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

Techniques d'enquête

Janvier 2014

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Programme des services de dépôt

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

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

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

Note de reconnaissance

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

Signes conventionnels

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

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

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

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

POLITIQUE DE RÉDACTION

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

Présentation de textes pour la revue

Techniques d'enquête est publiée en version électronique deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (rte@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca/Techniquesdenquete).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 39, numéro 2, décembre 2013

Table des matières

Article sollicité Waksberg

Articles réguliers

Techniques d'enquête, décembre 2013
Vol. 39, N° 2, p. 273
Statistique Canada, N° 12-001-X au catalogue

Série Waksberg d'articles sollicités

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

Veuillez consulter la section avis à la fin de la revue pour des informations sur le processus de nomination et de sélection du prix Waksberg 2015.

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

Communication sollicitée pour le prix Waksberg 2013

Auteur : Ken Brewer

Ken Brewer a été méthodologiste d'enquête au *Australian Bureau of Statistics* pendant une vingtaine d'années. Il a rejoint la Section de l'échantillonnage en 1954, deux ans après sa création. Il est ensuite devenu directeur de la Division de l'échantillonnage et de la méthodologie. Il a « pris sa retraite » en 1992 pour faire de la recherche à temps plein à l'*Australian National University*. Ken a toujours été fasciné par les fondements de l'inférence statistique, en général, ainsi que dans le contexte de l'échantillonnage. Il est bien connu pour son livre de 1983 « *Sampling with Unequal Probabilities* », écrit avec M. Hanif, mais il a également apporté plusieurs autres contributions importantes. Entres autres, il a été un pionnier dans l'utilisation de la modélisation pour l'inférence et l'analyse avec des données d'enquête. Il a grandement contribué à la discussion comparant, en échantillonnage, l'inférence basée sur un modèle et celle basée sur le plan. Il a présenté le concept de « calage cosmétique » pour tenter de concilier les deux approches.


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

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 275-289
Statistique Canada, N° 12-001-X au catalogue

Trois controverses dans l'histoire de l'échantillonnage

Ken Brewer[1]

Résumé

L'histoire de l'échantillonnage, qui remonte aux écrits de A.N. Kiaer (1897), a été marquée par d'importantes controverses. Avant tout, Kiaer lui-même a dû lutter pour convaincre ses contemporains que l'échantillonnage était, en soi, une procédure légitime. Il s'y est efforcé pendant plusieurs décennies et était un vieillard avant que l'échantillonnage devienne une activité honorable. A.L. Bowley a été le premier à fournir à la fois une justification théorique de l'échantillonnage (en 1906) et une démonstration pratique de sa faisabilité (dans un sondage réalisé à Reading, qui a été publié en 1912). En 1925, les membres de l'IIS réunis à Rome ont adopté une résolution témoignant de leur acceptation de l'utilisation de l'échantillonnage par randomisation ainsi que par choix raisonné. Bowley a utilisé les deux approches. Cependant, au cours des deux décennies suivantes, on a assisté à une tendance croissante à rendre la randomisation obligatoire. En 1934, Jerzy Neyman a profité de l'échec relativement récent d'un grand sondage par choix raisonné pour préconiser que les sondages subséquents fassent appel uniquement à l'échantillonnage aléatoire. Il a trouvé en M.H. Hansen, W.N. Hurwitz et W.G. Madow des disciples doués qui, ensemble, ont publié en 1953 un traité d'échantillonnage faisant autorité. Cet ouvrage est demeuré incontesté pendant près de deux décennies. Toutefois, dans les années 1970, R.M. Royall et ses coauteurs ont remis en cause le recours à l'inférence fondée sur l'échantillonnage aléatoire et recommandé d'utiliser plutôt l'échantillonnage fondé sur un modèle. Ce plaidoyer a, à son tour, déclenché la troisième controverse importante en un peu moins d'un siècle. Néanmoins, le présent auteur, comme plusieurs autres, est convaincu que l'inférence fondée sur le plan de sondage et celle fondée sur un modèle ont toutes deux un rôle utile à jouer.

Mots-clés : Règle de trois; méthode représentative; statistique *p*; prédiction; randomisation; modèle, Horvitz-Thompson.

1 Introduction

L'un des problèmes les plus difficiles qu'il m'a fallu résoudre en rédigeant le présent article a été de décider par où débuter. Au départ, j'avais l'intention de commencer par Laplace, comme je l'avais fait dans un article antérieur (Brewer et Gregoire 2009), qui mentionnait incorrectement qu'il n'avait pas pu réaliser son ambition d'estimer la population de la France en utilisant ce que nous décririons aujourd'hui comme un sondage. En fait, il y était arrivé dès le 22 septembre 1802 en utilisant un échantillon des petits districts administratifs appelés communes (Cochran 1978). Dans des comptes rendus ultérieurs, j'avais lu qu'il avait eu de la difficulté à répéter cet exploit, les frontières de la France ne cessant d'évoluer à l'époque, et j'avais conclu incorrectement qu'il ne l'avait jamais réalisé.

Cependant, j'ai rapidement été obligé de remonter plus loin dans l'histoire. Non, Laplace n'avait pas été la première personne à utiliser un estimateur par le ratio, pas même le premier Français (Stephan 1948). L'Anglais John Graunt avait utilisé l'estimateur par le ratio dans son estimation de la population de Londres (Graunt 1662). Enfin, il n'a peut-être pas réellement utilisé l'estimateur par le ratio (il n'a probablement rien utilisé qui serait reconnu aujourd'hui comme un estimateur par le ratio, certainement pas par un statisticien d'enquête pointilleux comme moi !), mais il est vrai qu'il a utilisé la règle de trois.

Je n'étais pas tombé sur cette règle auparavant, mais apparemment elle était bien connue et s'exprimait comme ceci : « Si AB = CD et que D est inconnu, alors D = AB/C. » Manifestement, l'estimateur par le

1. Ken Brewer, School of Finance, Actuarial Studies and Applied Statistics, College of Business and Economics, Australian National University, Australie. Courriel : ken.brewer@anu.edu.au.

ratio d'aujourd'hui était un cas particulier de la règle de trois. En fait, cette règle doit avoir été établie bien avant le XVII^{e} siècle, de sorte qu'elle pourrait présenter un réel intérêt lorsque l'on cherche la date initiale de l'échantillonnage.

Rapidement, l'idée m'est venue à l'esprit qu'il y a 4 000 ans, les astronomes de Hammurabi connaissaient certainement la règle de trois, car ils maîtrisaient très bien l'arithmétique, ayant inventé un système sexagésimal qui survit de nos jours dans les « heures » (et aussi les « degrés »), les « minutes » et les « secondes », ainsi que dans les triangles de 30^{o}-60^{o}-90^{o} [« 30-60-90 »].

Ayant pris conscience de ce fait, j'ai préféré commencer à chercher un point de départ plus récent pour le présent article, et j'ai fini par conclure qu'un bon choix consisterait à partir de l'«échantillonnage moderne », sujet que l'on m'avait suggéré auparavant. La structure de l'article est la suivante. À la section 2, je discute de la première controverse qui concerne la « méthode représentative » d'Anders Kiaer. La section 3 contient une discussion de la deuxième controverse, qui a trait à l'usage exclusif de la randomisation comme moyen de sélectionner les échantillons, ainsi que l'a préconisé Neyman (1934). Les arguments en faveur de l'adoption d'une approche assistée par modèle ou fondée sur un modèle comme moyen d'inférence par échantillonnage sont décrits à la section 4. La section 5 offre une approche intermédiaire intégrant les deux procédures. L'article se conclut par un résumé à la section 6.

2 La première controverse : Anders Kiaer et la « méthode représentative »

Anders Kiaer (1838-1919), a été le fondateur et le premier directeur de Statistics Norway. Nombreux sont ceux qui affirment aujourd'hui qu'il a été le premier statisticien d'enquête moderne, mais sa contribution à la statistique n'était pas sans détracteurs à l'époque. Ceux-ci affirmaient, par exemple, que son approche de l'échantillonnage péchait par un manque de description théorique. En outre, les références faisaient sérieusement défaut dans les articles de Kiaer. La plupart des accusations portées contre lui par ses contemporains étaient fondées, mais il est également vrai qu'en publiant pour la première fois ses idées en 1895, il a lancé un processus qui a abouti à l'élaboration de la théorie moderne de l'échantillonnage. Kiaer a également été le premier à utiliser un sondage de manière indépendante, plutôt que comme un sous-produit d'un dénombrement complet.

En 1895, Kiaer avait réalisé des sondages avec succès dans son propre pays depuis au moins 15 ans, constatant pour son plus grand plaisir qu'il n'était pas toujours nécessaire de dénombrer une population complète pour obtenir de l'information utile à son sujet. Il décida qu'il était temps de convaincre ses pairs de ce fait, ce qu'il a essayé de faire à la séance de l'Institut international de statistique (IIS) qui se tenait à Berne cette année-là. Kiaer y a soutenu que ce qu'il appelait une « investigation partielle », fondée sur un sous-ensemble des unités de la population, pouvait effectivement fournir ce genre d'information, à condition seulement que le sous-ensemble en question soit choisi avec prudence de manière à refléter en miniature l'ensemble de cette population. Il décrivit ce processus comme étant sa « méthode représentative » et réussit à gagner un certain appui, notamment de la part de ses collègues scandinaves. Malheureusement, son idée de « représentation » était trop subjective et (avec le recul) manquait trop de rigueur probabiliste pour progresser face à la conviction universelle que seuls les dénombrements complets, les « recensements », pouvaient fournir des renseignements utiles (Wright 2001, Lie 2002).

De surcroît, toutes les innovations de Kiaer, en particulier son idée d'un échantillonnage « représentatif », étaient suffisamment controversables pour déclencher une forte opposition à ses idées chez ses contemporains, ce qu'ont mis tout spécialement en évidence les réactions négatives suscitées par l'article qu'il a présenté à l'assemblée de l'IIS de 1895. Il a néanmoins persisté et continué de présenter des articles au sujet de ses sondages et des méthodes qu'il utilisait aux assemblées ultérieures.

Huit ans plus tard, à l'assemblée de l'IIS qui s'est tenue à Berlin en 1903, Lucien March a soutenu que la randomisation pourrait offrir un fondement objectif pour l'utilisation des « investigations partielles » (Wright 2001, Lie 2002).

Cette idée a été explorée plus en détail par Sir Arthur Lyon Bowley, d'abord dans un article théorique (Bowley 1906) et plus tard au moyen d'une démonstration pratique de sa faisabilité dans un sondage réalisé à Reading, en Angleterre (Bowley 1912).

En 1925, à l'assemblée qui s'est tenue à Rome, les membres de l'ISS étaient suffisamment convaincus (en grande partie par un rapport que l'Institut avait lui-même commandé !) pour adopter une résolution témoignant de leur acceptation de la notion d'échantillonnage. Toutefois, ils laissaient à la discrétion des chercheurs d'opter pour l'échantillonnage aléatoire ou par choix raisonné. Avec l'avantage du recul, nous pouvons conjecturer que, aussi vague qu'ait été leur conscience du fait, les auteurs de ce rapport avaient l'intuition que, si l'échantillonnage par choix raisonné pouvait parfois donner des estimations utiles, la base offerte par la randomisation était également souhaitable.

Au cours de l'années suivante, Bowley a lui-même publié une monographie importante (Bowley 1926) dans laquelle il exposait ce que l'on savait alors au sujet des approches par choix raisonné et aléatoire de sélection de l'échantillon, et a également proposé des idées en vue de poursuivre l'élaboration de l'une et de l'autre. Il a notamment émis l'idée de rassembler les unités similaires en groupes appelés « strates », et d'inclure les mêmes proportions d'unités provenant de chaque strate dans l'échantillon. En outre, il a essayé de rendre l'échantillonnage par choix raisonné plus rigoureux en tenant compte des corrélations entre les variables d'intérêt pour le sondage et toutes autres variables auxiliaires susceptibles d'être utiles dans le processus d'estimation.

3 La deuxième controverse : Neyman préconise l'usage exclusif de la randomisation

Dans les années 1920, la situation était claire, mais loin d'être idéale. L'échantillonnage n'était plus considéré comme une approche à écarter, mais il n'existait que peu de directives, voire aucune, pour décider si l'échantillon devait être sélectionné aléatoirement ou par choix raisonné. Au cours des deux décennies suivantes, on a constaté une tendance lente, mais régulière, à rendre obligatoire l'approche fondée sur la randomisation. Et il y avait une bonne raison à cela : il n'existait pas d'autres modèles intéressants susceptibles de donner envie aux statisticiens pratiquant l'échantillonnage de les utiliser.

Un article particulièrement influant préconisant l'usage exclusif de la randomisation a été l'attaque de 69 pages de Jerzy Neyman (1934) contre le sondage réalisé par Gini et Galvani (1929). Ces deux auteurs avaient sélectionné un échantillon « par choix raisonné » de 29 districts (circondari) sur 214 à partir du Recensement de la population de l'Italie de 1921. Ils avaient tiré leur échantillon de façon qu'il reflète presque exactement les valeurs moyennes pour l'ensemble de l'Italie, pour sept variables choisies en

raison de leur importance. Mais Neyman a montré que leur échantillon présentait des différences considérables pour d'autres variables importantes. Il a ensuite monté contre l'étude une attaque s'appuyant sur un triple argument.

1) Comme ils n'avaient pas utilisé la randomisation, les chercheurs n'avaient pas pu faire appel au théorème central limite. Par conséquent, ils avaient été incapables de se fonder sur la normalité des estimations pour construire les « intervalles de confiance » que Neyman lui-même avait inventés récemment. Cette idée a été présentée en anglais pour la première fois dans cet article.
2) Comme l'avaient admis Gini et Galvani, la difficulté qu'ils avaient eue à satisfaire à leurs exigences de « choix raisonné » (voulant que l'échantillon concorde étroitement avec la population pour sept variables) les avait obligés à limiter leur étude à 214 districts plutôt que les 8 354 communes en lesquelles l'Italie avait également été divisée. Par conséquent, leur échantillon de 15 % ne comprenait que 29 districts (au lieu, peut-être, de 1 200 ou 1 300 communes). Neyman a également montré qu'ils auraient pu s'attendre à un ensemble considérablement plus précis d'estimations si l'échantillon avait été constitué d'un nombre beaucoup plus grand de ces communes (d'ordre de grandeur plus petit).
3) Crucialement, le modèle de population utilisé par les chercheurs était irréaliste et inapproprié. (Neyman était convaincu que les modèles, de par leur nature même, risquaient systématiquement de représenter inadéquatement la situation réelle.) De surcroît, la randomisation éliminait la nécessité de ce genre de modélisation de la population. En utilisant l'inférence fondée sur la randomisation, les propriétés statistiques d'un estimateur pourraient être établies en utilisant la distribution de ses estimations à partir de tous les échantillons pouvant être sélectionnés. En outre, en utilisant la randomisation, ce même estimateur sous différents plans de sondage pouvait avoir des propriétés statistiques différentes. (Un bon exemple, quoique pas l'un de ceux de Neyman, est qu'un estimateur qui est biaisé sous un plan d'échantillonnage avec probabilités égales pourrait fort bien être sans biais sous un échantillonnage avec probabilités inégales.)

Ces trois arguments n'étaient pas tous aussi valables et convaincants les uns que les autres, mais Gini et Galvani étaient prêts à admettre que quelque chose clochait sérieusement dans leur approche. De plus, Neyman pouvait défendre facilement le deuxième argument (voulant que la taille d'échantillon de 29 soit trop petite). Il était irréfutable. Les concepteurs du sondage étaient également prêts à admettre le troisième argument, selon lequel la modélisation de la population n'était pas adéquate. Le premier argument (au sujet des intervalles de confiance) semble avoir été accepté tout simplement parce que c'était Neyman qui l'affirmait, et que, puisqu'il avait certainement raison au sujet des deux autres points, il avait probablement raison pour celui-là aussi.

3.1 Opposition de Bowley au premier argument de Neyman et les résultats

Un statisticien qui n'était pas prêt à accepter la façon de penser de Neyman était Bowley, qui avait proposé la motion de remerciement à Neyman pour son exposé de 1934. Nous pouvons donc citer les mots utilisés par les deux opposants. En fait, Bowley a commencé son argumentation en se demandant à haute voix si les intervalles de confiance n'étaient pas simplement une « illusion de confiance » (en anglais, « *confiance trick* »).

Il a demandé : [traduction] « Est-ce que [un intervalle de confiance] nous mène vraiment à ce dont nous avons besoin — la chance que, à l'intérieur de l'univers que nous échantillonnons, la proportion soit comprise entre certaines limites ? Je ne pense pas. Je pense que nous nous trouvons dans la situation où nous savons que *soit* un événement improbable a eu lieu *ou* que la proportion de la population est comprise entre ces limites… L'affirmation de la théorie n'est pas convaincante, et tant que je ne serai pas convaincu, je douterai de sa validité. »

Dans sa réponse, Neyman a soutenu que la question de Bowley (selon laquelle l'intervalle de confiance était une illusion de confiance) [traduction] « conten[ait] l'énoncé du problème sous forme bayésienne » et que par conséquent sa solution [traduction] « *doit* dépendre de la *loi de probabilité a priori* ». Il a ajouté : [traduction] « Dans la mesure où nous nous en tenons à l'ancienne forme du problème, tout progrès supplémentaire est impossible. » Il a donc conclu qu'il était nécessaire d'arrêter de poser la question « bayésienne » de Bowley et d'adopter plutôt la position que son propre énoncé « *soit… ou* » [c'est-à-dire *soit* qu'un événement improbable a eu lieu *ou* que la proportion de la population était comprise entre les limites énoncées] [traduction] « form[ait] un fondement pour le travail pratique d'un statisticien se penchant sur des problèmes d'estimation… »

Cependant, il n'en demeure pas moins que les intervalles de confiance ne sont pas faciles à comprendre. Un intervalle de confiance est, en fait, un intervalle propre à l'échantillon de valeurs réelles possibles du paramètre à estimer, qui a été construit de manière à posséder une propriété particulière. Cette propriété est que, sur un grand nombre d'observations d'échantillon, la proportion de fois que la vraie valeur du paramètre tombe à l'intérieur de cet intervalle (construit séparément pour chaque échantillon) est égale à une valeur prédéterminée appelée niveau de confiance. Ce niveau de confiance s'écrit conventionnellement sous la forme $p = 1 - \alpha$, où α est petit comparativement à l'unité. Les valeurs conventionnelles de α sont 0,05, 0,01 et, parfois, 0,001. Donc, si de nombreux échantillons de taille n sont tirés indépendamment d'une loi normale, la proportion de fois que la vraie valeur du paramètre se trouvera à l'intérieur de l'intervalle de confiance d'un échantillon donné, avant que cet échantillon soit sélectionné, est $[1 - \alpha]$.

[Traduction] « Toutefois, la probabilité que cette valeur vraie du paramètre se trouve à l'intérieur de l'intervalle de confiance tel qu'il est calculé pour tout échantillon individuel de taille n ne sera pas $[1 - \alpha]$. L'intervalle de confiance calculé pour tout échantillon individuel de taille n sera, en général, plus large ou plus étroit que la moyenne et son centre pourrait s'écarter de la valeur vraie du paramètre, surtout si n est petit. Il est également parfois possible de reconnaître quand un échantillon est atypique et, donc, d'émettre l'hypothèse fondée sur la connaissance de ce fait que, dans ce cas particulier, la probabilité que la valeur vraie soit comprise dans un intervalle de confiance à 95 % particulier diffère considérablement de 0,95. »

Considérons alors, en particulier, l'intervalle de confiance à 95 % le plus fréquemment utilisé, à savoir celui compris entre $p = 0,05$ et $p = 1,00$. (Fisher (1925) a en fait proposé d'utiliser l'intervalle compris entre $p = 1/22$ et $p = 1.$) Les rédacteurs en chef de publications portant sur une grande variété de domaines (la plupart d'entre eux n'étant pas eux-mêmes des statisticiens) estiment que cette définition de la « signification » est celle qui leur donne de manière fort commode la liberté de publier des valeurs p qui tombent en dehors de cet intervalle et de rejeter celles qui ne le sont pas. Je pense, personnellement, qu'il est grand temps d'examiner très attentivement la suggestion de Fisher.

Ce qu'affirmait Fisher (en utilisant $p = 1/22$ plutôt que $p = 0,05$) était « qu'en utilisant ce critère, nous ne devrions être conduits à suivre une fausse indication qu'une fois sur 22 essais ». Mais qu'entendait-il (et qu'entendons-nous aujourd'hui) par « suivre une fausse indication » ? Ce que nous *devrions* entendre est ceci : que si l'hypothèse nulle (H_0) est vraie, une « indication fausse », c'est-à-dire « une observation significative de façon trompeuse », sera observée, en moyenne, une fois sur 22 (ou 20). Mais ce n'est pas ce que de nombreux utilisateurs non statisticiens de la statistique p imaginent qu'elle signifie. Ces utilisateurs semblent penser qu'elle signifie que seulement une de leurs « observations significatives » sur 20 (c'est-à-dire seulement une observation sur 20 parmi toutes celles dont la valeur p est inférieure à 0,05) sera significative de *façon trompeuse*.

Il s'agit là de l'idée fausse bien connue au sujet de la statistique p ! (Voir Berger et Sellke (1987) pour des détails.) Dire « Si H_0 est vraie, les observations ne seront décrites incorrectement comme étant « significatives » qu'une seule fois sur 20 (ou 22) » est correct mais inutile, car si H_0 est vraie, il s'ensuit que *chaque* observation décrite comme étant « significative », quelle qu'en soit la raison, doit aussi avoir été décrite de cette façon erronément. Mais dire simplement « Que H_0 soit vraie ou non, $p < 0,05$ » est également erroné. Dans ces circonstances, un taux de fausses découvertes (FDR pour *False Discovery Rate*) significatif est (en fait) quelque chose qui s'approche de $p < 0,0025$ ou $p < 0,05^2$.

Il s'agit d'un sujet auquel j'ai réfléchi quelque peu ces derniers temps. En particulier, j'ai corédigé un article en quatre parties portant sur ce sujet.

La partie 1 (Brewer et Hayes 2011a) expose comment remédier à la parcimonie notoire du critère d'information bayésien (BIC) en ajoutant certains termes de pénalité manifestement nécessaires. Le critère d'information bayésien augmenté (ABIC) est presque toujours compris entre le BIC original et le critère d'information d'Akaike (AIC) (tout aussi notoirement *dépourvu* de parcimonie). Une autre caractéristique utile de l'ABIC est que, dans son cas univarié, il s'agit d'une simple fonction de T (le cas limite de grand échantillon du t de Student).

La partie 2 (Brewer et Hayes 2011b) comprend la dérivation d'un test d'hypothèse bayésien de référence qui est entièrement compatible avec l'ABIC de la partie 1. Une généralisation évidente de la loi des nombres (purement empirique) de Benford joue un rôle important dans l'obtention d'une loi a priori bayésienne objective (bien que non uniforme) sur l'intervalle entier allant de zéro (ou moins l'infini) à plus l'infini pour le test d'hypothèse pertinent. (Le problème qui se pose habituellement avec les probabilités a priori nulles est évité ici en utilisant à la place des mesures de type Lebesgue.) Fait important, quand $T = 1$, le test d'hypothèse bayésien pertinent donne une mesure a posteriori qui ne penche pas plus vers l'hypothèse nulle que vers l'hypothèse alternative. En outre, quand le critère ABIC est généralisé aux petits échantillons, sous forme d'une fonction de la statistique t, la statistique p de Fisher fixe une borne supérieure du taux de fausse découverte (FDR) quel que soit le nombre de degrés de liberté.

Dans la partie 3 (Brewer, Hayes et Gillison 2012), un jeu de quelque 1 300 pentes de régression, provenant d'un sondage sur la biodiversité de la mosaïque des paysages tropicaux, est utilisé pour établir un support empirique pour le critère ABIC, confirmant de cette façon les constatations théoriques antérieures.

À la partie 4 (Hayes et Brewer 2012), les résultats approximatifs dérivés aux parties 1 à 3 sont complétés par des résultats exacts qui peuvent être obtenus en utilisant une approche similaire, mais ne

nécessitant pas d'hypothèse nulle explicite. Enfin, nous énonçons certaines conséquences probables de la reconnaissance du fait que, si l'hypothèse nulle implicite est précise, des valeurs beaucoup plus petites de $|p|$ (habituellement de l'ordre de 0,0025 plutôt que 0,05) sont nécessaires pour fournir un FDR utile.

3.2 L'acceptation des deuxième et troisième arguments de Neyman

Même s'ils étaient pertinents à l'époque et bien présentés, les deuxième et troisième arguments qu'avançait Neyman dans son article (à savoir l'inefficacité de la procédure de sélection de Gini et Galvani (1929) et la nécessité d'utiliser uniquement l'échantillonnage aléatoire) n'ont été adoptés que progressivement au cours de la décennie suivante. W. Edwards Deming a entendu Neyman à Londres en 1936. Impressionné, il a pris des dispositions pour que Neyman donne des cours et que son approche soit enseignée aux statisticiens du gouvernement américain. Un événement qui a marqué un tournant dans l'acceptation de cette approche a été l'utilisation, dans le Recensement de la population et du logement des États-Unis de 1940, d'un échantillon conçu par Deming ainsi que Morris Hansen et d'autres, pour obtenir des réponses à des questions supplémentaires. Cependant, une fois pleinement acceptés, les deuxième et troisième arguments de Neyman ont écarté toutes les autres considérations pendant au moins deux décennies.

Ces quelque 20 années ont été une période de grand progrès. Dans les termes introduits par Kuhn (1962), l'échantillonnage en population finie a trouvé dans l'inférence fondée sur la randomisation « un paradigme » universellement accepté, et il s'en est suivi une période inhabituellement longue de « science normale » fondée sur l'« échantillonnage probabiliste ». (L'« échantillonnage probabiliste » requiert que tous les éléments de la population possèdent une probabilité connue et positive d'inclusion dans l'échantillon.)

3.3 L'apparition d'ouvrages pertinents

Ce consensus a rendu possible la publication de plusieurs traités d'échantillonnage influents. L'article historique de Kish (1995) en mentionne cinq parus dans des délais rapprochés : Yates (1949), Deming (1950), Cochran (1953), Hansen, Hurwitz et Madow (« HH et M ») (1953) et Sukhatme (1954).

Selon moi, les deux plus importants de ces ouvrages ont été ceux de Cochran et de HH et M, mais pour des raisons assez opposées. HH et M semblaient ne vouloir rien entendre de la modélisation de la population. (Je doute que le mot « model » soit même mentionné dans l'un de leurs deux volumes. Il ne figure pas dans l'index.) Cochran (1953), au contraire, avait découvert plusieurs usages de ce genre de modèles, même déjà en 1953.

En relisant Cochran (1953) récemment, j'ai eu la nette impression que plus il écrivait, plus il utilisait avec aise les modèles de population. Donc, j'ai commencé à compter les occurrences du mot. La première édition comprenait 316 pages de texte. Les mots « model » et « models » y étaient utilisés à 23 occasions. Dans la première moitié du livre, le mot « model » n'apparaissait qu'une seule fois (à la page 123) et le mot « models » n'y figurait pas du tout. Mais Cochran a utilisé ces mots de nouveau trois fois dans le troisième quart de l'ouvrage et 19 fois, dans le dernier quart. (Parfois, les chiffres parlent plus que les mots !)

Un autre fait étrange était que même si l'ouvrage en deux volumes de HH et M intitulé *Sample Survey Methods and Theory* semblait ne pas contenir du tout le mot « model », chacun de ces deux volumes

contenait un chapitre sur l'«estimation par la régression ». Je ne vois pas comment on peut avoir un estimateur par la régression sans avoir un modèle de régression, du moins en tête.

HH et M ont également défini quatre « estimations » dans le chapitre 11 de leur volume 1 : l'*estimation par différence*, l'*estimation par la régression*, l'*estimation par le ratio* et l'*estimation sans biais simple*. Au chapitre 11 du volume 2, seules l'*estimation par différence* et l'*estimation par la régression* sont définies, mais naturellement les deux autres auraient déjà été bien connues de toute personne déjà familiarisée avec le volume 1.

La question de savoir si HH et M auraient considéré l'estimation par la régression comme impliquant un modèle persiste. Je parie qu'ils auraient hésité à le faire !

3.4 Mes quinze mois aux États-Unis

En 1966-1967, j'ai eu le privilège de passer plus d'un an aux États-Unis et de visiter (par ordre) le U.S. Bureau of the Census à Washington DC, puis les universités Harvard et Princeton. Au Bureau of the Census, j'avais espéré pouvoir passer du temps avec Morris Hansen et j'attendais avec hâte l'occasion de lui dire que certaines choses utiles pouvaient vraiment être faites avec des modèles de population, mais lorsque celle-ci s'est présentée, il m'a interrompu en disant « Nous n'avons pas besoin de *modèles* », et il a immédiatement changé de sujet !

Au contraire, quand je suis allé à Harvard, où j'ai passé beaucoup de temps avec Cochran, nous avons pu examiner le sujet rationnellement tous les deux et convenir que les modèles avaient un rôle utile, bien que limité, à jouer. À Princeton, j'ai essayé de susciter l'intérêt pour le sujet chez plusieurs statisticiens bien connus à l'université, mais sans grand succès.

Une contestation d'un tout autre ordre de l'orthodoxie de l'approche dépourvue de modèle de Hansen avait été exprimée par Godambe (1955), lorsqu'il avait donné sa preuve de la non-existence d'un estimateur de la moyenne de population fondé sur la randomisation qui soit uniformément le meilleur. Une nouvelle notation et une nouvelle classe d'estimateurs étaient nécessaires pour appuyer l'argument, et, sous sa forme première, ce cadre a suscité une certaine résistance. À la Section 5 de son article, citant le traité de Yates (1949) et l'article de Cochran (1939) comme antécédents, Godambe proposait un critère d'optimalité de rechange, la minimisation de la variance d'échantillonnage attendue sous ce que l'on a appelé plus tard un modèle de superpopulation.

À l'époque, peu d'autres statisticiens travaillant dans ce domaine incroyablement novateur de l'échantillonnage semblaient être concernés par ce résultat. Je dois confesser que je n'étais moi-même pas concerné à l'époque, mais aujourd'hui, je pense que j'aurais peut-être dû l'être !

4 La troisième controverse, « Inférence par échantillonnage : assistée par modèle ou fondée sur un modèle? »

L'école partisane de l'échantillonnage en population finie a été stupéfaite lorsque Royall (1970) a publié son très intéressant plaidoyer pour le rétablissement de l'échantillonnage par choix raisonné et de l'inférence fondée sur la prédiction. Lire cet article équivalait à lire l'article de Neyman (1934) en le renversant. Alors que les mêmes questions étaient abordées, les conclusions opposées étaient tirées.

Toutefois, en 1973, Royall avait abandonné la plus extrême de ses recommandations, à savoir que le meilleur échantillon à sélectionner serait celui qui était optimal aux termes d'un modèle représenté par les équations suivantes :

$$Y_i = \beta X_i + U_i . \tag{4.1}$$

$$E\left(U_i\right) = 0 \tag{4.2}$$

$$E(U_i^2) = \sigma^2 X_i \tag{4.3}$$

et

$$E(U_i U_j) = 0. \tag{4.4}$$

Un tel échantillon aurait habituellement été constitué des n unités les plus grandes de la population, en prenant comme mesure leurs valeurs réalisées x_i, ce qui aurait causé des difficultés si la valeur du paramètre β n'avait pas été presque constante sur la plage complète des tailles des unités de la population.

Dans des articles ultérieurs (Royal et Herson 1973a, Royal et Herson 1973b, Cumberland et Royall 1981), Royall a proposé que l'échantillon choisi soit « équilibré », autrement dit que les moments de l'échantillon x_i soient aussi proches que possible des moments correspondants pour l'ensemble de la population. Cette proposition formalisait la notion beaucoup plus ancienne selon laquelle les échantillons devraient être choisis délibérément de manière à ressembler à la population en miniature. Les échantillons de Gini et Galvani avaient été choisis à peu près de la même façon — ce qui signifie ici « à peu près de la même façon en intention » — mais certainement pas avec le même succès d'exécution.

En majeure partie, la position originale de Royall est restée inébranlable. La fonction d'un statisticien spécialiste de l'échantillonnage était de créer un modèle réaliste de la population pertinente, de concevoir un échantillon pour en estimer les paramètres et de faire toutes les inférences concernant cette population en fonction de ces estimations des paramètres. Le concept, axé sur la randomisation, de la définition de la variance d'un estimateur en fonction de la variabilité des estimations produites sur tous les échantillons possibles devait être écarté au profit de la variance fondée sur la prédiction, qui était spécifique à l'échantillon et fondée sur la moyenne de toutes les réalisations possibles du modèle de prédiction choisi.

Peu importe l'échantillon tiré, l'estimateur de Royall pour un total de population $T_y = \sum_U y_i$ avait la forme prédictive suivante :

$$t_y = \sum_s y_i + \sum_{U-s} x_i \hat{\beta}_{\text{BLUE}} ,$$

où $\hat{\beta}_{\text{BLUE}} = \sum_s y_i \Big/ \sum_s x_i$ était le meilleur estimateur linéaire sans biais de β fondé sur l'échantillon sous le modèle représenté par l'équation (4.1). Il s'agit d'une forme prédictive, puisque les valeurs y de $U - s$ sont prédites par le modèle.

À aucune étape les statisticiens spécialistes de l'échantillonnage n'avaient mis beaucoup de temps à prendre parti dans ce débat. Maintenant, chacun avait choisi son camp. Le feu de l'argumentation semble avoir été exacerbé par des obstacles de langue; par exemple, les mots « espérance » et « variance » étaient associés à un ensemble de connotations dans le cas de l'inférence fondée sur la randomisation et à un

ensemble assez différent pour l'inférence fondée sur la prédiction. Toutes les assertions faites par un camp étaient perçues par l'autre camp comme un non-sens inintelligible.

Une importante contre-attaque a été lancée dans un article rédigé par Hansen, Madow et Tepping (1983). Ils ont montré qu'une divergence faible (et selon la plupart des critères, indécelable) par rapport au modèle de Royall pouvait néanmoins fausser considérablement l'inférence d'après l'échantillon. La réfutation évidente aurait été : « Mais cette distorsion n'aurait pas eu lieu si l'échantillon avait été tiré de manière équilibrée. »

5 Une troisième option, « Utiliser les deux approches ensemble »

Finalement, une troisième position a été proposée, celle que partage le présent auteur, à savoir que, puisque l'approche fondée sur le plan de sondage (ou fondée sur la randomisation) et celle fondée sur un modèle (ou fondée sur la prédiction) ont toutes deux des avantages, et qu'il est possible de les combiner, les deux devraient être utilisées ensemble. J'avais en fait laissé entrevoir cette possibilité dans Brewer (1963), article qui avait suscité peu d'intérêt à l'époque, mais qui a été repéré plus tard par J.N.K. Rao, qui lui a accordé une certaine reconnaissance, du moins dans la mesure où il m'a invité à lui rendre visite à Ottawa où je suis resté six semaines en 1974.

Combiner ces deux approches était relativement simple. Chacune d'elles comprenait une variable y qui était l'objet d'intérêt principal et une variable apparentée ou auxiliaire x, au sujet de laquelle étaient connus des renseignements supplémentaires susceptibles d'aider à estimer la valeur de la variable y. Ces « renseignements supplémentaires » étaient habituellement le total de population connu de toutes les valeurs de x, désigné par T_x. Par conséquent, la *relation* présentant le plus d'intérêt était celle qui reliait le paramètre crucial β dans l'équation (1a) à l'estimateur *cosmétique* $\hat{\beta}_{\text{COS}}$, à savoir

$$\hat{\beta}_{\text{COS}} = \frac{\sum_s (\pi_i^{-1} - 1) y_i}{\sum_s (\pi_i^{-1} - 1) x_i}, \tag{5.1}$$

où π_i est la probabilité que l'unité i soit sélectionnée dans l'échantillon, ou suivant la notation utilisée par Särndal (2011),

$$\hat{\beta}_{\text{COS}} = \frac{\sum_s (d_k - 1) y_i}{\sum_s (d_k - 1) x_i}, \tag{5.2}$$

où son d_k est identique à mon π_i^{-1}. L'estimateur du total $Y = \sum_U y_k$ est

$$\hat{Y}_{\text{COS}} = \sum_s d_k y_k + \left(\sum_U x_k - \sum_s d_k x_k \right) \frac{\sum_s (d_k - 1) y_k}{\sum_s (d_k - 1) x_k}. \tag{5.3}$$

Särndal (2011) a également montré que ces valeurs x et y peuvent être reliées l'une à l'autre de plusieurs façons, mais aussi qu'il existe un élément commun à toutes ces façons. Cet élément commun est

que y augmente linéairement quand x augmente, et que l'ampleur de cette linéarité est mesurée par le paramètre β dans l'équation (4.1). Fait important, toutefois, quand $\hat{\beta}_{\text{COS}}$ remplace $\hat{\beta}_{\text{BLUE}}$ dans l'estimateur par prédiction de Royall, on peut montrer que l'estimateur est presque sans biais sous le plan, quelle que soit la validité du modèle hypothétique.

L'équation (5.2) figure aussi explicitement à la page 569 de Brewer (2011), directement après sa formule plus générale en notation matricielle, c'est-à-dire

$$\hat{\beta}_{\text{COS}} = \left[X_s' Z_s^{-1} \left(\Pi_s^{-1} - I_n \right) X_s \right]^{-1} X_s' Z_s^{-1} \left(\Pi_s^{-1} - I_n \right) y_s . \tag{5.4}$$

Quand la question se pose de savoir combien de variables explicatives doivent être utilisées dans le modèle pertinent, Särndal (2011) fait une distinction qui semble désobligeante entre les pays « riches en variables explicatives » et « pauvres en variables explicatives ». Il considère certainement ces pays « pauvres en variables explicatives » comme étant considérablement désavantagés du fait d'avoir relativement peu d'« explicateurs ».

Au moins un pays « riche en variables explicatives » (l'Australie) semble avoir pris la décision délibérée d'ignorer les avantages dont pourraient bénéficier ceux qui sont « riches en variables explicatives ». La procédure australienne courante (celle servant principalement à produire les séries désaisonnalisées) consiste à utiliser une seule variable auxiliaire, à savoir le total de recensement le plus récent disponible, comme seul « explicateur ».

Avant cela, Brewer (1999a) avait également montré qu'il pourrait être préférable d'utiliser un estimateur par la régression cosmétique pour compenser tout manque d'équilibre au lieu de s'esquinter à sélectionner des échantillons équilibrés. Cependant, ceux qui préfèrent utiliser directement l'échantillonnage équilibré peuvent maintenant procéder à une sélection aléatoire parmi de nombreux échantillons équilibrés ou quasi équilibrés en utilisant la « méthode du cube » (Deville et Tillé 2004). Cet article contient aussi plusieurs références à des méthodes antérieures de sélection d'échantillons équilibrés, mais quelle que soit la manière dont l'échantillon équilibré pertinent est obtenu, la façon de l'utiliser est identique.

Dans Brewer et Gregoire (2009), chacune des trois approches pertinentes d'estimation (randomisation seulement, prédiction seulement et les deux combinées) est examinée. Arrivé ici, il est commode de citer un autre de mes articles (Brewer 2005, pages 390-391) qui expose les raisons pour lesquelles je souhaitais, et souhaite encore, utiliser les deux méthodes simultanément, et comme il est facile de le faire.

[Traduction] « Chaque approche a ses mérites, et il y a des avantages à les utiliser ensemble. Considérons comment chacune de ces inférences fonctionne.

Premièrement, l'inférence fondée sur le plan de sondage. Considérons le cas général où les probabilités d'inclusion π_i sont connues mais diffèrent d'une unité à l'autre. Dans ce cas, nous pouvons imaginer que le statisticien d'enquête construit un modèle de la population en examinant l'une après l'autre les unités échantillonnées et en disant *Ah oui, la première unité a été incluse avec une chance sur dix, donc mon modèle de la population inclut cette unité ainsi que neuf autres unités non échantillonnées ayant la même valeur Y_k que la première unité. Par contre, la deuxième unité a été incluse avec seulement une chance sur deux, de sorte que mon modèle inclut cette unité et seulement une autre unité qui lui est semblable.* »

Ici, la conséquence de l'utilisation de cette procédure est donc que, dans l'esprit de l'échantillonneur, le modèle de la population comprendrait deux unités échantillonnées réelles (une provenant de chaque strate de l'échantillon) ainsi que dix unités imaginaires (neuf provenant de la strate avec une fraction d'échantillonnage de un sur dix, et une provenant de la state avec une fraction d'échantillonnage de un sur deux) et, enfin, toutes les unités provenant de la strate entièrement dénombrée.

Brewer (2005, page 391) poursuit ainsi : [traduction] « Donc, même l'estimation fondée sur le plan de sondage peut être imaginée comme étant fondée sur un modèle, mais un modèle assez différent des modèles de prédiction… privilégiés par l'école dite de l'*estimation fondée sur un modèle*. Plus exactement, cette école devrait être décrite comme celle de l'estimation *fondée sur la prédiction* et l'école de l'estimation *fondée sur le plan de sondage* devrait être décrite comme celle de l'estimation *fondée sur la randomisation*. Chaque école utilise un modèle, mais dans un cas, il s'agit d'un modèle de prédiction, et dans l'autre, d'un modèle de randomisation. »

L'approche fondée sur la randomisation décrite plus haut est celle qui a été utilisée pour la sélection de deux unités de l'échantillon (une provenant de chaque strate échantillonnée) et de toutes les unités de la strate dénombrée entièrement. Elle donne aussi l'estimateur de Horvitz-Thompson bien connu, qui peut s'écrire

$$\hat{T}_{\mathrm{HT}} = \sum_{i \in s} \frac{Y_i}{\pi_i} = \sum_{i=1}^{N} \delta_i \frac{Y_i}{\pi_i} \tag{5.5}$$

où δ_i est un indicateur d'inclusion prenant la valeur « un » si la i^{e} unité se trouve dans l'échantillon ou dans la strate entièrement dénombrée, et la valeur « zéro » autrement. Dans ce cas particulier, il est défini sur les deux unités échantillonnées ainsi que sur toutes les unités dans la strate entièrement dénombrée. [Cette dernière phrase corrige l'erreur mentionnée plus haut.]

Les statisticiens appartenant à l'école de l'estimation fondée sur la prédiction ridiculisent l'usage de l'inférence fondée sur la randomisation parce que les probabilités d'inclusion sont choisies arbitrairement par le concepteur de l'échantillon et sont par conséquent incapables (disent-ils) d'indiquer quoi que ce soit de significatif au sujet de la population ! Ils préfèrent utiliser le meilleur estimateur linéaire sans biais (BLUE) du paramètre de régression β en guise d'étape vers l'obtention du meilleur prédicteur linéaire sans biais (BLUP) de T. Il s'agit d'un prédicteur, parce que T est une variable aléatoire sous le modèle, et non un paramètre.

Alors, quel est le meilleur estimateur de T, le HT ou le BLUP ? Le BLUP est le meilleur si le modèle de prédiction est vérifié exactement, et est nettement meilleur si l'échantillon ainsi que la population sont petits. Cependant, il existera toujours une taille d'échantillon au-delà de laquelle l'estimateur HT est le plus efficace, à moins que le modèle soit vérifié exactement.

6 Sommaire

En conclusion, nous pouvons constater que l'échantillonnage, au cours de son histoire relativement brève, a été étonnamment vulnérable aux controverses. Au départ, la notion même que l'on puisse procéder à toute forme d'échantillonnage suscitait de l'opposition. La seule source valide d'information statistique était considérée comme le dénombrement complet. Il a fallu la détermination de Kiaer, qui

occupait déjà une position d'autorité de haut rang, pour briser l'opposition à ce que l'on avait fini par démontrer être un outil valable.

La deuxième controverse était aussi attribuable à la détermination de quelques personnes seulement. Neyman a pris les rênes, mais cette fois-ci, d'autres sont intervenus. Bowley était certainement impliqué dès le début, mais Neyman semble avoir offert des arguments plus convaincants à un moment crucial. Ces arguments étaient controversables, même de prime à bord, et ils ne m'impressionnent certes guère aujourd'hui, mais à l'époque, Neyman a trouvé un disciple en la personne de Hansen, qui a dominé la collectivité de l'échantillonnage pendant des décennies, au moins jusqu'au milieu des années 1970.

La troisième controverse se poursuit et la façon dont elle se terminera n'est pas vraiment claire, mais la préférence à l'heure actuelle (du moins pour des échantillons de taille moyenne) serait d'utiliser de concert les estimateurs par prédiction et par randomisation.

En résumé, les estimateurs HT et BLUP peuvent tous deux être utiles dans différentes situations. Il est logique d'utiliser le BLUP lorsque la taille d'échantillon est petite et qu'un modèle est désespérément nécessaire. L'estimateur HT protège contre l'échec du modèle de prédiction lorsque l'échantillon devient grand. Un statisticien prudent combinerait les principes des deux estimateurs.

Bibliographie

Berger, J.O., et Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p*-values and evidence (with discussion). *Journal of the American Statistical Association*, 82, 112-139.

Bowley, A.L. (1906). Address to the economic and statistics section of the British association for the advancement of science. *Journal of the Royal Statistical Society*, 76, 672-701.

Bowley, A.L. (1912). Working class households in reading. *Journal of the Royal Statistical Society*, 76, 672-701.

Bowley, A.L. (1926). Measurement of the precision obtained in sampling. *Bulletin of the International Statistical Institute*, 22, 11-62 (supplement).

Brewer, K.R.W. (2011). Remarks on the paper on "Combined inference in survey sampling" by Carl-Erik Särndal. *Pakistan Journal of Statistics*, 27, 4, 567-572.

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 10, 213-233.

Brewer, K.R.W. (1999a). Design-based or model-based inference? Stratified random vs stratified balanced sampling. *International Statistical Review*, 67, 35-47.

Brewer, K.R.W. (1999b). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, 25, 205-212.

Brewer, K.R.W. (2005). Anomalies, probings, insights: Ken Foreman's role in the sampling inference controversy of the late 20th century. *Australian and New Zealand Journal of Statistics*, 47, 4, 385-399.

Brewer, K.R.W., et Gregoire, T.G. (2009). Introduction to survey sampling. Chapter 1 of *Handbook of Statistics* 29A, *Sample Surveys: Design, Methods and Applications*, (Éds., D. Pfefferman et C.R. Rao), Elsevier.

Brewer, K.R.W., et Hayes, G. (2011a). Understanding and using Fisher's *p*: Part 1: Countering the *p*-statistic Fallacy. *Mathematical Scientist*, 36, 107-116.

Brewer, K.R.W., et Hayes, G. (2011b). Understanding and using Fisher's *p*: Part 2: A Reference Bayesian Hypothesis Test. *Mathematical Scientist*, 36, 117-125.

Brewer, K.R.W., Hayes, G. et Gillison, A.N. (2012). Understanding and using Fisher's *p*: Part 3: Examining an Empirical Data Set. *Mathematical Scientist*, 37, 20-26.

Cochran, W.G. (1953). *Sampling Techniques.* First Edition, Wiley.

Cochran, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492- 510.

Cochran, W.G. (1978). Laplace's ratio estimator. In *Contributions to Survey Sampling and Applied Statistics*; papers in honor of H.O. Hartley; H.A. David (Editeur), 3-10.

Deming, W.E. (1950). *Some theory of sampling.* Dover books on mathematics.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling, the cube method. *Biometrika*, 91, 893-912.

Fisher, R.A. (1925). *Statistical methods for research workers.* 14th Edition (1970) Oliver et Boyd.

Gini, C., et Galvani, L. (1929). Di una applicazione del metodo rappresentativo all' ultimo censimento italiano della popolazione (1 dicembre 1921). *Annali di statistica* VI 4, 1-107.

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

Graunt, J. (1661/2). Natural and political observations made upon the Bills of Mortality. Reprinted (1939) Baltimore: The John Hopkins Press.

Hansen, M.H., Hurwitz W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory (2 vols.)* (Republished 1993) Wiley, New York.

Hansen, M.H., Madow W.G. et Tepping, B.J. (1983). An evaluation of dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Hayes, G., et Brewer, K.R.W. (2012). Understanding and using Fisher's *p*: Part 4: Do we even need to specify a prior measure at H_0? *Mathematical Scientist*, 37, 27-33. Sons, New York.

Kiaer, A.N. (1897). The representative method of statistical surveys. Papers from the Norwegian Academy of Science and Letters, II The Historical, philosophical Section, 1897 No. 4.

Kish, L. (2003). Selected Papers. Graham Kalton (Editor) Steven Heeringa (Editor) Wiley.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2 (5), 813- 830.

Kuhn, T.S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

Lie, E. (2002). The rise and fall of sample surveys in Norway, 1875-1906. *Science in context*, 15 (3), 385-1906.

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

Royall, R.M. (1970). On finite population sampling theory under certain regression models. *Biometrika*, 57, 377-387.

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

Royall, R.M., et Herson, J. (1973b). Robust estimation in finite population II: Stratification on a size variable. *Journal of the American Statistical Association*, 68, 890-893.

Särndal, C.-E. (2011). Combined inference in survey sampling. *Pakistan Journal of Statistics*, 27 (4) 359-370.

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

Sukhatme, P.V. (1954). Sampling theory of surveys: With applications. Asia Publishing House.

Wright, T. (2001). Selected moments in the development of probability sampling: Theory and practice. *Survey Research Methods Section Newsletter*, American Statistical Association, Alexandria, VA. Issue 13, 1-6.

Yates, F. (1949). Sampling Methods for Censuses and Surveys, Londres, C. Griffin.


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

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 291-311
Statistique Canada, N° 12-001-X au catalogue

Une approche d'inférence fondée sur la vraisemblance composite pondérée pour des modèles à deux niveaux issus de données d'enquête

J.N.K. Rao, François Verret et Mike A. Hidiroglou[1]

Résumé

Les modèles multiniveaux sont d'usage très répandu pour analyser les données d'enquête en faisant concorder la hiérarchie du plan de sondage avec la hiérarchie du modèle. Nous proposons une approche unifiée, basée sur une log-vraisemblance composite pondérée par les poids de sondage pour des modèles à deux niveaux, qui mène à des estimateurs des paramètres du modèle convergents sous le plan et sous le modèle, même si les tailles d'échantillon dans les grappes sont petites, à condition que le nombre de grappes échantillonnées soit grand. Cette méthode permet de traiter les modèles à deux niveaux linéaires ainsi que linéaires généralisés et requiert les probabilités d'inclusion de niveau 2 et de niveau 1, ainsi que les probabilités d'inclusion conjointe de niveau 1, où le niveau 2 représente une grappe et le niveau 1, un élément dans une grappe. Nous présentons aussi les résultats d'une étude en simulation qui donnent la preuve que la méthode proposée est supérieure aux méthodes existantes sous échantillonnage informatif.

Mots-clés : Vraisemblance composite; probabilités d'inclusion; échantillonnage informatif; modèles multiniveaux.

1 Introduction

Les données recueillies dans le cadre d'enquêtes socioéconomiques, sur la santé et autres à grande échelle sont utilisées abondamment à des fins analytiques, comme l'inférence sur les paramètres de modèles de régression linéaire et de régression logistique linéaire de populations. Ne pas tenir compte des caractéristiques du plan de sondage (comme la stratification, la mise en grappes et les probabilités de sélection inégales) peut donner lieu à des inférences incorrectes sur les paramètres du modèle, à cause du biais de sélection dans l'échantillon causé par l'échantillonnage informatif. Il est tentant d'étendre les modèles en incluant parmi les variables auxiliaires toutes les variables du plan de sondage qui définissent le processus de sélection à divers niveaux, puis d'ignorer le plan de sondage et d'appliquer des méthodes classiques au modèle étendu. Les principales difficultés de cette approche sont les suivantes (Pfeffermann et Sverchkov 2003) : 1) l'analyste pourrait ne pas connaître toutes les variables du plan ou ne pas avoir accès à toutes ces variables; 2) l'utilisation d'un trop grand nombre de variables du plan de sondage peut causer des difficultés d'inférence à partir du modèle étendu; 3) le modèle étendu pourrait ne plus présenter d'intérêt scientifique pour l'analyste. Par ailleurs, l'approche fondée sur le plan de sondage peut fournir des inférences par échantillonnage répété asymptotiquement valides sans modifier le modèle de l'analyste. Une approche unifiée, fondée sur des équations d'estimation pondérées par les poids de sondage conduit à des estimateurs convergents sous le plan des paramètres de « recensement », c'est-à-dire de population finie, qui à leur tour permettent d'estimer les paramètres associés du modèle. En outre, les méthodes de rééchantillonnage, comme le jackknife et le bootstrap pour données d'enquête, peuvent fournir des estimateurs de variance valides et des inférences connexes sur les paramètres de recensement. Les mêmes méthodes pourraient aussi être applicables à l'inférence sur les paramètres du modèle, dans de nombreux

1. J.N.K. Rao, École de mathématiques et de statistique, Université Carleton, Ottawa (Ontario), Canada, K1S 5B6. Courriel : jrao@math.carleton.ca; François Verret, Statistique Canada, 15 B, immeuble R.-H.-Coats, Ottawa (Ontario), Canada, K1A 0T6. Courriel : francois.verret@statcan.gc.ca; Mike A. Hidiroglou, Statistique Canada, 16 D, immeuble R.-H.-Coats, Ottawa (Ontario), Canada, K1A 0T6. Courriel : mike.hidiroglou@statcan.gc.ca.

cas d'enquêtes à grande échelle. Dans les autres cas, il est nécessaire d'estimer la variance sous le modèle des paramètres de recensement à partir de l'échantillon. L'estimateur de la variance totale est alors donné par la somme de cet estimateur et de l'estimateur de variance par rééchantillonnage. Beaumont et Charest (2010) ont étendu le bootstrap à l'estimation de la variance totale associée aux paramètres du modèle. Le lecteur est invité à consulter Rao et coll. (2010) pour un aperçu des méthodes d'inférence sur les paramètres de régression issus de données d'enquête complexes.

Dans le présent article, nous visons avant tout à faire des inférences fondées sur le plan de sondage sur les paramètres des composantes de la variance et sur les paramètres de régression de modèles multiniveaux en partant de données obtenues au moyen de plans d'échantillonnage à plusieurs degrés qui correspondent aux niveaux du modèle. Par exemple, dans une étude sur l'éducation menée auprès des élèves, les écoles (unités d'échantillonnage de premier degré) pourraient être sélectionnées avec probabilité proportionnelle à la taille de l'école et les élèves (unités d'échantillonnage de deuxième degré) pourraient être sélectionnés dans les écoles échantillonnées selon un plan d'échantillonnage aléatoire stratifié. De nouveau, ne pas tenir compte du plan de sondage et utiliser des méthodes classiques pour les modèles multiniveaux peut donner lieu à des inférences incorrectes en cas de biais de sélection dans l'échantillon. Dans l'approche fondée sur le plan de sondage, il est plus difficile d'estimer les paramètres des composantes de la variance du modèle que les paramètres de régression. Les travaux antérieurs sur les modèles multiniveaux pour données d'enquête sont résumés à la section 2. Notre objectif principal est de présenter une approche unifiée d'inférence pour des modèles multiniveaux provenant de données d'enquête, fondée sur une log-vraisemblance composite pondérée (section 4). La méthode proposée produit des inférences asymptotiquement valides sur les paramètres des composantes de la variance, même quand les tailles d'échantillon dans les grappes sont petites, à condition que le nombre de grappes échantillonnées soit grand, contrairement à certaines méthodes existantes résumées à la section 2. Les résultats d'une simulation limitée sont présentés à la section 5.

2 Modèles à deux niveaux : travaux antérieurs

2.1 Modèles à deux niveaux

Les modèles multiniveaux (ou modèles hiérarchiques) sont d'usage très répandu, notamment dans les domaines des sciences sociales, de l'éducation et de la santé, pour analyser les données d'enquête possédant une structure hiérarchique. Ici, nous nous concentrons sur les modèles à deux niveaux associés à l'échantillonnage à deux degrés de grappes (niveau 2) : un échantillon, s, d'unités de niveau 2, i, est sélectionné selon un plan spécifié, puis un échantillon, $s(i)$, d'éléments (ou unités de niveau 1), j, est sélectionné dans chacune des unités de niveau 2 échantillonnées i conformément à un autre plan spécifié. Nous supposons, en nous inspirant de la littérature sur les modèles multiniveaux pour données d'enquête, que le modèle concorde avec la hiérarchie du plan de sondage, comme dans l'exemple d'une enquête sur l'éducation réalisée auprès des élèves. Cependant, dans le cas de certaines enquêtes polyvalentes, la structure hiérarchique du plan de sondage pourrait être assez différente de la hiérarchie du modèle. Par exemple, l'Enquête longitudinale nationale auprès des enfants et des jeunes au Canada est réalisée selon un plan de sondage à plusieurs degrés où les degrés correspondent aux régions géographiques, aux

ménages dans une région et aux élèves dans un ménage, tandis qu'un modèle multiniveaux de l'éducation peut comprendre comme niveau les élèves, les classes, les écoles et les commissions scolaires (Rao et Roberts 1998). Puisque les grappes du plan de sondage recoupent les grappes du modèle pour ce genre d'enquête, il est difficile d'élaborer une méthode pondérée selon le plan de sondage appropriée d'inférence sur les paramètres du modèle qui permet de tenir compte de l'échantillonnage informatif des grappes et/ou des éléments dans les grappes échantillonnées. Sous échantillonnage informatif, le modèle supposé pour la population n'est pas nécessairement vérifié pour l'échantillon.

Soit N le nombre d'unités de niveau 2 dans la population et M_i, le nombre d'unités de niveau 1 dans l'unité i de niveau 2. Un modèle de superpopulation à deux niveaux est donné par

$$y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i \sim_{ind} f\left(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1\right), \ \mathbf{v}_i \sim_{iid} f\left(\mathbf{v}_i \mid \boldsymbol{\theta}_2\right), i = 1,...,N \ ; j = 1,...,M_i, \tag{2.1}$$

où y_{ij} et $\mathbf{x}_{ij} = \left(x_{ij0},...,x_{ij,p-1}\right)^T$ sont la réponse et le vecteur de dimension p des valeurs des covariables associés à l'élément j dans la grappe i et $x_{ij0} = 1$, $\mathbf{v}_i$ désigne un effet aléatoire de niveau 2, et $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$ désignent les paramètres associés aux deux degrés du modèle supposé. Ici $f\left(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1\right)$ et $f\left(\mathbf{v}_i \mid \boldsymbol{\theta}_2\right)$ sont les densités de probabilité spécifiées de y_{ij} sachant $\mathbf{x}_{ij}$ et $\mathbf{v}_i$, et de $\mathbf{v}_i$, respectivement. Notons, que, dans le modèle (2.1), les réponses y_{ij} d'une unité i donnée sont supposées être conditionnellement indépendantes sachant l'effet aléatoire $\mathbf{v}_i$, mais elles sont corrélées marginalement en raison de l'effet aléatoire $\mathbf{v}_i$ commun. La formulation du modèle (2.1) englobe à la fois les modèles à deux niveaux linéaires et les modèles à deux niveaux linéaires généralisés. Sous échantillonnage informatif des grappes et/ou des éléments dans les grappes échantillonnées, les méthodes classiques applicables aux modèles multiniveaux qui ne tiennent pas compte du plan de sondage et supposent que le modèle (2.1) est vérifié pour l'échantillon peuvent produire des estimateurs asymptotiquement biaisés des paramètres du modèle $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$ (Pfeffermann et coll. 1998).

Cas particuliers

1) Un simple modèle de la moyenne à erreurs emboîtées souvent utilisé dans les études en simulation portant sur les modèles à deux niveaux est donné par

$$y_{ij} = \mu + v_i + e_{ij}, e_{ij} \sim_{iid} N\left(0, \sigma_e^2\right), v_i \sim_{iid} N\left(0, \sigma_v^2\right), \tag{2.2}$$

où $i = 1,...,N; j = 1,...,M_i$. Le modèle (2.2) peut être écrit sous la forme (2.1) comme

$$y_{ij} \mid v_i \sim_{ind} N\left(\mu + v_i, \sigma_e^2\right), v_i \sim_{iid} N\left(0, \sigma_v^2\right), \boldsymbol{\theta}_1 = \left(\mu, \sigma_e^2\right), \theta_2 = \sigma_v^2.$$

Marginalement, $y_{ij} \sim N\left(\mu, \sigma_v^2 + \sigma_e^2\right)$ mais y_{ij} et y_{ik} $(j \neq k)$ sont corrélées : $\text{corr}\left(y_{ij}, y_{ik}\right) = \rho = \sigma_v^2 \Big/ \left(\sigma_v^2 + \sigma_e^2\right), \ j \neq k.$

2) Un modèle linéaire à deux niveaux, souvent utilisé en pratique, est donné par

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_i + e_{ij}, i = 1,...,N; j = 1,...,M_i, \tag{2.3}$$

où $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i, \mathbf{v}_i \sim_{iid} N_p(\mathbf{0}, \boldsymbol{\Sigma}_v), i = 1,...,N$ et $e_{ij} \sim_{iid} N(0, \sigma_e^2)$. Ce modèle peut également être exprimé sous la forme (2.1) comme

$$y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i \sim_{ind} N(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{v}_i, \sigma_e^2), \mathbf{v}_i \sim_{iid} N_p(0, \boldsymbol{\Sigma}_v) \tag{2.4}$$

où $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}^T, \sigma_e^2)^T$ et $\boldsymbol{\theta}_2$ est le vecteur des $p(p+1)/2$ éléments distincts de $\boldsymbol{\Sigma}_v$. Marginalement, $y_{ij} \sim N(\mathbf{x}_{ij}^T \boldsymbol{\beta}, \mathbf{x}_{ij}^T \boldsymbol{\Sigma}_v \mathbf{x}_{ij} + \sigma_e^2)$, mais y_{ij} et y_{ik} $(j \neq k)$ sont corrélées en raison de l'effet aléatoire commun $\mathbf{v}_i$. Cependant, dans le cas d'un modèle linéaire généralisé à deux niveaux, la loi marginale de y_{ij} ne donne généralement pas une expression analytique : par exemple, dans le cas d'un modèle linéaire logistique à deux niveaux pour réponses binaires.

2.2 Estimation ponctuelle

La log-vraisemblance de « recensement » ou de population sous le modèle à deux niveaux supposé (2.1) est donnée par

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log L_i(\boldsymbol{\theta}) \equiv \sum_{i=1}^{N} l_i(\boldsymbol{\theta}) = l(\boldsymbol{\theta}), \tag{2.5}$$

où $\boldsymbol{\theta}$ est le vecteur comprenant les éléments $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$, et

$$L_i(\boldsymbol{\theta}) = \int \exp\left[\sum_{j=1}^{M_i} \log f\left(y_{ij} \middle| \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1\right)\right] f\left(\mathbf{v}_i \mid \boldsymbol{\theta}_2\right) d\mathbf{v}_i \tag{2.6}$$

voir Asparouhov (2006) et Rabe-Hesketh et Skrondal (2006). La fonction de score de recensement $\mathbf{U}(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ satisfait $E_m\{\mathbf{U}(\boldsymbol{\theta})\} = \mathbf{0}$, où E_m désigne l'espérance sous le modèle. Le paramètre de recensement $\boldsymbol{\theta}_N$ est défini comme la solution unique de $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$ et $\boldsymbol{\theta}_N$ est convergent sous le modèle pour $\boldsymbol{\theta}$, où $\boldsymbol{\theta}_N$ est le vecteur des éléments $\boldsymbol{\theta}_{1N}$ et $\boldsymbol{\theta}_{2N}$.

Soit l'échantillon constitué de n grappes avec m_i éléments provenant de la grappe échantillonnée i. Soit π_i et $\pi_{j|i}$ les probabilités d'inclusion de niveau 2 et de niveau 1, respectivement, associées à la grappe i et à l'élément j dans la grappe i. Alors, les pondérations de niveau 2 et de niveau 1 sont données par $w_i = \pi_i^{-1}$ et $w_{j|i} = \pi_{j|i}^{-1}$, respectivement. Asparouhov (2006) et Rabe-Hesketh et Skrondal (2006) ont proposé une pseudo log-vraisemblance d'échantillon pondérée obtenue en remplaçant $\sum_{j=1}^{M_i}(.)$ dans (2.6) par $\sum_{j \in s_i} w_{j|i}(.)$ et $\sum_{i=1}^{N}(.)$ dans (2.5) par $\sum_{i \in s} w_i(.)$, où s désigne l'échantillon de grappes et $s(i)$ désigne l'échantillon d'éléments dans les grappes $i \in s$. Elle est donnée par

$$\tilde{l}_w(\boldsymbol{\theta}) = \sum_{i \in s} w_i \tilde{l}_{wi}(\boldsymbol{\theta}) \tag{2.7}$$

où $\tilde{l}_{wi}(\boldsymbol{\theta}) = \log \tilde{L}_{wi}(\boldsymbol{\theta})$ et

$$\tilde{L}_{wi}(\boldsymbol{\theta}) = \int \exp\left[\sum_{j \in s(i)} w_{j|i} \log f(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1) \right] f(\mathbf{v}_i \mid \boldsymbol{\theta}_2) d\mathbf{v}_i . \quad (2.8)$$

En maximisant la pseudo log-vraisemblance $\tilde{l}_w(\boldsymbol{\theta})$ donnée par (2.7), nous obtenons un estimateur du pseudo maximum de vraisemblance (PMV) $\tilde{\boldsymbol{\theta}}_w$. Les calculs sont exposés en détail dans Asparouhov (2006) et dans Rabe-Hesketh et Skrondal (2006). Dans le cas particulier des modèles linéaires à deux niveaux, Pfeffermann et coll. (1998) ont utilisé une méthode par les moindres carrés généralisés itérative proposée par Goldstein (1986). Notons que nous avons besoin des pondérations de niveau 1 et de niveau 2 pour calculer $\tilde{\boldsymbol{\theta}}_w$, contrairement au cas des modèles marginaux qui nécessitent seulement les pondérations non conditionnelles des éléments $w_{ij} = w_i w_{j|i}$.

La convergence sous le plan de sondage de l'estimateur PMV $\tilde{\boldsymbol{\theta}}_{2w}$ du paramètre de recensement $\boldsymbol{\theta}_{2N}$ ou la convergence sous le plan et sous le modèle de $\tilde{\boldsymbol{\theta}}_{2w}$ en tant qu'estimateur du paramètre du modèle $\boldsymbol{\theta}_2$ requiert que le nombre de grappes échantillonnées, n, ainsi que la taille d'échantillon dans les grappes, m_i, tendent vers l'infini, même dans le cas linéaire. En outre, le biais relatif des estimateurs sera important si les tailles d'échantillon m_i sont petites. Pour remédier à ce problème, plusieurs méthodes de rajustement des pondérations ont été proposées dans la littérature. En particulier, un facteur de mise à l'échelle k_{1i} est appliqué aux pondérations de niveau 1 $w_{j|i}$ dans (2.8) avant de maximiser la pseudo log-vraisemblance (2.7). Nous ne considérons ici que deux méthodes de rajustement des pondérations, désignées A et A1 (Asparouhov 2006). La méthode A utilise

$$k_{1i} = m_i \Big/ \sum_{j \in s(i)} w_{j|i} \quad (2.9)$$

Dans la méthode A1, k_{1i} est le même que dans la méthode A, mais les pondérations de niveau 2 w_i sont également rajustées au moyen du facteur $k_{2i} = 1/k_{1i}$ pour compenser le rajustement des pondérations de niveau 1. Asparouhov (2006) a mentionné l'utilisation d'un algorithme EM accéléré pour calculer l'estimateur PMV $\tilde{\boldsymbol{\theta}}_w$ avec Mplus 3 : www.statmodel.com : Muthén et Muthén, 1998-2005.

2.3 Estimation de la variance

En ce qui concerne l'estimation de la variance, Asparouhov (2006) a proposé un estimateur de variance « sandwich » par linéarisation de Taylor de $\tilde{\boldsymbol{\theta}}_w$, qui est donné par

$$v_L\left(\tilde{\boldsymbol{\theta}}_w\right) = \left(\tilde{\mathbf{I}}''_w\right)^{-1} \left[\sum_{i \in s} \left(k_{2i} w_i\right)^2 \tilde{\mathbf{I}}'_{wi} \left(\tilde{\mathbf{I}}'_{wi}\right)^T \right] \left(\tilde{\mathbf{I}}''_w\right)^{-1}, \quad (2.10)$$

où $\tilde{\mathbf{I}}'_w$ et $\tilde{\mathbf{I}}''_w$ désignent, respectivement, le vecteur des dérivées premières et la matrice des dérivées secondes de $\tilde{l}_w(\boldsymbol{\theta})$ évaluées à $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_w$, et $\tilde{\mathbf{I}}'_{wi}$ est la dérivée première de $\tilde{l}_{wi}(\boldsymbol{\theta})$ évaluée à $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_w$. Si la fraction d'échantillonnage de niveau 2 est faible, alors $v_L\left(\tilde{\boldsymbol{\theta}}_w\right)$ suit bien la variance de $\tilde{\boldsymbol{\theta}}_w$, mais non l'EQM de $\tilde{\boldsymbol{\theta}}_w$ si le biais relatif de $\tilde{\boldsymbol{\theta}}_w$ est grand.

Kovacevic et coll. (2006) ont étudié les estimateurs bootstrap de la variance de $\tilde{\boldsymbol{\theta}}_w$. Ils ont considéré deux options. L'option 1 consiste à utiliser les poids bootstrap de niveau 2 $w_i(b)$ basés sur la méthode de

Rao, Wu et Yue (1992) et à ne pas modifier les poids de niveau 1, c'est-à-dire $w_{j|i}(b) = w_{j|i}$, où $b = 1,...,B$ désigne les B échantillons bootstrap. L'option 2 consiste à appliquer la méthode du bootstrap de Rao, Wu et Yue (1992) au niveau 1 ainsi qu'au niveau 2, et à rajuster les poids bootstrap de niveau 1. En remplaçant les poids w_i et $w_{j|i}$ par $w_i(b)$ et $w_{j|i}(b)$ dans (2.7) et (2.8), on obtient les estimateurs bootstrap PMV $\tilde{\boldsymbol{\theta}}_w(b)$, $b = 1,...,B$ et l'estimateur bootstrap de la variance est donné par

$$v_{Boot}\left(\tilde{\boldsymbol{\theta}}_w\right) = \frac{1}{B}\sum_{b=1}^{B}\left[\tilde{\boldsymbol{\theta}}_w(b) - \tilde{\boldsymbol{\theta}}_w\right]\left[\tilde{\boldsymbol{\theta}}_w(b) - \tilde{\boldsymbol{\theta}}_w\right]^T. \tag{2.11}$$

Une étude en simulation de (2.11), fondée sur le simple modèle de la moyenne (2.2), a montré que l'option 1 peut donner lieu à une sous-estimation de la variance de $\tilde{\sigma}_{ew}^2$. L'option 2 a donné de meilleurs résultats que l'option 1. Grilli et Pratesi (2004) ont étudié une autre méthode bootstrap pour l'estimation de la variance.

3 Équations d'estimation pondérées par les poids de sondage

Aux sections 3 et 4, nous étudions les méthodes d'établissement des équations d'estimation pondérées par les poids de sondage pour les paramètres des modèles multiniveaux qui conduisent à des estimateurs convergents sous le plan et sous le modèle, même lorsque les tailles d'échantillon dans les grappes sont petites. Les méthodes proposées dépendent uniquement des probabilités d'inclusion d'ordre un π_i et $\pi_{j|i}$, et des probabilités d'inclusion conjointe $\pi_{jk|i}$ dans les grappes. À la section 3, nous présentons une approche simple, fondée sur les moments, des équations d'estimation pondérées, qui est applicable aux modèles de régression linéaires à erreurs emboîtées. À la section 4, nous proposons une méthode unifiée, fondée sur les log-vraisemblances composites pondérées. Cette méthode permet de traiter les modèles multiniveaux linéaires ainsi que linéaires généralisés, contrairement à la méthode fondée sur les moments, et elle aboutit à des estimateurs convergents sous le plan et sous le modèle. Elle ne dépend, elle aussi, que de π_i, $\pi_{j|i}$ et $\pi_{jk|i}$.

3.1 Estimation ponctuelle

Nous commençons par illustrer l'approche des équations d'estimation pondérées, en utilisant le simple modèle de la moyenne (2.2). Ici, nous voulons estimer $\boldsymbol{\theta} = \left(\mu, \sigma_v^2, \sigma_e^2\right)^T$ en partant d'un plan d'échantillonnage en grappes à deux degrés qui concorde avec la hiérarchie du modèle. Nous avons choisi pour cela les trois fonctions d'estimation (FE) suivantes :

$$u_1\left(y_{ij}, \boldsymbol{\theta}\right) = y_{ij} - \mu, \tag{3.1}$$

$$u_2\left(y_{ij}, \boldsymbol{\theta}\right) = \left(y_{ij} - \mu\right)^2 - \left(\sigma_v^2 + \sigma_e^2\right) \tag{3.2}$$

$$u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = \left[\left(y_{ij} - \mu\right) - \left(y_{ik} - \mu\right)\right]^2 - 2\sigma_e^2 = z_{ijk}{}^2 - 2\sigma_e^2, j \neq k, \tag{3.3}$$

où $z_{ijk} = y_{ij} - y_{ik}$. Les équations d'estimation de recensement correspondantes sont données par

$$U_1(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{M_i} u_1\left(y_{ij}, \boldsymbol{\theta}\right) = 0, U_2(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{M_i} u_2\left(y_{ij}, \boldsymbol{\theta}\right) = 0 \tag{3.4}$$

$$U_3(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j<k=1}^{M_i} u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = 0\,. \tag{3.5}$$

Le paramètre de recensement résultant, $\tilde{\boldsymbol{\theta}}_N$, est convergent sous le modèle pour $\boldsymbol{\theta}$ parce que les espérances sous le modèle des trois fonctions d'estimation (3.1) à (3.3) sont nulles. Il découle de (3.4) et (3.5) que les équations d'estimation pondérées par les poids de sondage (EEP) sont données par

$$\hat{U}_{w1}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i} u_1\left(y_{ij}, \boldsymbol{\theta}\right) \equiv \sum_{i\in s} w_i \hat{U}_{w1i}(\boldsymbol{\theta}) = 0 \tag{3.6}$$

$$\hat{U}_{w2}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i} u_2\left(y_{ij}, \boldsymbol{\theta}\right) \equiv \sum_{i\in s} w_i \hat{U}_{w2i}(\boldsymbol{\theta}) = 0 \tag{3.7}$$

$$\hat{U}_{w3}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) \equiv \sum_{i\in s} w_i \hat{U}_{w3i}(\boldsymbol{\theta}) = 0\,, \tag{3.8}$$

où $w_{jk|i} = \pi_{jk|i}^{-1}$. L'estimateur EEP, $\hat{\boldsymbol{\theta}}_w$, est obtenu en résolvant le système d'équations (3.6) à (3.8). Pour le modèle de la moyenne, nous obtenons les solutions explicites des EEP suivantes

$$\hat{\mu}_w = \left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij} y_{ij}\right) \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij} \equiv \bar{y}_w \tag{3.9}$$

$$\hat{\sigma}_{vw}^2 = \sum_{i\in s}\sum_{j\in s(i)} w_{ij}\left(y_{ij} - \bar{y}_w\right)^2 \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij} - \hat{\sigma}_{ew}^2 \tag{3.10}$$

$$\hat{\sigma}_{ew}^2 = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} z_{ijk}^2 \Big/ \left(2\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\right), \tag{3.11}$$

où $w_{ij} = w_i w_{j|i}$. Soulignons que la méthode des moments susmentionnés ne dépend pas de la loi de probabilité.

Nous notons que $\hat{U}_{wt}(\boldsymbol{\theta}), t = 1,2,3$ sont les fonctions d'estimation d'espérance nulle par rapport au plan de sondage et au modèle, c'est-à-dire $E_m E_p\left\{\hat{U}_{wt}(\boldsymbol{\theta})\right\} = 0$. En utilisant ce résultat, nous pouvons montrer que l'estimateur EEP $\hat{\boldsymbol{\theta}}_w = \left(\hat{\mu}_w, \hat{\sigma}_{vw}^2, \hat{\sigma}_{ew}^2\right)^T$ est convergent sous le plan et sous le modèle pour $\boldsymbol{\theta}$ à mesure que le nombre d'unités de niveau 2 dans l'échantillon, n, augmente, même si les tailles d'échantillon dans les grappes, m_i, sont petites. Cette propriété n'est pas nécessairement vérifiée pour les estimateurs présentés à la section 2. La méthode proposée nécessite toutefois les probabilités d'inclusion

conjointe dans les grappes $\pi_{jk|i}$. Ces probabilités sont obtenues facilement pour l'échantillonnage aléatoire simple ou stratifié dans les grappes, ou quand la fraction d'échantillonnage dans les grappes est faible. En outre, plusieurs bonnes approximations de $\pi_{jk|i}$ lorsque l'échantillonnage dans les grappes est effectué avec probabilités inégales sont disponibles, et ces approximations dépendent uniquement des probabilités d'inclusion marginales $\pi_{j|i}$ (Haziza, Mecatti et Rao 2008). L'estimateur EEP $\hat{\boldsymbol{\theta}}_w$ est également convergent sous le plan pour $\tilde{\boldsymbol{\theta}}_N$, en notant que $E_p\left\{\hat{U}_{wt}(\tilde{\boldsymbol{\theta}}_N)\right\} = 0$, $t = 1,2,3$.

Le choix des fonctions d'estimation (3.1) à (3.3) n'est pas forcément unique. Ainsi, nous pourrions remplacer l'équation précédente $u_2\left(y_{ij},\boldsymbol{\theta}\right)$ par $\tilde{u}_2\left(y_{ij},y_{ik},\boldsymbol{\theta}\right) = \left(y_{ij}-\mu\right)\left(y_{ik}-\mu\right) - \sigma_v^2$ dans (3.7) et garder (3.6) et (3.8). L'estimateur EEP résultant est également convergent sous le plan et sous le modèle pour $\boldsymbol{\theta}$ à mesure que le nombre d'unités de niveau 2 augmente. L'approche de la vraisemblance composite par paire pondérée décrite à la section 4 offre une méthode unifiée de génération des fonctions d'estimation.

Korn et Graubard (2003) ont utilisé pour le modèle de la moyenne une autre approche qui présente certaines similarités avec l'approche proposée. Sous leur approche, les « paramètres de recensement », S_e^2 et S_v^2, sont d'abord obtenus en supposant que le modèle est vérifié pour la population finie. Les estimateurs pondérés par les poids de sondage $\hat{S}_{ew}^2$ et $\hat{S}_{vw}^2$ des paramètres de recensement sont ensuite obtenus en supposant que M_i est connu pour les grappes échantillonnées. L'estimateur $\hat{S}_{ew}^2$ est donné par

$$\hat{S}_{ew}^2 = \left\{\frac{1}{2}\sum_{i\in s}\left(M_i-1\right)w_i\left[\sum_{j<k\in s(i)} w_{jk|i}\left(y_{ij}-y_{ik}\right)^2 \Big/ \sum_{j<k\in s(i)} w_{jk|i}\right]\right\}\left[\sum_{i\in s}\left(M_i-1\right)w_i\right]^{-1}, \qquad (3.12)$$

en supposant que $m_i > 1$ pour toutes les grappes échantillonnées. Notons que (3.12) nécessite les probabilités d'inclusion conjointe $\pi_{jk|i}$ comme la méthode proposée, mais qu'il induit un biais de ratio intra-grappe lorsque les tailles d'échantillon dans les grappes sont faibles, contrairement à notre méthode. L'expression pour $\hat{S}_{vw}^2$ est plus compliquée et nous invitons le lecteur à consulter Korn et Graubard (2003) pour la formule pertinente.

