Preuve. Le ratio des déterminants des matrices des covariances estimées dans (20) est

$$\frac{\left|\tilde{\tilde{\boldsymbol{\Sigma}}}_{\bar{x}\bar{x},(i)}\right|}{\left|\hat{\boldsymbol{\Sigma}}_{\bar{x}\bar{x}}\right|} = 1 + O_p(n^{-1}) \qquad (26)$$

en vertu des hypothèses (24) et (25). La différence $\tilde{\tilde{\mathbf{G}}}_{\bar{x}\bar{x},(i)} - \hat{\mathbf{G}}_{\bar{x}\bar{x}}$ est

$$(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)\left(\tilde{\tilde{\boldsymbol{\Sigma}}}^{-1}_{\bar{x}\bar{x},(i)} - \hat{\boldsymbol{\Sigma}}^{-1}_{\bar{x}\bar{x}}\right)(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)' - 2(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)\tilde{\tilde{\boldsymbol{\Sigma}}}^{-1}_{\bar{x}\bar{x},(i)}\mathbf{d}'_{x_i} + \mathbf{d}_{x_i}\tilde{\tilde{\boldsymbol{\Sigma}}}^{-1}_{\bar{x}\bar{x},(i)}\mathbf{d}'_{x_i}.$$

En vertu des hypothèses (23) et (24),

$$\exp\{0{,}5[(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)(\tilde{\tilde{\boldsymbol{\Sigma}}}^{-1}_{\bar{x}\bar{x},(i)} - \hat{\boldsymbol{\Sigma}}^{-1}_{\bar{x}\bar{x}})(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)']\} = 1 + O_p(n^{-1}). \qquad (27)$$

En utilisant les hypothèses (24) et (19), le développement en série de Taylor à $\mathbf{d}_{x_i} = 0$ donne

$$\begin{aligned}\exp[-(\bar{\mathbf{x}}_{\mathrm{HT}} - \bar{\mathbf{x}}_N)\tilde{\tilde{\boldsymbol{\Sigma}}}^{-1}_{\bar{x}\bar{x},(i)}\mathbf{d}'_{x_i} + 0.5\,\mathbf{d}_{x_i}\tilde{\tilde{\boldsymbol{\Sigma}}}^{-1}_{\bar{x}\bar{x},(i)}\mathbf{d}'_{x_i}] &= 1 + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\tilde{\tilde{\boldsymbol{\Sigma}}}^{-1}_{\bar{x}\bar{x},(i)}\mathbf{d}'_{x_i} + O_p(n^{-1}) \\ &= 1 + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\hat{\boldsymbol{\Sigma}}^{-1}_{\bar{x}\bar{x}}\mathbf{d}'_{x_i} + O_p(n^{-1}). \qquad (28)\end{aligned}$$

Donc, en vertu de (26), (27) et (28),

$$[N\tilde{\pi}_{i|x_{\mathrm{HT}}}]^{-1} = (N\pi_i)^{-1}[1+(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\,\hat{\boldsymbol{\Sigma}}^{-1}_{\tilde{x}\tilde{x}}\,\mathbf{d}'_{x_i}] + O_p(n^{-2}).$$

En vertu des hypothèses (18), (23) et (25), et en utilisant le fait que $E\{\tilde{\pi}^{-2}_{i|x_{\mathrm{HT}}}\}$ est bornée,

$$\begin{aligned}\bar{y}_{p\tilde{\pi}} &= \bar{y}_{\mathrm{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\,\hat{\boldsymbol{\theta}} + O_p(n^{-1}) \\ &= \bar{y}_{\mathrm{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\,\boldsymbol{\theta}_N + O_p(n^{-1}). \qquad (29)\end{aligned}$$

Si la valeur un est un élément de $\mathbf{x}_i$ ou que $\mathrm{Var}\{\sum_{i=1}^n \pi_i^{-1}\} = 0$, et si $\mathbf{M}_{\tilde{x}\tilde{y}} = \boldsymbol{\Sigma}_{\tilde{x}\tilde{y}}$, l'estimateur SCP de la moyenne de population du vecteur $\mathbf{q}_i = (1,\ \mathbf{x}_i)$ satisfait

$$\bar{\mathbf{q}}_{p\tilde{\pi}} = N^{-1}\sum_{i=1}^{n} \tilde{\pi}^{-1}_{i|\tilde{x}_{\mathrm{HT}}}\ \mathbf{q}_i = (1,\ \bar{\mathbf{x}}_N) + O_p(n^{-1}), \quad (30)$$

parce que le $\boldsymbol{\theta}$ pour $\mathbf{x}$ est la matrice identité. En vertu de (30),

$$\begin{aligned}(\bar{\mathbf{x}}_c,\ \bar{y}_c) &= N\left[\sum_{i=1}^{n} \tilde{\pi}^{-1}_{i|\tilde{x}_{\mathrm{HT}}}\right]^{-1} (\bar{\mathbf{x}}_{p\tilde{\pi}},\ \bar{y}_{p\tilde{\pi}}) \\ &= (\bar{\mathbf{x}}_{p\tilde{\pi}},\ \bar{y}_{p\tilde{\pi}}) + O_p(n^{-1}). \quad (31)\end{aligned}$$

Donc,

$$\begin{aligned}\bar{y}_{w\mathrm{reg}} &= \bar{y}_c + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_c)\,\hat{\boldsymbol{\beta}}_{c,1} \\ &= \bar{y}_{p\tilde{\pi}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{p\tilde{\pi}})\,\hat{\boldsymbol{\beta}}_{c,1} + (\bar{y}_c - \bar{y}_{p\tilde{\pi}}) + (\bar{\mathbf{x}}_{p\tilde{\pi}} - \bar{\mathbf{x}}_c)\,\hat{\boldsymbol{\beta}}_{c,1} \\ &= \bar{y}_{p\tilde{\pi}} + O_p(n^{-1}) \\ &= \bar{y}_{\mathrm{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\mathrm{HT}})\,\hat{\boldsymbol{\theta}} + O_p(n^{-1}),\end{aligned}$$

en vertu de (30), (31) et (29).

Remerciements

Cette étude a été financée aux termes de l'entente de coopération 43-3AEU-3-80088 entre la Iowa State University, le USDA National Agricultural Statistics Service et le U.S. Bureau of the Census, et aux termes de l'entente de coopération 68-3A75-14 entre le USDA Natural Resources Conservation Service et la Iowa State University. Nous remercions le rédacteur adjoint et les examinateurs de leurs commentaires qui nous ont permis d'améliorer l'article.

Bibliographie

Bardsley, P., et Chambers, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

Chen, J., et Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.

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

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

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Fuller, W.A. (2002). Estimation par regression appliqué à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

Huang, E.T., et Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section*, American Statistical Association, 300-305.

Husain, M. (1969). Construction of Regression Weights for Estimation in Sample Surveys. Thèse de maîtrise non-publiée, Iowa State University, Ames, Iowa.

Rao, J.N.K., et Singh, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the section on survey research methods*, American Statistical Association, 57-64.

Stephan, F.F. (1942). An alternative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.

Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. *Revue Internationale de Statistique*, 66, 303-322.

Tillé, Y. (1999). Estimation dans des enquêtes par sondage avec des probabilités d'inclusion conditionnelles : Enquêtes à plan d'échantillonnage complexe. *Techniques d'enquête*, 25, 61-71.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 103-107
Statistique Canada, N° 12-001-XIF au catalogue

Une distance de calage optimale menant à un estimateur par la régression optimal

Per Gösta Andersson et Daniel Thorburn [1]

Résumé

En échantillonnage, quand on dispose d'information auxiliaire, il est bien connu que l'« estimateur (par la régression) optimal » fondé sur le plan de sondage d'un total ou d'une moyenne de population finie est (du moins asymptotiquement) plus efficace que l'estimateur GREG correspondant. Nous illustrerons ce fait au moyen de simulations avec échantillonnage stratifié à partir de populations à distribution asymétrique. Au départ, l'estimateur GREG a été construit au moyen d'un modèle linéaire de superpopulation auxiliaire. Il peut aussi être considéré comme un estimateur par calage, c'est-à-dire un estimateur linéaire pondéré, où les poids obéissent à l'équation de calage et, sous cette contrainte, sont aussi proches que possible des « poids d'Horvitz-Thompson » originaux (d'après une mesure de distance appropriée). Nous montrons que l'estimateur optimal peut aussi être considéré comme un estimateur par calage à cet égard avec une mesure quadratique de distance étroitement liée à celle générant l'estimateur GREG. Nous donnons aussi des exemples simples révélant qu'il n'est pas toujours facile d'obtenir cette nouvelle mesure.

Mots clés : Estimateur d'Horvitz-Thompson; estimateur par la régression; théorie de l'échantillonnage.

1. Notation et notions élémentaires

Considérons une population finie U constituée de N objets étiquetés 1, ..., N avec les valeurs étudiées connexes $y_1, ..., y_N$ et les vecteur (colonnes) auxiliaires de dimension J $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$. Nous voulons estimer le total de population $t_y = \sum_{i \in U} y_i$ en tirant un échantillon aléatoire s de taille n (fixe ou aléatoire) à partir de U, avec les probabilités de sélection de premier et de deuxième ordre $\pi_i = P(i \in s), \pi_{ij} = P(i, j \in s), i, j = 1, ..., N$. Les valeurs étudiées et les vecteurs auxiliaires sont enregistrés pour les objets échantillonnés et nous supposons qu'au moins $\boldsymbol{t}_x = \sum_{i \in U} \boldsymbol{x}_i$ est connu avant que l'échantillon ne soit tiré.

Il s'agit des paramètres standard pour un estimateur par la régression. À la section 2, nous discutons de divers estimateurs par la régression, à savoir l'estimateur GREG ordinaire (Särndal, Swensson et Wretman 1992), l'estimateur optimal (Montanari 1987 et Andersson, Nerman et Westhall 1995) et les estimateurs par calage (Deville et Särndal 1992). Il est bien connu que l'estimateur GREG peut être obtenu sous forme d'un estimateur par calage. À la section 3, nous montrons que cela est également vrai pour l'estimateur optimal, mais avec une mesure de distance plus compliquée. Aux deux dernières sections, nous illustrons ceci et l'estimateur optimal, d'abord au moyen d'exemples théoriques, puis de simulations.

Enfin, suivent certains commentaires au sujet de la notation matricielle utilisée dans l'article. En général, la transposée d'une matrice $\boldsymbol{A}$ est notée $\boldsymbol{A}^T$ et, si $\boldsymbol{A}$ est carrée, l'inverse (inverse généralisée) s'écrit $\boldsymbol{A}^{-1}(\boldsymbol{A}^-)$. En outre, posons que les vecteurs colonnes $\boldsymbol{y} = (y_i)_{i \in s}$ et $\mathbf{w}_0 = (1/\pi_i)_{i \in s}$, $\boldsymbol{X}$ forme la matrice « du plan » de dimensions $J \times n$ de l'information auxiliaire tirée de s et, enfin, que $\boldsymbol{I}_n$ représente une matrice diagnonale unitaire de taille n.