La méthode EEP peut être étendue facilement au modèle de régression linéaire à erreurs emboîtées

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij};\ e_{ij} \sim_{iid} N\left(0,\sigma_e^2\right),\ v_i \sim_{iid} N\left(0,\sigma_v^2\right). \qquad (3.13)$$

Dans ce cas, la fonction d'estimation (3.1) devient

$$u_1\left(y_{ij},\boldsymbol{\theta}\right) = \mathbf{x}_{ij}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right), \qquad (3.14)$$

La fonction d'estimation (3.2) devient

$$u_2\left(y_{ij},\boldsymbol{\theta}\right) = \left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2 - \left(\sigma_v^2+\sigma_e^2\right) \qquad (3.15)$$

et la fonction d'estimation (3.3) devient

$$u_3\left(y_{ij}, y_{ik}, \boldsymbol{\theta}\right) = \left[z_{ijk} - \left(\mathbf{x}_{ij} - \mathbf{x}_{ik}\right)^T \boldsymbol{\beta}\right]^2 - 2\sigma_e^2, \; j \neq k, \tag{3.16}$$

où $\boldsymbol{\theta}$ est le vecteur des éléments $\boldsymbol{\beta}$, σ_v^2 et σ_e^2 et $z_{ijk} = y_{ij} - y_{ik}$. Les solutions explicites de $\hat{U}_{wt}(\boldsymbol{\theta}) = 0, t = 1,2,3$ correspondant aux équations (3.14) à (3.16) sont obtenues sous la forme

$$\hat{\boldsymbol{\beta}}_w = \left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T\right)^{-1}\left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}y_{ij}\right), \tag{3.17}$$

$$\hat{\sigma}_{vw}^2 = \sum_{i\in s}\sum_{j\in s(i)} w_{ij}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_w\right)^2 \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij} - \hat{\sigma}_{ew}^2 \tag{3.18}$$

et

$$\hat{\sigma}_{ew}^2 = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\left[z_{ijk} - \left(\mathbf{x}_{ij} - \mathbf{x}_{ik}\right)^T\hat{\boldsymbol{\beta}}_w\right]^2 \Big/ \left(2\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\right). \tag{3.19}$$

3.2 Estimation de la variance

Un estimateur sandwich par linéarisation de Taylor de la variance de l'estimateur EEP $\hat{\boldsymbol{\theta}}_w$ peut être obtenu de manière analogue à l'estimateur de variance (2.10), à condition que la fraction d'échantillonnage de niveau 2 soit faible. Soit $\hat{\mathbf{U}}_w(\boldsymbol{\theta})$ le vecteur colonne dont les composantes sont $\hat{U}_{w1}(\boldsymbol{\theta})$, $\hat{U}_{w2}(\boldsymbol{\theta})$ et $\hat{U}_{w3}(\boldsymbol{\theta})$, et similairement $\hat{\mathbf{U}}_{wi}(\boldsymbol{\theta})$ le vecteur colonne dont les composantes sont $\hat{U}_{w1i}(\boldsymbol{\theta})$, $\hat{U}_{w2i}(\boldsymbol{\theta})$ et $\hat{U}_{w3i}(\boldsymbol{\theta})$. Alors, l'estimateur de variance par linéarisation est donné par

$$v_L\left(\hat{\boldsymbol{\theta}}_w\right) = \left(\hat{\mathbf{U}}'_w\right)^{-1}\left(\sum_{i\in s} w_i^2\,\hat{\mathbf{U}}_{wi}\hat{\mathbf{U}}_{wi}^T\right)\left[\left(\hat{\mathbf{U}}'_w\right)^{-1}\right]^T, \tag{3.20}$$

où $\hat{\mathbf{U}}_{wi}$ et $\hat{\mathbf{U}}'_w$ désignent $\hat{\mathbf{U}}_{wi}(\boldsymbol{\theta})$ évalué à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_w$, et la dérivée première $\hat{\mathbf{U}}'_w(\boldsymbol{\theta})$ évaluée à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_w$, respectivement. Les propriétés de l'estimateur de variance (3.20) sont étudiées par simulation à la section 5.2.

4 Log-vraisemblance composite pondérée : une approche unifiée

À la présente section, nous proposons une approche unifiée applicable aux modèles multiniveaux linéaires ainsi que linéaires généralisés. Cette approche est fondée sur le concept de la vraisemblance composite qui a acquis de la popularité dans la littérature ne portant pas sur les sondages pour traiter les données en grappes ou les données spatiales (voir par exemple, Lindsay 1988, Lele et Taper 2002 et Varin, Reid et Firth 2011). Une vraisemblance composite marginale par paire s'obtient en multipliant les contributions à la vraisemblance de toutes les paires distinctes dans les grappes. Notons que la vraisemblance composite est obtenue en prétendant que les sous-modèles sont indépendants. Lorsque le modèle de superpopulation est vérifié pour l'échantillon, nous pouvons obtenir les estimateurs des

paramètres en maximisant la vraisemblance composite par paire. Ici, nous étendons cette approche aux plans de sondage informatifs en obtenant des équations d'estimation pondérées qui requièrent seulement les poids marginaux w_i et $w_{j|i}$ et les poids par paire $w_{jk|i}$, comme à la section 3.

La log vraisemblance composite par paire de recensement est donnée par

$$l_C(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j<k=1}^{M_i} \log f\left(y_{ij},\, y_{ik}\middle|\boldsymbol{\theta}\right), \tag{4.1}$$

où $f\left(y_{ij},\, y_{ik}\middle|\boldsymbol{\theta}\right)$ est la densité de probabilité conjointe marginale de y_{ij} et y_{ik}. Nous estimons (4.1) par la log-vraisemblance composite par paire pondérée par les poids de sondage

$$l_{wC}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} \log f\left(y_{ij},\, y_{ik}\middle|\boldsymbol{\theta}\right) \tag{4.2}$$

qui dépend seulement des probabilités d'inclusion de niveau 1 et de niveau 2 d'ordre 1 et de probabilités d'inclusion de niveau 1 d'ordre 2. Puis, nous résolvons les équations de score composite pondérées

$$\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta}) = \partial l_{wC}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}, \tag{4.3}$$

provenant de (4.2) pour obtenir un estimateur de la vraisemblance composite pondérée, $\hat{\boldsymbol{\theta}}_{wC}$, de $\boldsymbol{\theta}$. La méthode proposée est applicable aux modèles à deux niveaux linéaires et linéaires généralisés.

Nous notons que $\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta})$, donné par (4.3), est un vecteur de fonctions d'estimation d'espérance nulle par rapport au plan et au modèle, c'est-à-dire $E_m E_p\left\{\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta})\right\} = \mathbf{0}$. En utilisant ce résultat, on peut montrer que l'estimateur de la vraisemblance composite pondérée (VCP) $\hat{\boldsymbol{\theta}}_{wC}$ de $\boldsymbol{\theta}$ est convergent sous le modèle quand le nombre d'unités de niveau 2 dans l'échantillon, n, augmente, même si les tailles d'échantillon dans les grappes, m_i, sont petites. La preuve est exposée en détail dans Yi, Rao et Li (2012). Dans le contexte ne faisant pas appel au sondage, les preuves théoriques et empiriques que l'approche de la vraisemblance composite conduit à des estimateurs efficaces sont limitées (par exemple, Bellio et Varin 2005, Lindsay et coll. 2011). Notre étude en simulation (section 5) indique que l'approche de la vraisemblance composite pondérée donne de bons résultats en ce qui concerne l'efficacité, même si les tailles d'échantillon dans les grappes sont petites.

Dans le cas du modèle à erreurs emboîtées (3.13), en nous inspirant de Lele et Taper (2002), nous pouvons simplifier l'approche de la vraisemblance composite par paire en remplaçant la densité de probabilité bivariée $f\left(y_{ij}, y_{ik}\middle|\boldsymbol{\theta}\right)$ par les densités de probabilité univariées de y_{ij} et la différence $z_{ijk} = y_{ij} - y_{ik}$. Pour le modèle de la moyenne (2.2), nous avons $y_{ij} \sim N\left(\mu,\, \sigma_v^2 + \sigma_e^2\right)$ et $z_{ijk} \sim N\left(0,\, 2\sigma_e^2\right)$. En reparamétrisant $\boldsymbol{\theta} = \left(\mu, \sigma_v^2, \sigma_e^2\right)^T$ de manière que $\boldsymbol{\phi} = \left(\mu, \sigma^2, \sigma_e^2\right)^T$, où $\sigma^2 = \sigma_v^2 + \sigma_e^2$, nous voyons que les paramètres des deux densités de probabilité univariées sont distincts et que les log-vraisemblances composites correspondant à y_{ij} et z_{ijk} sont données par

$$l_{wCy}\left(\mu, \sigma^2\right) = \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i} \log f\left(y_{ij}\middle|\mu, \sigma^2\right)$$

et

$$l_{wCz}\left(\sigma_e^2\right) = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} \log f\left(z_{ijk}\middle|\sigma_e^2\right).$$

Nous résolvons alors le système d'équations de score composite pondérées résultantes

$$\hat{U}_{wCy1}\left(\mu,\sigma^2\right) = \partial l_{wCy}\left(\mu,\sigma^2\right)\Big/\partial\mu = \sum_{i\in s} w_i \sum_{j\in s_i} w_{j|i}\left(y_{ij}-\mu\right)\Big/\sigma^2 = 0,$$

$$\hat{U}_{wCy2}\left(\mu,\sigma^2\right) = \partial l_{wCy}\left(\mu,\sigma^2\right)\Big/\partial\sigma^2 = \frac{1}{2}\sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i}\left[-\frac{1}{\sigma^2}+\frac{\left(y_{ij}-\mu\right)^2}{\sigma^4}\right] = 0$$

$$\hat{U}_{wCz}\left(\sigma_e^2\right) = \partial l_{wCz}\left(\sigma_e^2\right)\Big/\partial\sigma_e^2 = \frac{1}{2}\sum_{i} w_i \sum_{j<k\in s(i)} w_{jk|i}\left(-\frac{1}{\sigma_e^2}+\frac{z_{ijk}^2}{2\sigma_e^4}\right) = 0$$

pour obtenir les estimateurs de la vraisemblance composite pondérée (VCP) $\hat{\mu}_{wC}, \hat{\sigma}_{vwC}^2$ et $\hat{\sigma}_{ewC}^2$. Les estimateurs VCP sont identiques aux estimateurs (3.9) à (3.11) obtenus par l'approche des équations d'estimation pondérées de la section 3.

Nous nous penchons maintenant sur le modèle de régression linéaire à erreurs emboîtées (3.13). Mentionnons pour commencer que $y_{ij} \sim N\left(\mathbf{x}_{ij}^T\boldsymbol{\beta},\, \sigma^2\right)$, où $\sigma^2 = \sigma_v^2 + \sigma_e^2$, et $z_{ijk} = y_{ij} - y_{ik} \sim N\left\{\left(\mathbf{x}_{ij}-\mathbf{x}_{ik}\right)^T\boldsymbol{\beta},\, 2\sigma_e^2\right\}$. Il s'ensuit que les équations de score composite pondérées sont données par

$$\begin{aligned}\hat{\mathbf{U}}_{wCy1}(\boldsymbol{\beta},\sigma^2) &= \partial l_{wCy}\left(\boldsymbol{\beta},\sigma^2\right)\Big/\partial\boldsymbol{\beta} \\ &= \sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i}\mathbf{x}_{ij}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right) = \mathbf{0}\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{U}}_{wCy2}(\boldsymbol{\beta},\sigma^2) &= \partial l_{wCy}\left(\boldsymbol{\beta},\sigma^2\right)\Big/\partial\sigma^2 \\ &= -\frac{1}{2}\sum_{i\in s} w_i \sum_{j\in s(i)} w_{j|i}\left[\frac{1}{\sigma^2}-\frac{\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2}{\sigma^4}\right] = 0\end{aligned}$$

et

$$\begin{aligned}\hat{\mathbf{U}}_{wCz}(\sigma_e^2) &= \partial l_{wCz}\left(\sigma_e^2\right)\Big/\partial\sigma_e^2 \\ &= -\frac{1}{2}\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\left\{\frac{1}{\sigma_e^2}-\frac{\left[z_{ijk}-\left(\mathbf{x}_{ij}-\mathbf{x}_{ik}\right)^T\boldsymbol{\beta}\right]^2}{2\sigma_e^4}\right\} = 0.\end{aligned}$$

Les estimateurs VCP résultants de $\boldsymbol{\beta}$, σ_v^2 et σ_e^2 sont donnés par

$$\hat{\boldsymbol{\beta}}_{wC} = \left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T\right)^{-1}\left(\sum_{i\in s}\sum_{j\in s(i)} w_{ij}\mathbf{x}_{ij}y_{ij}\right),$$

$$\hat{\sigma}_{wC}^2 = \sum_{i\in s}\sum_{j\in s(i)} w_{ij}\left(y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{wC}\right)^2 \Big/ \sum_{i\in s}\sum_{j\in s(i)} w_{ij},$$

et

$$\hat{\sigma}^2_{ewC} = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\left[z_{ijk} - \left(\mathbf{x}_{ij} - \mathbf{x}_{ik}\right)^T \hat{\boldsymbol{\beta}}_{wC} \right]^2 \Bigg/ \left(2\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i} \right).$$

L'estimateur de σ_v^2 est donné par $\hat{\sigma}^2_{vwC} = \hat{\sigma}^2_{wC} - \hat{\sigma}^2_{ewC}$. De nouveau, les estimateurs VCP $\hat{\boldsymbol{\beta}}_{WC}$, $\hat{\sigma}^2_{vWC}$ et $\hat{\sigma}^2_{eWC}$ sont identiques aux estimateurs (3.17) à (3.19) obtenus par l'approche des équations d'estimation pondérées de la section 3.

L'approche de la vraisemblance composite susmentionnée, fondée sur y_{ij} et $z_{ijk} = y_{ij} - y_{ik}$, n'est pas applicable au modèle à deux niveaux linéaire donné par (2.4), parce que le vecteur de paramètres, $\boldsymbol{\theta}$, n'est pas identifiable sous la vraisemblance composite obtenue à partir des y_{ij} et z_{ijk}. Nous devons faire appel à la méthode par paire pour traiter le modèle (2.4).

Marginalement, $\left(y_{ij}, y_{ik}\right)^T$ suit une loi normale bivariée de moyennes $\mathbf{x}_{ij}^T\boldsymbol{\beta}$ et $\mathbf{x}_{ik}^T\boldsymbol{\beta}$ et de matrice de covariance 2×2

$$\boldsymbol{\Sigma}_{i(jk)} = \begin{bmatrix} \sigma_e^2 + \mathbf{x}_{ij}^T\boldsymbol{\Sigma}_v\mathbf{x}_{ij} & \mathbf{x}_{ij}^T\boldsymbol{\Sigma}_v\mathbf{x}_{ik} \\ \mathbf{x}_{ik}^T\boldsymbol{\Sigma}_v\mathbf{x}_{ij} & \sigma_e^2 + \mathbf{x}_{ik}^T\boldsymbol{\Sigma}_v\mathbf{x}_{ik} \end{bmatrix}.$$

Maintenant, il découle de (4.3) que les équations de score composite pondérées sont données par

$$\boldsymbol{\beta}: \qquad \hat{\mathbf{U}}_{wC\boldsymbol{\beta}} = \sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\mathbf{X}_{i(jk)}^T\boldsymbol{\Sigma}_{i(jk)}^{-1}\left(\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T\boldsymbol{\beta}\right) = \mathbf{0} \tag{4.4}$$

et

$$\begin{aligned} \boldsymbol{\tau}: \qquad \hat{\mathbf{U}}_{wCl} = \frac{1}{2}\sum_{i\in s} w_i \sum_{j<k\in s(i)} w_{jk|i}\Bigg[& \left(\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_{i(jk)}^{-1}\frac{\partial\boldsymbol{\Sigma}_{i(jk)}}{\partial\tau_l}\boldsymbol{\Sigma}_{i(jk)}^{-1}\left(\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T\boldsymbol{\beta}\right) \\ & -\mathrm{tr}\left(\boldsymbol{\Sigma}_{i(jk)}^{-1}\frac{\partial\boldsymbol{\Sigma}_{i(jk)}}{\partial\tau_l}\right)\Bigg] = \mathbf{0}, \qquad l = 1,\ldots, p(p+1)/2 + 1 = P \end{aligned} \tag{4.5}$$

où $\mathbf{X}_{i(jk)}$ est la matrice de dimensions $2\times p$ contenant les lignes $\mathbf{x}_{ij}^T$ et $\mathbf{x}_{ik}^T$, $\mathbf{y}_{i(jk)} = \left(y_{ij}, y_{ik}\right)^T$, et $\boldsymbol{\tau}$ est le vecteur de dimension P contenant les éléments $\tau_1 = \sigma_e^2$ et les $p(p+1)/2$ éléments distincts de $\boldsymbol{\Sigma}_v$ désignés par $\tau_2,\ldots,\tau_P$. Nous pouvons résoudre les équations de score composite pondérées (4.4) et (4.5) itérativement en utilisant la méthode de Newton-Raphson ou une autre méthode itérative pour obtenir les estimateurs VCP $\hat{\boldsymbol{\beta}}_{wC}$ et $\hat{\boldsymbol{\tau}}_{wC}$.

Dans le cas particulier du modèle de régression linéaire à erreurs emboîtées (3.13), les équations de score de recensement, fondées sur la log-vraisemblance de recensement complète $l(\boldsymbol{\theta})$ donnée par (2.5), peuvent s'écrire sous une forme explicite. Les équations de score pondérées d'échantillon correspondantes ne dépendent que des poids de niveau 1 $w_{j|i}$ et $w_{jk|i}$ et des poids de niveau 2 w_i, comme les équations de score composite pondérées (voir l'annexe). Les estimateurs résultants sont convergents sous le modèle pour $\boldsymbol{\theta}$, contrairement aux estimateurs fondés sur la pseudo log-vraisemblance pondérée $l_w(\boldsymbol{\theta})$ donnés par (2.7) et (2.8). Cependant, pour des modèles plus complexes, comme les modèles à deux niveaux avec pentes aléatoires, les équations de score pondérées d'échantillon dépendront des probabilités d'inclusion de niveau 1 d'ordres 3 et 4, contrairement aux équations de score composite pondérées (4.3) qui ne

dépendent que des probabilités d'inclusion de niveau 1 d'ordres 1 et 2, même pour les modèles multiniveaux complexes. Par conséquent, nous n'avons pas inclus l'approche des équations de score pondérées fondée sur la log-vraisemblance de recensement complète dans l'étude en simulation.

5 Étude en simulation

Nous avons réalisé une petite étude en simulation de la performance des estimateurs EEP proposés sous le simple modèle de la moyenne à erreurs emboîtées, en utilisant $\mu = 0{,}5$, $\sigma_v^2 = 0{,}5$ et $\sigma_e^2 = 2{,}0$. La population est constituée de $N = 1\,000$ grappes, chacune contenant $M_i = M = 100$ éléments. Nous avons utilisé un plan d'échantillonnage à deux degrés avec $n = 50$ grappes échantillonnées et $m_i = m = 5$ éléments sélectionnés dans chaque grappe échantillonnées. Les grappes ont été sélectionnées par échantillonnage aléatoire simple, et les éléments dans les grappes, par la méthode d'échantillonnage avec probabilité proportionnelle à la taille (PPT) de Rao-Sampford (Rao 1965 et Sampford 1967) avec des mesures de taille spécifiées z_{ij}. Nous avons choisi les mesures de taille de manière à refléter les divers niveaux d'informativité de l'échantillonnage.

À l'instar d'Asparouhov (2006), nous prenons en considération la sélection invariante ainsi que non invariante. Pour la sélection invariante, la mesure de taille z_{ij} dépend seulement des erreurs de niveau 1 et ne varie pas d'une grappe à l'autre. En particulier, nous posons

$$z_{ij} = \left(1 + \exp\left\{-0.5\left[e_{ij}/\alpha + e_{ij}^{*}\left(1 - \alpha^{-2}\right)^{1/2}\right]\right\}\right)^{-1}, \tag{5.1}$$

où e_{ij}^{*} est indépendante de e_{ij} mais de même loi, $N\left(0, \sigma_e^2 = 2.0\right)$. Pour la sélection non invariante, la mesure de taille z_{ij} dépend des erreurs de niveau 1 ainsi que de niveau 2 et n'est donc pas invariante d'une grappe à l'autre. En particulier, dans (3.7), nous remplaçons e_{ij} et e_{ij}^{*} par $v_i + e_{ij}$ et $v_i^{*} + e_{ij}^{*}$, respectivement, où v_i^{*} est indépendant de v_i mais de même loi $N\left(0, \sigma_v^2 = 0{,}5\right)$. Nous considérons quatre valeurs de α dans (5.1) : $\alpha = 1, 2, 3, \infty$, où $\alpha = \infty$ correspond à l'échantillonnage non informatif dans chaque grappe, $\alpha = 1$ correspond à l'échantillonnage le plus informatif, et l'informativité diminue quand α augmente.

Nous avons utilisé l'approche fondée sur le plan de sondage et le modèle (*pm*) pour simuler $R = 1\,000$ échantillons pour chaque valeur spécifiée de α et séparément pour les sélections invariante et non invariante. Sous cette approche, nous avons généré une population correspondant à $N = 1000$ et $M_i = M = 100$ à partir du modèle, puis nous avons sélectionné un échantillon à deux degrés d'éléments comme il est spécifié plus haut. Le processus en deux étapes a été répété $R = 1000$ fois pour simuler 1000 échantillons.

5.1 Performance des estimateurs

Partant de chaque échantillon, nous avons calculé les estimations de μ, σ_v^2 et σ_e^2 en utilisant le maximum de vraisemblance restreint (REML), les méthodes d'ajustement des pondérations A et A1, la

méthode EEP proposée et la méthode de rechange de Korn et Graubard (appelée KG). Les biais et les variances des estimateurs ont été calculés en se servant des 1000 estimations. La performance des divers estimateurs est évaluée en utilisant deux mesures, à savoir le ratio du biais = RB = (biais)/ (racine carrée de la variance) et la racine carrée de l'erreur quadratique moyenne relative = REQMR = (racine carrée de l'EQM)/ (valeur réelle du paramètre). Les tableaux 5.1, 5.2 et 5.3 donnent, respectivement, les valeurs du RB des estimateurs de μ, σ_v^2 et σ_e^2. Les valeurs de la REQMR des estimateurs de μ, σ_v^2 et σ_e^2 sont présentées aux tableaux 5.4, 5.5 et 5.6, respectivement.

Tableau 5.1
Ratio du biais (%) des estimateurs de μ

α	Sélection invariante			Sélection non invariante		
	REML	**A**	**A1/EEP/KG**	**REML**	**A**	**A1/EEP/KG**
1	346,5	80,2	2,2	370,9	83,9	3,0
2	167,7	40,1	0,3	172,3	45,3	6,1
3	114,3	30,7	4,5	114,9	30,8	4,8
∞	2,0	2,5	2,1	-1,5	-2,4	-2,2

Le tableau 5.1 donne le ratio du biais (%) des estimateurs de μ fondés sur le REML, les méthodes d'ajustement des pondérations A et A1, et les méthodes KG et EEP. Notons que dans le cas de μ, les estimateurs A1, KG et EEP (VCP) sont identiques. Les résultats du tableau 5.1 montrent que le RB est le même pour les sélections invariante et non invariante, et que le RB des méthodes REML et A diminue lorsque α augmente. En outre, la méthode REML produit un biais plus important sous échantillonnage informatif, même pour $\alpha = 3$; par exemple, le RB pour la méthode REML varie de 114 % à 346 % sous sélection invariante. La méthode A conduit aussi à un RB important sous échantillonnage informatif; par exemple le RB pour la méthode A varie de 30,8 % à 83,9 % sous sélection non invariante. Par ailleurs, le RB des méthodes EEP, A1 et KG ne dépend pas de α et est faible ($|RB| < 6\%$). Sous échantillonnage non informatif, la méthode REML donne d'aussi bons résultats que prévus ($|RB| < 3\%$).

Pour l'estimation de σ_v^2, commençons par noter que la proportion de fois que l'estimation de σ_v^2 est négative est nulle dans les simulations pour les quatre valeurs de α et pour toutes les méthodes d'estimation (REML, A, A1, EEP et KG). Le tableau 5.2 donne les valeurs du RB pour les estimateurs de σ_v^2. Il montre que le RB de la méthode REML n'est pas affecté par α sous sélection invariante, mais qu'il l'est sous sélection non invariante. Dans ce dernier cas, le REML donne lieu à une sous-estimation importante pour $\alpha = 1$ (RB = -49 %), mais |RB| diminue à mesure que α augmente. Le tableau 5.2 montre aussi que les méthodes A et A1 ne donnent pas d'aussi bons résultats sous échantillonnage informatif (RB variant de 16 % à 60 %). La méthode KG n'a pas donné de bons résultats pour $\alpha = 1$ (RB = 33 % sous sélection invariante et RB = 24 % sous sélection non invariante). Par ailleurs, la méthode EEP donne de bons résultats pour toutes les valeurs de α (RB variant de -4 % à -13 %) mais la sous-estimation est systématique pour les diverses valeurs de α.

Le tableau 5.3 donne les valeurs du RB des estimateurs de σ_e^2. Il montre que ces valeurs sont les mêmes pour les sélections invariante et non invariante, comme dans le cas de μ. Les méthodes REML et KG donnent lieu à une sous-estimation importante quand $\alpha = 1$ (RB = -107 % pour la méthode REML et RB = -71 % pour la méthode KG sous sélection invariante), mais |RB| diminue quand α augmente et devient négligeable pour $\alpha = \infty$. La performance des estimateurs A et A1 est médiocre pour toutes les valeurs de α, y compris $\alpha = \infty$. Par ailleurs, la méthode EEP donne de bons résultats pour toutes les valeurs de α avec |RB| < 8 %. Il semble que l'instabilité introduite par le facteur d'échelle (2.9) pourrait avoir contribué à la grande valeur de |RB| pour les méthodes A et A1 même sous échantillonnage non informatif ($\alpha = \infty$).

Tableau 5.2
Ratio du biais (%) des estimateurs de σ_v^2

α	REML	A	A1	EEP	KG
			Sélection invariante		
1	0,6	59,5	59,3	-8,5	33,2
2	0,5	24,5	26,3	-10,0	8,0
3	-3,4	16,1	18,2	-13,6	0,4
∞	-0,1	14,8	17,1	-8,9	0,6
			Sélection non invariante		
1	-49,0	50,1	58,9	-4,4	24,0
2	-10,9	24,6	28,7	-7,0	7,1
3	-4,0	20,0	22,7	-7,8	4,6
∞	-1,3	12,8	13,9	-13,3	-1,6

Tableau 5.3
Ratio du biais (%) des estimateurs de σ_e^2

α	REML	A	A1	EEP	KG
			Sélection invariante		
1	-106,9	-118,4	-66,9	2,4	-71,2
2	-22,7	-43,6	-34,3	2,1	-16,5
3	-9,4	-31,7	-28,4	2,9	-6,5
∞	-0,4	-21,8	-23,8	0,3	0,4
			Sélection non invariante		
1	-115,3	-131,3	-79,6	-6,9	-82,6
2	-30,4	-51,1	-43,3	-7,6	-23,9
3	-12,5	-34,9	-32,2	-2,3	-10,3
∞	1,1	-20,2	-21,8	2,6	1,6

Tableau 5.4
Racine carrée de l'erreur quadratique moyenne relative (%) des estimateurs de μ

α	Sélection invariante			Sélection non invariante		
	REML	A	A1/EEP/KG	REML	A	A1/EEP/KG
1	93,3	35,9	29,4	92,5	35,4	29,2
2	51,6	29,3	27,8	52,8	30,4	28,9
3	40,5	28,2	27,5	40,8	28,7	28,1
∞	25,8	26,1	26,5	26,6	27,3	27,7

Racine carrée de l'erreur quadratique moyenne relative

Le tableau 5.4 montre que les valeurs de la REQMR (%) des estimations de μ sont les mêmes pour les sélections invariante et non invariante, et que la REQMR des méthodes REML et A diminue quand α augmente. Sous échantillonnage informatif avec $\alpha = 1$, la REQMR pour la méthode REML est grande comparativement à celle de la méthode EEP (A1 et KG) lorsque le RB est grand. Par exemple, la REQMR = 93 % pour la méthode REML comparativement à REQMR = 29 % pour la méthode EEP. Comme prévu, la méthode REML donne la plus petite REQMR sous échantillonnage non informatif, mais l'augmentation de la REQMR pour les autres méthodes est assez faible. En outre, la REQMR de la méthode EEP (A1 et KG) dépend de α.

Tableau 5.5
Racine carrée de l'erreur quadratique moyenne relative (%) des estimateurs de σ_v^2

α	REML	A	A1	EEP	KG
	Sélection invariante				
1	36,5	47,3	51,1	43,6	43,8
2	37,1	39,7	41,1	40,5	39,5
3	36,3	37,3	38,7	39,5	37,8
∞	35,8	36,9	38,1	38,7	37,2
	Sélection non invariante				
1	36,7	44,6	52,6	43,4	41,5
2	35,6	37,9	40,4	39,3	37,7
3	37,0	38,7	40,4	40,2	38,8
∞	36,6	37,2	38,0	39,0	37,8

Pour ce qui est de la REQMR des estimateurs de σ_v^2, le tableau 5.5 montre que la méthode REML donne de bons résultats pour toutes les valeurs de α sous sélection invariante parce que le RB est petit dans ce cas. Nous constatons aussi que les méthodes KG et EEP ont une REQMR comparable pour toutes les valeurs de α. Le tableau 5.5 révèle aussi que les méthodes A et A1 produisent une REQMR un peu plus grande pour $\alpha = 1$: 51 % pour A1 et 47 % pour A sous sélection invariante comparativement à 44 % pour la méthode EEP.

Tableau 5.6
Racine carrée de l'erreur quadratique moyenne relative (%) des estimateurs de σ_e^2

α	REML	A	A1	EEP	KG
			Sélection invariante		
1	13,5	14,5	12,8	13,9	12,9
2	9,7	10,4	10,4	11,0	10,0
3	9,5	10,0	10,1	10,7	9,8
∞	10,1	10,3	10,5	11,1	10,3
			Sélection non invariante		
1	13,7	14,8	12,9	13,2	13,0
2	10,0	10,9	10,9	11,3	10,3
3	9,7	10,4	10,7	11,2	10,2
∞	10,3	10,6	10,8	11,4	10,7

Le tableau 5.6 donne les valeurs de la REQMR pour les estimateurs de σ_e^2 et nous constatons que les valeurs sont similaires pour les sélections invariante et non invariante. Le tableau montre aussi que les valeurs de la REQMR sont comparables pour les méthodes EEP, A, Al et KG, même si, en ce qui concerne le ratio du biais, les méthodes A, Al et KG donnent de moins bons résultats que la méthode EEP. Cette situation est due au fait que la variance est plus grande pour EEP que pour les autres méthodes. Par exemple, dans le cas de la sélection invariante et $\alpha = 1$, nous obtenons les variances qui suivent pour les méthodes EEP, KG et REML: 0,0771, 0,0438 et 0,0339 avec des ratios du biais (%) correspondants provenant du tableau 5.3 : 2,4, -71,2 et -106,9.

5.2 Performance de l'estimateur de variance

Nous présentons maintenant certains résultats de simulation concernant le biais relatif de l'estimateur de variance par linéarisation (3.12) de l'estimateur EEP (VCP) $\hat{\boldsymbol{\theta}}_w$. Nous avons d'abord répété le processus en deux étapes $R_1 = 2000$ fois et calculé $v_L^{(r)}\left(\hat{\boldsymbol{\theta}}_w\right)$ en partant de chaque échantillon à deux

degrés $r = 1,..., 2000$. Les moyennes des éléments diagonaux de $E\left\{v_L\left(\hat{\boldsymbol{\theta}}_w\right)\right\} \approx v_L\left(\hat{\boldsymbol{\theta}}_w\right) = R_1^{-1}\sum_{r=1}^{R_1} v_L^{(r)}\left(\hat{\boldsymbol{\theta}}_w\right)$ sont désignées par $\overline{v}_L\left(\hat{\mu}_w\right)$, $\overline{v}_L\left(\hat{\sigma}_{vw}^2\right)$ et $\overline{v}_L\left(\hat{\sigma}_{ew}^2\right)$, respectivement. Nous avons ensuite généré $R_2 = 10000$ échantillons indépendants et calculé l'erreur quadratique moyenne (EQM) empirique des trois estimateurs $\hat{\mu}_w$, $\hat{\sigma}_{vw}^2$ et $\hat{\sigma}_{ew}^2$. Nous avons $\mathrm{EQM}\left(\hat{\mu}_w\right) \approx R_2^{-1}\sum_{r=1}^{R_2}\left(\hat{\mu}_w^{(r)} - \mu\right)^2$, où $\hat{\mu}_w^{(r)}$ est l'estimation de μ d'après le r^{e} échantillon simulé, et les expressions similaires pour $\mathrm{EQM}\left(\hat{\sigma}_{vw}^2\right)$ et $\mathrm{EQM}\left(\hat{\sigma}_{ew}^2\right)$.

Le biais relatif de $v_L\left(\hat{\mu}_w\right)$ est calculé comme

$$\mathrm{BR}\left\{v_L\left(\hat{\mu}_w\right)\right\} = \left[\overline{v}_L\left(\hat{\mu}_w\right)/\mathrm{EQM}\left(\hat{\mu}_w\right)\right] - 1$$

et $\mathrm{BR}\left\{v_L\left(\hat{\sigma}_{vw}^2\right)\right\}$ et $\mathrm{BR}\left\{v_L\left(\hat{\sigma}_{ew}^2\right)\right\}$ ont été calculés de la même façon. Le tableau 5.7 donne les valeurs du BR pour les trois estimateurs de variance sous sélections invariante et non invariante et $\alpha = 1, 2, 3, \infty$. L'examen du tableau 5.7 montre clairement que l'estimateur de variance par linéarisation donne de bons résultats pour toutes les combinaisons, avec $|\mathrm{BR}| < 10\%$.

Tableau 5.7
Biais relatif (%) des estimateurs de variance

α	$v_L(\hat{\mu}_w)$	$v_L(\hat{\sigma}_{vw}^2)$	$v_L(\hat{\sigma}_{ew}^2)$
	Sélection invariante		
1	-3,0	-6,2	-7,5
2	-5,2	-4,5	-3,1
3	-1,3	-3,8	-1,8
∞	-0,9	-2,5	-2,0
	Sélection non invariante		
1	-3,8	-8,3	-4,2
2	-4,5	-5,8	-7,3
3	-4,3	-4,6	-5,7
∞	-2,4	-2,7	-2,9

6 Conclusion

Dans le présent article, nous avons proposé une approche unifiée fondée sur la vraisemblance composite pondérée (VCP) pour les modèles à deux niveaux pour faire des inférences sur des données d'enquête complexes. Les méthodes VCP proposées sont asymptotiquement valides même quand les tailles d'échantillon dans les grappes échantillonnées (unités de niveau 1) sont petites, contrairement à

certaines méthodes existantes, mais il est nécessaire de connaître les probabilités d'inclusion conjointe dans les grappes échantillonnées. Souvent, il est possible de traiter l'échantillonnage dans les grappes comme étant effectué avec remise, en raison des petites fractions d'échantillonnage dans les grappes. En outre, d'excellentes approximations des probabilités d'inclusion conjointe, qui ne dépendent que des probabilités d'inclusion marginales, sont disponibles lorsque les fractions d'échantillonnage ne sont pas petites (Haziza et coll. 2008). Nous prévoyons examiner l'exactitude de ce genre d'approximations dans le cadre d'une future étude. Des études en simulation de la performance des estimateurs VCP (4.5) et (4.6) pour les modèles à deux niveaux (2.3) fondées sur la méthode par paire seront également réalisées.

Les méthodes fondées sur la vraisemblance composite sont utilisées principalement lorsque la vraisemblance complète est complexe. Notre développement dans le contexte des sondages donne la preuve que la méthode fondée sur la vraisemblance complète avec pondérations n'est pas faisable pour des modèles multiniveaux, tandis que la méthode fondée sur la vraisemblance composite pondérée facilite l'obtention d'inférences valides, même si les tailles d'échantillon de grappe sont petites.

7 Remerciements

Nous remercions deux examinateurs et le rédacteur associé de leurs suggestions et commentaires constructifs.

Annexe

Équations de score pondérées : modèle de régression linéaire à erreurs emboîtées

Pour le modèle de régression linéaire à erreurs emboîtées (2.3), une forme explicite de la log-vraisemblance complète de recensement s'obtient en utilisant la forme explicite de la matrice de covariance $\mathbf{V}_i$ de $\mathbf{y}_i = \left(y_{i1},...,y_{iM_i}\right)^T$. Nous avons $\mathbf{V}_i^{-1} = \sigma_e^{-2}\left[\mathbf{I}_i - \sigma_v^2 / \lambda_i \mathbf{1}_i \mathbf{1}_i^T\right]$, où $\lambda_i = \sigma_e^2 + M_i\sigma_v^2$, $\mathbf{I}_i$ est la matrice identité de dimensions $M_i \times M_i$ et $\mathbf{1}_i$ est le vecteur unité de dimension $M_i \times 1$. En utilisant l'expression pour $\mathbf{V}_i^{-1}$, les équations de score de recensement s'obtiennent sous la forme

$$\boldsymbol{\beta}: \quad \left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\mathbf{x}_{ij}y_{ij} - \sigma_v^2\sum_{i=1}^{N}\lambda_i^{-1}\left(\sum_{j=1}^{M_i}\sum_{k=1}^{M_i}\mathbf{x}_{ij}y_{ik}\right)\right] - \left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T - \sigma_v^2\sum_{i=1}^{N}\lambda_i^{-1}\left(\sum_{j=1}^{M_i}\sum_{k=1}^{M_i}\mathbf{x}_{ij}\mathbf{x}_{ik}^T\right)\right]\boldsymbol{\beta} = 0 \quad \text{(A.1)}$$

$$\sigma_v^2: \quad \sum_{i=1}^{N}\lambda_i^{-2}\left[\sum_{j=1}^{M_i}\sum_{k=1}^{M_i}\left(y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik} - \mathbf{x}_{ik}^T\boldsymbol{\beta}\right)\right] - \sum_{i=1}^{N}\lambda_i^{-1}M_i = 0 \quad \text{(A.2)}$$

$$\sigma_e^2: \quad \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2+\sum_{i=1}^{N}\left(M_i\sigma_v^4\lambda_i^{-2}-2\sigma_v^2\lambda_i^{-1}\right)\sum_{j,k=1}^{M_i}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik}-\mathbf{x}_{ik}^T\boldsymbol{\beta}\right) - \sigma_e^2\sum_{i=1}^{N}\left(1-\sigma_v^2\lambda_i^{-1}\right)M_i=0 \tag{A.3}$$

Partant de (A.1), nous obtenons les équations de score pondérées

$$\boldsymbol{\beta}: \quad \sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}\mathbf{x}_{ij}y_{ij}-\sigma_v^2\sum_{i\in s}w_i\lambda_i^{-1}\left(\sum_{j\in s(i)}\sum_{k\in s(i)}w_{jk|i}\mathbf{x}_{ij}y_{ik}\right) -\left[\sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T-\sigma_v^2\sum_{i\in s}w_i\lambda_i^{-1}\left(\sum_{j\in s(i)}\sum_{k\in s(i)}w_{jk|i}\mathbf{x}_{ij}\mathbf{x}_{ik}\right)\right]\boldsymbol{\beta}=0 \tag{A.4}$$

où $w_{jj|i}=w_{j|i}$. Notons que les tailles de grappe M_i pour $i\in s$ sont supposées connues. On ne doit pas remplacer M_i par son estimation $\sum_{j\in s(i)}w_{j|i}$, parce que celle-ci comprend un biais de ratio dû aux petites tailles d'échantillon dans les grappes. L'équation d'estimation (A.4) est sans biais sous le plan pour l'équation de recensement (A.1).

Passons maintenant à l'équation de score pondérée pour σ_v^2, nous obtenons en partant de (A.2)

$$\sigma_v^2: \quad \sum_{i\in s}w_i\lambda_i^{-2}\left[\sum_{j\in s(i)}\sum_{k\in s(i)}w_{jk|i}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik}-\mathbf{x}_{ik}^T\boldsymbol{\beta}\right)\right]-\sum_{i\in s}w_i\lambda_i^{-1}\sum_{j\in s(i)}w_{j|i}=0 \tag{A.5}$$

L'équation d'estimation (A.5) est sans biais pour (A.2). Enfin, l'équation de score pondérée pour σ_e^2 s'obtient à partir de (A.3) sous la forme

$$\sigma_e^2: \quad \sum_{i\in s}w_i\sum_{j\in s(i)}w_{j|i}\left(y_{ij}-x_{ij}^T\beta\right)^2+\sum_{i\in s}w_i\left(M_i\sigma_v^4\lambda_i^{-2}-2\sigma_v^2\lambda_i^{-1}\right)\sum_{j,k\in s(i)}w_{jk|i}\left(y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\left(y_{ik}-\mathbf{x}_{ik}^T\boldsymbol{\beta}\right) -\sigma_e^2\sum_{i\in s}w_i\left(1-\sigma_v^2\lambda_i^{-1}\right)\sum_{j\in s(i)}w_{j|i}=0 \tag{A.6}$$

Il découle des équations (A.4) à (A.6) que les équations de score pondérées dépendent uniquement des pondérations d'ordre 1 w_i et $w_{j|i}$ et des pondérations d'ordre 2 $w_{jk|i}$ dans le cas particulier d'un modèle de régression linéaire à erreurs emboîtées.

Bibliographie

Asparouhov, T. (2006). Generalized multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods*, 35, 439-460.

Beaumont, J.-F., et Charest, A.-S. (2010). Bootstrap variance estimation with survey data when estimating model parameters. Document non publié (fourni par les auteurs).

Bellio, R., et Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 3, 217-227.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.

Grilli, L., et Pratesi, M. (2004). Estimation pondérée dans le cadre de modèles multiniveaux ordinaux et binaires sous un plan d'échantillonnage informatif. *Techniques d'enquête*, 30, 103-114.

Haziza, D., Mecatti, F. et Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao Sampford unequal probability sampling design. *Metron*, 66, 91-108.

Korn, E.L., et Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society B*, 65, 175-190.

Kovacevic, M.S., Rong, H. et You, Y. (2006). Bootstrapping for variance estimation in multi-level models fitted to survey data. *Proceedings of ASA Section on Survey Research Methods*, American Statistical Association, 3260-3269.

Lele, S., et Taper, M.L. (2002). A composite likelihood approach to (co)variance components estimation. *Journal of Statistical Planning and Inference*, 103, 117-125.

Lindsay, B.G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes*, (Ed. N.U. Prabhu), Providence: American Mathematical Society, 221-239.

Lindsay, B.G., Yi, G.Y. et Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21, 71-105.

Muthén, L.K., et Muthén, B.O. (1998-2005). *Mplus User's Guide. 3rd ed. Los Angeles, CA: Muthén & Muthén*.

Pfeffermann, D., et Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. Dans *Analysis of Survey Data*, (Éds. R. Chambers et C.J. Skinner), Wiley, Chichester, 175-196.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society B*, 60, 23-56.

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

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

Rao, J.N.K., et Roberts, G. (1998). Discussion on the papers by Firth and Bennett and Pfeffermann *et al*. *Journal of the Royal Statistical Society B*, 60, 50-51.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.

Rao, J.N.K., Hidiroglou, M., Yung, W. et Kovacevic, M. (2010). Role of weights in descriptive and analytical inferences from survey data: An overview. *Journal of the Indian Society of Agricultural Statistics*, 64, 129-135.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

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

Yi, G.Y., Rao, J.N.K. et Li, H. (2012). A weighted composite likelihood approach for analysis of survey data under two level models. Disponible sur demande auprès de jrao@math.carleton.ca.


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

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 313-331
Statistique Canada, N° 12-001-X au catalogue

Comparaison de différents plans de sondage et construction de bandes de confiance pour l'estimation de la moyenne de données fonctionnelles : une illustration sur la consommation électrique

Hervé Cardot, Alain Dessertaine, Camelia Goga, Étienne Josserand et Pauline Lardin[1]

Résumé

Lorsque les variables étudiées sont fonctionnelles et que les capacités de stockage sont limitées ou que les coûts de transmission sont élevés, les sondages, qui permettent de sélectionner une partie des observations de la population, sont des alternatives intéressantes aux techniques de compression du signal. Notre étude est motivée, dans ce contexte fonctionnel, par l'estimation de la courbe de charge électrique moyenne sur une période d'une semaine. Nous comparons différentes stratégies d'estimation permettant de prendre en compte une information auxiliaire telle que la consommation moyenne de la période précédente. Une première stratégie consiste à utiliser un plan de sondage aléatoire simple sans remise, puis de prendre en compte l'information auxiliaire dans l'estimateur en introduisant un modèle linéaire fonctionnel. La seconde approche consiste à incorporer l'information auxiliaire dans les plans de sondage en considérant des plans à probabilités inégales tels que les plans stratifiés et les plans πps. Nous considérons ensuite la question de la construction de bandes de confiance pour ces estimateurs de la moyenne. Lorsqu'on dispose d'estimateurs performants de leur fonction de covariance et si l'estimateur de la moyenne satisfait un théorème de la limite centrale fonctionnel, il est possible d'utiliser une technique rapide de construction de bandes de confiance qui repose sur la simulation de processus Gaussiens. Cette approche est comparée avec des techniques de bootstrap qui ont été adaptées afin de tenir compte du caractère fonctionnel des données.

Mots-clés : Bonferroni; Bootstrap; estimateur de Horvitz-Thompson; fonction de covariance; estimateur model-assisted; modèle linéaire fonctionnel; formule de Hájek.

1 Introduction

Avec le développement de procédés automatiques d'acquisition de données à des échelles de temps fines, il n'est maintenant plus inhabituel de disposer de très grandes bases de données concernant des phénomènes qui évoluent au cours du temps. Par exemple, en France, dans les années à venir, environ 30 millions de compteurs électriques vont être remplacés par des compteurs communicants. Ceux-ci seront capables de mesurer la consommation de chaque ménage et de chaque entreprise à des pas de temps potentiellement très fins (seconde ou minute) et d'envoyer ces mesures une fois par jour à un serveur central. Un autre exemple concerne les mesures d'audience des différentes chaînes de télévision. Des boitiers, reliés à internet, permettent de mesurer en temps continu si la télévision est allumée et quelle chaîne est regardée.

L'unité statistique étudiée est alors une fonction (du temps, de l'espace), ce qui nécessite d'introduire des outils d'analyse fonctionnelle. Bien que présent dès les années 1970s (Deville 1974), Dauxois et Pousse (1976), ce domaine de la statistique s'est réellement développé au cours des années 1990, avec les

1. Hervé Cardot, Université de Bourgogne, Institut de Mathématiques de Bourgogne, 9 av. Alain Savary, 21078 DIJON, FRANCE; Alain Dessertaine, LA POSTE - DIRECTION DU COURRIER - DFI – DCPES, 2 Boulevard Newton 77543 MARNE LA VALLEE CEDEX 2 and EDF, R&D, ICAME-SOAD, 1 av. du Général de Gaulle, 92141 CLAMART, France; Camelia Goga, Université de Bourgogne, Institut de Mathématiques de Bourgogne. Courriel : camelia.goga@u-bourgogne.fr; Étienne Josserand, Université de Bourgogne, Institut de Mathématiques de Bourgogne; Pauline Lardin, Université de Bourgogne, Institut de Mathématiques de Bourgogne and EDF, R&D, ICAME-SOAD.

progrès de l'informatique. Les applications concernent des domaines divers tels que la climatologie, l'économie, la télédétection, la médecine ou encore la chimie quantitative. Le lecteur pourra se reporter aux références récentes Ramsay et Silverman (2005) et Ferraty et Romain (2011) pour un panorama des différentes techniques et des exemples d'applications.

Lorsque les bases de données potentielles sont très grandes, il peut être difficile et coûteux de collecter, de sauvegarder et d'analyser l'ensemble des données. Si de plus on s'intéresse à des indicateurs simples tels que la courbe moyenne sous des contraintes d'espace mémoire ou de coût de transmission, l'emploi de techniques de sondage afin d'extraire un échantillon peut fournir une estimation précise à un coût raisonnable (Dessertaine 2008).

Les travaux combinant analyse des données fonctionnelles et théorie des sondages sont encore peu nombreux dans la littérature statistique. Cardot, Chaouch, Goga et Labruère (2010) s'intéressent à l'analyse en composantes principales en vue de réduire la dimension des données tandis que Cardot et Josserand (2011) étudient des propriétés de convergence uniforme d'estimateurs de Horvitz-Thompson de courbes moyennes. On peut également citer Chaouch et Goga (2012) qui proposent un estimateur robuste de courbes centrales.

L'objectif de ce travail est de comparer, sur un exemple réel, différentes stratégies d'échantillonnage dans un contexte fonctionnel. Ces données réelles portent sur les consommations électriques, relevées toutes les demi-heures pendant deux semaines, d'une population test de $N = 15\,069$ compteurs électriques. Le profil temporel de consommation électrique des particuliers dépend de covariables telles que les caractéristiques météorologiques (température, *etc.*) ou géographiques (altitude, latitude ou longitude). Ces variables ne sont malheureusement pas disponibles pour cette étude et nous n'utilisons qu'une seule variable comme information auxiliaire : la consommation moyenne de chaque compteur lors de la semaine précédente. Cette information peut être facilement transmise par tous les compteurs de la population.

L'extension au cadre fonctionnel des méthodes d'estimation qui prennent en compte de l'information auxiliaire n'est pas toujours directe. Cardot et Josserand (2011) proposent de stratifier la population des courbes pour améliorer l'estimation de la courbe moyenne. Chaouch et Goga (2012), qui s'intéressent à la courbe médiane, suggèrent d'utiliser un plan proportionnel à la taille avec remise ainsi que l'estimateur poststratifié. Nous proposons dans cet article de comparer plusieurs stratégies qui permettent de prendre en compte l'information auxiliaire. Une première stratégie fait intervenir l'information auxiliaire au niveau de la sélection de l'échantillon : tirage avec un plan à probabilités inégales (stratifié, πps) et estimation avec l'estimateur de Horvitz-Thompson. La deuxième stratégie fait intervenir cette information au niveau de l'estimation : tirage avec un échantillonnage aléatoire simple sans remise et estimation en utilisant un modèle de régression linéaire (Särndal, Swensson et Wretman (1992) adapté au cadre fonctionnel (Faraway 1997).

Une nouvelle question liée au caractère fonctionnel des données apparaît alors de manière naturelle : comment quantifier l'incertitude liée à l'échantillonnage ? La question, centrale pour les sondeurs, de la construction d'intervalles de confiance, n'a été que peu abordée en statistique des données fonctionnelles où il faut alors construire des bandes de confiance. En nous inspirant de techniques basées sur l'estimation de la fonction de covariance de l'estimateur (voir Faraway (1997), Cuevas, Febrero et Fraiman (2006) ou plus récemment Degras (2011)), nous proposons tout d'abord de construire des bandes de confiance par simulation de processus gaussiens. Une justification asymptotique de la validité de ces techniques est

donnée dans Cardot, Degras et Josserand (2013) lorsque les hypothèses du théorème central limite sont vérifiées et que l'on dispose d'un estimateur précis de la fonction de covariance. Une deuxième méthode de construction, qui repose sur les techniques de bootstrap, est également mise en œuvre. Il existe essentiellement trois techniques de bootstrap en population finie : le bootstrap sans remise proposé par Gross (1980), le « rescaling » bootstrap (Rao et Wu 1988) et le « mirror-match » bootstrap (Sitter 1992). Dans ce travail, nous utilisons le bootstrap sans remise qui repose sur les adaptations pour les plans stratifiés et proportionnels à la taille proposées par Chauvet (2007).

Nous introduisons dans la seconde section les notations, les estimateurs de la courbe moyenne en présence d'information auxilaire ainsi que les estimateurs de leur fonction de covariance. Les algorithmes de construction des bandes de confiance, de type bootstrap ou par simulation de processus gaussiens, sont décrits dans la section 3. La section 4 propose ensuite une comparaison des différentes stratégies, en termes de précision des estimateurs, de largeur et de couverture des bandes de confiance et de temps de calcul, de l'estimation des courbes de charge de l'opérateur français EDF (Electricité de France). Nous considérons pour cela des échantillons de taille $n = 1\,500$ dans notre population test constituée de $N = 15\,069$ courbes. Pour finir, nous présentons quelques perspectives de recherche dans la section 5.

2 Données fonctionnelles en population finie

Considérons une population finie $U = \{1, ..., N\}$ de taille N et supposons que, pour chaque élément k de la population U, nous pouvons observer la courbe déterministe $Y_k = \left(Y_k(t)\right)_{t\in[0,T]}$. L'objectif est d'estimer la courbe moyenne de la population qui est définie pour tout instant $t \in [0, T]$, par

$$\mu(t) = \frac{1}{N}\sum_{k\in U} Y_k(t).$$

Soit s un échantillon de taille fixée n, choisi aléatoirement dans U, selon un plan de sondage $p(.)$. Soient $\pi_k = \Pr(k \in s)$ et $\pi_{kl} = \Pr(k \,\&\, l \in s)$ les probabilités d'inclusion d'ordre un et deux respectivement. On suppose que $\pi_k > 0$ pour tout élément k de la population U.

La courbe moyenne μ est estimée à l'aide de l'estimateur de Horvitz-Thompson (Cardot et coll. 2010) comme suit

$$\hat{\mu}(t) = \frac{1}{N}\sum_{k\in s}\frac{Y_k(t)}{\pi_k} = \frac{1}{N}\sum_{k\in U}\frac{Y_k(t)}{\pi_k}1_{k\in s}, \quad t \in [0, T], \tag{2.1}$$

où $1_{k\in s}$ est l'indicatrice d'appartenance de l'unité k à l'échantillon s. Pour chaque instant $t \in [0, T]$, l'estimateur $\hat{\mu}(t)$ est sans biais pour $\mu(t)$, c'est à dire $E(\hat{\mu}(t)) = \mu(t)$ où l'espérance est considérée par rapport au plan de sondage.

Généralement les trajectoires $Y_k(t)$ ne sont pas observées continûment pour $t \in [0, T]$ mais uniquement sur un ensemble de D instants de mesure $0 = t_1 < t_2 < \ldots < t_D = T$. Une stratégie classique en analyse des données fonctionnelles consiste à effectuer une interpolation ou un lissage des

trajectoires discrétisées afin d'obtenir des objets qui sont réellement des fonctions (Ramsay et Silverman 2005). Cela permet également de traiter des courbes dont les instants de mesure ne sont pas identiques. Dans le cadre des sondages, l'interpolation linéaire, lorsqu'il n'y a pas d'erreur de mesure aux points discrétisés, a été étudiée par Cardot et Josserand (2011) tandis que des procédures de lissage sont proposées dans Cardot et coll. (2013). Si le nombre de points de discrétisation est suffisant et les trajectoires sont assez régulières (mais pas nécessairement dérivables), l'erreur d'approximation due au lissage ou à l'interpolation est négligeable face à l'erreur d'échantillonnage. On suppose dans la suite que les trajectoires sont observées en tout point t de l'intervalle $[0,T]$.

La fonction de covariance de type Horvitz-Thompson $\gamma(r,t) = \text{cov}\left(\hat{\mu}(r), \hat{\mu}(t)\right)$ est donnée par

$$\gamma(r,t) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{Y_k(r)}{\pi_k}\frac{Y_l(t)}{\pi_l}$$

pour tout $(r,t) \in [0,T]\times[0,T]$ et $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$. Si on suppose que les probabilités d'inclusion d'ordre deux satisfont $\pi_{kl} > 0$, un estimateur sans biais de $\gamma(r,t)$ est donné par l'estimateur sans biais de la variance de type Horvitz-Thompson,

$$\hat{\gamma}(r,t) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{Y_k(r)}{\pi_k}\frac{Y_l(t)}{\pi_l} \tag{2.2}$$

pour tout $(r,t) \in [0,T]\times[0,T]$.

2.1 Prise en compte d'information auxiliaire pour l'estimation de la trajectoire moyenne

Il est bien connu que l'utilisation d'une information auxiliaire qui explique bien la variable d'intérêt peut beaucoup améliorer la précision de l'estimateur de Horvitz-Thompson. Dans le cas des données EDF, la température extérieure ou le type de contrat pourraient sans doute être des variables auxiliaires intéressantes. Une stratification selon la position géographique permettrait également d'obtenir des estimations pour les différentes régions. Dans cette étude, nous disposons comme variable auxiliaire de la consommation électrique totale de la semaine précédente. Nous supposons que cette variable (réelle) est observée pour tous les éléments de la population.

Nous présentons dans cette section l'estimateur de Horvitz-Thompson pour la courbe moyenne ainsi qu'une estimation de la fonction de covariance de cet estimateur pour le sondage stratifié avec échantillonnage aléatoire simple sans remise (ÉASSR) dans chaque strate, noté dans la suite STRAT, et pour l'échantillonnage proportionnel à la taille sans remise que l'on note πps. Nous considérons également un estimateur de la courbe moyenne assisté par un modèle linéaire fonctionnel.

2.1.1 Le sondage stratifié avec ÉASSR dans chaque strate (STRAT)

La population U est supposée être stratifiée en un nombre fixé H de strates $U_1,\ldots,U_H$ de tailles $N_1,\ldots,N_H$. À l'intérieur de chaque strate U_h, on tire un échantillon s_h de taille n_h selon un plan ÉASSR.

Notons $\mu_h(t) = \sum_{k \in U_h} Y_k(t) / N_h$, pour $t \in [0,T]$, la courbe moyenne dans chaque strate et $\hat{\mu}_h(t) = \sum_{k \in s_h} Y_k(t) / n_h$, son estimation. L'estimateur de la courbe moyenne μ est alors défini par

$$\hat{\mu}_{\text{strat}}(t) = \frac{1}{N}\sum_{h=1}^{H} N_h \hat{\mu}_h(t) = \sum_{h=1}^{H} \frac{N_h}{N}\left(\frac{1}{n_h}\sum_{k \in s_h} Y_k(t)\right), \; t \in [0,T]. \tag{2.3}$$

L'estimateur de Horvitz-Thompson de la fonction de covariance γ est alors

$$\hat{\gamma}_{strat}(r,t) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{Y(r)Y(t),s_h} \quad r,t \in [0,T], \tag{2.4}$$

où

$$S_{Y(r)Y(t),s_h} = \frac{1}{n_h - 1}\sum_{k \in s_h}\left(Y_k(r) - \hat{\mu}_h(r)\right)\left(Y_k(t) - \hat{\mu}_h(t)\right)$$

est l'estimateur de la fonction de covariance $S_{Y(r)Y(t),U_h}$ dans la strate h. Pour $r = t \in [0,T]$, on obtient l'estimateur de la fonction de variance comme suit

$$\hat{\gamma}_{strat}(r) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S^2_{Y(r),s_h},$$

où

$$S^2_{Y(r),s_h} = \frac{1}{n_h - 1}\sum_{k \in s_h}\left(Y_k(r) - \hat{\mu}_h(r)\right)^2$$

est l'estimateur de la variance $S^2_{Y(r),U_h}$ dans la strate h. Cardot et Josserand (2011) proposent une extension, au cadre fonctionnel, de l'allocation optimale de Neyman. Les tailles n_h des échantillons s_h vérifiant

$$n_h = n\frac{N_h\sqrt{\int_0^T S^2_{Y(r),U_h}\,dr}}{\sum_{h=1}^{H} N_h\sqrt{\int_0^T S^2_{Y(r),U_h}\,dr}}, \quad h = 1,\ldots,H, \tag{2.5}$$

permettent de rendre minimale la variance intégrée, $\int_0^T \hat{\gamma}_{\text{strat}}(t)\,dt$, de l'estimateur stratifié. Cette allocation est similaire à l'allocation obtenue dans le cadre multivarié par Cochran (1977). En remplaçant la variable Y par une autre variable X connue sur toute la population et très corrélée avec la variable d'intérêt, on obtient une allocation dite x- optimale.

Remarque 2.1 *Pour $H = 1$, nous obtenons le plan aléatoire simple sans remise (ÉASSR) et la courbe moyenne $\mu(t)$ est estimée par*

$$\hat{\mu}_{éassr}(t) = \frac{1}{n}\sum_{k \in s} Y_k(t), \; t \in [0,T]. \tag{2.6}$$

L'estimateur de la fonction de covariance défini en (2.2) est alors

$$\hat{\gamma}_{éassr}(r,t) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y(r)Y(t),s}. \tag{2.7}$$

2.1.2 L'échantillonnage proportionnel à la taille sans remise (πps)

Les plans d'échantillonnage proportionnels à la taille avec ou sans remise sont souvent utilisés en pratique car leur efficacité est supérieure à celle de plans à probabilités égales lorsque la variable d'intérêt est plus ou moins proportionnelle à une variable auxiliaire X qui a des valeurs strictement positives.

Dans le cas des échantillons de taille fixe n tirés sans remise, il est possible de donner l'équivalent de la formule de Yates et Grundy (1953) et Sen (1953). La fonction de covariance de $\hat{\mu}$ vérifie,

$$\gamma(r,t) = -\frac{1}{2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U, l\neq k}(\pi_{kl} - \pi_k\pi_l)\left(\frac{Y_k(r)}{\pi_k} - \frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k} - \frac{Y_l(t)}{\pi_l}\right),\ r,t \in [0,T]. \tag{2.8}$$

Supposons que les valeurs x_k de la variable X sont connues pour toutes les unités k de la population. Il est alors possible de définir les probabilités d'inclusion :

$$\pi_k = n\frac{x_k}{\sum_{k\in U} x_k}.$$

Des méthodes ont été proposées dans la littérature pour le cas $\pi_k > 1$ (Särndal et coll. 1992).

Les probabilités d'inclusion d'ordre deux sont en général très difficiles à calculer pour les plans πps et par conséquent, la formule (2.2) ne peut pas être utilisée. Il existe cependant une approximation asymptotique simple de la variance qui a été proposée par Hájek (1964) et qui ne fait intervenir que les probabilités d'inclusion d'ordre un. Cette approximation se révèle très performante lorsque la taille de l'échantillon est grande et l'entropie du plan de sondage proche de l'entropie maximale. Pour sélectionner l'échantillon s avec des probabilités d'inclusion π_k, l'algorithme du cube (Deville et Tillé 2004) équilibré sur la variable $\boldsymbol{\pi} = (\pi_k)_{k\in U}$ peut être utilisé. Deville et Tillé (2005) montrent que pour ce plan de sondage particulier la formule de Hàjek est très performante pour estimer la variance d'un total ou d'une moyenne. Cette formule d'approximation de la variance peut aussi être utilisée pour la covariance, qui est alors estimée par

$$\hat{\gamma}_{\pi\text{ps}}(r,t) = \frac{1}{N^2}\sum_{k\in s}(1-\pi_k)\left(\frac{Y_k(r)}{\pi_k} - \hat{R}(r)\right)\left(\frac{Y_k(t)}{\pi_k} - \hat{R}(t)\right),\ \ r,t \in [0,T], \tag{2.9}$$

où

$$\hat{R}(t) = \frac{\sum_{k\in s}\frac{Y_k(t)}{\pi_k}(1-\pi_k)}{\sum_{k\in s}(1-\pi_k)}.$$

Nous avons également utilisé le sondage systématique à probabilités inégales proposé par Madow (1949) en raison de sa simplicité d'utilisation. Il est malheureusement difficile d'estimer la variance pour ce type de plan et nous ne l'utiliserons donc pas pour construire les bandes de confiance.

2.2 L'estimateur assisté par un modèle (« model-assisted »)

Considérons p variables auxiliaires réelles $X_1, \ldots, X_p$ et soit x_{kj} la valeur de la variable X_j pour le $k^{\text{ème}}$ individu. Notons par $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})'$ le vecteur contenant les valeurs de p variables auxiliaires mesurées sur le $k^{\text{ème}}$ individu. On considère que la relation entre la variable d'intérêt et les variables auxiliaires est modélisée par le modèle de superpopulation suivant

$$\xi : Y_k(t) = \mathbf{x}_k'\boldsymbol{\beta}(t) + \varepsilon_{kt}, \quad t \in [0, T] \tag{2.10}$$

avec

$$E_\xi(\varepsilon_{kt}) = 0, E_\xi(\varepsilon_{kt}\varepsilon_{lt'}) = 0 \text{ pour } k \neq l \text{ et } E_\xi(\varepsilon_{kt}\varepsilon_{kt'}) = \sigma^2_{tt'} \text{ pour } k = l.$$

Ce modèle est une généralisation immédiate à plusieurs variables auxiliaires du modèle linéaire fonctionnel proposé par Faraway (1997).

L'estimation de $\boldsymbol{\beta}$ basée sur le modèle ξ et le plan de sondage $p(\cdot)$ est donnée par

$$\hat{\boldsymbol{\beta}}(t) = \left(\sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k}\right)^{-1}\sum_{k\in s}\frac{\mathbf{x}_k Y_k(t)}{\pi_k}, \quad t \in [0, T]. \tag{2.11}$$

Remarquons que les poids de sondage ne dépendent pas du temps $t \in [0, T]$. Soit $\hat{Y}_k(t) = \mathbf{x}_k'\hat{\boldsymbol{\beta}}(t)$ l'estimateur basé sur le plan de sondage de la prédiction sous le modèle ξ de $Y_k(t)$. Par analogie directe avec le cas univarié (Särndal et coll. 1992), nous obtenons finalement l'estimateur suivant pour la moyenne, pour $t \in [0, T]$,

$$\begin{aligned}\hat{\mu}_{MA}(t) &= \frac{1}{N}\sum_{k\in s}\hat{Y}_k(t) - \frac{1}{N}\sum_{k\in s}\frac{\left(\hat{Y}_k(t) - Y_k(t)\right)}{\pi_k} \\ &= \frac{1}{N}\sum_{k\in U}\frac{Y_k(t) - x_k'\hat{\boldsymbol{\beta}}(t)}{\pi_k} + \frac{1}{N}\left(\sum_{k\in U}\mathbf{x}_k'\right)\hat{\boldsymbol{\beta}}(t).\end{aligned} \tag{2.12}$$

Si le modèle ξ contient la variable constante 1, alors l'estimateur devient

$$\hat{\mu}_{MA}(t) = \frac{1}{N}\sum_{k\in U}\hat{Y}_k(t), \quad t \in [0, T]. \tag{2.13}$$

Pour r et t fixés, la covariance asymptotique de $\hat{\mu}_{MA}(r)$ et $\hat{\mu}_{MA}(t)$ peut être calculée selon la technique classique des résidus (Särndal et coll. 1992),

$$\gamma_{MA}(r,t) \simeq \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{\left(Y_k(r) - \tilde{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t) - \tilde{Y}_l(t)\right)}{\pi_l}, \tag{2.14}$$

où $\tilde{Y}_k(r) = \mathbf{x}_k'\tilde{\boldsymbol{\beta}}(t)$ est la prédiction de $Y_k(t)$ sous le modèle de superpopulation et $\tilde{\boldsymbol{\beta}}(t) = \left(\sum_U \mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum_U \mathbf{x}_k Y_k(t)\right)$ est l'estimation de $\boldsymbol{\beta}$ au niveau de la population et $r,t \in [0,T]$. Cardot, Goga et Lardin (2013) montrent que ce résultat reste valable uniformément en $r,t \in [0,T]$.

Nous proposons comme estimateur de la fonction de covariance $\gamma_{MA}(r,t)$ l'estimateur de Horvitz-Thompson de la covariance asymptotique donnée par (2.14) où $\tilde{\boldsymbol{\beta}}(t)$ est remplacé par son estimateur $\hat{\boldsymbol{\beta}}(t)$ basé sur le plan de sondage,

$$\hat{\gamma}_{MA}(r,t) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{\left(Y_k(r)-\hat{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t)-\hat{Y}_l(t)\right)}{\pi_l}, \quad r,t \in [0,T]. \tag{2.15}$$

Remarque 2.2 *Il est tout à fait possible de considérer un modèle de superpopulation ξ plus général que le modèle linéaire proposé ici. Des techniques d'estimation basées sur un lissage par des B-splines (Goga et Ruiz-Gazen 2012) peuvent alors être envisagées. Dans notre étude, la relation entre la consommation à l'instant t et la consommation moyenne de la semaine précédente est quasi linéaire (voir figure 4.1) ce qui justifie de ne pas employer ces approches nonparamétriques.*

3 Construction des bandes de confiance

Nous considérons ici des bandes de confiance pour la courbe moyenne μ qui sont de la forme

$$\mathbb{P}\left(\mu(t) \in \left[\hat{\mu}(t) \pm c_\alpha\hat{\sigma}(t)\right],\ \forall t \in [0,T]\right) = 1-\alpha, \tag{3.1}$$

où la valeur du cœfficient c_α est inconnue, et dépend du niveau de confiance $1-\alpha$ souhaité, et $\hat{\sigma}(t)$ est un estimateur de l'écart-type de $\hat{\mu}(t)$. Le calcul de c_α est basé sur le fait que sous certaines hypothèses (Cardot et coll. 2013), le processus

$$Z(t) = \left(\hat{\mu}(t)-\mu(t)\right) / \hat{\sigma}(t), \quad t \in [0,T],$$

converge vers un processus Gaussien dans l'espace des fonctions continues $\mathcal{C}([0,T])$. On a alors

$$\mathbb{P}\left(\sup_{t\in T}|Z(t)| \le c_\alpha\right) = \mathbb{P}\left(\mu(t) \in \left[\hat{\mu}(t) \pm c_\alpha\hat{\sigma}(t)\right], \forall t \in [0,T]\right) \tag{3.2}$$

et il suffit donc de déterminer c_α, le quantile d'ordre $1-\alpha$ de la variable aléatoire réelle $\sup_{t\in[0,T]}|Z(t)|$ pour construire complètement la bande de confiance. La distribution du sup de processus Gaussiens n'est connue explicitement que pour quelques cas particuliers, le mouvement brownien par exemple.

Nous proposons deux approches pour déterminer la valeur de c_α. La première repose sur une estimation directe de l'écart-type et la simulation des processus Gaussiens $Z(t)$. La seconde, qui ne nécessite pas de disposer d'estimateur de la variance, repose sur des techniques de ré-échantillonnage où à la fois l'écart-type et la valeur de c_α sont obtenus à partir des réplications bootstrap.

3.1 Construction de bandes de confiance par simulation de processus Gaussiens

Les étapes de l'algorithme sont les suivantes :

1) Tirer l'échantillon s de taille n à l'aide du plan de sondage p et calculer l'estimateur $\hat{\mu}$ ainsi que l'estimateur $\hat{\gamma}(r,t)$ de la fonction de covariance $\gamma(r,t)$, $r,t \in [0,T]$.
2) Simuler M courbes Z_m, $m = 1\ldots, M$, de même loi que Z où Z est un processus Gaussien d'espérance 0 et de fonction de covariance ρ où $\rho(r,t) = \hat{\gamma}(r,t) / \left(\hat{\gamma}(r)\hat{\gamma}(t)\right)^{1/2}$, $r,t \in [0,T]$.
3) Déterminer c_α, le quantile d'ordre $1-\alpha$ des variables, $\left(\sup_{t\in[0,T]} \left|Z_m(t)\right|\right)_{m=1,\ldots,M}$.

Cet algorithme, très rapide et facile à mettre en œuvre, a déjà été proposé, dans le cadre d'observations i.i.d. par Faraway (1997), Cuevas et coll. (2006) et Degras (2011) pour construire des bandes de confiance. On trouvera une justification asymptotique rigoureuse de cette approche dans Cardot et coll. (2013) pour l'échantillonnage dans des populations finies.

3.2 Construction des bandes de confiance par bootstrap

Dans ce travail, nous utilisons la méthode de bootstrap proposée par Gross (1980) pour l'ÉASSR et les extensions proposées par Chauvet (2007) pour les plans STRAT et πps. Elle repose sur le principe suivant : l'échantillon s est utilisé pour simuler une population fictive U^* dans laquelle nous sélectionnons plusieurs échantillons bootstrappés. La mise en œuvre de cet algorithme n'est pas immédiate lorsque le rapport $1/\pi_k$ n'est pas entier. De nombreuses variantes ont été proposées dans la litérature pour tenir compte du cas général et nous avons décidé d'adopter celle initialement proposée par Booth, Butler et Hall (1994) pour le plan d'ÉASSR.

Considérons que l'échantillon s de taille n a été sélectionné à l'aide du plan de sondage p et soit $\hat{\mu}$ l'estimateur de μ calculé à partir de s.

Algorithme général du bootstrap

1) Dupliquer chaque individu $k \in s$, $\lfloor 1/\pi_k \rfloor$ fois, où [.] désigne la partie entière. On complète la population ainsi obtenue en sélectionnant un échantillon dans s avec une probabilité d'inclusion $\alpha_k = 1/\pi_k - \lfloor 1/\pi_k \rfloor$. Soit Y_k^*, $k \in U^*$ la valeur de la variable d'intérêt sur la pseudo-population.
2) Tirer M échantillons s_m^*, $m = 1,\ldots,M$, de taille n dans la pseudo-population U^* à l'aide du plan de sondage p^* avec des probabilités d'inclusion π_k^* et calculer

$$\hat{\mu}_m^*(t) = \frac{1}{N}\sum_{k\in s_m^*} \frac{Y_k^*(t)}{\pi_k^*},\ t \in [0,T] \text{ et } m = 1,\ldots,M.$$

3) Estimer la fonction $\hat{\sigma}(t)$ par l'écart-type empirique corrigé des $\hat{\mu}_m^*(t)$, $m = 1, \ldots, M$,

$$\hat{\sigma}^2(t) = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\mu}_m^*(t) - \hat{\mu}_\bullet^*(t)\right)^2,$$

où

$$\hat{\mu}_\bullet^*(t) = \frac{1}{M}\sum_{m=1}^{M}\hat{\mu}_m^*(t) \text{ et } t \in [0,T].$$

4) Choisir c_α comme le quantile d'ordre $1-\alpha$ des variables

$$\left(\sup_{t\in[0,T]}\frac{|\hat{\mu}_m^*(t) - \hat{\mu}(t)|}{\hat{\sigma}(t)}\right)_{m=1,\ldots,M}.$$

Une technique similaire à celle utilisée lors de l'étape 4 de l'algorithme a été utilisée par Bickel et Krieger (1989) pour construire des bandes de confiance de la fonction de répartition.

Le plan d'ÉASSR utilise l'algorithme général du bootstrap pour $\pi_k^* = n/N$, et pour le plan STRAT, nous appliquons dans chaque strate U_h, pour $h = 1, \ldots, H$, l'algorithme utilisé pour le plan d'ÉASSR avec $\pi_k^* = n_h / N_h \;\; k \in U_h$. On retrouve dans ce cas, l'algorithme proposé par Booth et coll. (1994).

L'adaptation de l'algorithme du bootstrap au plan πps a été proposée par Chauvet (2007). Elle consiste à sélectionner lors de l'étape 2 de l'algorithme général, l'échantillon s^* dans U^* avec les probabilités d'inclusion

$$\pi_k^* = \frac{nx_k}{\sum_{k\in U^*} x_k}.$$

Cette modification est nécéssaire pour respecter la contrainte de taille fixe lors du rééchantillonnage. Les probabilités d'inclusion π_k^* sont également utilisées pour estimer $\hat{\mu}_m^*$ lors de l'étape 2 de l'algorithme général. La séléction d'un échantillon πps peut être réalisée en utilisant l'algorithme du cube avec la variable d'équilibrage π. Dans ces conditions, un tri aléatoire dans la population U (resp. U^*) avant le tirage de s (resp. s_m^*) est souhaitable afin d'obtenir un plan de sondage proche de l'entropie maximale (Chauvet 2007, Tillé 2011). Chauvet (2007) donne également des résultats asymptotiques concernant la convergence de l'estimateur de la variance obtenu dans le cas du bootstrap du plan πps.

Enfin, il est également possible d'adapter cet algorithme général pour estimer la fonction de variance de l'estimateur $\hat{\mu}_{MA}$. Lors de l'étape 1 de l'algorithme, on calcule également les valeurs $\mathbf{x}_k^*$ de $\mathbf{x}_k$ dans la pseudo-population U^*. En utilisant le fait que l'estimateur assisté par un modèle linéaire est une fonction nonlinéaire d'estimateurs de type Horvitz-Thompson, la valeur bootstrappée $\hat{\mu}_{MA}^*$ de $\hat{\mu}_{MA}$ sur l'échantillon s_m^* est calculée selon

$$\hat{\mu}_{MA}^*(t) = \frac{1}{N}\sum_{k\in s_m^*}\frac{Y_k^*(t) - \mathbf{x}_k^{*'}\hat{\boldsymbol{\beta}}^*(t)}{\pi_k^*} + \frac{1}{N}\left(\sum_{k\in U}\mathbf{x}_k'\right)\hat{\boldsymbol{\beta}}^*(t)$$

où $\hat{\boldsymbol{\beta}}^{*}(t) = \left(\sum_{s_m^*} \mathbf{x}_k^* \mathbf{x}_k^{*\prime}\right)^{-1} \sum_{s_m^*} \mathbf{x}_k^* Y_k^*(t)$. Comme le remarquent Canty et Davison (1999) le fait d'utiliser le total de la variable $\mathbf{x}_k$ sur la population U au lieu de la pseudo-population U^* conduit à de meilleurs résultats en particulier quand cette variable présente des valeurs extrêmes.

4 Étude de la courbe de consommation moyenne d'électricité

Nous disposons d'une population U composée de $N = 15\,069$ courbes de consommation électrique mesurées toutes les demi-heures pendant deux semaines consécutives. Nous avons $D = 336$ points de mesure pour chaque semaine et nous souhaitons estimer la courbe moyenne de consommation de la deuxième semaine. On note $\mathbf{Y}'_k = \left(Y_k(t_1), \ldots, Y_k(t_D)\right)$, la consommation d'électricité de l'individu $k \in U$ mesurée la deuxième semaine et $\mathbf{X}'_k = \left(X_k(t_1), \ldots, X_k(t_D)\right)$ sa consommation au cours de la première semaine. La consommation moyenne de chaque individu k durant la première semaine, $x_k = \sum_{d=1}^{D} X_k(t_d) / D$, qui est une information simple et peu coûteuse à transmettre, sera utilisée comme information auxiliaire. Cette variable (réelle) qui est connue pour tous les éléments k de la population est fortement liée à la courbe de consommation courante. On note sur la figure 4.1 que la consommation courante en chaque t est quasiment proportionnelle à la consommation moyenne de la semaine précédente.

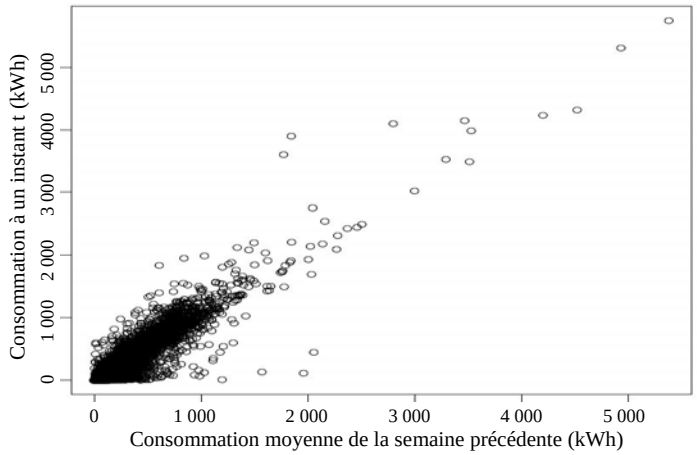


Figure 4.1 Représentation de la consommation à un instant t en fonction de la consommation moyenne de la semaine précédente

4.1 Description des stratégies utilisées

Nous considérons des échantillons de taille fixe $n = 1\,500$ obtenus selon différents plans de sondage. Les stratégies présentées sont répétées I fois afin d'évaluer et de comparer leurs performances.

1. *ÉASSR et estimateur de Horvitz-Thompson.*

La mise en œuvre de ce plan est simple, l'estimateur de Horvitz-Thompson de la courbe moyenne est donné par (2.6) et l'estimateur de sa covariance par (2.7).

2. *Sondage stratifié STRAT et estimateur de Horvitz-Thompson*.

Le plan stratifié est très efficace si les strates sont homogènes par rapport à la variable d'intérêt. Dans ce travail, nous avons utilisé l'algorithme des k -means afin de constituer les strates et nous avons considéré $H = 10$ strates. Une première stratification (STRAT 1) a été effectuée à partir de la classification des trajectoires discrétisées $\mathbf{X}'_k$ de la première semaine. Une seconde stratification, qui utilise uniquement l'information agrégée x_k a également été considérée. Elle est notée STRAT 2.

Les tailles des strates N_h obtenues en utilisant les deux stratifications ainsi que les tailles n_h optimales, selon (2.5), des échantillons à sélectionner dans chaque strate sont données dans les tableaux 4.1 et 4.2. Dans les deux cas, les strates sont numérotées en ordre croissant par rapport à la consommation moyenne de chaque strate. Plus précisement, la strate 1 correspond aux faibles consommateurs et la strate 10 est composée des 10 plus gros consommateurs d'électricité. Notons que la première stratification, qui nécessite de connaître la consommation d'électricité à chaque instant de mesure t, exige plus d'information que la deuxième stratification. La courbe moyenne est construite en utilisant (2.3) et sa covariance est estimée par (2.4).

Tableau 4.1
STRAT 1 : stratification à partir des courbes. Les strates sont construites à partir des courbes de la semaine 1. L'allocation n_h optimale est calculée à partir des courbes de la semaine 1.

h	1	2	3	4	5	6	7	8	9	10
N_h	3 866	4 769	623	2 690	664	1 251	806	328	62	10
n_h	212	345	87	242	117	179	172	101	35	10

Tableau 4.2
STRAT 2 : stratification à partir de la consommation moyenne x_k. L'allocation optimale n_h est calculée à partir de la consommation moyenne de la semaine 1.

h	1	2	3	4	5	6	7	8	9	10
N_h	3 257	4 236	3 139	1 937	1 189	731	415	125	30	10
n_h	260	293	248	204	159	133	111	56	26	10

3. *Sondage πps et estimateur de Horvitz-Thompson.*

Nous avons utilisé l'algorithme du cube proposé par Deville et Tillé (2004) et Chauvet et Tillé (2006) où les probabilités d'inclusion sont proportionnelles à x_k, $k \in U$. Afin d'avoir un plan de sondage proche de l'entropie maximale, un tri aléatoire de la population est effectué avant le tirage de l'échantillon s. La covariance de l'estimateur de la moyenne est estimée à l'aide de la formule (2.9). L'algorithme du cube est disponible sous R dans le package *sampling*, fonction *samplecube* et une macro SAS est disponible sur le site web de l'INSEE (Institut National de Statistique et des Etudes Economiques).

4. *ÉASSR et estimateur MA.*

L'estimateur $\hat{\mu}_{MA}$ assisté par le modèle ξ est construit à l'aide de l'information auxiliaire donnée par $\mathbf{x}'_k = (1, x_k)$ où x_k est la consommation moyenne de la semaine précédente. Dans ces conditions, $\hat{\mu}_{MA}$ est la somme sur toute la population U des valeurs estimées $\hat{Y}_k$ par le modèle (voir formule (2.13)). La covariance de l'estimateur de la moyenne est estimée à l'aide de la formule (2.15).

4.2 Erreur d'estimation de la courbe moyenne

L'erreur d'estimation de la courbe moyenne μ aux instants $t_1, \ldots, t_{336}$, est évaluée selon le critère

$$R_2\left(\hat{\mu}\right) = \frac{1}{336}\sum_{i=1}^{336}\left(\hat{\mu}\left(t_i\right) - \mu\left(t_i\right)\right)^2 \approx \frac{1}{T}\int_0^T \left(\hat{\mu}\left(t\right) - \mu\left(t\right)\right)^2 dt.$$

Les résultats sont présentés dans le tableau 4.3 pour $I = 10\,000$ simulations (réplications). Ils montrent clairement que, pour cette étude, la prise en compte de la consommation totale de la semaine précédente permet d'améliorer de manière importante la précision de l'estimation de la moyenne par rapport à l'échantillonnage aléatoire simple sans remise en divisant l'erreur quadratique moyenne R_2 par 5. Parmi les différentes stratégies, les plus performantes semblent être celles qui prennent en compte l'information auxiliaire via les probabilités d'inclusion (STRAT, πps et systématique proportionnel à la taille).

Tableau 4.3
Erreur quadratique R_2 d'estimation de la moyenne μ, avec $I = 10\,000$ réplications.

Stratégie	moyenne	1er quartile	médiane	3ème quartile
ÉASSR	40,53	10,82	22,16	51,09
STRAT (1)	5,78	3,68	5,08	7,07
STRAT (2)	6,49	4,03	5,48	7,88
πps	7,06	3,99	5,52	8,16
$\pi - ps$ systématique	6,73	3,85	5,20	8,07
MA	8,29	5,24	7,14	10,06

4.3 Taux de couverture et largeur des bandes de confiance

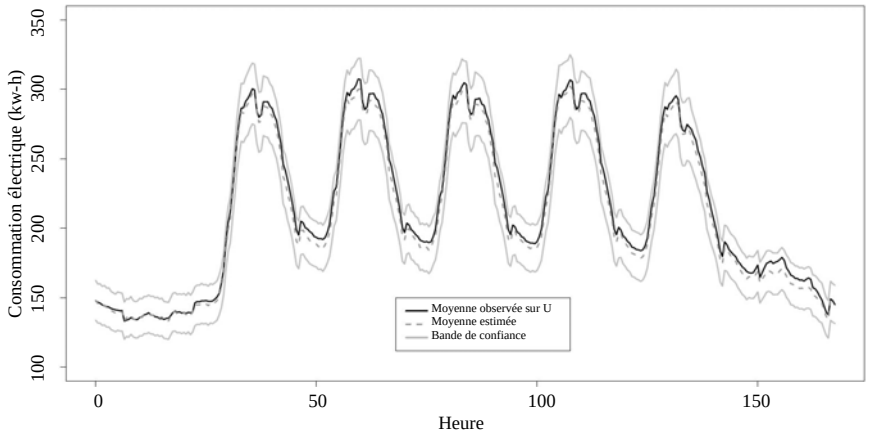
La construction des bandes de confiance de niveau $1 - \alpha$ nécessite le calcul des quantiles d'ordre $1 - \alpha$ du supremum de processus Gaussiens.

Pour ne pas privilégier une méthode de construction de bande de confiance par rapport à l'autre, nous avons appliqué les deux algorithmes sur un même échantillon s et nous avons considéré le même nombre M de processus. Ce nombre M varie d'un estimateur à l'autre en raison des temps de calculs nécessaires pour les approches de type bootstrap (voir Section 4.4).

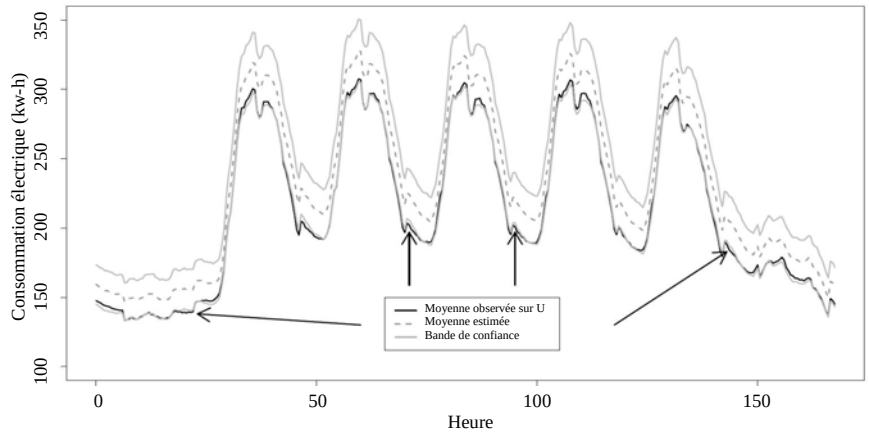
Le taux de couverture empirique est la proportion de fois, parmi les $I = 2\,000$ réplications, où la vraie courbe moyenne μ se trouve, pour tous les instants t, à l'intérieur de la bande de confiance construite à partir d'une estimation $\hat{\mu}$. Nous avons représenté sur la figure 4.2 deux exemples de bandes de confiance (courbes grises continues) construites à partir des courbes estimées (courbes grises pointillées). Sur la figure 4.2(A), nous constatons que la vraie courbe moyenne sur la population (courbe noir continue) est à l'intérieur de la bande de confiance à chaque instant. À l'opposé, sur la figure 4.2(B), nous constatons que la courbe moyenne de la population est en général surestimée et qu'il existe quelques instants (indiqués par les flèches) où la courbe observée sort de la bande de confiance. Les taux de couverture empiriques sont présentés dans le tableau 4.4.

Les deux méthodes de construction des bandes de confiance donnent des taux de couverture similaires et assez proches des taux nominaux souhaités (95 % et 99 %). Les résultats semblent cependant

légèrement moins satisfaisants pour les plans πps et pour l'approche MA pour lesquels la variance de l'estimateur est complexe et plus difficile à estimer précisément.



(A) La courbe moyenne observée est à l'intérieur de la bande de confiance.



(B) La moyenne observée est à l'extérieur de la bande de confiance aux instants indiqués par les flèches.

Figure 4.2 Exemples de bande de confiance

Tableau 4.4
Taux de couverture empirique (en %), pour $I = 2\,000$ réplications.

Méthodes	Nombre M de processus	Bootstrap		Processus Gaussien	
		$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$
ÉASSR	5 000	94,95	98,85	94,80	98,70
STRAT (1)	5 000	93,92	98,34	94,09	98,43
STRAT (2)	5 000	94,3	98,45	94	98,55
πps	1 000	94,73	98,77	93,87	98,61
MA	5 000	94,3	98,5	92,85	98,15

Un autre indicateur intéressant est la largeur moyenne de la bande de confiance,

$$\frac{1}{336}\sum_{i=1}^{336} 2c_\alpha \hat{\sigma}\left(t_i\right) \approx \frac{1}{T}\int_0^T 2c_\alpha \hat{\sigma}\left(t\right) dt$$

dont les valeurs sont présentées dans le tableau 4.5. Les deux méthodes fournissent des bandes de confiance dont les largeurs sont similaires. On note également que l'utilisation de la variable auxiliaire permet de diminuer sensiblement la largeur moyenne des bandes, celle-ci étant divisée par deux si on considère un des plans stratifiés plutôt qu'un plan d'ÉASSR.

Tableau 4.5
Largeur moyenne des bandes de confiance, pour $I = 2\,000$ réplications.

Méthodes	Nombre M de processus	Bootstrap		Processus gaussien	
		$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$
ÉASSR	5 000	35,98	43,35	35,99	43,19
STRAT (1)	5 000	16,64	18,92	16,62	18,88
STRAT (2)	5 000	17,58	19,99	17,55	19,94
πps	1 000	17,85	20,31	17,62	19,93
MA	5 000	19,88	22,65	19,75	22,44

Les figures 4.3 et 4.4 présentent les largeurs des bandes de confiance pour un niveau $\alpha = 0,05$, pour chaque instant, selon qu'elles soient ponctuelles ($c_\alpha = 1,96$), estimées par simulations de processus gaussiens ou bien obtenues en considérant l'approche basée sur l'inégalité de Bonferroni appliquée en chaque point de mesure. On a alors, dans ce dernier cas, $c_\alpha = 3,793048$, le quantile d'ordre $1 - 0,05 / (336 \times 2)$ d'une loi $N(0,1)$. Les bandes obtenues par Bonferroni sont conservatives et considèrent en quelque sorte le pire des cas en termes d'information, celui de l'indépendance des intervalles ponctuels. On peut remarquer que l'approche par simulation permet de réduire sensiblement la largeur moyenne des bandes en comparaison avec Bonferroni lorsque le plan ne permet pas de prendre en compte toute l'information temporelle des données (figure 4.3). À l'opposé, pour le plan stratifié (figure 4.4) qui permet une estimation précise de la courbe moyenne, la bande de confiance construite par simulation est proche de celle de Bonferroni, ce qui s'interprète intuitivement comme le fait que quasiment toute l'information a été capturée par le plan de sondage.

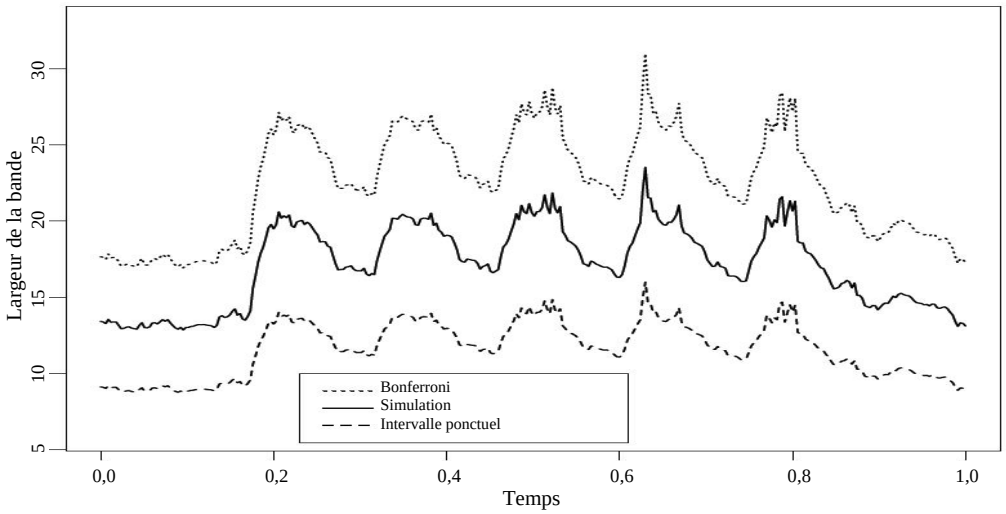


Figure 4.3 Échantillonnage aléatoire simple sans remise. Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni ($\alpha = 0,05$)

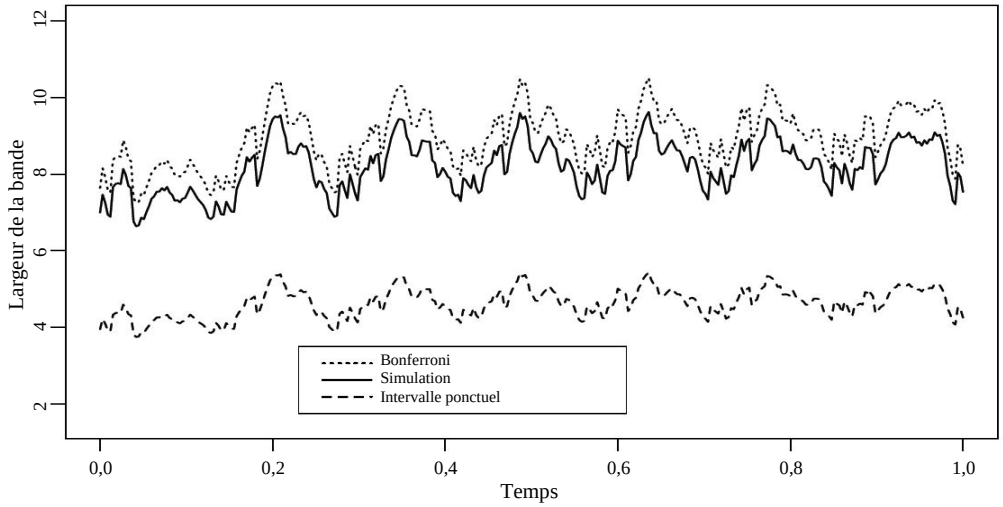


Figure 4.4 Sondage stratifié (STRAT 1). Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni (avec $\alpha = 0,05$)

4.4 Temps de calcul

Les temps de calcul avec la méthode par bootstrap sont largement supérieurs, de l'ordre d'un facteur de 1 à 1 000, à ceux de la méthode par simulations de processus gaussiens (voir tableau 4.6). Cette différence importante provient du fait que les méthodes de bootstrap nécessitent de répéter tout le processus d'estimation pour chaque échantillon bootstrapé : construction de la population fictive, tirage d'un nouvel échantillon, calcul de l'estimateur. On remarque également que les plans qui font intervenir de l'information auxiliaire sont moins rapides que le plan d'ÉASSR même si utilisés individuellement leur temps de calcul reste tout à fait raisonnable.

Tableau 4.6
Temps d'exécution d'une simulation en secondes pour M = 5 000 réplications. Les stratégies ÉASSR, MA et STRAT ont été programmés avec R et πps avec SAS.

Stratégie	Bootstrap	Processus gaussiens
ÉASSR	1 170,6	1,0
STRAT	1 839,5	1,4
πps	5 020,0	7,3
MA	3 156	1,4

5 Conclusion et perspectives

Nous avons, dans ce travail, mis en œuvre et comparé différentes stratégies permettant de prendre en compte de l'information auxiliaire pour l'estimation, et la construction de bandes de confiance, de la moyenne de données qui sont des courbes. Cette information peut être prise en compte au moment de l'échantillonnage en considérant des plans à probabilités inégales ou bien lors de l'estimation avec un

sondage aléatoire simple sans remise assisté par un modèle de régression à réponse fonctionnelle. Il apparaît clairement, sur notre exemple de courbes de charge d'électricité, que la connaissance des consommations totales une semaine avant, permet d'améliorer de manière importante la précision des estimateurs de la moyenne par rapport à un sondage de type ÉASSR.

Par ailleurs, dans ce contexte d'échantillons de taille importante et de données de grande dimension, il semble aussi possible de construire, pour ces différentes stratégies, des bandes de confiance qui ont des taux de couverture empiriques proches des taux souhaités. Les performances des deux approches proposées, estimation de la fonction de covariance et simulation de processus Gaussiens ou Bootstrap, semblent comparables en termes de largeur des bandes de confiance et la principale différence porte sur les temps de calcul. Le bootstrap qui semble plus général, puisqu'il ne nécessite pas de disposer d'un estimateur performant de la fonction de covariance, se révèle beaucoup plus lent en pratique.

Il y a parfois, dans ces flux de données de grande taille, des pertes d'information qui proviennent de problèmes de transmission du signal. L'opérateur observe donc au final certaines trajectoires de manière incomplète. Cette question, de non réponse partielle, peut sans doute être abordée en considérant des adaptations des techniques classiques de non réponse (Haziza 2009) au cadre fonctionnel. Une question primordiale concerne alors la construction de bons estimateurs de la fonction de covariance.

Remerciements

Nous remercions les arbitres anonymes ainsi que Guillaume Chauvet et Jean-Claude Deville pour leurs remarques fructueuses qui ont permis d'améliorer ce travail.

Bibliographie

Bickel, P., et Krieger, A. (1989). Confidence bands for a distribution function using the bootstrap. *Journal of the American Statistical Association*, 84, 95-100.

Booth, J., Butler, R. et Hall, P. (1994). Bootstrap methods for finite population. *Journal of the American Statistical Association*, 89, 1282-1289.

Canty, A.J., et Davison, A.C. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician*, 48, 379-391.

Cardot, H., Chaouch, M., Goga, C. et Labruère, C. (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference*, 140, 75-91.

Cardot, H., Degras, D. et Josserand, E. (2013). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*, 19, 2067-2097.

Cardot, H., Goga, C. et Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic J. of Statistics*, 7, 562-596.

Cardot, H., et Josserand, E. (2011). Horvitz-thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98, 107-118.

Chaouch, M., et Goga, C. (2012). Using complex surveys to estimate the l1-median of a functional variable: Application to electricity load curves. *Revue Internationale de Statistique*, 80, 40-59.

Chauvet, G. (2007). Méthodes de bootstrap en population finie. Thèse de doctorat, Université de Rennes II.

Chauvet, G., et Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21, 53-61.

Cochran, W. (1977). Sampling techniques. New York: John Wiley & sons, Inc., 3ième Édition.

Cuevas, A., Febrero, M. et Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, 51, 1063-1074.

Dauxois, J., et Pousse, A. (1976). Les analyse factorielles en calcul des probabilités et en statistique : essai d'étude synthétique. Thèse de doctorat, Université Paul Sabatier, Toulouse.

Degras, D. (2011). Simultaneous confidence bands for parametric regression with functional data. *Statistica Sinica*, 21(4), 1735-1765.

Dessertaine, A. (2008). Estimation de courbes de consommation électrique à partir des mesures synchrones. Dans *Méthodes de Sondages* (Éds., P. Guibert, D. Haziza, A. Ruiz-Gazen et Y. Tillé), Dunod, France, 353-357.

Deville, J. (1974). Méthodes statistiques et numériques de l'analyse harmonique. Ann. Insee, 15, 3-104.

Deville, J., et Tillé, Y. (2004). Efficient balanced sampling: The cube algorithm. *Biometrika*, 91, 893-912.

Deville, J., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Faraway, J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3), 254-261.

Ferraty, F., et Romain, Y., editors (2011). *Oxford Handbook of Functional Data Analysis*. Oxford University Press.

Goga, C., et Ruiz-Gazen, A. (2013). Efficient estimation of nonlinear finite population parameters using nonparametrics, à paraitre dans le *Journal of the Royal Statistical Society*, Series B, DOI: 10.111/rssb.12024.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, 181-184.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

Haziza, D. (2009). Imputation and inference in the presence of missing data. Dans *Sample Surveys: Theory Methods and Inference*, volume 29 de *Handbook of Statistics*, (Éds., C. Rao et D. Pfeffermann), North-Holland, 215-246.

Madow, W. (1949). *On the theory of systematic sampling*, ii. Annals of Mathematical Statistics, 19, 535-545.

Ramsay, J., et Silverman, B. (2005). *Functional data analysis*. Springer, New York, deuxième édition.

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

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model assisted survey sampling.* Springer.

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Tillé, Y. (2011). Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation. *Techniques d'enquête*, 37, 233-246.

Yates, F., et Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 235-261.


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

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 333-354
Statistique Canada, N° 12-001-X au catalogue

Critère d'information bayésien fondé sur la pseudo-vraisemblance pour la sélection de variables dans les données d'enquête

Chen Xu, Jiahua Chen et Harold Mantel[1]

Résumé

Les modèles de régression sont utilisés couramment pour analyser les données d'enquête lorsque l'on souhaite déterminer quels sont les facteurs influents associés à certains indices comportementaux, sociaux ou économiques au sein d'une population cible. Lorsque des données sont recueillies au moyen d'enquêtes complexes, il convient de réexaminer les propriétés des approches classiques de sélection des variables élaborées dans des conditions i.i.d. ne faisant pas appel au sondage. Dans le présent article, nous dérivons un critère BIC fondé sur la pseudo vraisemblance pour la sélection des variables dans l'analyse des données d'enquête et proposons une approche de vraisemblance pénalisée dans des conditions de sondage pour sa mise en œuvre. Les poids de sondage sont attribués comme il convient pour corriger le biais de sélection causé par la distorsion entre l'échantillon et la population cible. Dans un cadre de randomisation conjointe, nous établissons la cohérence de la procédure de sélection proposée. Les propriétés en échantillon fini de l'approche sont évaluées par des analyses et des simulations informatiques en se servant de données provenant de la composante de l'hypertension de l'Enquête sur les personnes ayant une maladie chronique au Canada de 2009.

Mots-clés : Sélection des variables; poids de sondage; inférence sous le modèle et sous le plan; BIC; vraisemblance pénalisée; cohérence de sélection.

1 Introduction

La détermination des facteurs influents associés à certains indices comportementaux, sociaux ou économiques dans une population cible est un sujet d'intérêt commun à de nombreux domaines de recherche scientifique. Par exemple, les sociologues souhaitent cerner les facteurs importants qui ont une incidence sur le taux de chômage dans une région particulière et les épidémiologistes cherchent à découvrir les comportements à risque associés aux maladies. Dans ce genre d'études, les chercheurs commencent souvent par effectuer un sondage auprès de la population cible (par exemple, Rahiala et Teräsvirta 1993; Korn et Graubard 1999; Wolfson 2004). Pour cela, un échantillon représentatif est sélectionné et des mesures des variables d'intérêt sont recueillies auprès des unités échantillonnées. Un modèle de régression est habituellement employé pour résumer l'information contenue dans les données. Ce modèle explique les variations de la variable réponse au moyen d'une fonction simple des variables explicatives (covariables). Lorsqu'ils ne disposent pas d'information a priori, les chercheurs peuvent recueillir des renseignements sur de nombreuses variables explicatives possibles. Ils peuvent ensuite atteindre l'objectif consistant à déterminer quels sont les facteurs influents en appliquant une procédure de sélection de variables.

La sélection des variables est un aspect fondamental de la modélisation statistique. Dans des conditions ne faisant pas appel au sondage, on a élaboré des critères de sélection classiques pour évaluer et sélectionner les variables possibles. La statistique C_p de Mallow (Mallows 1973), la validation croisée

1. Chen Xu et Jiahua Chen, Département de statistique, Université de la Colombie-Britannique, Vancouver (C.-B.), Canada, V6T 1Z4. Courriel : chen.xu@stat.ubc.ca et jhchen.stat.ubc.ca; Harold Mantel, Division de la recherche et de l'innovation en statistique, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6. Courriel : Harold.Mantel@statcan.gc.ca.

(généralisée) (CV/ GCV; Stone 1974; Craven et Wahba 1979), le critère d'information d'Akaike (AIC; Akaike 1973) et le critère d'information bayésien (BIC; Schwarz 1978) en sont des exemples. Tous ces critères sont fort utiles et produisent des inférences significatives en pratique.

Malgré l'abondance de la littérature sur la sélection des variables, peu d'attention a été accordée à ce sujet dans le contexte du sondage. L'application des méthodes de sélection de variables à des données d'enquête peut donner lieu à de nombreuses complications. Nous nous concentrons sur les problèmes qui découlent de caractéristiques particulières des enquêtes. Premièrement, les données recueillies par sondage sont habituellement obtenues auprès d'une population finie sans remise des unités échantillonnées, de sorte qu'elles possèdent une structure de dépendance intrinsèque. Deuxièmement, dans les plans de sondage complexes, les probabilités d'inclusion des unités échantillonnées varient souvent dans la population cible. Par conséquent, la corrélation entre la réponse et les covariables dans l'échantillon peut être faussée comparativement à celle observée dans la population. Cela pourrait être le cas lorsque certaines parties de la population sont échantillonnées de manière plus intensive que d'autres. Ne pas tenir compte du plan de sondage dans le processus de sélection peut donner des résultats biaisés pour la population cible.

Dans la littérature, les poids de sondage sont souvent utilisés pour estimer les paramètres des modèles de régression fondés sur des données d'enquête. Les estimations pondérées des coefficients de régression aident à éviter que l'inférence soit biaisée par l'échantillonnage informatif (Pfeffermann 1993; Fuller 2009, section 6.3; Skinner 2012). Même si l'estimation et la sélection du modèle ont chacune leurs propres objectifs, elles présentent souvent un lien cohérent dans un processus de modélisation. Il est donc naturel de conjecturer que l'utilisation de poids de sondage a un effet positif sur la sélection des variables.

Dans cet esprit, nous étudions l'utilisation de la pseudo-vraisemblance pour tenir compte des poids de sondage, et nous dérivons un critère BIC fondé sur la pseudo-vraisemblance pour la sélection des variables dans les données d'enquête. Nous proposons en outre une procédure fondée sur la pseudo-vraisemblance pénalisée (PVP) pour la mise en œuvre numérique du critère proposé. Dans un cadre de randomisation conjointe, nous prouvons que la nouvelle procédure permet systématiquement de repérer les variables influentes. Nous évaluons la méthode de sélection pondérée au moyen d'études en simulation en utilisant des données provenant de l'Enquête sur les personnes ayant une maladie chronique au Canada réalisée en 2009.

La présentation de l'article est la suivante. À la section 2, nous décrivons le mécanisme de randomisation conjointe et le modèle de superpopulation. À la section 3, nous dérivons le critère BIC fondé sur la pseudo-vraisemblance pour l'analyse des données d'enquête et proposons de l'appliquer au moyen de la procédure PVP. À la section 4, nous étudions le comportement asymptotique de la procédure BIC proposée. À la section 5, nous faisons appel à des études numériques pour évaluer plus en détail les résultats de notre approche et à la section 6, nous présentons nos conclusions. Nous donnons les preuves des théorèmes dans un supplément technique distinct [Xu et Chen (2012)], dans lequel figure également la dérivation du critère BIC proposé.

2 Inférence conjointe et superpopulation

Le comportement aléatoire d'une procédure d'inférence découle principalement du caractère aléatoire des données. Dans le contexte des enquêtes, l'ensemble d'unités échantillonnées est aléatoire en raison du

plan d'échantillonnage probabiliste. Parallèlement, la valeur de chaque unité échantillonnée peut être considérée comme un résultat aléatoire provenant d'une superpopulation conceptuellement infinie (Royall 1976).

Dans une analyse fondée sur le plan de sondage, la population finie est considérée comme non aléatoire et toutes les mesures des unités d'échantillonnage sont constantes. Les paramètres d'intérêt sont les quantités dans la population finie, telles que le total ou la médiane de la population. L'inférence statistique est évaluée en se basant sur le caractère aléatoire découlant du plan de sondage probabiliste.

On peut également considérer le caractère aléatoire induit par le plan de sondage comme un artéfact. Les mesures des unités échantillonnées sont alors des réalisations indépendantes d'une variable aléatoire provenant d'un modèle probabiliste de la superpopulation postulée. Des paramètres d'intérêt sont reliés au modèle hypothétique et les inférences sous le modèle sont évaluées uniquement en se basant sur la randomisation introduite par le modèle.

Une troisième approche, appelée inférence sous le modèle et le plan, incorpore la randomisation venant du plan de sondage ainsi que du modèle. Sous un tel mécanisme de randomisation conjointe, la population finie est considérée comme un échantillon aléatoire tiré d'une superpopulation. L'échantillon d'enquête est considéré comme résultant d'un échantillonnage de deuxième phase de la superpopulation. Les paramètres d'intérêt peuvent être des paramètres du modèle ou des paramètres de population finie. Sous ce mécanisme, les inférences au sujet des paramètres de la population finie sont motivées par le modèle de superpopulation. L'inférence sous le modèle et le plan de sondage peut être plus efficace que les approches fondées purement sur le plan lorsque la population finie est bien décrite par le modèle de superpopulation. Comparativement aux approches fondées purement sur le modèle, elle protège contre la violation du modèle et est par conséquent généralement plus robuste (voir, par exemple, Binder et Roberts 2003; Kalton 1983).

Nous étudions le problème de la sélection des variables sous le mécanisme de randomisation conjointe. Soit $\mathcal{D} = \{1, \ldots, N\}$ une population finie constituée de N unités échantillonnées. Les mesures faites sur la i^{e} unité sont désignées $(y_i, \mathbf{x}_i)$, où y_i est la réponse d'intérêt et $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ est un vecteur de variables explicatives de dimension p (vecteur de covariables). Ces éléments sont considérés comme des réalisations indépendantes de $(Y, \mathbf{X})$ provenant d'une superpopulation. Nous postulons un modèle linéaire généralisé (MLG) sur la superopulation de la façon suivante. Conditionnellement à $\mathbf{X}$, la loi de Y appartient à une famille exponentielle naturelle, dont la densité prend la form

$$f(y;\theta) = c(y)\exp\{\theta y - b(\theta)\}. \tag{2.1}$$

θ est connu comme étant le paramètre naturel de $f(y;\theta)$ tel que $b'(\theta) = E[Y|\mathrm{X}] \equiv \mu$ et $b''(\theta) = \mathrm{Var}[Y|\mathrm{X}] \equiv \sigma^2$, et $c(y)$ est une mesure de base non négative. L'influence de la variable explicative $\mathbf{X}$ sur Y est exprimée par $g(\mu) = \mathbf{X}^T\boldsymbol{\beta}$ pour une certaine fonction de lien supposée $g(.)$, où le vecteur $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^T$ est le coefficient de régression de dimension p. Si $g(.)$ est le lien canonique, c'est-à-dire $g(\mu) = \theta$, alors nous avons $\theta = \mathbf{X}^T\boldsymbol{\beta}$. Pour simplifier, nous nous concentrons sur le lien canonique dans le présent article.

Sur la base de ce modèle, l'effet de la variable explicative est caractérisé par la taille du coefficient de régression correspondant. Dans les applications, un modèle complexe contenant de nombreuses variables aboutit souvent à un surajustement et à une médiocre capacité d'interprétation. Donc, il est souhaitable d'ajuster les données au moyen d'un modèle parcimonieux dans lequel de nombreux coefficients de régression sont estimés être nuls. Les variables explicatives dont les coefficients ne sont pas nuls sont alors considérées comme influant sur la réponse. À cette fin, nous supposons que $\boldsymbol{\beta}$ est idéalement parcimonieux et nous abordons le problème de sélection des variables en déterminant un modèle parcimonieux formé par les covariables dont les coefficients ne sont pas nuls.

3 Sélection fondée sur la pseudo-vraisemblance avec le BIC

3.1 Le BIC dans les enquêtes

Sous la spécification du modèle décrite à la section 2, il est clair que, si la mesure $\left(y_i, \mathbf{x}_i\right)$ est observée pour chaque unité de la population $\mathcal{D}$, le caractère aléatoire des données introduit par le plan de sondage probabiliste a complètement disparu. Dans cette situation, la sélection de variables influentes est fondée sur la population complète et les critères de sélection classiques élaborés dans des conditions ne faisant pas appel au sondage (fondées purement sur le modèle) demeurent valides pour l'inférence sous le modèle et le plan. En particulier, soit $s \subseteq \{1, \ldots, p\}$ un ensemble arbitraire de $\tau(s)$ covariables, qui correspond à un modèle possible de la forme (2.1). Le BIC fondé sur la population complète (Schwarz 1978) sélectionne le modèle (covariables) qui minimise

$$\mathrm{BIC}_N(s) = -2l_N\left(\breve{\boldsymbol{\beta}}_s\right) + \tau(s)\log N, \tag{3.1}$$

où $l_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log f\left(y_i; \mathbf{x}_i\boldsymbol{\beta}\right)$ est la fonction de vraisemblance pour la population complète et $\breve{\boldsymbol{\beta}}_s$ est le maximiseur de $l_N(\boldsymbol{\beta})$ fondé sur s. On peut constater que le BIC (3.1) est une fonction décroissante de la vraisemblance maximisée et une fonction croissante du nombre de variables incluses dans le modèle. Donc, un plus petit BIC implique un modèle plus simple (moins de variables explicatives), un meilleur ajustement (vraisemblance maximisée plus élevée), ou les deux. La préférence est donnée à un modèle présentant un équilibre entre la complexité et la qualité de l'ajustement.

Nous notons que le BIC sous population complète (3.1) est conceptuel, parce que l'observation de $\left(y_i, \mathbf{x}_i\right)$ pour toutes les unités de $\mathcal{D}$ est habituellement impossible dans les applications. Souvent, on tire plutôt de $\mathcal{D}$ un échantillon représentatif $d = \{i_1, \ldots, i_n\} \subset \{1, \ldots, N\}$ contenant n unités et les mesures sont observées en se basant sur les unités échantillonnées. En raison de la structure de dépendance intrinsèque des unités échantillonnées, il n'est généralement pas possible de calculer une vraisemblance complète sur d. Comme solution de rechange, pour l'inférence sous le modèle et le plan, on utilise fréquemment une fonction de pseudo-log-vraisemblance, qui prend la forme

$$l_n(\boldsymbol{\beta}) = \sum_{i \in d} w_i \log f(y_i; \boldsymbol{\beta}) \tag{3.2}$$

où $w_i = k / P(i \in d)$ désigne le poids de sondage de la i^{e} unité. Le paramètre d'échelle k dans w_i n'a aucune incidence analytique sur l'inférence fondée sur la pseudo-vraisemblance. Pour simplifier l'exposé, nous choisissons $k = n / N$ tel que $n^{-1} l_n(\boldsymbol{\beta})$ est sans biais sous le plan jusqu'à $N^{-1} l_N(\boldsymbol{\beta})$. La maximisation de $l_n(\boldsymbol{\beta})$ sur $\boldsymbol{\beta}$ mène à un estimateur du maximum de pseudo-vraisemblance (EMPV) $\hat{\boldsymbol{\beta}}$ pour $\boldsymbol{\beta}$, c'est-à-dire

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta} l_n(\boldsymbol{\beta}).$$

Sous les plans de sondage appropriés, $\hat{\boldsymbol{\beta}}$ est souvent convergent en $n^{-1/2}$ vers $\boldsymbol{\beta}$ dans le cadre de randomisation conjointe. L'idée d'utiliser la pseudo-vraisemblance pour l'inférence sur les paramètres du modèle est largement répandue dans la littérature (voir, par exemple, Binder 1983; Godambe et Thompson 1986; Molina et Skinner 1992).

Dans le présent article, nous tentons d'élaborer un analogue du critère BIC fondé sur la pseudo-vraisemblance. Partant de la formulation de la super-population décrite à la section 2, soit $\boldsymbol{\beta}_s$, le coefficient $\tau(s)$- dimensionnel du modèle s et soit ν_s, la densité a priori de $\boldsymbol{\beta}_s$. Alors, une fonction de pseudo-densité marginale des données est donnée par

$$P_n(\mathbf{y}|s) = \int L_n(\mathbf{y}; \boldsymbol{\beta}_s) \nu_s(\boldsymbol{\beta}_s) d\boldsymbol{\beta}_s$$

avec $L_n(\mathbf{y}; \boldsymbol{\beta}_s) = \exp\{l_n(\mathbf{y}; \boldsymbol{\beta}_s)\}$. Donc, nous pouvons considérer l'expression qui suit comme étant la pseudo-probabilité a posteriori du modèle s :

$$P_n(s|\mathbf{y}) = \frac{P_n(\mathbf{y}|s) P(s)}{\sum_{s \in S} P(s) P_n(\mathbf{y}|s)}, \tag{3.3}$$

où S désigne l'ensemble de tous les modèles possibles. Dans l'esprit de l'analyse bayésienne, le modèle ayant la $P_n(s|\mathbf{y})$ la plus élevée est considéré comme étant celui que les données soutiennent le plus. Puisque $\sum_{s \in S} P(s) P_n(\mathbf{y}|s)$ ne dépend d'aucun modèle particulier, la $P_n(s|\mathbf{y})$ la plus élevée est donnée par le modèle qui maximise la $P_n(\mathbf{y}|s) P(s)$ correspondante. Lorsque l'on utilise le prior uniforme $P(s) = \zeta$ et que l'on choisit le facteur d'échelle de pondération comme étant $k = n / N$, on obtient une approximation de Laplace sous certaines conditions de régularité (voir Xu et Chen 2012) :

$$-2 \log\{P_n(\mathbf{y}|s)\} = -2 l_n(\hat{\boldsymbol{\beta}}_s) + \tau(s) \log n + O_p(1).$$

D'où, nous choisissons le modèle s qui minimise

$$\text{BIC}_n(s) = -2 l_n(\hat{\boldsymbol{\beta}}_s) + \tau(s) \log n. \tag{3.4}$$

Comparativement au BIC sous population complète (3.1), le premier terme du BIC (3.4) est la pseudo-vraisemblance pondérée par les poids de sondage maximale, qui pourrait être utile pour éviter les erreurs dues à l'échantillonnage susceptibles de donner lieu à des inférences biaisées pour la population cible. Nous considérons (3.4) comme une version du BIC fondée sur la pseudo-vraisemblance dans le contexte des sondages. Dans le cadre de randomisation conjointe, nous établissons la cohérence de sélection lorsqu'on utilise le BIC (3.4) par une procédure d'application via la pseudo-vraisemblance pénalisée (PVP), comme nous le verrons à la section 4.

3.2 Application du BIC au moyen de la pseudo-vraisemblance pénalisée

Dans la pratique, un moyen simple d'appliquer le BIC consiste à sélectionner le meilleur sous-ensemble, en évaluant et comparant le BIC pour chaque modèle possible. Cependant, cette procédure peut aboutir à des calculs impossibles quand le nombre de covariables est grand. Pour la remplacer, des méthodes basées sur la vraisemblance pénalisée ont été utilisées récemment comme procédures de calcul efficaces pour appliquer un critère de sélection. Pour exclure des variables du modèle, ces méthodes estiment que les coefficients de ces variables sont nuls et réduisent les autres coefficients en conséquence. En faisant varier la pénalité appliquée à la vraisemblance, nous pouvons obtenir une série de modèles de parcimonie variable. Afin d'éviter une recherche exhaustive sur l'entièreté de l'espace des modèles, on utilise un critère de sélection pour choisir un modèle optimal parmi ces modèles parcimonieux. L'efficacité de cette stratégie a été illustrée dans un contexte ne faisant pas appel au sondage pour le critère BIC (Wang, Li et Tsai 2007; Liu, Wang et Liang 2011) et pour le critère GCV (Fan et Li 2001; Xie, Pan et Shen 2008) entre autres.

Dans le même esprit, nous proposons une procédure fondée sur la pseudo-vraisemblance pénalisée (PVP) pour appliquer le BIC (3.4) aux données d'enquête. En particulier, partant de la pseudo-vraisemblance (3.2) avec $k = n / N$, nous définissons l'estimateur pénalisé pondéré par les poids de sondage $\hat{\boldsymbol{\beta}}_\lambda$, qui minimise la fonction de pseudo-vraisemblance pénalisée.

$$Q_n(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}) - n\sum_{j=1}^{p} \phi_\lambda\left(|\beta_j|\right), \tag{3.5}$$

où $\phi_\lambda(.)$ est une fonction de pénalité indicée par un paramètre d'ajustement λ qui contrôle la taille de la pénalité. Moyennant un choix approprié de $\phi_\lambda(.)$, $\hat{\boldsymbol{\beta}}_\lambda$ contient des estimations nulles pour certains coefficients et produit donc automatiquement un modèle parcimonieux. La parcimonie souhaitable de $\hat{\boldsymbol{\beta}}_\lambda$ exige habituellement que la fonction $\phi_\lambda(.)$ correspondante soit singulière à l'origine. Certains choix fréquents de $\phi_\lambda(.)$ comprennent la pénalité L_γ (Frank et Friedman 1993; Tibshirani 1996), c'est-à-dire $\phi_\lambda\left(|\beta|\right) = \lambda|\beta|^\gamma$ avec $\gamma \in (0,1]$, et la pénalité SCAD (Fan et Li 2001), qui est définie par la dérivée suivante :

$$\phi'_\lambda\left(|\beta|\right) = \lambda\left\{I\left(|\beta| \le \lambda\right) + \frac{\left(a\lambda - |\beta|\right)_+}{(a-1)\lambda} I\left(|\beta| > \lambda\right)\right\} \tag{3.6}$$

pour laquelle $a = 3{,}7$ est un choix fréquent.

En utilisant des valeurs différentes de λ pour une fonction $\phi_\lambda(.)$ spécifiée correctement, $\hat{\boldsymbol{\beta}}_\lambda$ produit des modèles de parcimonie variable. Ces modèles parcimonieux (par rapport à λ) forment naturellement une série des modèles possibles. Le BIC (3.4) peut alors être utilisé pour choisir un modèle optimal parmi cette série. Plus précisément, soit Ω l'intervalle de valeurs de λ et soit s_λ un modèle produit par $\hat{\boldsymbol{\beta}}_\lambda$. Nous traitons $S_\Omega = \{s_\lambda : \lambda \in \Omega\}$ comme la série de modèles possibles prise en considération et nous choisissons le modèle $s^* \in S_\Omega$ tel que $\text{BIC}_n(s^*) = \min_{\lambda\in\Omega} \text{BIC}(s_\lambda)$. Nous appelons cette procédure de sélection la méthode du BIC fondée sur la pseudo-vraisemblance pénalisée (BIC-PVP). Comparativement à la sélection classique du meilleur sous-ensemble, la procédure BIC-PVP est axée sur les modèles qui sont produits par les estimateurs analysés pondérés par les poids de sondage et, par conséquent, peut demander nettement moins de calculs.

4 Convergence de la procédure BIC-PVP

Nous examinons maintenant le comportement asymptotique de la procédure BIC-PVP dans le cadre de randomisation conjointe. Nous supposons qu'il existe une série de populations finies, disons $\mathcal{D}_r$ avec $r \to \infty$. Chaque $\mathcal{D}_r$ est un échantillon indépendant et identiquement distribué (i.i.d.) de taille N_r tiré d'une superpopulation modélisée par (2.1) avec la variable aléatoire $\left(Y, \mathbf{X} = \{X_1, \ldots, X_p\}\right)$. Dans chaque population $\mathcal{D}_r$, on tire un échantillon d_r de taille n_r selon un certain plan d'échantillonnage. Nous supposons que N_r ainsi que n_r tendent vers l'infini quand $r \to \infty$, la fraction d'échantillonnage n_r / N_r étant bornée par une constante $C < 1$. Pour simplifier la notation, nous abandonnons l'indice r dans la suite de la discussion.

Sans perte de généralité, nous supposons que les q premiers coefficients ne sont pas nuls et nous désignons la valeur réelle de $\boldsymbol{\beta}$ par $\boldsymbol{\beta}_0 = \{\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}\}$ avec $\boldsymbol{\beta}_{02} = 0$. En outre, nous utilisons s_0 pour désigner le modèle réel $\{1, \ldots, q\}$ qu'il faut identifier. Nous établissons la convergence de sélection de la procédure BIC-PVP en deux étapes. À la première étape, nous montrons que, pour des choix appropriés de $\phi_\lambda(.)$, la PVP peut systématiquement identifier le vrai s_0 de sorte que $s_0 \in S_\Omega$ avec la probabilité tendant vers 1. À la deuxième étape, nous vérifions que le BIC (3.4) sélectionne systématiquement s_0 parmi S_Ω.

Pour l'analyse asymptotique, nous définissons $\varphi_\lambda = \max\left\{\phi'_\lambda\left(|\beta_{0j}|\right) \text{ pour } j \in s_0\right\}$ et associons λ à n pour faire de φ_λ une séquence. Sous le cadre de randomisation conjointe, nous montrons l'allégation de l'étape 1 sous la forme du théorème suivant.

Théorème 1 Sous des conditions de régularité appliquées au modèle (2.1) et d'autres exigences spécifiées dans le supplément en ligne, si $\varphi_\lambda \to 0$ quand $n \to \infty$, il existe alors un maximiseur local $\hat{\boldsymbol{\beta}}_\lambda = \left(\hat{\boldsymbol{\beta}}_{\lambda 1}, \hat{\boldsymbol{\beta}}_{\lambda 2}\right)$ de la fonction de pseudo-vraisemblance pénalisée (3.5) tel que

$$\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0\right\| = O_p\left(n^{-1/2} + \varphi_\lambda\right) \quad \text{et} \quad P\left\{\hat{\boldsymbol{\beta}}_{\lambda 2} = 0\right\} \to 1$$

avec $\|.\|$ désignant la norme euclidienne.

Le résultat de convergence du théorème 1 est vérifié pour les fonctions de pénalité non convexes fréquemment utilisées. Par exemple, pour la pénalité L_γ avec $\gamma \in (0,1)$, la convergence est vérifiée si $\lambda \to 0$; pour la pénalité SCAD, la convergence est vérifiée si $\lambda \to 0$ et $\sqrt{n}\lambda \to \infty$. Cela implique aussi que, si la probabilité tend vers 1, le modèle réel s_0 est inclus dans S_Ω, ce qui sert de condition préalable pour la convergence de sélection du BIC sur S_Ω.

Nous établissons maintenant la convergence en utilisant le BIC sur S_Ω avec une fonction $\phi_\lambda(.)$ spécifiée qui satisfait le théorème 1. En se servant de la notation de la section 3.2, soit s_λ le modèle correspondant à l'estimateur PVP $\hat{\boldsymbol{\beta}}_\lambda$, et soit Ω l'intervalle de valeurs de λ pris en considération. Nous définissons deux séries de modèles possibles comme il suit :

- modèles surajustés : $S_+ = \{s : s_0 \subset s, s \neq s_0\}$;
- modèles sous-ajustés : $S_- = \{s : s_0 \not\subset s\}$.

La notation $\not\subset$ indique ici qu'il existe au moins un élément différent entre deux ensembles, de sorte que S_- est la série de modèles possibles qui ne comprend pas toutes les variables figurant dans le modèle réel. Alors, Ω peut être partitionné en conséquence en

$$\Omega_+ = \{\lambda : s_\lambda \in S_+\}, \quad \Omega_- = \{\lambda : s_\lambda \in S_-\}, \quad \Omega_0 = \{\lambda : s_\lambda = s_0\}. \tag{4.1}$$

Au moyen du théorème 1, nous avons montré que $P(\Omega_0 \neq \varnothing) \to 1$. Par conséquent, la convergence de sélection du BIC sur S_Ω est atteinte si le BIC est capable d'identifier s_0 pour tout modèle s_λ avec $\lambda \in \Omega_+ \cup \Omega_-$. Nous utilisons le théorème qui suit pour établir ce résultat de convergence.

Théorème 2 Sous les mêmes conditions qu'au théorème 1,

$$P\left\{\min_{\lambda \in \Omega_+ \cup \Omega_-} \mathrm{BIC}_n(s_\lambda) \leq \mathrm{BIC}_n(s_0)\right\} \to 0,$$

où Ω_+ et Ω_- sont définis dans (4.1).

5 Études numériques

Afin d'évaluer les résultats de la procédure BIC-PVP sur échantillon fini, nous avons procédé à des études numériques approfondies en nous servant de données tirées de l'Enquête sur les personnes ayant une maladie chronique au Canada (EPMCC, Statistique Canada 2009). En particulier, nous comparons la procédure proposée aux méthodes classiques ne faisant pas appel au sondage en nous basant sur des modèles de régression postulés entre les variables de l'EPMCC et des réponses hypothétiques (simulées).

Nous livrons provisoirement certaines données sur l'utilisation de la sélection fondée sur la pseudo-vraisemblance sous deux scénarios de simulation. Dans le premier scénario, les populations sont générées à partir de modèles présumés et les échantillons sont obtenus en se servant de plans de sondage susceptibles de créer de fausses corrélations entre les variables de l'EPMCC. Dans le deuxième scénario, les populations ne sont pas générées exactement à partir des modèles présumés et les échantillons sont obtenus au moyen d'un plan de sondage en rapport avec la réponse ainsi que les covariables possibles. En outre, nous présentons l'analyse des données originales de l'EPMCC de 2009 à titre d'exemple d'utilisation de la procédure BIC-PVP dans des applications réelles.

5.1 Données de l'EPMCC

L'EPMCC est une étude transversale parrainée par l'Agence de la santé publique du Canada conçue pour recueillir des renseignements concernant les expériences des Canadiens atteints de maladies chroniques. L'un des principaux objectifs de l'EPMCC est de déterminer les comportements influant sur la santé qui ont une incidence sur les résultats de la maladie, afin que le gouvernement puisse planifier des services de santé pour les personnes atteintes de maladie chronique et leur fournir ces services.

L'EPMCC est réalisée tous les deux ans et chaque cycle de l'enquête porte sur deux maladies chroniques. L'enquête de 2009 portait sur l'arthrite et l'hypertension. Nous nous limitons ici aux données sur l'hypertension. La population cible pour l'enquête sur l'hypertension correspond aux Canadiens de 20 ans et plus des dix provinces ayant reçu un diagnostic d'hypertension et vivant dans les logements privés. Pour faciliter le processus d'enquête, les unités d'échantillonnage de l'EPMCC de 2009 sont les personnes atteintes d'hypertension qui avaient participé à l'Enquête sur la santé dans les collectivités canadiennes (ESCC) de 2008. Pour les besoins de l'EPMCC, la population correspondant aux répondants de l'ESCC est d'abord stratifiée selon le sexe et quatre groupes d'âge, à savoir 20 à 44 ans, 45 à 64 ans, 65 à 74 ans et 75 ans et plus. Par conséquent, la population finie formée par les répondants de l'ESCC a été subdivisée en huit catégories d'âge (4 niveaux) selon le sexe (2 niveaux). L'EPMCC est réalisée selon un plan d'échantillonnage stratifié avec répartition proportionnelle à la taille de l'échantillon. Un échantillon global de 9 005 personnes a été sélectionné parmi les 17 437 répondants à l'ESCC, et 6 142 de ces personnes ont participé à l'EPMCC.

Nous avons cerné 40 variables en rapport avec l'hypertension en nous basant sur les données originales de l'EPMCC et, pour 7 de ces variables des données complètes existaient pour les 6 142 répondants. Pour les 33 autres variables, une certaine quantité de valeurs manquaient en raison des cas de non-réponse au questionnaire original (voir le tableau 5.5. en annexe pour la liste des variables et les taux de non-réponse correspondants). Il n'existe aucune raison systématique évidente expliquant la non-réponse totale. La variable pour laquelle il manque le plus de données est INCDRPR (revenu du ménage), le taux de non-réponse étant de 9,6 %, tandis que la quantité de données manquantes est relativement faible pour les autres variables. Afin de faciliter l'analyse, pour remplacer les données manquantes, nous avons utilisé des méthodes d'imputation simples décrites ci-après. Pour une variable catégorique, nous avons imputé une valeur aux cas de non-réponse en nous servant d'une valeur aléatoire provenant du jeu de réponses; pour une variable continue, nous avons imputé une valeur aux cas de non-réponse en nous servant de la valeur moyenne des réponses. Deux exceptions à ces méthodes d'imputation ont été faites pour les variables BMHX_02 et CNHX_05. La première sert de variable de réponse dans l'analyse des données de la seconde, tandis que la seconde présente des contraintes naturelles sur l'intervalle de ces valeurs. Nous

avons supprimé les 274 observations pour lesquelles des valeurs manquaient pour ces deux variables, ce qui donne un jeu de données de travail de base contenant 5 868 observations. La procédure d'imputation/suppression n'a aucun effet sur l'évaluation de la procédure BIC fondée sur la population simulée. Par contre, elle pourrait introduire un biais dans l'analyse des données réelles. Pourtant, étant donné le faible taux de données manquantes, et la plausibilité que les données manquent au hasard dans le cas particulier, il est peu probable que la conclusion soit gravement affectée.

Puisque l'EPMCC est une enquête de suivi à l'ESCC, les poids d'échantillonnage pour l'EPMCC ont été obtenus au départ d'après les poids des données de l'ESCC. Ils ont ensuite été corrigés pour s'assurer que les répondants à l'EPMCC soient représentatifs de la population cible. Par conséquent, les poids corrigés présentent une variation importante d'une unité échantillonnée à l'autre. Après mise à l'échelle au moyen de $k = n / N \approx 10^{-3}$, les poids corrigés varient entre 0,01 et 33,62 avec un intervalle interquartile de 0,76.

5.2 Scénario 1 : Corrélation faussée

Comme il est mentionné plus haut, sous des plans de sondage complexes, la structure de corrélation entre les variables reflétée par l'échantillon peut être faussée comparativement à la population. Dans le premier scénario de simulation, nous évaluons la méthode BIC proposée lorsque les données sont recueillies au moyen de plans de sondage susceptibles de créer des corrélations faussées entre les covariables possibles. En particulier, nous traitons les 40 variables cernées comme des covariables possibles pour une réponse hypothétique Y, et pour simplifier, nous les indiçons de X_1 à X_{40}. Nous considérons des réponses continues ainsi que des réponses binaires dans nos simulations. Pour les cas continus, nous générons les valeurs de Y selon les modèles

- Modèle 1 : $Y = 0,7X_6 + 0,7X_{10} + 0,6X_{18} - 0,6X_{22} + \varepsilon,$
- Modèle 2 : $Y = 0,7X_6 + 0,6X_{10} + 0,6X_{18} - 0,5X_{22} + 0,3X_{30} - 0,3X_{34} + \varepsilon,$

avec $\varepsilon \sim N(0,1)$. Pour les cas binaires, où $Y \in \{0,1\}$, nous générons les valeurs de Y selon les modèles logistiques

- Modèle 3 : $\text{logit}\left(\Pr\{Y = 1 | \mathbf{X}\}\right) = 0,7X_7 - 0,6X_8 + 0,5X_{26},$
- Modèle 4 : $\text{logit}\left(\Pr\{Y = 1 | \mathbf{X}\}\right) = 0,8X_7 - 0,7X_8 + 0,6X_{26} - 0,5X_{28} + 0,4X_{36}.$

Les modèles spécifiés comprennent l'un des identificateurs de strate de l'EPMCC (c'est-à-dire X_6 ou X_7) avec une structure emboîtée pour chaque contexte de modélisation.

La population finie utilisée dans la simulation a été créée comme il suit. Le jeu de données de travail de base de 5 868 répondants a été reproduit 10 fois proportionnellement aux valeurs entières arrondies des poids de sondage de l'EPMCC, ce qui a donné une population pseudo-finie ayant une taille de 55 950 avec information complète sur $X_1, \ldots, X_{40}$. Les valeurs de la réponse Y ont ensuite été générées en se basant sur les modèles 1 à 4 respectivement. Nous considérons le problème de sélection des variables comme consistant à déterminer le modèle postulé qui génère les valeurs de Y.

Nous étudions les résultats de la procédure proposée sous deux plans d'échantillonnage stratifiés. En particulier, nous créons quatre strates en nous basant sur les variables X_6 (âge, moins de 55 ans/55 ans et plus) et X_7 (sexe, Hommes/Femmes), ce qui donne le groupe (Femmes, moins de 55 ans) de taille 7 120, le groupe (Femmes, 55 ans et plus) de taille 19 199, le groupe (Hommes, moins de 55 ans) de taille 6 187 et le groupe (Hommes, 55 ans et plus) de taille 23 458. Sous le premier plan, on tire de chaque strate un échantillon aléatoire simple sans remise (EASSR) avec répartition égale des tailles d'échantillon. L'inférence est basée sur les quatre EASSR regroupés. Sous le deuxième plan, nous construisons en outre trois sous-groupes dans chaque strate en nous basant sur la somme de deux covariables binaires des deux modèles postulés. En particulier, nous construisons les sous-groupes d'après $X_{18} + X_{22}$ pour les données générées par les modèles 1 et 2, et nous construisons de la même façon les sous-groupes fondés sur $X_8 + X_{26}$ pour les données générées par les modèles 3 et 4. Puis, nous effectuons l'inférence en nous basant sur les EASSR tirés de chaque sous-groupe des quatre strates. La taille globale d'échantillon est répartie de manière égale au niveau de la strate avec une proportion de 2 pour 1 pour 2 pour les trois sous-groupes à l'intérieur d'une même strate. Un simple calcul Monte Carlo révèle que la corrélation d'échantillon entre X_{18} et X_{22} (pour les données provenant des modèles 1 et 2) peut être aussi élevé que 0,5, tandis que leur corrélation dans la population est à peine de l'ordre de 0,02. Nous observons un phénomène similaire pour les variables X_8 et X_{26} (des données provenant des modèles 3 et 4). Par conséquent, nous nous attendons à ce que la sélection des variables sous le deuxième plan d'échantillonnage soit plus difficile en raison de cette augmentation systématique. Dans les simulations, nous fixons la taille globale d'échantillon à $n = 500$ pour les modèles 1 et 2 et à $n = 1\,500$ pour les modèles 3 et 4. Un résumé des variables influant sur la réponse et des variables de plan ayant une incidence sur les probabilités d'échantillonnage figure en annexe (tableau A.2).

Nous avons exécuté la procédure de sélection BIC-PVP sur les échantillons probabilistes tirés de la population finie. En particulier, nous avons mis à l'échelle les poids de sondage comme il est mentionné dans (3.2) et nous avons choisi la pénalité SCAD pour la fonction de vraisemblance pénalisée (3.5). Nous avons résolu le maximiseur correspondant de (3.5) en nous servant de l'algorithme de seuillage itératif (She 2011). Aux fins de comparaison, nous utilisons également les critères AIC et GCV comme autres options pour le BIC (3.4) proposé. Partant de la discussion de la section 3, nous définissons les critères AIC et GCV fondés sur la pseudo-vraisemblance comme étant

$$\mathrm{AIC}_n(s) = -2l_n\left(\hat{\beta}_s\right) + 2\tau(s),$$

$$\mathrm{GVC}_n(s) = -\frac{1}{n}\frac{l_n\left(\hat{\beta}_s\right)}{\left(1 - \tau(s)/n\right)^2},$$

qui sont appliqués de la même façon via la procédure fondée sur la PVP. En outre, pour chaque spécification, nous répétons la procédure de sélection en ignorant tous les poids de sondage (en fixant leur valeur à l'unité). Les résultats de la sélection non pondérée correspondent aux inférences fondées purement sur le modèle tel que discuté à la section 2. En particulier, le BIC fondé sur la pseudo-vraisemblance se réduit au BIC classique (3.1) utilisé dans les situations ne faisant pas appel au sondage.

Dans les tableaux 5.1 et 5.2, nous résumons les résultats des simulations fondées sur 1 000 répétitions en ce qui concerne le taux de sélections positives (TSP), le taux de fausses découvertes (TFD), le taux de

sélections correctes (TSC) et la taille moyenne du modèle (TMM). En particulier, soit s_0 un modèle réel qui génère la population finie et s'_j le modèle sélectionné en se basant sur le j^{e} échantillon, $j = 1, \ldots,$ 1 000. Nous estimons les TSP, TFD, TSC et TMM comme il suit

$$\text{TSP} = \frac{\sum_{j=1}^{1\,000} \tau\left(s_0 \cap s'_j\right)}{1\,000\,\tau\left(s_0\right)}, \quad \text{TFD} = \frac{\sum_{j=1}^{1\,000} \tau\left(s'_j \,/\, s_0\right)}{1\,000\,\tau\left(s'_j\right)},$$

$$\text{TSC} = \frac{\sum_{j=1}^{1\,000} I\left(s'_j = s_0\right)}{1\,0}, \quad \text{TMM} = \frac{\sum_{j=1}^{1\,000} \tau\left(s'_j\right)}{1\,0},$$

où $\tau\left(s\right)$ désigne la taille du modèle s et $I\left(\cdot\right)$ est la fonction indicatrice. De plus, nous évaluons comme suit l'exactitude prédictive du modèle sélectionné. Pour chaque spécification, nous générons un échantillon de test de taille 200 par EASSR à partir de la même population finie que celle dont a été tiré l'échantillon d'apprentissage. Pour les modèles 1et 2, nous utilisons la somme des carrés des résidus (SCR) moyenne calculée sur les données de test comme mesure de la capacité prédictive du modèle sélectionné. Pour les modèles 3 et 4, nous calculons les taux de prédictions positives ainsi que négatives. Plus précisément, soit π^* une valeur repère spécifiée et $\hat{\pi}_i$, la probabilité de succès estimée du i^{e} échantillon de test, $i = 1, \ldots, 200$. Nous prédisons alors la i^{e} réponse y_i par $\hat{y}_i = 1$ si $\hat{\pi}_i > \pi^*$ et $\hat{y}_i = 0$ autrement. Les taux de prédictions correctes sont estimés par

$$\text{TPP} = \frac{\sum_{i \in \{i: y_i = 1\}}^{200} I\left(\hat{y}_i = 1\right)}{\sum_{i=1}^{200} I\left(y_i = 1\right)}, \quad \text{TPN} = \frac{\sum_{i \in \{i: y_i = 0\}}^{200} I\left(\hat{y}_i = 0\right)}{\sum_{i=1}^{200} I\left(y_i = 0\right)}.$$

Nous prenons ensuite la moyenne des TPP et TPN finaux sur 1 000 répliques. Notons qu'ici, le TPP et le TPN sont semblables à la sensibilité et à la spécificité dans les études cliniques, qui indiquent la capacité d'une approche de prédiction 0-1 en ce qui concerne les prédictions positives et négatives correctes. En général, une grande valeur de π^* donne lieu à un TPN élevé mais à un TPP faible. La valeur de π^* doit être spécifiée prudemment dans les applications. Dans nos études en simulation, nous fixons la valeur à $\pi^* = 0{,}5$ pour simplifier.

Les résultats sont encourageants pour la méthode BIC proposée. D'après les tableaux 5.1 et 5.2, nous constatons que pour les modèles sélectionnés en appliquant le critère AIC,le TSP et le TFD sont tous deux élevés, ce qui indique l'inclusion d'un nombre excessif de variables non pertinentes. Comparativement, le BIC réduit significativement le TFD des modèles sélectionnés en sacrifiant légèrement le TSP, et donne lieu à la sélection du modèle dont les tailles sont plus proches de la réalité. Bien que le critère GCV se comporte de la même façon que le BIC sous le modèle linéaire, il donne des résultats concordant avec ceux de l'AIC pour les modèles logistiques pour lesquels les réponses binaires fournissent moins d'information.

Sous le premier plan d'échantillonnage, les probabilités d'inclusion ne sont reliées à Y qu'au moyen d'une seule covariable dans le modèle (c'est-à-dire X_6 ou X_7). La structure de corrélation entre la

réponse et les covariables de la population finie est maintenue en grande partie dans l'échantillon. Par conséquent, on n'observe aucune différence importante entre les procédures de sélection pondérées et non pondérées dans le tableau 5.1.

Nous livrons provisoirement les informations concernant l'utilisation des poids de sondage découlant de l'application du deuxième plan de sondage, où la structure de corrélation dans l'échantillon est systématiquement faussée. Clairement, la corrélation faussée entre les covariables pour les unités échantillonnées détériore l'efficacité des méthodes de sélection, comme en témoignent les valeurs plus faibles du TSP et plus élevées du TFD comparativement à celles obtenues pour les procédures non pondérées. L'intégration des poids de sondage dans le processus de sélection aide à corriger le biais résultant. En particulier, nous avons observé des améliorations appréciables pour la sélection fondée sur le BIC. Dans le cas le plus remarquable (c'est-à-dire modèle 3 du tableau 5.2), le BIC fondé sur la pseudo-vraisemblance donne des résultats considérablement meilleurs que ceux fournis par le BIC classique en faisant passer le TSP de 0,65 à 0,89, ce qui réduit le TFD correspondant qui passe de 0,62 à 0,50. Notre observation fait écho à la justification de la pondération voulant que celle-ci élimine le biais dû à l'échantillonnage informatif (section 6.3, Fuller 2009).

Tableau 5.1
Sélection pour le plan de sondage ne produisant pas de corrélation fortement faussée (premier plan). Les résultats sont résumés en fonction du taux de sélections positives (TSP), du taux de fausses découvertes (TFD), du taux de sélections correctes (TSC) et de la taille moyenne du modèle (TMM); les évaluations de la prédiction pour les modèles 1 et 2 sont fondées sur le test de la somme des carrés des résidus (SCR), et pour les modèles 3 et 4, sur le taux de prédictions positives/négatives (TPP, TPN) avec une valeur repère de 0,5.

Pondérations	Critère	TSP	TFD	TSC	TMM	Prédiction
			Modèle 1			
Ignorée	GCV	0,96	0,19	0,28	4,9	1,04
	AIC	0,99	0,48	0,05	8,7	1,08
	BIC	0,96	0,19	0,28	4,9	1,04
Incluse	GCV	0,95	0,24	0,19	5,2	1,05
	AIC	0,99	0,61	0,01	11,4	1,11
	BIC	0,95	0,24	0,20	5,3	1,05
			Modèle 2			
Ignorée	GCV	0,72	0,19	0,02	5,5	1,07
	AIC	0,89	0,44	0,01	10,3	1,09
	BIC	0,73	0,19	0,03	5,6	1,07
Incluse	GCV	0,74	0,24	0,02	6,1	1,08
	AIC	0,89	0,54	0,01	12,5	1,12
	BIC	0,74	0,24	0,03	6,1	1,08
			Modèle 3			
Ignorée	GCV	0,99	0,59	0,00	7,8	(0,71; 0,45)
	AIC	0,99	0,62	0,00	8,4	(0,69; 0,49)
	BIC	0,96	0,43	0,00	5,1	(0,72; 0,44)
Incluse	GCV	0,99	0,67	0,00	9,9	(0,71; 0,47)
	AIC	0,99	0,70	0,00	10,7	(0,68; 0,48)
	BIC	0,94	0,45	0,00	5,3	(0,71; 0,45)
			Modèle 4			
Ignorée	GCV	0,97	0,44	0,01	9,4	(0,66; 0,55)
	AIC	0,98	0,47	0,01	9,8	(0,65; 0,56)
	BIC	0,87	0,26	0,07	6,0	(0,69; 0,53)
Incluse	GCV	0,98	0,54	0,01	11,4	(0,66; 0,54)
	AIC	0,98	0,56	0,00	11,9	(0,66; 0,55)
	BIC	0,86	0,30	0,05	6,2	(0,68; 0,53)

Tableau 5.2
Sélection pour le plan générant des corrélations fortement faussées (2^e plan). Les résultats sont résumés en fonction du taux de sélections positives (TSP), du taux de fausses découvertes (TFD), du taux de sélections correctes (TSC) et de la taille moyenne du modèle (TMM); les évaluations de la prédiction pour les modèles 1 et 2 sont fondées sur le test de la somme des carrés des résidus (SCR), et pour les modèles 3 et 4, sur le taux de prédictions positives/négatives (TPP, TPN) avec une valeur repère de 0,5.

Pondérations	Critère	TSP	TFD	TSC	TMM	Prédiction
			Modèle 1			
Ignorée	GCV	0,83	0,23	0,17	4,6	1,09
	AIC	0,97	0,49	0,04	8,6	1,10
	BIC	0,83	0,23	0,17	4,6	1,09
Incluse	GCV	0,95	0,31	0,13	5,9	1,07
	AIC	0,99	0,65	0,00	12,5	1,12
	BIC	0,95	0,30	0,14	5,9	1,07
			Modèle 2			
Ignorée	GCV	0,62	0,22	0,02	5,0	1,13
	AIC	0,88	0,45	0,01	10,3	1,14
	BIC	0,62	0,22	0,02	5,1	1,12
Incluse	GCV	0,72	0,28	0,01	6,5	1,10
	AIC	0,89	0,59	0,00	13,7	1,12
	BIC	0,72	0,27	0,01	6,5	1,10
			Modèle 3			
Ignorée	GCV	0,87	0,62	0,00	7,3	(0,66; 0,44)
	AIC	0,88	0,63	0,00	7,6	(0,65; 0,45)
	BIC	0,65	0,62	0,00	4,5	(0,68; 0,42)
Incluse	GCV	0,97	0,74	0,00	11,9	(0,70; 0,46)
	AIC	0,97	0,75	0,00	12,4	(0,68; 0,46)
	BIC	0,89	0,50	0,00	5,6	(0,70; 0,44)
			Modèle 4			
Ignorée	GCV	0,94	0,48	0,00	9,5	(0,62; 0,51)
	AIC	0,95	0,50	0,00	10,0	(0,62; 0,52)
	BIC	0,72	0,41	0,00	6,1	(0,64; 0,49)
Incluse	GCV	0,93	0,61	0,00	12,5	(0,64; 0,53)
	AIC	0,94	0,62	0,00	12,9	(0,64; 0,53)
	BIC	0,82	0,34	0,01	6,4	(0,67; 0,54)

5.3 Scénario 2 : Spécification incorrecte du modèle

Une raison bien connue de l'utilisation des poids de sondage est qu'elle protège contre la spécification incorrecte du modèle (Pfeffermann et Holmes 1985; Kott 1991) : les inférences fondées sur les estimations pondérées peuvent demeurer valides pour la population sondée, même si le modèle n'est pas correct. Afin de mieux comprendre le rôle de la pondération dans la sélection des variables, nous comparons la méthode fondée sur le critère BIC proposé aux méthodes non pondérées classiques dans la simulation où le modèle supposé est mal spécifié à partir du modèle qui génère les données. Dans de telles situations, il n'existe pas de modèle « vrai » postulé et l'objectif de la sélection des variables est de trouver un modèle optimal qui décrit bien la population finie. Nous utilisons la population pseudo-finie stratifiée de la section 5.2, mais générerons la variable réponse Y en fonction des strates. Plus précisément, nous avons généré les valeurs de Y pour les unités dans les strates (Hommes, 55 ans et plus) et (Femmes, 55 ans et plus) par

$$Y = 0,6X_6 + 0,4X_{18} + 0,4X_{20} + 0,6X_{38} + \varepsilon,$$

et les valeurs de Y pour les unités dans les strates (Hommes, moins de 55 ans) et (Femmes, moins de 55 ans) par

$$Y = 0,6X_6 + 0,4X_{18} + 0,4X_{20} + \varepsilon$$

avec $\varepsilon \sim N(0,1)$ désignant une erreur aléatoire. Autrement dit, nous supposons que la variable X_{38} est influente seulement pour les personnes de 55 ans et plus, mais non pour les personnes de moins de 55 ans. En outre, nous violons aussi le modèle 1 présumé en excluant X_6 de l'ensemble de covariables possibles, ce qui limite la situation où une caractéristique importante du plan de sondage est omise dans la modélisation. Nous tirons un échantillon EASSR stratifié de taille 500 ou 1 000 en utilisant le premier plan de sondage de la section 5.2. Puis nous testons les procédures pondérées et non pondérées de sélection de variables en nous basant sur les unités échantillonnées.

Nous résumons les résultats de la simulation au tableau 5.3 en estimant les taux de sélection de X_{18}, X_{20} et X_{38} sur la base de 1 000 répliques. Comme pour les simulations précédentes, nous incluons aussi la taille moyenne du modèle (TMM) et la SCR de test des modèles sélectionnés (c'est-à-dire la SCR moyenne fondée sur les données d'un échantillon de test de taille 200) dans le résumé. Le tableau 5.3 nous permet de constater que, si les hypothèses de modélisation sont violées, le BIC fondé sur la pseudo-vraisemblance produit encore une exactitude de prédiction assez élevée en proposant des variables pertinentes avec une forte probabilité. En revanche, le fait d'ignorer les poids de sondage entraîne une perte relative de près de 9 % sur la SCR de test à cause de l'exclusion de X_{38}. Apparemment, l'accroissement de la taille de l'échantillon contribue à l'amélioration de la qualité de l'ajustement des modèles mal spécifiés, mais l'amélioration est obtenue au prix de l'inclusion d'un plus grand nombre de variables.

Tableau 5.3
Fréquence de sélection des variables influentes dans le cas d'un modèle mal spécifié; la taille moyenne du modèle (TMM) et la somme des carrés des résidus (SCR) de test sont également présentées.

Pondération	Critère	X_{18}	X_{20}	X_{38}	TMM	CR de test
			$n = 500$			
Ignorée	GCV	0,78	0,95	0,56	5,9	1,93
	AIC	0,95	0,99	0,73	12,5	1,95
	BIC	0,83	0,97	0,60	6,6	1,93
Incluse	GCV	0,73	0,92	0,84	6,3	1,77
	AIC	0,91	0,99	0,85	12,5	1,79
	BIC	0,78	0,94	0,83	6,9	1,77
			$n = 1\,000$			
Ignorée	GCV	0,96	1,00	0,79	7,6	1,87
	AIC	0,99	1,00	0,87	13,1	1,88
	BIC	0,97	1,00	0,80	7,9	1,87
Incluse	GCV	0,93	1,00	0,94	7,6	1,71
	AIC	0,98	1,00	0,96	13,0	1,72
	BIC	0,94	1,00	0,94	7,7	1,71

5.4 Analyse des données de l'EPMCC

Afin d'illustrer l'application du critère BIC proposé, nous l'utilisons pour déterminer les comportements influant sur la santé qui ont une incidence sur le contrôle de la pression artérielle en utilisant les données de l'EPMCC de 2009. La variable réponse est BMHX_02 provenant du jeu de données de travail obtenu à partir de l'EPMCC, qui comporte deux niveaux indiquant si la pression artérielle du répondant est ou non sous contrôle, selon la dernière mesure faite par un professionnel de la santé. Nous traitons les 39 autres variables du jeu de données de travail comme des covariables possibles et notre objectif est de repérer les covariables influentes qui sont associées au contrôle de la pression artérielle. Nous construisons une régression logistique de BMHX_02 sur les covariables possibles et utilisons la procédure BIC-PVP avec la pénalité SCAD pour sélectionner les covariables influentes (les poids sont rééchelonnés par le facteur $k = 10^{-3}$). En guise d'étape préliminaire, chaque covariable est normalisée de manière à ce que les premier et deuxième moments correspondants dans l'échantillon pondéré soient égaux à 0 et à l'unité, respectivement. À titre de comparaison, les critères AIC et GCV sont également utilisés dans l'analyse.

Dans la figure 5.1, nous représentons les scores du critère en fonction du degré de parcimonie du modèle. Nous voyons que le BIC sélectionne un modèle contenant 11 covariables, tandis que les critères GCV et AIC sélectionnent le modèle contenant 24 covariables. Si l'on ignore les poids de sondage dans la procédure de sélection, des modèles avec 7 ou 21 covariables sont suggérés par le critère BIC standard ou par les critères GCV ou AIC. La distinction entre les résultats des sélections pondérées et non pondérées reflète le faussement possible de la structure de corrélation des unités échantillonnées. Ce genre de distinctions peut également s'expliquer par la spécification incorrecte du modèle pour une partie de la population de l'EPMCC (Lohr et Liu 1994). Étant donné le biais possible des méthodes non pondérées, les résultats de la sélection pondérée sont plus plausibles dans l'analyse.

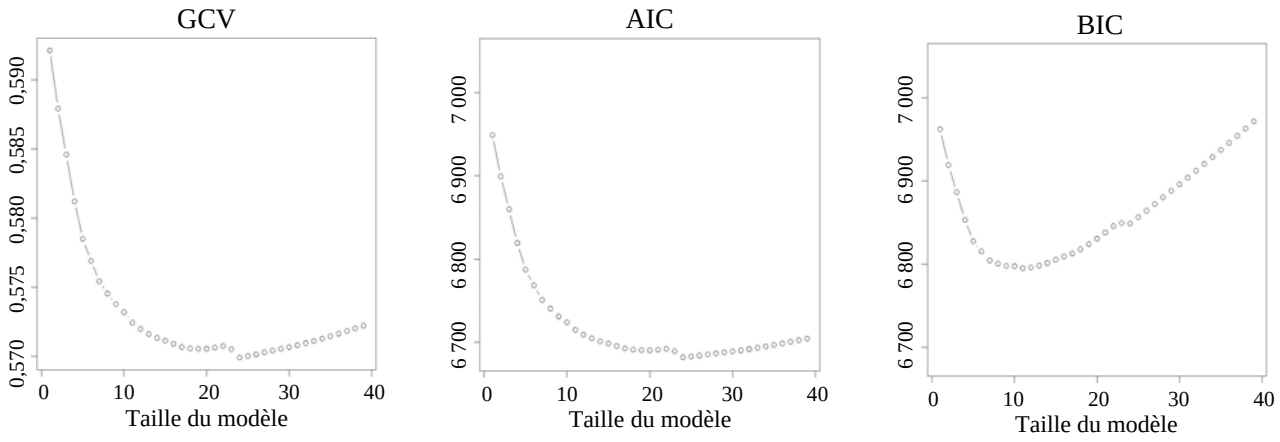


Figure 5.1 Valeurs des critères de sélection fondées sur les modèles possibles

Nous évaluons aussi les modèles sélectionnés en ce qui a trait à l'exactitude de la prédiction comme il suit. Premièrement, nous tirons 500 jeux indépendants de 5 868 échantillons bootstrap (avec remise) du

jeu de données de travail de l'EPMCC. Pour le t^{e} échantillon bootstrap d_t, $t = 1, \ldots, 500$, le poids de sondage w_i pour la i^{e} unité est ajusté selon $\tilde{w}_{ti} = v_{ti} w_i$ avec v_{ti} désignant le nombre de fois que la i^{e} unité est sélectionnée dans d_t. Puis, nous ajustons les modèles sélectionnés à chaque échantillon bootstrap (en tenant compte des poids en conséquence) et nous évaluons les taux pondérés de prédictions positives et de prédictions négatives (TPPP, TPPN) par

$$\text{TPPP} = \frac{\sum_{i \notin d_t} w_i I\left(\hat{y}_i = 1, y_i = 1\right)}{\sum_{i \notin d_t} w_i I\left(y_i = 1\right)}, \quad \text{TPPN} = \frac{\sum_{i \notin d_t} w_i I\left(\hat{y}_i = 0, y_i = 0\right)}{\sum_{i \notin d_t} w_i I\left(y_i = 0\right)},$$

où y_i et $\hat{y}_i$ désigne la i^{e} réponse dans BMHX_02 et sa valeur prédite. Nous résumons les TPPP et TPPN moyens calculés sur 500 échantillons bootstrap au tableau 5.4 pour trois valeurs repères différentes (c'est-à-dire 0,25, 0,35, 0,45).

D'après le tableau 5.4, nous constatons que les modèles sélectionnés d'après l'analyse non pondérée ont généralement un TPPP plus faible, ce qui offre un argument supplémentaire en faveur de l'utilisation des poids de sondage dans la procédure de sélection. Comparativement aux critères GCV/AIC, le critère BIC choisit le modèle ayant un TPPP légèrement plus prudent, mais un TPPN plus élevé. Néanmoins, l'écart n'est pas significatif. La taille du modèle sélectionné en appliquant le BIC est appréciablement plus faible que celle du modèle sélectionné par les critères GCV/AIC, ce qui permet d'interpréter plus facilement la relation entre la réponse BMHX_02 et les covariables.