2. Estimateurs par la régression et par calage

Un estimateur simple sans biais de t_y est l'estimateur d'Horvitz-Thompson $\hat{t}_y = \sum_{i \in s} y_i / \pi_i = \boldsymbol{y}^T \boldsymbol{w}_0$. Cependant, on peut obtenir des estimateurs plus efficaces en utilisant l'information auxiliaire, par exemple, l'estimateur GREG assisté par modèle bien connu (voir Särndal et coll. (1992). Par exemple, sous l'hypothèse d'un modèle de régression linéaire homoscédastique de superpopulation, l'estimateur GREG est

$$\hat{t}_{yr} = \boldsymbol{y}^T \boldsymbol{w}_0 + (\boldsymbol{y}^T \boldsymbol{R}_r \boldsymbol{X}^T)(\boldsymbol{X} \boldsymbol{R}_r \boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x) \qquad (1)$$

$$= \boldsymbol{y}^T \boldsymbol{g}, \qquad (2)$$

où $\boldsymbol{R}_r = \boldsymbol{w}_0 \boldsymbol{I}_n$, $\hat{\boldsymbol{t}}_x = \sum_{i \in s} \boldsymbol{x}_i / \pi_i$ et

$$\boldsymbol{g} = \left(\frac{1}{\pi_i} (1 + \boldsymbol{x}_i^T (\boldsymbol{X} \boldsymbol{R}_r \boldsymbol{X}^T)^{-1} (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)) \right)_{i \in s}.$$

Or, l'expression (2) de l'estimateur GREG est intéressante, puisque nous avons aussi l'expression

$$\boldsymbol{x}^T \boldsymbol{g} = \boldsymbol{t}_x, \qquad (3)$$

1. Per Gösta Andersson, Mathematical Statistics, Department of Mathematics, Linköping University, SE-581 83 Linköping, Suède; Daniel Thorburn, Department of Statistics, Stockholm University, SE-106 91 Stockholm, Suède.

qui est appelée équation de calage. Ceci nous amène à un autre moyen possible d'obtenir l'estimateur GREG, conformément à Deville et Särndal (1992). Supposons que nous cherchions un estimateur $\boldsymbol{y}^T\boldsymbol{w}$ de t_y avec un vecteur $\boldsymbol{w}$ de poids dépendant de l'échantillon $(w_i)_{i\in s}$, qui respecte l'équation de calage correspondante, tout en minimisant la distance entre $\boldsymbol{w}$ et $\boldsymbol{w}_0$ conformément à la mesure de distance quadratique

$$(\boldsymbol{w}-\boldsymbol{w}_0)^T\,\boldsymbol{R}\,(\boldsymbol{w}-\boldsymbol{w}_0),$$

où $\mathbf{R}=(\boldsymbol{w}_0\ \boldsymbol{I}_n)^{-1}$.

Ceci nous donne

$$\boldsymbol{w}=\boldsymbol{w}_0+\boldsymbol{R}^{-1}\boldsymbol{x}^T\,(\boldsymbol{X}\ \boldsymbol{R}^{-1}\boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x-\hat{\boldsymbol{t}}_x), \tag{4}$$

qui signifie que $\boldsymbol{w}=\boldsymbol{g}$, puisque ici $\boldsymbol{R}=\boldsymbol{R}_r^{-1}$.

En ce qui concerne l'estimateur optimal, considérons d'abord le vecteur $(\hat{t}_y,\ \hat{\boldsymbol{t}}_{\boldsymbol{x}}^{\boldsymbol{T}})$ et posons que $\sum_{y,x}$ est le vecteur (ligne) des covariances de $\hat{t}_y$ et $\hat{\boldsymbol{t}}_{\boldsymbol{x}}$, et que $\sum_{x,x}$ est la matrice des covariances de $\hat{\boldsymbol{t}}_{\boldsymbol{x}}$. Maintenant, l'estimateur linéaire sans biais (en $\hat{t}_y$ et $\hat{\boldsymbol{t}}_{\boldsymbol{x}}$) à variance minimale (voir Montanari 1987) de t_y est l'estimateur par différence

$$\hat{t}_y+\sum_{y,x}\sum_{x,x}^{-1}(\boldsymbol{t}_x-\hat{\boldsymbol{t}}_{\mathbf{x}}). \tag{5}$$

Puisqu'en pratique, $\sum_{y,x}$ et $\sum_{x,x}$ sont inconnus, posons que l'estimateur optimal est

$$\begin{aligned}\hat{t}_{y\mathrm{opt}}&=\boldsymbol{y}^T\ \boldsymbol{w}_0+\hat{\sum}_{y,x}\hat{\sum}_{x,x}^{-1}(\boldsymbol{t}_x-\hat{\boldsymbol{t}}_x)\\&=\hat{t}_y+(\boldsymbol{y}^T\ \boldsymbol{R}_{\mathrm{opt}}\ \boldsymbol{X}^T)\,(\boldsymbol{X}\ \boldsymbol{R}_{\mathrm{opt}}\ \boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x-\hat{\boldsymbol{t}}_x),\end{aligned} \tag{6}$$

où $\boldsymbol{R}_{\mathrm{opt}}=((\pi_{ij}-\pi_i\ \pi_j)/(\pi_{ij}\ \pi_i\ \pi_j))_{i,\ j\in s}$.

Dans un contexte asymptotique, où $n\to\infty$ et $N\to\infty$, $\hat{\sum}_{x,y}$ et $\hat{\sum}_{x,x}$ peuvent être considérés comme des composantes de la matrice asymptotique des covariances de $(\hat{t}_y,\ \hat{\boldsymbol{t}}_x^T)$. Sous l'hypothèse de convergence de $\hat{\sum}_{x,y}$ et $\hat{\sum}_{x,x}$, qui est vérifiée sous des conditions très peu contraignantes (voir Andersson et coll. 1995), l'estimateur optimal a la même variance asymptotique que l'estimateur par différence (5). En particulier, il s'ensuit que l'estimateur optimal est asymptotiquement meilleur que l'estimateur GREG habituel (voir Rao 1994, Montanari 2000 et Andersson 2001), c'est-à-dire sa variance asymptotique n'est jamais plus grande et est habituellement plus faible. À la section 5, nous présentons certaines simulations simples montrant que l'estimateur optimal peut être nettement plus efficace que l'estimateur GREG. Cependant, on ne sait rien de l'efficacité des échantillons finis, puisque l'estimateur de la covariance peut converger lentement. La vitesse de convergence est illustrée à la section 5. Notons aussi que, dans certains cas, il existe des estimateurs non linéaires qui sont asymptotiquement encore meilleurs.

Donc, le fait que l'estimateur GREG soit aussi un estimateur par calage dont la mesure de distance est

$$(\boldsymbol{w}-\boldsymbol{w}_0)^T\,\boldsymbol{R}_r^{-1}\,(\boldsymbol{w}-\boldsymbol{w}_0) \tag{7}$$

et la comparaison de (1) et (6) portent à croire que remplacer $\boldsymbol{R}_r$ par $\boldsymbol{R}_{\mathrm{opt}}$ dans (7) devrait impliquer que c'est plutôt l'estimateur par la régression optimale que nous décrivons sous forme d'un estimateur par calage. Nous montrons plus loin que cela est vrai.

3. Le résultat principal

Afin de montrer l'existence d'une mesure de distance correspondant à l'estimateur optimal, nous commencerons par énoncer et prouver un résultat dans le cas général.

Lemme : Avec $\boldsymbol{R}$ dénotant une matrice définie positive arbitraire de dimensions $n\times n$,

$$(\boldsymbol{w}-\boldsymbol{w}_0)^T\,\boldsymbol{R}\,(\boldsymbol{w}-\boldsymbol{w}_0) \tag{8}$$

soumise à la contrainte $\boldsymbol{X}\,\boldsymbol{w}=\boldsymbol{t}_x$, est minimisée par

$$\boldsymbol{w}=\boldsymbol{w}_0+\boldsymbol{R}^{-1}\ \boldsymbol{X}^T\,(\boldsymbol{X}\ \boldsymbol{R}^{-1}\ \boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x-\hat{\boldsymbol{t}}_x).$$

Démonstration : En introduisant le vecteur $\boldsymbol{\lambda}$ de dimensions $J\times 1$ de multiplicateurs de Lagrange, après dérivation, nous obtenons le système d'équations

$$2\boldsymbol{R}(\boldsymbol{w}-\boldsymbol{w}_0)+\boldsymbol{X}^T\,\boldsymbol{\lambda}=0 \tag{9}$$

$$\boldsymbol{X}W-t_x=0 \tag{10}$$

En multipliant (9) par $\boldsymbol{X}\boldsymbol{R}^{-1}$, en utilisant (10) et en isolant $\boldsymbol{\lambda}$, nous obtenons, avec $\boldsymbol{X}\,\boldsymbol{w}_0=\hat{\boldsymbol{t}}_x$:

$$\boldsymbol{\lambda}=2(\boldsymbol{X}\,\boldsymbol{R}^{-1}\boldsymbol{X}^T)^{-1}(\hat{\boldsymbol{t}}_x-\boldsymbol{t}_x). \tag{11}$$

En introduisant cette expression dans (9) et en isolant $\boldsymbol{w}$ nous obtenons finalement

$$\boldsymbol{w}=\boldsymbol{w}_0+\boldsymbol{R}^{-1}\ \boldsymbol{X}^T\,(\boldsymbol{X}\ \boldsymbol{R}^{-1}\ \boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x-\hat{\boldsymbol{t}}_x).$$

D'après le lemme, nous avons donc le résultat principal suivant :

Théorème : En posant que $\boldsymbol{R}_{\mathrm{opt}}$ est une (semi-)définie positive et en utilisant la mesure de distance de calage optimale, que nous obtenons en permettant que $\boldsymbol{R}=\boldsymbol{R}_{\mathrm{opt}}^{-1}\,(\boldsymbol{R}_{\mathrm{opt}}^{-})$ dans (8), l'estimateur par calage deviendra l'estimateur par la régression optimale.

Remarque : Dans certains cas, $\boldsymbol{R}_{\mathrm{opt}}$ peut être indéfinie (voir plus loin). La seule chose que nous savons est qu'il s'agit d'un estimateur sans biais d'une matrice de covariance. Si elle n'est pas une semi-définie positive, il

existe aussi des valeurs de x telles que $\boldsymbol{X}\boldsymbol{R}_{\text{opt}}\boldsymbol{X}^T$ ne soit pas semi-définie positive, mais la probabilité de l'existence de telles valeurs de x tend vers zéro quand les tailles de la population et de l'échantillon augmentent (et si $\sum_{x,x}$ est une définie positive). Une minimisation stricte d'une distance ayant une « composante négative » entraînerait des corrections infiniment grandes. Autant que nous sachions, ce problème posé par l'estimateur optimal n'a encore jamais été souligné.