Tableau 5.4
Exactitude de prédiction des modèles sélectionnés : (TPPP, TPPN) fondés sur différentes valeurs repères

Pondérations	Critère	≥ 0,25	≥ 0,35	≥ 0,45
Ignorée	AIC/GCV	(0,646; 0,525)	(0,460; 0,688)	(0,299; 0,811)
	BIC	(0,649; 0,513)	(0,445; 0,705)	(0,265; 0,818)
Incluse	AIC/GCV	(0,645; 0,523)	(0,488; 0,682)	(0,338; 0,790)
	BIC	(0,654; 0,532)	(0,485; 0,706)	(0,322; 0,830)

Pour évaluer la stabilité de la sélection, nous répétons la procédure de sélection pondérée fondée sur les 500 échantillons bootstrap. Au tableau 5.5, nous donnons le taux de sélection bootstrap pour les sept covariables les plus significatives en fonction de leur EMV dans le jeu de données de travail original de l'EPMCC. Les estimations des coefficients et les erreurs-types correspondantes fondées sur les échantillons bootstrap sont également incluses. D'après le tableau 5.5, nous constatons que quatre variables significatives seulement (c'est-à-dire DHHX_AGE, GENXDHMH, INHX_06, HWTDBMI) sont systématiquement sélectionnées en appliquant le critère BIC, tandis que les critères GCV/AIC ont tendance à sélectionner des variables moins fiables dans le modèle. Les résultats de la sélection fondés sur le critère BIC donnent à penser que le contrôle de la pression artérielle est fortement associé à l'âge, au poids corporel, à la santé mentale et à l'information concernant les médicaments. Nos observations

correspondent à celles de nombreuses études sur l'hypertension publiées (voir, par exemple, Gelber, Gaziano, Manson, Buring et Sesso 2007; Yan, Liu, Matthews, Daviglus, Ferguson et Kiefe 2003).

Tableau 5.5
Résultats de sélection bootstrap pour les variables significatives : (coefficient estimé, erreur-type, taux de sélections)

Variable	GCV	AIC	BIC
GEO_ON	(0,14; 0,09; 0,86)	(0,16; 0,09; 0,92)	(0,09; 0,09; 0,58)
DHHX_AGE	(-0,29; 0,09; 1,0)	(-0,32; 0,09; 1,0)	(-0,27; 0,08; 1,0)
GENXDHMH	(-0,15; 0,05; 0,99)	(-0,15; 0,05; 0,99)	(-0,14; 0,06; 0,92)
SMHXDSLT	(0,11; 0,07; 0,76)	(0,12; 0,07; 0,84)	(0,08; 0,09; 0,47)
MOHXDBPM	(-0,08; 0,07; 0,67)	(-0,09; 0,06; 0,81)	(-0,05; 0,07; 0,35)
INHX_06	(0,18; 0,06; 0,97)	(0,18; 0,06; 0,99)	(0,18; 0,07; 0,91)
HWTDBMI	(0,14; 0,06; 0,95)	(0,14; 0,06; 0,97)	(0,13; 0,06; 0,91)
Taille moyenne du modèle	23,1	27,8	10,3

6 Conclusion

Dans le présent article, nous avons abordé le problème de la sélection des variables dans l'analyse de données d'enquêtes complexes. Lorsque les unités sont sélectionnées selon un plan d'échantillonnage non proportionnel, la structure de corrélation des données reflétée par l'échantillon peut être faussée. L'intégration des poids de sondage dans le processus de sélection protège contre l'obtention de résultats de sélection biaisés. Dans cet esprit, nous avons dérivé un critère BIC pondéré par les poids de sondage fondé sur la pseudo-vraisemblance et proposé en outre une procédure efficace (pseudo-vraisemblance pénalisée) pour son application. Sous certaines conditions de régularité, nous avons montré que notre critère repère systématiquement les variables influentes sous un cadre de randomisation conjoint modèle-plan. Les résultats acceptables de la méthode proposée ont été confirmés par des études numériques.

Remerciements

Les auteurs remercient le rédacteur associé et les deux examinateurs anonymes de leurs commentaires judicieux et de leurs suggestions utiles. Les auteurs remercient le professeur J.N.K. Rao de l'Université Carleton de ses commentaires constructifs concernant un manuscrit antérieur. Les présents travaux ont été financés par Statistique Canada et par le MITACS.

Annexe

Tableau A.1
Variables pour l'analyse des données de l'EPMCC avec ajustement de la non-réponse : A : affectée à d'autres catégories; S : supprimée des données; M : imputée par les valeurs moyennes; NA : non ajustée pour la non-réponse

	Variable	Description	Niveaux	Manquante	Ajustement
1	BMHX_02	État de contrôle de la pression artérielle	2	1,6 %	S
2	GEO_QB	Provinces groupées par région – Québec	2	- -	NA
3	GEO_ON	Provinces groupées par région – Ontario	2	- -	NA
4	GEO_BC	Provinces groupées par région – Colombie-Britannique	2	- -	NA
5	GEO_PR	Provinces groupées par région – Prairies	2	- -	NA
6	DHHX_AGE	Âge	Cont.	- -	NA
7	DHHX_SEX	Sexe	2	- -	NA
8	GENXDHMH	Santé mentale perçue	2	0,2 %	A
9	CNHX_05	Hypertension – âge au diagnostic	Cont.	2,7 %	S
10	MEHX_02	Nombre de médicaments pris	Cont.	0,3 %	M
11	MEHX_03	Nombre de fois par jour que les médicaments sont pris	Cont.	0,1 %	M
12	MEHXGMED	Nombre de médicaments pour l'hypertension	Cont.	2,0 %	M
13	MEHX_06	Nombre de fois par jour que des médicaments pour l'hypertension sont pris	Cont.	1,0 %	M
14	MEHXDMCO	Respect des prescriptions concernant la prise de médicaments – global	2	0,2 %	A
15	HUHXDHP	A consulté un médecin de famille au sujet de l'hypertension	2	0,1 %	A
16	SMHX_11A	A fumé à n'importe quel moment depuis le diagnostic	2	0,1 %	A
17	SMHX_13A	A bu de l'alcool depuis le diagnostic	2	0,2 %	A
18	SMHXDSLT	Apport quotidien en sel	2	0,2 %	A
19	SMHXDFDC	Aliments de régime	2	0,1 %	A
20	SMHXDPAC	Exercice/activité physique	2	0,1 %	A
21	SMHXDBW	Contrôle du poids corporel	2	0,2 %	A
22	MOHXDBPM	Autosurveillance de la pression artérielle	2	0,3 %	A
23	MOHX_02	Usage correct de l'appareil de mesure de la pression artérielle	2	0,5 %	A
24	INHX_01A	Information fournie par le médecin de famille	2	2,4 %	A
25	INHX_01F	Information fournie par un membre de la famille/ami	2	2,4 %	A
26	INHX_02A	Information tirée de livres, brochures, dépliants	2	1,5 %	A
27	INHX_02C	Information tirée de la notice figurant dans l'emballage du médicament	2	1,5 %	A
28	INHX_02G	Information provenant des médias	2	1,5 %	A
29	INHX_02H	Information provenant d'Internet	2	1,5 %	A
30	INHX_04	Information reçue – effet émotionnel de l'hypertension	2	0,8 %	A
31	INHX_06	Information reçue – usage correct des médicaments	2	0,6 %	A
32	INHX_07	Information reçue – information supplémentaire	2	0,9 %	A
33	CPGFGAM	Activités de jeux de hasard	2	0,5 %	A
34	DHHDECF	Type de logement	2	0,2 %	A
35	EDUDH04	Niveau d'études le plus élevé dans le ménage	2	3,4 %	A
36	FGVCTOT	Consommation quotidienne – fruits et légumes	2	5,2 %	A
37	GEODUR2	Régions urbaines et rurales	2	- -	NA
38	HWTDBMI	Indice de masse corporelle (IMC), données autodéclarées	Cont.	2,1 %	M
39	INCDRPR	Revenu du ménage – niveau provincial	10	9,6 %	A
40	SACDTOT	Nombre total d'heures – activités sédentaires	Cont.	1,5 %	M

Tableau A.2
Variables influentes et variables du plan de sondage dans les simulations : ∗ - variable influant sur la réponse; • - variable du plan affectant les probabilités d'échantillonnage dans le premier plan; ◊ - variable du plan affectant les probabilités d'échantillonnage dans le deuxième plan.

	Variable	Modèle 1	Modèle 2	Modèle 3	Modèle 4
6	DHHX_AGE	* • ◊	* • ◊	• ◊	• ◊
7	DHHX_SEX	• ◊	• ◊	* • ◊	* • ◊
8	GENXDHMH			* ◊	* ◊
10	MEHX_02	*	*		
18	SMHXDSLT	* ◊	* ◊		
22	MOHXDBPM	* ◊	* ◊		
26	INHX_02A			* ◊	* ◊
28	INHX_02G				*
30	INHX_04		*		
34	DHHDECF		*		
36	FVCGTOT				*

Bibliographie

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Dans le *2nd International Symposium on Information Theory*, (Éds., B.N. Petrox et F. Caski), 267-281.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D., et Roberts, G. (2003). *Analysis of Survey Data*, Chapter: Design-based and model-based methods for estimating model parameters. Wiley Series in Survey Methodology, Chichester.

Craven, P., et Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31, 377-403.

Fan, J., et Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Frank, I.E., et Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109-148.

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

Gelber, R.P., Gaziano, J.M., Manson, J.E., Buring, J.E. et Sesso, H.D. (2007). A prospective study of body mass index and the risk of developing hypertension in men. *American Journal of Hypertension*, 20, 370-377.

Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Kalton, G. (1983). Models in the practice of survey sampling. *Revue Internationale de Statistique*, 51, 175-188.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys.* New York: John Wiley & Sons, Inc.

Kott, P.S. (1991). A model-based look at linear regression with survey data. *The American Statistician*, 45, 107-112.

Liu, X., Wang, L. et Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21, 1225-1248.

Lohr, S.L., et Liu, J. (1994). A comparison of weighted and unweighted analyses in the NCVS. *Journal of Quantitative Criminology*, 10, 343-360.

Mallows, C.L. (1973). Some Comments on C_p. *Technometrics*, 15, 661-675.

Molina, E.A., et Skinner, C.J. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics & Data Analysis*, 13, 395-405.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

Pfeffermann, D., et Holmes, D.J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, Séries A, 148, 268-278.

Rahiala, M., et Teräsvirta, T. (1993). Business survey data in forecasting the output of Swedish and Finnish metal and engineering industries: A Kalman filter approach. *Journal of Forecasting*, 12, 255-271.

Royall, M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.

She, Y. (2011). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*, in press.

Skinner, C. (2012). Weighting in the regression analysis of survey data with a cross-national application. *Canadian Journal of Statistics*, manuscript.

Statistics Canada (2009). Enquête sur les personnes ayant une maladie chronique au Canada – Guide de l'utilisateur 2009. Documentation supplémentaire.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (avec discussion). *Journal of the Royal Statistical Society,* Séries B, 39, 111-147.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Séries B, 58, 267-288.

Wang, H., Li, R. et Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568.

Wolfson, W.G. (2004). Analysis of labour force survey data for the information technology occupations 2000-2003. *Report for the Software Human Resource Council*, WGW Services Ltd., Ottawa, Ontario.

Xie, B., Pan, W. et Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, 64, 921-930.

Xu, C., et Chen, J. (2012). Technical supplement to "Pseudo-Likelihood-Based Bayesian Information Criterion for Variable Selection in Survey Data". Disponible auprès du premier auteur.

Yan, L.L., Liu, K., Matthews, K.A., Daviglus, M., Ferguson, T.F. et Kiefe, C.I. (2003). Psychosocial factors and risk of hypertension: The coronary artery risk development in young adults (CARDIA) study. *The Journal of the American Medical Association*, 290, 2138-2148.

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 355-383
Statistique Canada, N° 12-001-X au catalogue

Analyse fondée sur le plan de sondage de plans d'expérience factoriels intégrés dans des échantillons probabilistes

Jan A. van den Brakel[1]

Résumé

Les instituts nationaux de statistique intègrent fréquemment des expériences dans les enquêtes par sondage courantes, par exemple pour déterminer si des modifications du processus d'enquête ont un effet sur les estimations des principaux paramètres de cette dernière, pour quantifier l'effet de diverses mises en œuvre de l'enquête sur ces estimations, ou pour mieux comprendre les diverses sources d'erreur non due à l'échantillonnage. Le présent article propose une procédure d'analyse fondée sur le plan de sondage pour des plans factoriels complètement randomisés et des plans factoriels en blocs randomisés intégrés dans des échantillons probabilistes. Une statistique de Wald fondée sur le plan de sondage est élaborée pour vérifier si les paramètres de population, comme les moyennes, les totaux et les ratios de deux totaux de population, estimés sous les différentes combinaisons de traitements de l'expérience diffèrent de manière significative. Les méthodes sont illustrées au moyen d'une application réelle d'une expérience intégrée dans l'Enquête sur la population active des Pays-Bas.

Mots-clés : Plans complètement randomisés; inférence fondée sur le plan de sondage; expériences intégrées; modèles d'erreur de mesure; inférence assistée par un modèle; plans en blocs randomisés.

1 Introduction

Les études expérimentales randomisées et l'échantillonnage probabiliste forment habituellement deux domaines distincts de la statistique appliquée. Cependant, ces domaines se joignent lorsqu'on intègre des expériences dans des enquêtes par sondage courantes. Cette intégration d'expériences randomisées dans des enquêtes courantes est fréquente pour comparer et évaluer les effets de différents modes de mise en œuvre sur les résultats de l'enquête. Ce genre d'études empiriques a pour objectif d'améliorer la qualité et l'efficacité des processus d'enquête sous-jacents ou d'obtenir des renseignements plus quantitatifs sur les diverses sources d'erreurs non dues à l'échantillonnage. Bon nombre d'expériences faites dans ce contexte sont réalisées à petite échelle ou sur des groupes particuliers. Or, la recherche empirique sur les méthodes d'enquête possède plus de valeur si les conclusions sont généralisables à de plus grandes populations que l'échantillon inclus dans l'expérience. Comme l'ont souligné Fienberg et Tanur (1987, 1988, 1989 et 1996), la sélection aléatoire d'unités expérimentales dans une population cible de plus grande taille est un moyen important de s'assurer que les résultats d'une expérience puissent être généralisés à une plus grande population que le groupe de personnes inclus dans l'expérience. Cela mène naturellement à l'intégration d'expériences randomisées dans des enquêtes par sondage courantes. Dans la littérature sur les sondages, les expériences de ce genre, aussi appelées plans à échantillon fractionné (*split-ballot designs*), ou sous-échantillons « interpénétrants » (*interpenetrating subsampling*), remontent à Mahalanobis (1946).

Dans les instituts nationaux de statistique, ces expériences sont particulièrement utiles pour quantifier les ruptures dues à des ajustements du processus d'enquête dans les séries issues d'enquêtes répétées. La répétition d'une enquête produit une série qui décrit l'évolution des paramètres cibles. On peut recourir à

1. Jan A. van den Brakel, Département des méthodes statistiques, Statistics Netherlands, C.P. 4481, 6401 CZ Heerlen, Pays-Bas et Département d'économie quantitative, Maastricht University School of Business and Economics, C.P. 616, 6200 MD, Maastricht, Pays-Bas. Courriel : jbrl@cbs.nl.

des expériences intégrées pour éviter qu'une ou plusieurs modifications du processus d'enquête ne donnent lieu à des différences inexpliquées dans la série chronologique.

Une difficulté importante que pose l'analyse de ce genre d'expériences consiste à trouver le mode approprié d'inférence. Dans le contexte des sondages, l'inférence statistique est habituellement fondée sur le plan de sondage ou assitée par modèle. Elle s'appuie donc principalement sur la structure stochastique induite par le plan de sondage. Un estimateur fondé sur le plan de sondage bien connu est l'estimateur de Horvitz-Thompson (HT), élaboré par Narain (1951) et par Horvitz et Thompson (1952) pour l'échantillonnage avec probabilités inégales sans remise à partir de populations finies. Sous l'approche assistée par modèle élaborée par Särndal, Swensson et Wretman (1992), l'exactitude de l'estimateur HT est améliorée en tirant parti de l'information auxiliaire disponible au sujet de l'ensemble de la population cible, ce qui donne l'estimateur par la régression généralisée (GREG). De nombreux instituts nationaux de statistique font appel à cette approche fondée sur le plan et assistée par modèle pour produire les statistiques officielles.

L'inférence statistique sur laquelle s'appuie habituellement la théorie de la conception et de l'analyse d'expériences randomisées repose principalement sur un modèle. On suppose que les observations obtenues durant l'expérience sont des réalisations d'un modèle linéaire. Pour tester les hypothèses au sujet des effets du traitement, on dérive des tests F en supposant que les observations suivent des lois normales indépendantes. Fait exception Kempthorne (1955), qui propose une approche de randomisation semblable à l'approche de l'inférence fondée sur le plan en théorie de l'échantillonnage. Le test F est utilisé comme approximation du test sous randomisation. Dans le cas des expériences randomisées, l'inférence fondée sur un modèle ne convient pas nécessairement pour l'analyse des expériences intégrées, particulièrement si l'on recourt à l'inférence fondée sur le plan ou assistée par un modèle pour l'enquête courante afin de produire les statistiques officielles.

Dans une expérience intégrée, l'échantillon probabiliste de l'enquête courante est divisé aléatoirement en divers sous-échantillons conformément au plan d'expérience. Chaque sous-échantillon peut être considéré comme un échantillon probabiliste tiré de la population finie cible et être utilisé pour estimer les paramètres, comme les moyennes, les totaux et les ratios, observés sous les diverses mises en œuvre de l'enquête ou les divers traitements de l'expérience en utilisant la procédure d'estimation appliquée à l'enquête ordinaire pour produire les statistiques officielles. L'objectif de ce genre d'expériences intégrées est de comparer l'effet de diverses mises en œuvre de l'enquête sur les estimations des principaux paramètres de l'enquête courante et de vérifier si les différences observées entre ces estimations sont statistiquement significatives. On recourt pour cela à une approche fondée sur le plan de sondage où les estimations ponctuelles et les estimations de variance des paramètres de population sont (approximativement) sans biais par rapport au plan, en ce qui concerne le plan de sondage utilisé pour tirer un échantillon probabiliste initial de la population cible ainsi que le plan d'expérience utilisé pour répartir aléatoirement cet échantillon entre les divers sous-échantillons. L'analyse doit également refléter les détails particuliers de l'approche d'estimation ordinaire utilisée pour les statistiques officielles, dans la mesure où cela est possible compte tenu de la taille d'échantillon disponible pour les différents traitements.

Dans le cadre d'études antérieures, une théorie fondée sur le plan de sondage a été élaborée pour l'analyse d'expériences à un seul facteur conçues comme des plans complètement randomisés (PCR) ou des plans en blocs randomisés (PBR) pour évaluer l'effet d'un facteur sur $K \geq 2$ niveaux

(van den Brakel (2008), van den Brakel et Renssen (1998, 2005); van den Brakel et van Berkel (2002)). L'approche proposée consiste à appliquer l'estimateur GREG pour dériver la statistique de Wald et la statistique t fondées sur le plan pour vérifier si les écarts entre les estimations des paramètres de population finie observés sous les diverses mises en œuvre de l'enquête sont statistiquement significatifs. Cette théorie est en outre étendue aux expériences intégrées dans les plans de sondage à panel rotatif par Chipperfield et Bell (2010).

La théorie classique des plans expérimentaux montre clairement qu'il est efficace de tester divers facteurs de traitement simultanément dans un seul plan factoriel au lieu de réaliser des expériences à un seul facteur distinctes (Hinkelmann et Kempthorne (1994); Montgomery (2001)). On peut s'attendre à ce que, dans un processus d'enquête, les divers paramètres du plan de sondage interagissent, p. ex. si l'on compare empiriquement différentes conceptions de questionnaires et différents modes de collecte des données. L'utilisation de plans factoriels est indiquée si l'on teste l'ajustement de plus d'un facteur de l'enquête dans une expérience intégrée, car on peut ainsi évaluer les effets principaux des facteurs de traitement sur un moins grand nombre d'unités expérimentales, tout en étant capable d'analyser les interactions entre les facteurs. Un autre avantage de l'évaluation simultanée de différents traitements dans un plan factoriel est que la validité des résultats observés s'accroît, puisque les effets sont observés sous une plus grande gamme de conditions (Hinkelmann et Kempthorne 1994). Par conséquent, dans le présent article, nous étendons aux plans factoriels la théorie fondée sur le plan de sondage pour l'analyse des expériences intégrées.

À la section 2, nous élaborons la théorie pour les plans factoriels quand l'effet de deux facteurs est testé simultanément. Ensuite, à la section 3, nous étendons la méthodologie à des plans factoriels d'ordre plus élevé. À la section 4, nous étendons la méthodologie aux tests d'hypothèse au sujet des ratios des totaux de population et aux plans dans lesquels les grappes d'unités d'échantillonnage sont réparties aléatoirement entre les combinaisons de traitements. À la section 5, nous appliquons ces méthodes à une expérience factorielle avec diverses lettres d'introduction dans l'Enquête sur la population active (EPA) des Pays-Bas. L'article se termine par une discussion à la section 6.

2 Analyse d'expériences factorielle $K \times L$ intégrées

2.1 Plans d'expérience intégrés dans des échantillons probabilistes

Dans un plan factoriel $K \times L$, les effets des deux facteurs sont évalués simultanément. Le premier facteur, désigné A, contient $K \geq 2$ niveaux. Le deuxième facteur, désigné B, contient $L \geq 2$ niveaux. Le but de l'expérience est de tester les effets principaux des deux facteurs et des interactions entre les deux facteurs sur les estimations des principaux paramètres de l'enquête en cours. Pour cela, on tire un échantillon probabiliste s de taille n d'une population cible finie U de taille N selon le plan de sondage de l'enquête ordinaire. Le plan de sondage, qui peut être complexe, est décrit par les probabilités d'inclusion de premier ordre π_i de l'unité i et les probabilités d'inclusion de second ordre $\pi_{ii'}$ des unités i et i'.

Ensuite, cet échantillon est divisé aléatoirement en KL sous-échantillons selon un plan d'expérience randomisé. Dans le cas d'un plan complètement randomisé (PCR), l'échantillon s de taille n est divisé

aléatoirement en KL sous-échantillons s_{kl}, contenant chacun n_{kl} unités d'échantillonnage. Les unités d'échantillonnage de chaque sous-échantillon sont assignées à l'une des KL combinaisons de traitements. Sous un plan PCR, $n_{++} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl}$ désigne le nombre total d'unités d'échantillonnage dans l'échantillon s. La probabilité que l'unité d'échantillonnage i soit assignée au sous-échantillon s_{kl}, conditionnellement à la réalisation de s, est égale à n_{kl}/n_{++}. La probabilité non conditionnelle que l'unité d'échantillonnage i soit sélectionnée dans le sous-échantillon s_{kl} est égale à $\pi_i^* = \pi_i \left(n_{kl}/n_{++}\right)$.

La puissance d'une expérience peut être accrue en utilisant des structures d'échantillonnage telles que les strates, les grappes ou les intervieweurs comme variables de bloc dans un plan en blocs randomisés (PBR), puisque la randomisation restreinte élimine la variance inter-blocs de l'analyse de l'expérience (Fienberg et Tanur (1987, 1988)). Dans le cas d'un plan PBR, les unités d'échantillonnage sont groupées de manière déterministe en B blocs plus ou moins homogènes s_b. Dans chaque bloc, les unités d'échantillonnage sont assignées aléatoirement à une des KL combinaisons de traitements. Soit n_{bkl} le nombre d'unités d'échantillonnage dans le bloc b assignées à la combinaison de traitements kl, et $n_{b++} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{bkl}$, le nombre d'unités d'échantillonnage dans le bloc b. La probabilité que l'unité d'échantillonnage i soit assignée au sous-échantillon s_{kl}, conditionnellement à la réalisation de s et $i \in s_b$, est égale à $n_{bkl}/n_{b++}, i \in s_b$. La probabilité non conditionnelle que l'unité d'échantillonnage i soit sélectionnée dans le sous-échantillon s_{kl} est égale à $\pi_i^* = \pi_i \left(n_{bkl}/n_{b++}\right)$.

Dans de nombreuses applications pratiques, on assigne à l'enquête ordinaire l'un des KL sous-échantillons qui, en plus d'être utilisé pour produire les publications ordinaires, sert de groupe témoin dans l'expérience. Le cas échéant, la taille de ce sous-échantillon sera considérablement plus grande que celle des autres sous-échantillons.

De nombreuses questions doivent être résolues à l'étape de la planification et de la conception des expériences intégrées. Par exemple, une attention particulière doit être accordée au personnel sur le terrain, puisqu'une expérience intégrée peut avoir sur la routine quotidienne de collecte des données d'importantes répercussions auxquelles ils doivent s'habituer. Voir van den Brakel et Renssen (1998) et van den Brakel (2008) pour des renseignements plus détaillés sur ces problèmes de conception.

Même si les plans factoriels sont efficaces d'un point de vue statistique, il pourrait exister de puissants arguments à leur encontre. Dans les plans factoriels complets, le nombre de combinaisons de traitements augmente rapidement avec le nombre de facteurs, ce qui peut rendre la mise en œuvre difficile à l'étape de la collecte des données d'un processus d'enquête. Une solution générale, venant de la théorie classique des plans expérimentaux, consiste à faire coïncider les interactions d'ordre plus élevé avec des blocs ou à appliquer des plans factoriels fractionnaires (Hinkelmann et Kempthorne (2005); Montgomery (2001)). Toutefois ces plans équilibrés sont généralement difficiles à combiner avec les contraintes du travail sur le terrain qui se présentent dans la pratique quotidienne des sondages. Dans de nombreuses applications, les facteurs qui ont été modifiés durant le remaniement d'une enquête sont par conséquent combinés en un seul traitement. L'effet total de ces modifications est évalué en regard de la situation classique dans une expérience comprenant deux traitements. Cela signifie que les effets de tous les facteurs compris dans l'expérience sont confondus et ne peuvent pas être estimés séparément.

2.2 Tests d'hypothèse au sujet de paramètres de population finie

L'objectif des expériences intégrées est de tester si diverses mises en œuvre d'une enquête produisent des estimations significativement différentes des paramètres de population finie. Ces différences sont le résultat des erreurs non dues à l'échantillonnage, comme les erreurs de mesure et le biais de réponse. Un modèle d'erreur de mesure est nécessaire pour relier les différences systématiques entre les paramètres de population finie dues à des mises en œuvre d'enquête ou à des traitements différents. Par conséquent, nous étendons aux plans factoriels le modèle d'erreur de mesure proposé pour les expériences à un seul facteur par van den Brakel et Renssen (2005) et van den Brakel (2008).

Soit y_{iqkl} l'observation obtenue auprès du i^{e} individu observé sous la kl^{e} combinaison de traitements par le q^{e} intervieweur. Nous supposons que les observations sont une réalisation du modèle d'erreur de mesure

$$y_{iqkl} = u_i + \beta_{kl} + \gamma_q + \varepsilon_{ikl}. \tag{2.1}$$

Ici, u_i est la valeur intrinsèque réelle du i^{e} individu, β_{kl} est l'effet de la kl^{e} combinaison de traitements et ε_{ikl} est une composante d'erreur. Le modèle tient également compte des effets d'intervieweur, c'est-à-dire $\gamma_q = \psi + \xi_q$, où ψ désigne un biais systématique d'intervieweur et ξ_q, l'effet aléatoire du q^{e} intervieweur, respectivement. Soit E_m et cov_m l'espérance et la covariance sous le modèle d'erreur de mesure. Nous supposons que $\text{E}_m\left(\varepsilon_{ikl}\right) = 0$, $\text{var}_m\left(\varepsilon_{ikl}\right) = \sigma^2_{ikl}$, et que les erreurs de mesure entre les unités d'échantillonnage sont indépendantes. En outre, nous supposons que $\text{E}_m\left(\xi_q\right) = 0$, $\text{var}_m\left(\xi_q\right) = \tau^2_q$ et que les effets d'intervieweur aléatoire entre les intervieweurs sont indépendants. Par conséquent, le modèle tient compte de la corrélation des réponses entre les unités d'échantillonnage interviewées par le même intervieweur. Le modèle d'erreur de mesure permet de considérer des variances distinctes pour les erreurs de mesure sous diverses combinaisons de traitements et des variances distinctes pour les intervieweurs.

Les effets du traitement β_{kl} peuvent être interprétés comme étant le biais dans le paramètre de population estimé si la vraie valeur intrinsèque de population de u est mesurée au moyen de la kl^{e} mise en œuvre de l'enquête. Nous pouvons décomposer l'effet du traitement à la manière classique d'une analyse de variance à deux critères de classification :

$$\beta_{kl} = u + A_k + B_l + AB_{kl}, \tag{2.2}$$

où u est l'effet global, A_k et B_l sont les effets principaux des facteurs de traitement A et B, et AB_{kl} représente les interactions entre les facteurs de traitement A et B. Si les effets du traitement sont définis comme étant des écarts fixes par rapport à la valeur intrinsèque u_i des individus, alors la moyenne globale u est nulle. Dans ce cas, A_k correspond à la moyenne du biais associé au k^{e} niveau du facteur A calculée sur tous les niveaux du facteur B, B_l est la moyenne du biais associé au l^{e} niveau du facteur B calculée sur tous les niveaux du facteur A, et AB_{kl} le biais supplémentaire associé à la combinaison du k^{e} niveau de facteur A et du l^{e} niveau du facteur B qui s'ajoute à A_k et B_l.

Les contraintes qui suivent sont nécessaires pour spécifier le modèle (2.2) :

$$\sum_{k=1}^{K} A_k = 0, \ \sum_{l=1}^{L} B_l = 0, \tag{2.3}$$

et

$$\sum_{k=1}^{K} AB_{kl} = 0, l = 1, 2, \ldots, L, \sum_{l=1}^{L} AB_{kl} = 0, k = 1, 2, \ldots, K. \tag{2.4}$$

Pour chaque unité d'échantillonnage, nous définissons une variable réponse possible sous chacune des KL combinaisons de traitements. Par conséquent, le modèle d'erreur de mesure peut être exprimé en notation matricielle sous la forme :

$$\mathbf{y}_{iq} = \mathbf{j}_{KL} u_i + \boldsymbol{\beta} + \mathbf{j}_{KL} \gamma_q + \boldsymbol{\varepsilon}_i, \tag{2.5}$$

où $\mathbf{y}_{iq} = \left(y_{iq11}, \ldots, y_{iqkl}, \ldots, y_{iqKL}\right)^t$, $\boldsymbol{\beta} = \left(\beta_{11}, \ldots, \beta_{kl}, \ldots, \beta_{KL}\right)^t$, $\mathbf{j}_{KL}$ est un vecteur d'ordre KL avec chaque traitement égal à l'unité et $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i11}, \ldots, \varepsilon_{ikl}, \ldots, \varepsilon_{iKL}\right)^t$. Les unités d'échantillonnage ne sont assignées qu'à une seule combinaison de traitements, de sorte qu'une seule des réponses $\mathbf{y}_{iq}$ est effectivement observée. Les hypothèses qui sous-tendent le modèle spécifié plus haut sont énoncées commesuit :

$$\mathrm{E}_m\left(\boldsymbol{\varepsilon}_i\right) = \mathbf{0}, \tag{2.6}$$

$$\mathrm{cov}_m\left(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}\right) = \begin{cases} \boldsymbol{\Sigma}_i & : \ i = i' \\ \mathbf{O} & : \ i \neq i' \end{cases}, \tag{2.7}$$

$$\mathrm{E}_m\left(\xi_q\right) = 0, \tag{2.8}$$

$$\mathrm{cov}_m\left(\xi_q, \xi_{q'}\right) = \begin{cases} \tau_q^2 & : \ q = q' \\ 0 & : \ q \neq q' \end{cases}, \tag{2.9}$$

$$\mathrm{cov}_m\left(\varepsilon_{ikl}, \xi_q\right) = 0, \tag{2.10}$$

où $\mathbf{0}$ est un vecteur d'ordre KL dont chaque élément est nul, $\boldsymbol{\Sigma}_i$ est une matrice d'ordre $KL \times KL$ contenant les variances des erreurs de mesure σ_{ikl}^2, et $\mathbf{O}$ est une matrice d'ordre $KL \times KL$ dont chaque élément est nul.

Soit $\bar{\mathbf{Y}} = \left(\bar{Y}_{11}, \ldots, \bar{Y}_{1L}, \ldots, \bar{Y}_{kl}, \ldots, \bar{Y}_{K1}, \ldots, \bar{Y}_{KL}\right)^t$ le vecteur KL-dimensionnel des moyennes de population de $\mathbf{y}_{iq}$ défini par (2.5). Il s'agit de valeurs obtenues sous un dénombrement complet de la population finie sous chacune des combinaisons de traitements, et définies comme étant :

$$\bar{\mathbf{Y}} = \mathbf{j}_{KL} \frac{1}{N} \sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}_{KL} \psi + \mathbf{j}_{KL} \sum_{q=1}^{Q} \frac{N_q}{N} \xi_q + \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\varepsilon}_i, \tag{2.11}$$

où Q désigne le nombre total d'intervieweurs disponibles pour la collecte des données et N_q, le nombre d'unités assignées au q^{e} intervieweur dans le cas d'un dénombrement complet.

Seules les différences systématiques entre les paramètres de population qui sont reflétées par les effets du traitement $\boldsymbol{\beta}$ doivent mener au rejet de l'hypothèse nulle de l'absence d'effets du traitement. Cette condition est remplie en formulant les hypothèses au sujet de $\bar{\mathbf{Y}}$ en espérance sur le modèle d'erreur de mesure, c'est-à-dire

$$\mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{j}_{KL} \frac{1}{N} \sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}_{KL} \psi. \tag{2.12}$$

Par conséquent, l'hypothèse au sujet des effets principaux et les interactions sont formulées comme il suit

$$\begin{aligned} H_0&: \mathbf{C}\mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{0}, \\ H_1&: \mathbf{C}\mathrm{E}_m \bar{\mathbf{Y}} \neq \mathbf{0}, \end{aligned} \tag{2.13}$$

où $\mathbf{C}$ désigne une matrice de contrastes appropriée et $\mathbf{0}$, un vecteur dont les éléments sont égaux à un et dont la dimension est égale au nombre de contrastes (lignes) définis par $\mathbf{C}$. La matrice de contrastes pour l'hypothèse au sujet des effets principaux du facteur A est définie par

$$\mathbf{C}_{\mathbf{A}} = \frac{1}{L}\left(\mathbf{j}_{(K-1)} \mid -\mathbf{I}_{(K-1)}\right) \otimes \mathbf{j}_L^t \equiv \frac{1}{L} \tilde{\mathbf{C}}_{\mathbf{A}} \otimes \mathbf{j}_L^t \tag{2.14}$$

avec $\mathbf{I}_{(K-1)}$ la matrice identité d'ordre $K-1$. La matrice $\tilde{\mathbf{C}}_{\mathbf{A}}$ définit les $K-1$ contrastes entre les K niveaux du facteur A, en prenant la moyenne sur les L niveaux du facteur B. De l'équation (2.12) et des contraintes (2.3) et (2.4) il découle que les contrastes entres les paramètres de population correspondent exactement aux contrastes entres les effets principaux du premier facteur :

$$\tilde{\mathbf{C}}_{\mathbf{A}} \mathrm{E}_m \bar{\mathbf{Y}} = \tilde{\mathbf{C}}_{\mathbf{A}} \boldsymbol{\beta} = (A_1 - A_2, ..., A_1 - A_K)^t.$$

La matrice des contrastes pour l'hypothèse au sujet des effets principaux des facteurs B est définie par

$$\mathbf{C}_{\mathbf{B}} = \frac{1}{K} \mathbf{j}_K^t \otimes \left(\mathbf{j}_{(L-1)} \mid -\mathbf{I}_{(L-1)}\right) \equiv \frac{1}{K} \mathbf{j}_K^t \otimes \tilde{\mathbf{C}}_{\mathbf{B}}. \tag{2.15}$$

Cette matrice définit les $L-1$ contrastes entre les L niveaux du facteur B, en prenant la moyenne sur les K niveaux du facteur A. De l'équation (2.12) et des contraintes (2.3) et (2.4), il découle que les contrastes entre les paramètres de population correspondent exactement aux contrastes entre les effets principaux du deuxième facteur :

$$\tilde{\mathbf{C}}_{\mathbf{B}} \mathrm{E}_m \bar{\mathbf{Y}} = \tilde{\mathbf{C}}_{\mathbf{B}} \boldsymbol{\beta} = (B_1 - B_2, ..., B_1 - B_L)^t.$$

Les matrices des contrastes pour les effets principaux utilisent le premier niveau des facteurs A et B comme catégorie de référence. Cela implique que la combinaison de traitements $A_1 \times B_1$ est considérée comme le groupe témoin dans l'expérience.

Les interactions entre les deux facteurs de traitement sont définies comme étant les $L-1$ contrastes du facteur B entre les $K-1$ contrastes du facteur A ou, de façon équivalente, comme les $K-1$ contrastes du facteur A entre les $L-1$ contrastes du facteur B (Hinkelmann et Kempthorne, 1994,

chapitre 11). Par conséquent, la matrice des contrastes pour l'hypothèse au sujet des interactions entre les facteurs A et B peut être définie comme

$$\mathbf{C}_{\mathbf{AB}} = \left(\mathbf{j}_{(K-1)}\Big| - \mathbf{I}_{(K-1)}\right) \otimes \left(\mathbf{j}_{(L-1)}\Big| - \mathbf{I}_{(L-1)}\right) = \tilde{\mathbf{C}}_{\mathbf{A}} \otimes \tilde{\mathbf{C}}_{\mathbf{B}}. \tag{2.16}$$

Cette matrice contient les $(K-1)(L-1)$ contrastes qui définissent les interactions entre les facteurs A et B. Les contrastes entre les paramètres de population correspondent exactement aux interactions entre le premier et le deuxième facteur, puisque

$$\begin{aligned}\tilde{\mathbf{C}}_{\mathbf{AB}}\mathrm{E}_m\bar{\mathbf{Y}} = \tilde{\mathbf{C}}_{\mathbf{AB}}\boldsymbol{\beta} = (&AB_{11} - AB_{12} - AB_{21} + AB_{22}, \ldots,\\ &AB_{11} - AB_{1L} - AB_{21} + AB_{2L}, \ldots,\\ &AB_{11} - AB_{12} - AB_{K1} + AB_{K2}, \ldots,\\ &AB_{11} - AB_{1L} - AB_{K1} + AB_{KL})^t.\end{aligned}$$

Chaque élément de ce vecteur de dimension $(K-1)(L-1)$ définit l'une des $(K-1)(L-1)$ interactions, ce qui correspond précisément aux contrastes entre les effets des interactions définies par (2.2). Par exemple, le premier élément peut être interprété comme étant l'écart de l'effet du traitement et de la combinaison particulière du facteur A au niveau 2 et du facteur B au niveau 2 par rapport aux deux effets principaux de ces facteurs.

2.3 Test de Wald

Les hypothèses spécifiées à la section 2.2 peuvent être testées au moyen d'un test de Wald (Wald 1943), dont l'application est fréquente dans les procédures de test fondées sur le plan de sondage (voir par exemple Skinner, Holt et Smith (1989) ;Chambers et Skinner (2003)). Si $\hat{\bar{\mathbf{Y}}}$ désigne un estimateur sans biais sous le plan pour $\bar{\mathbf{Y}}, \mathbf{C}$ désigne la matrice de contrastes $\mathbf{C}_{\mathbf{A}}, \mathbf{C}_{\mathbf{B}},$ ou $\mathbf{C}_{\mathbf{AB}}$ définie dans (2.14), (2.15) ou (2.16), et $\mathrm{cov}\left(\mathbf{C}\hat{\bar{\mathbf{Y}}}\right)$ désigne la matrice de covariance des contrastes entre $\hat{\bar{\mathbf{Y}}}$, alors les hypothèses peuvent être testées au moyen de la statistique de Wald $W = \hat{\bar{\mathbf{Y}}}^t\mathbf{C}^t\left\{\mathrm{cov}\left(\mathbf{C}\hat{\bar{\mathbf{Y}}}\right)\right\}^{-1}\mathbf{C}\hat{\bar{\mathbf{Y}}}$. À la présente section, nous étendons aux plans factoriels intégrés les estimateurs GREG proposés par van den Brakel et Renssen (2005) et par van den Brakel (2008) pour les expériences à un seul facteur. Pour simplifier la notation, l'indice inférieur q sera omis dans y_{iqkl}, puisqu'il n'est pas nécessaire de faire une sommation explicite sur l'indice inférieur d'intervieweur dans la plupart des formules élaborées dans la suite de l'article.

Afin d'appliquer le mode d'inférence assisté par un modèle à l'analyse des expériences intégrées, nous supposons pour chaque unité de la population que, dans le modèle d'erreur de mesure (2.5), la valeur intrinsèque u_i est une réalisation indépendante du modèle de régression linéaire suivant :

$$u_i = \beta^t\mathbf{x}_i + e_i, \tag{2.17}$$

où e_i est un vecteur d'information auxiliaire de dimension H, β est un vecteur de coefficients de régression de dimension H et e_i désigne les résidus, qui sont des variables aléatoires indépendantes de

variance ω_i^2. Il est nécessaire que les variances ω_i^2 soient connues jusqu'à un facteur d'échelle commun, c'est-à-dire $\omega_i^2 = \omega^2 \nu_i$, avec ν_i connu. L'estimateur GREG de $\bar{Y}_{kl}$, fondé sur les n_{kl} observations du sous-échantillon s_{kl}, est défini comme étant (Särndal et coll. 1992)

$$\hat{\bar{Y}}_{kl;\text{greg}} = \hat{\bar{Y}}_{kl} + \hat{\mathbf{b}}_{kl}^t \left(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}\right), k = 1, 2, \ldots, K, \text{ et } l = 1, 2, \ldots, L, \tag{2.18}$$

où

$$\hat{\bar{Y}}_{kl} = \frac{1}{N} \sum_{i=1}^{n_{kl}} \frac{y_{ikl}}{\pi_i^*}, \tag{2.19}$$

désigne l'estimateur HT de $\bar{Y}_{kl}$, $\bar{\mathbf{X}}$ désigne la moyenne de population finie des variables auxiliaires $\mathbf{x}$, et $\hat{\bar{\mathbf{X}}}$ désigne l'estimateur HT de $\bar{\mathbf{X}}$ basé sur les n_{kl} unités échantillonnées du sous-échantillon s_{kl}. En outre,

$$\hat{\mathbf{b}}_{kl} = \left(\sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2 \pi_i^*} \right)^{-1} \sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i y_{ikl}}{\omega_i^2 \pi_i^*}, \tag{2.20}$$

désigne l'estimateur de type HT pour les coefficients de régression dans (2.17) fondé sur les n_{kl} unités d'échantillonnage dans le sous-échantillon s_{kl}. Dans (2.19) et (2.20), π_i^* représentent les probabilités d'inclusion de premier ordre des unités d'échantillonnage dans les KL différents sous-échantillons dérivées à la sous-section 2.1. Maintenant, $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}} = \left(\hat{\bar{Y}}_{11;\text{greg}}, \ldots, \hat{\bar{Y}}_{KL;\text{greg}}\right)^t$ est un estimateur approximativement sans biais sous le plan pour $\bar{\mathbf{Y}}$, ainsi que pour $\mathrm{E}_m \bar{\mathbf{Y}}$ par définition.

Sous l'hypothèse nulle selon laquelle il n'existe pas d'effet du traitement ni d'interaction, il s'ensuit que $\mathbf{b}_{kl} = \mathbf{b}_{k'l'}$. Dans ce cas, il pourrait être efficace de substituer à $\hat{\mathbf{b}}_{kl}$ dans l'estimateur GREG (2.18) l'estimateur groupé

$$\hat{\mathbf{b}} = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2 \pi_i^*} \right)^{-1} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i y_{ikl}}{\omega_i^2 \pi_i^*}. \tag{2.21}$$

Puisque l'on doit estimer H au lieu de $KL \times H$ coefficients de régression, les estimations groupées des coefficients de régression $\hat{\mathbf{b}}$ seront plus précises, surtout si les sous-échantillons sont petits. Notons, toutefois, que de nombreux schémas de pondération utilisés fréquemment satisfont à la condition qu'il existe un vecteur constant λ tel que $\omega_i^2 = \lambda \mathbf{x}_i$ pour tout $i \in U$. Dans cette situation, l'estimateur GREG se réduit à la forme simplifiée $\hat{\bar{Y}}_{kl;\text{greg}} = \hat{\mathbf{b}}_{kl}^t \bar{\mathbf{X}}$ (Särndal et coll. 1992, section 6.5). Sous cette forme simplifiée, les effets du traitement sont complètement inclus dans les coefficients de régression. Dans le cas de l'estimateur groupé (2.21), les KL estimateurs GREG sont exactement égaux par définition, puisque $\hat{\bar{Y}}_{kl;\text{greg}} = \hat{\mathbf{b}}^t \bar{\mathbf{X}}$ pour tout k et l.

Une expression de la matrice de covariance des contrastes entre les éléments de $\hat{\mathbf{Y}}_{\mathbf{GREG}}$ où la covariance est calculée sous le plan d'échantillonnage, le plan d'expérience et le modèle d'erreur de mesure, est donnée par

$$\text{cov}\left(\mathbf{C}\hat{\mathbf{Y}}_{\mathbf{GREG}}\right) = \text{E}_m\text{E}_s\mathbf{C}\mathbf{D}\mathbf{C}^t, \tag{2.22}$$

où E_s désigne l'espérance par rapport au plan d'échantillonnage et $\mathbf{D}$, une matrice diagonale de dimensions $KL \times KL$ dont les éléments diagonaux sont

$$d_{kl} = \frac{1}{n_{kl}\left(n_{++}-1\right)}\sum_{i=1}^{n_{++}}\left(\frac{n_{++}\left(y_{ikl}-\mathbf{b}_{kl}^t\mathbf{x}_i\right)}{N\pi_i}-\frac{1}{n_{++}}\sum_{i'=1}^{n_{++}}\frac{n_{++}\left(y_{i'kl}-\mathbf{b}_{kl}^t\mathbf{x}_{i'}\right)}{N\pi_{i'}}\right)^2, \tag{2.23}$$

dans le cas d'un plan PCR et

$$d_{kl} = \sum_{b=1}^{B}\frac{1}{n_{bkl}\left(n_{b++}-1\right)}\sum_{i=1}^{n_{b++}}\left(\frac{n_{b++}\left(y_{ikl}-\mathbf{b}_{kl}^t\mathbf{x}_i\right)}{N\pi_i}-\frac{1}{n_{b++}}\sum_{i'=1}^{n_{b++}}\frac{n_{b++}\left(y_{i'kl}-\mathbf{b}_{kl}^t\mathbf{x}_{i'}\right)}{N\pi_{i'}}\right)^2, \tag{2.24}$$

dans le cas d'un plan PBR. Un estimateur de $\mathbf{D}$ peut être dérivé du plan d'expérience, conditionnellement au modèle d'erreur de mesure et au plan de sondage. Donc, la matrice de covariance (2.22) est, de manière commode, énoncée implicitement comme étant l'espérance sous le modèle d'erreur de mesure et sous le plan de sondage. Un estimateur fondé sur le plan de sondage de cette matrice de covariance est donné par

$$\hat{\text{cov}}\left(\mathbf{C}\hat{\mathbf{Y}}_{\mathbf{GREG}}\right) = \text{E}_m\text{E}_s\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t, \tag{2.25}$$

où $\hat{\mathbf{D}}$ est une matrice diagonale de dimensions $KL \times KL$ dont les éléments sont

$$\hat{d}_{kl} = \frac{1}{n_{kl}(n_{kl}-1)}\sum_{i=1}^{n_{kl}}\left(\frac{n_{++}(y_{ikl}-\hat{\mathbf{b}}_{kl}^t\mathbf{x}_i)}{N\pi_i}-\frac{1}{n_{kl}}\sum_{i'=1}^{n_{kl}}\frac{n_{++}(y_{i'kl}-\hat{\mathbf{b}}_{kl}^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2, \tag{2.26}$$

dans le cas d'un plan PCR et

$$\hat{d}_{kl} = \sum_{b=1}^{B}\frac{1}{n_{bkl}(n_{bkl}-1)}\sum_{i=1}^{n_{bkl}}\left(\frac{n_{b++}(y_{ikl}-\hat{\mathbf{b}}_{kl}^t\mathbf{x}_i)}{N\pi_i}-\frac{1}{n_{bkl}}\sum_{i'=1}^{n_{bkl}}\frac{n_{b++}(y_{i'kl}-\hat{\mathbf{b}}_{kl}^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2, \tag{2.27}$$

dans le cas d'un plan PBR. Les preuves de (2.22) et (2.25) sont données par van den Brakel (2010) et ressemblent au calcul de la matrice de covariance pour les expériences à un seul facteur donné dans van den Brakel et Renssen (2005) et dans van den Brakel (2008).

Les résultats pour (2.22) et (2.25) sont obtenus sous la condition qu'il existe un vecteur $\mathbf{a}$ de dimension H constant tel que $\mathbf{a}^t\mathbf{x}_i = 1$ pour tout $i \in U$. Cette condition est assez faible, puisqu'elle implique l'application d'un modèle de pondération qui utilise au moins la taille de la population finie comme information a priori. Voir van den Brakel et Renssen (2005) ou van den Brakel (2008) pour une discussion plus détaillée.

Puisque les KL sous-échantillons sont tirés sans remise d'une population finie, il existe une covariance sous le plan de sondage non nulle entre les éléments de $\hat{\hat{\mathbf{Y}}}_{\mathbf{GREG}}$. De ce point de vue, il est remarquable que (2.25) possède une structure comme celle que l'on obtiendrait si les sous-échantillons étaient tirés indépendamment par échantillonnage avec remise en utilisant des probabilités de sélection inégales. Cela donne une procédure d'estimation de la variance intéressante pour les expériences intégrées, puisqu'aucune covariance sous le plan entre les estimations sur les sous-échantillons ne figure dans (2.25) et qu'aucune probabilité d'inclusion de second ordre n'est requise dans les estimateurs de variance (2.26) et (2.27). Nous obtenons ce résultat parce que nous dérivons la covariance des contrastes entre $\hat{\hat{\mathbf{Y}}}_{\mathbf{GREG}}$ au lieu de la matrice de covariance de $\hat{\hat{\mathbf{Y}}}_{\mathbf{GREG}}$ proprement dite. Une interprétation détaillée de ce résultat figure dans van den Brakel et Renssen (2005) ou dans van den Brakel (2008). Voir van den Brakel et Binder (2000) ainsi que Hidiroglou et Lavallée (2005) pour les approximations de la matrice de covariance de $\hat{\hat{\mathbf{Y}}}_{\mathbf{GREG}}$.

Les estimateurs sous le plan de sondage $\hat{\hat{\mathbf{Y}}}_{\mathbf{GREG}}$ et $\hat{\text{cov}}\left(\mathbf{C}\hat{\hat{\mathbf{Y}}}_{\mathbf{GREG}}\right)$ peuvent être utilisés pour construire une statistique de Wald fondée sur le plan pour tester les hypothèses décrites à la section 2.2 :

$$W = \hat{\hat{\mathbf{Y}}}_{\mathbf{GREG}}^{t}\mathbf{C}^{t}\left(\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^{t}\right)^{-1}\mathbf{C}\hat{\hat{\mathbf{Y}}}_{\mathbf{GREG}}. \tag{2.28}$$

Les inférences sous le plan sont généralement fondées sur des approximations normales en grand échantillon pour construire les intervalles de confiance des estimations ponctuelles ou les valeurs p et les régions critiques pour les statistiques de test. Sous cette approche, il s'ensuit, sous l'hypothèse nulle, que la statistique de Wald est asymptotiquement distribuée comme une variable aléatoire du ki-carré centrée, où le nombre de degrés de liberté est égal au nombre de contrastes spécifiés dans l'hypothèse.

La statistique de Wald pour les hypothèses au sujet des effets principaux et des interactions est donnée par (2.28) en utilisant la matrice de contrastes $\mathbf{C}_{\mathbf{A}}, \mathbf{C}_{\mathbf{B}}$ ou $\mathbf{C}_{\mathbf{AB}}$. Sous l'hypothèse nulle, il s'ensuit que $W \to \chi^2_{[K-1]}$ pour le test au sujet des effets principaux du facteur $A, W \to \chi^2_{[L-1]}$ pour le test au sujet des effets principaux du facteur B et $W \to \chi^2_{[(K-1)(L-1)]}$ pour le test au sujet des interactions, où $\chi^2_{[p]}$ désigne une variable aléatoire qui suit une loi du ki-carré centrée avec p degrés de liberté.

Le test de Wald pour les effets principaux peut encore être simplifié. Nous développons les expressions pour le test de Wald pour les effets principaux du facteur A. Des expressions similaires peuvent être dérivées pour les effets principaux du facteur B. Soit

$$\begin{aligned}\hat{\hat{\mathbf{Y}}}_{\mathbf{A;GREG}} &= \left(\hat{\hat{Y}}_{1.;\text{greg}}, \ldots, \hat{\hat{Y}}_{K.;\text{greg}}\right)^{t}, \text{ avec } \hat{\hat{Y}}_{k.;\text{greg}} = \frac{1}{L}\sum_{l=1}^{L}\hat{\hat{Y}}_{kl;\text{greg}}, \\ \hat{\mathbf{D}}_{\mathbf{A}} &= \text{Diag}\left(\hat{d}_{1.}, \ldots, \hat{d}_{K.}\right), \text{ avec } \hat{d}_{k.} = \frac{1}{L^2}\sum_{l=1}^{L}\hat{d}_{kl}.\end{aligned} \tag{2.29}$$

Il s'ensuit que $\mathbf{C}_{\mathbf{A}}\hat{\hat{\mathbf{Y}}}_{\mathbf{GREG}} = \tilde{\mathbf{C}}_{\mathbf{A}}\hat{\hat{\mathbf{Y}}}_{\mathbf{A;GREG}}$ et $\mathbf{C}_{\mathbf{A}}\hat{\mathbf{D}}\mathbf{C}_{\mathbf{A}}^{t} = \tilde{\mathbf{C}}_{\mathbf{A}}\hat{\mathbf{D}}_{\mathbf{A}}\tilde{\mathbf{C}}_{\mathbf{A}}^{t}$. En appliquant le lemme d'inversion de matrice, la statistique de Wald pour les effets principaux du facteur A peut se simplifier comme il suit :

$$
\begin{aligned}
W &= \hat{\tilde{\mathbf{Y}}}^t_{\mathbf{A;GREG}} \tilde{\mathbf{C}}^t_{\mathbf{A}} (\tilde{\mathbf{C}}_{\mathbf{A}} \hat{\mathbf{D}}_{\mathbf{A}} \tilde{\mathbf{C}}^t_{\mathbf{A}})^{-1} \tilde{\mathbf{C}}_{\mathbf{A}} \hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}} \\
&= \hat{\tilde{\mathbf{Y}}}^t_{\mathbf{A;GREG}} \left(\hat{\mathbf{D}}^{-1}_{\mathbf{A}} - \frac{1}{\text{Trace}(\hat{\mathbf{D}}^{-1}_{\mathbf{A}})} \hat{\mathbf{D}}^{-1}_{\mathbf{A}} \mathbf{j}_{(K-1)} \mathbf{j}^t_{(K-1)} \hat{\mathbf{D}}^{-1}_{\mathbf{A}} \right) \hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}} \qquad (2.30) \\
&= \sum_{k=1}^{K} \frac{\hat{\tilde{Y}}^2_{k.;greg}}{\hat{d}_{k.}} - \left(\sum_{k=1}^{K} \frac{1}{\hat{d}_{k.}} \right)^{-1} \left(\sum_{k=1}^{K} \frac{\hat{\tilde{Y}}^2_{k.;greg}}{\hat{d}_{k.}} \right)^2 .
\end{aligned}
$$

Enfin, notons que l'estimateur HT (2.19) ne satisfait pas à la condition voulant qu'il existe un vecteur $\mathbf{a}$ de dimension H tel que $\mathbf{a}^t \mathbf{x}_i = 1$ pour tout $i \in U$. L'usage minimal de l'information auxiliaire entrant dans l'estimateur GREG s'obtient au moyen d'un schéma de pondération qui n'utilise que la taille de la population finie comme information a priori, c'est-à-dire $(x_i) = 1$ et $\omega_i^2 = \omega^2$ (Särndal et coll. 1992, section 7.4). Sous ce schéma de pondération, il s'ensuit que

$$
\hat{\tilde{Y}}_{kl;greg} = \left(\sum_{i=1}^{n_{kl}} \frac{1}{\pi_i^*} \right)^{-1} \left(\sum_{i=1}^{n_{kl}} \frac{y_{ikl}}{\pi_i^*} \right) \equiv \tilde{y}_{kl} \qquad (2.31)
$$

et $(\hat{\mathbf{b}}_{kl}) = \tilde{y}_{kl}$. On reconnaîtra que l'expression (2.31) est l'estimateur par le ratio de Hájek pour une moyenne de population (Hájek 1971). Ce schéma de pondération satisfait à la condition qu'il existe un vecteur $\mathbf{a}$ de dimension H tel que $\mathbf{a}^t \mathbf{x}_i = 1$ pour tout $i \in U$. Par conséquent, un estimateur approximativement sans biais sous le plan de la matrice de covariance des contrastes entre les estimations sur les sous-échantillons est donné par (2.26) et (2.27) pour un plan PCR et un plan PBR, respectivement, où $\hat{\mathbf{b}}^t_{kl} \mathbf{x}_i = \tilde{y}_{kl}$. L'estimateur (2.31) est préférable à l'estimateur HT (2.19), puisque (2.31) est plus stable et que la matrice de covariance des contrastes entre (2.31) prend toujours la forme relativement simple de (2.25).

2.4 Cas particuliers

Pour deux cas particuliers, nous allons montrer que la statistique de Wald fondée sur le plan de sondage est égale à la statistique F de l'analyse de variance classique. Par conséquent, il faut considérer le remplacement de (2.26) ou (2.27) par un estimateur de variance groupé de type ANOVA pour les éléments diagonaux de la matrice $\hat{\mathbf{D}}$. Pour un plan PCR, un estimateur de variance groupé de ce type est donné par

$$
\hat{d}^p_{kl} = \frac{1}{n_{kl}(n_{++} - KL)} \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{k'l'}} \left(\frac{n_{++}(y_{ik'l'} - \hat{\mathbf{b}}^t_{k'l'} \mathbf{x}_i)}{N \pi_i} - \frac{1}{n_{k'l'}} \sum_{i'=1}^{n_{k'l'}} \frac{n_{++}(y_{i'k'l'} - \hat{\mathbf{b}}^t_{k'l'} \mathbf{x}_{i'})}{N \pi_{i'}} \right)^2 , \qquad (2.32)
$$

et pour un plan PBR, par

$$
\hat{d}^p_{kl} = \sum_{b=1}^{B} \frac{1}{n_{bkl}(n_{b++} - KL)} \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{bk'l'}} \left(\frac{n_{b++}(y_{ik'l'} - \hat{\mathbf{b}}^t_{k'l'} \mathbf{x}_i)}{N \pi_i} - \frac{1}{n_{bk'l'}} \sum_{i'=1}^{n_{bk'l'}} \frac{n_{b++}(y_{i'k'l'} - \hat{\mathbf{b}}^t_{k'l'} \mathbf{x}_{i'})}{N \pi_{i'}} \right)^2 . \qquad (2.33)
$$

Considérons maintenant un plan PCR intégré dans un échantillon autopondéré, c'est-à-dire $\pi_i = n_{++}/N$, avec des sous-échantillons de taille égale, c'est-à-dire $n_{kl} = n_{k'l'} = n_s$. Pour toutes les unités des KL sous-échantillons, les probabilités d'inclusion sont données par $\pi_i^* = n_s/N$. Soit $\bar{y} = \left(1/n_s\right)\sum_{i=1}^{n_s} y_{ikl}$. Sous l'estimateur par le ratio de Hájek (2.31) et l'estimateur de variance groupé (2.32), il s'ensuit que $\hat{\bar{Y}}_{kl;\text{greg}} = \bar{y}_{kl}$, $\hat{\mathbf{b}}_{kl} = \bar{y}_{kl}$, et

$$\hat{d}_{kl}^{\,p} = \frac{1}{n_s\left(n_{++} - KL\right)}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\sum_{i=1}^{n_s}\left(y_{ik'l'} - \bar{y}_{k'l'}\right)^2 \equiv \frac{\hat{S}_{p;\,\text{PCR}}^2}{n_s}.$$

Les estimations des paramètres des K niveaux du facteur A en prenant la moyenne sur les L niveaux du facteur B sont données par

$$\bar{y}_{k.} = \frac{1}{L}\sum_{l=1}^{L}\bar{y}_{kl} = \frac{1}{n_{k+}}\sum_{l=1}^{L}\sum_{i=1}^{n_s} y_{ikl}, k = 1,\ldots,K, \tag{2.34}$$

avec $n_{k+} = \sum_{l=1}^{L} n_{kl}$. Les éléments diagonaux de $\hat{\mathbf{D}}_{\mathbf{A}}$ sont maintenant donnés par

$$\hat{d}_{k.}^{\,p} = \frac{1}{L^2}\sum_{l=1}^{L}\hat{d}_{kl}^{\,p} = \frac{1}{L^2}\sum_{l=1}^{L}\frac{\hat{S}_{p;\,\text{PCR}}^2}{n_s} = \frac{\hat{S}_{p;\,\text{PCR}}^2}{n_{k+}}, k = 1,\ldots,K. \tag{2.35}$$

Soit $\bar{y}_{..} = \left(1/n_{++}\right)\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n_s} y_{ikl}$. En insérant (2.34) et (2.35) dans (2.30), on obtient l'expression qui suit pour la statistique de Wald des effets principaux du facteur A

$$W = \frac{1}{\hat{S}_{p;\,\text{PCR}}^2}\left(\sum_{k=1}^{K} n_{k+}\bar{y}_{k.}^2 - n_{++}\bar{y}_{..}^2\right). \tag{2.36}$$

Notons que, dans (2.36), $W/\left(K-1\right)$ correspond à la statistique F pour les effets principaux issus d'une analyse de variance à deux critères de classification avec les interactions (Scheffé 1959, chapitre 4). Sous l'hypothèse nulle et l'hypothèse que les erreurs suivent des lois normales indépendantes, la statistique F dans le cas de deux critères de classification suit une loi F avec $\left(K-1\right)$ et $\left(n_{++} - KL\right)$ degrés de liberté, ce qui est désigné par $F_{[n_{++}-KL]}^{[K-1]}$. Si $n_{++} \to \infty$, alors $F_{[n_{++}-KL]}^{[K-1]} \to \chi_{[K-1]}^2/(K-1)$. D'où, la statistique F et la statistique de Wald ont la même loi limite.

Considérons maintenant un plan PBR intégré dans un plan de sondage autopondéré avec sous-échantillons de taille égale, d'où $\pi_i = n_{+++}/N$ et $n_{kl} = n_{k'l'} = n_s$, avec $n_{+++} = \sum_{b=1}^{B} n_{b++}$. Soit $\bar{y}_{bkl} = \left(1/n_{bkl}\right)\sum_{i=1}^{n_{bkl}} y_{ikl}$. En outre, nous supposons que la fraction d'unités d'échantillonnage assignée à chaque combinaison de traitements dans chaque bloc est égale, c'est-à-dire $n_{bkl}/n_{b++} = n_s/n_{+++}$, et que les blocs sont de suffisamment grande taille pour supposer que $n_{b++}/\left(n_{b++} - KL\right) \approx 1$. Sous l'estimateur par le ratio de Hájek (2.31) et l'estimateur de variance groupé (2.33), il s'ensuit que $\hat{\bar{Y}}_{kl;\text{greg}} = \bar{y}_{kl}$, $\hat{\mathbf{b}}_{kl} = \bar{y}_{kl}$, et

$$\hat{d}_{kl}^{\,p} = \sum_{b=1}^{B} \frac{1}{n_{bkl}\left(n_{b++} - KL\right)} \left(\frac{n_{b++}}{n_{+++}}\right)^2 \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{bk'l'}} \left(y_{ik'l'} - \bar{y}_{bk'l'}\right)^2$$
$$\approx \frac{1}{n_s n_{+++}} \sum_{b=1}^{B} \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{bk'l'}} \left(y_{ik'l'} - \bar{y}_{bk'l'}\right)^2 \equiv \frac{\hat{S}^2_{p;\,\mathrm{PBR}}}{n_s}.$$

Les estimations des paramètres des K niveaux du facteur A en prenant la moyenne sur les L niveaux du facteur B et les blocs sont désignées par

$$\bar{y}_{.k.} = \frac{1}{L} \sum_{l=1}^{L} \bar{y}_{kl} = \frac{1}{n_{+k+}} \sum_{b=1}^{B} \sum_{l=1}^{L} \sum_{i=1}^{n_{bkl}} y_{ikl}, k = 1, \ldots, K, \tag{2.37}$$

où $n_{+k+} = \sum_{b=1}^{B} \sum_{l=1}^{L} n_{bkl}$. Les éléments diagonaux de $\hat{\mathbf{D}}_{\mathbf{A}}$ sont donnés par

$$\hat{d}_{k.}^{\,p} = \frac{1}{L^2} \sum_{l=1}^{L} \hat{d}_{kl}^{\,p} = \frac{\hat{S}^2_{p;\,\mathrm{PBR}}}{n_{+k+}}, k = 1, \ldots, K. \tag{2.38}$$

Soit $\bar{y}_{...} = \left(1/n_{+++}\right) \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n_{bkl}} y_{ikl}$. Si ces résultats sont insérés dans (2.30), l'expression de la statistique de Wald pour les effets principaux du facteur A peut se simplifier en

$$W = \frac{1}{\hat{S}^2_{p;\,\mathrm{PBR}}} \left(\sum_{k=1}^{K} n_{+k+} \bar{y}^2_{.k.} - n_{+++} \bar{y}^2_{...} \right). \tag{2.39}$$

On peut voir que, dans (2.39), $W/(K-1)$ correspond à la statistique F des effets principaux d'une analyse de variance à deux critères de classification avec les interactions (Scheffé1959, chapitre 4). Dans le cas d'un plan PCR, ces statistiques de Wald et F ont la même loi limite.

3 Plans factoriels avec plus de deux facteurs

Les résultats élaborés pour les plans factoriels $K \times L$ sont maintenant étendus à des plans comprenant plus de deux facteurs. Commençons par introduire une notation plus générale pour les facteurs de traitement. Soit A_g le g^{e} facteur de traitement dans l'expérience avec les niveaux $a_g = 1, \ldots, M_g$. Dans le cas général, g = 1,…,G facteurs sont inclus dans l'expérience. Les paramètres de population observés sous les $M_1 M_2 \ldots M_G$ combinaisons de traitements sont rassemblés dans le vecteur $\bar{\mathbf{Y}} = \left(\bar{Y}_{11\ldots1}, \ldots, \bar{Y}_{a_1 a_2 \ldots a_G}, \ldots, \bar{Y}_{M_1 M_2 \ldots M_G}\right)^t$. L'indice indiquant les niveaux d'un facteur s'applique à l'intérieur de chaque niveau du facteur qui le précède. Donc, l'indice a_g va de $a_g = 1, \ldots, M_g$ dans chaque niveau de $a_{(g-1)}$. Les hypothèses au sujet des effets principaux et des interactions sont, comme il est motivé à la section 2.2, au sujet de $\bar{\mathbf{Y}}$ en espérance sous le modèle d'erreur de mesure.

Les matrices des contrastes pour les effets principaux et les interactions dans (2.13) sont élaborées pour le cas général d'un plan factoriel $M_1 \times M_2 \times \ldots \times M_G$. Soit $\mathcal{A} = \{1, \ldots, G\}$ l'ensemble d'étiquettes pour les facteurs et $\tilde{\mathbf{C}}_{\mathbf{A_g}} = (\mathbf{j}_{(M_g - 1)} \mid -\mathbf{I}_{(M_g - 1)})$ Commençons par définir les trois fonctions suivantes :

$$\mathbf{J}_{\mathbf{1_g}} = \begin{cases} \mathbf{j}^t_{M_1} \otimes \ldots \otimes \mathbf{j}^t_{M(g-1)} & : \ g > 1 \\ 1 & : \ g = 1 \end{cases},$$

$$\mathbf{J}_{\mathbf{2_g}} = \begin{cases} \mathbf{j}^t_{M(g+1)} \otimes \ldots \otimes \mathbf{j}^t_{M_G} & : \ g < G \\ 1 & : \ g = G \end{cases},$$

$$\mathbf{J}_{\mathbf{3_{g,g'}}} = \begin{cases} \mathbf{j}^t_{M(g+1)} \otimes \ldots \otimes \mathbf{j}^t_{M(g'-1)} & : \ g' - g > 1 \\ 1 & : \ g' = g + 1 \end{cases}.$$

L'effet principal du facteur A_g est défini comme étant la moyenne des $M_g - 1$ contrastes entre les M_g niveaux sur les niveaux des $G - 1$ autres facteurs et est donné par :

$$\mathbf{C}_{\mathbf{A_{g_1}}} = \left(\prod\nolimits_{g \in \mathcal{A} \setminus \{g_1\}} M_g \right)^{-1} \mathbf{J}_{\mathbf{1_{g_1}}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_1}}} \otimes \mathbf{J}_{\mathbf{2_{g_1}}}, g_1 = 1, \ldots, G.$$

La post-multiplication de $\tilde{\mathbf{C}}_{\mathbf{A_{g_1}}}$ par $\mathbf{J}_{\mathbf{2_{g_1}}}$ somme sur les niveaux des facteurs $A_{(g_1+1)} \ldots A_G$ qui sont emboîtés dans chaque niveau A_{g_1}. Subséquemment, $\tilde{\mathbf{C}}_{\mathbf{A_{g_1}}}$ définit les $M_{g_1} - 1$ contrastes entre les niveaux de A_{g_1} qui sont emboîtés dans chaque combinaison des niveaux de $A_1 \ldots A_{(g_1-1)}$. La prémultiplication de $\tilde{\mathbf{C}}_{\mathbf{A_{g_1}}}$ par $\mathbf{J}_{\mathbf{1_{g_1}}}$ ajoute les matrices de contrastes $\tilde{\mathbf{C}}_{\mathbf{A_{g_1}}}$ qui sont emboitées dans toutes les combinaisons des niveaux de $A_1 \ldots A_{(g_1-1)}$.

L'interaction entre A_{g_1} et A_{g_2} est définie comme celle des $M_{g_2} - 1$ contrastes du facteur A_{g_2} avec les $M_{g_1} - 1$ contrastes de A_{g_1} en prenant la moyenne sur les niveaux des $G - 2$ autres facteurs et est donnée par :

$$\mathbf{C}_{\mathbf{A_{g_1}A_{g_2}}} = \left(\prod\nolimits_{g \in \mathcal{A} \setminus \{g_1, g_2\}} M_g \right)^{-1} \mathbf{J}_{\mathbf{1_{g_1}}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_1}}} \otimes \mathbf{J}_{\mathbf{3_{g_1,g_2}}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_2}}} \otimes \mathbf{J}_{\mathbf{2_{g_2}}},$$
$$g_1 = 1, \ldots, G-1, g_2 = 1, \ldots, G, g_1 < g_2.$$

La post-multiplication de $\tilde{\mathbf{C}}_{\mathbf{A_{g_2}}}$ by $\mathbf{J}_{\mathbf{2_{g_2}}}$ ajoute les niveaux des facteurs $A_{(g_2+1)} \ldots A_G$ qui sont emboîtés dans chaque niveau de A_{g_2}. $\tilde{\mathbf{C}}_{\mathbf{A_{g_2}}}$ définit les contrastes de l'effet principal du facteur A_{g_2} qui sont emboîtés dans chaque combinaison des niveaux $A_1 \ldots A_{(g_2-1)}$. La post-multiplication de $\tilde{\mathbf{C}}_{\mathbf{A_{g_1}}}$ par $\mathbf{J}_{\mathbf{3_{g_1,g_2}}}$ somme les matrices de contrastes $\tilde{\mathbf{C}}_{\mathbf{A_{g_2}}}$ sur les niveaux de $A_{(g_1+1)} \ldots A_{(g_2-1)}$ qui sont emboîtés dans chaque combinaison des niveaux de $A_1 \ldots A_{g_1}$. La prémultiplication de $\mathbf{J}_{\mathbf{3_{g_1,g_2}}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_2}}} \otimes \mathbf{J}_{\mathbf{2_{g_2}}}$ par $\tilde{\mathbf{C}}_{\mathbf{A_{g_1}}}$ définit les contrastes des interactions entre A_{g_1} et A_{g_2}, à l'intérieur de chaque combinaison des niveaux de

$A_1 \ldots A_{(g_1-1)}$. Enfin, la prémultiplication de $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ par $\mathbf{J}_{\mathbf{1}_{g_1}}$ somme les contrastes des interactions entre A_{g_1} et A_{g_2} sur les niveaux de $A_1 \ldots A_{(g_1-1)}$.

L'interaction entre A_{g_1}, A_{g_2} et A_{g_3} est définie comme étant les $M_{g_3} - 1$ contrastes du facteur A_{g_3} entre les interactions de A_{g_1} et A_{g_2} en prenant la moyenne sur les niveaux des $G - 3$ autres facteurs. Ce processus s'étend d'une manière comparable aux interactions d'ordre plus élevé, ce qui donne les définitions qui suivent de ces interactions :

$$
\begin{aligned}
\mathbf{C}_{\mathbf{A}_{g_1}\mathbf{A}_{g_2}\mathbf{A}_{g_3}} &= \left(\prod\nolimits_{g\in\mathcal{A}\setminus\{g_1,g_2,g_3\}} M_g\right)^{-1} \mathbf{J}_{\mathbf{1}_{g_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_1}} \otimes \mathbf{J}_{\mathbf{3}_{g_1,g_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_2}} \otimes \mathbf{J}_{\mathbf{3}_{g_2,g_3}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_3}} \otimes \mathbf{J}_{\mathbf{2}_{g_3}}, \\
&\quad g_1 = 1,\ldots,G-2, g_2 = 2,\ldots,G-1, g_3 = 3,\ldots,G, g_1 < g_2 < g_3, \\
\mathbf{C}_{\mathbf{A}_{g_1}\mathbf{A}_{g_2}\mathbf{A}_{g_3}\mathbf{A}_{g_4}} &= \left(\prod\nolimits_{g\in\mathcal{A}\setminus\{g_1,g_2,g_3,g_4\}} M_g\right)^{-1} \mathbf{J}_{\mathbf{1}_{g_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_1}} \otimes \mathbf{J}_{\mathbf{3}_{g_1,g_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_2}} \otimes \mathbf{J}_{\mathbf{3}_{g_2,g_3}} \otimes \\
&\quad \tilde{\mathbf{C}}_{\mathbf{A}_{g_3}} \otimes \mathbf{J}_{\mathbf{3}_{g_3,g_4}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_4}} \otimes \mathbf{J}_{\mathbf{2}_{g_3}}, \\
&\quad g_1 = 1,\ldots,G-3, g_2 = 2,\ldots,G-2, g_3 = 3,\ldots,G-1, \\
&\quad g_4 = 4,\ldots,G, g_1 < g_2 < g_3 < g_4, \\
&\vdots \\
\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3\ldots\mathbf{A}_G} &= \tilde{\mathbf{C}}_{\mathbf{A}_{g_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_3}} \otimes \ldots \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_4}}
\end{aligned}
$$

Le nombre de lignes de chaque matrice de contrastes coïncide avec le nombre de contrastes qui définissent les divers effets principaux et interactions. Le nombre de colonnes de ces matrices est égal à $M_1 M_2 \ldots M_G$.

Ces matrices de contrastes sont insérées dans (2.13) pour définir les diverses hypothèses au sujet des effets principaux et des interactions entre les G facteurs de traitement. Les unités d'échantillonnage présentes dans l'échantillon initial sont réparties aléatoirement entre toutes les combinaisons de traitements possibles conformément à un plan PCR ou un plan PBR, pour donner $M_1 M_2 \ldots M_G$ sous-échantillons différents. Soit $n_{a_1\ldots a_G}$ le nombre d'unités d'échantillonnage assignées à la combinaison de traitements $a_1 \ldots a_G$ dans le sous-échantillon $s_{a_1\ldots a_G}$ et $n_{+\ldots+}$ la taille de l'échantillon initial. Dans le cas d'un plan PCR, les probabilités d'inclusion de premier ordre des unités dans le sous-échantillon $s_{a_1\ldots a_G}$ sont maintenant données par $\pi_i^* = \pi_i \left(n_{a_1\ldots a_G} / n_{+\ldots+}\right)$. Dans le cas d'un plan PBR, les probabilités d'inclusion de premier ordre des unités dans l'échantillon $s_{a_1\ldots a_G}$ sont données par $\pi_i^* = \pi_i \left(n_{ba_1\ldots a_G} / n_{b+\ldots+}\right)$, où $n_{ba_1\ldots a_G}$ désigne le nombre d'unités d'échantillonnage assignées à la combinaison de traitements $a_1 \ldots a_G$ dans le bloc b et $n_{b+\ldots+}$ le nombre total d'unités d'échantillonnage dans le bloc b.

Maintenant, $\hat{\bar{Y}}_{a_1\ldots a_G;\text{greg}}$ désigne l'estimateur GREG de $\bar{Y}_{a_1\ldots a_G}$ fondé sur les observations obtenues dans le sous-échantillon $s_{a_1\ldots a_G}$ et est défini de manière analogue à l'expression (2.18). Ces $M_1 M_2 \ldots M_G$

estimateurs GREG sont regroupés dans le vecteur $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}} = \left(\hat{\bar{Y}}_{1\ldots1;\mathrm{greg}}, \ldots, \hat{\bar{Y}}_{M_1\ldots M_G;\mathrm{greg}}\right)^t$ qui est un estimateur approximativement sans biais sous le plan de $\bar{\mathbf{Y}}$ et $\mathrm{E}_m\bar{\mathbf{Y}}$. Les estimateurs sous le plan de sondage des matrices de covariance des contrastes entre les éléments de $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ sont définis par (2.25), où les éléments diagonaux de $\hat{\mathbf{D}}$ sont définis de manière analogue à l'expression (2.26) dans le cas d'un plan PCR ou à l'expression (2.27) dans le cas d'un plan PBR.

Enfin, les hypothèses au sujet des effets principaux et des interactions sont testées en se servant de la statistique de Wald (2.28), qui suit asymptotiquement la même loi qu'une variable aléatoire du ki-carré dont le nombre de degrés de liberté est égal au nombre de contrastes spécifiés dans les diverses hypothèses. À titre d'exemple, les matrices de contrastes des effets principaux et des interactions d'un plan factoriel comprenant quatre facteurs sont présentées au tableau 3.1.

Tableau 3.1
Contrastes dans un plan factoriel $M_1 \times M_2 \times M_3 \times M_4$

Matrice de contrastes	Nombre de contrastes (degrés de liberté)
$\mathbf{C}_{\mathbf{A}_1} = 1/\left(M_2M_3M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$M_1 - 1$
$\mathbf{C}_{\mathbf{A}_2} = 1/\left(M_1M_3M_4\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$M_2 - 1$
$\mathbf{C}_{\mathbf{A}_3} = 1/\left(M_1M_2M_4\right)\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$M_3 - 1$
$\mathbf{C}_{\mathbf{A}_4} = 1/\left(M_1M_2M_3\right)\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$M_4 - 1$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2} = 1/\left(M_3M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$\left(M_1 - 1\right)\left(M_2 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_3} = 1/\left(M_2M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$\left(M_1 - 1\right)\left(M_3 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_4} = 1/\left(M_2M_3\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_2\mathbf{A}_3} = 1/\left(M_1M_4\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$\left(M_2 - 1\right)\left(M_3 - 1\right)$
$\mathbf{C}_{\mathbf{A}_2\mathbf{A}_4} = 1/\left(M_1M_3\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_2 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_3\mathbf{A}_4} = 1/\left(M_1M_2\right)\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_3 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3} = 1/\left(M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$\left(M_1 - 1\right)\left(M_2 - 1\right)\left(M_3 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_4} = 1/\left(M_3\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_2 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_3\mathbf{A}_4} = 1/\left(M_2\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_3 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_2\mathbf{A}_3\mathbf{A}_4} = 1/\left(M_1\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_2 - 1\right)\left(M_3 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3\mathbf{A}_4} = \tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_2 - 1\right)\left(M_3 - 1\right)\left(M_4 - 1\right)$

4 Autres extensions

Jusqu'à présent, nous avons pris en considération les plans expérimentaux dans lesquels les unités finales d'échantillonnage du plan de sondage sont réparties aléatoirement entre les traitements. À cause de contraintes de travail sur le terrain, il pourrait exister des raisons pratiques de répartir aléatoirement des grappes d'unités d'échantillonnage entre les différents traitements, au prix de réduire la puissance des tests d'hypothèse au sujet des effets du traitement. Par exemple, il pourrait être intéressant d'attribuer la même combinaison de traitements aux unités d'échantillonnage qui appartiennent à un même ménage ou qui sont assignées à un même intervieweur. Van den Brakel (2008) a élaboré une procédure d'analyse fondée sur le plan de sondage pour les expériences à un seul facteur, conçue selon un plan PCR ou un plan PBR, dans laquelle les grappes d'unités d'échantillonnage sont réparties aléatoirement entre les traitements. Ces méthodes peuvent être étendues directement à l'analyse des plans factoriels examinés dans le présent article.