Le moyen le plus simple de trouver une distance qui donne l'estimateur optimal sous forme d'un estimateur par calage consiste à trouver une matrice $\boldsymbol{R}_{\text{dist}}$ ayant les mêmes vecteurs propres que $\boldsymbol{R}_{\text{opt}}$, mais où les valeurs propres sont remplacées par leurs valeurs absolues (ce résultat peut être démontré de la même façon que la preuve du lemme qui précède. La distance peut être considérée comme la somme des produits des valeurs propres et des carrés des vecteurs propres. Écrire que les dérivées sont nulles signifie que, dans la proposition, nous trouvons les extrêmes, c'est-à-dire les minima pour les valeurs propres positives et les maxima pour les valeurs propres négatives. Si nous changeons tous les signes négatifs, tous les extrêmes seront des minima).

4. Exemples

Définie positive $\boldsymbol{R}_{\text{opt}}$: Supposons que les objets compris dans U sont sélectionnés indépendamment avec probabilité de sélection $\pi_1, ..., \pi_N$ (échantillonnage de Poisson), ce qui sous-entend une taille d'échantillon aléatoire n, où $E[n] = \sum_{i \in U} \pi_i$. Étant donné l'indépendance des tirages, $\boldsymbol{R}_{\text{opt}}$ est diagonale et plus précisément

$$\boldsymbol{R}_{\text{opt}}^{-1} = \boldsymbol{I}_n \left(\frac{\pi_i^2}{1-\pi_i} \right)_{i \in s}.$$

Semi-définie positive $\boldsymbol{R}_{\text{opt}}$: Supposons que n objets sont tirés selon un plan d'échantillonnage aléatoire simple, c'est-à-dire que chaque objet a une probabilté de sélection $\pi_i = n/N$. Les éléments de $\boldsymbol{R}_{\text{opt}}$ sont

$$i = j : \left(\frac{N}{n} \right)^2 \frac{N-n}{N}$$

$$i \neq j : \left(\frac{N}{n} \right)^2 \frac{n-N}{N(n-1)}.$$

Cela signifie que $\boldsymbol{R}_{\text{opt}}$ est une matrice singulière de rang $n-1$.

Supposons plutôt (comme dans l'étude par simulation suivante) que U est partitionnée en L strates de tailles $N_1, ..., N_L$, à partir desquelles nous tirons des échantillons aléatoires simples indépendants de tailles $n_1, ..., n_L$. Les éléments de $\boldsymbol{R}_{\text{opt}}$ sont alors

$$i = j : \left(\frac{N_h}{n_h} \right)^2 \frac{N_h - n_h}{N_h}$$

$$i \neq j : \left(\frac{N_h}{n_h} \right)^2 \frac{n_h - N_h}{N_h (n_h - 1)},$$

où, dans le dernier cas, i et j appartiennent tous deux à la strate h, $h = 1, ..., L$ et 0 autrement. Par conséquent, $\boldsymbol{R}_{\text{opt}}$ est de rang $N - h$.

Semi-définie non positive $\boldsymbol{R}_{\text{opt}}$: Soit U constituée de quatre éléments et s constitué de deux éléments. Supposons qu'un échantillon systématique est tiré avec probabilité 0,94 et un échantillon aléatoire simple, avec probabilité de 0,06, c'est-à-dire $\pi_{13} = \pi_{24} = 0{,}48$ et $\pi_{12} = \pi_{14} = \pi_{23} = \pi_{34} = 0{,}01$. Dans ce cas,

$$\boldsymbol{R}_{\text{opt}} = \begin{pmatrix} 2 & 23/12 \\ 23/12 & 2 \end{pmatrix} \tag{12}$$

avec probabilité de 0,96 et

$$\boldsymbol{R}_{\text{opt}} = \begin{pmatrix} 2 & -96 \\ -96 & 2 \end{pmatrix} \tag{13}$$

avec probabilité de 0,04. La deuxième matrice a une valeur propre négative.

Le problème ne disparaît pas nécessairement si N est grand. Considérons, au lieu de la précédente, une population comprenant $N/4$ strates contenant chacune quatre éléments. Supposons que la procédure d'échantillonnage susmentionnée soit utilisée indépendamment dans chaque strate. Dans ce cas, $\boldsymbol{R}_{\text{opt}}$ sera une matrice constituée des matrices 2×2 susmentionnées sur la diagonale et de zéro ailleurs.

5. Une étude par simulation

5.1 Notation et aperçu

Afin de comparer empiriquement l'estimateur optimal (OPT) à l'estimateur GREG (GREG), et de comparer également ces deux estimateurs à l'estimateur d'Horvitz-Thompson (HT), nous avons réalisé une petite étude par simulation. Aux sections précédentes, nous avons mentionné que OPT est l'estimateur linéaire asymptotiquement le plus efficace et un estimateur par calage. Bien qu'il possède de nombreuses propriétés intéressantes, il peut être inefficace pour des tailles d'échantillon raisonnables. Ici, nous allons montrer, au moyen de certaines situations simulées, que l'estimateur optimal peut aussi représenter une amélioration considérable par rapport à GREG pour des tailles d'échantillon moyennes, quand la population est (délibérément) choisie de façon telle qu'elle soit défavorable pour GREG. Une situation simple, mais non triviale, pour laquelle OPT n'est pas égal à GREG se produit en cas

d'échantillonnage aléatoire simple stratifié, notamment si les pentes ne sont pas les mêmes pour les diverses strates et la population non stratifiée. Considérons, par conséquent, une population de taille N, qui est partitonnée en L strates de tailles N_1, ..., N_L. Tirons, à partir de chaque strate, h, un échantillon aléatoire simple s_h de taille n_h, où $s_1 + \ldots s_L = s$ et $n_1 + \ldots + n_L = n$. Par souci de simplicité, supposons aussi que l'information auxiliaire est unidimensionnelle et globale, c'est-à-dire que seul t_x est connu d'avance. Pour GREG, nous avons choisi le modèle de régression linéaire simple homoscédastique (voir Särndal et coll. 1992).

Les expressions résultantes pour HT, OPT et GREG, respectivement sont

$$\hat{t}_y = N\bar{y}_{st}$$
$$\hat{t}_{y\,\mathrm{opt}} = N(\bar{y}_{st} + \hat{B}_{\mathrm{opt}}(\bar{x} - \bar{x}_{st}))$$
$$\hat{t}_{yr} = N(\bar{y}_{st} + \hat{B}_r(\bar{x} - \bar{x}_{st})),$$

où $\bar{x} = (1/N)\sum_{i=1}^{N} x_i$, $\bar{y}_{st} = (1/N)\sum_{h=1}^{L} N_h \bar{y}_{s_h}$, (analogue à $\bar{x}_{st}$) et

$$\hat{B}_{\mathrm{opt}} = \frac{\sum_{h=1}^{L} \frac{N_h^2}{n_h - 1}\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\sum_{i \in s_h}(x_i - \bar{x}_{s_h})(y_i - \bar{y}_{s_h})}{\sum_{h=1}^{L} \frac{N_h^2}{n_h - 1}\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\sum_{i \in s_h}(x_i - \bar{x}_{s_h})^2}$$

$$\hat{B}_r = \frac{\sum_{h=1}^{L} \frac{N_h}{n_h}\sum_{i \in s_h}(x_i - \bar{x}_{st})(y_i - \bar{y}_{st})}{\sum_{h=1}^{L} \frac{N_h}{n_h}\sum_{i \in s_h}(x_i - \bar{x}_{st})^2}.$$

Il est facile de voir d'après ces formules que le coefficient de régression optimal est la moyenne des pentes dans les strates et que le coefficient de régression GREG est la pente globale. Si l'écart entre ces pentes est important, la correction GREG devient mauvaise. Ici, nous nous intéressons tout spécialement à la comparaison des qualités de ces estimateurs lorsque le modèle (linéaire) auxiliaire pour GREG échoue. Nous avons donc généré des valeurs de x et de y à partir de variables aléatoires X et Y corrélées suivant des lois log-normales, où $\ln X$ suit une loi normale d'espérance 0 et de variance σ_1^2 ($N(0, \sigma_1)$), et $\ln Y$ est $N(0, \sigma_2)$. Les variances σ_1^2 et σ_2^2, et la corrélation entre $\ln X$ et $\ln Y$, peuvent alors être choisies de façon à obtenir des valeurs préétablies des variances σ_x^2 de X et σ_y^2 de Y, et de leur corrélation $\rho(X, Y)$. Les valeurs générées à partir des lois normales bivariées ont été obtenues au moyen de MATLAB (version 6.0). Nous avons créé de cette façon 12 populations, chacune de taille N = 10 000, y compris quatre combinaisons de variances σ_x^2 et σ_y^2 (10 et 100) et trois valeurs de corrélation $\rho(X, Y)$ (0,5, 0,7 et 0,9). Pour ces populations, une variance de 10 implique une asymétrie de 9,37 et une variance de 100 donne une asymétrie de 38,59.

Maintenant, avant la stratification, nous ordonnons les objets de chaque population en fonction des valeurs de y croissantes. Le nombre de strates $L = 5$ est partout avec les tailles $N_1 = 4\,000$, $N_2 = 2\,500$, $N_3 = 2\,000$, $N_4 = 1\,000$ et $N_5 = 500$. Ces données sont construites de telle façon que les objets ayant les valeurs de y les plus petites constituent la strate 1, ainsi de suite. À partir de chaque population stratifiée, nous avons tiré des échantillons de taille $n = 250, 1\,000$ et $2\,500$, où pour chaque échantillon $n_1 = \ldots = n_5$. Autrement dit, nous avons créé un plan d'échantillonnage approximativement PPT (probabilité proportionnelle à la taille) avec, par exemple, les objets dans la strate 5 ayant la probabilité de sélection la plus grande (n_5 / N_5). Le nombre d'échantillons simulés était K = 25 000 pour chacun des $12 \times 3 = 36$ cas, et nous avons alors calculé les estimateurs HT, OPT et GREG pour chaque échantillon.

En général, les mesures courantes de la qualité d'un estimateur $\hat{t}$ d'un total t pour une série $\hat{t}_1$, ..., $\hat{t}_L$ sont le biais relatif estimé

$$\frac{\bar{\hat{t}} - t}{t}$$

et la variance estimée

$$S^2 = \frac{1}{K-1}\sum_{i=1}^{K}(\hat{t}_i - \bar{\hat{t}})^2,$$

où $\bar{\hat{t}} = (1/K)\sum_{i=1}^{K}\hat{t}_i$.

Puisque nous nous intéressons principalement aux comparaisons entre OPT et GREG, nous ne présenterons que les résultats des mesures relatives de la variance (ou, de façon équivalente, l'écart-type)

$$\frac{S_{y\,\mathrm{opt}}^2}{S_{y\,\mathrm{HT}}^2} \text{ et } \frac{S_{yr}^2}{S_{y\,\mathrm{HT}}^2},$$

d'après lesquelless nous pouvons comparer les variances estimées d'OPT et de GREG à celles de HT, et déterminer lequel, d'OPT et de GREG, a la variance estimée la plus faible.

5.2 Résultats

Premièrement, à titre de référence, la valeur absolue du biais relatif estimé de l'estimateur HT sans biais n'a excédé dans aucun cas $4 \cdot 10^{-4}$. Les valeurs maximales correspondantes pour OPT et GREG étaient $6 \cdot 10^{-3}$, ce qui signifie que nous pouvons nous concentrer sur les ratios des variances estimées pour évaluer les efficacités relatives de HT, OPT et GREG.

Comme le montre l'examen du tableau 1, OPT est supérieur à HT ainsi qu'à GREG (à une exception près :

$\rho(X, Y) = 0{,}9$, $\sigma_x^2 = 10$, $\sigma_y^2 = 100$ et $n = 250$, où la variance estimée de GREG est un peu plus faible). Toutefois, pour la corrélation la plus faible, la diminution de la variance estimée d'OPT comparativement à HT n'est pas importante. Par ailleurs, GREG ne concurrence pas bien les deux autres estimateurs et cette anomalie est particulièrement prononcée pour la plus grande taille d'échantillon $n = 2\,500$. Donner à $\rho(X, Y)$ la valeur 0,7 améliore OPT ainsi que GREG, mais ce dernier est de nouveau, dans ces conditions, inférieur à HT dans la plupart des cas. Enfin, pour $\rho(X, Y) = 0{,}9$, GREG continue de présenter des propriétés médiocres comparativement à HT pour $n = 2\,500$ (à l'exception de $\sigma_x^2 = 100$ et $\sigma_y^2 = 10$). En général, GREG se rapproche d'OPT quand la valeur de $\rho(X, Y)$ augmente (le modèle linéaire auxiliaire devenant moins incorrectement spécifié), tandis que, par ailleurs, la supériorité d'OPT par rapport à GREG s'accroît quand la taille d'échantillon augmente, ce qui ne devrait pas être étonnant, puisque OPT est asymptotiquement bien motivé.

Tableau 1
Efficacité relative estimée (en pourcentage) d'OPT $(S_{y\,\text{opt}}^2 / S_{y\,\text{HT}}^2)$ et de GREG $(S_{y\,r}^2 / S_{y\,\text{HT}}^2)$ par rapport à HT, basée sur 25 000 échantillons simulés pour chaque taille d'échantillon

	$\sigma_x^2 = 10$ $\sigma_y^2 = 10$		$\sigma_x^2 = 10$ $\sigma_y^2 = 100$		$\sigma_x^2 = 100$ $\sigma_y^2 = 10$		$\sigma_x^2 = 100$ $\sigma_y^2 = 100$	
	OPT	GREG	OPT	GREG	OPT	GREG	OPT	GREG
$\rho(X, Y) = 0{,}5$								
$n = 250$	99,1	232,8	97,4	176,8	93,9	179,4	91,4	122,3
$n = 1\,000$	98,3	247,1	98,0	193,7	97,5	183,5	99,9	141,9
$n = 2\,500$	96,8	756,7	96,8	1 455,0	97,8	534,7	96,8	1 625,5
$\rho(X, Y) = 0{,}7$								
$n = 250$	89,7	197,6	83,8	101,2	73,6	120,4	64,3	72,9
$n = 1\,000$	91,0	227,5	89,8	117,2	81,2	120,5	71,7	84,0
$n = 2\,500$	93,8	648,2	91,5	1 308,6	93,1	218,6	93,1	673,5
$\rho(X, Y) = 0{,}9$								
$n = 250$	56,5	76,1	41,2	38,8	27,2	43,4	40,4	41,4
$n = 1\,000$	61,8	87,3	44,1	44,2	27,6	44,1	41,5	45,4
$n = 2\,500$	77,0	237,4	59,8	335,4	63,6	66,0	74,6	259,8

Bibliographie

Andersson, P.G. (2001). Improving estimation quality in large sample surveys. Thèse de doctorat, Department of Mathematics, Chalmers University of Technology and Göteborg University.

Andersson, P.G., Nerman, O. et Westhall J. (1995). Auxiliary information in survey sampling. *Technical Report NO 1995:3*, Department of Mathematics, Chalmers University of Technology and Göteborg University.

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

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

Montanari, G.E. (2000). Conditioning on auxiliary variable means in finite population inference. *Australian & New Zealand Journal of Statistics*, 42, 407-421.

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

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


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 109-113
Statistique Canada, N° 12-001-XIF au catalogue

Approximations de b^* dans la prévision des effets du plan dus à la mise en grappes

Peter Lynn et Siegfried Gabler [1]

Résumé

Il est fréquent de se servir de l'expression bien connue de l'effet du plan dû à la mise en grappes élaborée par Kish pour éclairer le processus d'élaboration du plan d'échantillonnage en utilisant une approximation telle que $\bar{b}$ à la place de b. Cependant, si le plan comprend une pondération ou une variation de la taille des grappes, cette approximation peut être médiocre. Dans le présent article, nous discutons de la sensibilité de l'approximation aux écarts par rapport aux hypothèses implicites et proposons une approximation de rechange.

Mots clés : Plan d'échantillonnage complexe; coefficient de corrélation intragrappe; probabilité de sélection; pondération.

1. Présentation d'autres fonctions de la taille des grappes

Kish (1965) a utilisé une expression de l'effet du plan (facteur d'accroissement de la variance) dû à la mise en grappes de l'échantillon, $\text{deff} = 1 + (b-1)\,\rho$, où b est le nombre d'observations dans chaque grappe (unité primaire d'échantillonnage) et ρ est le coefficient de corrélation intragrappe. Cette expression bien connue est enseignée dans les cours sur la théorie de l'échantillonnage et est utilisée par les statisticiens d'enquête pour concevoir et évaluer les échantillons.

L'expression est vérifiée si la taille des grappes ne varie pas et que l'échantillonnage est fait avec probabilités égales (autopondération). Nous pouvons exprimer ces deux critères formellement par :

$$b_c = b \;\forall_c \tag{1}$$

où $c = 1, \ldots, C$ représente les grappes, et

$$w_i = w \;\forall i \tag{2}$$

où $i = 1, \ldots, I$ représente les classes de pondération et w_i, les poids de sondage connexes.

Cependant, la plupart des enquêtes s'écartent des conditions (1) et (2). Dans le cas général, c'est-à-dire l'élimination des contraintes (1) et (2), Gabler, Häder et Lahiri (1999) ont montré que, sous un modèle approprié, $\text{deff}_c = 1 + (b^* - 1)\,\rho$, où

$$b^* = \sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_i b_{ci}\right)^2 \Big/ \sum_{i=1}^{I} w_i^2 b_i = \sum_{c=1}^{C}\left(\sum_{j=1}^{b_c} w_{cj}\right)^2 \Big/ \sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2 \tag{3}$$

et b_{ci} est le nombre d'observations dans la classe de pondération i dans la grappe c, $b_i = \sum_{c=1}^{C} b_{ci}$ (nous avons modifié la notation utilisée par Gabler et coll. (1999) par souci de cohérence) et w_{cj} est le poids associé à la j^{e} observation dans la grappe c, $j = 1, \ldots, b_c$.

La quantité b^* peut alors être calculée d'après les microdonnées d'enquête, à condition de connaître, pour chaque observation, le poids de sondage et la grappe concernée. Cependant, à l'étape de l'élaboration du plan d'échantillonnage, la façon de prédire b^* n'est pas claire. Gabler et coll. (1999) ont interprété le b de Kish comme étant une forme de taille pondérée moyenne de grappe :

$$\begin{aligned}\bar{b}_w &= \sum_{c=1}^{C} b_c \left(\sum_{i=1}^{I} w_i^2 b_{ci}\right) \Big/ \sum_{c=1}^{C}\sum_{i=1}^{I} w_i^2\, b_{ci} \\ &= \sum_{c=1}^{C} \left(b_c \sum_{j=1}^{b_c} w_{cj}^2\right) \Big/ \sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2 \end{aligned} \tag{4}$$

où b_c est le nombre d'observations dans la grappe c, $b_c = \sum_{i=1}^{I} b_{ci}$. Cependant, (4) n'est pas plus facile à prédire que (3) à l'étape de la conception du plan d'échantillonnage. Une interprétation plus simple, qui est peut-être adoptée fréquemment pour élaborer les plans d'échantillonnage, est la taille de grappe moyenne non pondérée :

$$\bar{b} = \sum_{c=1}^{C} b_c \Big/ C = m/C. \tag{5}$$

Il est beaucoup plus facile de prédire $\bar{b}$ que $\bar{b}_w$ ou b^* à l'étape de l'élaboration du plan d'échantillonnage, car il suffit de connaître le nombre total d'observations, m et le nombre total de grappes, C.

1. Peter Lynn, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, Royaume-Uni. Courriel : p.lynn@essex.ac.uk; Siegfried Gabler, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Postfach 12 21 55, 68072 Mannheim, Allemagne. Courriel : gabler@zuma-mannheim.de.

2. Relation entre b^*, $\bar{b}_w$ et $\bar{b}$ sous d'autres hypothèses

Soit

$$\overline{w}_c = \frac{1}{b_c}\sum_{j=1}^{b_c} w_{cj} = \sum_{i=1}^{I} w_i \frac{b_{ci}}{b_c},$$

$$\mathrm{Cov}(b_c,\ b_c\overline{w}_c^2) = \frac{1}{C}\sum_{c=1}^{C} b_c^2\ \overline{w}_c^2 - \frac{m}{C^2}\sum_{c=1}^{C} b_c\ \overline{w}_c^2$$

et

$$\mathrm{Var}(w_{cj}) = \frac{1}{b_c}\sum_{j=1}^{b_c}(w_{cj}-\overline{w}_c)^2 = \sum_{i=1}^{I}\frac{b_{ci}}{b_c}(w_i-\overline{w}_c)^2\ \forall c.$$

Alors

$$b^* = \frac{C\cdot\mathrm{Cov}(b_c,\ b_c\ \overline{w}_c^2) + \bar{b}\sum_{c=1}^{C} b_c\ \overline{w}_c^2}{\sum_{c=1}^{C} b_c\cdot\mathrm{Var}(w_{cj}) + \sum_{c=1}^{C} b_c\ \overline{w}_c^2}. \quad (6)$$

Si l'expression (1) est vérifiée, alors (6) devient :

$$b^* = \bar{b}\left(\frac{\sum_{c=1}^{C}\overline{w}_c^2}{\sum_{c=1}^{C}\mathrm{Var}(w_{cj}) + \sum_{c=1}^{C}\overline{w}_c^2}\right). \quad (7)$$

Par conséquent, dans ce cas, $b^* \le \bar{b}$. Si, en outre, les poids sont égaux dans les grappes, c'est-à-dire :

$$w_{cj} = w_c\ \forall j \in c \quad (8)$$

alors $b^* = \bar{b}$.