Considérons le cas général d'un plan factoriel $M_1 \times M_2 \times \ldots \times M_G$. Les grappes d'unités d'échantillonnage dans l'échantillon initial sont réparties aléatoirement entre les différentes combinaisons de traitements. La probabilité conditionnelle qu'une unité d'échantillonnage soit attribuée à un sous-échantillon est calculée à partir des fractions de grappes attribuées aux différentes combinaisons de traitements à l'intérieur de l'échantillon ou à l'intérieur de chaque bloc. Voir van den Brakel (2008) pour les détails. L'estimateur GREG pour $\bar{Y}_{a_1 \ldots a_G}$ est défini de manière analogue à l'expression (2.18). Les estimateurs sous le plan de sondage pour les matrices de covariance des contrastes entre les éléments de $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ sont définis par (2.25), où les éléments diagonaux de $\hat{\mathbf{D}}$ sont définis comme dans l'expression (4.6) donnée dans van den Brakel (2008), qui est fondée sur la variance entre les estimations des totaux de grappes.

Les paramètres cibles d'une enquête sont souvent définis comme un ratio de deux totaux de population. Van den Brakel (2008) a élaboré une procédure d'analyse fondée sur le plan pour tester les hypothèses au sujet des ratios dans le cas d'expériences à un seul facteur selon un plan PCR ou un plan PBR. Ces résultats peuvent être étendus à l'analyse des plans factoriels traités dans le présent article. Pour chaque sous-échantillon, on peut construire un ratio de deux estimateurs GREG pour chaque combinaison de traitements. Les estimateurs sous le plan de sondage des matrices de covariance des contrastes entre les ratios sont définis par (2.25), où les éléments diagonaux de $\hat{\mathbf{D}}$ sont définis de manière analogue à l'expression (4.11) donnée dans van den Brakel (2008), qui est un estimateur de la variance du ratio de deux estimateurs GREG. La statistique de Wald (2.28) est utilisée pour tester les hypothèses au sujet des effets principaux et des interactions.

5 Tester de nouvelles lettres d'introduction pour l'Enquête sur la population active des Pays-Bas

La présente section décrit une expérience portant sur différentes lettres d'introduction intégrées dans l'Enquête sur la population active (EPA) des Pays-Bas, en guise d'exemple numérique illustrant la méthode élaborée dans l'article.

5.1 Plan de sondage

L'EPA est une enquête à panel rotatif. Chaque mois, un échantillon en grappes à deux degrés stratifié d'environ 6 000 adresses est tiré d'un registre de toutes les adresses connues aux Pays-Bas. Les strates correspondent aux régions géographiques, et les municipalités sont considérées comme les unités primaires d'échantillonnage, et les adresses, comme les unités secondaires d'échantillonnage. Toutes les personnes résidant à une adresse particulière, jusqu'à un maximum de trois, sont incluses dans l'échantillon. Durant la première vague, les données sont recueillies par interview sur place assistée par ordinateur. Puis, les répondants sont de nouveau interviewés quatre fois, à intervalle trimestriel, par interview téléphonique assistée par ordinateur.

La procédure de pondération de l'EPA est fondée sur l'estimateur GREG de Särndal et coll. (1992). Les probabilités d'inclusion reflètent le plan de sondage utilisé pour sélectionner les ménages, ainsi que les différents taux de réponse selon la région géographique. Le schéma de pondération est fondé sur une combinaison de différentes variables sociodémographiques catégoriques. L'un des paramètres les plus importants de l'EPA est la population active en chômage, qui est définie comme le ratio du nombre total de chômeurs à la population active totale.

5.2 Plan d'expérience

La lettre d'introduction est l'un des paramètres de conception d'une enquête qui a une incidence sur les taux de réponse et la coopération des répondants (De Leeuw, et coll. (2007)). La lettre d'introduction standard de l'EPA est adressée à tous les occupants du logement et le ton est formel et autoritaire. Par conséquent, cette lettre n'est pas conforme aux théories de la psychologie sociale concernant la participation aux enquêtes proposées par Groves, et coll. (1992) et par Groves et Couper (1998). Pour essayer d'améliorer les taux de réponse à l'EPA, Luiten, et coll. (2008) ont proposé pour l'enquête différentes lettres de rechange qui ont été étudiées empiriquement au moyen d'une expérience sur le terrain à grande échelle intégrée dans l'EPA.

Le premier facteur pris en considération dans cette expérience, disons A, est celui de la salutation du répondant qui comporte deux niveaux, c'est-à-dire l'approche standard où la lettre est adressée à l'occupant du logement (A_1) par opposition à une lettre nominative (A_2). En principe, les lettres nominatives devraient avoir plus de chance d'être lues et, par conséquent, augmenter les taux de réponse et la participation aux enquêtes. Le deuxième facteur, disons B, est celui du contenu de la lettre et comporte trois niveaux, c'est-à-dire la lettre officielle standard (B_1) par opposition à deux autres lettres $(B_2$ et $B_3)$. Dans la première lettre de rechange, le contenu de la lettre standard est adapté afin d'expliquer pourquoi l'enquête est réalisée, quels sont les bénéfices de la participation pour le répondant et pourquoi il est important pour Statistics Netherlands que le répondant participe à l'enquête. Dans la deuxième lettre de rechange, on tente d'atténuer le ton formel de la lettre standard. Les trois versions de la lettre d'introduction figurent dans van den Brakel (2010).

La mise en œuvre d'une nouvelle lettre en tant que lettre standard pour l'EPA ne sera prise en considération que si une expérience randomisée a permis de prouver son effet positif sur le comportement de réponse et de quantifier son effet sur les estimations des principaux paramètres. Les deux facteurs ont été testés dans un plan factoriel 2×3 qui a produit six combinaisons de traitements. Cette expérience a

été intégrée dans la première vague de l'EPA pour une période de cinq mois (de décembre 2007 à avril 2008). Durant cette période, la taille brute mensuelle de l'échantillon a été répartie aléatoirement entre six sous-échantillons selon un plan PBR où les intervieweurs représentaient les variables de bloc. Environ 220 intervieweurs étaient disponibles pour le travail sur le terrain. Dans l'analyse, les régions d'intervieweur adjacentes ont été fusionnées en 13 blocs. Une fraction de l'échantillon de 0,8 a été assignée à la lettre d'introduction ordinaire, c'est-à-dire la combinaison de traitements $A_1 \times B_1$. Une fraction de l'échantillon de 0,04 a été assignée à chacune des cinq autres combinaisons de traitements.

La répartition des unités d'échantillonnage entre les traitements est fondée principalement sur des arguments pratiques. L'intégration d'expériences dans des enquêtes par sondage courantes présente deux objectifs concurrents. Pour estimer les statistiques officielles aussi précisément que possible, il est souhaitable d'affecter autant d'unités d'échantillonnage que possible au groupe témoin, puisque ce sous-échantillon est également utilisé pour produire les statistiques publiées régulièrement. Pour estimer les contrastes dans l'expérience aussi précisément que possible, il est par contre souhaitable de diviser l'échantillon total de manière égale entre les différentes combinaisons de traitements. Dans l'application décrite ici, il a été décidé que la perte maximale acceptable de taille de l'échantillon pour la production des statistiques publiées régulièrement était de 20 %, ce qui a conduit à la répartition susmentionnée entre les combinaisons de traitements. Pour un taux de réponse de 56 % et une taille mensuelle d'échantillon de 6 000 ménages, il est prévu qu'environ 13 440 ménages soient observés dans le groupe témoin $A_1 \times B_1$ et 670 ménages pour chacune des combinaisons de traitements de rechange.

Même si la répartition est fondée sur des considérations pratiques, il est important d'avoir une idée de la puissance de l'expérience planifiée. La variable cible analysée dans le présent article est la proportion de chômeurs dans la population active, exprimée en pourcentage. Si l'on ne tient pas compte du plan en blocs de l'expérience, il s'ensuit que la variance de traitement est égale à $\hat{d}_{kl} = \hat{S}_{kl}^2 / n_{kl}$, où $\hat{S}_{kl}^2$ est défini implicitement par (2.26). On suppose que $\hat{S}_{kl}^2$ est égal, disons, à $\hat{S}^2$ pour chaque combinaison de traitements. Au moyen des données d'échantillon disponibles, on obtient pour la population active en chômage que $\hat{S}^2 = 285$. Maintenant, l'écart minimal observable pour un contraste qui donnerait lieu au rejet de l'hypothèse nulle sous un seuil de signification et un niveau de puissance spécifiés est égal à

$$\Delta = \sqrt{\text{var}(\Delta)}\left(Z_{(1-\alpha/2)} + Z_{(1-\beta)}\right), \tag{5.1}$$

où $Z_{(\gamma)}$ désigne le γ^{e} percentile de la loi normale centrée réduite, α désigne le seuil de signification du test et $(1-\beta)$, la puissance. L'effet principal du facteur A concerne un contraste $\hat{\Delta}_A = \left(\hat{\bar{Y}}_{1.;\text{greg}} - \hat{\bar{Y}}_{2.;\text{greg}}\right)$. De (2.29) il découle que la variance de ce contraste est égale à $\hat{\text{var}}\left(\hat{\Delta}_A\right) = \left(\hat{S}^2/9\right)\sum_{l=1}^{3}\left(1/n_{1l} + 1/n_{2l}\right)$. L'effet principal du facteur B concerne deux contrastes $\hat{\Delta}_{B_l} = \left(\hat{\bar{Y}}_{.1;\text{greg}} - \hat{\bar{Y}}_{.l;\text{greg}}\right), l = 2,3$ avec les variances $\hat{\text{var}}\left(\hat{\Delta}_{B_l}\right) = \left(\hat{S}^2/4\right)\sum_{k=1}^{2}\left(1/n_{k1} + 1/n_{kl}\right), l = 2,3.$ Les interactions entre les facteurs A et B concernent les deux contrastes $\hat{\Delta}_{AB_l} = \left(\hat{\bar{Y}}_{11;\text{greg}} - \hat{\bar{Y}}_{1l.;\text{greg}} - \hat{\bar{Y}}_{21;\text{greg}} + \hat{\bar{Y}}_{2l.;\text{greg}}\right)$ avec les variances $\hat{\text{var}}\left(\hat{\Delta}_{AB_l}\right) = \hat{S}^2\left(1/n_{11} + 1/n_{1l} + 1/n_{21} + 1/n_{2l}\right)$, $l = 2,3.$

En insérant les variances des différents contrastes dans (5.1), on obtient les valeurs minimales des différences qui donneraient lieu au rejet de l'hypothèse nulle pour les effets principaux et les interactions pour les tailles d'échantillon, les seuils de signification et les niveaux de puissance spécifiés. Dans le tableau 5.1, ces différences sont calculées pour la population active en chômage sous l'application de la répartition susmentionnée et d'un plan équilibré, où la taille de l'échantillon pour chaque combinaison de traitements est égale à 2 800. Les valeurs sont présentées pour les hypothèses alternatives non spécifiées à un seuil de signification de 5 % et une puissance de 50 %, 80 % et 90 %. Dans la théorie des plans d'expérience, un niveau de puissance de 80 % est généralement accepté pour la détermination de la taille de l'échantillon. Dans le cas des sondages, les exigences quant à la taille minimale d'échantillon sont généralement fondées uniquement sur les exigences relatives au seuil de signification, ce qui correspond à un niveau de puissance de 50 %. Les différences sont spécifiées pour l'application de tests distincts aux contrastes. L'effet principal du facteur B et l'effet des interactions contiennent l'un et l'autre deux contrastes. Afin de maintenir le seuil de signification global de 5 %, on a également calculé les différences pour les deux tests en utilisant la procédure de comparaison simultanée de Bonferroni.

Tableau 5.1
Différence observable pour la population active en chômage exprimée en pourcentage au seuil de signification de 5 % et à différents niveaux de puissance

Contraste	Nombre de Contrastes	Puissance, tests t distincts			Puissance, test t de Bonferroni		
		50 %	80 %	90 %	50 %	80 %	90 %
Plan appliqué							
Effet principal A	1	0,96	1,36	1,58	0,96	1,36	1,58
Effet principal B	2	1,12	1,59	1,85	1,27	1,75	2,00
Interaction	2	2,23	3,19	3,69	2,55	3,51	4,00
$A_1 \times B_1 - A_k \times B_l$	5	1,31	1,87	2,17	1,72	2,28	2,57
Plan équilibré							
Effet principal A	1	0,51	0,73	0,84	0,51	0,73	0,84
Effet principal B	2	0,63	0,89	1,03	0,71	0,98	1,12
Interaction	2	1,25	1,79	2,07	1,43	1,97	2,25
$A_1 \times B_1 - A_k \times B_l$	5	0,88	1,26	1,46	1,16	1,54	1,74

Le tableau 5.1 illustre les différents aspects des expériences intégrées et des plans d'expérience factoriels. Premièrement, il illustre les coûts-avantages d'un plan factoriel. Deux fois plus d'unités expérimentales sont nécessaires si les effets principaux des deux facteurs sont évalués avec la même précision au moyen de deux expériences à un seul facteur distinctes. Le tableau 5.1 montre aussi que la puissance est nettement plus faible pour le test des interactions que pour les tests des deux effets principaux. Plus le nombre de facteurs de traitement combinés dans une seule expérience est grand, plus la

taille d'échantillon assignée à chaque combinaison de traitements est petite et plus la puissance est faible pour les tests d'interaction. Cette constatation oblige à mettre en perspective l'avantage souvent mentionné voulant que les plans d'expérience factoriels permettent aussi de tester les interactions entre les différents facteurs de traitement. En pratique, les tailles d'échantillon sont fondées sur les calculs de puissance faits pour les tests portant sur les effets principaux. Par conséquent, seuls les effets d'interaction importants peuvent être décelés avec suffisamment de puissance. Un plan factoriel offre néanmoins l'avantage d'accroître la validité des effets principaux observés, puisqu'ils sont testés pour une grande gamme de conditions.

En cas de rejet de l'hypothèse nulle selon laquelle il n'existe pas d'interaction, les effets principaux sont difficiles à interpréter. Dans cette situation, il est plus utile de comparer le groupe témoin, c'est-à-dire $A_1 \times B_1$, aux cinq combinaisons de traitements de rechange. Les différences minimales observables pour ces cinq contrastes donnant lieu au rejet de l'hypothèse nulle au seuil de signification de 5 % et à différents niveaux de puissance sont également incluses dans le tableau 5.1.

La comparaison des valeurs minimales des différences sous le plan d'expérience appliqué et sous le plan d'expérience équilibré illustre la perte de puissance lorsqu'on choisit une répartition fortement asymétrique entre les combinaisons de traitements. La minimisation du risque de perdre trop de précision pour les statistiques publiées régulièrement est la raison qui motive le choix de ce genre de répartition. Cela illustre clairement la dualité qui découle de la combinaison de deux objectifs concurrents dans une expérience intégrée; d'une part, produire des estimations pour publication ordinaire et, d'autre part, tester les contrastes de différentes combinaisons de traitements.

Afin d'évaluer la valeur des résultats que permet d'obtenir cette expérience, les différences minimales observables dans le contexte de l'expérience sont liées aux erreurs-types des estimations ordinaires d'après les données de l'enquête. Généralement, les erreurs-types des estimations d'après les données d'enquête au niveau national sont beaucoup plus petites que les différences minimales observables dans une expérience, puisque la taille d'échantillon assignée aux différents traitements est généralement nettement plus petite que la taille d'échantillon ordinaire. Toutefois, si l'on adopte l'hypothèse selon laquelle les différences observées dans une expérience au niveau national s'appliquent aussi aux estimations d'après l'enquête pour les domaines importants, les différences observables au moyen de l'expérience pourraient devenir comparables aux erreurs-types de ces estimations de domaine. Cela suppose qu'il n'existe pas d'interaction entre les domaines et les effets de traitement. Au niveau national, l'erreur-type des chiffres mensuels concernant la population active en chômage est égale à 0,15 point de pourcentage. Les erreurs-types pour les domaines varient entre 0,3 et 1,0 point de pourcentage. La comparaison de ces erreurs-types avec les différences présentées au tableau 5.1 montre que les effets principaux demeurent plus importants que les erreurs-types au niveau national, mais deviennent comparables avec la précision des estimations mensuelles ordinaires par domaine.

5.3 Résultats

Le tableau 5.2 donne un aperçu des taux de réponse des ménages dans les six sous-échantillons de l'expérience. Il permet de conclure que les différentes lettres d'introduction donnent lieu à des différences assez faibles entre les taux de réponse. Le facteur A produit une augmentation du taux de réponse de 2,4 points de pourcentage, grâce à l'utilisation d'une lettre personnalisée (après correction des proportions

pour tenir compte de la répartition non équilibrée de l'échantillon entre les combinaisons de traitements). Les autres lettres prises en considération dans le facteur B ont donné lieu à une diminution du taux de réponse de 1,5 point de pourcentage (option B_2) et de 1,9 point de pourcentage (option B_3).

Tableau 5.2
Taux de réponse – Expérience avec différentes lettres d'introduction

Traitement	Réponse		Refus		Reste		Total
$A_1 \times B_1$	13 234	56,69 %	5 127	21,96 %	4 985	21,35 %	23 346
$A_1 \times B_2$	604	53,59 %	271	24,05 %	252	22,36 %	1 127
$A_1 \times B_3$	635	56,34 %	254	22,54 %	238	21,12 %	1 127
$A_2 \times B_1$	662	59,00 %	256	22,82 %	204	18,18 %	1 122
$A_2 \times B_2$	663	59,09 %	236	21,03 %	223	19,88 %	1 122
$A_2 \times B_3$	627	55,64 %	259	22,98 %	241	21,38 %	1 127

Un modèle de régression logistique est utilisé pour modéliser le comportement de réponse afin de tester les hypothèses au sujet de l'effet des deux facteurs de traitement. Il s'agit d'une analyse conditionnelle type qui ne tient pas compte des caractéristiques du plan de sondage telles que les probabilités de sélection inégales et le groupement des ménages (grappes) à l'intérieur des municipalités. La mise en grappes induite par le plan de sondage à deux degrés n'est pas prise en compte, puisque les ménages sont répartis aléatoirement entre les traitements dans l'expérience. Cette analyse par régression logistique est axée sur les différences dans l'échantillon observé, dues ici à des différences de non-réponse sélective. Elle fournit des renseignements supplémentaires indiquant si les facteurs accroissent la réponse à l'échelle de la population cible complète ou si des groupes particuliers réagissent différemment aux divers traitements. Les interactions de deuxième ordre et d'ordre plus élevé entre les deux facteurs de traitement et les variables sociodémographiques catégoriques dans le modèle de régression logistique indiquent que la variation de la réponse entre les différentes sous-populations augmente et que ces dernières réagissent différemment aux traitements.

Dans le modèle de régression logistique, la variable dépendante binaire indique si un ménage a répondu complètement comparativement aux autres catégories de réponse. Il est supposé que le comportement de réponse dépende :

- d'une moyenne générale;
- des facteurs de traitement A (nom) et B (contenu);
- d'une variable de bloc comprenant 13 catégories;
- des variables auxiliaires :
    - niveau d'urbanisation en cinq catégories;
    - sexe, en trois catégories, spécifiant si un ménage comprend seulement des hommes, seulement des femmes, ou un mélange d'hommes et de femmes;

- l'âge en tant que variable quantitative contenant l'âge moyen des membres du ménage;
- l'ethnicité en sept catégories spécifiant si un ménage est composé de personnes nées aux Pays-Bas, de personnes d'origine occidentale, de personnes d'origine non occidentale, et toutes les combinaisons possibles;
- la composition de la famille, en quatre catégories à savoir conjoints, famille monoparentale, personne seule, et une catégorie Autre.

Toutes les interactions de troisième ordre entre les variables sont prises en considération au départ pour la sélection descendante (*backwards*) des variables du modèle. Le modèle final sélectionné contient les termes qui sont présentés dans la première colonne du tableau 5.3. Par souci de concision, les coefficients de régression, assortis de leurs erreurs-types et des statistiques de test pour les catégories distinctes, sont donnés pour les facteurs de traitement seulement. L'hypothèse selon laquelle il n'existe pas d'interaction entre les deux facteurs de traitement ne peut pas être rejetée (valeur p pour la statistique de Wald égale 0,121). Du tableau 5.3, il découle que le facteur A, c'est-à-dire l'utilisation d'une lettre adressée nominativement à une personne, a un effet positif, mais non significatif sur le taux de réponse. Le facteur B, c'est-à-dire deux lettres de rechange dont le contenu est amélioré, a même un effet légèrement négatif, mais non significatif, sur les taux de réponse. Ce résultat est remarquable, puisque les deux lettres de rechange ont pour objectif d'améliorer le ton formel de la lettre standard, mais il est en harmonie avec les résultats d'une expérience antérieure dans laquelle la réaction à une lettre d'introduction à l'EPA plus informelle a également donné lieu à des taux de réponse significativement plus faibles (van den Brakel 2008). Puisqu'il n'y a pas d'interaction entre les facteurs de traitement et les variables auxiliaires, il n'existe également aucune indication que les facteurs de traitement induisent une réaction chez des sous-populations particulières.

Tableau 5.3
Analyse des taux de réponse par régression logistique

Paramètre	Coefficient	Erreur-type	Statistique de Wald	d.d.l.	Valeur p
Moyenne	0,287	0,078	13,604	1	0,000
Bloc			212,425	12	0,000
Traitement A (nom, A_2)	0,083	0,045	3,394	1	0,065
Traitement B (contenu)			2,965	2	0,227
Option 1 (B_2)	-0,046	0,051	0,816	1	0,366
Option 2 (B_3)	-0,083	0,051	2,678	1	0,102
Urbanisation			16,589	4	0,002
Ethnicité			127,734	6	0,000
Sexe			48,076	2	0,000
Composition de la famille			27,339	3	0,000

La deuxième étape de l'analyse a pour objectif de déterminer si les estimations de la population active en chômage obtenues sur les six sous-échantillons en utilisant les différentes lettres d'introduction

diffèrent significativement. La méthode d'analyse fondée sur le plan de sondage élaborée dans le présent article est utilisée afin de tenir compte du plan de sondage, du plan d'expérience et de la procédure d'estimation de l'EPA. Les estimations de la population active en chômage sous les six combinaisons de traitements différentes sont obtenues au moyen de l'estimateur GREG. Cette analyse non conditionnelle a pour but de déterminer si les diverses lettres d'introduction donnent lieu à des différences dans le biais de sélection, après correction pour tenir compte des différences de taux de réponse en utilisant la procédure d'estimation fondée sur le plan de sondage appliquée pour l'EPA ordinaire.

Dans cette analyse, le modèle d'erreur de mesure linéaire (2.1) est appliqué à une variable de réponse binaire. Cela pourrait sembler strict, puisque les modèles logistiques sont plus naturels ici. Toutefois, sous l'approche assistée par modèle, il est fréquent d'appliquer des modèles de régression linéaire pour obtenir un estimateur GREG pour des variables de réponse binaire. En outre, dans le cas de l'EPA des Pays-Bas, on émet l'hypothèse d'un modèle de régression linéaire pour arriver à un estimateur GREG pour produire les chiffres officiels sur la population active. Afin d'élaborer pour les expériences intégrées dans l'enquête une méthode d'analyse fondée sur le plan de sondage qui tient également compte de l'estimateur GREG utilisé pour l'enquête ordinaire, on émet de la même façon l'hypothèse d'un modèle d'erreur de mesure linéaire. Une discussion détaillée de l'utilisation et de l'interprétation d'un modèle d'erreur de mesure linéaire appliqué à une variable de réponse binaire figure dans van den Brakel (2008).

Dans l'estimateur GREG (2.18), les probabilités d'inclusion reflètent le plan de sondage de l'EPA et le plan d'expérience utilisé pour diviser l'échantillon initial en six sous-échantillons. Le schéma de pondération appliqué pour caler les poids de sondage était le suivant : *âge* + *région* + *état matrimonial* + *sexe* +*niveau d'urbanisation*, où les cinq variables sont catégoriques. Il s'agit d'une version réduite du schéma de pondération ordinaire de l'EPA.

Les résultats des estimations pour les six sous-échantillons sont résumés au tableau 5.4, où la proportion de la population active en chômage est exprimée en pourcentage. Il semble ne se dégager aucune tendance systématique entre les estimations calculées sous les sous-échantillons. Ces estimations et les estimations de leur variance indiquent qu'il n'y a pas de différence significative entre le groupe témoin et les cinq combinaisons de traitements de rechange. Enfin, les effets principaux et les effets d'interaction des deux facteurs de traitement sont évalués, en tenant compte du fait que l'expérience a été conçue selon un plan PBR dans lequel les régions d'intervieweur adjacentes sont fusionnées pour former 13 blocs. Les résultats de l'analyse sont résumés au tableau 5.5.

Tableau 5.4
Estimations ponctuelles et erreurs-types, population active en chômage (exprimée en pourcentage)

Combinaison de traitements		Estimation $\hat{\bar{Y}}_{kl;greg}$	Erreur-type $\sqrt{\hat{d}_{kl}}$
$k\ (A_k)$	$l\ (B_l)$		
1	1	4,100 %	0,145 %
1	2	3,761 %	0,646 %
1	3	5,264 %	0,753 %
2	1	3,609 %	0,608 %
2	2	4,546 %	0,666 %
2	3	3,385 %	0,664 %

Tableau 5.5
Analyse des effets principaux et des interactions, population active en chômage (exprimée en pourcentage)

Source	Estimation $C\hat{\bar{Y}}_{greg}$	Statistique de Wald	d.d.l.	Valeur *p*
Traitement A (nom) $A_1 - A_2$	0,528	1,109	1	0,292
Traitement B (contenu)		0,732	2	0,694
$B_1 - B_2$	-0,300			
$B_1 - B_3$	-0,471			
Interaction		3,801	2	0,150
$AB_{11} - AB_{12} - AB_{21} + AB_{22}$	1,276			
$AB_{11} - AB_{13} - AB_{21} + AB_{23}$	-1,388			

Les résultats d'analyse résumés au tableau 5.5 permettent de conclure que rien n'indique que les différentes lettres d'introduction donnent lieu à des estimations différentes des paramètres. Cette constatation concorde avec celle découlant des résultats de l'analyse des taux de réponse. S'il n'existe aucune preuve empirique que les différentes lettres d'introduction affectent les taux de réponse de la population complète ou d'une sous-population, on pourrait s'attendre à ce qu'il n'y ait aucune différence significative entre les estimations des paramètres.

Rien n'indique que les lettres de rechange examinées dans la présente expérience améliorent le comportement de réponse ou ont des effets systématiques sur les estimations des variables cibles, telles que la population active en chômage. Par conséquent, il a été décidé de ne pas adapter la lettre d'introduction standard de l'EPA.

6 Discussion

Dans les plans d'expérience factoriels, on fait varier les niveaux de deux ou de plusieurs facteurs de traitement et l'on examine simultanément toutes les combinaisons de traitements possibles. Les plans d'expérience sont d'usage très répandu dans le domaine de l'expérimentation scientifique pour plusieurs raisons. Les effets principaux des facteurs sont examinés en prenant leur moyenne sur les niveaux de tous les autres facteurs. Par conséquent, les conclusions au sujet des divers effets sont fondées sur une grande gamme de conditions, ce qui augmente la validité des résultats. En outre, il est possible d'analyser l'interaction entre les différents facteurs de traitement, quoique la puissance de ces tests diminue à mesure qu'augmente le nombre de facteurs combinés dans une expérience. Enfin, les plans factoriels sont plus efficaces que les plans à un seul facteur, car un moins grand nombre d'unités expérimentales est nécessaire pour estimer les effets principaux avec la même précision.

Dans le présent article, une théorie fondée sur le plan de sondage est élaborée pour l'analyse des résultats de plans d'expérience factoriels intégrés dans des échantillons probabilistes. Cette approche convient particulièrement bien pour quantifier les effets des différents paramètres de conception d'un processus d'enquête sur l'estimation des paramètres d'une enquête par sondage. On trouve des applications de cette approche dans les domaines de la conception complète d'enquête, de la recherche

empirique sur les pratiques d'enquête et de la quantification des ruptures dans les séries issues d'enquêtes répétées. En outre, des procédures d'analyse fondées sur le plan de sondage sont élaborées pour tester les hypothèses au sujet des moyennes de population pour les plans d'expérience factoriels dans lesquels les unités finales d'échantillonnage sont réparties aléatoirement entre les différentes combinaisons de traitements selon un plan complètement randomisé (PCR) ou un plan en blocs randomisés (PBR). Les procédures, pour les plans factoriels dans lesquels les grappes d'unités d'échantillonnage sont réparties aléatoirement entre les combinaisons de traitements ou pour les tests d'hypothèse au sujet de ratios de totaux de population, sont obtenues de manière analogue aux méthodes présentées dans van den Brakel (2008) pour des expériences à un seul facteur.

L'estimateur de variance fondé sur le plan de sondage ne nécessite ni les probabilités d'inclusion conjointes ni les covariances sous le plan entre les différents sous-échantillons. D'où on obtient une méthode d'analyse fondée sur le plan de sondage applicable aux plans d'expérience factoriels intégrés dans des échantillons probabilistes complexes possédant une structure relativement simple intéressante, comme celle que l'on obtiendrait si les unités d'échantillonnage étaient tirées avec probabilité de sélection inégale avec remise. Les avantages classiques des plans d'expérience factoriels, résumés dans le premier paragraphe de la discussion, s'appliquent encore sous cette approche fondée sur le plan de sondage. Comme l'illustre l'expression de la variance (2.29), un moins grand nombre d'unités expérimentales est nécessaire pour estimer les effets principaux avec la même précision sous un plan factoriel que sous des plans à un seul facteur.

Les avantages d'un plan PBR par rapport à un plan PCR tient au fait que la variance inter-blocs est éliminée de l'estimation des effets du traitement. Dans la théorie classique, fondée sur un modèle, de l'analyse des expériences randomisées, il existe un test F pour les blocs ainsi que pour les facteurs de traitement. Par contre, sous les conditions de randomisation restreintes d'un plan PBR, d'aucuns soutiennent généralement qu'un test F pour les effets de bloc n'est pas valide. Dans ces cas, il existe des mesures de rechange pour évaluer l'efficacité d'un plan PBR; voir, par exemple, Montgomery (2001). Dans la théorie fondée sur le plan de sondage élaborée pour les plans PBR dans le présent article, une asymétrie existe entre le bloc et les facteurs de traitement, comme dans le cas de l'approche de randomisation suivie par Hinkelmann et Kempthorne (1994). En raison de la randomisation restreinte à l'intérieur des blocs, il n'existe aucun test pratique pour l'effet principal du facteur de bloc.

Remerciements

L'auteur remercie le rédacteur associé et les examinateurs anonymes de leurs commentaires constructifs au sujet d'une version antérieure du présent article. Les opinions exprimées dans l'article sont celles de l'auteur et ne reflètent pas forcément la politique de Statistics Netherlands.

Bibliographie

Chambers, R.L., et Skinner, C.J. (2003), *Analysis of Survey Data*, Chichester: John Wiley.

Chipperfield, J., et Bell, P. (2010). Embedded experiments in repeated and overlapping surveys. *Journal of the Royal Statistical Society*, Series A, 173, 51-66.

De Leeuw, E., Callegaro, M., Hox, J., Korendijk, E. et Lensvelt-Mulders, G. (2007). The influence of advance letters in response in telephone surveys. *Public Opinion Quarterly*, 71, 413-443.

Fienberg, S.E., et Tanur, J.M. (1987). Experimental and Sampling Structures: Parallels Diverging and Meeting. *Revue Internationale de Statistique*, 55, 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, 135-151.

Fienberg, S.E., et Tanur, J.M. (1989). Combining Cognitive and Statistical Approaches to Survey Design. *Science*, 243, 1017-1022.

Fienberg, S.E., et Tanur, J.M. (1996). Reconsidering the Fundemental Contributions of Fisher and Neyman on Experimentation and Sampling. *Revue Internationale de Statistique*, 64, 237-253.

Groves, R.M., Cialdini R.B. et Couper, M.P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56, 475-495.

Groves, R.M., et Couper, M.P. (1998). *Nonresponse in household interview surveys*, New York: John Wiley.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling" par D. Basu, dans *Foundations of Statistical Inference* (Éds., V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart, et Winston.

Hidiroglou, M.A., et Lavallée, P. (2005). Sondage indirect à deux phases : une application à l'essai de questionnaires sur le terrain. *Actes du Symposium de Statistique Canada de 2005 : Défis méthodologiques reliés aux besoins futurs d'information*.

Hinkelmann, K., et Kempthorne, O. (1994). *Design and Analysis of Experiments, Volume 1: Introduction to experimental design*, New York: John Wiley.

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

Jäckle, A., Roberts, C. et Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *Revue Internationale de Statistique*, 78, 3-20.

Kempthorne, O. (1955). The Randomization Theory of Experimental Inference. *Journal of the American Statistical Association*, 50, 946-967.

Luiten, A., Campanelli, P., Klaasen, D. et Beukenhorst, D. (2008). Advance letters and the language and behaviour profile, papier présenté au 19th International Workshop on Household Survey Nonresponse.

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.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

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.

Skinner, C.J., Holt, D. et Smith, T.M.F. (1989). *Analysis of Complex Surveys*, Chichester: John Wiley.

van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statistical Society*, Series A, 171, 581-613.

van den Brakel, J.A. (2010). Design-based analysis of factorial designs embedded in probability samples. Discussion paper 201014, Statistics Netherlands, Heerlen.

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, 805-810.

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, 217-231.

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


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

Techniques d'enquête, décembre 2013
Vol. 39, N° 2, pp. 385-403
Statistique Canada, N° 12-001-X au Catalogue

Estimation des déciles et estimation de la variance par rééchantillonnage dans le cas de données d'enquêtes complexes provenant de populations présentant une asymétrie positive

Stephen J. Kaputa et Katherine Jenny Thompson[1]

Résumé

Thompson et Sigman (2000) ont proposé une méthode d'estimation des médianes applicable à des données provenant de populations à forte asymétrie positive. Cette méthode comprend une interpolation sur des intervalles (classes) qui dépendent des données. Ils ont démontré qu'elle possède de bonnes propriétés statistiques pour les médianes calculées d'après un échantillon fortement asymétrique. La présente étude étend les travaux antérieurs aux méthodes d'estimation des déciles pour une population à asymétrie positive en utilisant des données d'enquête complexes. Nous présentons trois méthodes d'interpolation, ainsi que la méthode classique d'estimation des déciles (sans classes) et nous évaluons chaque méthode empiriquement au moyen d'une étude en simulation en utilisant les données sur les logements résidentiels provenant de l'Enquête sur la construction (Survey of Construction). Nous avons constaté qu'une variante de la méthode courante en utilisant le 95^{e} centile comme facteur d'échelle produit les estimations des déciles ayant les meilleures propriétés statistiques.

Mots clés : Médiane; méthode du demi-échantillon répété modifiée; interpolation; déciles.

1 Introduction

L'élaboration d'estimations valables des déciles pour des populations présentant une asymétrie positive à partir de données complexes pose des défis intéressants. Deux approches distinctes d'estimation des centiles au moyen de données d'enquête complexes sont décrites dans la littérature. Selon la première méthode (la méthode « classique »), on obtient les estimations des déciles en partant de fonctions de répartition empiriques, en sélectionnant la valeur de l'élément qui correspond au centile d'échantillon calculé et en additionnant les poids de sondage associés. Cette approche produit des estimations des déciles qui sont « presque sans biais », mais instables. Une autre approche consiste à grouper les données continues en intervalles disjoints (classes), puis à effectuer une interpolation linéaire sur la classe contenant le décile. Si les classes sont définies de manière appropriée, cette approche produit aussi des estimations des déciles quasi sans biais, tout en améliorant leur stabilité – du moins pour les rangs centiles éloignés de la queue de la distribution. Pour les centiles supérieurs, les données groupées dans les classes contiennent souvent très peu d'observations et n'ont que peu d'uniformité, voire aucune. Donc, la fiabilité des estimations des déciles élevés (par exemple 90^{e} centile ou supérieur) est rarement comparable à celle des estimations des autres déciles.

Bien que le recours à l'interpolation soit avantageux pour produire des estimations stables, l'élaboration d'un jeu optimal de classes pour une caractéristique donnée n'est pas toujours facile. Souvent, la distribution change au cours du temps et les largeurs et localisations des classes dans

1. Stephen J. Kaputa et Katherine Jenny Thompson, Office of Statistical Methods and Research for Economic Programs, US Census Bureau, 4600 Silver Hill RD, Washington, DC 20233. Courriel : Stephen.kaputa@census.gov.

l'échantillon doivent refléter ce changement d'échelle. Ainsi, le prix de vente moyen des logements unifamiliaux dans une région géographique donnée pourrait augmenter au cours du temps à cause de l'inflation, alors que la population de logements unifamiliaux dans cette région reste caractérisée par une distribution asymétrique, avec quelques logements chers situés dans la queue. De nombreux programmes fournissant des données économiques ont ce trait en commun. Par conséquent, dans le cas d'une enquête permanente, établir un jeu fixe de classes pour l'interpolation n'est pas une bonne idée. Pour résoudre ce problème, Thompson et Sigman (2000) ont adopté une méthode d'estimation des *médianes* pour les données provenant de populations ayant une asymétrique fortement positive. Leur méthode fait appel à l'interpolation sur des intervalles (classes) qui dépendent des données, après mise à l'échelle en fonction du 75^{e} percentile. L'étude antérieure portait sur les propriétés d'estimation et d'estimation de la variance des méthodes considérées, en utilisant la méthode des demi-échantillons répétés modifiée (MHS pour *modified half sample*) pour estimer la variance (Fay 1989; Judkins 1990).

La présente étude étend les travaux antérieurs aux méthodes d'estimation des déciles en utilisant des données d'enquête complexes échantillonnées dans une population à asymétrie positive. Nous présentons trois méthodes d'interpolation distinctes, ainsi que la méthode d'estimation classique des déciles (sans classes) et nous évaluons chaque méthode empiriquement, en utilisant les données sur les logements résidentiels provenant de la *Survey of Construction* (SOC) menée par le U.S. Census Bureau dans une étude en simulation. Nos travaux ont été motivés par une demande récente des utilisateurs des données de la SOC de produire et de publier des jeux complets d'estimations des déciles pour plusieurs caractéristiques des logements. Donc, l'étude a été menée sous les contraintes consistant à produire des estimations des médianes aussi fiables que celles publiées à l'heure actuelle et à estimer les variances par la méthode de rééchantillonnage MHS.

À la section 2, nous présentons les méthodes d'estimation des déciles proposées et donnons un aperçu de la méthode des demi-échantillons répétés modifiée. À la section 3, nous évaluons ces méthodes, en utilisant des données empiriques et simulées provenant de la Survey of Construction (SOC). Enfin, nous concluons par des recommandations à la section 4.

2 Méthodologie

2.1 Estimation des déciles

Nous considérons deux approches pour estimer les déciles dans le cas de données continues : la méthode des déciles d'échantillon (méthode DE) et l'interpolation. La méthode DE consiste à utiliser les poids d'échantillonnage classés par ordre pour situer l'estimation (Rao et Shao 1996). Pour ce faire, les valeurs des caractéristiques sont triées par ordre croissant et les poids d'échantillonnage sont cumulés jusqu'à ce que la valeur soit supérieure au pourcentage du poids total du décile souhaité.

Les méthodes d'interpolation consistent à regrouper les données continues en classes, puis à interpoler sur la classe contenant le décile. Pour obtenir l'estimation du décile $\left(\xi^{d}\right)$, nous utilisons la formule d'interpolation de Woodruff (Woodruff 1952) :

$$\xi^{d} = F^{-1}\left(d\hat{N}\right) \approx ll + \left(\frac{d\hat{N} - cf}{f_i}\right) * (i) \qquad (2.1)$$

où

F = fréquence cumulée de la caractéristique en utilisant les poids d'échantillonnage;
ll = limite inférieure de la classe contenant le décile;
$\hat{N}$ = nombre total estimé d'éléments dans la population;
cf = fréquence cumulée dans tous les intervalles précédant la classe contenant le décile d'échantillon;
f_i = fréquence de classe du décile (nombre total estimé d'éléments dans la population de l'intervalle contenant le décile d'échantillon);
i = largeur de la classe contenant le décile d'échantillon;
d = décile souhaité (0,1, 0,2, 0,3, …, 0,9).

On remarquera que cette formule n'exige pas que les classes soient toutes de même longueur. Cependant, elle requiert que les données soient distribuées uniformément dans chaque classe. Cette dernière exigence est celle qui pose vraiment un défi dans le cas d'une population fortement asymétrique, surtout dans la queue supérieure de la distribution.

La figure 2.1 qui suit illustre comment utiliser la méthode de Woodruff pour estimer le 80^e décile. Les données d'échantillon ont été groupées en 12 classes distinctes. La fonction de répartition empirique est produite au moyen du jeu complet de données d'échantillon pondérées. (Comme l'a fait remarquer un examinateur, la fonction de répartition empirique est très lisse pour des données d'enquête par sondage; en pratique, la courbe présenterait des échelons discrets. Néanmoins, la procédure de la méthode de Woodruff demeurerait la même.) L'estimation du décile est située à l'intersection de la courbe de la fonction de répartition empirique et de l'asymptote en rouge à $Y = 0{,}80$. Le 80^e décile est $F^{-1}(0{,}80)$, contenu dans la 5^e classe; l'estimation interpolée du 80^e centile s'obtient par conséquent en utilisant (2.1) sur la cinquième classe.

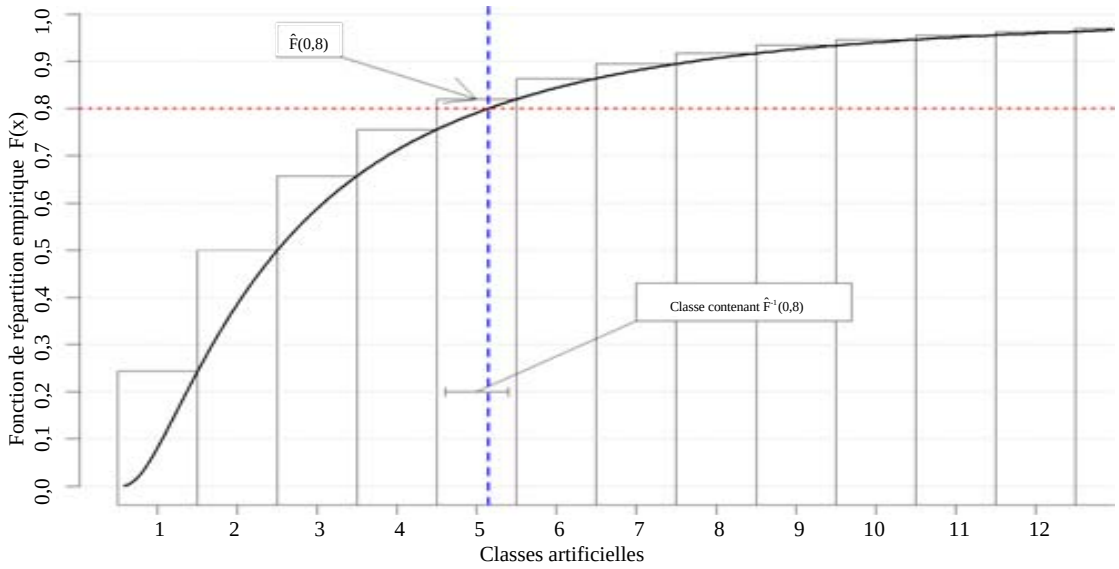


Figure 2.1 Illustration de la méthode de Woodruff

Il est parfois difficile de déterminer la taille de classe optimale à la fois pour l'estimation et pour l'estimation de la variance. À mesure que les classes rétrécissent (approchant une largeur de 1), les estimations de la variance deviennent plus instables. Le lissage des estimations par interpolation réduit l'instabilité de la variance, mais augmente le biais dans l'estimation. La composante de biais s'accroît à mesure que les largeurs de classe deviennent plus grandes.

En général, les données économiques ont une distribution à asymétrie positive. En outre, les distributions des caractéristiques des sous-domaines varient, et leurs moments respectifs changent au cours du temps à mesure qu'évolue l'économie. Par conséquent, il est presque impossible d'établir pour l'interpolation un jeu standard de classes fixes produisant des résultats cohérents au cours du temps. Au lieu de procéder ainsi, Thompson et Sigman (2000) ont élaboré une méthode de groupement par classe « dépendant des données », où la largeur de chaque classe est déterminée séparément par la cellule d'estimation. La méthode qu'ils recommandent comprend la conversion de chaque caractéristique à une échelle standard par transformation linéaire, puis l'utilisation d'un jeu standard de classes pour chaque caractéristique. Ils utilisent la transformation linéaire suivante

$$X_i' = X_i \times \frac{1\,000}{Q_{75}}$$

où Q_{75} est le 75^{e} centile (3^{e} quartile) de la distribution dans l'échantillon, obtenue en appliquant la méthode DE. L'estimation de la médiane des X' obtenue par interpolation est multipliée par $\left(Q_{75}/1\,000\right)$ pour obtenir une valeur sur l'échelle originale. Cette procédure équivaut simplement à diviser, dans chaque cellule d'estimation, la partie de l'échantillon original allant de 0 à Q_{75} en Z classes de largeur égale et en plaçant le reste de l'échantillon dans une classe qui, par conception, est beaucoup plus grande que les autres. Dans le cas des données sur les logements à asymétrie fortement positive, cette transformation donne de bons résultats pour l'estimation de la médiane, parce que celle-ci est éloignée du paramètre d'échelle Q_{75}. Cependant, elle ne permet pas d'estimer le 80^{e} ni le 90^{e} décile. Donc, si nous voulons poursuivre en utilisant une méthode d'interpolation, nous devons considérer d'autres transformations.

L'approche la plus simple consiste à utiliser la méthode des classes dépendantes des données originale avec un paramètre d'échelle plus élevé, c'est-à-dire en utilisant toute valeur de centile supérieure à 90 %. Nous utilisons le 95^{e} centile comme facteur d'échelle et dans la suite de l'exposé, nous donnons à cette méthode le nom de « méthode C95 ».

La méthode C95 crée des distributions uniformes dans la majorité des classes, mais pose encore des problèmes à l'extrémité supérieure de la distribution pour deux raisons. Premièrement, la dernière classe ne contient que 5 % de la distribution de l'échantillon et les valeurs dans cette classe sont généralement très différentes. Deuxièmement, la procédure de groupement par classe dépendante des données requiert que chaque décile soit « éloigné » de la grande classe finale; sinon, les estimations des déciles présentent la même instabilité que celles obtenues en utilisant la « méthode DE ». Malheureusement, la classe contenant le 90^{e} centile est souvent proche de la classe finale lorsqu'on utilise un paramètre d'échelle de 95 %.

Pour résoudre ce second problème, nous considérons une autre approche de groupement des données par classe, nommée « méthode C75 ». Nous créons pour cela deux jeux de classes par cellule d'estimation,

chacune avec différentes largeurs au-dessus et en dessous de la valeur Q_{75} de la cellule. Cette tâche requiert deux transformations linéaires distinctes par cellule d'estimation, données par

$$X_i' = X_i \times \frac{1\,000}{X_{75}} \text{ quand } X_i < X_{75}$$

$$X_i'' = \left(X_i - X_{75}\right) \times \frac{1\,000}{\left(X_{100} - X_{75}\right)}$$

quand $X_i \geq X_p$ et X_{100} = valeur maximale dans l'échantillon.

Les X_i' sont alors placés dans Z classes de longueur égale, et les X_i'', dans K classes de longueur égale, où $Z \neq K$. L'interpolation est exécutée indépendamment pour chaque décile, en appliquant la transformation inverse appropriée à chaque décile interpolé. Cette procédure fait en sorte que les estimations de la médiane concordent exactement avec celles obtenues par la procédure courante.

La troisième approche d'interpolation que nous considérons comprend des hypothèses paramétriques au sujet des caractéristiques. Souvent, les données économiques suivent approximativement une loi log-normale (par exemple Steel et Fay 1995). La méthode de groupement par classe normale (que nous désignons par « CN ») s'appuie sur les propriétés de la distribution normale appliquée aux données log-transformées pour obtenir des classes dépendantes des données. La technique de groupement par classe fait en sorte que les zones à forte probabilité possèdent une plus petite largeur de classe afin de limiter le nombre d'observations par classe et que les zones à faible probabilité possèdent une plus grande largeur de classe afin d'accroître le nombre d'observations par classe.

Dans la méthode CN, les données log-transformées sont centrées autour de la médiane d'échantillon pondérée, puis les données centrées sont réduites en se servant d'une estimation de l'écart-type de la population. Nous utilisons la médiane d'échantillon, parce qu'elle est plus robuste que la moyenne aux valeurs aberrantes. Naturellement, si les données suivent une loi normale, la moyenne et la médiane sont équivalentes. Étant donné une loi normale centrée réduite, où $\mu = 0$ et $\sigma = 1$, l'intervalle interquartile IQR est donné par

$$IQR = Q_3 - Q_1$$

$$IQR = \left(0{,}67449*\sigma\right) - \left(-0{,}67449*\sigma\right) = \sigma\left(0{,}67449 + 0{,}67449\right) = \sigma * 1{,}34898.$$

Nous avons estimé l'écart-type (sigma) comme étant le ratio $\sigma \approx IQR/1{,}34898$, où l'$IQR$ est obtenu à partir de la fonction de répartition empirique dans la cellule d'estimation. Pour normaliser les données, nous avons appliqué la transformation suivante

$$Y_i = \text{Log}\left(X_i\right) \; Y_i' = \frac{Y_i - Y_{\text{med}}}{\sigma_y} = \frac{Y_i - Y_{\text{med}}}{IQR_y/1{,}34898}$$

où

Y_{med} = la médiane d'échantillon log-transformée sur le domaine i;

IQR_y = l'intervalle interquartile d'échantillon log-transformé sur le domaine i.

De nouveau, les déciles d'échantillon et les intervalles interquartiles sont obtenus par la méthode DE. Si les données suivent une loi log-normale, Y_i' devrait suivre une loi normale centrée réduite, de sorte qu'environ 68,3 % des données se trouvent dans l'intervalle de plus ou moins un écart-type par rapport à la moyenne et 95,4 % des données se trouvent dans l'intervalle de plus ou moins deux écarts-types par rapport à la moyenne. En utilisant ces propriétés, nous avons réparti les Y_i' transformés entre les cinq zones distinctes et créé les 45 classes illustrées au tableau 2.1.

Tableau 2.1
Classes pour la transformation log-normale (méthode normale)

Zone	1	2	3	4	5
Intervalle	[Faible, -2)	[-2, -1)	[-1, 1)	[1, 2)	[2, Élevé]
Pourcentage dans la zone	2,3	13,6	68,2	13,6	2,3
Classes	1	6	31	6	1
Pourcentage moyen d'unités échantillonnées par classe	2,3	2,3	2,2	2,3	2,3

Il existe quatre largeurs de classe différentes avec à peu près le même pourcentage moyen d'unités échantillonnées par classe. Nous appliquons la méthode de Woodruff aux données transformées pour obtenir les déciles, puis nous prenons l'exponentielle de ces estimations des déciles pour obtenir les valeurs sur l'échelle originale. Contrairement aux méthodes de changement d'échelle linéaires présentées plus haut, cette approche induit un biais d'estimation supplémentaire causé par la transformation puissance. Il aurait peut-être été possible de corriger ce biais de transformation au moyen d'un développement en série de Taylor, comme l'a suggéré un examinateur, mais nous n'avons pas envisagé cette approche.

2.2 Estimation de la variance

La méthode de rééchantillonnage MHS (dite « méthode de Fay ») est un « compromis » entre le jackknife stratifié et la méthode BRR (Fay 1989). Rao et Shao (1999) démontrent que l'estimateur MHS de la variance est asymptotiquement convergent pour les statistiques lisses, telles que les estimateurs par le ratio, et pour les statistiques non lisses, telles que les quantiles d'échantillon estimés par la méthode DE décrites à la section 2.1. Dans leur article, ils n'étendent pas cette propriété aux estimations interpolées des déciles, bien qu'il découle de leur démonstration que ces estimations de variance devraient être convergentes à mesure que la largeur des classes s'approche de 1. Comme la méthode BRR, la méthode MHS s'appuie sur une matrice de Hadamard pour former les répliques, mais utilise des poids de rééchantillonnage de 1,5 et 0,5 au lieu des valeurs de 2 et 0 utilisées dans la méthode BRR. La formule MHS pour l'estimation de l'erreur-type de toute estimation $\hat{\theta}$ est

$$\hat{S}\left(\hat{\theta}\right) = \sqrt{\frac{4}{R} * \sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}_0\right)^2} \qquad (2.2)$$

où $\hat{\theta}_r$ est la r^{e} estimation répétée $(r = 1, 2, \ldots, R)$ et $\hat{\theta}_0$ est l'estimation pour l'échantillon complet. La somme des carrés des termes d'erreur est rajustée par un facteur de $4 = 1/(1 - 0,5)^2$ pour éviter la présence d'un biais négatif dans l'estimation de la variance (Judkins 1990).

3 Analyse empirique

3.1 Plan de sondage de la SOC

Comme il est mentionné dans l'introduction, notre étude a été motivée par une demande des utilisateurs des données de la Survey of Construction (SOC). La SOC est une enquête nationale conçue pour recueillir des renseignements sur les caractéristiques des logements résidentiels neufs aux États-Unis. Les données de la SOC sont utilisées pour produire trois indicateurs économiques importants publiés chaque mois par le U.S. Census Bureau, à savoir les mises en chantier, les logements achevés et les ventes de logements (logements unifamiliaux seulement). En outre, le programme de la SOC publie des estimations mensuelles, trimestrielles et annuelles pour diverses caractéristiques des logements, telles que le prix de vente, le prix de vente moyen par pied carré de logements vendus, le délai entre l'obtention du permis de bâtir et la mise en chantier, et le délai entre la mise en chantier et l'achèvement de la construction du logement. Dans le présent article, nous examinons deux caractéristiques importantes des logements qui sont publiées annuellement, à savoir le prix de vente et le prix par pied carré de logements vendus. Pour les deux caractéristiques, les données sont recueillies mensuellement à mesure qu'elles deviennent disponibles auprès des constructeurs. À l'heure actuelle, les estimations moyennes et médianes des deux caractéristiques sont incluses dans les rapports annuels; les prix de vente moyens et médians des logements vendus sont également publiés mensuellement.

L'univers de la SOC comprend deux sous-populations : les régions où un permis de bâtir est exigé et les régions où il ne l'est pas. Les régions qui exigent un permis de bâtir sont couvertes par la *Survey of the Use of Permits* (SUP) et celles qui ne délivrent pas de permis de bâtir sont couvertes par la *Nonpermit Survey* (NP). La grande majorité de l'échantillon provient de la SUP. Les deux populations sont échantillonnées à partir des mêmes UPE, mais les échantillons sont indépendants aux degrés d'échantillonnage ultérieurs. Puisque la majorité de l'échantillon de la SOC est constitué de permis échantillonnés, nous nous concentrons entièrement sur le volet SUP de la SOC dans notre étude.

L'échantillon de la SUP est sélectionné en trois étapes. L'échantillon de premier degré est un sous-échantillon sélectionné avec probabilité proportionnelle à la taille (PPT) des unités primaires d'échantillonnage (UPE) définies dans le plan de sondage de la Current Population Survey (CPS) de 2000 et dont le tirage a lieu tous les dix ans. Les UPE de la CPS sont des zones de territoire telles que les comtés ou les cantons. L'échantillon de deuxième degré de la SUP est un échantillon systématique stratifié de localités délivrant des permis et comprises dans les UPE, et il est également tiré tous les dix ans. L'échantillonnage de troisième degré est effectué mensuellement dans chacune des localités émettrices de permis échantillonnées. Chaque mois, les agents de terrain dressent les listes complètes des nouveaux permis de bâtir d'après les données des bureaux des permis dans les localités échantillonnées et sélectionnent un échantillon systématique de permis de bâtir. Des taux d'échantillonnage sont appliqués aux bureaux des permis de manière à obtenir un taux d'échantillonnage global d'un sur cinquante pour les

immeubles comptant d'une à quatre unités. Les immeubles plus grands comptant cinq unités et plus sont inclus avec certitude (c'est-à-dire qu'ils sont autoreprésentatifs).

Le programme de la SOC utilise la méthode de rééchantillonnage MHS pour estimer les variances à l'aide d'une matrice de Hadamard de dimensions 200×200, en attribuant un total de 198 lignes aux groupes répétés. Puisque la SOC n'est pas réalisée selon un plan de sondage à deux UPE par strate, une approche de regroupement de strates est adoptée pour créer les répliques : voir Thompson (1998) pour des renseignements détaillés.

3.2 Analyse empirique des données de la SOC

Nos analyses empiriques portent sur des données de la SOC recueillies de 2006 à 2009. Dans le cas de la SOC, on utilise la méthode de groupement par classe dépendante des données décrite à la section 2.1 avec 41 classes pour produire les estimations de la médiane. Nous utilisons 51 classes pour la méthode C95, avec 95 % de l'échantillon réparti sur 50 classes de taille égale; la 51^{e} classe contient toutes les données supérieures au 95^{e} centile. Pour la méthode C75, nous utilisons 40 classes de taille égale pour toutes les valeurs inférieures au 75^{e} centile et 10 classes de taille égale pour toutes les valeurs comprises entre le 75^{e} centile et la valeur maximale de la distribution d'échantillon. Enfin, pour la méthode CN, nous utilisons un total de 45 classes.

Pour le prix de vente et le prix de vente par pied carré, toutes les estimations des déciles obtenues par les méthodes C75, C95 et DE étaient assez comparables du 10^{e} au 70^{e} décile; les déciles CN étaient généralement un peu plus élevés que leurs homologues pour les autres méthodes. Cependant, les estimations pour les 80^{e} et 90^{e} déciles par la méthode C75 étaient systématiquement plus élevées que celles obtenues par les trois autres méthodes. L'explication est simple : dans le cas de la méthode C75, les 80^{e} et 90^{e} déciles sont tous deux presque toujours situés dans la même classe. Les distributions des caractéristiques sont toutes deux relativement asymétriques. Par conséquent, la majorité des 25 % supérieurs de l'échantillon est contenue dans la classe la plus proche du 75^{e} centile.

Les courbes des estimations des déciles obtenues par chaque méthode étaient très cohérentes tant aux niveaux national que régional. Contrairement aux estimations des caractéristiques, les comparaisons des estimations de variance révèlent moins de tendances claires. Les estimations de variance pour le 80^{e} décile obtenues par la méthode C75 étaient considérablement plus grandes que celles obtenues par les trois autres méthodes et celles pour le 90^{e} décile étaient, de même, considérablement plus petites. En ce qui concerne les trois autres méthodes, les variances pour la méthode CN avaient tendance à être plus faibles que les variances correspondantes obtenues par les méthodes C95 et DE; ces différences étaient plus prononcées pour les estimations des déciles du prix de vente par pied carré.

Trois des quatre méthodes considérées ont donné des ensembles comparables d'estimations des déciles. La méthode C75 s'est avérée impossible à appliquer en raison la forte asymétrie des distributions considérées; nous n'avons tout simplement pas pu trouver une « largeur de classe » adéquate pour le quartile supérieur des données. Donc, à la suite de l'évaluation empirique des données, notre jeu de méthodes d'estimation proposées a été réduit à trois. Cependant, même si ces trois méthodes donnaient des estimations fort semblables, les estimations de variance étaient clairement différentes. Par conséquent, nous avons décidé de réaliser une étude en simulation pour évaluer les propriétés statistiques des divers estimateurs sur des échantillons répétés.

3.3 Étude en simulation

3.3.1 Modélisation et procédure de sélection de l'échantillon

Pour notre simulation, nous avons élaboré une population qui imite les qualités de la majorité de la population de la SUP. Autrement dit, nous avons créé des populations stratifiées de bureaux des permis (UPE), à partir desquelles nous avons sélectionné des échantillons de permis (USE). La création de conditions de simulation aussi complexes offrait plusieurs avantages. D'un point de vue purement pratique, ces conditions étaient avantageuses pour l'interprétation des résultats empiriques présentés à la section 3.2. Fait plus important, les travaux de recherche antérieurs menés par Thompson et Sigman (2000) avaient donné des résultats presque parfaits pour les médianes interpolées dépendantes des données sur une population simulée qui ne comprenait pas de mise en grappes; la distinction entre les propriétés statistiques de chaque méthode n'est devenue évidente qu'après avoir intégré la mise en grappes dans le plan de sondage.

Nous avons utilisé une approche « ascendante » pour créer des données de population simulées valables. Premièrement, nous avons modélisé des populations multivariées de données sur les permis dans chaque région. Ensuite, nous avons combiné les données sur les permis modélisées pour former des « grappes » représentant les bureaux des permis (les unités primaires d'échantillonnage). En guise de protection contre les erreurs de spécification du modèle, nous avons créé indépendamment deux populations artificielles de données sur les permis pour lesquelles chaque enregistrement de permis contenait le prix de vente et le prix par pied carré, puis nous avons modélisé une population à distribution log-normale dans chaque région en utilisant l'algorithme décrit dans Lienhard (2004) et l'autre en utilisant un algorithme SIMDATA non paramétrique (Thompson 2000) dans chaque région.

En général, les données sur les permis modélisées dans la population non paramétrique donnent une meilleure représentation des niveaux correspondants à chaque décile du jeu de données d'apprentissage. En revanche, les distributions des permis dans la population log-normale sont assez lisses, tandis que les distributions dans la population non paramétrique sont « irrégulières » et présentent des discontinuités (escalier) importantes entre les estimations ponctuelles adjacentes. Le tableau 3.1 donne les principaux centiles et les moyennes des populations simulées pour les deux caractéristiques modélisées, en les comparant aux valeurs empiriques provenant des données de la SOC (figurant dans la colonne intitulée « Données d'apprentissage »).

Notre méthode de simulation comprenait la création de populations de permis (les USE), puis la création de grappes de premier degré artificielles (bureaux des permis) et enfin leur stratification. La création de grappes en deux étapes et le processus de stratification décrits plus bas reposent sur l'hypothèse que les permis dans une strate de population sont hétérogènes en ce qui concerne le prix de vente et le prix par pied carré, et que les permis émis par un même bureau des permis ont des caractéristiques de logement similaires. Ces critères ont été fournis par les spécialistes du domaine, qui pensent que la modélisation des variations entre bureaux des permis est plus réaliste que la modélisation des variations à l'intérieur de ces bureaux. La population log-normale multivariée se prête mieux à cet exercice que la population non paramétrique; les éléments assignés dans chaque grappe ont tendance à être homogènes en raison de la distribution plus lisse.

Nous avons recouru à l'analyse discriminante pour grouper les permis simulés en strates disjointes. Après avoir appliqué la même fonction discriminante à chaque population de données sur les permis

simulées (log-normale et non paramétrique), nous avons regroupé les permis dans les strates pour former environ 14 000 bureaux des permis actifs par population. L'application de classification automatique a créé des bureaux des permis de taille variable à l'intérieur desquels les caractéristiques étaient homogènes.

Tableau 3.1
Statistiques des populations simulées et statistiques des données empiriques

			Population simulée		Données d'apprentissage (pondérées)
			Log-normale	Non paramétrique	
Prix de vente	Décile	1	74 747,74	94 323,27	95 000,00
		5	105 009,64	120 073,33	120 000,00
		10	126 492,43	136 009,85	140 000,00
		20	158 517,61	158 931,57	160 000,00
		30	186 927,58	181 941,12	180 000,00
		40	215 346,44	205 003,46	210 000,00
		50	245 502,91	230 188,69	230 000,00
		60	280 064,08	260 524,68	260 000,00
		70	322 790,68	301 374,90	300 000,00
		80	381 501,24	359 138,64	360 000,00
		90	482 209,62	488 517,45	490 000,00
		95	586 359,10	622 185,68	630 000,00
		99	855 983,47	1 167 704,85	1 300 000,00
		Moyenne	283 085,94	287 134,16	290 000,00
			Population simulée		Données d'apprentissage (pondérées)
			Log-normale	Non paramétrique	
Prix de vente par pied carré	Décile	1	32,76	32,68	35,00
		5	43,11	47,27	47,00
		10	49,86	54,70	55,00
		20	59,26	63,74	64,00
		30	67,22	70,71	72,00
		40	74,91	77,14	78,00
		50	83,11	83,60	84,00
		60	92,42	90,60	91,00
		70	103,75	98,70	99,00
		80	119,61	109,57	110,00
		90	147,62	130,02	130,00
		95	178,94	155,08	160,00
		99	265,10	262,68	270,00
		Moyenne	93,58	94,90	96,00

Nous avons sélectionné 5 000 échantillons répétés dans notre population simulée en utilisant une version très simplifiée du plan de sondage de la SOC décrit plus haut. Le premier degré d'échantillonnage correspond à la sélection des bureaux des permis. Les 250 bureaux les plus grands à l'échelle des États-Unis ont été sélectionnés avec une probabilité de un (certitude), de manière que chaque échantillon répété contienne les mêmes bureaux autoreprésentatifs. Donc, nous avons sélectionné avec probabilité proportionnelle à la taille un échantillon de deux bureaux des permis non autoreprésentatifs dans chaque strate, en attribuant à chaque bureau son propre poids de bureau des permis.

Au deuxième degré d'échantillonnage, nous avons sélectionné des enregistrements de permis dans chacun des bureaux des permis échantillonnés. Nous avons sélectionné un échantillon aléatoire simple

(EAS) de permis dans chaque bureau, avec un taux d'échantillonnage de bureau obtenu en divisant le poids du bureau des permis par 50, afin d'obtenir un échantillon global de permis de 1 sur 50 (si le poids du bureau de permis est supérieur à 50, tous les permis émis par ce bureau sont échantillonnés). Pour calculer le poids final de chaque enregistrement, nous avons multiplié le poids du bureau des permis par le poids du permis. Les permis sélectionnés dans les bureaux échantillonnés avec certitude varient dans chaque échantillon répété en raison de l'échantillonnage indépendant, à moins que le bureau contienne plus de 50 permis.

Enfin, dans chaque échantillon, nous avons assigné des permis ou des bureaux des permis aux répliques. Nous n'avons pas imité l'application à demi-échantillon partiellement équilibré de la SOC décrite à la section 3.1. Le regroupement des strates induit un biais dans les estimations de variance. Pour éliminer cette composante du biais de notre simulation, nous avons utilisé un plan à deux UPE par strate et 572 répliques (c'est-à-dire une matrice de Hadamard de dimensions 572×572), de sorte que chacun des 250 bureaux autoreprésentatifs et chacune des 321 strates non autoreprésentatives (paires de bureaux des permis échantillonnés) reçoive sa propre ligne dans la matrice de Hadamard. En imitant la méthode de production de la SOC, chaque bureau autoreprésentatif a été traité comme une « pseudo-strate », et les panels de répliques ont été obtenus en répartissant aléatoirement les permis dans chaque bureau.

Dans chaque échantillon, nous avons calculé un jeu d'estimations aux échelles nationale et régionale pour les trois méthodes d'estimation des déciles prises en considération dans chaque réplique et nous avons calculé les estimations de la variance par la méthode MHS pour chaque décile en utilisant (2.2) avec $R = 572$.

3.3.2 Méthodologie d'évaluation

L'étude en simulation a pour but d'examiner les propriétés statistiques de chaque méthode d'estimation des déciles et les estimations de variance connexes sur des échantillons répétés. Soit ζ_m^d l'estimation du décile d calculée par la méthode $m\,(m = \text{DE}, \text{C95}, \text{CN})$.

Pour évaluer les propriétés d'*estimation* de la méthode m pour le décile d sur des échantillons répétés, nous avons calculé le biais relatif et l'erreur quadratique moyenne empirique. Le biais relatif de chaque estimation de décile pour chaque méthode d'estimation est donné par

$$\hat{B}(\xi_m^d) = 100 \times \left[\overline{\xi}_m^d \big/ \xi_p^d\right] - 1$$

où ξ_{ms}^d est l'estimation du décile d par la méthode m dans l'échantillon s, $\overline{\xi}_m^d$ est la moyenne sur les 5 000 échantillons, et ξ_p^d est le décile de population [les mesures d'évaluation pour les estimations et les estimations de la variance pour le domaine i (Nord-Est, Midwest, Sud et Ouest dans l'étude en simulation présentée à la section 4) peuvent être obtenues sur demande auprès des auteurs, mais sont omises par souci de concision].

L'erreur quadratique moyenne (EQM) empirique de chaque estimation de décile pour chaque méthode d'estimation est donnée par

$$\text{EQM}\left(\xi_m^d\right) = \frac{1}{5\,000}\sum_{s=1}^{5\,000}\left(\xi_{ms}^d - \xi_p^d\right)^2 = \frac{1}{5\,000}\sum_{s=1}^{5\,000}\left(\xi_{ms}^d - \overline{\xi}_m^d\right)^2 + (\overline{\xi}_m^d - \xi_p^d)^2$$

$$\text{EQM}\left(\xi_m^d\right) = \hat{\sigma}\left(\xi_m^d\right) + \hat{B}^2\left(\xi_m^d\right).$$

Pour évaluer les propriétés d'*estimation de la variance* de la méthode d'estimation pour le décile d sur les échantillons répétés, nous avons calculé les statistiques suivantes :

Biais relatif de la variance $100 \times \left[\hat{\bar{v}}\left(\xi_m^d\right) \Big/ \mathrm{EQM}\left(\xi_m^d\right) \right] - 1$ où $\hat{\bar{v}}\left(\xi_m^d\right)$ est l'estimation de *variance* moyenne du décile d par la méthode m sur 5 000 échantillons, c'est-à-dire $\hat{\bar{v}}\left(\xi_m^d\right) = \hat{v}\left(\xi_{ms}^d\right) \Big/ 5\,000$. Dans notre étude de cas, les estimations de variance dans chaque échantillon $s, \hat{v}\left(\xi_{ms}^d\right)$, sont des estimations de variance par la méthode des demi-échantillons répétés modifiée décrite à la section 3.3.1.

Stabilité de l'estimation de la variance $\sqrt{\sum_{s=1}^{5\,000} \left(\hat{v}\left(\xi_{ms}^d\right) - \mathrm{EQM}\left(\xi_m^d\right)\right)^2 \Big/ 5\,000} \Big/ \mathrm{EQM}\left(\xi_m^d\right)$.

Taux de couverture (TC) = la proportion de l'intervalle de confiance à 90 % pour une méthode donnée qui contient le décile de population réel ξ_p^d.

La stabilité de la variance est une mesure de la variance des estimations de la variance. Idéalement, tant le biais relatif que les mesures de stabilité devraient s'approcher de zéro. Les taux de couverture montrent l'effet combiné de l'estimation et de l'estimation de la variance sur l'inférence.

3.3.3 Résultats de l'étude en simulation

Les sections qui suivent résument les résultats de notre étude en simulation, présentés dans des graphiques illustratifs (les tableaux peuvent être obtenus sur demande auprès des auteurs).

3.3.3.1 Propriétés d'estimation de chaque méthode

La figure 3.1 représente les biais relatifs des estimations des déciles à l'échelle nationale selon la méthode d'estimation pour la population log-normale. Rappelons que des estimations sans biais auront un biais relatif nul indiqué par l'asymptote horizontale grise sur chaque figure. La comparaison visuelle des niveaux de biais doit se faire avec prudence, car les graphiques pour les deux caractéristiques étudiées pourraient ne pas être à la même échelle.

La méthode DE produit les estimations des déciles les moins biaisées tant pour le prix de vente que pour le prix par pied carré. Cela dit, les biais des estimations des déciles pour les deux caractéristiques obtenus en utilisant les méthodes C95 et CN sont triviaux. Les biais les plus importants s'observent au 10^{e} centile et au 90^{e} centile, c'est-à-dire près des queues de la distribution, où l'échantillon est en principe moins stable. Bien que les estimations des déciles par la méthode DE soient moins biaisées que celles obtenues par les méthodes C95 et CN, elles sont moins précises. En général, les déciles par la méthode C95 ont l'EQM la plus faible parmi les trois méthodes concurrentes, quoique dans de nombreux cas, l'écart entre les EQM des méthodes C95 et CN soit négligeable.

La figure 3.2 représente le biais relatif des estimations des déciles à l'échelle nationale selon la méthode d'estimation dans la population non paramétrique. Pour le prix par pied carré, les courbes de biais suivent les mêmes tendances que plus haut, de même que celles des EQM. Par contre, pour le prix de vente, les tendances du biais et de l'EQM sont différentes. Ici, les estimations par la méthode DE sont les moins biaisées, mais le biais le plus important a lieu à la médiane (0,005). Il en est de même pour les deux méthodes d'interpolation, les médianes C95 et CN ayant un biais relatif positif de sept dixièmes de pour

cent. Pour les 50e et 60e déciles, l'EQM de l'estimation C95 est un peu plus grande que celle des autres estimations correspondantes, ce qui reflète l'effet de cet estimateur.

Dans le cas de la population non paramétrique, certains biais sont suffisamment grands pour être préoccupants, surtout pour l'estimation de la médiane. Cela dit, la population log-normale semble imiter plus fidèlement les données réelles de la SOC, si bien que les résultats non paramétriques ne sont pas nécessairement un reflet de la « réalité » de la SOC. Ces résultats traduisent l'effet du terme de biais constant causé par l'interpolation dans les estimations des déciles.

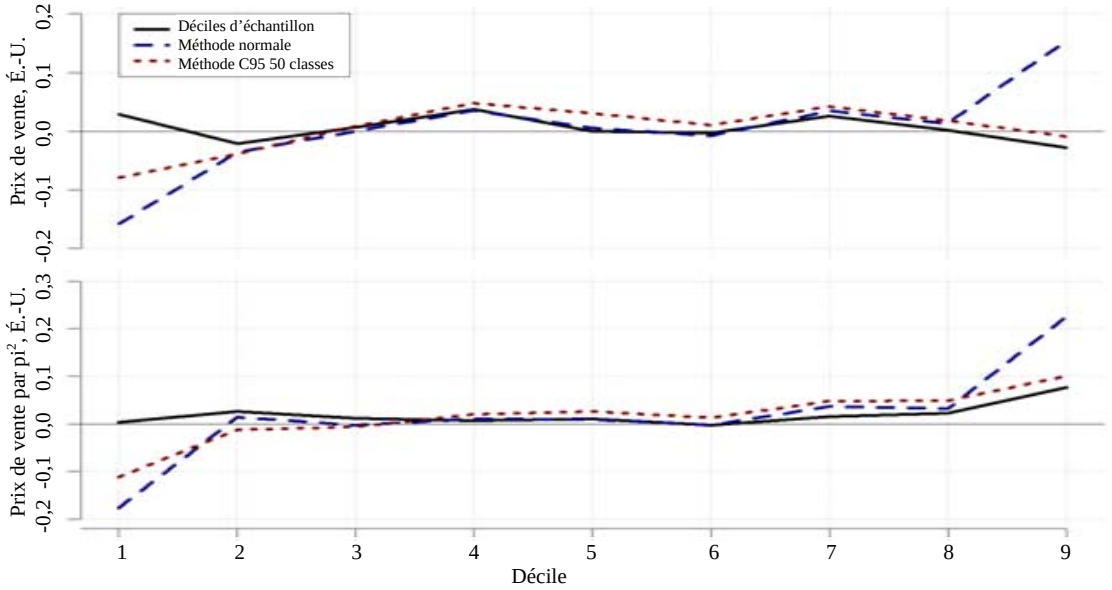


Figure 3.1 Biais relatif des estimations du prix de vente et du prix de vente par pied carré pour la population log-normale (exprimé en pourcentage)

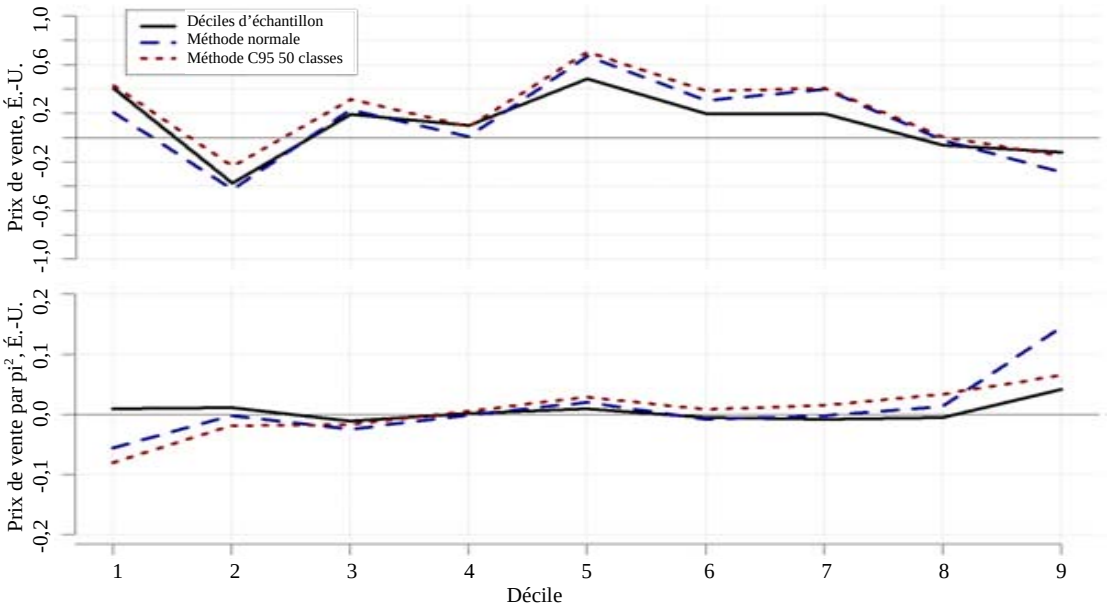


Figure 3.2 Biais relatif des estimations du prix de vente et du prix de vente par pied carré pour la population non paramétrique (exprimé en pourcentage)

Dans l'ensemble, les EQM suivent des tendances similaires pour les deux populations et les deux caractéristiques. Les EQM minimales s'observent autour des déciles centraux. Les déciles de la queue inférieure de la distribution ont des EQM légèrement plus grandes et les EQM des déciles de la queue supérieure augmentent rapidement.

3.3.3.2 Propriétés d'estimation de la variance (par rééchantillonnage MHS) pour chaque méthode

Comme le montre la figure 3.3, *tous* les biais relatifs de variance des estimations de variance MHS dans la population log-normale sont positifs quel que soit l'estimateur, et les estimations de variance DE sont celles dont le biais est le plus important. Les méthodes C95 et CN donnent des biais relatifs similaires pour toutes les caractéristiques, le biais étant plus faible dans tous les cas que pour la méthode DE. Dans l'ensemble, ce sont les estimations de variance C95 qui sont les moins biaisées. Notons que, pour le prix de vente, *toutes* les estimations de variance présentent un biais positif avec tous les estimateurs; pour économiser l'espace, l'axe des y commence à 5 %. Les mêmes mises en garde concernant les comparaisons visuelles que celles énoncées à la section 3.3.3.1 s'appliquent aux figures de la présente section.

Les estimations de la variance sous la méthode DE sont de loin les moins stables pour les deux caractéristiques. Ce résultat est prévisible, puisque les estimations de variance sous interpolation bénéficient du lissage. Des deux méthodes d'interpolation, la méthode C95 est celle qui donne les estimations de variance les plus stables pour tous les déciles, sauf pour quelques-uns situés dans la queue supérieure de la distribution. Les variances plus stables pour les déciles supérieurs sous la méthode CN découlent vraisemblablement de l'utilisation des propriétés d'une distribution normale pour obtenir des pourcentages égaux de l'échantillon dans chaque classe.

Aucune des trois méthodes étudiées n'a donné des taux de couverture de 90 % pour le prix de vente ni pour le prix par pied carré (figure 3.4). La plupart des taux de couverture sont légèrement anticonventionnels (sous l'asymptote horizontale de 90 %), aucune méthode d'estimation des déciles ne paraît offrir de meilleures propriétés de couverture que les autres.