Si l'expression (8) est vérifiée, mais non (1), alors

$b^* \ge \bar{b}$ si, et uniquement si, $\mathrm{Cov}(b_c,\ b_c\ \overline{w}_c^2) \ge 0$ puisque

$$b^* - \bar{b} = \frac{C\cdot\mathrm{Cov}(b_c,\ b_c\ \overline{w}_c^2)}{\sum_{c=1}^{C} b_c\ \overline{w}_c^2}.$$

La covariance sera négative uniquement si de petites tailles de grappes coïncident avec de grands poids moyens dans les grappes et inversement. À la section 4, nous observons que cette situation ne s'est présentée dans aucun pays lors du premier cycle de l'Enquête sociale européenne. De surcroît, de (3) et (4), nous tirons :

$$b^* = \bar{b}_w = \sum_{c=1}^{C}(w_c\ b_c)^2 \Big/ \sum_{c=1}^{C} w_c^2\ b_c. \quad (9)$$

Si nous imposons en outre la contrainte (1), alors nous obtenons le résultat évident $b^* = \bar{b}_w = \bar{b} = b_c\ \forall c$.

Le résultat donné par (9) s'appliquerait aux enquêtes où la seule variation des probabilités de sélection est celle due à un échantillonnage disproportionné entre domaines pour lesquels il n'y a pas de recoupement de grappes. Un exemple courant serait la stratification disproportionnée selon la région, avec les UPE correspondant à des zones géographiques contenues dans les régions.

Un assouplissement pratique de la contrainte imposée à la variation des poids est :

$$b_{ci} = b_c\left(\frac{b_i}{m}\right)\forall i, c. \quad (10)$$

Autrement dit, nous permettons aux poids de varier dans les grappes, mais nous contraignons la distribution des fréquences relatives à être la même dans toutes les grappes, c'est-à-dire que les moyennes et les variances des poids dans les grappes ne dépendent pas des grappes.

Maintenant, (3) se simplifie comme suit :

$$\begin{aligned} b^* &= \sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_i\ b_c\ \frac{b_i}{m}\right)^2 \Big/ \sum_{i=1}^{I} w_i^2\ b_i \\ &= \sum_{c=1}^{C}\left(b_c^2\left(\sum_{i=1}^{I} w_i\ b_i\right)^2\right) \Big/ m^2\ \sum_{i=1}^{I} w_i^2 b_i \\ &= \frac{\left(\sum_{i=1}^{I} w_i b_i\right)^2}{\sum_{i=1}^{I} w_i^2\ b_i}\ \frac{\sum_{c=1}^{C} b_c^2}{m^2}. \end{aligned} \quad (11)$$

Notons que $((\sum_{i=1}^{I} w_i\ b_i)^2) / \sum_{i=1}^{I} w_i^2\ b_i = m/(1+c_w^2)$, où c_w^2 est le carré du coefficient de variation, sur l'ensemble des observations, des poids. En outre, $(\sum_{c=1}^{C} b_c^2)/m^2 = (1+c_b^2)/C$, où c_b^2 est le carré du coefficient de variation, sur l'ensemble des grappes, des tailles de grappe. Donc, (11) devient :

$$b^* = \frac{m}{(1+c_w^2)}\frac{(1+c_b^2)}{C} = \bar{b}\ \frac{(1+c_b^2)}{(1+c_w^2)} = \vec{b},\ \text{disons.} \quad (12)$$

Par conséquent, $\bar{b}$ sous-estimera b^* si $c_b^2 > c_w^2$ et inversement. Plus précisément, si $w_{cj} = w\ \forall j, c$ et $c_b^2 > 0$, alors $b^* > \bar{b}$. La mesure dans laquelle $\bar{b}$ sous-estimera b^* sera d'autant plus importante que la variation de b_c sera grande.

Il est rare que l'hypothèse (10) soit parfaitement vérifiée, mais ce résultat pourrait être utile dans des situations où la distribution des poids devrait, en principe, être la même dans les diverses grappes. Les échantillons fondés sur les adresses où on sélectionne une personne par adresse pourraient en être un exemple. Si la distribution du nombre de personnes par adresse est approximativement constante dans les UPE (dans la population), alors la distribution des poids ne variera d'une grappe à l'autre de l'échantillon qu'à

cause de la variation d'échantillonnage et de la non-réponse disproportionnée (l'effet de celle-ci, pourrait, naturellement, être considérable si la taille des grappes échantillonnées est faible).

Si la variation des poids n'est soumise à aucune contrainte, mais que $\mathrm{Var}(w_{cj}) > 0$ pour au moins une grappe c, alors, d'après (6),

$$b^* \geq \bar{b} \text{ si, et uniquement si } \zeta = \frac{C^2 \mathrm{Cov}(b_c, b_c \bar{w}_c^2)}{m \sum_{c=1}^{C} b_c \mathrm{Var}(w_{cj})} \geq 1. \quad (13)$$

Si l'expression (10) est vérifiée, alors $\zeta = c_b^2 / c_w^2$.

3. Incidence sur le plan d'échantillonnage

L'expression (12) donne à penser qu'on peut prédire b^* en prédisant les grandeurs relatives de c_b^2 et c_w^2. Cependant, ce résultat s'applique à une situation particulière, où

$$\begin{aligned}
\mathrm{Cov}(w_{cj},\ b_c) &= \frac{1}{m} \sum_{c=1}^{C} \sum_{j=1}^{b_c} (w_{cj} - \bar{w}) \left(b_c - \bar{\bar{b}}\right) \\
&= \frac{1}{m} \sum_{c=1}^{C} \left(b_c - \bar{\bar{b}}\right) \left(\sum_{i=1}^{I} w_i\ b_{ci} - b_c\ \bar{w} \right) \\
&\overset{\text{d'après (10)}}{=} \frac{1}{m^2} \sum_{c=1}^{C} \left(b_c - \bar{\bar{b}}\right) b_c \left(\sum_{i=1}^{I} w_i\ b_i - m\bar{w} \right) \\
&= 0
\end{aligned}$$

où

$$\bar{w} = \frac{1}{m} \sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj} = \frac{1}{m} \sum_{c=1}^{C} b_c\ \bar{w}_c$$

$$\bar{\bar{b}} = \frac{1}{m} \sum_{c=1}^{C} \sum_{j=1}^{b_c} b_c = \frac{1}{m} \sum_{c=1}^{C} b_c^2 = \frac{m}{C} (1 + c_b^2).$$

Si l'on s'attend à ce que cette covariance soit faible, il pourrait être approprié de prédire b^* comme suit :

$$\hat{b}^* = \hat{\bar{b}} = \hat{\bar{b}} \frac{(1 + \hat{c}_b^2)}{(1 + \hat{c}_w^2)}. \quad (14)$$

Les deux coefficients de variation peuvent être estimés en connaissant le plan d'échantillonnage proposé. À la section suivante, nous étudions la sensibilité de prévisions obtenues de cette façon à l'hypothèse (10) en utilisant des données réelles produites au moyen de divers plans d'échantillonnage avec $\mathrm{Cov}(w_{cj},\ b_c) > 0$.

4. Exemple : Enquête Sociale Européenne

L'Enquête sociale européenne (ESE) est une enquête transnationale qui a été conçue en s'efforçant par tous les moyens d'établir une équivalence fonctionnelle approximative entre les plans d'échantillonnage utilisés par les divers pays participants (Lynn, Häder, Gabler et Laaksonen 2004). Malgré cela, les types de plans d'échantillonnage utilisés sont très variés, avant tout à cause de la diversité des bases de sondage disponibles et des objectifs locaux, tels que le souhait de procéder à une analyse infranationale, qui peut nécessiter une stratification disproportionnée par domaine. Nous utilisons ici les données provenant du premier cycle de l'ESE, dont les travaux sur le terrain ont eu lieu en 2002–2003. Des 22 pays participants, 17 ont utilisé un plan d'échantillonnage avec mise en grappes. De ceux-ci, deux n'avaient pas encore fourni de données-échantillons utilisables au moment de la rédaction de l'article. Au tableau 1, nous présentons les valeurs d'échantillon de b^*, $\bar{b}$, c_b^2, c_w^2, $\vec{b}$, $|\vec{b} - b^*|$, $|\bar{b} - b^*|$, $\mathrm{Corr}(w_{cj},\ b_c)$, et ζ pour les 15 autres pays. Il convient de souligner que le Royaume-Uni et la Pologne avaient tous deux un plan d'échantillonnage à 2 domaines avec mise en grappes uniquement dans un domaine, à savoir la Grande-Bretagne (c'est-à-dire Irlande du Nord non comprise) et les régions moins densément peuplées (c'est-à-dire toutes sauf les 42 plus grandes villes), respectivement. Les chiffres présentés au tableau 1 ont trait uniquement au domaine mis en grappes.

Selon (12), nous devrions observer $\bar{b} > b^*$ quand $\hat{c}_w^2 > \hat{c}_b^2$. Un plan d'échantillonnage courant pour lequel on peut s'attendre à cette inégalité est un plan où a) la taille des grappes sélectionnées est constante, si bien que la variation de b_c est limitée à celle causée par les différences de non-réponse et b) les échantillons sont obtenues par sélection d'adresses avec probabilités égales, puis sélection aléatoire subséquente d'une personne par adresse, ce qui entraîne une variation des poids de sondage reflétant la variation de la taille du ménage. En tout, six pays ont adopté un plan de sondage de ce genre (AT, CH, ES, GB, GR, IL). Nous observons effectivement que, pour tous ces pays, $\zeta < 1$ et $\bar{b} > b^*$. En outre, pour cinq de ces pays (AT, CH, ES, GB, GR, $h = 1, \ldots, 5$), nous pourrions nous attendre à ce que (10) soit une approximation raisonnable, car la seule variation des poids de sondage est celle imputable à la sélection dans un ménage/une adresse. Pour ces pays, nous nous attendrions à ce que $\hat{\bar{b}}$ donne de meilleurs résultats que $\bar{b}$. En effet, $|\vec{b} - b^*| < |\bar{b} - b^*|$ pour quatre des cinq pays, et $(\sum_{h=1}^{5} |\vec{b} - b^*|) / \sum_{h=1}^{5} |\bar{b} - b^*| = 0{,}48$. Le seul pays pour lequel $\hat{\bar{b}}$ ne représente pas une amélioration est l'Espagne, ce qui était prévisible, puisque $\bar{b}$ est faible. Les grappes de petite taille sont relativement plus sensibles aux effets de la non-réponse et de la variance d'échantillonnage, ce qui entraîne la violation de (10). En Israël, la stratification disproportionnée selon la région géographique est une autre source de variation des poids de sondage. Puisque cette variation cause aussi une violation de (10), nous ne nous attendrions pas nécessairement à ce que $\hat{\bar{b}}$ représente une amélioration par rapport à $\bar{b}$ en tant que prédicteur de b^*.