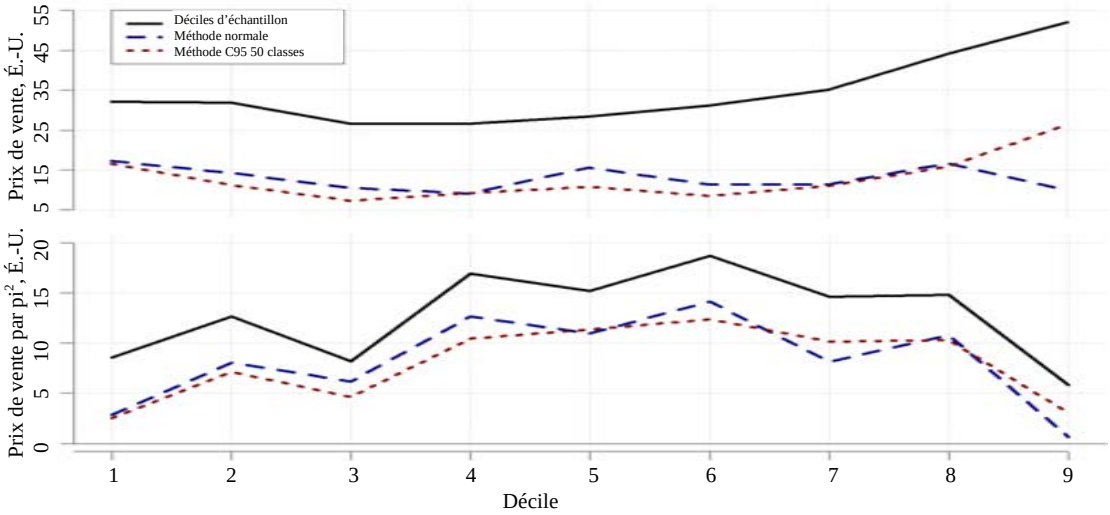


Figure 3.3 Biais relatif de la variance pour le prix de vente et le prix de vente par prix carré pour la population log-normale (exprimé en pourcentage)

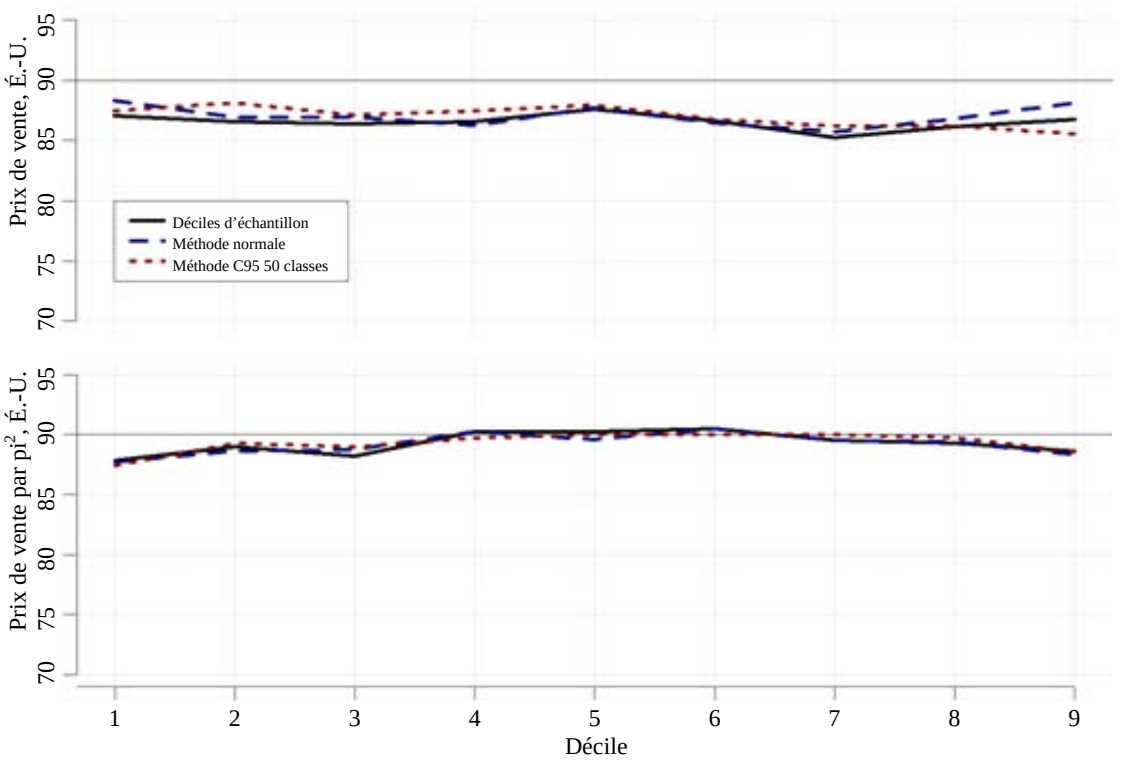


Figure 3.4 Taux de couverture pour le prix de vente et pour le prix par pied carré (population log-normale)

La figure 3.5 illustre les biais relatifs des variances MHS obtenues pour la population non paramétrique. Ces biais relatifs sont généralement beaucoup plus faibles dans le cas des méthodes d'interpolation C95 et CN que dans le cas de la méthode DE, et la méthode C95 a tendance à produire des estimations de variance moins biaisée pour les deux caractéristiques. Le biais relatif du prix de vente ne suit pas la même courbe pour la population non paramétrique que pour la population log-normale. Les biais relatifs sous la méthode DE sont toujours positifs et plus élevés que dans le cas des deux méthodes d'interpolation. Ces dernières produisent des résultats différents selon la population de prix de vente. Dans le cas de la population non paramétrique, de nombreux biais relatifs sont négatifs au lieu d'être tous positifs. Pour le prix de vente par pied carré, les biais relatifs présentent la même tendance que pour la population log-normale, de grands biais positifs étant observés pour la méthode DE et des biais positifs similaires plus faibles, pour les deux méthodes d'interpolation.

La stabilité des estimations de variance non paramétrique concordait bien avec celle obtenue pour la population log-normale, à quelques différences près pour les estimations des déciles du prix de vente. Dans le cas du prix de vente, les estimations de la stabilité pour la méthode DE restent plus grandes que pour les deux méthodes d'interpolation, mais suivent une courbe plus irrégulière. Dans le cas de la méthode CN, on observe pour le 40^{e} décile une grande estimation de la stabilité qui ne suit pas la tendance prévue.

Pour le prix de vente par pied carré, les taux de couverture affichent la même tendance que dans le cas de la population log-normale (figure 3.6). Toutefois, pour le prix de vente, la courbe des taux est plus variable.

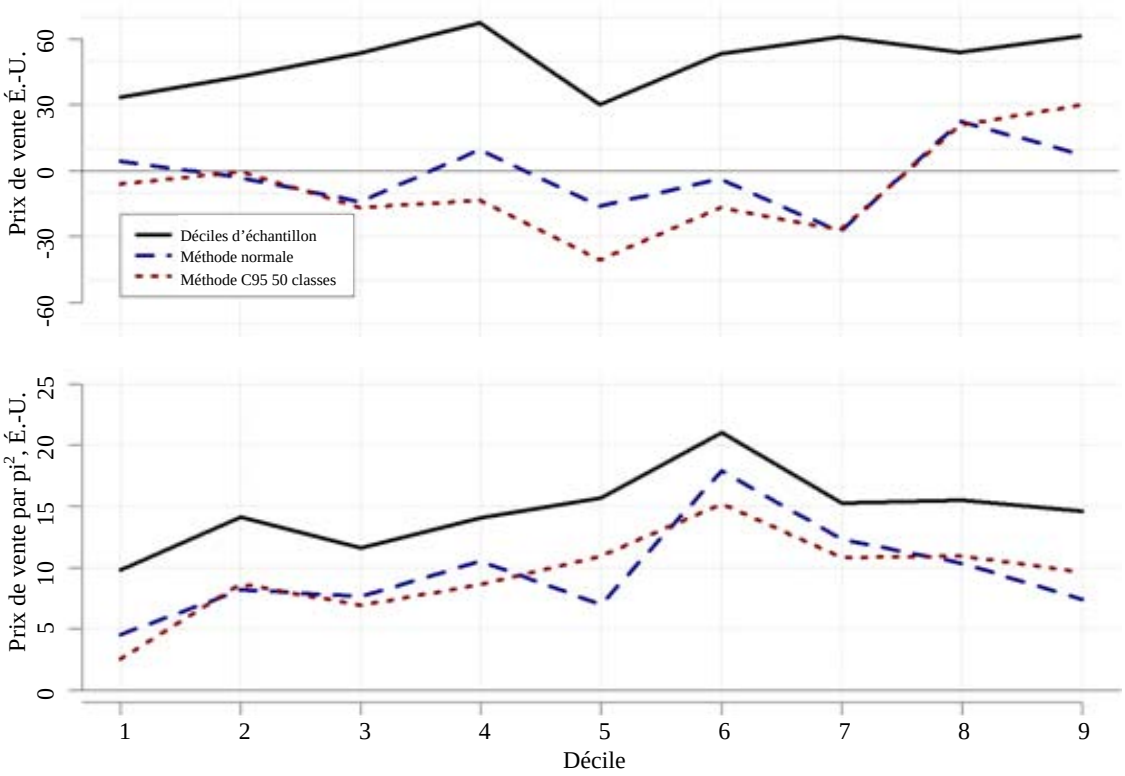


Figure 3.5 Biais relatif de la variance du prix de vente et du prix de vente par pied carré (population non paramétrique)

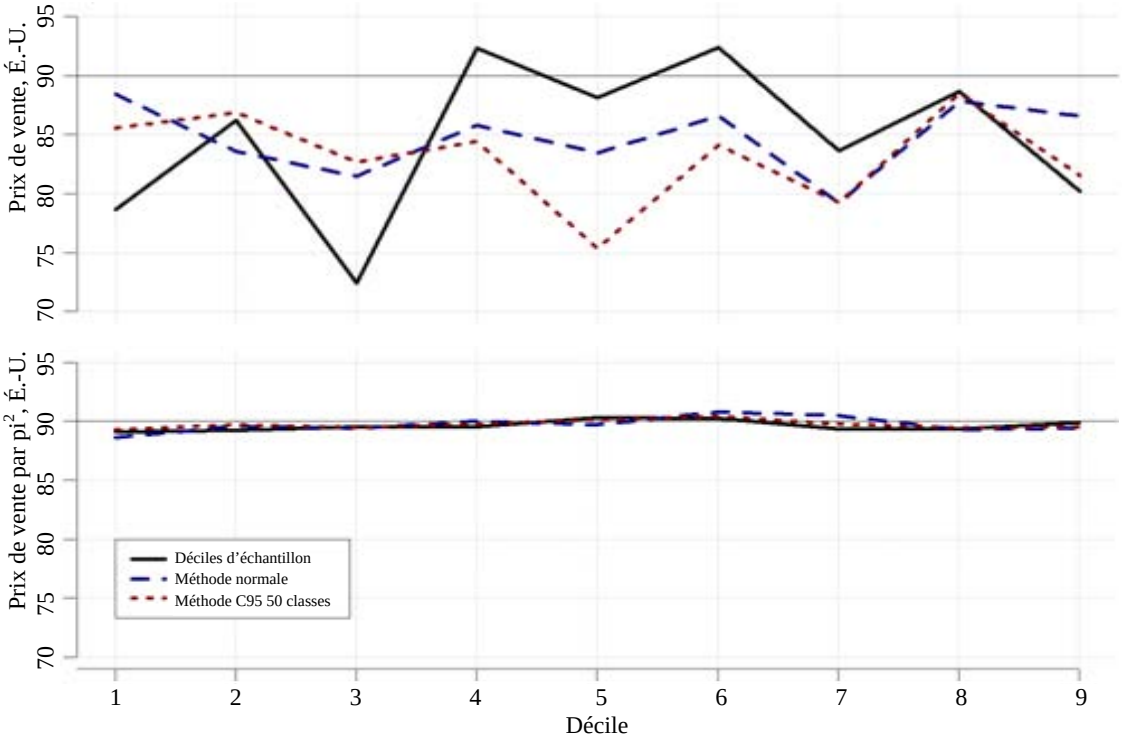


Figure 3.6 Taux de couverture du prix de vente et du prix de vente par pied carré (population non paramétrique)

3.3.3.3 Simulations supplémentaires pour évaluer les effets de taille de classe

Dans l'ensemble, les propriétés statistiques des estimations C95 et des estimations de variance obtenues pour la population log-normale (pour les deux caractéristiques) et pour la population non paramétrique pour le prix de vente par pied carré sont assez prometteuses. Cependant, aucune des méthodes examinées ne possède des propriétés qui sont près d'être aussi solides pour le prix de vente dans la population non paramétrique. Cette constatation est préoccupante, malgré les mises en garde faites plus haut au sujet de la modélisation de la population non paramétrique.

Dans la population non paramétrique, les classes de prix de vente proches de la médiane contenaient un plus grand nombre d'observations que celles se trouvant dans la queue de la distribution [remarque : cela est le cas pour les méthodes C95 et CN]. Ces grandes classes peuvent lisser exagérément la distribution, donnant lieu à des estimations très stables. Le lissage excessif se manifeste dans la méthode d'estimation de la variance par rééchantillonnage sous forme d'une sous-estimation due au manque de variabilité entre les estimations répétées pour les déciles « du milieu ». Inversement, comme prévu, la méthode DE produit des estimations instables tout au long de la distribution et, par conséquent, surestime la variance (biais positif).

Le but de la transformation des données avant le groupement par classe est d'obtenir des distributions uniformes dans les classes. Pour le prix de vente, ni l'une ni l'autre transformation ne donne une distribution uniforme dans les classes, ce qui donne lieu à un biais d'interpolation non négligeable, qui à son tour affecte les estimations de l'EQM.

Afin de mieux comprendre comment la méthode d'estimation influe sur la méthode d'estimation de la variance, rappelons que nos estimations de la variance sont évaluées d'après l'erreur quadratique moyenne (EQM) obtenue sous la méthode d'estimation. La méthode DE donne des estimations essentiellement sans biais, mais au prix d'une variance importante et instable. Le recours à l'interpolation réduit la variance d'échantillonnage et améliore sa stabilité, mais peut accroître considérablement le terme du carré du biais de l'EQM.

En dernière analyse, la méthode C95 est celle qui donne les résultats les plus prometteurs pour la plupart des caractéristiques. Elle suscite néanmoins plusieurs préoccupations quant au biais de l'estimation médiane pour le prix de vente dans une population non paramétrique. Pour aborder ces préoccupations, nous avons exécuté des simulations supplémentaires sur les deux populations et les deux caractéristiques, en utilisant la méthode C95 avec 50 classes, 75 classes et 100 classes.

Dans la plupart des cas, l'utilisation de 75 classes avec la méthode C95 réduit généralement le biais de l'estimation et l'EQM sans nuire aux propriétés des estimations de variance. Il s'agit définitivement d'un exercice d'équilibre. À mesure que le nombre de classes augmente, les statistiques d'évaluation correspondantes commencent à ressembler à celles obtenues avec la méthode DE. Cela améliore les estimations interpolées dans les cas où les déciles calculés par la méthode DE ont de meilleures propriétés statistiques. Cependant, augmenter le nombre de classes a un effet nuisible sur les propriétés statistiques des estimations des déciles et des estimations de variance dans les cas où la méthode C95 avec 50 classes donnait des estimations ayant un biais plus faible et des estimations de variance plus stables.

4 Conclusion

Dans Thompson et Sigman (2000), la constatation fondamentale était que des méthodes d'interpolation peuvent être utilisées pour produire des estimations stables de la *médiane* pour des échantillons tirés de populations présentant une asymétrie positive, mais que l'efficacité de l'interpolation dépend fortement de la largeur des classes ainsi que de leur localisation dans l'échantillon. La contribution principale de l'étude était l'élaboration d'une approche de groupement par classe dépendante des données en utilisant la distribution dans chaque cellule d'estimation individuelle.

Notre approche en vue d'établir une méthode d'estimation des déciles pour des échantillons complexes provenant d'une population à asymétrie positive vient étoffer ces premiers résultats, en reconnaissant à la fois que le groupement par classe dépendante des données est une nécessité et que la méthode de groupement par classe choisie doit tenir compte de l'asymétrie positive de la distribution pour faciliter l'obtention du jeu complet d'estimations des déciles. Nous avons pris en considération trois méthodes d'interpolation, suivant chacune une approche différente pour résoudre le problème du petit nombre de données au 90^{e} décile que posent les distributions asymétriques. Notre analyse empirique a montré que toutes les approches étudiées donnaient des jeux complets d'estimations des déciles ayant des propriétés statistiques raisonnables, du moins pour la Survey of Construction. Cependant, les propriétés des estimations de variance par la méthode MHS correspondante n'étaient pas aussi bonnes et présentaient des tendances différentes. À l'échelle des États-Unis, les résultats de nos simulations donnent la preuve que l'utilisation de la transformation C95 et de 75 classes dans une population simulée produit des estimations des déciles et des estimations de la variance dont les propriétés statistiques sont systématiquement bonnes, pour ce qui est du biais de l'estimation, de l'EQM, du biais et de la stabilité des estimations de variance, mais qui atteignent rarement un taux de couverture de 90 %.

Naturellement, il est beaucoup plus difficile d'estimer un jeu complet de déciles qu'une seule médiane, surtout à partir de distributions présentant une asymétrie positive. Cependant, la méthode que nous recommandons semble donner d'assez bons résultats pour la plupart des estimations de déciles et pourrait certainement être modifiée afin de produire des estimations valables des quartiles si les estimations de déciles dans des conditions de production s'avèrent trop instables. Entretemps, le programme de la SOC a décidé de mettre en œuvre la méthode d'interpolation C95 et de produire des jeux complets d'estimations de déciles pour certaines caractéristiques annuelles dans de futurs rapports.

Nous pensons que nos résultats peuvent être étendus à d'autres plans de sondage, mais nous reconnaissons que nos travaux de recherche ont été menés dans des conditions très contraignantes, à savoir un échantillonnage en grappes à plusieurs degrés à partir d'une population fortement asymétrique, avec un plan à deux UPE par strate au premier degré. Dans d'autres applications, l'interpolation sur des classes dépendantes des données pourrait être combinée à l'estimateur de variance proposé dans l'article publié par Woodruff en 1952, comme l'a suggéré J.N.K. Rao. Pour les enquêtes qui ne se prêtent pas bien à la méthode de rééchantillonnage BRR ou MHS et qui publient des estimations de déciles, notre approche de groupement par classe dépendante des données et d'interpolation pourrait être conjuguée à une méthode de rééchantillonnage bootstrap, tel que le bootstrap de Rao-Wu (Rao et Wu 1988).

Remerciements

Le présent rapport est diffusé afin de tenir les parties intéressées au courant des travaux de recherche et de favoriser la discussion. Les opinions exprimées sur les questions d'ordre statistique, méthodologique ou technique sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau. Les auteurs remercient Erica Filipek, Bonnie Kegan, Amy Newman-Smith de leur contribution précieuse à ce projet de recherche. En outre, ils remercient Wan-Ying Chang, Laura Bechtel, Xijian Liu, J.N.K. Rao, le rédacteur associé et deux examinateurs anonymes de leurs commentaires utiles au sujet de versions antérieures du présent manuscrit.

Bibliographie

Fay, R.E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.

Lienhard, S. (2004). Multivariate Lognormal Simulation with Correlation. http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=6426&objectType=File.

Rao, J.N.K., et Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

Rao, J.N.K., et Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

Rao, J.N.K., et Wu, C.F.J. (1988). Re-sampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Steel, P., et Fay, R.W. (1995). Variance estimation for finite populations with imputed data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Thompson, J.R. (2000). Simulation: A Modeler's Approach. New York : John Wiley & Sons, Inc., 87-110.

Thompson, K.J. (1998). Evaluation of Modified Half Sample Replication for Estimating Variances for the Survey of Construction (SOC). Rapport technique #ESM9801, disponible sur demande auprès de l'Office of Statistical Methods and Research for Economic Programs from the U.S. Census Bureau.

Thompson, K.J., et Sigman, R.S. (2000). L'estimation et l'estimation de la variance par répliques des prix de vente médians des maisons vendues. *Techniques d'enquête*, 26, 2, 173-183.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.


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

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 405-432
Statistique Canada, N° 12-001-X au catalogue

Détermination conjointe de la stratification et de la répartition optimales de l'échantillon en utilisant un algorithme génétique

Marco Ballin et Giulio Barcaroli[1]

Résumé

Le présent article offre une solution au problème de la détermination de la stratification optimale de la base de sondage de la population disponible en vue de minimiser le coût de l'échantillon requis pour satisfaire aux contraintes de précision sur un ensemble d'estimations cibles différentes. La solution est recherchée en explorant l'univers de toutes les stratifications qu'il est possible d'obtenir par classification croisée des variables auxiliaires catégoriques disponibles dans la base de sondage (les variables auxiliaires continues peuvent être transformées en variables catégoriques par des méthodes appropriées). Par conséquent, l'approche suivie est multivariée en ce qui concerne les variables cibles ainsi que les variables auxiliaires. L'algorithme proposé est fondé sur une approche évolutionniste non déterministe qui fait appel au paradigme de l'algorithme génétique. La caractéristique principale de l'algorithme est que l'on considère chaque stratification possible comme un individu susceptible d'évoluer dont l'adaptation est mesurée par le coût de l'échantillon associé requis pour satisfaire à un ensemble de contraintes de précision, ce coût étant calculé en appliquant l'algorithme de Bethel pour une répartition multivariée. Cet algorithme de stratification optimale, implémenté dans un module (ou package) R (SamplingStrata), a été appliqué jusqu'à présent à un certain nombre d'enquêtes courantes à l'Institut national de statistique de l'Italie : les résultats montrent systématiquement une amélioration importante de l'efficacité des échantillons obtenus comparativement aux stratifications adoptées antérieurement.

Mots-clés : Algorithme génétique; stratification optimale; plan de sondage; répartition de l'échantillon; module (ou package) R.

1 Introduction

L'optimalité d'un échantillon peut être définie en fonction des coûts (associés au travail sur le terrain, en particulier au nombre d'unités à interviewer) et de la précision (reliée à la variance d'échantillonnage des estimations cibles). Les plans d'échantillonnage stratifiés sont des plans d'usage très répandu qui permettent de réduire les coûts et d'augmenter la précision des estimations lorsque des variables de stratification sont disponibles dans la base de sondage.

De nombreuses études traitant de l'optimisation des plans d'échantillonnage stratifiés ont été publiées. Nous pouvons les classer comme suit selon l'objet de l'optimisation :

1. la répartition de l'échantillon doit être optimisée, tandis que la stratification est considérée comme telle;
2. la stratification doit être optimisée, tandis que la question de la répartition de l'échantillon est reportée à une étape ultérieure;
3. la stratification et la répartition de l'échantillon sont optimisées en une seule étape.

1. Marco Ballin et Giulio Barcaroli, Istituto Nazionale di Statistica, via C.Balbo 16 - 00184 Rome (Italie). Courriel : ballin@istat.it, barcarol@istat.it.

Dans le premier groupe, nous pouvons inclure Cochran (1977), Bethel (1985, 1989), Chromy (1987), Huddleston, Claypool et Hocking (1970), Kish (1976), Stokes et Plummer (2004), Day (2006, 2010), Díaz-García et Cortez (2008), Kozak, Zieliński et Singh (2008), Khan, Maiti et Ahsan (2010), Kozak et Wang (2010). Bethel (1985, 1989) et Chromy (1987) proposent des algorithmes similaires pour l'extension de la répartition de Neyman au cas multivarié en utilisant des méthodes de programmation convexes. Stokes et Plummer (2004) montrent comment utiliser l'outil de programmation non linéaire disponible dans les chiffriers Excel pour résoudre le même problème. Dans Day (2006, 2010), l'approche de l'algorithme évolutionnaire est proposée pour résoudre le problème de répartition multivariée de l'échantillon sous les mêmes conditions que celles indiquées par Bethel et Chromy. Dans Díaz-García et Cortez (2008), le problème de la répartition multivariée optimale de l'échantillon est résolu sous forme d'un problème d'optimisation multi-objectifs de nombres entiers. Kozak et coll. (2008) étudient le cas de l'échantillonnage stratifié à deux degrés.

Dans le deuxième groupe, nous pouvons considérer Dalenius et Hodges (1959), Singh (1971), Hidiroglou (1986), Lavallée et Hidiroglou (1988), Gunning et Horgan (2004), et Khan, Nand et Ahmad (2008). En général, le problème traité se rapporte à l'optimisation de la stratification que l'on peut obtenir en fonction d'une ou de plusieurs variables continues, corrélées à une ou à plusieurs variables cibles.

Un certain nombre d'articles décrivent le traitement simultané des deux problèmes (stratification et répartition de l'échantillon). Kozak, Verma et Zieliński (2007) proposent une méthode en vue d'obtenir une stratification multivariée tout en minimisant la taille globale de l'échantillon. La méthode est définie uniquement sur une base théorique, et les auteurs affirment que, dans le cas univarié, l'optimisation n'est pas difficile, tandis que dans le cas multivarié, les recherches doivent se poursuivre. Keskintürk et Er (2007) utilisent l'algorithme génétique pour résoudre simultanément les problèmes de répartition de l'échantillon et de détermination des limites des strates dans le cas d'une seule variable de stratification continue en considérant le nombre de strates et la taille totale de l'échantillon comme pré-déterminés. La proposition de Benedetti, Espa et Lafratta (2008) est fondée sur l'utilisation d'une approche arborescente : leur procédure définit un chemin allant de la stratification nulle vers la stratification dite atomique (caractérisée par le nombre maximal de strates obtenu en utilisant toutes les variables auxiliaires, avec les classifications les plus détaillées), généralement sans l'atteindre, étant donné qu'un nombre de règles d'arrêt sont appliquées. Baillargeon et Rivest (2009, 2011) proposent une méthode qui permet d'optimiser conjointement les limites de strate et la taille de l'échantillon en utilisant un algorithme itératif : les limites de strate (reliées à une seule variable de stratification) sont obtenues en minimisant la taille attendue d'échantillon pour estimer le total de population d'une seule variable étudiée (de sorte que cette approche est univariée en ce qui concerne tant la stratification que les variables cibles). En conclusion, la plupart des contributions de ce groupe sont consacrées à la résolution du problème consistant à trouver les meilleures limites de strate pour une seule variable auxiliaire continue; seuls Benedetti et coll. (2008) traitent le cas de la stratification multivariée.

Dans le cas de variables de stratification catégoriques, nous pourrions considérer la stratification donnée par leur produit cartésien; mais, si le nombre de strates créées est grand, on pourrait aboutir à une taille d'échantillon énorme, dépassant de loin celle qui est abordable ou nécessaire pour être certain d'obtenir les niveaux de précision requis. Donc, une tâche cruciale consiste à choisir le « meilleur » produit croisé de variables auxiliaires, c'est-à-dire la meilleure partition de la base de sondage qui n'entraîne pas simultanément une explosion du nombre de strates.

Le présent article propose une solution au problème de détermination conjointe de la stratification optimale d'une base de sondage, et de la taille et de la répartition optimales de l'échantillon dans des conditions entièrement multivariées (c'est-à-dire en ce qui concerne tant les variables de stratification que les variables cibles). La seule restriction a trait à la nature des variables de stratification qui doivent être catégoriques (mais nous donnons des renseignements sur un moyen approprié de transformer les variables continues en variables catégoriques). La solution proposée est fondée sur l'utilisation de l'algorithme génétique. La procédure générale a été implémentée dans un module R, nommé `SamplingStrata`, disponible dans le CRAN (Barcaroli, Pagliuca et Willighagen 2013a). Ce module fait appel à une version modifiée de certaines fonctions d'un autre module de R, `genalg` (Willighagen 2012).

La présentation de l'article est la suivante : la section 2 contient une formalisation du problème d'optimisation. La section 3 décrit en détails l'utilisation de l'algorithme génétique afin de résoudre de manière optimale le problème de la recherche de la meilleure stratification donnant l'échantillon requis dont le coût est minimal. Pour mieux illustrer la méthode, la section 4 donne un exemple basé sur un jeu de données bien connu (les données sur les « fleurs d'iris »). La section 5 décrit la présentation et l'analyse des résultats de l'application de l'algorithme à une enquête réelle, l'*Enquête italienne sur la structure des exploitations agricoles*, et les compare à la solution pratique adoptée par les statisticiens d'enquête. Une autre application, à l'*Enquête mensuelle sur le lait et les produits laitiers*, est présentée à la section 6. Les conclusions finales sont exposées à la section 7.

2 Formalisation du problème d'optimisation

Univers des diverses options de stratification

Nous définissons comme étant une *base de sondage* F un ensemble de N enregistrements contenant de l'information (organisée en variables) en rapport avec N individus de la population de référence. Certaines variables sont utiles pour l'identification des unités, tandis que d'autres peuvent être utilisées pour définir la stratégie d'échantillonnage. Les valeurs de ces dernières (appelées à partir d'ici *variables auxiliaires*) peuvent être observées au moyen d'un recensement ou d'autres sources de données telles que des registres administratifs.

Nous supposons qu'un ensemble de M variables auxiliaires $X_m \left(m = 1, \ldots, M\right)$ sont disponibles dans la base de sondage. Cet ensemble peut contenir différents types de variables (nominales, ordinales ou continues). Nous supposons également que les variables auxiliaires continues sont subdivisées en classes en appliquant des algorithmes de transformation appropriés.

Toutes ces variables peuvent potentiellement être utilisées pour stratifier les unités qui figurent dans la base de sondage.

Sous ces hypothèses, nous pouvons associer à chaque variable auxiliaire un vecteur $d_m = \left\{x_1, \ldots, x_{k_m}\right\}$ de valeurs entières contiguës, qui représentent chacune une valeur originale dans l'ensemble de domaines.

Alors, la stratification la plus détaillée de F peut être considérée comme le résultat du produit cartésien $PC = X_1 \times X_2 \times \ldots \times X_M$.

Le nombre maximal de strates sera $K = \prod_{m=1}^{M} k_m - I^*$, où I^* est le nombre de combinaisons impossibles ou absentes de valeurs dans la base de sondage. Donc, la stratification la plus détaillée de la base de sondage est telle qu'elle contient K strates, qui correspondent à toutes les combinaisons possibles des valeurs dans les M variables auxiliaires. Nous appelons *strates atomiques* les strates qui appartiennent à cette stratification particulière. Chaque strate atomique est caractérisée par une combinaison *unique* de valeurs des M variables auxiliaires. Nous pouvons attribuer une étiquette $l_k \left(k = 1, \ldots, K\right)$ à chaque strate atomique.

Si nous considérons l'ensemble étiqueté de strates atomiques $L = \left\{l_1, l_2, \ldots, l_K\right\}$, nous pouvons définir l'ensemble de toutes les partitions possibles $P_1, P_2, \ldots, P_B$, où B peut être calculé en utilisant la formule de Bell :

$$B_K = \sum_{i=0}^{K-1} \binom{K-1}{i} \cdot B_i \qquad \left(B_0 = 1\right)$$

Nous définissons l'ensemble $\left\{P_1, P_2, \ldots, P_B\right\}$ de partitions de L comme étant l'*univers (ou espace) des stratifications.*

Évaluation d'une stratification donnée

Étant donné une partition P_i de L, caractérisée par H strates, soit N_h et $S_{h,g}^2, h = 1, \ldots, H$, $g = 1, \ldots, G$ respectivement, le nombre d'unités et les variances dans la strate h des G variables cibles de l'enquête $Y_1, \ldots, Y_G$. En supposant un échantillonnage aléatoire simple de n_h unités sans remise dans chaque strate, la variance de l'estimateur de Horvitz-Thompson du total de la g^e variable cible $\left(\hat{T}_g\right)$ est

$$\mathrm{Var}\left(\hat{T}_g\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{h,g}^2}{n_h} \qquad g = 1, \ldots, G \tag{2.1}$$

Considérons la fonction de coût suivante

$$C\left(n_1, \ldots, n_H\right) = C_0 + \sum_{h=1}^{H} C_h n_h \tag{2.2}$$

où C_0 indique un coût fixe (qui ne dépend pas de la taille de l'échantillon) et C_h représente le coût moyen de l'observation d'une unité dans la strate h.

Sachant $V_g \left(g = 1, \ldots, G\right)$, les limites supérieures des variances d'échantillonnage prévues de $\hat{T}_1, \ldots, \hat{T}_G$, le problème de répartition multivariée optimale classique de l'échantillon (Bethel 1985) peut être défini comme la recherche de la solution du minimum (par rapport à n_h) de la fonction linéaire C sous les contraintes convexes $\mathrm{Var}\left(\hat{T}_g\right) \leq V_g \quad g = 1, \ldots, G$:

$$\begin{cases} \min C\left(n_1,\ldots,n_H\right) = C_0 + \sum_{h=1}^{H} C_h n_h \\ \mathrm{Var}(\hat{T}_g) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{h,g}^2}{n_h} \le V_g \qquad g = 1,\ldots,G \end{cases} \tag{2.3}$$

Bethel (1989) a suggéré qu'il était plus facile de résoudre le problème en considérant la fonction suivante de n_h :

$$x_h = \begin{cases} 1/n_h & \text{si } n_h \ge 1 \\ \infty & \text{autrement} \end{cases} \tag{2.4}$$

En utilisant x_h, la fonction de coût peut s'écrire

$$C\left(x_1,\ldots,x_H\right) = C_0 + \sum_{h=1}^{H} \frac{C_h}{x_h} \tag{2.5}$$

et la variance,

$$\mathrm{Var}\left(\hat{T}_g\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{1}{x_h N_h}\right) S_{h,g}^2 x_h = \sum_{h=1}^{H} N_h^2 S_{h,g}^2 x_h - N_h S_{h,g}^2 \quad g = 1,\ldots,G \tag{2.6}$$

Par conséquent, le problème de répartition multivariée de l'échantillon peut être défini comme la recherche du minimum (par rapport à x_h) de la fonction convexe (2.5) sous un ensemble de contraintes linéaires

$$\sum_{h=1}^{H} N_h^2 S_{h,g}^2 x_h - N_h S_{h,g}^2 \le V_g \quad g = 1,\ldots,G \tag{2.7}$$

Bethel a fourni un algorithme, dont il est prouvé qu'il converge vers la solution (si elle existe), en appliquant la méthode des multiplicateurs de Lagrange à ce problème (un algorithme plus facile avait été proposé antérieurement par Chromy (1987); comme l'a fait remarquer Bethel, l'algorithme de Chromy fonctionne dans la plupart des cas pratiques, mais il n'existe pas de preuve qu'il converge si une solution existe).

L'approche de l'optimisation illustrée ici donne une solution continue, qui doit être arrondie pour fournir des tailles entières d'échantillons de strates. Notre mise en œuvre de l'algorithme de Bethel fournit les valeurs de n_h comme étant les valeurs $1/x_h$ arrondies à l'entier supérieur.

Il convient de souligner que la même approche peut être suivie pour traiter le problème des domaines multiples. Considérons la transformation habituelle pour le problème d'estimation par domaine :

$$Y_i^d = \begin{cases} Y_i & \text{si l'unité } i \text{ appartient au domaine } d \\ 0 & \text{autrement} \end{cases}$$

Si les quantités définies antérieurement pour décrire l'approche de Bethel sont calculées en utilisant les variables $Y^d \left(d = 1,\ldots,D\right)$, la solution de la répartition multivariée est la solution pour le cas de plusieurs domaines.

Sélection de la meilleure stratification sur la base d'un dénombrement complet

Afin de choisir la meilleure stratification d'une base de sondage donnée, c'est-à-dire celle qui garantit le coût minimal $C\left(n_1,\ldots,n_H\right)$ associé à un échantillon dont la taille totale et la répartition sont conformes aux contraintes de précision, il est possible de procéder comme il suit :

- générer la stratification la plus détaillée associée à F, c'est-à-dire l'ensemble L de strates atomiques;
- dénombrer toutes les partitions P_i de L;
- pour chaque partition P_i, résoudre le problème de répartition de l'échantillon correspondant, ce qui équivaut à déterminer le vecteur $\left(n_1,\ldots,n_H\right)$, et calculer la valeur $C_i\left(n_1,\ldots,n_H\right)$ associée à P_i;
- choisir la partition P_i pour laquelle $C_i\left(n_1,\ldots,n_H\right)$ est minimisé.

Ce faisant, l'optimisation de la solution est obtenue en tenant compte de l'univers complet des stratifications.

Malheureusement, cette procédure ne s'applique qu'aux situations où la dimension K de L est faible : en fait, le nombre de partitions (donné par la formule de Bell) augmente très rapidement (par exemple, $B_4 = 15$, $B_{10} = 115\,975$ et $B_{100} \approx 4,76 \times 10^{115}$). Par conséquent, dans la plupart des cas, le dénombrement complet de l'espace des solutions n'est pas faisable. La présente proposition, fondée sur l'algorithme génétique, permet d'explorer l'univers des stratifications et de déterminer lesquelles ne devraient pas être éloignées de la stratification optimale.

L'algorithme génétique

L'algorithme génétique $\left(AG\right)$ est une technique de recherche utilisée en calcul pour trouver des solutions exactes ou approximatives aux problèmes d'optimisation et de recherche. Les algorithmes génétiques représentent une classe particulière d'algorithmes évolutionnaires qui font appel à des techniques inspirées de la biologie de l'évolution, telles que l'*hérédité,* la *mutation,* la *sélection* et le *croisement* (également appelé *recombinaison*) (Vose 1999) (Schmitt 2001 et 2004).

On implémente un AG dans une simulation informatique itérative dans laquelle un ensemble initial d'*individus,* qui sont chacun une solution potentielle du problème courant (représentés par un vecteur appelé *génome*), évolue par *hérédité, mutation, sélection* et *croisement,* de manière à accroître la *valeur d'adaptation* (en anglais, *fitness*) moyenne des *générations* suivantes. Ici, la *valeur d'adaptation* correspond à la fonction objectif définie dans le problème d'optimisation de manière que l'évolution ait pour résultat la maximisation (ou la minimisation) de la fonction objectif.

L'ensemble d'individus traités à chaque itération de l'*AG* est appelé *génération*. L'*évolution* est l'ensemble des changements qui ont lieu durant la production des *générations* consécutives par itération du processus.

À chaque itération de l'*AG*, après avoir évalué la valeur d'adaptation de chaque individu de la génération, un ensemble d'individus sont sélectionnés de manière stochastique (en privilégiant ceux ayant la valeur d'adaptation la plus élevée) et modifiés (recombinés et parfois soumis aléatoirement à des

mutations) pour former une nouvelle génération. Cette nouvelle génération est alors évaluée durant l'itération suivante de l'algorithme. Puisque les individus ayant la meilleure valeur d'adaptation sont plus susceptibles que les autres d'être sélectionnés pour engendrer les individus de la génération suivante, l'*AG* produit une augmentation de la valeur moyenne d'adaptation au cours de l'évolution.

Le paramètre *taux de mutations* est exprimé comme le taux de *chromosomes* (les éléments du *génome*) qui peuvent subir une mutation pour chaque individu au moment où sont générés les *enfants* destinés à former la génération suivante. Une valeur élevée garantit de grandes différences entre les générations successives. Il convient de souligner qu'un taux élevé de mutations rend l'*AG* plus susceptible de ne pas stagner à des optima locaux, au prix d'une convergence plus lente vers la solution optimale, tandis qu'une valeur faible accélère la convergence, mais augmente le risque d'optima locaux.

Habituellement, l'algorithme s'arrête quand un nombre maximal d'itérations a été atteint ou que la poursuite des itérations n'améliore pas la solution courante. Dans les deux cas, la solution optimale peut ou non avoir été atteinte.

3 Application de l'algorithme génétique au problème de stratification optimale

Dans le cadre de l'*AG*, le problème de stratification et de répartition peut être représenté comme il suit :

- une stratification donnée est considérée comme un *individu*;
- le *génome* d'un individu est un vecteur dont la dimension est donnée par le nombre K de strates atomiques;
- chaque position $i\left(i = 1, \ldots, K\right)$ dans le vecteur est associée à une strate atomique donnée, et contient une valeur entière $v_i \left(1 < v_i < U\right)$ avec $U \leq K$, où U est défini comme le nombre maximal de strates dans la solution finale : si certains éléments du vecteur ont la même valeur, cela signifie que les strates atomiques correspondantes sont fusionnées en une nouvelle strate désignée par cette valeur;
- de cette façon, une stratification $P\left(\nu\right)$ peut être identifiée par un vecteur $\boldsymbol{v} = \left[v_1, \ldots, v_K\right]$, où chaque valeur v_i est positionnellement associée à la strate atomique identifiée par l'étiquette l_i, et peut prendre une valeur entière à l'intérieur d'un intervalle $\left[1, U\right]$. L'*espace de toutes les stratifications (ou partitions) possibles* $P\left(\nu\right)$ (espace des solutions) est donné par tous les vecteurs possibles $\boldsymbol{v}$;
- la fonction d'*adaptation* d'un individu $P\left(\nu\right)$ est la valeur de la fonction de coût $C\left(n_1, \ldots, n_{H_{P(v)}}\right) = C_0 + \sum_{h=1}^{H_{P(v)}} C_h n_h$, où les termes C_0 et C_h sont des constantes données, et les $n_1, \ldots, n_{H_{P(v)}}$ sont calculés par application de l'algorithme de Bethel à la stratification sous l'ensemble de contraintes de précision sur les variables cibles.

Mentionnons que, si nous posons que $C_0 = 0,$ et $C_h = 1$ pour toutes les strates atomiques, alors la valeur de la fonction de coût coïncide simplement avec la taille d'échantillon requise pour satisfaire les contraintes de précision.

Maintenant que nous avons défini une représentation appropriée du domaine de toutes les solutions possibles, ainsi que la fonction d'adaptation qu'il faut calculer pour chaque solution nous allons montrer comment fonctionne l'*AG*.

Étape 0 : Création de la première génération d'individus

La première étape consiste à former un premier ensemble de stratifications distinctes (la première génération d'individus) : sur la base de la valeur du paramètre de *taille des générations,* nous générons p individus différents. Cela signifie que, pour le j^{e} individu, K valeurs entières (une pour chaque élément du vecteur représentant le génome) sont générées aléatoirement à partir d'une loi uniforme dans l'intervalle $\left[1, U\right]$. En fixant $U \leq K$, nous pouvons déterminer une limite supérieure du nombre maximal de strates agrégées distinctes.

Étape 1 : Évaluation de la valeur d'adaptation de chaque individu dans la population

Pour chaque individu de la population (c'est-à-dire pour chacune des p stratifications), la valeur d'adaptation est déterminée en calculant le coût total requis pour satisfaire les contraintes de précision sur les G estimations $\hat{T}_g$ différentes (afin d'éliminer la dépendance sur l'échelle (ou l'intervalle) des valeurs associées aux G variables cibles, au lieu de considérer les contraintes exprimées par (2.7) comme une limite supérieure de la variance des variables cibles, nous déterminons les contraintes sur leur coefficient de variation $\text{CV} = \sqrt{\text{var}(\hat{T}_G)} \big/ \hat{T}_G$). L'évaluation est effectuée en appliquant l'algorithme de Bethel, qui nécessite comme données d'entrée pour chaque strate de la solution courante :

- les moyennes et les écarts-types des variables cibles;
- le coût d'interview par unité;
- le chiffre de population (nombre d'unités).

Chacun des éléments susmentionnés est calculé sur la base des valeurs correspondantes dans les strates atomiques.

Considérons une partition particulière $P(\mathrm{v})$ de L déterminée par une solution donnée $\boldsymbol{v} = \left[v_1, \ldots, v_K\right]$. Soit $D_i \left(i = 1, 2, \ldots, Q_{P(v)}\right)$ une strate dans cette partition. Il existe deux possibilités :

1. D_i coïncide avec une strate atomique l_k;
2. $D_i = \left\{l_j^i, \ldots, l_l^i\right\}$ est le résultat de l'agrégation d'un sous-ensemble $\left\{l_j^i, \ldots, l_l^i\right\}$ de strates atomiques.

Dans le premier cas, les moyennes et les variances des variables cibles dans la strate sont connues. Dans le deuxième, les moyennes et les variances de D_i peuvent être calculées en utilisant les formules suivantes :

$$\bar{Y}_{g,D_i} = \frac{\sum_{l_k \in D_i} \bar{Y}_{g,l_k} N_{l_k}}{\sum_{l_k \in D_i} N_{l_k}} \tag{3.1}$$

$$S^2_{g,D_i} = \left(\sum_{l_k \in D_i} N_{l_k} - 1 \right)^{-1} \left\{ \sum_{l_k \in D_i} \left(N_{l_k} - 1 \right) S^2_{g,l_k} + \sum_{l_k \in D_i} N_{l_k} \left(\bar{Y}_{g,l_k} - \bar{Y}_{g,D_i} \right)^2 \right\} \tag{3.2}$$

où :

$\bar{Y}_{g,D_i}$ et $\bar{Y}_{g,l_k}$ sont les valeurs moyennes dans la strate agrégée D_i et dans les strates atomiques l_k;

N_{l_k} est le nombre d'unités dans la strate atomique l_k;

S^2_{g,D_i} et S^2_{g,l_k} sont les variances dans la strate agrégée D_i et les strates atomiques l_k.

Le coût d'observation prévu d'une unité dans une strate agrégée est obtenu en calculant la moyenne des coûts dans chacune des strates atomique participante, pondérés par leur population :

$$C_{D_i} = \frac{\sum_{l_k \in D_i} C_{l_k} N_{l_k}}{\sum_{l_k \in D_i} N_{l_k}} \tag{3.3}$$

Enfin, nous pouvons calculer la population dans toute strate agrégée comme étant la somme des unités dans les strates atomiques participantes :

$$N_{D_i} = \sum_{l_k \in D_i} N_{l_k} \tag{3.4}$$

Donc, en correspondance avec chaque solution possible, nous sommes capables de calculer dynamiquement toute l'information requise pour appliquer l'algorithme de répartition optimale, qui produit le coût total

$$C\left(n_1, \ldots, n_{H_{P(v)}}\right) = C_0 + \sum_{h=1}^{H_{P(v)}} C_h n_h$$

qui est la valeur d'adaptation de l'individu.

Étape 2 : Production d'une nouvelle génération

Une fois que l'adaptation de chaque individu est évaluée, une proportion d'entre eux est sélectionnée pour produire une nouvelle génération. Les individus sont sélectionnés selon le processus basé sur l'adaptation, en vertu duquel les individus les mieux adaptés sont plus susceptibles d'être sélectionnés que les autres, tandis qu'une faible proportion seulement d'individus moins bien adaptés sont sélectionnés. La présence de cette deuxième composante contribue au maintien d'une diversité suffisante de la nouvelle génération, afin d'éviter une convergence prématurée sur de mauvaises solutions. On peut aussi choisir d'indiquer le nombre d'individus les mieux adaptés (exprimé en pourcentage de la taille p de la génération) qui, en toutes circonstances, doivent être également présents dans la génération suivante (paramètre d'*élitisme*).

La génération suivante sera donc composée d'un certain nombre d'individus provenant de la génération précédente (les meilleurs), ainsi que d'un certain nombre d'« enfants », obtenu en sélectionnant et en croisant des « parents » provenant de la génération courante. Dans l'approche de l'*AG*, le *génome* d'un « enfant » est formé en utilisant les opérateurs de *croisement* et de *mutation* :

- *croisement :* de nombreuses techniques de croisement, faisant appel à différentes structures de données et différents critères de sélection des chromosomes, sont appliquées aux AG, mais l'approche générale consiste à échanger un sous-ensemble de chromosomes entre deux parents. Dans notre application, une fois que les deux parents ont été sélectionnés avec une probabilité proportionnelle à leur valeur d'adaptation, un *point de croisement* est généré, de nouveau aléatoirement. Ce point de croisement est un nombre entier compris dans l'intervalle $[1, K]$.

  Soit c ce point de croisement généré : alors, l'individu enfant sera formé en héritant des c premiers chromosomes du premier parent et des chromosomes restants $(K - c)$ du second parent;
- *mutation :* sachant la probabilité qu'une valeur arbitraire dans une séquence génétique soit modifiée par rapport à son état original *(chances de mutation)*, pour chaque chromosome du génome, l'*AG* tire une valeur aléatoire pour décider si la valeur en question sera modifiée ou non.

En appliquant les méthodes susmentionnées de croisement et de mutation, on crée un nouvel individu qui habituellement partage nombre des caractéristiques de ses « parents ». De nouveaux parents sont sélectionnés pour produire de nouveaux enfants, et le processus se poursuit jusqu'à ce qu'une nouvelle génération d'individus (stratification) de taille appropriée soit générée.

Étape 3 : Critères d'itération et d'arrêt

Habituellement, la valeur d'adaptation moyenne augmente lorsqu'on passe d'une génération à la suivante. Les étapes 1 et 2 sont répétées jusqu'à ce qu'une condition d'arrêt soit atteinte. Les conditions d'arrêt habituelles sont les suivantes :

1. le nombre maximal d'itérations a été atteint;
2. un « plateau » a été atteint, de sorte que des itérations successives n'améliorent plus les résultats;
3. une combinaison des deux points précédents.

Dans notre cas, la condition d'arrêt peut être considérée comme une combinaison des points susmentionnés. En fait, la règle appliquée est le nombre maximal d'itérations, mais ce nombre est déterminé en analysant les exécutions précédentes, afin de détecter le « plateau » et d'être certain que des itérations supplémentaires n'amélioreront vraisemblablement pas la solution finale.

Paramètres critiques de l'algorithme de stratification optimale

Ici, nous faisons une distinction entre les paramètres qui sont des paramètres communs de l'algorithme génétique et ceux qui sont particuliers au problème auquel l'algorithme est appliqué, c'est-à-dire la

stratification optimale d'une base de sondage de population (les noms des paramètres sont ceux utilisés dans le module R `SamplingStrata`).

Parmi les premiers, mentionnons :

- la taille de la génération d'individus *(pop)*;
- le nombre d'itérations *(iterations)*;
- les chances de mutation *(mut_chance)*;
- élitisme (elitism_rate).

Les paramètres contextuels qui les remplacent sont :

- nombre minimal d'unités par strate *(minnumstrat)* (l'algorithme de Bethel s'efforce d'attribuer à chaque strate au moins le nombre d'unités indiquées par ce paramètre);
- le nombre initial de strates *(initialStrata)*;
- la possibilité d'accroître le nombre maximal de strates *(addStrataFactor).*

Comme pour le premier groupe, il n'existe pas de règles strictes pour attribuer les valeurs à ces paramètres. Étant donné un problème particulier, il est suggéré d'exécuter un certain nombre d'essais afin d'évaluer la sensibilité des solutions aux valeurs des paramètres.

Il est important de tenir compte du fait que des paramètres tels que *taille de la génération* et *élitisme* influent généralement sur la rapidité de convergence, mais moins sur la solution finale, à condition qu'un nombre « raisonnable » d'itérations soit donné.

Le caractère raisonnable du paramètre *nombre d'itérations* peut être évalué en analysant le comportement de la fonction d'adaptation : si les valeurs de cette fonction ne diminuent plus après un certain nombre d'itérations, il est raisonnable de s'attendre à ce que l'augmentation du nombre d'itérations ne produira pas de meilleurs résultats.

En revanche, la valeur de *chances de mutation* a des effets sur la rapidité de convergence ainsi que sur la qualité de la solution finale : des chances élevées de mutation permettent d'éviter les minima locaux, au prix d'une convergence plus lente.

Inversement, les valeurs des paramètres du deuxième groupe doivent être données sur la base de considérations pratiques, reliées aux caractéristiques et aux exigences de l'enquête que l'on conçoit.

Pour ce qui est du paramètre *nombre minimal d'unités par strate*, s'il faut s'assurer que toutes les strates contiennent un nombre adéquat d'observations (afin de tenir compte de la non-réponse prévue, de la nécessité de calculer la variance d'échantillonnage, de problèmes sur le terrain, *etc.*), on peut fixer une valeur plus élevée que la valeur par défaut (laquelle est fixée à 2).

Le paramètre *nombre initial de strates* est très important. Avant tout, si elle est associée à une valeur nulle du paramètre *addStrataFactor*, sa valeur détermine le nombre maximal acceptable de strates dans la solution finale. Cette possibilité peut être utile non seulement pour des raisons de travail sur le terrain (si, par exemple, pour des raisons organisationnelles, le nombre de strates doit être limité), mais surtout parce que la solution finale est très sensible à la valeur de ce paramètre. Nous avons constaté que si l'on exécute l'algorithme avec différentes valeurs du paramètre *initialStrata*, allant de faibles valeurs jusqu'au

maximum donné par le nombre de strates atomiques, les solutions peuvent être fort différentes. Il est possible de permettre à l'algorithme de choisir pour nous, comme il suit : nous fixons *initialStrata* en lui attribuant une faible valeur, et en attribuant en même temps une valeur élevée au paramètre *addStrataFactor* (ce dernier est utilisé pour augmenter dynamiquement la valeur établie par le paramètre *initialStrata* : chaque fois qu'une mutation a lieu, un nombre aléatoire compris entre 0 et 1 est généré, et s'il est plus grand que la quantité (1-*addStrataFactor*), le nombre maximal de strates est augmenté d'une unité) (par défaut, il est égal à 0). La manipulation de ces deux paramètres donne différentes possibilités :

1. pour toute valeur donnée de *initialStrata*, si *addStrataFactor* est fixé à 0, l'algorithme doit considérer cette valeur comme une limite fixe, et toutes les solutions à explorer seront caractérisées par ce nombre maximal de strates;
2. autrement, si *addStrataFactor* est fixé à une valeur supérieure à 0, l'algorithme peut explorer les solutions en faisant varier le nombre de strates, d'une valeur initiale donnée par *initialStrata* jusqu'à un nombre maximal donné par le nombre de strates atomiques.

4 Un exemple fondé sur le jeu de données *Fleurs d'iris*

Afin d'illustrer comment appliquer l'algorithme pour trouver la stratification optimale, nous pouvons nous servir du jeu de données *Fleurs d'iris* bien connu. Ce jeu de données comprend un total de 150 observations, réparties de manière égale entre les trois espèces de fleurs d'iris (*setosa, virginica* et *versicolor*). Quatre caractéristiques sont mesurées pour chaque observation (c'est-à-dire la longueur et la largeur du sépale et du pétale, en centimètres).

Nous considérerons ce jeu de données comme une base de sondage possible de laquelle nous pouvons tirer un échantillon, sous un plan stratifié, afin d'estimer deux variables cibles :

- Y_1 : Pétale.longueur;
- Y_2 : Pétale.largeur.

Pour simplifier, nous supposons que deux variables auxiliaires seulement sont disponibles dans la base de sondage :

- X_1 : Sépale.longueur;
- X_2 : Espèce.

Alors que la deuxième variable auxiliaire est catégorique, la première est continue et doit être transformée en une variable catégorique ordonnée. Pour cela, nous utilisons la *méthode de classification automatique univariée à k-moyennes* (Hartigan et Wong 1979), et obtenons les intervalles suivants : [4,3; 5,5], (5,5; 6,5], (6,5; 7,9].

Le produit cartésien des deux variables auxiliaires devrait produire $3 \times 3 = 9$ strates distinctes. En réalité, celle correspondant à Espèce = « *setosa* » et Sépale.longueur $\in$ (6,5; 7,9] ne contient aucune unité. Donc, les strates présentées au tableau 4.1 seront considérées comme représentant la stratification atomique initiale.

Tableau 4.1
Information sur les strates atomiques

strate	X_1 = Sépale.longueur	X_2 = Espèce	N	Y_1 = Pétale.longueur		Y_2 = Pétale.largeur		Coût
				Moyenne	Écart-type	Moyenne	Écart-type	
1	[4,3; 5,5] (1)	Setosa (1)	45	1,47	0,17	0,24	0,11	1
2	[4,3; 5,5] (1)	Versicora (2)	6	3,58	0,49	1,17	0,21	1
3	[4,3; 5,5] (1)	Virginica (3)	1	4,50	0,00	1,70	0,00	1
4	[5,5; 6,5] (2)	Setosa (1)	5	1,42	0,17	0,26	0,08	1
5	[5,5; 6,5] (2)	Versicora (2)	35	4,27	0,37	1,32	0,19	1
6	[5,5; 6,5] (2)	Virginica (3)	23	5,23	0,32	1,95	0,29	1
7	[6,5; 7,9] (3)	Versicora (2)	9	4,68	0,19	1,46	0,11	1
8	[6,5; 7,9] (3)	Virginica (3)	26	5,88	0,49	2,11	0,23	1

Pour simplifier, nous supposons que le coût fixe C_0 est nul et que la valeur de tous les C_h est fixée à 1 : de la sorte, le coût d'une solution coïncide avec la somme des unités d'échantillonnage affectées aux strates, c'est-à-dire avec la taille totale d'échantillon $\left(C = n = \sum_{h=1}^{H} n_h\right)$.

Nous imposons comme contraintes de précision sur les estimations des deux variables cibles une borne supérieure de 0,05 (5 %) de leur coefficient de variation prévu.

Enfin, nous fixons le nombre minimal d'unités qui doit être sélectionné dans chaque strate à 2 (le minimum requis pour calculer la variance d'échantillonnage).

Sous ces hypothèses, et en utilisant la stratification atomique, l'algorithme de Bethel résout le problème de répartition optimale de l'échantillon en définissant une taille minimale d'échantillon de 17 unités, avec un vecteur de répartition $\mathbf{a} = (2, 2, 1, 2, 3, 3, 2, 2)$.

Si nous procédons à la partition de l'ensemble de strates atomiques, le nombre total résultant de stratifications possibles (donné par la formule de Bell) est $B_8 = 4\,140$. Ce nombre est tel qu'il est possible de dénombrer toutes les partitions des strates atomiques et, pour chacune d'elles, de calculer la taille minimale d'échantillon en appliquant l'algorithme de Bethel (pour dénombrer toutes les partitions dans cet exemple, nous avons utilisé la fonction `setparts()`, contenue dans le module R `partitions` (Hankin 2011)).

L'intervalle de tailles d'échantillon va d'un minimum de 11 à un maximum de 78 (cette dernière valeur correspond à la solution « *pas de stratification* ») (voir figure 4.1).

Nous constatons que la valeur minimale $(n = 11)$ trouvée est considérablement plus faible que la valeur calculée correspondant à la stratification atomique $(n = 17)$. Cette valeur minimale caractérise 8 partitions seulement sur 4 140.

Maintenant, nous appliquons l'algorithme génétique pour évaluer sa capacité à trouver la solution optimale (ou du moins une solution qui ne s'en écarte pas trop), sans être obligés d'explorer toutes les solutions, mais uniquement un sous-ensemble strict de celles-ci.

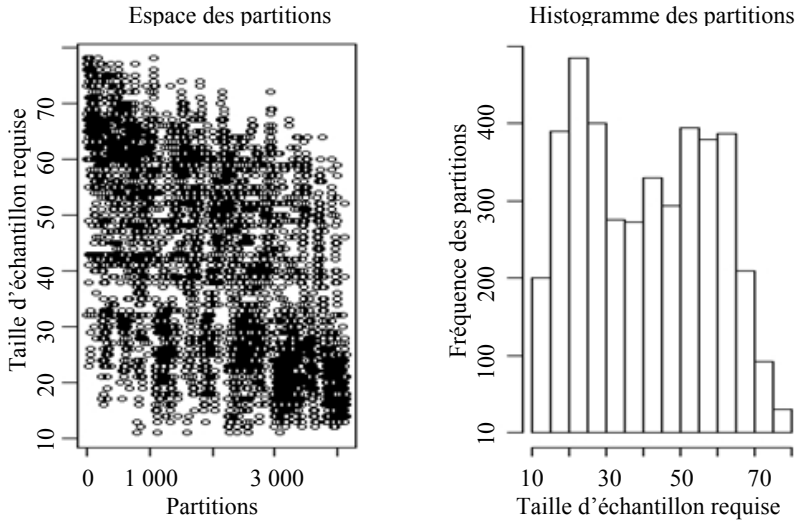


Figure 4.1 Espace des partitions

Étape 0 : Création de la première génération

Pour commencer, nous posons que $U = 8$ (nous pouvons accepter un nombre de strates final qui est égal au nombre de strates atomiques, de sorte que $U = K$). Le paramètre *pop* de *taille de génération* est fixé à 10. Donc, un premier ensemble contenant dix individus (stratifications) distincts est généré. Chacun d'eux est représenté par un vecteur de huit éléments, c'est-à-dire le nombre de strates atomiques différentes. Un individu $\boldsymbol{v} = (1, 2, 3, 4, 5, 6, 7, 8)$ ou, de façon équivalente, $\boldsymbol{v} = (3, 6, 4, 2, 1, 8, 7, 5)$ correspond à la stratification la plus détaillée (car chacune des strates porte une étiquette différente), tandis que $\boldsymbol{v} = (1, 1, 1, 1, 1, 1, 1, 1)$ ou de façon équivalente $\boldsymbol{v} = (4, 4, 4, 4, 4, 4, 4, 4)$ correspond à la « stratification nulle » (car les strates atomiques portent des étiquettes identiques).

Étape 1 : Évaluation de l'adaptation de chaque individu de la génération

Nous appliquons l'algorithme de Bethel à chacun des dix individus de la génération courante afin de trouver le coût de l'échantillon requis pour satisfaire les contraintes de précision fixes.

Pour cela, nous commençons par calculer les strates et l'information pour chaque individu. Par exemple, pour un individu généré $\boldsymbol{v} = (4, 1, 1, 4, 8, 7, 8, 1)$, l'information est obtenue au moyen de l'une des strates atomiques disponibles, en appliquant (3.1) et (3.2) (voir le tableau 4.2).

Tableau 4.2
Information sur les strates agrégées générées

Strate agrégée	Strates atomiques originales	(X_1, X_2)	N	Y_1		Y_2	
				Moyenne	Écart-type	Moyenne	Écart-type
1	2,3,8	(1,2) ou (1,3) ou (3,3)	33	5,41	1,01	1,92	0,44
2	1,4	(1,1) ou (2,1)	50	1,46	0,17	0,25	0,10
3	6	(2,3)	23	5,23	0,31	1,95	0,28
4	5,7	(2,2) ou (3,2)	44	4,35	0,37	1,35	0,18

La valeur d'adaptation de cet individu est mesurée par la taille d'échantillon requise correspondante, qui s'avère être 14, avec un vecteur de répartition $\mathbf{a} = (6, 2, 3, 3)$.

Tous les individus sont triés en fonction de leur performance : l'individu en première position est celui qui requiert la taille d'échantillon minimale, et en 10e position, celui nécessitant la taille d'échantillon maximale.

Étape 2 : Production d'une nouvelle génération

Si nous fixons la valeur du paramètre *élitisme* à 20 % (une valeur par défaut fréquente), nous prenons systématiquement les deux meilleurs individus de la génération courante et nous les transférons directement à la génération suivante, sans aucune modification de leur génome.

Puis, nous procédons à la génération de nouveaux individus de la façon suivante :

1. nous sélectionnons des couples d'individus de la génération courante avec une probabilité proportionnelle à leur valeur d'adaptation : par exemple, supposons que nous sélectionnions $\mathrm{v}_k = (1, 1, 3, 4, 3, 2, 2, 2)$ et $\mathrm{v}_j = (2, 2, 2, 2, 2, 1, 1, 1)$;
2. un point de croisement est généré aléatoirement, c'est-à-dire un nombre entier à l'intérieur de l'intervalle $[1, 8]$: supposons qu'il est égal à 3;
3. le croisement est exécuté en attribuant à l'enfant les trois premiers éléments du parent v_k et les cinq derniers éléments du parent v_j, pour obtenir de cette façon $\mathrm{v}_{\text{new}} = (1, 1, 3, 2, 2, 1, 1, 1)$;
4. en ayant fixé la valeur du paramètre de *taux de mutations* à 0,05, pour chaque élément de l'enfant, un nombre aléatoire est généré dans l'intervalle $[0, 1]$: si ce nombre est inférieur à 0,05, la valeur de l'élément est modifiée (en générant une nouvelle valeur comprise entre 1 et 9), sinon elle n'est pas modifiée.

Étape 3 : Critères d'itération et d'arrêt

Le nombre d'itérations a été fixé à 25. Donc, les étapes 1 et 2 sont répétées 25 fois. L'individu qui, de toutes les générations, possède la meilleure valeur d'adaptation est retenu comme étant la meilleure solution.

Le graphique de la figure 4.2, obtenu durant l'exécution du programme, montre la convergence de l'algorithme. Deux courbes différentes sont présentées : la courbe inférieure est reliée à la meilleure solution trouvée jusqu'à la k^{e} itération (comme la meilleure solution est mémorisée, la courbe ne peut

que diminuer à mesure que l'exécution de l'algorithme se poursuit); la courbe supérieure donne la moyenne des 10 solutions évaluées à chaque itération.

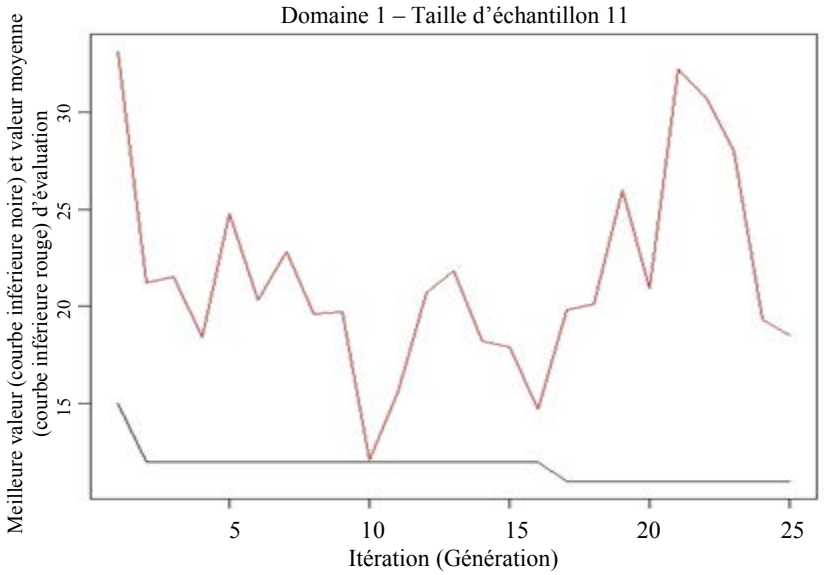


Figure 4.2 Meilleure valeur et valeur moyenne d'évaluation durant l'exécution de l'AG

La meilleure solution résultante est $\boldsymbol{v} = (4,1,3,4,1,3,3,2)$. Elle correspond à la stratification présentée dans le tableau 4.3, avec un vecteur de répartition $\mathbf{a} = (3,2,4,2)$.

Tableau 4.3
Information sur les strates finales

Strate agrégée	Strates atomiques originales	(X_1, X_2)	N	Y_1		Y_2	
				Moyenne	Écart-type	Moyenne	Écart-type
1	2,5	(1,2) ou (2,2)	41	4,16	0,45	1,30	0,19
2	8	(3,3)	26	5,88	0,49	2,10	0,22
3	3,6,7	(1,3) ou (2,3) ou (3,2)	33	5,06	0,38	1,80	0,33
4	1,4	(1,1) ou (2,1)	50	1,46	0,17	0,25	0,10

Brièvement, en appliquant l'algorithme génétique, nous réussissons à trouver la solution optimale en n'explorant que $25 \times 10 = 250$ stratifications différentes au lieu des 4 140 appartenant à l'univers des partitions.

Afin de confirmer que ce résultat n'est pas dû à un « coup de chance », nous effectuons différentes exécutions de l'algorithme : chaque exécution comporte 10 répétitions de l'application de l'algorithme génétique, en faisant varier les valeurs du paramètre « nombre d'itérations ». Les résultats sont présentés au tableau 4.4.

Tableau 4.4
Capacité de l'AG à trouver la solution optimale

Exécution de l'AG (10 répétitions chacune)	Valeur du paramètre « nombre d'itérations » dans l'AG	Solutions avec n = 11 (optimale)	Solutions avec n = 12	Solutions avec n = 14
(a)	25	5	4	1
(b)	50	7	3	-
(c)	100	9	1	-
(d)	200	10	-	-

Dans l'exécution (a), nous découvrons que, avec 25 itérations seulement, arriver à trouver la solution optimale est effectivement un « coup de chance », car dans la moitié des essais, la solution trouvée est plus grande que la solution optimale. Toutefois, si l'on fait augmenter le nombre d'itérations jusqu'à 200 (exécution (d)), l'algorithme génétique s'avère fiable en ce qui a trait à sa capacité d'atteindre l'optimalité, car dans tous les essais, la solution optimale est trouvée.

En ce qui concerne le nombre de strates correspondant à la solution optimale trouvée, en moyenne, il est de 4, avec un intervalle de $\left[3, 5\right]$.

Enfin, nous voulons aussi vérifier que les solutions trouvées sont conformes aux contraintes de précision (CV maximal égal à 5 % pour les deux variables cibles). Donc, dans l'exécution (d) (itérations = 200), pour chacune des 10 solutions produites, nous procédons au tirage de 1 000 échantillons dans la base de sondage et nous calculons les CV associés. Les résultats correspondants sont présentés à la figure 4.3 : la moyenne des CV pour la première variable cible (pétale.longueur) est de l'ordre de 3 %, tandis que pour la deuxième, elle est de l'ordre de 5 %. Donc, nous pouvons dire qu'en moyenne, les contraintes de précision n'ont pas été violées.

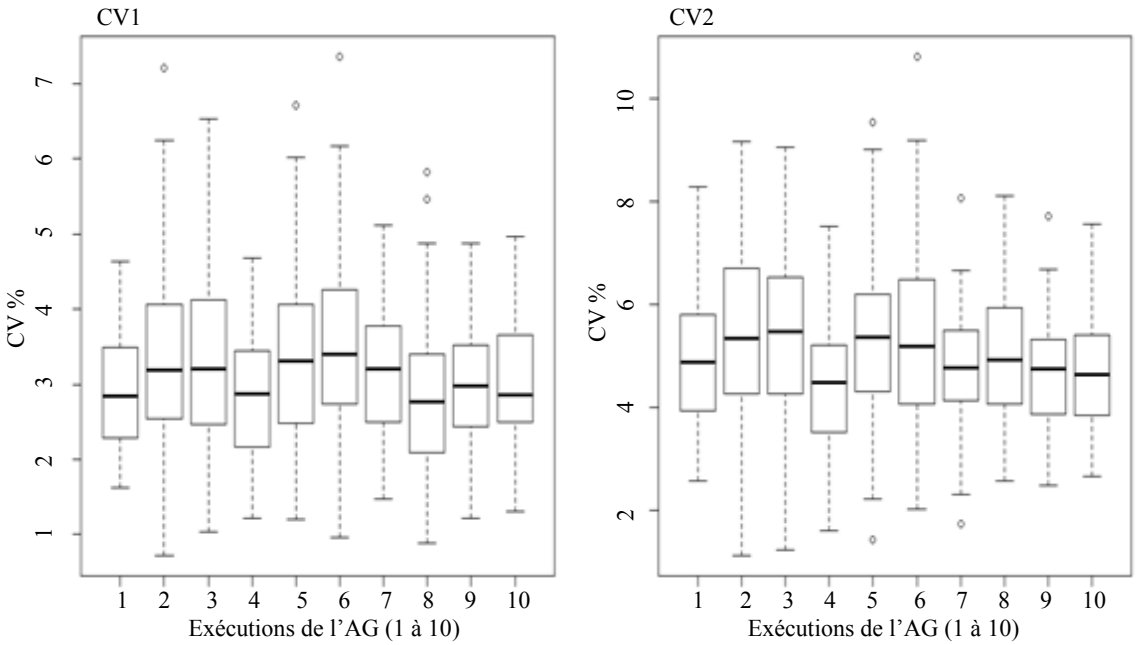


Figure 4.3 Distributions des CV pour les variables cibles dans les simulations

Un exemple plus complet comprenant l'utilisation de toutes les fonctions du module `SamplingStrata` est présenté dans Barcaroli (2013b).

5 Une application : l'*Enquête italienne sur la structure des exploitations agricoles* (ESEA)

La base de sondage utilisée pour la sélection de l'échantillon de l'*Enquête italienne sur la structure des exploitations agricoles de 2003* (ESEA) contient 2 153 710 exploitations agricoles. Pour l'établissement du plan de sondage de l'ESEA, les variables auxiliaires prises en considération sont les suivantes :

1. régions (21 valeurs différentes);
2. provinces (103 valeurs différentes);
3. statut juridique (2 classes);
4. secteur d'activité économique (9 classes);
5. unités de dimension économique (3 classes);
6. superficie agricole utilisée (3 classes);
7. unités de bétail (3 classes);
8. altimétrie du siège social de l'exploitation agricole (5 classes).

Quatorze variables cibles distinctes ont été prises en considération comme étant la cible principale de l'ESEA, pour lesquelles les niveaux de précision requis (en ce qui concerne la valeur maximale du coefficient de variation) ont été fixés à l'échelle régionale (domaines d'intérêt). La liste des variables et des contraintes de précision connexe est présentée au tableau 5.1.

Les 8 variables auxiliaires ainsi que les 14 variables cibles ont été observées durant le recensement de l'agriculture précédent de 2000, de sorte que leurs valeurs sont disponibles pour chaque unité présente dans la base de sondage. Il est donc possible de calculer les moyennes et les écarts-types se rapportant à n'importe quelle strate définie.

Pour commencer, nous décrivons la procédure « manuelle » courante suivie en 2003 pour choisir la stratification la plus appropriée pour sélectionner l'échantillon.

Configuration manuelle des strates de 2003 pour sélectionner l'échantillon de l'ESEA

À la première étape, on a défini une strate à tirage complet dans chaque région sur la base des caractéristiques locales. Les seuils pour la définition des strates à tirage complet ont été déterminés en appliquant la méthode de Hidiroglou (1986).

À la deuxième étape, on a effectué un choix entre une stratification fondée sur les provinces ou sur la région dans son ensemble, région par région, en se basant sur des considérations organisationnelles locales.

À la troisième étape, les six autres variables ont été utilisées l'une après l'autre dans chaque région ou province (selon le résultat obtenu à la deuxième étape) comme variables de stratification. Pour chacune de

ces options de stratification, on a calculé la taille optimale d'échantillon (la taille minimale d'échantillon dans chaque strate a été fixée à 50) (dans la fonction de coût, le coût fixe a été égalé à 0 et les coûts variables ont été fixés à 1 dans chaque strate atomique : donc, la fonction de coût coïncide avec la taille totale d'échantillon). La stratification donnant lieu à la taille globale d'échantillon minimale dans chaque région (habituellement définie sur différentes variables) a été considérée comme la sortie de cette étape.

À la quatrième étape, les cinq variables restantes ont été utilisées séparément pour affiner la stratification obtenue antérieurement. Pour chacune de ces spécifications affinées, la taille optimale d'échantillon a été calculée en considérant les mêmes contraintes que celles utilisées à l'étape 3.

Cette procédure par étape a été répétée sur une base régionale, en affinant la meilleure stratification obtenue à chaque étape en se servant des variables disponibles restantes jusqu'à ce que la stratification obtenue s'avère être moins efficace que la stratification de l'étape précédente.

De cette façon, la valeur totale de la taille d'échantillon planifiée a été fixée à 42 465 unités (en fait, la taille d'échantillon utilisée pour l'ESEA de 2003 a été portée à 52 713 afin d'obtenir de meilleures estimations au niveau national. Ici, nous considérons le chiffre de 42 465 afin que soit correcte la comparaison avec les résultats obtenus au moyen de l'algorithme génétique).

Utilisation de l'algorithme génétique pour déterminer les strates optimales et la meilleure répartition de l'échantillon

La stratification la plus détaillée disponible de la base de sondage, obtenue sous forme du produit cartésien de toutes les variables auxiliaires, comprend 24 454 strates distinctes, dont 1 787 sont définies comme étant des strates à tirage complet. Donc, les strates atomiques sont données par les 22 667 strates d'échantillonnage obtenues en soustrayant les 1 787 strates à tirage complet. Ces dernières sont regroupées en une seule strate, dont les 6 971 unités seront toujours sélectionnées quel que soit l'échantillon.

En fait, l'une des variables auxiliaires, *région*, est considérée comme la variable de domaine. Donc, notre tâche consiste à optimiser la stratification de la base de sondage et la répartition de l'échantillon séparément pour chacune des 21 régions de l'Italie. Par exemple, la première région (Piémont) est caractérisée par 105 074 unités dans 1 646 strates d'échantillonnage, et 597 unités dans 129 strates à tirage complet.

Les contraintes de précision (de nouveau exprimées en fonction des limites supérieures des coefficients de variation) ont été fixées, pour chacune des 14 variables cibles distinctes, aux mêmes valeurs que celles choisies durant la configuration manuelle des strates exécutée pour l'enquête de 2003 : ces limites sont 5 %, 6 % ou 10 % pour les variables les plus importantes dans chaque région. Le tableau 5.1 donne le jeu complet de coefficients de variation utilisé pour planifier l'ESEA de 2003.

Le tableau 5.2 donne les résultats des deux solutions en ce qui concerne la taille requise d'échantillon : celle prévue en 2003 par le spécialiste chargé de la conception de l'échantillon de l'ESEA (colonne 6) et celle obtenue en appliquant l'algorithme génétique (colonne 7).

Comme la détermination de la meilleure stratification a été effectuée séparément pour chaque région, 21 résultats indépendants attestent de la grande commodité de l'algorithme dans la plupart des domaines. On constate une diminution spectaculaire de la taille globale d'échantillon requise, comme en témoigne l'économie de 38,17 % par rapport au total antérieur. Ce résultat varie de région en région, la diminution

maximale étant observée pour la Sardaigne (-57,85 %) et la diminution minimale, pour la Sicile (-20,61 %). En outre, en ce qui concerne les strates, en partant du nombre initial de strates atomiques (22 667), on observe une réduction énorme à l'étape de la stratification finale, qui est caractérisée par 213 strates distinctes seulement (nombre variant d'un minimum de 6 strates dans la région de Frioul à 22 strates en Sicile).

Tableau 5.1
Coefficients de variation maximaux prévus (%) utilisés pour l'ESEA de 2003

Région	Céréales	Cultures industrielles	Légumes frais	Fleurs	Vignobles	Olives	Agrumes	Fruits	Bovins	Porcins	Ovins	Unités de dimension économique	Superficie agricole utilisée	Unités de bétail
Piémont	5,0	10,0			5,0				5,0			5,0	6,0	6,0
Vallée d'Aoste									5,0			5,0	6,0	6,0
Lombardie	5,0	10,0							5,0	5,0		5,0	6,0	6,0
Bolzano								5,0				5,0	6,0	6,0
Trente								5,0				5,0	6,0	6,0
Vénétie	5,0	10,0			5,0					5,0		5,0	6,0	6,0
Frioul-VJ	5,0	10,0										5,0	6,0	6,0
Ligurie				5,0								5,0	6,0	6,0
Émilie-Romagne	5,0	10,0			5,0			5,0	5,0	5,0		5,0	6,0	6,0
Toscane	5,0	10,0			5,0							5,0	6,0	6,0
Ombrie						5,0						5,0	6,0	6,0
Marches												5,0	6,0	6,0
Latium	5,0		5,0		5,0	5,0						5,0	6,0	6,0
Abruzzes						5,0						5,0	6,0	6,0
Molise						5,0						5,0	6,0	6,0
Campanie	5,0	10,0	5,0			5,0		5,0				5,0	6,0	6,0
Pouilles	5,0		5,0		5,0	5,0						5,0	6,0	6,0
Basilicate	5,0											5,0	6,0	6,0
Calabre	5,0					5,0	5,0					5,0	6,0	6,0
Sicile	5,0		5,0		5,0	5,0	5,0				5,0	5,0	6,0	6,0
Sardaigne	5,0										5,0	5,0	6,0	6,0

Pour ce qui est des paramètres utilisés pour obtenir le résultat susmentionné, les plus importants étaient les suivants :

1. nombre d'itérations (ou de générations);
2. taille de la génération (nombre d'individus, ou de solutions, évalué à chaque itération);
3. chances de mutation;
4. nombre initial de strates;
5. facteur d'accroissement du nombre initial de strates.