Tableau 1
Valeurs d'échantillon de b^*, $\bar{b}$, c_b^2, c_w^2, $\vec{b}$ $|\vec{b}-b^*|$, $|\bar{b}-b^*|$, Corr(w_{cj}, b_c) et ζ pour 15 enquêtes

Pays		b^*	$\bar{b}$	c_b^2	c_w^2	$\vec{b}$	$\lvert\vec{b}-b^*\rvert$	$\lvert\bar{b}-b^*\rvert$	Corr(w_{cj}, b_c)	ζ
Autriche	AT	6,49	7,08	0,08	0,25	6,15	0,34	0,58	0,0036	0,4549
Belgique	BE	6,56	5,79	0,13	0,00	6,56	0,00	0,77	.	.
Suisse	CH	8,83	9,23	0,12	0,21	8,50	0,34	0,40	0,0223	0,7060
République tchèque	CZ	2,94	2,70	0,24	0,25	2,68	0,26	0,24	0,0225	1,7350
Allemagne	DE	18,85	18,13	0,07	0,11	17,42	1,43	0,72	−0,2287	.
Espagne	ES	4,96	5,04	0,17	0,22	4,80	0,15	0,08	−0,0767	0,8757
Grande-Bretagne	GB	11,11	12,27	0,08	0,22	10,90	0,21	1,16	0,0114	0,4198
Grèce	GR	5,47	5,86	0,09	0,22	5,25	0,22	0,39	−0,0280	0,5207
Hongrie	HU	8,68	8,18	0,06	0,00	8,68	0,00	0,50	.	.
Irlande	IE	12,09	11,18	0,13	0,04	12,05	0,05	0,91	0,0006	3,1054
Israël	IL	11,79	12,82	0,12	0,56	9,27	2,53	1,02	−0,1271	0,4401
Italie	IT	10,98	10,87	0,26	0,16	11,80	0,83	0,10	−0,5589	1,3018
Norvège	NO	44,09	18,68	1,33	0,01	43,32	0,77	25,41	0,0807	.
Pologne (rurale)	PL	10,07	9,45	0,06	0,01	9,88	0,19	0,62	0,2923	.
Slovénie	SI	10,76	10,13	0,06	0,00	10,76	0,00	0,63	.	.

Parmi les pays où $c_b^2 < c_w^2$, le seul pour lequel $\bar{b} < b^*$ et $\zeta > 1$ est la République tchèque (CZ). Il s'agit aussi du pays pour lequel la valeur de $\bar{b}$ est la plus faible. Quand la taille des grappes est particulièrement faible, l'effet du plan deff est faible et le choix de l'estimateur de b^* pourrait être moins important.

Dans cinq pays, les unités d'échantillonnage ont été sélectionnées individuellement avec probabilités égales (dans les grappes) à partir de registres de la population (BE, DE, HU, PL, SI). Dans ce cas, l'expression (8) (et, par conséquent, (10)) est strictement vérifiée, si bien que nous avons $\bar{b} < b^*$. Pour trois de ces pays (BE, HU, SI), l'échantillon est sélectionné avec probabilités égales, de sorte que nous observons $\vec{b} = b^*$. Il est clair que $\hat{\bar{b}}$ est supérieur à $\bar{b}$ en cas d'échantillonnage avec probabilités égales. Dans le cas de l'Allemagne et de la Pologne, il existe une certaine variation inter (mais non intra) grappes des poids de sondage. Cette variation est modeste en Pologne, et $|\vec{b}-b^*| < |\bar{b}-b^*|$, mais il n'en est pas ainsi en Allemagne, où l'ancienne Allemagne de l'Est a été échantillonnée à un taux nettement plus élevé que l'ancienne Allemagne de l'Ouest.

Le plan d'échantillonnage norvégien est le seul qui donne lieu à une variation importante de la taille des grappes à l'étape de la sélection. L'effet très important de cette variation sur $\bar{b}-b^*$ est nettement visible. De nouveau, il s'agit d'une situation où $\hat{\bar{b}}$ est probablement préférable à $\bar{b}$ en tant que prédicteur de b^*.

Les plans de l'échantillonnage de l'Irlande et de l'Italie comportent, tous deux, la sélection d'adresses à partir de registres électoraux avec probabilité proportionnelle au nombre d'électeurs, puis la sélection d'un résident au hasard pour chaque adresse sélectionnée. Ce genre de plans ne reposent pas sur la sélection avec probabilités égales, mais produisent vraisemblablement des poids de sondage nettement moins variables que les plans d'échantillonnage fondés sur les adresses dont nous avons discuté antérieurement (Lynn et Pisati 2005). Dans les deux cas, $\hat{c}_w^2 < \hat{c}_b^2$, la différence étant plus importante pour l'Italie, où la taille de certaines grappes (dans les municipalités les plus grandes) est nettement plus importante que pour d'autres (en Irlande, toutes les tailles sont égales à l'étape de la sélection). À part la République tchèque, nous n'observons $\zeta > 1$ que pour deux pays.

5. Conclusion

Pour faciliter la prévision de l'effet du plan dû à la mise en grappes, nous pensons que $\hat{\bar{b}}$ est probablement un meilleur choix que $\bar{b}$ en tant que prédicteur de b^* dans les situations où l'on peut raisonnablement s'attendre à ce que l'expression (10) soit approximativement vérifiée. Ces situations incluent, sans s'y limiter, les types courants de plan d'échantillonnage suivants :

- plan d'échantillonnage avec probabilités égales où la taille des grappes varie en raison du plan;
- plan d'échantillonnage avec probabilités égales où la taille des grappes ne varie pas en raison du plan, mais varie vraisemblablement à cause de la non-réponse;
- échantillon fondé sur les adresses, où une personne est sélectionnée à chaque adresse, aucune autre source importante de variation des probabilités de sélection n'existe et la taille des grappes ne varie pas en raison du plan de sondage.

Remerciements

Cette étude a été réalisée pendant que le premier auteur était professeur invité au centre de recherche sur les méthodes d'enquête (ZUMA), à Mannheim, en Allemagne.

Bibliographie

Gabler, S., Häder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.

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

Lynn, P., Häder, S., Gabler, S. et Laaksonen, S. (2004). Methods for Achieving Equivalence of Samples in Cross-National Surveys. ISER Working Paper 2004-09. Disponible au http://www.iser.essex. ac.uk/pubs/workpaps/pdf/2004-09.pdf.

Lynn, P., et Pisati, M. (2005). Improving the quality of sample design for social surveys in Italy: Lessons from the European Social Survey. À paraître.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2005
Vol. 31, N° 1, pp. 115-120
Statistique Canada, N° 12-001-XIF au catalogue

Une note sur la statistique C_p sous un modèle de régression à erreur emboîtée

Jane L. Meza et P. Lahiri [1]

Résumé

Les modèles de régression à erreur emboîtée sont utilisés fréquemment pour l'estimation par petits domaines et les problèmes connexes. Cependant, l'application des critères standard de sélection du modèle de régression aux modèles à erreur emboîtée donne parfois lieu à des méthodes de sélection du modèle inefficaces. Nous illustrons ce point en examinant les propriétés de la statistique C_P au moyen d'une étude par simulation de Monte Carlo. L'inefficacité de la statistique C_P peut, cependant, être corrigée grâce à une transformation appropriée des données.

Mots clés : Statistiques C_P ; modèle de regression à erreur emboîtée; simulation de Monte Carlo.

1. Introduction

Nous examinons les limites d'un critère de sélection standard du modèle de régression, c'est-à-dire la statistique C_P, quand on l'applique au modèle de régression à erreur emboîtée. La statistique C_P (Mallows 1973) est définie par

$$C_P = \frac{\text{SCR}_P}{\hat{\sigma}^2} - n + 2p \qquad (1)$$

où SCR_P est la somme des carrés des résidus et p est le nombre de paramètres du modèle P, n est le nombre d'observations et $\hat{\sigma}^2$ est une estimation de σ^2. Si le modèle est correct, la valeur de C_P doit être semblable ou inférieure à p. Le critère de sélection du modèle C_P est sensible aux valeurs aberrantes et aux écarts par rapport à l'hypothèse d'erreurs i.i.d. suivant une loi normale. La statistique C_P ne peut, par conséquent, être appliquée directement au modèle de régression à erreur emboîtée, pour lequel la structure de l'erreur n'est pas i.i.d.

Nous proposons une transformation des données qui corrige la corrélation intragrappes et permet d'utiliser le critère standard de sélection du modèle C_P. La méthode que nous présentons ici peut être appliquée pour choisir des covariables dans l'analyse des données d'enquête complexes et aux modèles d'estimation par petits domaines. Par exemple, elle pourrait être utilisée pour sélectionner les covariables dans le modèle de régression à erreur emboîtée utilisé par Battese, Harter et Fuller (1988) pour estimer la superficie (en hectares) des cultures de maïs ou de soja pour douze comtés de l'Iowa. Ces auteurs ont utilisé le modèle suivant :

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij}, \qquad (2)$$

pour l'unité $j = 1, \ldots, n_i$ dans le comté $i = 1, \ldots, m$, où n_i est la taille de l'échantillon pour le petit domaine i et la taille totale de l'échantillon est $n = \sum_{i=1}^{m} n_i$. Les effets de comté, v_i, suivent une loi $N(0, \sigma_v^2)$ indépendante des erreurs aléatoires e_{ij}, qui suivent une loi $N(0, \sigma_e^2)$. La superficie (en hectares) dans l'unité j du comté i est dénotée y_{ij} et $x_{ij} = (1, x_{ij1}, \ldots, x_{ijp})$ est un vecteur de dimension $p+1$ des valeurs des covariables $x_1, \ldots, x_p$ pour l'unité j dans le comté i. Le vecteur $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ est un vecteur de dimension $p+1$ de paramètres inconnus.

Le modèle de régression à erreur emboîtée peut être exprimé sous la forme matricielle suivante

$$y = X\beta + \varepsilon \qquad (3)$$

où $y = (y_1', \ldots, y_m')'$, $y_i' = (y_{i1}, \ldots, y_{in_i})$, $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_m')'$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})'$, $\varepsilon_{ij} = v_i + e_{ij}$. En outre, $X' = (X_1', \ldots, X_m')$ où X_i est une matrice de dimensions $n_i \times (p+1)$ avec les lignes x_{ij} pour $j = 1, \ldots, n_i$, $\varepsilon \sim N(0, \sigma^2 V)$ où $\sigma^2 = \sigma_v^2 + \sigma_e^2$, V a la forme d'une matrice diagonale par blocs $\oplus_1^m V_i$ avec $V_i = (1-\rho) I_{n_i} + \rho J_{n_i}$ où $\rho = \sigma_v^2 / \sigma^2$ est le coefficient de corrélation intrastrate courant, I_{n_i} est la matrice identité de dimensions $n_i \times n_i$ et J_{n_i} est la matrice unitaire de dimensions $n_i \times n_i$.

Puisque les erreurs du modèle à erreur emboîtée ne sont pas i.i.d., nous ne pouvons appliquer les procédures de régression standards. L'étude par simulation décrite à la section 3 montre que le critère C_P donne de mauvais résultats sous le modèle de régression à erreur emboîtée. Les transformations envisagées à la section suivante sont utilisées pour transformer le modèle de régression à erreur emboîtée en un modèle de régression standard à erreurs i.i.d. Appliqué à ces observations transformées, le critère C_P donne de nettement meilleurs résultats.

1. Jane L. Meza, University of Nebraska Medical Center, 984350 Nebraska Medical Center, Omaha, NE 68198-4350. Courriel : jmeza@unmc.edu; P. Lahiri, University of Maryland at College Park, 1218 Le Frak Hall, College Park, MD 20742-8241. Courriel : Plahiri@survey.umd.edu.

2. Correction pour les corrélations intradomaines

Comme nous l'avons mentionné à la section précédente, les méthodes classiques de sélection du modèle, telles que l'application du critère C_P, ne conviennent pas, puisqu'elles ne tiennent pas compte des corrélations intrastrates. Wu, Holt et Holmes (1988), ainsi que Rao, Sutradhar et Yue (1993) ont étudié l'effet des méthodes classiques dans le cas du modèle de régression à erreur emboîtée dans un contexte différent.

Considérons le modèle de régression à erreur emboîtée et posons que $\sigma^2 = \sigma_v^2 + \sigma_e^2$ et que ρ est le coefficient de corrélation intradomaine ordinaire $\rho = \sigma_v^2 / \sigma^2$. Comme dans Fuller et Battese (1973) et dans Rao et coll. (1993), transformons le modèle de régression à erreur emboîtée en un modèle de régression standard avec erreur i.i.d.

Soit

$$\alpha_i = 1 - \left[\frac{1-\rho}{1+(n_i-1)\rho} \right]^{1/2}, \qquad (4)$$

$$y_{ij}^* = y_{ij} - \alpha_i \bar{y}_i, \qquad (5)$$

$$x_{ij}^* = x_{ij} - \alpha_i \bar{x}_i, \qquad (6)$$

où $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ et $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij} / n_i$. Le modèle transformé devient alors

$$y_{ij}^* = x_{ij}^* \beta + e_{ij}^*, \qquad (7)$$

pour $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ et les e_{ij}^* sont indépendantes et de même loi $N(0, \sigma_e^2)$. Maintenant, nous pouvons appliquer le critère de sélection du modèle standard C_P aux données transformées.

En pratique, ρ est généralement inconnu et doit être estimé d'après les données. Rao et coll. (1993) ont utilisé la méthode de Henderson (1953) pour obtenir les estimateurs quadratiques sans biais $\hat{\sigma}_v^2$ et $\hat{\sigma}_e^2$ des composantes de la variance σ_v^2 et σ_e^2. Une fois ces estimateurs obtenus, $\rho = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2)$ peut être estimé par

$$\hat{\rho} = \max\left[0, \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2} \right]. \qquad (8)$$

Pour obtenir les estimateurs des composantes de la variance, représentons par $\{u_{ij}\}$ les résidus de la régression par les moindres carrés ordinaires de $\{y_{ij} - \bar{y}_{i.}\}$ sur $\{x_{ij1} - \bar{x}_{i.1}, \ldots, x_{ijp} - \bar{x}_{i.p}\}$ sans le terme d'ordonnée à l'origine, où $x_{i.l} = \sum_{j=1}^{n_i} x_{ijl} / n_i$ pour $l = 1, \ldots, p$. Soit $\{r_{ij}\}$ les résidus de la régression par les moindres carrés ordinaires de y_{ij} sur $\{x_{ij0}, \ldots, x_{ijp}\}$ avec le terme d'ordonnée à l'origine.

Les estimateurs de σ_v^2 et σ_e^2 sont donnés par

$$\hat{\sigma}_e^2 = (n - m - p - 1 - \lambda)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} e_{ij}^2, \qquad (9)$$

$$\hat{\sigma}_v^2 = n_*^{-1} \left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} r_{ij}^2 - (n - p - 1)\hat{\sigma}_e^2 \right], \qquad (10)$$

$$n_* = n - tr\left[(X'X)^{-1} \sum_{i=1}^{m} n_i^2 \bar{x}_{i.} \bar{x}_{i.}' \right] \qquad (11)$$

où $\lambda = 0$ si le modèle ne contient pas de terme d'ordonnée à l'origine et $\lambda = 1$ autrement. Nous proposons d'appliquer le critère standard de sélection du modèle C_P à ces observations transformées y_{ij}^* et x_{ij}^*.

3. Une étude par simulation

Nous avons réalisé une étude par simulation pour examiner le comportement du critère de sélection du modèle C_P et des transformations proposées pour le modèle de régression à erreur emboîtée. Nous avons considéré le modèle suivant :

$$y_{ij} = \beta_0 x_{ij0} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + v_i + e_{ij} \qquad (12)$$

pour $i = 1, \ldots, 10$, $n_i \in \{2, \ldots, 5\}$, $j = 1, \ldots, n_i$ et $n = 40$. Les v_i suivent une loi $N(0, \sigma_v^2)$ indépendante des e_{ij} qui suivent une loi $N(0, 1)$. Les données x_{ijl} sont tirées d'un exemple donné par Gunst et Mason (1980) et inclus dans Shao (1993) (tableau 1). La valeur de x_{ij0} est 1 pour tous $i = 1, \ldots, 10$, $j = 1, \ldots, n_i$.

Comme certains coefficients β_k peuvent être nuls, nous avons choisi, à partir de $(x_0, x_1, x_2, x_3, x_4)$, diverses combinaisons de variables comme prédicteurs pour générer les données provenant d'un modèle de régression à erreur emboîtée. Il existe $2^p - 1 = 31$ modèles possibles. Chacun est dénoté par un sous-ensemble de (0, 1, 2, 3, 4) qui contient les indices des variables x_l qui y sont incluses.

Pour générer les données, nous avons exécuté 1 000 simulations pour plusieurs valeurs de σ_v^2 afin d'estimer la probabilité de sélection de chaque modèle au moyen du critère C_P. Nous avons donné la valeur 1 à σ_e^2 pour toutes les simulations. Les résultats des simulations sont présentés au tableau 2. Nous avons considéré les valeurs 0, 1, 2, 5, 10 et 16 pour σ_v^2 et fixé les valeurs de β' à (2, 0, 0, 4, 0), (2, 0, 0, 4, 8), (2, 9, 0, 4, 8) et (2, 9, 6, 4, 8) comme dans Shao (1993). Les modèles ont été répartis en trois catégories, à savoir optimal, catégorie II (correct mais non optimal) ou catégorie 1 (incorrect).

Le critère C_P a donné de mauvais résultats pour les grandes valeurs de σ_v^2. Pour le modèle $\beta' = (2, 0, 0, 4, 0)$ avec $\sigma_v^2 = 1$, les probabilités de sélection estimées étaient : modèle optimal, 0,54; modèle correct, 0,46; modèle incorrect, 0. Par contre, pour $\sigma_v^2 = 16$, les probabilités de

Tableau 1
Données pour la simulation de l'erreur emboîtée

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
0,3600	0,5300	1,0600	0,5326	0,0900	0,1800	0,5900	0,1855
1,3200	2,5200	5,7400	3,6183	0,0200	0,1600	0,2400	0,1572
0,0600	0,0900	0,2700	0,2594	0,0200	0,1100	0,2100	0,0998
0,1600	0,4100	0,8300	1,0346	0,0500	0,2400	0,4300	0,2804
0,0100	0,0200	0,0700	0,0381	0,1100	0,3900	0,2900	0,2879
0,0200	0,0700	0,0700	0,3440	0,1800	0,1100	0,4300	0,6810
0,5600	0,6200	2,1200	1,4559	0,0400	0,0900	0,2300	0,3242
0,9800	1,0600	2,8900	4,0182	0,8500	1,3300	2,7000	2,6013
0,3200	0,2000	0,7600	0,4600	0,1700	0,3200	0,6600	0,4469
0,0100	0,0000	0,0700	0,1540	0,0800	0,1200	0,4900	0,2436
0,1500	0,2500	0,5000	0,6516	0,3800	0,1800	0,4900	0,4400
0,2400	0,2800	0,5900	0,0611	0,1100	0,1300	0,1800	0,3351
0,1100	0,3500	0,4000	0,1922	0,3900	0,3800	0,9900	1,3979
0,0800	0,1300	0,2800	0,0931	0,4300	0,4600	1,4700	2,0138
0,6100	0,8500	0,4900	0,0538	0,5700	1,1600	1,8200	1,9356
0,0300	0,0300	0,2300	0,0199	0,1300	0,0300	0,0800	0,1050
0,0600	0,1100	0,5000	0,0419	0,0400	0,0500	0,1400	0,2207
0,0200	0,0800	0,2500	0,1093	0,1300	0,1800	0,2800	0,0180
0,0400	0,2400	0,0800	0,0328	0,2000	0,9500	0,4100	0,1017
0,0000	0,0200	0,0400	0,0797	0,0700	0,0600	0,1800	0,0962

sélection estimées étaient : modèle optimal, 0,43; modèle correct, 0,35; modèle incorrect, 0,22. Le critère C_P n'a pas donné de bons résultats non plus pour les modèles plus grands avec de grandes valeurs de σ_v^2. En revanche, il a donné de très bons résultats pour les grands modèles avec de petites valeurs de σ_v^2. Pour le modèle complet $\beta' = (2, 9, 6, 4, 8)$, avec $\sigma_v^2 = 1$, les probabilités de sélection estimées étaient : modèle optimal, 0,98; modèle correct, 0,02; modèle incorrect, 0. Comparativement, pour $\sigma_v^2 = 16$, les probabilités de sélection estimées étaient : modèle optimal, 0,11; modèle incorrect, 0,89. Il convient de souligner que, dans ce scénario, le seul modèle correct est le modèle optimal.