Tableau 5.2
Détermination de la taille de l'échantillon de l'ESEA de 2003 : comparaison des résultats

(1) Domaine (région)	(2) Nombre total d'unités dans la base de sondage	(3) Nombre de strates atomiques d'échantil-lonnage dans la base de sondage	(4) Nombre d'unités dans les strates d'échantil-lonnage	(5) Nombre d'unités dans les strates à tirage complet	(6) Taille de l'échantillon selon la stratification de 2003	(7) Taille de l'échantillon selon la solution de l'algorithme génétique	(8) Nombre de strates dans la solution de l'AG	(9) Différence relative en % (7) c. (6)
Piémont	105 671	1 646	105 074	597	2 687	1 497	9	-44,29
Vallée d'Aoste	6 125	65	6 074	51	408	317	7	-22,30
Lombardie	71 257	1 902	69 495	1 762	3 428	2 151	7	-37,25
Bolzano	23 362	127	23 202	160	692	430	7	-37,86
Trente	30 021	124	29 908	113	676	523	7	-22,63
Vénétie	176 999	1 450	176 064	935	3 531	1 868	11	-47,10
Frioul	32 981	638	32 805	176	807	498	6	-38,29
Ligurie	29 992	584	29 967	25	766	485	7	-36,68
Émilie-Romagne	103 702	2 157	102 922	780	2 584	2 022	11	-21,75
Toscane	107 288	1 959	106 964	324	2 099	1 337	16	-36,30
Ombrie	46 074	435	45 897	177	1 354	751	7	-44,53
Marches	60 439	1 005	60 271	168	918	488	8	-46,84
Latium	162 109	1 304	161 801	308	3 233	2 216	14	-31,46
Abruzzes	67 117	888	66 941	176	1 035	743	10	-28,21
Molise	28 890	375	28 834	56	1 190	630	6	-47,06
Campanie	212 145	1 271	211 833	312	2 559	1 883	13	-26,42
Pouilles	288 087	1 026	287 877	210	4 712	2 009	14	-57,36
Basilicate	68 470	504	68 355	115	703	493	7	-29,87
Calabre	145 812	1 624	145 654	158	2 798	1 792	17	-35,95
Sicile	295 637	2 345	295 472	165	3 955	3 140	22	-20,61
Sardaigne	91 532	1 238	91 329	203	2 330	982	7	-57,85
Italie	**2 153 710**	**22 667**	**2 146 739**	**6 971**	**42 465**	**26 255**	**213**	**-38,17**

Leurs valeurs finales ont été déterminées, après de nombreux essais, sur la base de l'analyse des exécutions pour chaque région.

En particulier, en inspectant le graphique de convergence, il est possible de voir si le nombre d'itérations est suffisant pour avoir la certitude que la solution finale est définitivement la meilleure qu'il est possible d'obtenir, ou si un nombre plus élevé d'itérations est nécessaire. Pour cela, on peut analyser le comportement des deux courbes du graphique : la courbe inférieure donne la *meilleure* valeur d'évaluation, tandis que la courbe supérieure donne la valeur *moyenne* d'évaluation. Lorsque la valeur moyenne d'évaluation continue à diminuer, de même que la meilleure valeur d'évaluation, cela vaut la peine de poursuivre les itérations. Lorsque la courbe de la meilleure valeur devient stablement constante (et, habituellement, que la courbe de la valeur moyenne commence à fluctuer vers le haut et le bas), aucun gain supplémentaire ne peut être attendu de nouvelles itérations. C'est ce que montre, par exemple, le graphique de convergence pour la région de Trente, à la figure 5.1.

Pour le paramètre *iterations*, une valeur de 5 000 s'est révélée commode. Pour les chances de mutation, nous avons constaté que 0,001 était une valeur appropriée : cela signifie que, pour tout chromosome dans le génome (toute valeur dans le vecteur $\boldsymbol{v}$), une mutation n'a lieu, en moyenne, qu'une fois sur mille. Un élément critique consiste à fixer le nombre initial de strates. Puisque la solution finale est très sensible au nombre de strates, nous avons décidé de laisser l'algorithme faire le choix. On peut, pour cela, comme nous l'avons déjà expliqué à la section 4, attribuer une faible valeur à *initialStrata*, et donner une valeur plus grande que 0 à *addStrataFactor* : cela permet à l'algorithme d'explorer les solutions correspondant à une grande gamme de nombre de strates. Dans notre expérience, nous avons fixé le nombre initial de strates à la valeur 5 et avons attribué une valeur de 0,01 au facteur d'accroissement du nombre initial de strates (cela signifie que, chaque fois qu'une mutation a lieu, il existe une probabilité de 1 % d'augmenter de 1 le nombre courant de strates).

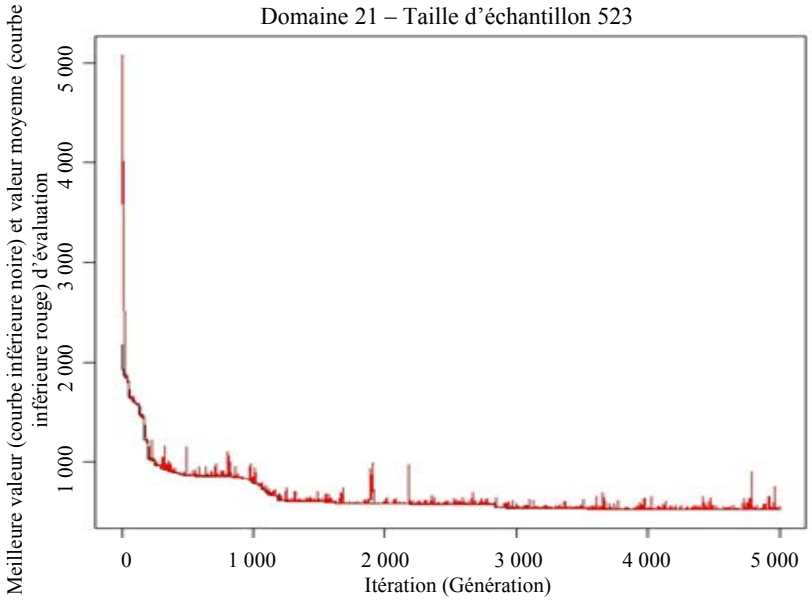


Figure 5.1 Meilleure valeur et valeur moyenne d'évaluation pour la région de Trente

Du point de vue des calculs, l'exécution de la tâche globale a pris 641 820 secondes (plus de 178 heures, près d'une semaine) (la tâche a été exécutée sur un ordinateur de bureau AMD Athlon 64 × 2 (2,90 Ghz, 3 GB RAM)).

6 Une autre application : l'*Enquête mensuelle sur le lait et les produits laitiers*

Notre algorithme a également été appliqué à l'*Enquête mensuelle sur le lait et les produits laitiers de 2010*. Il s'agit d'une enquête par sondage qui dépend strictement de l'« Enquête annuelle sur le lait et les produits laitiers », qui est un recensement de toutes les exploitations agricoles italiennes produisant du lait

et des produits laitiers. Les deux enquêtes recueillent la même information : la quantité de lait recueillie au niveau national et son utilisation (dans la transformation des produits laitiers : lait, fromage, beurre, *etc*.); l'objectif de l'enquête par sondage mensuelle est d'obtenir des renseignements à jour avant que les résultats de l'enquête annuelle (réalisée l'année précédente) soient disponibles. L'échantillon de 2010 a été planifié comme il suit :

1. l'information recueillie auprès des 2 250 unités qui avaient répondu au cycle de 2008 de l'enquête annuelle a été structurée comme une base de sondage : en particulier, quatre des variables cibles de l'enquête annuelle, qui sont continues, ont été transformées en variables catégoriques (facteurs ordonnés) en utilisant la méthode de classification automatique à k-moyennes, et ont été considérées comme information auxiliaire dans le base de sondage;
2. le produit croisé des variables catégoriques obtenues a donné une stratification de la base de sondage consistant en 152 strates (atomiques);
3. l'information reliée aux moyennes et aux écarts-types des quatre variables cibles de l'enquête mensuelle a été calculée pour chacune des strates atomiques en utilisant les données de l'enquête annuelle.

Les contraintes sur les coefficients de variation des estimations des totaux sont présentées au tableau 6.1.

Tableau 6.1
Coefficients de variation (%) utilisés pour planifier l'Enquête mensuelle sur le lait de 2010

Variable	CV maximal acceptable pour les estimations du total (%)
Lait recueilli	1
Lait	15
Beurre	3,8
Fromages de lait de vache	3

Après cela, l'algorithme de Bethel a été appliqué afin de vérifier quelle était la taille d'échantillon requise pour la stratification (atomique) initiale disponible pour la base de sondage (en outre, dans cette application, la fonction de coût coïncide avec la taille totale de l'échantillon, car le coût fixe a été fixé à 0, et les coûts variables ont été fixés à 1 dans chaque strate atomique) : cela a donné 290 unités à interviewer, réparties entre les 152 strates différentes. La procédure habituelle se termine ici : à ce stade, les 290 unités seraient sélectionnées dans la base de sondage représentée par l'enquête annuelle, puis l'enquête mensuelle débuterait.

Au lieu de cela, l'application de l'algorithme génétique a suggéré un regroupement des 152 strates atomiques initiales en 88 strates agrégées, nécessitant une taille d'échantillon de 247 seulement pour satisfaire les mêmes contraintes, c'est-à-dire une diminution d'environ 15 %.

Après de très nombreuses essais, les valeurs suivantes ont été données aux paramètres les plus importants :

1. la taille de la génération a été fixée à 50;
2. le nombre d'itérations a été fixé à 4 000;
3. un minimum de deux unités par strate a été exigé;
4. le nombre initial de strates (coïncidant avec le nombre maximal de celles-ci, parce que le paramètre *addStrataFactor* a été fixé à 0) a été pris égal au nombre de strates atomiques (152);
5. les chances de mutation ont été fixées à 0,0005.

La combinaison des paramètres « taille de la génération » et « nombre d'itérations » a déterminé l'évaluation de 200 000 $(50 \times 4\,000)$ solutions. Le graphique de convergence présenté à la figure 6.1 montre qu'après 2 700/2 800, plus aucune amélioration de la meilleure solution identifiée n'a eu lieu.

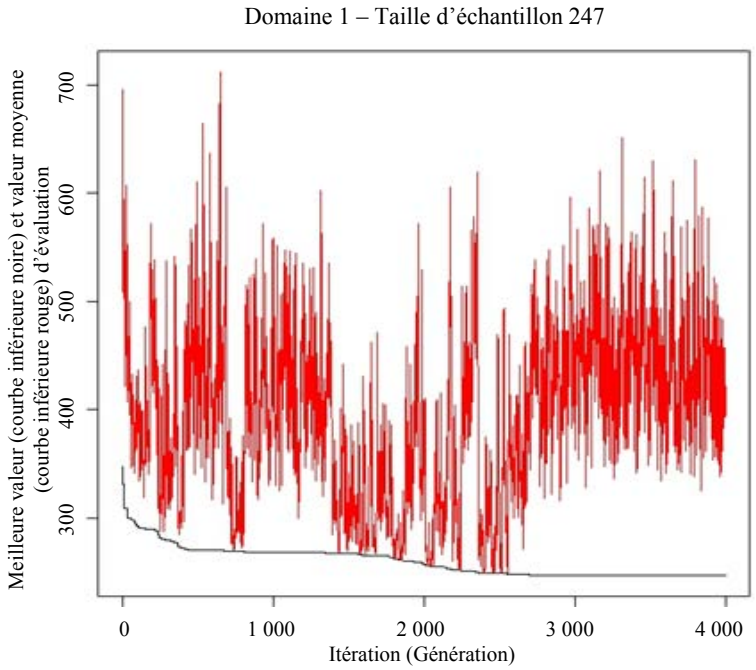


Figure 6.1 Meilleure valeur et valeur moyenne d'évaluation dans l'optimisation de l'Enquête mensuelle sur le lait

7 Conclusion et futurs travaux

Pour toute enquête par sondage polyvalente et à domaines multiples, la stratification optimale de la base de sondage peut être déterminée en même temps que la taille optimale de l'échantillon et la répartition optimale des unités entre les strates, en combinant l'utilisation de l'algorithme de Bethel (ou,

plus généralement, d'un résolveur de programmation non linéaire) pour déterminer la taille minimale d'échantillon requise pour satisfaire les contraintes de précision, et de l'algorithme génétique pour l'exploration de l'univers des stratifications potentielles, générées de façon rigoureuse conformément à la théorie des partitions. L'information requise est presque la même que celle nécessaire pour le problème de répartition de l'échantillon, à savoir la précision souhaitée pour les estimations du total (ou des moyennes) des variables cibles, et l'information concernant la distribution de chaque variable cible dans les strates de population. La stratification initiale doit être considérée au niveau le plus détaillé (stratification atomique), c'est-à-dire celle déterminée par le produit cartésien des valeurs de toutes les variables de stratification disponibles.

L'exploration complète de l'ensemble de toutes les stratifications possibles entraîne, dans certains cas, des calculs prohibitifs. L'utilisation de l'algorithme génétique permet d'explorer l'espace des solutions d'une manière très efficace. En ajustant minutieusement les paramètres d'exécution, il est possible de déterminer la solution optimale, ou du moins une solution s'écartant vraisemblablement peu de la solution optimale.

L'application de cet algorithme à deux enquêtes différentes (l'*Enquête italienne sur la structure des exploitations agricoles* de 2003 et l'*Enquête mensuelle sur le lait et les produits laitiers* de 2010) a montré que les solutions obtenues sont nettement meilleures, en ce qui concerne l'efficacité de l'échantillon, que celles produites manuellement par les méthodologistes spécialisés (à Istat, l'algorithme a été appliqué à trois autres enquêtes : *« Economic outcomes of agricultural holdings », « Structure and production of main wooden cultivations », « Survey on forecasting of some herbal crops sowing »*).

Dans tous les cas mentionnés, il a été possible de calculer les valeurs nécessaires comme données d'entrée dans notre algorithme (en particulier, les moyennes et les écarts-types des variables cibles dans les différentes strates atomiques), parce que les valeurs connexes figuraient dans la base de sondage pour chaque unité. Dans des situations plus réalistes, ce genre d'information n'est pas directement disponible. À sa place, nous pourrions utiliser des estimations produites en partant d'autres sources, comme des données administratives, d'autres enquêtes ou des cycles antérieurs de la même enquête, voire même d'hypothèses (habituellement prudentes) sur la variabilité des variables cibles dans les strates. Selon Rivest (2002), il est possible de modéliser les variables cibles en se servant des variables auxiliaires X comme variables explicatives, afin d'estimer les moyennes et les écarts-types sur la base des valeurs prédites de Y. Naturellement, la méthode proposée sera d'autant moins robuste que l'information sur les variables cibles est moins « directe », à cause de l'incertitude due à l'utilisation d'information indirecte ou de prédictions fondées sur un modèle.

Une autre limite de l'approche est celle liée au traitement des variables auxiliaires continues. Dans notre approche, nous suggérons simplement de les transformer en variables catégoriques afin de pouvoir en tenir compte dans la détermination de l'univers de toutes les stratifications possibles de la base de sondage. Un premier élément des prochains travaux consistera à donner des indications sur la façon de transformer ces variables afin d'obtenir la meilleur forme possible. Un deuxième élément tient au fait que certaines strates contenues dans la solution optimale peuvent être caractérisées par des valeurs non contiguës des variables continues transformées ou des variables catégoriques ordinales, une situation bizarre qui ne devrait pas être permise. Elle pourrait être évitée en imposant des contraintes sur la génération des solutions possibles.

Bibliographie

Baillargeon, S., et Rivest, L.-P. (2009). A general algorithm for univariate stratification. *Revue Internationale de Statistique,* 77, 3, 331-344.

Baillargeon, S., et Rivest, L.-P. (2011). Élaboration de plans stratifiés en R à l'aide du programme *stratification*. *Techniques d'enquête,* 37, 1, 59-72.

Barcaroli, G., Pagliuca, D. et Willighagen, E. (2013a). SamplingStrata: Optimal stratification of sampling frames for multipurpose sampling surveys. R package version 1.0-1. http://cran.r-project.org/web/packages/SamplingStrata/index.html.

Barcaroli, G. (2013b). Optimization of sampling strata with the SamplingStrata package. http://cran.r-project.org/web/packages/SamplingStrata/vignettes/SamplingStrataVignette.pdf.

Benedetti, R., Espa, G. et Lafratta, G. (2008). Une approche arborescente de la formation de strates dans les enquêtes-entreprises polyvalentes. *Techniques d'enquête,* 34, 2, 217-226.

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. *American Statistical Proceedings of the Survey Research Methods Section,* 209-212.

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête,* 15, 1, 49-60.

Chromy, J.B. (1987). Design optimization with multiple objectives. *Proceedings of the American Statistical Association Section on Survey Research Methods 1987*, 194-199.

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

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

Day, C.D. (2006). Application of an evolutionary algorithm to multivariate optimal allocation in stratified sampling designs. *Proceedings of the American Statistical Association Section on Survey Research Methods 2006* [CD-ROM].

Day, C.D. (2010). A multi-objective evolutionary algorithm for multivariate optimal allocation. *Section on Survey Research Methods - JSM 2010,* 3351-3358.

Díaz-García, J.A., et Cortez, L.U. (2008). Optimisation multi-objective pour une répartition optimale dans l'échantillonnage stratifié multivarié. *Techniques d'enquête,* 34, 2, 237-245.

Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête,* 30, 2, 177-185.

Hankin, R.K.S., et West, L.J. (2007). Set Partitions in R. *Journal of Statistical Software,* Code Snippet 2. December 2007, 23, http://www.jstatsoft.org/.

Hankin, R.K.S. (2011). Partitions: Additive partitions of integers. R package version 1.9-19. http://cran.r-project.org/web/packages/partitions/index.html.

Hartigan, J.A., et Wong, M.A. (1979). A k-means clustering algorithm. *Applied Statistics,* 28, 100-108.

Hidiroglou, M.A. (1986). The costruction of self-representing stratum of large units in survey design. *The American Statistician,* 40, 27-31.

Huddleston, H.F., Claypool, P.L. et Hocking, R.R. (1970). Optimal sample allocation to strata using convex programming. *Applied Statistics,* 19, 273-278.

Keskintürk, T., et Er, S. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics and Data Analysis,* 15 September 2007, 52, 1, 53-67.

Khan, M.G.M., Nand, N. et Ahmad, N. (2008). Détermination des bornes optimales de strate au moyen de la programmation dynamique. *Techniques d'enquête,* 34, 2, 227-236.

Khan, M.G.M., Maiti, T. et Ahsan, M.J. (2010). An optimal multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach. *Journal of Official Statistics,* 26, 4, 695-708.

Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society,* Series A, 159, 80-95.

Kozak, M., Verma, M.R. et Zieliński, A. (2007). Modern approach to optimum stratification: Review and perspectives. *Statistics in Transition,* 8(2), 223-250.

Kozak, M., Zieliński, A. et Singh, S. (2008). Stratified two-stage sampling in domains: Sample allocation between domains, strata, and sampling stages. *Statistics & Probability Letter,* Juin 2008, 78, 8, 970-974.

Kozak, M., et Wang, H.Y. (2010). On stochastic optimization in sample allocation among strata. *Metron - International Journal of Statistics,* LXVIII, 1, 95-103.

Lavallée, P., et Hidiroglou, M.A. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête,* 14, 1, 35-45.

Rivest, L.-P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête,* 28, 2, 207-214.

Schmitt, L.M. (2001). Theory of genetic algorithms. *Theoretical Computer Science,* 259, 1-61.

Schmitt, L.M. (2004). Theory of genetic algorithms II: Models for genetic operators over the string-tensor representation of populations and convergence to global optima for arbitrary fitness function under scaling. *Theoretical Computer Science,* 310, 181-231.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variables. *Journal of the American Statistical Association,* 66, 829-833.

Stokes, L., et Plummer, J. (2004). Using spreadsheet solvers in sample design. *Computational Statistics & Data Analysis,* 44, 527-546.

Vose, M.D. (1999). The Simple Genetic Algorithm: Foundations and Theory, MIT Press, Cambridge, MA.

Willighagen, E. (2012). Genalg: R Based Genetic Algorithm. R package version 0.1.1. http://cran.r-project.org/web/packages/genalg/index.html.

Techniques d'enquête, décembre 2013
Vol. 39, N° 2, pp. 433-457
Statistique Canada, N° 12-001-X au Catalogue

Un estimateur par la régression généralisée de la variation des prix des logements fondé sur des évaluations foncières

Jan de Haan et Rens Hendriks[1]

Résumé

Statistics Netherlands s'appuie sur la méthode du ratio prix de vente-évaluation ou méthode SPAR (pour *Sale Price Appraisal Ratio*) pour produire son indice des prix des logements. Cette méthode combine les prix de vente aux évaluations foncières faites par l'administration publique. Le présent article décrit une approche de rechange dans laquelle les évaluations foncières servent d'information auxiliaire dans un cadre de régression généralisée (GREG). Une application aux données des Pays-Bas montre que, même si l'indice GREG est plus lisse que le ratio des moyennes d'échantillon, il donne une série très semblable à la série SPAR. Pour expliquer ce résultat, nous montrons que l'indice SPAR est un estimateur de notre indice GREG plus général et qu'en pratique, il est presque aussi efficace.

Mots clés : Estimation par la régression généralisée; indice des prix des logements; évaluations foncières; échantillonnage.

1 Introduction

Lorsqu'ils essaient de construire des indices des prix des logements de qualité constante, les organismes statistiques doivent résoudre plusieurs problèmes. Premièrement, l'appariement exact des biens immobiliers au fil du temps est problématique, car la qualité de ces biens aura probablement évolué; les logements se déprécient et peuvent également avoir subi des réparations, des rénovations ou des rajouts importants. Autrement dit, chaque bien immobilier peut être considéré à chaque période comme un bien unique. Deuxièmement, le roulement des logements est généralement faible comparativement au parc de logements et la composition des biens immobiliers vendus évolue au fil du temps, ce qui pose un problème de qualité. Troisièmement, les données sur les caractéristiques des logements font souvent défaut, ce qui a des répercussions sur le choix de la méthode de mesure.

Trois grandes catégories d'indices des prix des logements sont décrites dans la littérature, à savoir les indices médians ou moyens, les indices fondés sur la méthode des ventes répétées et les indices hédoniques. Un indice médian (moyen) suit l'évolution du prix du logement médian (moyen) négocié d'une période à la suivante. Cette méthode pose problème du fait que les caractéristiques, disons, du logement médian évolue au cours du temps. On contourne souvent ce problème en stratifiant les échantillons selon la région, le type de logement, *etc*., procédure que l'on appelle aussi ajustement de la composition (*mix adjustment*). Évidemment, la stratification nécessite des données supplémentaires.

Les *méthodes des ventes répétées* abordent le problème de la composition qualitative (*quality mix*) en limitant le jeu de données aux logements qui ont été vendus au moins deux fois au cours de la période d'échantillonnage. On s'assure de cette façon de « comparer les mêmes choses », en supposant que la qualité des logements individuels ne change pas. Les méthodes des ventes répétées sont fondées sur des régressions dans lesquelles sont regroupées les données sur les ventes répétées pour différentes périodes.

1. Jan de Haan, OTB Research Institute for the Built Environment, Delft University of Technology and Division of Process Development, IT and Methodology, Statistics Netherlands, P.O. Box 24500, 2490 HA La Haye, Pays-Bas. Courriel : j.dehaan@cbs.nl; Rens Hendriks, Division of Economic and Business Statistics and National Accounts, Statistics Netherlands. Courriel : r.hendriks@cbs.nl.

Un inconvénient éventuel de cette approche est la révision; lorsque de nouvelles données sont ajoutées à l'échantillon, les indices calculés antérieurement changent. La méthode des ventes répétées a été proposée au départ par Bailey, Muth et Nourse (1963). Case et Shiller (1987, 1989) soutiennent que les variations des prix des logements comprennent des composantes dont la variance augmente avec l'intervalle entre les ventes et proposent une approche par les moindres carrés pondérés pour corriger ce genre d'hétéroscédasticité. Une autre méthode pondérée a été proposée par Calhoun (1996). Jansen, de Vries, Coolen, Lamain et Boelhouwer (2008), comparent, en se servant de données pour les Pays-Bas, la méthode des ventes répétées non pondérée à diverses méthodes pondérées et concluent que la méthode non pondérée donne des résultats satisfaisants.

Contrairement aux méthodes des ventes répétées, les *méthodes de régression hédoniques* permettent, en principe, de corriger les données pour tenir compte des changements de qualité des biens individuels (en plus des variations de composition qualitative). Ces méthodes requièrent des données sur les caractéristiques des logements, comme le nombre de chambres à coucher, et la taille et la localisation du terrain, pour estimer par la régression des indices de prix corrigés pour tenir compte de la qualité. Aujourd'hui, de nombreux pays calculent des indices hédoniques des prix des logements. Ainsi, un indice hédonique est produit par l'institut de statistique de la France (INSEE) en collaboration avec le Conseil supérieur du notariat (Gouriéroux et Laferrère 2009), ainsi que par Statistics Finland (Saarnio 2006). Au Royaume-Uni, trois indices hédoniques des prix des logements sont produits par différents instituts. En Australie, RPData-Rismark calcule des indices hédoniques pour les capitales (Hardman 2011). On distingue deux grandes catégories d'indices hédoniques. La méthode des variables indicatrices temporelles (*time dummies*) consiste à modéliser le logarithme du prix en fonction des caractéristiques du bien et d'un jeu de variables indicatrices représentant les périodes. Comme les données pour toutes les périodes sont regroupées, cette méthode souffre aussi de révisions. Les méthodes d'imputation hédoniques, qui consistent à estimer les « prix manquants », n'ont pas cet inconvénient. Hill et Melser (2008) discutent de nombreuses méthodes d'imputation hédoniques dans le contexte du logement. Diewert, Heravi et Silver (2009) et de Haan (2010) donnent une comparaison des indices de prix fondés sur la méthode des variables indicatrices temporelles et sur la méthode d'imputation hédonique.

Une quatrième approche en vue d'estimer l'indice des prix des logements est celle de l'*utilisation de données d'évaluation foncière*. Une option consiste à augmenter le jeu de données sur les ventes répétées en utilisant des données d'évaluation foncière comme estimation des valeurs passées ou courantes des biens immobiliers qui n'ont pas été revendus durant la période d'échantillonnage. Certaines données sur lesquelles repose l'indice fondé sur les ventes répétées seraient alors des pseudo-données plutôt que des données réelles sur ces ventes. Pour en savoir davantage sur l'utilisation des valeurs d'évaluation dans un indice des prix fondés sur les ventes répétées et sur l'élimination du biais d'évaluation, voir, par exemple, Geltner (1996), Edelstein et Quan (2006), et Leventis (2006). Une autre option, qui tient également compte des effets des variations de composition qualitative, consiste à combiner les prix de vente de la période courante avec les évaluations pour une période antérieure afin de calculer des prix relatifs (ratio de prix) dans un cadre classique d'appariement de biens. L'un des avantages par rapport à l'approche des ventes répétées tient au fait que les indices ne seront pas révisés. Cette méthode dite du ratio prix de vente-évaluation (*Sale Price Appraisal Ratio* ou SPAR) est appliquée depuis longtemps en Nouvelle-Zélande et est également utilisée maintenant aux Pays-Bas et dans quelques autres pays européens. Bourassa, Hoesli et Sun (2006) décrivent l'indice SPAR de la Nouvelle-Zélande qui est produit par Quotable Value, une

entreprise publique d'évaluation foncière. D'autres études de la méthode SPAR comprennent Rossini et Kershaw (2006), van der Wal, ter Steege et Kroese (2006), de Vries, de Haan, van der Wal et Mariën (2009), de Haan, van der Wal et de Vries (2009), Shi, Young et Hargreaves (2009), et Grimes et Young (2010).

Dans le présent article, nous décrivons une autre méthode fondée sur les évaluations foncières pour mesurer la variation des prix des logements. Les évaluations foncières servent d'information auxiliaire dans un cadre d'estimation par la *régression généralisée* (GREG). La régression GREG est une technique assistée par modèle qui peut être utilisée pour augmenter l'efficacité comparativement aux estimateurs plus simples tels que les moyennes d'échantillon (Särndal, Swensson et Wretman 1992), à condition que l'information sur la population soit connue pour une ou plusieurs variables présentant une forte corrélation linéaire avec la variable étudiée. Dans notre cas, nous calculons la régression des prix de vente à chaque période sur les évaluations foncières. Les valeurs d'évaluation sont disponibles aux Pays-Bas pour tous les biens faisant partie du parc immobiliers durant une période de référence donnée, et nous nous attendons à ce qu'elles présentent une forte colinéarité avec les prix de vente. Bien que la méthode repose sur la régression, l'indice des prix résultant n'est pas un indice hédonique, car le modèle de régression est descriptif plutôt qu'explicatif.

La présentation de l'article est la suivante. Pour préparer le terrain, à la section 2, nous décrivons la méthode SPAR et les liens entre cette dernière et les moyennes d'échantillon des prix de vente ainsi que les évaluations foncières. En raison du changement de composition et du nombre relativement faible de transactions, la série SPAR des Pays-Bas est caractérisée par une forte volatilité, surtout pour les petits créneaux du marché. À la section 3, nous décrivons un estimateur GREG simple et deux options de rechange. La première option est une version stratifiée de l'indice original, tandis que la deuxième repose sur une autre spécification du modèle. À la section 4, nous présentons des constatations empiriques obtenues en utilisant les données des Pays-Bas. Les indices GREG s'avèrent fort semblables aux indices SPAR et sont tout aussi volatils. À la section 5, nous expliquons ce résultat en montrant que l'indice SPAR est en fait un estimateur de l'indice GREG et qu'il est presque aussi efficace. À la section 6, nous présentons nos conclusions et proposons un sujet de future étude dans ce domaine.

2 Estimateurs de Horvitz-Thompson et l'indice SPAR

Habituellement, l'objectif de l'échantillonnage est d'estimer le total ou la moyenne (arithmétique) d'une variable donnée pour une population finie. Dans le contexte du logement, nous pourrions vouloir estimer la valeur totale du parc de logements, disons, à la période 0. Soit U^0 le parc de logements de taille N^0 et p_n^0 la valeur du logement $n\left(n = 1,\ldots, N^0\right)$. La valeur cible que l'on veut estimer est

$$V^0 = \sum_{n\in U^0} p_n^0. \tag{2.1}$$

Supposons que nous ayons un échantillon S^0 comprenant n^0 logements vendus durant la période de référence. Si les logements sont sélectionnés par échantillonnage aléatoire simple dans le parc de

logements U^0, où chaque logement possède la même probabilité d'inclusion, alors l'estimateur de Horvitz-Thompson

$$\hat{V}^0 = \left(N^0/n^0\right)\sum_{n=1}^{n^0} p_n^0 \tag{2.2}$$

est un estimateur sans biais de (2.1); voir, par exemple, Cochran (1977).

Une cible naturelle – quoi qu'il ne s'agisse pas de la seule possibilité – pour un indice des prix des logements serait la variation de la valeur d'un parc de logements fixe. Le conditionnement sur le *parc à la période de référence* a deux implications : les ajouts au parc (principalement des logements neufs) doivent être exclus et les variations des biens immobiliers existants doivent être ajustées pour tenir compte des changements de qualité, c'est-à-dire l'effet de la dépréciation, des rénovations et des rajouts. Pour simplifier, nous supposons que ces changements de qualité sont négligeables. Dans ces conditions, l'indice des prix cible en passant de la période de référence 0 à la période de comparaison $t\,(>0)$ est défini comme

$$P^{0t} = \frac{\sum_{n\in U^0} p_n^t}{\sum_{n\in U^0} p_n^0}, \tag{2.3}$$

la notation étant évidente. Supposons que nous ayons aussi un échantillon S^t, constitué de n^t logements vendus à la période t et supposons qu'il s'agit d'un tirage aléatoire indépendant fait dans le parc de logements à la période de référence. Le ratio des estimateurs de Horvitz-Thompson (les moyennes d'échantillon) aux deux périodes

$$\hat{P}^{0t} = \frac{\left(N^0/n^t\right)\sum_{n\in S^t} p_n^t}{\left(N^0/n^0\right)\sum_{n\in S^0} p_n^0} = \frac{\sum_{n\in S^t} p_n^t/n^t}{\sum_{n\in S^0} p_n^0/n^0} \tag{2.4}$$

peut sembler être un estimateur naturel de notre indice cible (2.3). Toutefois, si les échantillons S^0 et S^t sont tirés indépendamment, la variance de l'estimateur (2.4) risque d'être considérable. En outre, un ratio estimé tel que (2.4) présente un biais qui dépend de la variance du numérateur et de la covariance du numérateur et du dénominateur (Cochran 1977). Dans la perspective d'un indice, le problème important est que la composition des biens négociés à la période t n'est pas la même qu'à la période 0. Autrement dit, nous ne comparons pas les mêmes choses.

L'approche classique d'estimation des indices de prix repose sur les méthodes d'appariement de modèles où les prix p_n^0 et p_n^t sont observés pour un panel fixe d'articles. L'utilisation de données de panel permet de s'assurer que l'on compare des articles qui sont les mêmes, ce qui réduit la variance de l'estimateur par le ratio, parce que p_n^0 et p_n^t sont habituellement corrélés positivement. Cependant, à moins que les échantillons S^0 et S^t soient extraordinairement grands, on n'obtient que quelques appariements de logements, si tant qu'il y en ait. Donc, alors que les prix p_n^t sont observés pour les logements appartenant à S^t, les prix à la période de référence p_n^0 « manquent » pour la plupart de ces

logements. Les données qui pourraient par contre être disponibles sont les évaluations foncières de l'administration publique a_n^0. Nous pourrions utiliser ces évaluations comme valeurs à la période de référence et construire l'estimateur par « pseudo » appariement de modèles qui suit de la variation des prix des logements :

$$\tilde{P}^{0t} = \frac{\sum_{n \in S^t} p_n^t / n^t}{\sum_{n \in S^t} a_n^0 / n^t}. \tag{2.5}$$

Un problème que pose l'estimateur (2.5) est que l'indice à la période de référence ne sera pas égal à 1, parce que les évaluations a_n^0 diffèrent des prix de vente p_n^0. Le rééchelonnement de l'estimateur (2.5) en le divisant par sa valeur à la période de référence est une solution évidente, qui donne

$$\hat{P}_{\text{SPAR}}^{0t} = \frac{\sum_{n \in S^t} p_n^t / n^t}{\sum_{n \in S^t} a_n^0 / n^t} \left[\frac{\sum_{n \in S^0} p_n^0 / n^0}{\sum_{n \in S^0} a_n^0 / n^0} \right]^{-1} = \frac{\sum_{n \in S^t} p_n^t / n^t}{\sum_{n \in S^0} p_n^0 / n^0} \left[\frac{\sum_{n \in S^0} a_n^0 / n^0}{\sum_{n \in S^t} a_n^0 / n^t} \right]. \tag{2.6}$$

Notons que le facteur de rééchelonnement est stochastique, car il s'agit d'un ratio de moyennes d'échantillon pour la période de référence, ce qui augmentera la variance de (2.6) comparativement à l'estimateur donné par (2.5), en fonction des corrélations entre les évaluations foncières et les prix de vente. Des renseignements détaillées figurent dans de Haan (2007). Cependant, nous ne pouvons pas contourner le rééchelonnement, puisqu'un indice de prix dont la valeur initiale n'est pas égale à 1 n'aurait pas de sens.

L'expression (2.6) est appelée indice du ratio prix de vente-évaluation (*sale price appraisal ratio*) ou indice SPAR. La méthode SPAR est appliquée aux Pays-Bas depuis janvier 2008 pour mesurer le changement de prix des logements occupés par le propriétaire. Comme il est mentionné plus haut, nous supposons que l'indice SPAR a pour objectif de suivre l'évolution du prix du *parc de logements*, qui est une mesure de la variation du patrimoine. Par ailleurs, dans le contexte de l'Indice harmonisé des prix à la consommation, l'indice des prix des logements doit mesurer le changement de prix des *logements vendus* durant la période de référence (Makaronidis et Hayes 2006; Eurostat 2010). Sous ce dernier concept, aucun échantillonnage ne doit être effectué si toutes les transactions sont enregistrées et utilisées dans le calcul de l'indice, comme cela est le cas aux Pays-Bas.

Le deuxième membre de l'équation (2.6) exprime l'indice SPAR sous la forme du produit de deux facteurs, le ratio des moyennes d'échantillon et un facteur entre crochets. Comme l'indice SPAR est essentiellement fondé sur la méthode d'appariement de modèles (en utilisant des évaluations à la place des prix de vente à la période de référence), ce facteur rajuste le ratio des moyennes d'échantillon pour tenir compte des changements de composition qualitative des échantillons qui ont lieu entre la période 0 et la période t. Un problème éventuel est que l'indice SPAR *n'est pas un estimateur de type panel*. Par conséquent, une série chronologique SPAR, disons pour les périodes $t = 0, \ldots, T$, pourrait souffrir d'une volatilité dans le court terme due à des changements de composition, surtout si le nombre de ventes est faible.

3 Estimation par la régression généralisée

3.1 Une méthode GREG simple

A la présente section, nous décrivons une approche de rechange pour mesurer le changement de prix des logements qui s'appuie sur des données d'évaluation. Les évaluations foncières servent maintenant d'information auxiliaire dans un cadre de régression généralisée (GREG). Considérons le simple modèle de régression linéaire à deux variables suivant :

$$p_n^0 = \alpha^0 + \beta^0 a_n^0 + \varepsilon_n^0, \tag{3.1}$$

où ε_n^0 est le terme d'erreur. Contrairement aux modèles de régression hédonique, qui postulent une relation causale entre le prix de vente p_n^0 et un jeu de caractéristiques ayant trait à la structure et à la localisation des unités de logement, ce modèle ne dit rien sur la façon dont les prix des logements sont produits; l'équation (3.1) est simplement un modèle descriptif.

L'estimation du modèle (3.1) par la régression par les moindres carrés sur les données de l'échantillon S^0 donne les prix prédits

$$\hat{p}_n^0 = \hat{\alpha}^0 + \hat{\beta}^0 a_n^0. \tag{3.2}$$

Les résidus de la régression pour $n \in S^0$ sont $e_n^0 = p_n^0 - \hat{p}_n^0$. En supposant un échantillonnage aléatoire, comme auparavant, nous pouvons écrire l'estimateur de Horvitz-Thompson $\sum_{n\in S^0} p_n^0 / n^0$ de la valeur moyenne $\sum_{n\in U^0} p_n^0 / N^0$ sous la forme

$$\sum_{n\in S^0} p_n^0 / n^0 = \sum_{n\in S^0} \hat{p}_n^0 / n^0 + \sum_{n\in S^0} e_n^0 / n^0 = \hat{\alpha}^0 + \hat{\beta}^0 \sum_{n\in S^0} a_n^0 / n^0 + \sum_{n\in S^0} e_n^0 / n^0. \tag{3.3}$$

Le remplacement de la moyenne d'échantillon des évaluations, $\sum_{n\in S^0} a_n^0 / n^0$, par son équivalent pour la population, $\sum_{n\in U^0} a_n^0 / N^0$ donne l'estimateur par la régression généralisée (GREG) :

$$\hat{\bar{P}}_{\text{GREG}}^0 = \hat{\alpha}^0 + \hat{\beta}^0 \sum_{n\in U^0} a_n^0 / N^0 + \sum_{n\in S^0} e_n^0 / n^0 = \sum_{n\in U^0} \hat{p}_n^0 / N^0 + \sum_{n\in S^0} e_n^0 / n^0. \tag{3.4}$$

La théorie de l'échantillonnage assisté par modèle montre que les estimateurs GREG sont *asymptotiquement sans biais sous le plan de sondage* (Särndal et coll. 1992), quel que soit le choix des variables explicatives. À moins que l'échantillon soit petit, le biais peut être négligé. Il est évident que l'estimateur GREG (3.4) sera plus efficace – au sens où sa variance est plus faible – que l'estimateur de Horvitz-Thompson (3.3). Par conséquent, l'estimateur GREG donnera habituellement de meilleurs résultats que l'estimateur de Horvitz-Thompson en termes d'erreur quadratique moyenne (la somme de la variance et du carré du biais).

La même procédure peut être appliquée à la période de comparaison t. Après avoir estimé le modèle

$$p_n^t = \alpha^t + \beta^t a_n^0 + \varepsilon_n^t \tag{3.5}$$

par la régression par les moindres carrés sur les données de l'échantillon de la période courante S^t, nous obtenons les prix prédits

$$\hat{p}_n^t = \hat{\alpha}^t + \hat{\beta}^t a_n^0, \tag{3.6}$$

ce qui mène à l'estimateur GREG de la valeur moyenne du parc de logements à la période t :

$$\hat{\bar{p}}_{\text{GREG}}^t = \hat{\alpha}^t + \hat{\beta}^t \sum_{n\in U^t} a_n^0 / N^t + \sum_{n\in S^t} e_n^t / n^t = \sum_{n\in U^t} \hat{p}_n^t / N^t + \sum_{n\in S^t} e_n^t / n^t, \tag{3.7}$$

où $e_n^t = p_n^t - \hat{p}_n^t$ désigne les résidus de la régression à la période t. Pour un parc de logements fixe, nous avons $U^t = U^0$, d'où $\sum_{n\in U^t} a_n^0 / N^t = \sum_{n\in U^0} a_n^0 / N^0$, et il s'ensuit que

$$\hat{\bar{p}}_{\text{GREG}}^t = \hat{\alpha}^t + \hat{\beta}^t \sum_{n\in U^0} a_n^0 / N^0 + \sum_{n\in S^t} e_n^t / n^t = \sum_{n\in U^0} \hat{p}_n^t / N^0 + \sum_{n\in S^t} e_n^t / n^t. \tag{3.8}$$

L'estimateur GREG du changement de prix des logements s'obtient simplement en prenant le ratio des équations (3.8) et (3.4):

$$\hat{P}_{\text{GREG}}^{0t} = \frac{\hat{\bar{p}}_{\text{GREG}}^t}{\hat{\bar{p}}_{\text{GREG}}^0} = \frac{\hat{\alpha}^t + \hat{\beta}^t \bar{a}^0 + \sum_{n\in S^t} e_n^t / n^t}{\hat{\alpha}^0 + \hat{\beta}^0 \bar{a}^0 + \sum_{n\in S^0} e_n^0 / n^0} = \frac{\sum_{n\in U^0} \hat{p}_n^t / N^0 + \sum_{n\in S^t} e_n^t / n^t}{\sum_{n\in U^0} \hat{p}_n^0 / N^0 + \sum_{n\in S^0} e_n^0 / n^0}, \tag{3.9}$$

où $\bar{a}^0 = \sum_{n\in U^0} a_n^0 / N^0$. Un certain biais de petit échantillon supplémentaire sera introduit en raison de la structure non linéaire (ratio). Lorsque l'on utilise la régression par les moindres carrés ordinaires (MCO) pour estimer les modèles (3.1) et (3.5), les moyennes d'échantillon non pondérées des résidus de la régression dans (3.9), $\sum_{n\in S^0} e_n^0 / n^0$ et $\sum_{n\in S^t} e_n^t / n^t$, sont égales à 0 et l'indice GREG se réduit à

$$\hat{P}_{\text{GREG,MCO}}^{0t} = \frac{\sum_{n\in U^0} \hat{p}_n^t / N^0}{\sum_{n\in U^0} \hat{p}_n^0 / N^0} = \frac{\hat{\alpha}^t + \hat{\beta}^t \bar{a}^0}{\hat{\alpha}^0 + \hat{\beta}^0 \bar{a}^0} = \frac{\hat{\alpha}^t / \bar{a}^0 + \hat{\beta}^t}{\hat{\alpha}^0 / \bar{a}^0 + \hat{\beta}^0}. \tag{3.10}$$

Comme l'indique la première expression dans le membre de droite de l'équation (3.10), l'approche GREG (MCO) consiste essentiellement à imputer les prix pour la période de référence et pour la période courante en utilisant les équations (3.2) et (3.6). La différence par rapport à la méthode hédonique d'*imputation double* tient à deux aspects : nous utilisons un modèle descriptif, et non un modèle hédonique, pour estimer les prix prédits – de sorte que nous ne pouvons pas parler de prix prédits sans biais – et nous imputons les prix de tous les logements faisant partie du parc au lieu de ceux du sous-ensemble de logements échantillonnés.

3.2 Propriétés de l'indice GREG

L'indice GREG (MCO) possède plusieurs propriétés intéressantes. Premièrement, son calcul est très simple. Une fois que l'on a calculé la moyenne de population des évaluations foncières $\bar{a}^0$ et les coefficients de régression pour la période de référence $\hat{\alpha}^0$ et $\hat{\beta}^0$, il suffit d'exécuter chaque mois une

régression des prix de vente en fonction des évaluations, puis d'introduire les valeurs des coefficients $\hat{\alpha}^t$ et $\hat{\beta}^t$ dans (3.10). Notons que l'indice GREG peut s'écrire sous la forme d'un pseudo indice-chaîne :

$$\hat{P}^{0t}_{\text{GREG,MCO}} = \frac{\hat{\alpha}^t/\bar{a}^0 + \hat{\beta}^t}{\hat{\alpha}^0/\bar{a}^0 + \hat{\beta}^0} = \prod_{\tau=1}^{t} \frac{\hat{\alpha}^\tau/\bar{a}^0 + \hat{\beta}^\tau}{\hat{\alpha}^{\tau-1}/\bar{a}^0 + \hat{\beta}^{\tau-1}}. \tag{3.11}$$

Cela peut être utile en pratique, surtout quand de nouvelles données d'évaluation deviennent disponibles. Les nouvelles évaluations sont souvent fournies à l'organisme statistique avec un délai important, pouvant dépasser un an. Les évaluations les plus récentes doivent être utilisées pour deux raisons. La qualité des évaluations peut s'améliorer avec le temps, ce qui semble avoir été le cas aux Pays-Bas (de Vries et coll. 2009). En outre, l'hypothèse d'un parc de logements fixe peut être relâchée afin que les logements nouvellement construits puissent être intégrés par enchaînement; l'indice GREG chaîné tient compte de la dynamique du parc de logements. Les mêmes avantages de l'enchaînement s'appliquent à la méthode SPAR. Supposons que de nouvelles évaluations foncières, se rapportant à la période $T\,(0 < T \leq t)$, soient disponibles à la période $t+1$. La série chronologique peut alors être mise à jour par enchaînement, c'est-à-dire en multipliant $\hat{P}^{0t}_{\text{GREG,MCO}}$ par la variation d'un mois à l'autre $\left(\tilde{\alpha}^{t+1}/\bar{a}^T + \tilde{\beta}^{t+1}\right)/\left(\tilde{\alpha}^t/\bar{a}^T + \tilde{\beta}^t\right)$, où les coefficients sont maintenant ceux d'une régression des prix de vente sur les évaluations foncières à la période T.

Deuxièmement, les *erreurs-types* de l'indice GREG peuvent être estimées assez facilement en utilisant la matrice de variance-covariance des coefficients de régression, qui est une sortie standard de la plupart des progiciels statistiques. Une expression de l'erreur-type approximative est dérivée en annexe. L'erreur-type de l'indice GEEG dépend de la qualité de l'ajustement (R^2) du modèle de régression. Il est fort probable que la valeur de R^2 pour la régression à la période de référence soit plus élevée que pour les régressions à la période courante. Nous nous attendons en effet à observer une forte relation linéaire entre les évaluations foncières et les prix de vente à la période de référence des évaluations, mais une relation probablement plus faible aux périodes ultérieures en raison des différences de tendance des prix selon le type de logement ou la région. Il est un peu plus compliqué d'établir une expression pour les erreurs-types approximatives dans le cas de l'indice SPAR, parce que la variabilité d'échantillonnage des évaluations moyennes est une source additionnelle d'erreur d'échantillonnage; voir de Haan (2007).

Cette dernière remarque nous mène à la troisième propriété de l'indice GREG, c'est-à-dire sa dépendance à l'égard de la *qualité des données d'évaluation*. Pour au moins deux raisons, il peut arriver que les évaluations foncières ne représentent pas exactement les prix de transaction durant la période de référence, de sorte que l'ajustement du modèle n'est pas parfait $(R^2 < 1)$. Les organismes chargés des évaluations pourraient ne pas avoir accès (en temps réel) aux prix de vente réels et, par conséquent, être obligés d'exercer leur propre jugement en se basant sur d'autres renseignements. Toutefois, même s'ils connaissaient les prix de vente, ces organismes pourraient encore décider de faire des ajustements lorsqu'ils déterminent la valeur des biens immobiliers. On peut soutenir que le prix de vente ne mesure pas toujours correctement la valeur de marché inconnue – laquelle peut être considérée comme une variable latente – et a tendance à être plus volatile. À cet égard, Francke (2010) et d'autres ont utilisé le terme de bruit de transaction.

La manière dont les évaluations foncières ont été déterminées aura une incidence sur l'erreur-type de l'indice GREG. À condition que la qualité des données d'évaluation soit la même pour tous les logements

compris dans le parc, il n'existe aucun biais, puisque les évaluations servent seulement de variables auxiliaires dans les régressions exécutées sur les échantillons S^0 et S^t de biens immobiliers vendus aux périodes 0 et $t\left(t = 1, \ldots, T\right)$. Cependant, en général, nous nous attendons à ce que la qualité des évaluations soit meilleure pour les biens appartenant à l'échantillon de la période de référence où a eu lieu l'évaluation S^0, quoique cela varie fort probablement en fonction de la méthode d'évaluation. Aux Pays-Bas, les biens immobiliers sont évalués aux fins de l'impôt (impôt sur le revenu ainsi que les impôts municipaux). Les municipalités sont chargées des évaluations. Plusieurs d'entre elles évaluent les logements qui sont vendus durant la période de référence (janvier) au moyen du prix de vente. Les logements qui n'ont pas été vendus sont parfois évalués en les comparant à des logements négociés similaires. Il semble que certaines municipalités utilisent une forme de régression hédonique pour évaluer les logements, mais la méthodologie n'a malheureusement pas été rendue publique. Pour plus de renseignements sur le système d'évaluation foncière des Pays-Bas, voir de Vries et coll. (2009).

Jusqu'à présent, nous avons supposé que la qualité des logements individuels ne varie pas au fil du temps. Cette hypothèse est forte. Donc, la quatrième propriété – et l'inconvénient le plus important – de la méthode GREG est que l'indice des prix résultants est entaché d'un *biais de changement de qualité* puisque l'on n'effectue pas d'ajustement explicite de la qualité. La méthode SPAR ainsi que la méthode classique fondée sur les ventes répétées présentent le même inconvénient. En principe, les méthodes de régression hédonique permettent de traiter le problème du changement de qualité, quoi qu'il puisse s'avérer difficile d'utiliser des variables de contrôle pour toutes les caractéristiques influant sur le prix pertinentes, en particulier la microlocalisation. La méthode SPAR tient compte automatiquement de la microlocalisation, à condition naturellement que les évaluations foncières en tiennent suffisamment compte, puisqu'elle est basée sur la méthode d'appariement de modèles pour laquelle l'appariement est effectué au niveau de l'adresse.

3.3 Estimateur GREG de rechange

Statistics Netherlands calcule les indices des prix des logements non seulement pour l'ensemble du pays, mais aussi pour certains créneaux du marché du logement, selon le type de logement (logements familiaux et appartements) et la région (provinces et grandes villes), principalement pour répondre aux besoins des utilisateurs. L'échantillon peut aussi être stratifié afin d'atténuer l'effet du *biais de sélection dans l'échantillon*. Ce type de biais peut survenir si l'ensemble de logements vendus durant une période particulière n'est pas une sélection aléatoire provenant du parc de logements. L'indice national doit alors être calculé indirectement sous forme d'une moyenne pondérée des indices de strate plutôt que directement d'après toutes les observations.

Supposons que le parc total de logements U^0 est subdivisé en K strates non chevauchantes U_k^0 de taille $N_k^0 \left(\sum_{k=1}^K N_k^0 = N^0\right)$. L'indice des prix cible (2.3) peut alors être réécrit sous la forme

$$P^{0t} = \frac{\sum_{n\in U^0} p_n^t}{\sum_{n\in U^0} p_n^0} = \frac{\sum_{k=1}^K \sum_{n\in U_k^0} p_n^t}{\sum_{k=1}^K \sum_{n\in U_k^0} p_n^0} = \sum_{k=1}^K s_k^0 P_k^{0t}, \tag{3.12}$$

où $P_k^{0t} = \sum_{n\in U_k^0} p_n^t \Big/ \sum_{n\in U_k^0} p_n^0$ est l'indice des prix cible pour la strate $U_k^0 \left(k = 1,\ldots,K\right)$. Les parts de la valeur du parc de logements à la période de référence $s_k^0 = \sum_{n\in U_k^0} p_n^0 \Big/ \sum_{n\in U^0} p_n^0$, qui servent de pondérations pour les indices de strate, sont inconnues et doivent être estimées. En supposant que l'on connaît les variables qui définissent les strates pour tout $n \in U^0$, un choix naturel pour les pondérations serait les parts fondées sur l'évaluation foncière $\hat{s}_k^0 = \sum_{n\in U_k^0} a_n^0 \Big/ \sum_{n\in U^0} a_n^0 = \left(N_k^0 \Big/ N^0\right)\left(\overline{a}_k^0 \Big/ \overline{a}^0\right)$. Manifestement, les variables de logement qui définissent les strates doivent être incluses dans le jeu de données d'évaluation. Aux Pays-Bas, l'adresse et le type de logement sont inclus. Cela permet une subdivision de la population en strates obtenues par classification croisée de la localisation et du type de logement. Les évaluations foncières ne sont peut-être pas toujours des estimations exactes de la valeur de marché « réelle » des biens immobiliers individuels, mais au niveau de la strate, nous nous attendons à ce que l'exactitude des évaluations moyennes soit suffisante pour le calcul des pondérations.

Des techniques statistiques telles que l'estimation GREG sont habituellement appliquées pour estimer les totaux ou les moyennes pour de petits domaines pour lesquels le nombre d'observations est si faible que les erreurs-types lorsque l'on utilise les estimateurs classiques (de Horvitz-Thompson) – ici le ratio des moyennes d'échantillon – deviendraient inacceptablement grandes. Il convient de mentionner que, même avec la méthode GREG, le schéma de stratification ne doit pas être trop détaillé, car cela pourrait accroître excessivement la variance des indices de strate, et donc, de l'indice agrégé. Fait peut-être encore plus important, le biais de petit échantillon augmentera au point de devenir éventuellement non négligeable pour les très petits échantillons.

Les régressions par les MCO des prix de vente sur les évaluations foncières doivent maintenant être exécutées à chaque période pour chaque strate afin de calculer l'indice GREG agrégé. L'indice GREG (MCO) stratifié est donné par

$$\hat{P}_{\text{StrGREG}}^{0t} = \sum_{k=1}^{K} \hat{s}_k^0 \hat{P}_{k,\text{GREG,MCO}}^{0t} = \sum_{k=1}^{K} \hat{s}_k^0 \left(\frac{\hat{\alpha}_k^t \Big/ \overline{a}_k^0 + \hat{\beta}_k^t}{\hat{\alpha}_k^0 \Big/ \overline{a}_k^0 + \hat{\beta}_k^0} \right); \tag{3.13}$$

Les écarts entre les coefficients de pente $\hat{\beta}_k^s \left(s = 0, t\right)$ d'une strate à l'autre pourraient résulter de l'erreur d'échantillonnage ou refléter un phénomène réel. Celui-ci peut avoir une importance particulière pour les périodes t très éloignées de la période 0, car les différents créneaux du marché du logement ont tendance à présenter des tendances des prix variables. On pourrait effectuer un test afin de savoir si tout écart entre les coefficients de pente reflète un phénomène réel.

Un modèle de rechange, à estimer sur le jeu complet de données, comprendrait un terme d'ordonnée à l'origine unique, mais des coefficients β pouvant varier d'une strate à l'autre. Soit $D_{n,k}$ une variable indicatrice binaire qui prend la valeur 1 si le bien immobilier n appartient à la strate k et 0 autrement. À la période $s\left(s = 0, t\right)$, le modèle

$$p_n^s = \alpha^s + \sum_{k=1}^{K} \beta_k^s D_{n,k} a_n^0 + \varepsilon_n^s \tag{3.14}$$

est estimé par la régression par les MCO sur les données de l'échantillon S^s, ce qui donne les prix prédits $\tilde{p}_n^s = \tilde{\alpha}^s + \tilde{\beta}_k^s a_n^0$ pour $n \in U_k^0$. De nouveau, la somme des résidus est égale à zéro et le nouvel indice GREG (OMC) (non stratifié) devient

$$\tilde{P}_{\text{GREG,MCO}}^{0t} = \frac{\sum_{n\in U^0} \tilde{p}_n^t / N^0}{\sum_{n\in U^0} \tilde{p}_n^0 / N^0} = \frac{\sum_{k=1}^{K}\sum_{n\in U_k^0} \tilde{p}_n^t / N^0}{\sum_{k=1}^{K}\sum_{n\in U_k^0} \tilde{p}_n^0 / N^0} = \frac{\tilde{\alpha}^t + \sum_{k=1}^{K}\left(\frac{N_k^0}{N^0}\right)\tilde{\beta}_k^t \bar{a}_k^0}{\tilde{\alpha}^0 + \sum_{k=1}^{K}\left(\frac{N_k^0}{N^0}\right)\tilde{\beta}_k^0 \bar{a}_k^0}. \quad (3.15)$$

Le modèle (3.14) est plus souple que le modèle original donnée par les équations (3.1) et (3.5), et pourrait être utile si la proportionnalité entre les prix de vente et les évaluations foncières n'est pas respectée. L'estimateur (3.15) se réduit à l'indice GREG original (3.10) si les coefficients $\tilde{\beta}_k^s$ sont tous égaux. En pratique, cela n'arrivera pas et (3.15) et (3.10) donneront des réponses différentes. Une raison fréquemment avancée pour justifier l'utilisation des estimateurs GREG est que, étant asymptotiquement sans biais, ils sont relativement *robustes au choix du modèle*. Donc, nous nous attendrions à ce que l'effet de la spécification du modèle de rechange (3.15) soit modéré. Par ailleurs, il est généralement reconnu dans la littérature que l'indépendance à l'égard du modèle peut être un problème dans des circonstances particulières, notamment lorsqu'on a affaire à des populations très variables et ayant tendance à présenter des valeurs aberrantes. Par exemple, Hedlin, Falvey, Chambers et Kokic (2001) soulignent qu'il est important de procéder à une recherche minutieuse des spécifications du modèle, tandis que Beaumont et Alavi (2004) se concentrent sur le traitement des valeurs aberrantes. Il serait donc utile d'examiner l'effet de la spécification de ce modèle de rechange.

4 Illustration empiriques

Pour l'étude empirique, nous avons utilisé deux jeux de données de sources différentes. Le premier contient les prix de vente pour presque toutes les transactions concernant des logements existants (à l'exclusion des logements neufs) aux Pays-Bas effectuées entre janvier 2003 et mars 2009, telles qu'elles ont été enregistrées par le bureau du cadastre des Pays-Bas. Le nombre total d'observations est de 1 126 242, soit environ 15 000 par mois. Les ventes ont été enregistrées au moment de la signature de l'acte de vente au cabinet du notaire, en moyenne six semaines après la signature du compromis de vente. Le deuxième jeu de données contient les évaluations foncières de l'administration publique ayant trait à janvier 2003 pour tous les logements occupés par le propriétaire compris dans le parc de logements. Comme les adresses sont disponibles dans les deux jeux de données, nous connaissons le prix de vente et la valeur d'évaluation pour chaque transaction. Comme le type de logement est également disponible, nous avons pu stratifier l'échantillon selon le type et la localisation du logement.

Pour commencer, nous avons exécuté des régressions par les MCO non stratifiées des prix de vente sur les évaluations foncières, en utilisant le modèle (3.8), pour chacun des 75 mois. Certains de ces résultats sont présentés au tableau 4.1; les résultats empiriques détaillés peuvent être obtenus sur demande auprès des auteurs. Évidemment, les coefficients $\hat{\beta}^t$ ne sont pas nuls à de très faibles seuils de signification. Dans la plupart des cas, les ordonnées à l'origine $\hat{\alpha}^t$ diffèrent de manière significative de zéro au seuil de

signification de 5 %. Environ 80 % à 90 % de la variation des prix de vente est « expliquée » par la variation des évaluations, comme l'indique les valeurs de R^2. Autrement dit, le coefficient de corrélation entre les prix de vente et les évaluations foncières à la période de référence varie de 0,89 à 0,95. La figure 4.1 montre que R^2 diminue légèrement au cours du temps. Comme nous l'avons mentionné plus haut, cela pourrait tenir au fait que les prix varient différemment dans différents créneaux du marché. Nous avons été un peu surpris de constater que la valeur de R^2 n'était pas la plus élevé en janvier 2003, qui est la période de référence.

Sur la base des résultats de régression susmentionnés, nous avons calculé les indices de prix GREG en utilisant l'équation (3.10). De janvier 2003 jusqu'au milieu 2008, les prix des logements ont augmenté de quelque 25 % aux Pays-Bas, puis ont commencé à baisser, probablement en raison de la crise financière et économique. Fait important, l'indice GREG est beaucoup plus lisse que le simple ratio des moyennes d'échantillon, comme en témoigne la figure 4.2, ce qui est précisément ce pourquoi l'indice a été conçu.

Tableau 4.1
Résultats de régression

Mois	Alpha	t	Bêta	t	R carré
Janvier 2003	1 900,49	2,26	0,98	275,19	0,87
Janvier 2004	5 039,16	5,96	1,01	269,26	0,88
Janvier 2005	-2 555,12	2,43	1,08	237,54	0,84
Janvier 2006	1 282,14	1,41	1,11	286,39	0,87
Janvier 2007	-7 567,99	6,36	1,19	243,72	0,83
Janvier 2008	11 007,39	8,48	1,26	231,93	0,83
Janvier 2009	16 677,31	9,83	1,30	184,24	0,81

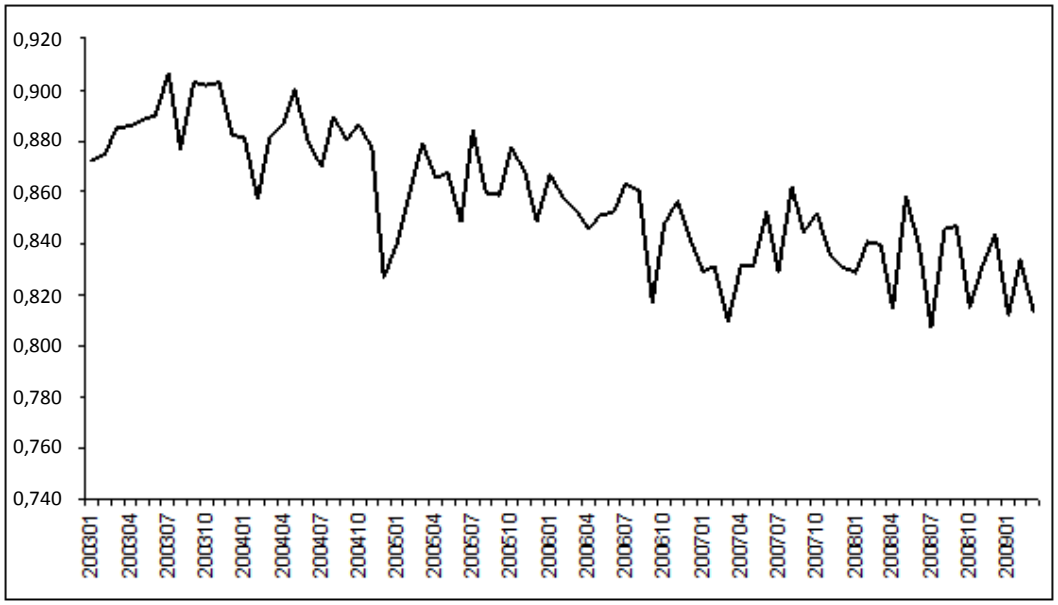


Figure 4.1 Valeur de R carré

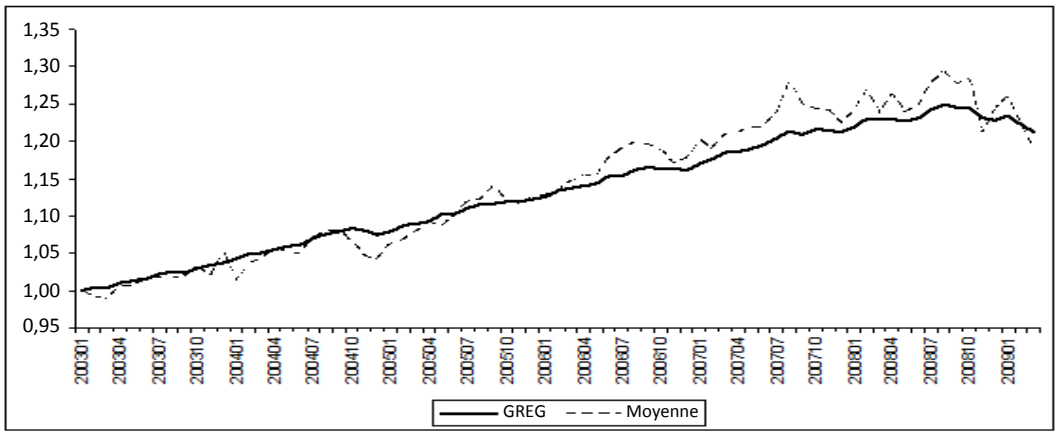


Figure 4.2 Indice GREG et ratio des moyennes d'échantillon

À la figure 4.3, l'indice GREG est comparé à l'indice SPAR. En général, la tendance des deux indices est très semblable, quoi qu'il semble exister une petite différence à la fin de la période. La figure 4.4 montre que les variations d'un mois à l'autre des indices GREG et SPAR ne diffèrent pas beaucoup non plus, l'indice GREG étant juste un peu moins volatil. Donc, nous pouvons conclure qu'à l'échelle nationale, les deux méthodes produisent des résultats plus ou moins équivalents. Notons que dans les figures 4.3 et 4.4, l'indice SPAR n'est pas l'indice SPAR officiel publié par Statistics Netherlands. Nous avons calculé un indice à base fixe en utilisant les évaluations foncières pour janvier 2003 uniquement, alors que l'indice officiel est un indice-chaîne, basé sur les évaluations pour diverses périodes de référence; voir aussi la section 5.3.

Ensuite, nous avons stratifié les données en fonction des 13 provinces et de 5 types de logements, avant d'exécuter les régressions par les MCO pour chaque mois pour les 65 strates résultantes et de calculer les indices GREG ainsi que les ratios des moyennes d'échantillon. La figure 4.5 donne les résultats pour une strate, celle des appartements dans la province de Frise. Étant donné le nombre relativement faible d'observations, on observe quelques pics importants, par exemple en septembre 2009, quand le ratio des moyennes d'échantillon a augmenté de 50 %. De nouveau, l'indice GREG est plus lisse que le ratio des moyennes d'échantillon (mais néanmoins très volatil) et étonnamment similaire à l'indice SPAR. Le même tableau se dégage pour les autres strates, de sorte que nous ne présentons pas ces résultats.

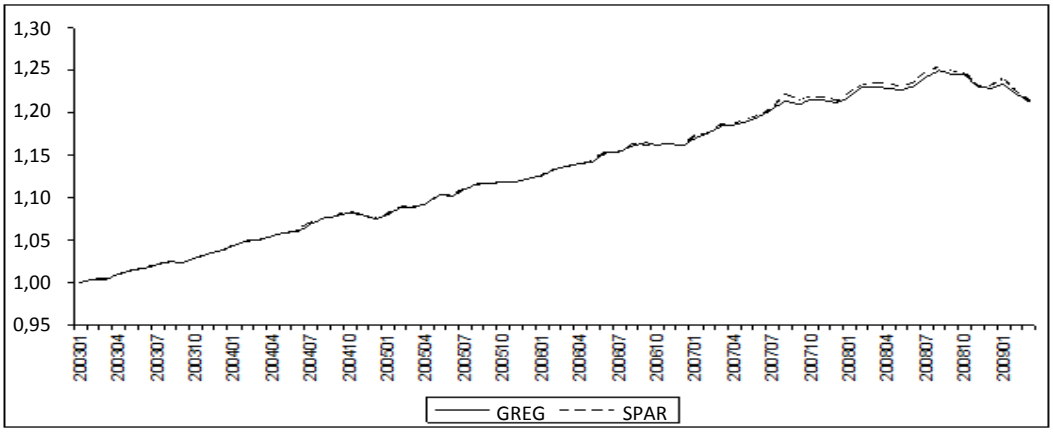


Figure 4.3 Indices GREG et SPAR

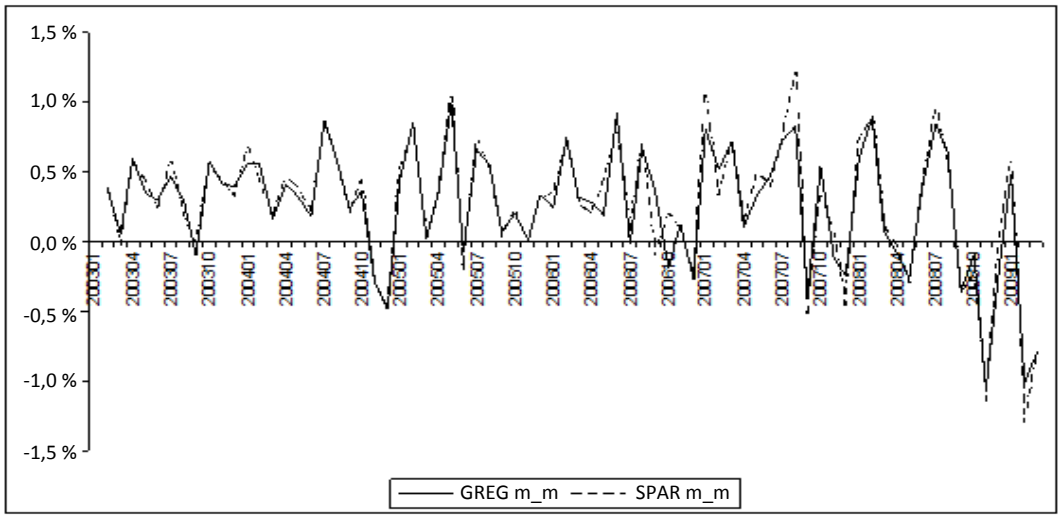


Figure 4.4 GREG et SPAR : variations en pourcentage d'un mois à l'autre

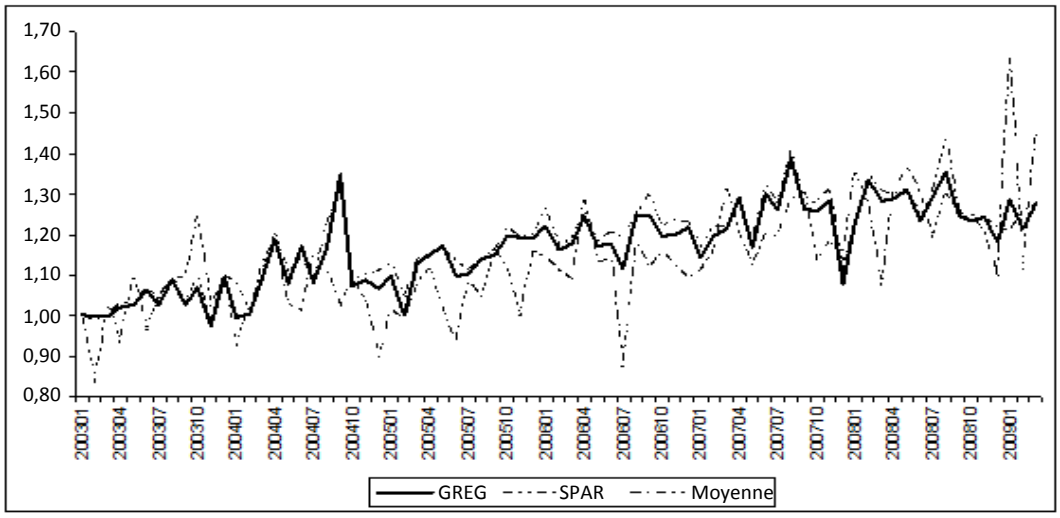


Figure 4.5 Indices GREG et SPAR et ratio des moyennes d'échantillon, appartements dans la province de Frise

Enfin, en utilisant les résultats par strate, nous avons calculé des indices GREG stratifié pour l'ensemble du pays en appliquant l'équation (3.13), où les parts de la valeur d'évaluation à la période de référence servent de pondérations pour la valeur du parc de logements. Comme le montre la figure 4.6, il n'y a pratiquement aucune différence entre les indices GREG stratifié et non stratifié, ce qui donne à penser que le biais de sélection dans l'échantillon n'est pas un problème important. La figure 4.6 montre aussi un deuxième indice de prix GREG de rechange, calculé selon l'équation (3.15), qui est fondé sur des

régressions par les MCO du modèle avec les variables indicatrices temporelles (3.14). De nouveau, les différences par rapport à l'indice GREG original paraissent faibles.

Il convient de mentionner que, même à l'intérieur des strates, certains logements étaient plus susceptibles d'être vendus que d'autres, en particulier durant la crise d'après 2008, de sorte qu'un certain biais de sélection dans l'échantillon persiste dans les indices GREG et SPAR. La direction et la grandeur de ce biais ne peuvent être prédites que si l'on dispose de données sur les caractéristiques des biens immobiliers pour estimer la probabilité que les logements soient vendus. En outre, comme nous l'avons mentionné plus haut, une stratification trop détaillée augmente à la fois la variance d'échantillonnage et le biais d'échantillon si le nombre de logements vendus est extrêmement faible, et peut accroître plutôt que réduire l'erreur quadratique moyenne des estimateurs.

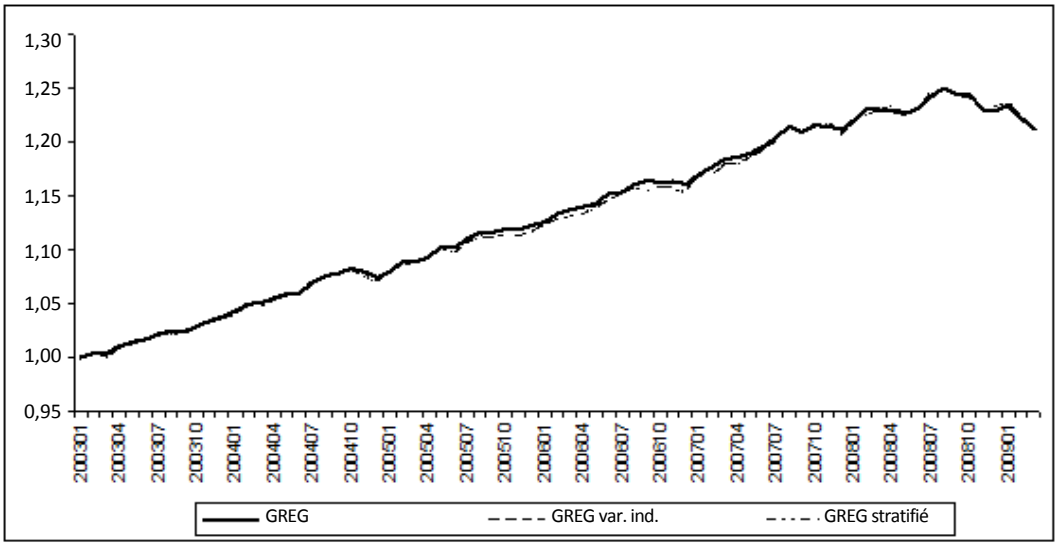


Figure 4.6 Indices GREG, GREG stratifié et GREG avec variables indicatrices

5 Discussion

5.1 Comparaisons de GREG et SPAR

La question la plus intéressante que suscite la section 4 est celle de savoir pourquoi les indices GREG et SPAR sont si semblables malgré leurs méthodes de construction très différentes. Il n'est pas étonnant que les tendances soient similaires : même si l'indice GREG ne s'appuie pas sur la méthode d'appariement de modèles, sa cible est la même que celle de l'indice SPAR. Si les tailles d'échantillon n^0 et n^t s'approchaient de la taille de la population N^0 – ce qui naturellement n'arrive jamais dans la réalité – les deux indices des prix s'approcheraient du changement de valeur du parc fixe de logements. Autrement dit, les deux méthodes sont asymptotiquement sans biais ou « convergentes ».

Ce qui peut paraître surprenant est que le degré de volatilité l'indice GREG au cours du temps est à peu près le même que celui de l'indice SPAR. Pour en comprendre la raison, rappelons que, sous les MCO, la somme des résidus de régression est nulle à chaque période. Cela implique que $\sum_{n\in S^0} p_n^0/n^0 = \sum_{n\in S^0} \hat{p}_n^0/n^0$ et $\sum_{n\in S^t} p_n^t/n^t = \sum_{n\in S^t} \hat{p}_n^t/n^t$. Pour les modèles de régression élémentaires (3.1) et (3.5), l'indice SPAR peut donc s'écrire aussi sous la forme

$$\hat{P}_{\text{SPAR}}^{0t} = \frac{\sum_{n\in S^t} \hat{p}_n^t/n^t}{\sum_{n\in S^0} \hat{p}_n^0/n^0}\left[\frac{\sum_{n\in S^0} a_n^0/n^0}{\sum_{n\in S^t} a_n^0/n^t}\right] = \frac{\left(\hat{\alpha}^t + \hat{\beta}^t\bar{a}^{0(t)}\right)/\bar{a}^{0(t)}}{\left(\hat{\alpha}^0 + \hat{\beta}^0\bar{a}^{0(0)}\right)/\bar{a}^{0(0)}} = \frac{\hat{\alpha}^t/\bar{a}^{0(t)} + \hat{\beta}^t}{\hat{\alpha}^0/\bar{a}^{0(0)} + \hat{\beta}^0}, \tag{5.1}$$

en utilisant (3.2) et (3.6) pour $n \in S^0$ et $n \in S^t$, respectivement, où $\bar{a}^{0(0)} = \sum_{n\in S^0} a_n^0/n^0$ et $\bar{a}^{0(t)} = \sum_{n\in S^t} a_n^0/n^t$ pour être bref. Il existe une similarité frappante entre la dernière expression des deuxièmes membres de (5.1) et (3.10). La seule différence est que, dans l'indice SPAR (5.1), les coefficients $\hat{\alpha}^0$ et $\hat{\alpha}^t$ sont divisés par les moyennes d'échantillon des évaluations foncières, $\bar{a}^{0(0)}$ et $\bar{a}^{0(t)}$, tandis que dans l'indice GREG (3.10), ils sont tous les deux divisés par la moyenne de population non stochastique, fixe, $\bar{a}^0$. Essentiellement, l'indice SPAR est un estimateur entièrement fondé sur échantillon de l'indice GREG.

Comparativement à la méthode SPAR, l'approche GREG élimine une source d'erreur d'échantillonnage, c'est-à-dire la variabilité d'échantillonnage des évaluations moyennes. Conformément à la théorie de la régression généralisée, nous nous attendrions intuitivement à ce que la méthode GREG réduise l'erreur d'échantillonnage de l'indice des prix et produise une série chronologique moins volatile (sous l'hypothèse raisonnable que $\bar{a}^{0(t)}$ et $\hat{\alpha}^t$ ne sont pas corrélées entre les périodes $t = 0,\ldots,T$). En d'autres mots, alors que la méthode GREG a été conçue comme une amélioration du ratio des moyennes d'échantillon, nous aurions pu nous attendre également à ce qu'elle joue le rôle de procédure de lissage de l'indice SPAR. Toutefois, comme nous l'avons montré à la section 4, en pratique, cela n'est guère le cas. Ce résultat peut s'expliquer comme il suit.

La réduction de la variance de l'indice GREG comparativement à l'indice SPAR dépend de la valeur des termes d'ordonnée à l'origine des régressions aux périodes 0 et t. Si les droites de régression passaient exactement par l'origine $(\hat{\alpha}^t = \hat{\alpha}^0 = 0)$, les indices GREG et SPAR seraient tous deux égaux au ratio des coefficients de pente $\hat{\beta}^t/\hat{\beta}^0$ et aucune réduction de la variance n'aurait lieu. Dans le cas moins extrême où $\hat{\alpha}^t$ et $\hat{\alpha}^0$ sont proches de 0 et où les ratios $\hat{\alpha}^t/\bar{a}^0$, $\hat{\alpha}^t/\bar{a}^{0(t)}$, $\hat{\alpha}^0/\bar{a}^0$ et $\hat{\alpha}^0/\bar{a}^{0(0)}$ dans (3.10) et (5.2) sont très faibles comparativement à $\hat{\beta}^t$ et $\hat{\beta}^0$, les indices GREG et SPAR ne différeront que légèrement, et la réduction de la variance sera marginale; voir aussi l'annexe.