En résumé, quand on applique le critère C_P à des données obéissant au modèle de régression à erreur emboîtée :

1. pour n'importe quel modèle, la probabilité estimée de sélection du modèle *optimal* diminue quand σ_v^2 augmente;
2. pour n'importe quel modèle, la probabilité estimée de sélection d'un modèle *incorrect* augmente quand σ_v^2 augmente;
3. à mesure que le nombre de variables incluses dans le modèle augmente et que σ_v^2 augmente, la probabilité estimée de sélection du modèle *optimal* diminue;
4. à mesure que le nombre de variables incluses dans le modèle augmente et que σ_v^2 augmente, la probabilité estimée de sélection d'un modèle *incorrect* augmente.

Nous avons alors utilisé les données pour estimer la probabilité de sélectionner chaque modèle en utilisant le critère C_P sous la transformation pour un coefficient de corrélation ρ connu. Les résultats de la simulation sont donnés dans le tableau 3. Pour le modèle $\beta' = (2, 0, 0, 4, 0)$ avec $\sigma_v^2 = 0$ (modèle de régression standard), les probabilités de sélection estimées étaient : modèle optimal, 0,62; modèle correct, 0,38; modèle incorrect, 0 (tableau 2). Pareillement, sous la transformation pour ρ connu avec $\sigma_v^2 = 16$, les probabilités de sélection estimées étaient : modèle optimal, 0,60; modèle correct, 0,40; modèle incorrect, 0 (tableau 3). Pour le modèle complet $\beta' = (2, 9, 6, 4, 8)$, la probabilité estimée de sélectionner le modèle optimal était 1 pour le modèle de régression standard (tableau 2, $\sigma_v^2 = 0$), ainsi que sous la transformation avec ρ connu pour toutes les valeurs de σ_v^2 envisagées (tableau 3).

En pratique, le coefficient de corrélation ρ est inconnu et doit être estimé d'après les données. Par conséquent, la transformation est plus utile pour les praticiens quand ρ est inconnu. Les résultats de la transformation dans ces conditions sont présentés au tableau 4. Quand nous avons estimé ρ, la probabilité estimée de sélectionner le modèle optimal ou un modèle correct n'a diminué que légèrement. La diminution la plus importante de la probabilité estimée de sélectionner le modèle optimal était 0,03 pour le modèle avec $\beta' = (2, 0, 4, 0)$ et $\sigma_v^2 = 1$, soit 0,61 pour ρ connu (tableau 3) comparativement à 0,58 pour ρ inconnu (tableau 4).

Tableau 2
Probabilités de sélection du modèle avant transformation

				$\beta = (2, 0, 0, 4, 0)'$			
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0,62	0,54	0,49	0,46	0,45	0,43
0, 2, 3	II	0,11	0,09	0,09	0,10	0,07	0,06
0, 1, 3	II	0,09	0,14	0,19	0,17	0,15	0,12
0, 3, 4	II	0,09	0,13	0,13	0,14	0,11	0,10
0, 1, 2, 3	II	0,03	0,05	0,06	0,05	0,04	0,04
0, 1, 3, 4	II	0,02	0,03	0,02	0,02	0,02	0,01
0, 2, 3, 4	II	0,02	0,01	0,02	0,02	0,01	0,02
0, 1, 2, 3, 4	II	0,02	0,01	0,00	0,00	0,01	0,00
0, 1	I	0,00	0,00	0,00	0,01	0,07	0,09
0, 2	I	0,00	0,00	0,00	0,01	0,03	0,05
0, 4	I	0,00	0,00	0,00	0,00	0,01	0,04
0, 1, 2	I	0,00	0,00	0,00	0,01	0,01	0,01
0, 1, 4	I	0,00	0,00	0,00	0,01	0,02	0,03
0, 1, 2, 4	I	0,00	0,00	0,00	0,00	0,00	0,00
				$\beta = (2, 0, 0, 4, 8)'$			
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0,72	0,67	0,63	0,61	0,58	0,49
0, 2, 3, 4	II	0,12	0,12	0,14	0,14	0,11	0,09
0, 1, 3, 4	II	0,12	0,16	0,18	0,14	0,12	0,11
0, 1, 2, 3, 4	II	0,04	0,05	0,05	0,05	0,04	0,04
0, 4	I	0,00	0,00	0,00	0,00	0,01	0,06
0, 1, 4	I	0,00	0,00	0,00	0,02	0,05	0,10
0, 2, 4	I	0,00	0,00	0,00	0,03	0,07	0,10
0, 1, 2, 4	I	0,00	0,00	0,00	0,00	0,01	0,01
				$\beta = (2, 9, 0, 4, 8)'$			
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0,83	0,78	0,75	0,63	0,39	0,25
0, 1, 2, 3, 4	II	0,17	0,20	0,18	0,13	0,09	0,07
0, 3, 4	I	0,00	0,01	0,03	0,13	0,29	0,35
0, 1, 4	I	0,00	0,00	0,00	0,03	0,11	0,15
0, 2, 3, 4	I	0,00	0,01	0,03	0,07	0,06	0,09
0, 2, 4	I	0,00	0,00	0,00	0,00	0,02	0,05
0, 1, 2, 4	I	0,00	0,00	0,00	0,02	0,04	0,04
				$\beta = (2, 9, 6, 4, 8)'$			
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1,00	0,98	0,90	0,60	0,29	0,11
0, 2, 3, 4	I	0,00	0,02	0,07	0,24	0,32	0,28
0, 1, 3, 4	I	0,00	0,00	0,02	0,11	0,18	0,23
0, 1, 2, 4	I	0,00	0,00	0,01	0,06	0,13	0,17
0, 3, 4	I	0,00	0,00	0,00	0,00	0,03	0,09
0, 2, 4	I	0,00	0,00	0,00	0,00	0,03	0,10
0, 1, 4	I	0,00	0,00	0,00	0,00	0,01	0,03
0, 1, 3	I	0,00	0,00	0,00	0,00	0,00	0,00

Tableau 3
Probabilités de sélection du modèle après transformation, ρ connu

$\beta = (2, 0, 0, 4, 0)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0,61	0,60	0,61	0,61	0,60
0, 3, 4	II	0,11	0,10	0,11	0,11	0,11
0, 2, 3	II	0,10	0,11	0,11	0,10	0,11
0, 1, 3	II	0,09	0,10	0,08	0,09	0,09
0, 1, 2, 3	II	0,04	0,04	0,04	0,04	0,04
0, 1, 3, 4	II	0,03	0,03	0,03	0,02	0,02
0, 2, 3, 4	II	0,02	0,02	0,02	0,02	0,02
0, 1, 2, 3, 4	II	0,01	0,01	0,01	0,01	0,01
$\beta = (2, 0, 0, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0,71	0,71	0,73	0,72	0,71
0, 2, 3, 4	II	0,13	0,12	0,11	0,12	0,13
0, 1, 3, 4	II	0,11	0,12	0,10	0,11	0,11
0, 1, 2, 3, 4	II	0,05	0,05	0,05	0,05	0,05
$\beta = (2, 9, 0, 4, 8)'$						
Modèle	Categorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0,82	0,83	0,83	0,82	0,83
0, 1, 2, 3, 4	II	0,18	0,17	0,17	0,18	0,17
$\beta = (2, 9, 6, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1,00	1,00	1,00	1,00	1,00

Tableau 4
Probabilités de sélection du modèle après transformation, ρ inconnu

$\beta = (2, 0, 0, 4, 0)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0,58	0,59	0,60	0,61	0,60
0, 3, 4	II	0,11	0,10	0,11	0,10	0,10
0, 2, 3	II	0,11	0,10	0,11	0,11	0,11
0, 1, 3	II	0,08	0,09	0,10	0,09	0,09
0, 1, 2, 3	II	0,04	0,04	0,03	0,04	0,04
0, 1, 3, 4	II	0,03	0,03	0,02	0,02	0,02
0, 2, 3, 4	II	0,03	0,03	0,02	0,02	0,03
0, 1, 2, 3, 4	II	0,02	0,02	0,01	0,01	0,01
$\beta = (2, 0, 0, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0,70	0,70	0,70	0,71	0,70
0, 2, 3, 4	II	0,13	0,14	0,13	0,13	0,13
0, 1, 3, 4	II	0,13	0,11	0,12	0,11	0,12
0, 1, 2, 3, 4	II	0,04	0,05	0,05	0,05	0,05
$\beta = (2, 9, 0, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0,82	0,82	0,81	0,83	0,83
0, 1, 2, 3, 4	II	0,18	0,18	0,19	0,17	0,17
$\beta = (2, 9, 6, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1,00	1,00	1,00	1,00	1,00

D'après les résultats de nos simulations, quand le critère de sélection C_P est appliqué à des données obéissant au modèle de régression erreur emboîtée :

1. sous les deux transformations (ρ connu et ρ inconnu), la probabilité estimée de sélectionner un modèle *incorrect* est 0;
2. sous la transformation pour ρ connu, la probabilité de sélectionner le modèle *optimal* est semblable à celle du modèle de régression standard;
3. quand on doit estimer ρ, la probabilité estimée de sélectionner le modèle optimal ou un modèle correct ne diminue que légèrement;
4. sous les deux transformations (ρ connu et ρ estimé), le critère C_P donne de bons résultats, même pour des modèles plus grands avec grande valeur de σ_v^2;
5. les propriétés du critère C_P pour le modèle de régression à erreur emboîtée ressemblent à celles du critère C_P pour le modèle de régression standard.

En résumé, le critère C_P donne de mauvais résultats sous le modèle de régression à erreur emboîtée quand la valeur de σ_v^2 est grande. Quand on applique la transformation pour ρ inconnu (ou ρ connu), le modèle devient un modèle de régression standard et la statistique C_P se comporte en conséquence.

Remerciements

L'étude a été financée partiellement par une bourse de l'organisation Gallup.

Bibliographie

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

Fuller, W.A., et Battese, G.E. (1973). Transformations for estimation of linear models with nested error structures. *Journal of the American Statistical Association*, 68, 626-632.

Gunst, G.F., et Mason, R.L. (1980). *Regression Analysis and Its Application.* New York : Marcel Dekker.

Henderson, C.R. (1953). Estimation of variance and variance components. *Biometrics*, 9, 226-252.

Mallows, C.L. (1973). Some comments on C_P. *Technometrics*, 15, 661-675.

Rao, J.N.K., Sutradhar, B.C. et Yue, K. (1993). Generalized least squares *F* test in regression analysis with two-stage cluster samples. *Journal of the American Statistical Association*, 88, 1388-1391.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88, 486-494.

Wu, C.F.J., Holt, D. et Holmes, D.J. (1988). The effect of two-stage sampling on the *F* Statistic. *Journal of the American Statistical Association*, 83, 150-159.

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 20, No. 3, 2004

Volume 20, No. 4, 2004

Volume 21, No. 1, 2005

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 33, No. 1, March/mars 2005, 1-148


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, N° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, 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 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, ; 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).
4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

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.