Cette dernière situation est ce que l'on constate effectivement en pratique, comme le montrent les figures 5.1 et 5.2, où les valeurs de $\hat{\alpha}^t/\bar{a}^0$ et $\hat{\alpha}^t/\bar{a}^{0(t)}$ et celles de $\hat{\beta}^t$ sont représentées en fonction du temps. Les ratios $\hat{\alpha}^t/\bar{a}^0$ et $\hat{\alpha}^t/\bar{a}^{0(t)}$ sont remarquablement similaires et petits comparativement aux $\hat{\beta}^t$. Bien que nous ne puissions pas ignorer ces ratios, c'est la variation de $\hat{\beta}^t$ qui dicte principalement les indices GREG et SPAR. L'indice SPAR est non seulement un estimateur entièrement fondé sur échantillon de l'indice GREG, comme nous l'avons mentionné plus haut, mais il semble être presque aussi efficace.

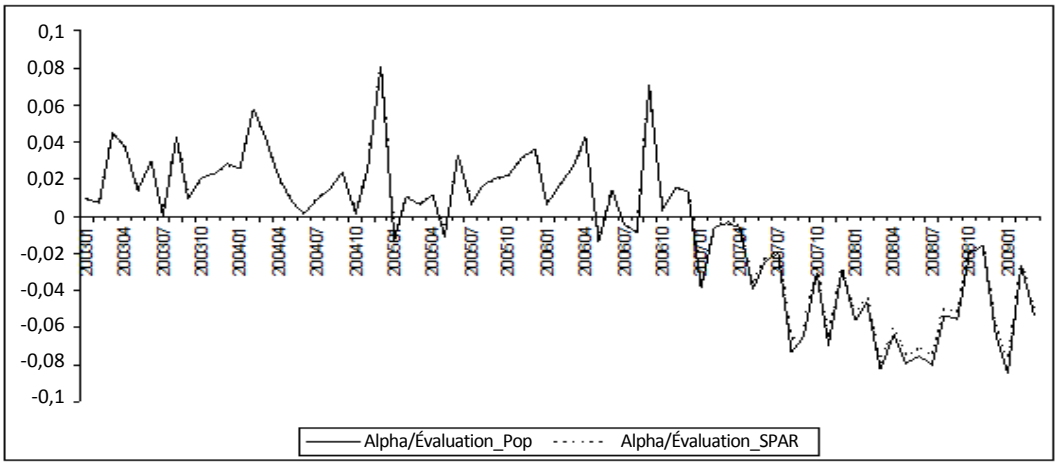


Figure 5.1 Ordonnées à l'origine divisées par les moyennes des évaluations

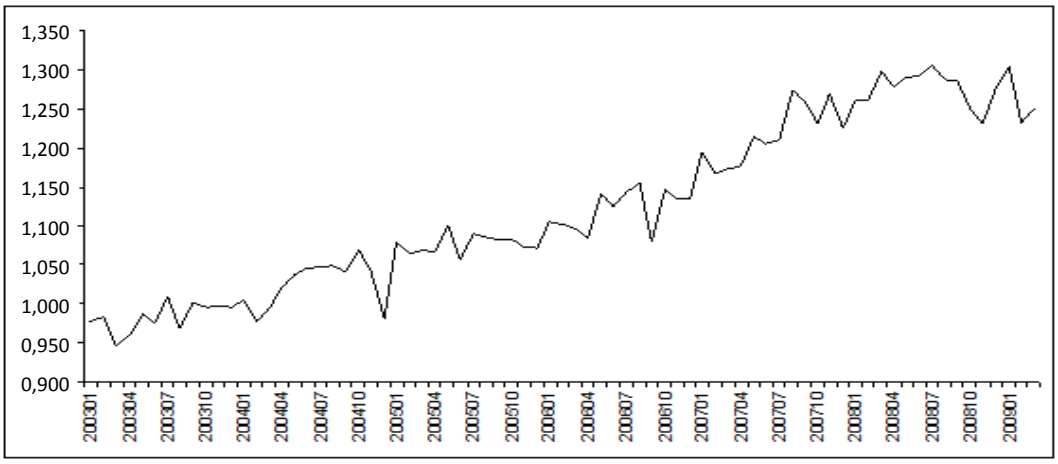


Figure 5.2 Coefficients de pente

5.2 Volatilité du coefficient de pente

Plusieurs facteurs peuvent avoir contribué à la volatilité des coefficients de pente $\hat{\beta}^t$ dans nos régressions des prix de vente sur les évaluations foncières, et donc sur les indices de GREG et SPAR. Nous allons discuter brièvement de trois de ces facteurs, à savoir le changement de composition de l'échantillon, l'hétéroscédasticité et les valeurs aberrantes.

Un échantillon de logements peut être considéré comme un échantillon de localisations, ou adresses, puisque les logements sont attachés au terrain sur lequel ils sont construits. Un changement de composition de l'échantillon n'est rien d'autre qu'un changement dans les localisations au niveau le plus bas. Un *changement de composition des localisations* influe sur la composition de l'échantillon en ce qui concerne les caractéristiques de qualité moyennes des biens, telles que le nombre de pièces, la superficie, *etc*. Dans notre cadre simple, où nous observons une seule caractéristique (non physique), à savoir la valeur d'évaluation, un changement de composition des localisations se résume à un changement de la distribution d'échantillon des évaluations. Cela, conjugué à toute variation des changements de prix selon

le créneau du marché, induit un changement dans la distribution d'échantillons des ratios p_n^t / a_n^0, qui à son tour entraîne un changement de $\hat{\beta}^t$ dans le modèle de régression à deux variables (3.5).

Hormis la stratification, nous ne pouvons pas faire grand-chose quant à l'effet des changements de composition des localisations dans l'échantillon (mais la stratification par province et par type de logement n'a pas été très utile), de sorte qu'il est difficile de réduire la volatilité de $\hat{\beta}^t$ et, par conséquent, des indices GREG et SPAR. Il est également impossible d'introduire une variable de contrôle pour la localisation au niveau de l'adresse dans les méthodes d'imputation hédoniques. Dans ces dernières, l'effet du changement de composition (des localisations) est atténué par l'ajout de variables de contrôle pour la région ainsi qu'une gamme de caractéristiques physiques. Cependant, cela ne signifie pas nécessairement que l'imputation hédonique produira une série d'indices plus stable que les méthodes GREG ou SPAR. La plupart des modèles hédoniques classiques sont moins bien ajustés aux données transversales que notre modèle, et les coefficients des caractéristiques présentent habituellement une forte variabilité au cours du temps. Donc, il n'est peut-être pas étonnant que Bourassa, Hoesli et Sun (2006) constatent que [traduction] « l'indice SPAR […] suit fiablement les variations de prix des logements mais est moins volatil que les indices produits par des méthodes qui requièrent plus d'estimations de paramètres. »

Nous pouvons aussi examiner la variabilité du coefficient de pente d'un point de vue purement statistique. Il est bien connu que, dans un modèle à deux variables, l'estimateur par les MCO $\hat{\beta}^t$ peut s'écrire sous la forme

$$\hat{\beta}^t = r\left(p^t, a^0\right)\frac{s\left(p^t\right)}{s\left(a^0\right)}, \tag{5.2}$$

où $r\left(p^t, a^0\right)$ désigne le coefficient de corrélation dans l'échantillon à la période t entre les prix de vente et les évaluations foncières, qui est égal à la racine carrée de R^2; $s\left(p^t\right)$ et $s\left(a^0\right)$ sont les écarts-types d'échantillon correspondants. Une comparaison des figures 4.1 et 5.2 laisse entendre que des variations subites de R^2 sont en grande partie responsables de la volatilité de $\hat{\beta}^t$. Ainsi, en décembre 2004, une diminution importante de R^2 coïncide avec une diminution importante de $\hat{\beta}^t$ (et avec une diminution des indices GREG et SPAR, comme le montre la figure 4.4).

La régression par les moindres carrés peut être pondérée ou non pondérée. En l'absence d'*hétéroscédasticité*, c'est-à-dire quand la variance des erreurs est constante, il faut utiliser les MCO. En présence d'hétéroscédasticité, la préférence va aux moindres carrés pondérés (MCP); si l'on utilise les poids appropriés, les MCP donnent des coefficients plus stables que les MCO. Dans ce cas, la somme des résidus dans l'échantillon pondéré diffère de zéro, de sorte que l'estimateur (3.9), doit être utilisé. Pour faciliter l'interprétation de l'indice GREG et la comparaison avec l'indice SPAR, à la section 3, nous avons supposé qu'il n'y avait pas de problème d'hétéroscédasticité et nous nous sommes limités aux MCO. Donc, l'estimateur GREG (MCO) donné par (3.10) demeure asymptotiquement sans biais sous le plan en présence d'hétéroscédasticité.

La forme la plus intéressante d'hétéroscédasticité (classique) – et, étant donné notre jeu de données, la seule forme que nous serions capables de réduire – se présenterait si la variance des erreurs de notre modèle de régression (3.5) dépendait de la valeur d'évaluation, celle-ci étant la seule variable explicative. Cependant, les résidus de nos régressions par les MCO n'indiquent par la présence d'une

hétéroscédasticité de ce type importante. Cela est illustré à la figure 5.3, qui représente les prix de vente en fonction des évaluations, pour trois mois y compris la période de référence (janvier 2003); les droites de régression sont également données. En guise de confirmation, nous avons également effectué le test de White (1980), qui n'a pas indiqué cette forme d'hétéroscédasticité.

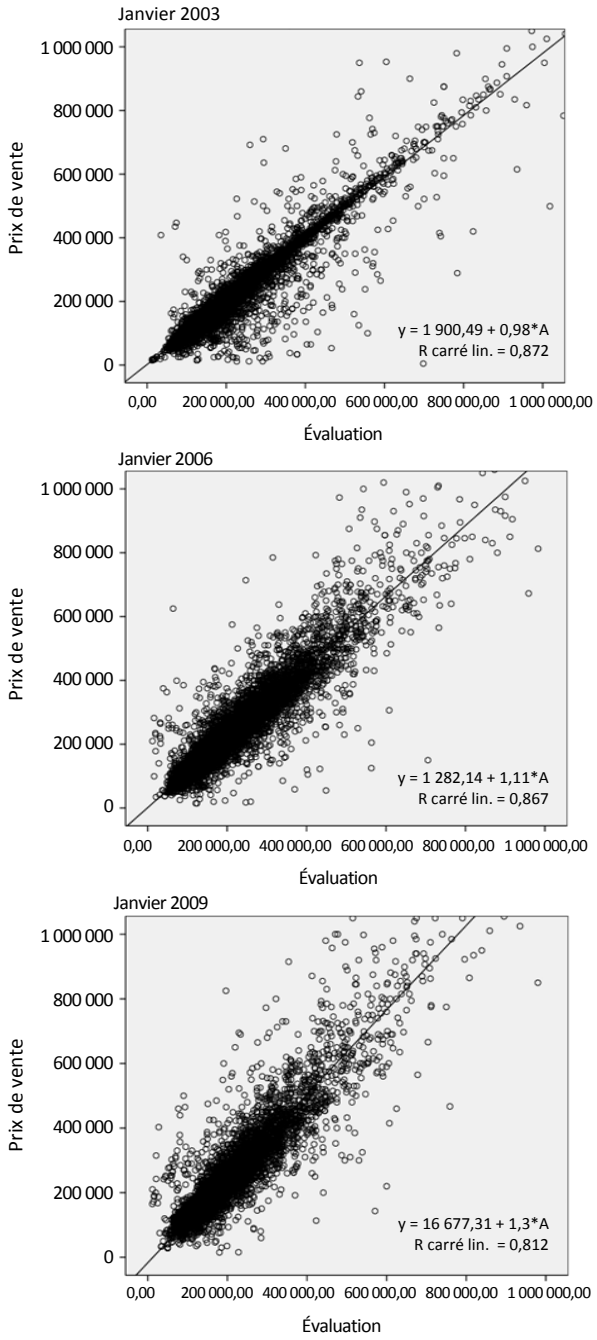


Figure 5.3 Nuages de points et droites de régression

Notre jeu de données initiales de prix de vente et d'évaluations foncières comprenait certaines *valeurs aberrantes* évidentes. Pour estimer l'indice GREG, nous avons par conséquent utilisé un jeu de données nettoyées qui a été préparé pour calculer l'indice officiel des prix des logements aux Pays-Bas. Statistics Netherlands applique plusieurs procédures de nettoyage des données. Les logements qui ont été vendus plus d'une fois durant un mois donné sont exclus du jeu de données. Pour éliminer les erreurs de saisies et les valeurs aberrantes qui pourraient influencer excessivement les résultats, les biens dont le prix de vente ou l'évaluation foncière est inférieur à 10 000 € ou supérieur à 5 000 000 € et ceux dont le ratio prix de vente-évaluation est « irréaliste » sont également supprimés. La suppression des observations « irréalistes » est faite en examinant la distribution du logarithme des ratios prix de vente-évaluation; sont supprimées toutes les observations pour lesquelles l'écart du logarithme du ratio par rapport à la moyenne est de plus de 5 écarts-types. Pour plus de renseignements, voir Statistics Netherlands (2008).

Ces procédures sont assez arbitraires. Pour les estimateurs par la régression, tels que l'estimateur GREG, il est plus approprié de supprimer les observations dont l'effet de levier est important, c'est-à-dire d'éliminer de l'échantillon les unités dont l'exclusion a un effet important sur les coefficients de régression. Une mesure bien connue dans ce contexte est le DFBETA d'une unité de l'échantillon (Cook et Weisberg 1982). Puisque l'indice SPAR peut s'écrire sous forme d'un indice fondé sur la régression, cette mesure pourrait également être utilisée pour déceler et supprimer les valeurs aberrantes. Les nuages de points de la figure 5.3 montrent que le jeu de données nettoyé contient encore certaines valeurs aberrantes importantes. Il reste à déterminer si ces valeurs ont un effet de levier important et si leur élimination réduira la volatilité des $\hat{\beta}^t$ dans les indices GREG et SPAR.

5.3 Certaines autres remarques

La méthode GREG part du principe que le parc de logements est fixe. Autrement dit, nous avons supposé qu'il ne se produit pas d'entrées (par exemple logements nouvellement construits) ni de sorties (logements mis aux rebuts) et que la qualité des logements demeure constante au cours du temps. Notre approche n'est pas symétrique en ce sens que nous nous conditionnons sur le parc de logements à la *période de référence*. Dans la perspective d'un indice, nous estimons un indice des prix de Laspeyres pour le parc de logements où les quantités sont toutes égales à 1 parce que chaque logement est traité comme un bien unique. Une approche tout aussi justifiable consisterait à mesurer la variation du parc de logements à la période courante, qui comprend les ajouts au parc durant chaque période, en utilisant un indice de Paasche. En calculant la moyenne géométrique des deux indices, on obtiendrait l'indice de Fisher. Ce dernier est une mesure privilégiée de la variation des prix en raison de sa forme symétrique. La construction d'un indice GREG de type Fisher est toutefois impossible, puisque la composante de Paasche requiert des valeurs d'évaluation en temps réel pour les logements neufs dans le parc, alors qu'elles ne sont manifestement pas disponibles.

L'hypothèse d'un parc de logements fixe (à la période de référence) peut être relâchée par enchaînement annuel, à condition que le parc de logements soit réévalué annuellement. Il s'agit de la situation actuelle aux Pays-Bas; dans le passé, les évaluations foncières étaient effectuées tous les trois ou quatre ans. Une mise à jour annuelle des évaluations pourrait également comprendre une correction pour les changements de qualité des biens, du moins dans une certaine mesure, parce que les évaluations mises

à jour tiennent vraisemblablement compte des réparations importantes, des rénovations et de la dépréciation.

Une remarque finale s'impose. À certaines fins, il est souhaitable de décomposer l'indice des prix des logements global en deux composantes : l'une qui mesure la variation de prix du bâtiment, et l'autre, la variation de prix du terrain. Ni notre méthode GREG ni les méthodes SPAR et des ventes répétées ne conviennent pour cela. Les méthodes d'imputation hédoniques pourraient convenir, malgré des problèmes pratiques tels que la multicolinéarité; voir Diewert, de Haan et Hendriks (2012) pour une première tentative. Si les données sur la taille du bâtiment, la taille du terrain et d'autres attributs déterminant le prix devenaient disponibles pour tous les biens inclus dans le parc de logements, nous serions capables d'estimer un « indice GREG avec imputation hédonique », comprenant la décomposition terrain-bâtiment. Les chances d'obtenir ce genre de données aux Pays-Bas sont malheureusement minces.

6 Conclusion

La simple méthode GREG décrite dans le présent article, qui est fondée sur la régression des prix de vente sur les évaluations foncières par les MCO, réduit considérablement la volatilité d'un indice des prix des logements comparativement aux ratios des moyennes d'échantillon. L'indice SPAR peut être considéré comme un estimateur de l'indice GREG (MCO) (lui-même un estimateur, évidemment) dans lequel la moyenne de population des évaluations à la période de référence est remplacée par les moyennes d'échantillon à la période de référence et à la période comparée. Nos résultats empiriques pour les Pays-Bas indiquent que l'indice SPAR est presque aussi efficace que l'indice GREG, même pour de petites sous-populations. Nous avons vérifié cela en tirant un échantillon aléatoire de 50 observations chaque mois du nombre total de ventes mensuelles (15 000 en moyenne). Les variations d'un mois à l'autre de l'indice SPAR sont à peine plus importantes que celles de l'indice GREG.

En raison du changement de composition de l'ensemble de logements vendus, la série chronologique GREG (et SPAR) présente une forte volatilité à court terme. Une augmentation durant un mois particulier est habituellement suivie d'une diminution le mois suivant. Autrement dit, les variations d'un mois à l'autre ne disent pas grand-chose au sujet de la variation réelle du prix du parc de logements qui, sauf dans des circonstances inhabituelles, devrait être harmonieuse. Une méthode améliorée de détection des valeurs aberrantes aiderait peut-être à réduire la volatilité de l'indice, mais l'effet serait vraisemblablement limité. L'application d'une procédure de lissage semble être une option. Cependant, ce lissage entraîne habituellement des révisions des indices de prix publiés antérieurement et l'absence de révision est l'un des points forts des approches GREG et SPAR. Une autre option consisterait à réduire la fréquence d'observation, en la choisissant par exemple trimestrielle, mais cela pourrait être indésirable également.

D'un point de vue purement statistique, dans notre modèles à deux variables, la variabilité de R^2 semble être en grande partie à l'origine de la volatilité du coefficient de pente et, par conséquent, de celle de la série d'indices de prix. De futures études pourraient porter sur la relation entre les changements de composition ayant trait aux caractéristiques des biens immobiliers et les variations de R^2. Comme les données sur de nombreuses caractéristiques des logements ne sont pas disponibles, nous ne pouvons pas étudier cette question au moyen de nos données. Heureusement, Statistics Netherlands a accès à un jeu de données produit par la plus grande association d'agents immobiliers aux Pays-Bas qui pourrait être utile. Ce jeu de données couvre environ 70 % des ventes de logements qui ont eu lieu aux Pays-Bas de 1999 à

2008, comprend des données sur de nombreuses caractéristiques des biens et a été enrichi par les données d'évaluation foncière. Dans le passé, nous avons utilisé ce jeu de données pour comparer l'indice SPAR à divers types d'indices hédoniques.

Remerciements

Les auteurs remercient les participants à l'Economic Measurement Group Workshop, qui s'est déroulé du 1er au 3 décembre 2010 à l'Université de New South Wales, à Sydney, en Australie, et les participants à un séminaire d'économie appliquée, tenu le 22 novembre 2011 à l'Université du Queensland, à Brisbane, en Australie, de leurs commentaires constructifs concernant des versions préliminaires de l'article. Les commentaires et suggestions faits par le rédacteur et deux examinateurs anonymes ont également contribué à améliorer l'article. Les auteurs remercient de son aide Erna van der Wal, qui leur a fourni les données. Les opinions exprimées dans l'article sont celles des auteurs et ne représentent pas forcément celles de Statistics Netherlands.

Annexe

Erreurs-types approximatives de l'indice GREG

L'indice GREG défini par l'équation (3.10) dans le corps du texte est un ratio de deux estimateurs, $\hat{\bar{p}}^{t}_{\text{GREG}}$ et $\hat{\bar{p}}^{0}_{\text{GREG}}$; pour simplifier, nous supprimons la notation « MCO ». En utilisant un développement du premier degré en séries de Taylor, la variance de l'indice peut être approximée par (voir, par exemple, Kendall et Stuart 1976)

$$\text{var}\left(\hat{P}^{0t}_{\text{GREG}}\right) \cong \left[\frac{E\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)}{E\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)}\right]^2 \left[\frac{\text{var}\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)}{\left\{E\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)\right\}^2} + \frac{\text{var}\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)}{\left\{E\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)\right\}^2} + \frac{\text{cov}\left(\hat{\bar{p}}^{t}_{\text{GREG}}, \hat{\bar{p}}^{0}_{\text{GREG}}\right)}{E\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)E\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)}\right], \qquad \text{(A.1)}$$

où $E\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)$ et $E\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)$ désignent les valeurs espérées.

Le terme de covariance dans (A.1) est égal à 0 puisque, par hypothèse, les échantillons aux périodes 0 et t sont tirés indépendamment. Le remplacement des valeurs espérées dans (A.1) par les estimateurs et le calcul subséquent de la racine carré donne l'expression qui suit pour l'erreur-type de $\hat{P}^{0t}_{\text{GREG}}$:

$$se\left(\hat{P}^{0t}_{\text{GREG}}\right) \cong \hat{P}^{0t}_{\text{GREG}}\left[\frac{\text{var}\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)}{\left(\hat{\bar{p}}^{t}_{\text{GREG}}\right)^2} + \frac{\text{var}\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)}{\left(\hat{\bar{p}}^{0}_{\text{GREG}}\right)^2}\right]^{1/2}. \qquad \text{(A.2)}$$

L'équation (A.2) peut être estimée en pratique en utilisant $\hat{\bar{p}}^{s}_{\text{GREG}} = \hat{\alpha}^{s} + \hat{\beta}^{s}\bar{a}^{0}\ (s = 0, t)$, d'où $\text{var}\left(\hat{\bar{p}}^{s}_{\text{GREG}}\right) = \text{var}\left(\hat{\alpha}^{s}\right) + \left(\bar{a}^{0}\right)^2 \text{var}\left(\hat{\beta}^{s}\right) + 2\bar{a}^{0}\,\text{cov}\left(\hat{\alpha}^{s}, \hat{\beta}^{s}\right)$. Les estimations des (co)variances sont obtenues facilement dans la plupart des progiciels statistiques à partir de la matrice de variance-covariance.

La division de (A.2) par $\hat{P}^{0t}_{\text{GREG}}$ donne une expression pour l'erreur-type relative ou coefficient de variation, $CV\left(\hat{P}^{0t}_{\text{GREG}}\right) = se\left(\hat{P}^{0t}_{\text{GREG}}\right)\Big/\hat{P}^{0t}_{\text{GREG}}$, de l'indice GREG :

$$CV\left(\hat{P}^{0t}_{\text{GREG}}\right) \cong \left[\frac{\text{var}\left(\hat{\bar{P}}^{t}_{\text{GREG}}\right)}{\left(\hat{\bar{P}}^{t}_{\text{GREG}}\right)^2} + \frac{\text{var}\left(\hat{\bar{P}}^{0}_{\text{GREG}}\right)}{\left(\hat{\bar{P}}^{0}_{\text{GREG}}\right)^2}\right]^{1/2} = \left[\left\{CV\left(\hat{\bar{P}}^{t}_{\text{GREG}}\right)\right\}^2 + \left\{CV\left(\hat{\bar{P}}^{0}_{\text{GREG}}\right)\right\}^2\right]^{1/2}. \quad \text{(A.3)}$$

Un élément plus important est l'erreur-type relative de la *variation en pourcentage* de l'indice, c'est-à-dire $CV\left(\hat{P}^{0t}_{\text{GREG}} - 1\right) = se\left(\hat{P}^{0t}_{\text{GREG}} - 1\right)\Big/\left(\hat{P}^{0t}_{\text{GREG}} - 1\right)$. Celle-ci est généralement plus grande que $CV\left(\hat{P}^{0t}_{\text{GREG}}\right)$, étant donné que $se\left(\hat{P}^{0t}_{\text{GREG}} - 1\right) = se\left(\hat{P}^{0t}_{\text{GREG}}\right)$ et $\hat{P}^{0t}_{\text{GREG}} - 1 < \hat{P}^{0t}_{\text{GREG}}$.

Si les deux droites de régression passent presque par l'origine, donc que $\hat{\alpha}^s \cong 0\left(s = 0, t\right)$, nous avons $\hat{P}^{0t}_{\text{GREG}} \cong \hat{\beta}^t\big/\hat{\beta}^0$ et (A.2) se simplifie pour donner

$$se\left(\hat{P}^{0t}_{\text{GREG}}\right) = se\left(\hat{P}^{0t}_{\text{GREG}} - 1\right) \cong \hat{P}^{0t}_{\text{GREG}}\left[\frac{\text{var}\left(\hat{\beta}^t\right)}{\left(\hat{\beta}^t\right)^2} + \frac{\text{var}\left(\hat{\beta}^0\right)}{\left(\hat{\beta}^0\right)^2}\right]^{1/2}. \quad \text{(A.4)}$$

Dans ce cas particulier, les indices GREG et SPAR coïncident presque, de sorte que l'expression (A.4) est également vérifiée pour l'indice SPAR (en utilisant $\hat{P}^{0t}_{\text{SPAR}}$ au lieu de $\hat{P}^{0t}_{\text{GREG}}$).

Bibliographie

Bailey, M.J., Muth, R.F. et Nourse, H.O. (1963). A regression method for real estate price construction. *Journal of the American Statistical Association*, 58, 933-942.

Beaumont, J.-F., et Alavi, A. (2004). Estimation robuste par la régression généralisée. *Techniques d'enquête*, 30, 2, 217-231.

Bourassa, S.C., Hoesli, M. et Sun, J. (2006). A simple alternative house price index method. *Journal of Housing Economics*, 15, 80-97.

Calhoun, C.A. (1996). OFHEO House Price Indexes: HPI Technical Description. Office of Federal Housing Enterprise Oversight, Washington, DC.

Case, K.E., et Shiller, R.J. (1987). Prices of single-family homes since 1970: New indexes for four cities. *New England Economic Review*, Septembre-Octobre, 45-56.

Case, K.E., et Shiller, R.J. (1989). The efficiency of the market for single family homes. *The American Economic Review*, 79, 125-137.

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

Cook, R.D., et Weisberg, S. (1982). *Residuals and Influence in Regression*, New York : Chapman and Hall.

de Haan, J. (2007). Formulae for the Variance of (Changes in) the SPAR Index. Manuscrit non publié, Statistics Netherlands, Voorburg (Pays-Bas seulement, disponible auprès de l'auteur sur demande).

de Haan, J. (2010). Hedonic price indexes: A comparison of imputation, time dummy and 'Re-Pricing' methods, *Journal of Economics and Statistics* (Jahrbucher fur Nationalokonomie und Statistik), 230, 772-791.

de Haan, J., van der Wal, E. et de Vries, P. (2009). The measurement of house prices: A review of the sale price appraisal method. *Journal of Economic and Social Measurement*, 34, 51-86.

de Vries, P., de Haan, J., van der Wal, E. et Mariën,G. (2009). A house price index based on the SPAR method. *Journal of Housing Economics*, 18, 214-223.

Diewert, W.E., de Haan, J. et Hendriks, R. (2012). The decomposition of a house price index into land and structures components: A hedonic regression approach. *Econometric Reviews* (à venir).

Diewert, W.E., Heravi, S. et Silver, M. (2009). Hedonic imputation versus time dummy hedonic indexes. Dans *Price Index Concepts and Measurement*, (Éds., W.E. Diewert, J. Greenlees et C. Hulten), NBER Studies in Income and Wealth, Chicago: Chicago University Press, 70, 161-196.

Edelstein, R.H., et Quan, D.C. (2006). How does appraisal smoothing bias real estate returns measurement? *Journal of Real Estate Finance and Economics*, 32, 41-60.

Eurostat (2010). *Technical Manual on Owner-Occupied Housing for Harmonised Index of Consumer Prices*, Version 1.9. Disponible au www.epp.eurostat.ec.europa.eu/portal/page/portal/hicp/documents/Tab/Tab/03_METH-OOH-TECHMANUAL_V1-9.pdf.

Francke, M.K. (2010). Repeat sales index for thin markets: A structural time series approach. *Journal of Real Estate Finance and Economics*, 41, 24-52.

Geltner, D. (1996). The repeated-measures regression-based index: A better way to construct appraisal-based indexes of commercial property value. *Real Estate Finance*, 12, 29-35.

Gouriéroux, C., et Laferrère, A. (2009). Managing hedonic house price indexes: The french experience. *Journal of Housing Economics*, 18, 206-213.

Grimes, A., et Young, C. (2010). A Simple Repeat Sales House Price Index: Comparative Properties Under Alternative Data Generation Processes. Motu Working Paper 10-10, Motu Economic and Public Policy Research, New Zealand.

Hardman, M. (2011). Calculating High Frequency Australian Residential Property Price Indices. Rismark Document technique, disponible au www.rpdata.com/images/stories/content/PDFs/technical_method_paper.pdf.

Hedlin, D., Falvey, H., Chambers, R. et Kokic, P. (2001). Does the model matter for GREG estimation? A business survey example. *Journal of Official Statistics*, 17, 527-544.

Hill, R.J., et Melser, D. (2008). Hedonic imputation and the price index problem: An application to housing. *Economic Inquiry*, 46, 593-609.

Jansen, S.J.T., de Vries, P., Coolen, H.C.C.H., Lamain, C.J.M. et Boelhouwer, P. (2008). Developing a house price index for the Netherlands: A practical application of weighted repeat sales. *Journal of Real Estate Finance and Economics*, 37, 163-186.

Kendall, M., et Stuart, A. (1976). *The Advanced Theory of Statistics – Volume 1: Distribution Theory*, 4e Édition, Londres : Charles Griffin & Company.

Leventis, A. (2006). Removing Appraisal Bias from a Repeat Transactions House Price Index: A Basic Approach. Document présenté à l'atelier OECD-IMF on Real Estate Price Indexes, Paris, 6 au 7 novembre 2006.

Makaronidis, A., et Hayes, K. (2006). Owner Occupied Housing for the HICP. Document présenté à l'atelier OECD-IMF on Real Estate Price Indexes, Paris, 6 au 7 novembre 2006.

Rossini, P., et Kershaw, P. (2006). Developing a Weekly Residential Price Index Using the Sales Price Appraisal Ratio. Document présenté à la twelfth Annual Pacific Rim Real Estate Society Conference, Auckland, 22 au 25 janvier 2006.

Saarnio, M. (2006). Housing Price Statistics at Statistics Finland. Document présenté à l'atelier OECD-IMF on Real Estate Price Indexes, Paris, 6 au 7 novembre 2006.

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

Shi, S., Young, M. et Hargreaves, B. (2009). Issues in measuring a monthly house price index in New Zealand. *Journal of Housing Economics*, 18, 336-350.

Statistics Netherlands (2008). Price Index Owner-occupied Existing Dwellings; Method Description. Statistics Netherlands, La Haye, disponible au www.cbs.nl/NR/rdonlyres/A49D8542-26EC-40FD-9093-82A519247F4B/0/MethodebeschrijvingPrijsindexBestaandeKoopwoningene.pdf.

van der Wal, E., ter Steege, D. et Kroese B. (2006). Two Ways to Construct a House Price Index for the Netherlands: The Repeat Sale and Sale Price Appraisal Ratio. Document présenté à l'atelier OECD-IMF on Real Estate Price Indexes, Paris, 6 au 7 novembre 2006.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-838.


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

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 459-476
Statistique Canada, N° 12-001-X au catalogue

La première impression a-t-elle de l'importance ? Examen de l'effet de la conception de l'écran d'accueil sur le taux de réponse

Roos Haer et Nadine Meidert[1]

Résumé

Les sondages en ligne sont généralement caractérisés par de faibles taux de réponse. Les suggestions habituelles que l'on trouve dans les manuels sur la recherche par sondage en ligne soulignent le rôle important que joue l'écran d'accueil en vue d'encourager les répondants à participer au sondage. Les travaux de recherche ont donné la preuve empirique de l'importance de cet écran, montrant que la plupart des répondants interrompent la communication à l'étape de l'écran d'accueil. Cependant, peu d'études ont eu pour sujet l'effet de la conception de cet écran sur le taux d'interruption. Dans le cadre d'une étude réalisée à l'Université de Constance, trois traitements expérimentaux ont été ajoutés à un sondage auprès de la population d'étudiants de première année (2 629 étudiants) afin d'évaluer l'effet de diverses caractéristiques de conception de l'écran sur les taux d'interruption. Les expériences méthodologiques comprenaient la variation de la couleur de fond de l'écran d'accueil, la variation de la durée promise de la tâche sur le premier écran et la variation de la longueur de l'information fournie sur l'écran d'accueil pour expliquer aux répondants leurs droits à la protection de la vie privée. Les analyses montrent que plus la durée indiquée de la tâche était longue et plus l'attention donnée à l'explication des droits à la protection de la vie privée sur l'écran d'accueil était importante, plus le nombre d'étudiants qui commençaient à répondre au sondage et achevaient de le faire était faible. Par contre, l'utilisation d'une couleur de fond différente n'a pas produit la différence significative attendue.

Mots-clés : Sondages en ligne; écrans d'accueil; interruptions; conception.

1 Introduction

Étant donné le nombre croissant d'utilisateurs d'Internet et la popularité grandissante des plus grandes largeurs de bande, les sondages en ligne en vue de recueillir des données se multiplient rapidement (Vicente et Reis 2010). Les avantages de ce mode de sondage ont été bien décrits; ils permettent d'économiser beaucoup de temps et d'argent. Cependant, parallèlement aux aspects positifs, les sondages en ligne suscitent des préoccupations méthodologiques que l'on ne peut pas ignorer si l'on veut garantir la qualité de ces sondages (Vicente et Reis 2010). Ces préoccupations ont trait principalement à la non-réponse et à la couverture. Tandis que cette dernière est une cause moins importante de souci dans les sondages auprès de personnes spécifiquement désignées, comme les étudiants, la non-réponse reste un grand sujet de préoccupation dans le domaine de la recherche par sondage en ligne (Crawford, Couper et Lamias 2001).

Les sondages en ligne sont associés à des taux de réponse relativement faibles comparativement à d'autres modes d'étude par sondage (Lozar Manfreda Bosnjak, Berzelak, Haas et Vehovar 2008). Cette situation touche tous les types de sondages en ligne, des échantillons tirés de listes aux panels probabilistes prérecrutés et aux panels à option de participation ou à participation volontaire (Couper et Miller 2008, page 833). Le taux de non-réponse est égal à la somme des répondants qui n'ont pas participé au sondage en ligne alors qu'ils avaient été invités à le faire, et des répondants qui ont interrompu la

1. Roos Haer et Nadine Meidert, Département de science politique et d'administration publique, Université de Constance, Universitätsstraße 10, 78464 Constance, Allemagne. Courriel : Roos.vanderHaer@uni-konstanz.de, Nadine.Meidert@uni-konstanz.de.

communication et ont abandonné prématurément. Autrement dit, les non-répondants sont les répondants qui ne voient pas toutes les questions et ne répondent pas à toutes les questions (Bosjnak et Tuten 2001). Il est important de noter que nous utilisons les termes « abandon » et « interruption » comme des synonymes tout au long de la présente étude. La non-réponse est un aspect particulièrement important pour les chercheurs, parce que les caractéristiques et attitudes inconnues des non-répondants peuvent causer des inexactitudes dans les résultats de l'étude (Bosnjak et Tuten 2001). Ce problème constitue un défi pour tous les modes de sondage, mais surtout pour les sondages en ligne (Galesic 2006). En général les taux d'abandon aux sondages en ligne peuvent être aussi élevés que 80 %, avec une moyenne d'environ 30 %. Dans le cas des sondages en ligne ciblant individuellement les répondants, ces taux sont plus faibles, mais demeurent néanmoins de l'ordre de 15 % (Galesic 2006, page 313; Peytchev 2009).

Le nombre le plus important d'abandons est, de loin, celui observé à la première page, c'est-à-dire l'écran d'accueil (Couper 2008). Sur cet écran de démarrage, l'invité reçoit confirmation qu'il est arrivé à la bonne place, est informé du contenu du sondage et de ses droits à la protection de la vie privée (compte tenu du contexte du pays), et est encouragé à passer au sondage proprement dit. Par conséquent, cette page particulière est un élément important en vue de « sceller l'affaire », c'est-à-dire à transformer l'invité en un répondant (Couper 2008, page 330). Toutefois, malgré son importance, les chercheurs n'ont accordé que peu d'attention, voire aucune, à l'influence de l'écran d'accueil sur les taux d'interruption (Couper 2008, page 330).

Cette situation est encore plus étonnante si l'on considère les riches fonctionnalités multimédias des sondages en ligne qui permettent de compléter le texte par toute une gamme d'éléments visuels, comme la couleur, les graphiques, la typographie et l'animation. La recherche expérimentale a montré que le contenu du texte de même que ses caractéristiques auxiliaires peuvent être des outils puissants en vue de soutenir l'intérêt des répondants durant le sondage et favoriser l'achèvement de la réponse au questionnaire (Couper, Traugott et Lamias 2001). Même si les répondants sont exposés à ces caractéristiques de conception dès l'apparition du premier écran, la plus grande partie de la recherche expérimentale a été limitée à l'influence de la conception du sondage dans son ensemble sur les taux d'interruption et n'a pas porté sur l'influence de la disposition et des caractéristiques de conception particulière de l'écran d'accueil.

Cela étant, l'objectif de la présente étude est d'explorer systématiquement certains des facteurs reliés à l'écran d'accueil qui pourraient influer sur la décision d'interrompre prématurément la participation au sondage. La présente étude rentre donc dans la catégorie des travaux de recherche axés sur les moyens d'augmenter les taux de réponse (Bosjnak et Tuten 2001). En déterminant ces facteurs, une attention particulière est accordée à l'utilisation de la couleur, à la longueur annoncée du sondage et aux variantes de la description des droits à la protection de la vie privée du répondant. Ces trois facteurs sont des éléments standard de l'écran d'accueil dont il est souvent question dans les manuels, mais qui sont négligés en recherche empirique.

La suite de l'article est présentée comme il suit. Premièrement, les données empiriques provenant des sondages en ligne sont utilisées pour élaborer nos hypothèses au sujet de l'influence de caractéristiques de conception particulière de l'écran d'accueil sur les taux d'interruption. Ensuite, nous décrivons l'étude expérimentale que nous avons réalisée en vue de tester ces hypothèses. Enfin, nous concluons l'article par une discussion et un énoncé des implications pour la conception des sondages en ligne.

2 Contexte théorique

Les interruptions ou ce qu'on appelle aussi les abandons représentent l'une des menaces les plus importantes en ce qui concerne l'inférence basé sur les données de sondage en ligne. Il s'agit de répondants qui abandonnent avant de répondre complètement au sondage (Bosjnak et Tuten 2001). Les interruptions, en tant qu'éléments du taux de non-réponse, peuvent nuire à la qualité des statistiques fondées sur le sondage; plus le taux d'interruption est élevé, plus le risque d'une erreur due à la non-réponse est grand. Donc, les spécialistes de la recherche par sondage ont consacré beaucoup d'effort à la réduction de ce taux (Groves, Fowler, Couper, Lepkowski, Singer et Tourangeau 2004). Dans la littérature sur la méthodologie des sondages en ligne, la plus grande partie de la recherche sur la façon d'améliorer la qualité des données en réduisant ce taux est axée sur le suivi des cas de non-réponse, sur les primes d'incitation ainsi que sur la longueur, le libellé et la présentation du questionnaire (Deutskens, De Ruyter, Wetzels et Oosterveld 2004, page 22). La plupart de ces caractéristiques visant à améliorer la réponse se concentrent sur les modifications apportées au contenu et à sa présentation. Autant que nous sachions, peu d'attention a été accordée jusqu'à présent aux lignes directrices concernant la disposition et le libellé de l'écran d'accueil. Par exemple, la recommandation de Dillman (2007, page 377) sur la façon de construire un écran d'accueil efficace souligne seulement que cette page particulière doit être motivante, en insistant sur la facilité de réponse, et qu'elle doit indiquer aux répondants comment aller à la page suivante. Des instructions détaillées et pratiques quant à la façon de concevoir un écran d'accueil efficace font défaut. Cela est étonnant, étant donné que la plupart des répondants abandonnent après le premier écran (c'est-à-dire ce que l'on appelle la non-réponse totale) (Couper 2008; Bosjnak et Tuten 2001). En outre, cet écran de démarrage particulier donne aux répondants une première impression du sondage. Il suscite des émotions à l'égard du questionnaire qui pourraient pousser le répondant non seulement à commencer à répondre au sondage en ligne, mais aussi à fournir les réponses plus rapidement, à passer sur les imperfections de la conception du sondage et peut-être même à répondre plus honnêtement (Dillman, Gertseva et Mahon-Haft 2005). En outre, il s'agit aussi d'une question d'esthétique, car les traits visuels déterminent les sentiments et la réaction émotionnelle d'un individu. Dans les sondages en ligne, l'écran d'accueil est le premier contact visuel du répondant et la disposition a donc une incidence sur les sentiments du répondant à l'égard du sondage dans son ensemble. On peut même supposer qu'une conception de sondage attrayante peut distraire de la mauvaise qualité du questionnaire proprement dit (Mahon-Haft et Dillman 2010).

Pour combler cette lacune dans la recherche empirique, nous testons trois facteurs intégrés dans l'écran d'accueil susceptibles d'avoir un effet sur les taux d'interruption de la participation à un sondage en ligne. Ces facteurs sont l'utilisation d'une couleur de fond, le nombre de mots utilisés pour expliquer les droits à la protection de la vie privée et la sécurité des données aux répondants prospectifs, et la durée annoncée du sondage en ligne. Ces trois éléments de l'écran d'accueil sont choisis parce qu'ils sont non seulement des éléments essentiels des écrans d'accueil, mais aussi parce qu'ils influent sur la première impression que les répondants pourraient avoir du sondage.

2.1 Couleur de fond

Contrairement aux sondages avec questionnaire papier, l'utilisation d'Internet ouvre toute une gamme de possibilités visuelles. Ce potentiel visuel est important, car les répondants font attention à de

nombreuses caractéristiques des questions du sondage et non pas seulement au libellé qui communique la question ou à la signification littérale de ces mots (Tourangeau, Couper et Conrad 2007). Ces caractéristiques non verbales comprennent le langage numérique, symbolique et graphique (Redline et Dillman 2002; Dillman 2007). Comme les nombres ou les symboles ne sont pour ainsi dire jamais utilisés dans la conception d'écrans d'accueil efficaces, les éléments graphiques non verbaux (c'est-à-dire luminosité, taille, forme, disposition spatiale, contraste, fond/forme, et même couleur) pourraient jouer un rôle important dans l'accroissement des taux de réponse.

Grâce à la flexibilité d'Internet, il est, par exemple, simple pour le concepteur du sondage de créer des combinaisons de texte et de fond dans une variété de couleurs différentes (Hall et Hanna 2004). Par conséquent, on observe dans les sondages en ligne des myriades de combinaisons de couleurs. Le choix d'une couleur particulière est relié au contraste visuel entre l'information verbale présentée et le fond coloré. Ce contraste est déterminé partiellement par les longueurs d'onde des couleurs. Ainsi, les couleurs saturées possèdent des longueurs d'onde différentes qui doivent être focalisées à différentes profondeurs derrière la lentille de l'œil, ce qui entraîne une fatigue visuelle (Couper 2008, page 164). En outre, les travaux de recherche ont montré que les répondants ont tendance à trouver les couleurs ayant une courte longueur d'onde (les bleus et les verts) plus plaisantes que celles ayant une grande longueur d'onde (les rouges et les jaunes) (Hall et Hanna 2004). Par exemple, Pope et Baker (2005) ont fait varier la couleur de fond d'un sondage auprès d'étudiants collégiaux en utilisant un fond bleu ou rose pour un sondage sur les problèmes liés à l'alcool. La réponse au sondage avec le fond bleu était plus rapide (mais les différences n'étaient pas statistiquement significatives).

Outre la longueur d'onde, les couleurs peuvent influencer la communication d'autres façons. La couleur possède une signification, qui peut tenir à des conventions culturelles, à des associations apprises ou aux actions associées à la couleur dans l'instrument proprement dit (Couper 2008, page 168). En d'autres termes, la couleur peut affecter émotionnellement les répondants. Ainsi, la couleur rouge est souvent associée au danger ou à la chaleur, surtout lorsqu'elle est alliée au bleu pour le froid (voir, par exemple, Gorn, Chattopadhyay, Yi et Dahl 1997). Quelques études se sont concentrées sur les émotions des utilisateurs lorsqu'ils répondent à des sondages en ligne colorés. Par exemple, Weller et Livingston (1988) ont constaté que la couleur du questionnaire avait effectivement un effet sur les réponses reçues. Spécifiquement, le rose produisait une réponse moins émotionnelle que le bleu.

Certaines études ont porté sur l'influence de la couleur sur les taux de réponse. Par exemple, Etter, Cucherat et Perneger (2002) ont conclu dans leur méta-analyse de dix études expérimentales que l'impression des questionnaires sur du papier coloré n'influençait pas considérablement la vitesse de réponse ni la proportion d'items manquants. Fait plus important encore, lorsque toutes les couleurs (bleu, vert ou jaune) ont été regroupées, aucune étude incluse dans la méta-analyse ne révélait un effet statistiquement significatif du papier de couleur (par opposition au papier blanc) sur le taux de réponse. La seule couleur qui avait un effet mineur (comparativement au blanc) était le rose. Les études qui ont été réalisées afin d'examiner l'influence de la couleur des questionnaires en ligne autoadministrés sur les taux de réponse montrent aussi que la couleur de fond peut avoir un certain effet, mais que celui-ci n'est pas systématiquement présent et n'est pas toujours très important (Couper 2008). Par exemple, Dillman, Conradt et Bowker (1998) et Hall et Hanna (2004) montrent dans leurs études que l'utilisation de lettres noires sur un fond blanc est la conception la plus efficace en ce qui concerne les taux de réponse. Une méta-analyse récente effectuée par Edwards, Roberts, Clarke, Diguiseppi, Wentz, Kwan, Cooper, Felix et

Pratap (2009) pour les sondages par la poste a également confirmé ce point. Ils ont constaté que les chances de réponse augmentaient d'un tiers en utilisant un fond blanc. En étendant cet argument à l'utilisation de couleurs dans les écrans d'accueil, nous nous attendons à ce que le groupe de répondants recevant un écran d'accueil dont les couleurs ont une grande longueur d'onde et qui intensifient les réactions émotionnelles négatives ait un taux d'interruption plus élevé que le groupe recevant un écran d'accueil simple ne comportant pas de nombreuses couleurs.

2.2 Droits à la protection de la vie privée

L'une des raisons possibles pour lesquelles les sondages en ligne donnent lieu à de faibles taux de réponse comparativement aux autres modes d'étude par sondage pourrait être les préoccupations concernant la confidentialité liée au courrier électronique et à Internet en général (Couper 2000). Alors que les questionnaires de sondage en ligne autoadministrés offrent la capacité de recueillir des renseignements de nature délicate entachés d'un plus faible biais de désirabilité, les préoccupations au sujet de la sécurité d'Internet peuvent annuler cet avantage, et éventuellement produire des taux plus élevés de non-réponse (ou des réponses moins honnêtes).

Il n'est donc par étonnant que, selon les règlements de leur pays, la plupart des chercheurs faisant appel au sondage en ligne renseignent les répondants prospectifs sur l'usage qui sera fait des renseignements qu'ils fourniront. En outre, ils insistent sur l'aspect volontaire de la participation au sondage et donnent souvent l'assurance que les noms des répondants ne seront jamais appariés d'aucune façon aux résultats de l'étude. Ces droits des répondants à la protection de la vie privée sont non seulement mentionnés dans le courriel d'invitation, mais aussi dans l'écran d'accueil. Ce dernier joue aussi un rôle important pour ce qui est de rassurer les répondants et de les motiver à commencer à répondre au sondage en ligne.

Quelques études ont eu pour objet d'examiner quels sont les effets des promesses concernant la protection de la vie privée et de la confidentialité sur les taux de réponse. Les premières études avaient trait, pour la plupart, au recensement décennal des ménages aux États-Unis et étaient fondées sur l'hypothèse que la promesse de protection de la vie privée était une « bonne » chose, c'est-à-dire qu'elle augmentait les taux de réponse en faisant disparaître les préoccupations des répondants (Singer, Hippler et Schwarz 1992, page 258; Singer, Von Thurn, Miller 1995, pages 66-67). Cependant, ces premières études ont prouvé le contraire; ces promesses réduisent la volonté de participer au sondage (Fay, Bates et Moore 1991; Singer, Mathiowetz et Couper 1993; Singer, Van Hoewyk et Neugebauer 2003; Hillygus, Nie, Prewitt et Pals 2006). Par exemple, Singer et coll. (1992) ont montré que la mention des droits à la protection de la vie privée avait un effet négatif sur la réponse, qu'il soit mesuré par la non-réponse partielle, par la non-réponse totale, par le taux de réponse ou par la qualité de la réponse. Ces études ont révélé des conséquences inattendues. Les promesses de protection de la confidentialité et de la vie privée pourraient en fait intensifier les préoccupations des participants au sujet du contenu du sondage. Ces promesses semblent modifier la perception qu'ont les répondants de la menace liée au sondage : elles suggèrent que le questionnaire pourrait contenir des questions désagréables, difficiles ou même embarrassantes. Autrement dit, ces promesses ont un effet d'amorçage de la réponse, c'est-à-dire qu'elles activent le concept de droit à la protection de la confidentialité et de la vie privée dans la mémoire du répondant, qui accorde alors plus de poids à ce concept dans la décision subséquente de participer ou non. En étendant cette constatation à la question de la mention des droits à la protection de la vie privée dans

l'écran d'accueil, nous nous attendons à ce que plus les chercheurs utilisent de mots pour expliquer ces droits, plus les répondants prospectifs seront susceptibles de prendre conscience de problèmes possibles au sujet de ces droits et moins ils seront susceptibles de vouloir participer au sondage en ligne. Cependant, nous n'avons aucune attente précise en ce qui concerne l'effet sur les taux d'interruption durant le sondage en ligne.

2.3 Durée annoncée

La décision de répondre à un sondage en ligne et d'aller jusqu'au bout est influencée en grande partie par l'effort demandé au répondant (Vicente et Reis 2010). Cet effort est déterminé partiellement par la longueur perçue du sondage (Bradburn 1978). Le bon sens nous dit que les longs sondages augmentent le coût de participation perçu et accroissent la probabilité que les gens interrompent prématurément leur participation.

Plusieurs études qui avaient pour objet d'examiner l'effet de la longueur du questionnaire sur les taux de réponse aux sondages en ligne ont donné des résultats contradictoires. Ainsi, la méta-analyse réalisée par Cook, Heath et Thomson (2000) n'a indiqué aucune corrélation significative entre la longueur du questionnaire et les taux de réponse aux sondages en ligne. Par contre, des études subséquentes ont indiqué que la longueur du questionnaire affectait les taux de réponse (Vicente et Reis 2010, page 256). Par exemple, Deutskens et coll. (2004) et Ganassali (2008) ont affirmé que le taux d'interruption était plus élevé pour la version longue de leur sondage en ligne que pour la version courte. En outre, Marcus, Bosnjak, Linder, Pilishenko et Schütz (2007) ont testé la relation entre la longueur du sondage en ligne et les taux de réponse dans le cadre d'une expérience sur le terrain. Ils ont constaté un effet significatif : 30,8 % de personnes répondaient au sondage court, mais seulement 18,6 % à la version plus longue. Cet effet important était significatif pour plusieurs autres modèles dans lesquels étaient inclus des contrôles pour diverses explications, comme l'importance du sujet du sondage ou l'utilisation de primes d'incitation.

Une question connexe est l'annonce a priori de la longueur du questionnaire. La relation entre cette annonce et le taux de réponse dépend toutefois davantage de la longueur perçue que de la longueur réelle du sondage. La durée annoncée est aussi un indicateur du fardeau perçu de réponse et elle influence la décision de participer et de continuer à participer. Les auteurs de quelques études ont expérimenté avec cette annonce. Par exemple, Crawford et coll. (2001) ont réalisé une expérience afin de déterminer si l'annonce préalable de la longueur du questionnaire aurait une incidence sur le pourcentage de personnes qui commencent à répondre au sondage et si le taux d'interruption serait plus élevé quand le sondage dépassait la durée promise. Comme ils l'avaient supposé, ils ont constaté que les répondants qui avaient été informés que la réponse au sondage ne prendrait que 8 à 10 minutes présentaient un taux de non-réponse global plus faible que ceux à qui l'on avait dit que cela prendrait 20 minutes. Toutefois, le taux d'interruption après avoir commencé à répondre au sondage était plus faible pour le groupe des 20 minutes. Ces résultats ont été confirmés par ceux d'autres études, comme celles de Hogg et Mill (2003), de Baker et Prewitt (2003) et de Galesic (2006).

La littérature traitant de l'effet de la durée annoncée sur les taux de réponse s'apparente étroitement à la discussion sur les avantages et les inconvénients de l'utilisation d'un indicateur de progression (voir, par exemple, Galesic et Bosnjak 2009; Heerwegh 2004). Par exemple, Yan, Conrad, Tourangeau et

Couper (2010) ont constaté que l'effet des indicateurs de progression dépend des attentes des répondants et de la mesure dans laquelle elles se concrétisent; la présence d'un indicateur de progression donnait lieu à moins d'interruptions lorsque les répondants s'attendaient à une tâche courte sur la base de l'invitation et quand le questionnaire était effectivement plus court que la longueur de la tâche à laquelle ils s'attendaient.

À l'instar de Crawford et coll. (2001), et contrairement aux deux autres facteurs de conception possibles, nous nous attendons à ce que l'annonce de la durée du sondage dans l'écran d'accueil influence non seulement la non-réponse initiale, mais aussi le taux d'interruption plus loin dans le sondage. Plus précisément, nous nous attendons à ce qu'un moins grand nombre de personnes commencent à répondre à un sondage en ligne quand la durée annoncée dans l'écran d'accueil est plus longue. En outre, ces répondants sont moins susceptibles d'abandonner le sondage une fois qu'ils ont commencé à répondre, puisque la durée réelle du sondage ne dépassera pour ainsi dire pas la durée perçue.

3 Conception et mise en œuvre de l'étude

Les expériences que nous décrivons ici ont été incluses dans un sondage auprès des étudiants de première année des programmes de baccalauréat et de maîtrise à l'Université de Constance par la section de la gestion de la qualité de l'université. Cette section voulait savoir pourquoi les étudiants choisissent d'étudier à Constance. Nous avons conçu le questionnaire en étroite collaboration avec la section de la gestion de la qualité, tandis que nous avons conceptualisé seuls les divers designs de l'écran d'accueil.

Les différentes conceptions de l'écran d'accueil ont été ajoutées après que le contenu de l'étude ait été déterminé. Nous avons testé les trois caractéristiques au moyen d'un plan d'expérience 2 x 2 x 2. Le tableau 3.1 donne un aperçu des groupes de contrôle et de traitement. Voir l'annexe A pour un exemple de l'un des six écrans d'accueil possibles.

Tableau 3.1
Plan de l'étude

		Droits à la protection de la vie privée			
		Consultables au moyen d'un lien		**Dans l'écran d'accueil**	
		Couleur de fond			
		blanc	**rouge**	**blanc**	**rouge**
Durée annoncée du sondage	courte (8 min.)	courte blanc lien	courte rouge lien	courte blanc écran	courte rouge écran
	longue (20 min.)	longue blanc lien	longue rouge lien	longue blanc écran	longue rouge écran

Pour tester l'influence de la couleur de fond sur la probabilité d'une interruption, nous avons sélectionné deux présentations : l'une comprenant un texte en noir sur un fond blanc et l'autre comprenant un texte en noir sur un fond rouge. Nous sommes conscients du fait que le rouge n'est pas une couleur de

fond très réaliste. Néanmoins, étant donné les résultats contradictoires des études antérieures, nous avons choisi cette couleur comme cas le plus probable d'interruption directement après l'écran d'accueil. Le rouge est une couleur saturée ayant une grande longueur d'onde. En outre, il peut avoir un effet négatif sur la réponse émotionnelle du répondant, puisqu'il est habituellement utilisé comme signal d'avertissement. Cependant, nous sommes conscients du fait que la conception de notre étude ne permet pas de déterminer clairement lequel des mécanismes discutés (c'est-à-dire longueur d'onde, saturation ou réponse émotionnelle) peut avoir un effet. Toutefois, nous pouvons donner certains premiers indices pour ce qui est de savoir si la couleur de l'écran d'accueil importe vraiment. Il convient de souligner que nous avons vérifié si l'affichage des couleurs de fond était le même dans les différents navigateurs.

Afin d'examiner l'effet de l'énoncé des droits à la protection de la vie privée sur le taux d'interruption, nous avons de nouveau utilisé deux conceptions : une version dans laquelle les droits à la protection de la vie privée étaient décrits en détail directement dans l'écran d'accueil, et une autre dans laquelle ils n'étaient mentionnés que brièvement et le répondant pouvait utiliser un lien Internet qui ouvrait une nouvelle fenêtre où ces droits à la protection de la vie privée étaient énoncés clairement, de la même façon que dans la première version.

Pour tester l'effet de la perception de la durée, c'est-à-dire de la durée annoncée du sondage en ligne dans l'écran d'accueil, nous avons annoncé deux durées différentes pour répondre au questionnaire. Nous avons utilisé le résultat du prétest comme guide pour estimer la durée. Ce résultat indiquait qu'il fallait environ 12 minutes pour répondre complètement au questionnaire. Par conséquent, nous avons décidé d'informer les personnes échantillonnées pour une version du sondage qu'il ne leur faudrait que 8 minutes environ pour répondre, ce qui correspondait à la durée minimale prévue pour répondre au questionnaire, tandis qu'à un autre groupe de répondants, nous avons indiqué que cela prendrait environ 20 minutes. Le temps nécessaire pour répondre au sondage dépendait beaucoup du nombre de réponses que les répondants donneraient au sondage, puisque le questionnaire contenait de nombreux filtres. Le temps moyen réel pour répondre complètement au questionnaire était de 17,81 minutes (avec une écart-type de 9,01), durée considérablement plus longue que ne l'indiquait le prétest.

L'invitation à participer au sondage en ligne a été envoyée à chacun des 2 629 comptes de courrier électronique de l'université appartenant aux étudiants de première année (voir l'annexe B). Nous nous sommes concentrés sur cette population d'étudiants, parce que nous avons supposé qu'ils n'avaient pas été exposés fréquemment aux sondages en ligne réalisés par l'université et qu'ils seraient par conséquent plus enclins à répondre à ce genre de sondage sans opter pour la « suffisance » (en anglais, *satisficing*) (Toepoel, Das et Van Soest 2008). Une fois que les étudiants ont cliqué sur le lien vers le sondage (n = 1 419), ils ont été assignés aléatoirement à l'un des huit groupes, avec un minimum de 151 et un maximum de 185 étudiants dans chaque groupe de traitement. En moyenne, le nombre d'étudiants par groupe de traitement était de 177 environ avec un écart-type de 8,5 (voir le tableau 4.1 pour le nombre exact de répondants par groupe de traitement).

Puisque l'information fournie dans le courriel d'invitation recoupait souvent celle figurant à l'écran d'accueil, nous avons limité dans la mesure du possible les instructions dans le courriel afin d'isoler les effets possibles de l'écran d'accueil. En outre, pour que la présentation du courriel soit aussi simple que possible, nous n'avons utilisé aucun format HTML. Avec le courriel d'invitation, les étudiants ont reçu une adresse URL unique dans laquelle était intégré leur mot de passe personnel, ceci afin d'éviter qu'un étudiant réponde plus d'une fois au sondage. Ces courriels invitant les étudiants à participer au sondage

ont été envoyés le 8 novembre 2011 après la réalisation d'un prétest qualitatif auprès de 15 personnes qui travaillaient dans le service d'administration de l'université, qui étaient des étudiants ou qui avaient de l'expérience en recherche par sondage. Ce prétest était axé sur les aspects techniques du sondage et sur le libellé du texte et des questions.

Cinq jours après l'envoi du premier courriel, un rappel a été envoyé à ceux qui n'avaient pas encore participé au sondage. Un dernier rappel a été envoyé le 18 novembre 2011 non seulement à ceux qui n'avaient pas encore participé, mais aussi à ceux qui avaient commencé à répondre au sondage, mais n'avaient pas terminé. Le sondage a été clôturé le 5 décembre 2011. Des 2 629 étudiants, 1 419 ont commencé à répondre au sondage, et de ceux-ci, 1 118 ont répondu complètement. L'achèvement du sondage signifie que le participant est arrivé à la dernière page. Comme toutes les questions importantes étaient mises en œuvre sous forme d'un choix forcé, la non-réponse partielle ne s'applique pas. Ces chiffres donnent un taux de réponse de 43 % selon la norme RR1 de l'*American Association for Public Opinion Research* (AAPOR) et de 54 % selon la norme RR2 de l'AAPOR, qui tient compte de la réponse partielle (AAPOR 2011).

Le sondage en ligne a été mis en œuvre en utilisant le programme Unipark, qui est un logiciel de sondage en ligne qui permet aux utilisateurs de créer ce genre de sondage avec un minimum d'effort. Le programme permet une programmation simple et, comparativement aux produits d'autres fabricants, son prix est faible pour l'usage scientifique. Unipark est assez souple à différents égards. Par exemple, les participants peuvent arrêter de répondre au questionnaire et continuer plus tard. En outre, le système enregistre l'information lorsque le participant interrompt la session et le temps dont ils ont besoin pour remplir le questionnaire et les écrans individuels. En outre, il permet d'intégrer des outils complexes et non standardisés dans la conception du questionnaire. Malgré l'usage croissant d'appareils mobiles, nous n'avons pas implémenté de page Web mobile, ce qui s'est avéré être une bonne décision, car 65 participants seulement ont utilisé un appareil mobile pour répondre au questionnaire. Cependant, s'ils étaient plus susceptibles d'avoir accès au sondage, ces participants étaient aussi plus susceptibles d'interrompre leur participation (21 % de membres du groupe sans appareil mobile et 35 % groupe avec appareil mobile, $p < 0{,}01$).

4 Résultats

Avant de présenter les résultats statistiques, nous comparons les proportions de notre échantillon (c'est-à-dire les étudiants qui ont participé au sondage) avec celles dans la population (c'est-à-dire tous les étudiants qui ont reçu une invitation). L'invitation a été envoyée à 2 629 étudiants, dont un peu plus de 47 % étaient de sexe masculin. De ceux qui ont participé (1 419), un peu plus de 44 % étaient de sexe masculin. Il ne semble pas exister de différence entre les comportements de réponse des hommes et des femmes. Cependant, certaines différences se dégagent lorsque l'on compare les différentes facultés. Alors que la faculté des sciences semble être représentée adéquatement dans l'échantillon (30 % environ dans l'échantillon ainsi que dans la population), la faculté des sciences humaines semble être surreprésentée (43 % dans l'échantillon, 29 % dans la population), et la faculté de science politique, de droit et d'économie semble sous-représentée (25 % dans l'échantillon, 40 % dans la population).

Dans le tableau 4.1, les taux d'interruption et le nombre absolu d'étudiants qui ont participé au sondage en ligne sont présentés par groupe de traitement. Notons que 83 étudiants ont abandonné après avoir vu le premier écran, tandis que 218 autres l'ont fait à d'autres pages du sondage en ligne, pour un total de 301 interruptions. Ce tableau descriptif montre aussi que les taux d'interruption (directement après l'écran d'accueil ou pour toutes les autres pages confondues) sont généralement plus faibles pour les répondants qui ont reçu un écran d'accueil dans lequel les droits à la protection de la vie privée n'étaient pas mis en relief et dans lequel la durée annoncée du sondage était sous-estimée. Par contre, l'influence de la couleur de l'écran d'accueil sur les taux d'interruption semble être variable.

Afin de vérifier si les tendances observées au tableau 4.1 sont robustes et statistiquement significatives, nous avons procédé à des régressions logit avec trois variables dépendantes différentes. Premièrement, une variable dichotomique indiquant si le répondant s'était interrompu directement à l'écran d'accueil (code 1, ou 0 autrement). Deuxièmement, une variable dichotomique prenant la valeur 1 si le répondant abandonnait à n'importe quelle autre page que l'écran d'accueil (code 0 autrement). Troisièmement, une variable dichotomique indiquant si le répondant avait abandonné à n'importe quelle page du sondage (aussi bien l'écran d'accueil que toute autre page, code 1, ou 0 autrement). Cependant, dans le cas de la dernière mesure, nous ne pouvons pas prouver clairement que l'effet est dû principalement à l'écran d'accueil ou à une interaction entre l'écran d'accueil et l'ensemble du sondage en ligne (ou de pages particulières).

Tableau 4.1
Interruption dans les différents groupes expérimentaux

	Interruption à l'écran d'accueil		Taux d'interruption global		Nombre total n de répondants par groupe de traitement
	n	%	*n*	%	
Blanc, courte, lien	2	1,07	35	18,72	187
Blanc, courte, écran	4	2,30	31	17,83	174
Blanc, longue, lien	15	7,89	49	25,79	190
Blanc, longue, écran	18	10,40	50	28,90	173
Rouge, courte, lien	3	1,79	18	10,71	168
Rouge, courte, écran	12	6,78	35	19,77	177
Rouge, longue, lien	13	7,10	37	20,22	183
Rouge, longue, écran	16	9,58	46	27,54	167
Total des n	83		301		1 419
Moyenne des n (écart-type)	10,4 (6,4)		37,6 (10,7)		177,4 (8,5)

Les résultats des régressions logit sont présentés au tableau 4.2. La deuxième colonne de ce tableau donne les effets des différents traitements sur la probabilité que les répondants abandonnent à l'étape de l'écran d'accueil. La troisième colonne montre l'effet des différents traitements sur la probabilité d'une interruption durant le sondage, en excluant les répondants qui ont abandonné à l'écran d'accueil. La quatrième colonne donne l'effet des caractéristiques de conception sur la probabilité globale d'interruption. Nous avons également estimé les modèles comprenant tous les effets d'interactions possibles entre les variables expérimentales. Toutefois, les résultats n'indiquaient pas clairement qu'une combinaison de divers traitements augmentait considérablement les effets. En outre, nous avons inclus les

effets des interactions entre les variables expérimentales et les variables de sous-groupe telles que le sexe ou la faculté. Comme nous n'avons pas pu dégager de différence non ambiguë entre les sous-groupes ou que les variations à l'intérieur des sous-groupes étaient trop faibles pour estimer le modèle, les résultats de ces interactions ne sont pas présentés ni discutés à la section suivante où nous donnons seulement les résultats des modèles parcimonieux.

Conformément aux études antérieures, nous nous attendons à ce que les répondants qui reçoivent un écran d'accueil rouge soient plus susceptibles d'abandonner que ceux qui reçoivent un écran blanc. Bien que le coefficient logit positif dans la deuxième colonne du tableau indique qu'il existe effectivement une relation positive entre le fond rouge et un plus haut niveau d'interruption à l'étape de l'écran d'accueil, cette relation n'est pas statistiquement significative. Toutefois, l'écran d'accueil rouge a un effet négatif statistiquement significatif sur le taux d'interruption à n'importe quelle page sauf l'écran d'accueil, ce qui signifie que la combinaison d'un écran d'accueil rouge et des autres écrans blancs du questionnaire semble encourager les participants à poursuivre. Si l'on examine l'effet de l'écran d'accueil rouge sur le taux global d'interruption, le coefficient présenté dans la quatrième colonne du tableau donne à penser que la couleur de l'écran d'accueil n'a pas d'effet significatif. Cette observation indique que même s'il est important, l'écran d'accueil est simplement un des écrans du sondage en ligne. Comme l'a suggéré l'un des participants au prétest, il se pourrait que la couleur rouge suscite un sentiment tellement négatif que les répondants cliquent immédiatement pour aller plus loin sans regarder l'écran. Cette idée a été testée en exécutant une régression par la méthode des moindres carrés ordinaires (MCO) de la couleur et des autres traitements utilisés comme contrôle sur la quantité de temps passé à l'écran d'accueil. Cependant, les résultats (disponibles sur demande) n'ont pas fourni la preuve d'un effet statistiquement significatif. Notons que les valeurs du pseudo R-carré présentées pour le modèle (et pour l'ensemble des modèles) sont assez faibles. Toutefois, il s'agit d'un résultat donné fréquemment par les régressions logit en vue d'analyser des résultats expérimentaux. Par exemple, Marcus et coll. (2007) signalent un R-carré de Nagelkerke de 0,041 et Bandilla, Couper et Kaczmirekt (2012), un pseudo R-carré de 0,05.

Tableau 4.2
Régression logit

	(1) Interruption à l'écran d'accueil	(2) Interruption à n'importe quelle page sauf l'écran d'accueil	(3) Interruption à n'importe quel moment du sondage
Couleur de fond : rouge	0,17	-0,33**	-0,20
	(0,23)	(0,15)	(0,13)
Durée annoncée : 20 minutes	1,15***	0,23	0,53***
	(0,26)	(0,15)	(0,13)
Information sur la sécurité des données :	-0,52**	-0,14	-0,28**
consultable au moyen d'un lien	(0,23)	(0,15)	(0,13)
Constante	-3,34***	-1,61***	-1,37***
	(0,27)	(0,15)	(0,13)
N	1 419	1 419	1 419
Pseudo R-carré	0,04	0,01	0,02
Prob > khi^2	0,00	0,05	0,00

Erreurs-types entre parenthèses
*** p<0,01, ** p<0,05, * p<0,1

Nous soupçonnions aussi que les répondants recevant un écran d'accueil annonçant que le sondage prendrait seulement 20 minutes étaient moins susceptibles de commencer à y répondre que ceux qui recevaient un écran d'accueil indiquant que le sondage prendrait seulement 8 minutes. Cette attente théorique est étayée statistiquement par un coefficient logit positif et statistiquement significatif de 1,15. En outre, nous avons supposé que les répondants qui commençaient à répondre au « long » sondage étaient moins enclins à abandonner pendant le sondage en ligne. Cependant, nous n'avons dégagé aucune preuve à l'appui de cette hypothèse. Le coefficient non significatif de 0,23 signifie qu'il n'y a pas de différence significative de taux d'interruption à n'importe quel écran sauf la première page entre les répondants auxquels il a été annoncé que le sondage durerait « 8 minutes » et ceux auxquels il a été annoncé qu'il leur faudrait 20 minutes pour répondre. Dans l'ensemble, le coefficient positif et significatif de 0,53 dans la cinquième colonne du tableau indique que les répondants qui ont vu un écran d'accueil sur lequel il était indiqué que le sondage en ligne prendrait 20 minutes étaient plus susceptibles d'interrompre le sondage que ceux auxquels on avait annoncé que cela prendrait seulement 8 minutes. Dans l'ensemble, le coefficient de la durée annoncée est celui qui est le plus grand dans les différents modèles, ce qui signifie que le facteur le plus important expliquant le taux d'interruption est la durée annoncée à l'écran d'accueil. Ce résultat correspond aux résultats de l'étude de Galesic et Bosnjak (2009), qui ont montré que plus la durée annoncée était grande, plus faible était le nombre de répondants qui commençaient à répondre au questionnaire et y répondaient au complet.

La dernière caractéristique de conception que nous avons fait varier à l'écran d'accueil était l'importance accordée à l'énoncé des droits à la protection de la vie privée des répondants. Nous nous attendions à ce que plus les droits seraient soulignés, plus les répondants auraient conscience des problèmes éventuels concernant ces droits et moins ils seraient disposés à commencer à répondre au sondage en ligne. Les résultats des modèles logit appuient cette notion. Le coefficient négatif de -0,52 indique que, quand les droits à la protection de la vie privée sont expliqués par la voie d'un lien figurant à l'écran du sondage en ligne (six répondants seulement ont effectivement ouvert ce lien), c'est-à-dire quand moins de mots sont utilisés pour expliquer ces droits sur l'écran proprement dit, les taux d'interruption à l'étape de l'écran d'accueil diminuent. Autrement dit, la mise en évidence des droits à la protection de la vie privée à l'écran d'accueil augmente la non-réponse. En outre, l'explication détaillée des droits à la protection de la vie privée à l'écran d'accueil a aussi une influence sur les taux d'interruption durant tout le sondage. Cependant, nous ne savons pas avec certitude si la diminution du taux de réponse est due à l'importance accordée aux droits à la protection de la vie privée ou à la longueur de l'écran d'accueil (expliquer les droits à la protection de la vie privée à l'écran d'accueil produit un plus long écran). La recherche devra se poursuivre afin de faire la distinction entre ces deux processus apparentés.

5 Conclusion et discussion

L'un des défis les plus importants pour les concepteurs de sondage en ligne est d'arriver à ce que les répondants se connectent au site du sondage et à soutenir leur motivation à achever le sondage une fois qu'ils l'ont commencé. Cependant, alors que la plupart des concepteurs ont examiné la disposition et la conception du sondage en ligne proprement dit et la façon dont elles affectent les taux d'interruption, ils ont accordé moins d'attention au rôle important que joue l'écran d'accueil. Ce premier écran d'un sondage

en ligne transforme l'invité en un répondant et influe sur ses premières impressions du sondage. En outre, la recherche empirique a montré que la plupart des répondants abandonnent après cette première page.

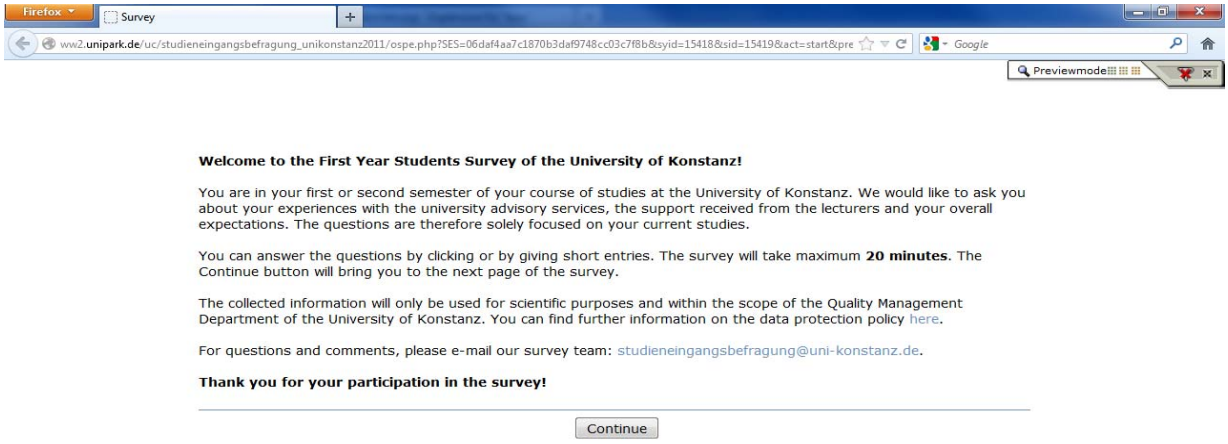
L'objectif de la présente étude était d'évaluer certains facteurs reliés à la conception de l'écran d'accueil des sondages en ligne qui ont une incidence sur les taux de réponse dans l'environnement électronique. Afin d'examiner cette influence, nous avons intégré un plan d'expérience 2 x 2 x 2 dans un sondage en ligne. Les résultats donnent à penser que la présentation de l'écran d'accueil joue un rôle important dans la communication avec le répondant. Plus la durée du sondage annoncée à l'écran d'accueil est longue et plus l'attention qui est accordée à l'explication des droits à la protection de la vie privée du répondant est grande, plus faible est le nombre de répondants qui commencent à répondre au sondage en ligne et qui achèvent de le faire. Cependant, la couleur de fond de l'écran d'accueil n'avait pas d'effets statistiquement significatifs sur le niveau des taux d'interruption à l'étape de l'écran d'accueil. Seul un effet durant le sondage en ligne proprement dit a été observé, mais, comme on ne constate pas d'effet significatif sur le taux global d'interruption, cette caractéristique de la conception de l'écran d'accueil n'est pas considérée comme pertinente. Dans l'ensemble, sur la base de ces résultats, nous pouvons énoncer certaines considérations, qui pourraient aider à améliorer la pratique des sondages en ligne : 1) Veiller à ce que le sondage en ligne soit aussi court que possible; 2) Procéder à des prétests élaborés pour obtenir de l'information fiable sur le temps nécessaire pour répondre au sondage; 3) Fournir des renseignements sur les droits à la protection de la vie privée dans l'écran d'accueil, mais tout en tenant compte du fait que la plupart des répondants préfèrent une courte description de ces droits et ne veulent pas passer trop de temps à les lire. Un moyen approprié de satisfaire à ces souhaits consiste à fournir aux répondants un lien vers une description plus détaillée de ces droits.

Une limite importante de la présente étude tient au fait qu'elle a été réalisée auprès d'un échantillon tiré d'une population très particulière, à savoir des étudiants universitaires de première année. Il est fort probable que le sujet du sondage en ligne, c'est-à-dire le choix de poursuivre des études, ait une grande importance pour les étudiants que d'autres sujets, ce qui pourrait accroître le niveau général de réponse. La recherche par sondage en ligne sur d'autres populations que les étudiants devrait permettre de déterminer si les résultats présentés ici sont robustes. En outre, il serait intéressant de déterminer l'effet précis de l'importance accordée à l'énoncé des droits à la protection de la vie privée sur les taux de réponse. Est-il attribuable au nombre de mots consacrés à ce sujet ou s'agit-il de l'accent mis sur les problèmes éventuels liés à la protection de la vie privée ? Les mécanismes particuliers de l'effet observé ne sont pas clairs et une étude plus approfondie est par conséquent nécessaire.

Remerciements

Nous remercions Christine Abele, Valentin Gold, Katharina Holzinger et Elena Sewelies de leurs suggestions utiles, de leurs commentaires, de leur soutien (pour réunir les données) et de leur collaboration.

Annexe A



Annexe B

M...,

Vous commencez maintenant le premier ou le deuxième semestre de votre programme d'études à l'Université de Constance. Au cours de ces quelques premières semaines, vous avez appris à connaître avec votre département, l'université et la ville. Nous aimerions savoir quelle a été votre expérience en ce qui concerne les services consultatifs de l'université, le soutien offert par les chargés de cours et vos attentes globales. Nous vous invitons donc à participer à notre sondage auprès des étudiants de première année. Votre participation à ce sondage nous aidera à améliorer les conditions d'études à l'Université de Constance.

Veuillez cliquer sur le lien vers le sondage (version anglaise) qui suit :

http://personalizedlink

Si vous ne pouvez pas accéder au sondage au moyen du lien, veuillez copier et coller l'adresse dans votre navigateur Web.

La participation au sondage est volontaire. Le sondage est assujetti au règlement concernant la protection des données et l'information que vous fournirez sera utilisée uniquement par le département de la gestion de la qualité de l'Université de Constance et à des fins scientifiques. D'autres renseignements sur la protection des données figurent à l'écran d'accueil du sondage.

Si vous avez des questions ou des commentaires, veuillez communiquer par courriel avec l'équipe du sondage à l'adresse : studieneingangsbefragung@uni-konstanz.de.

Nous vous remercions de votre participation et nous vous présentons nos vœux de réussite durant vos études !

Cordialement,

Bibliographie

American Association for Public Opinion Research (AAPOR) (2011). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Récupérée le 27 février 2013 à partir du http://aapor.org/Content/NavigationMenu/AboutAAPOR/StandardsampEthics/StandardDefinitions/StandardDefinitions2011.pdf.

Baker-Prewitt, J. (2003). *All Web Surveys are not Created Equal: Your Design Choices Can Impact Results.* Document présenté à la SumIT03 Global Market Research Symposium, Montréal.

Bandilla, W., Couper, M.P. et Kaczmirekt, L. (2012). The mode of invitation for web surveys. *Survey Practice*, 5.

Bosnjak, M., et Tute, T.L. (2001). Classifying response behavior in web surveys. *Journal of Computer Medicated Communication*, 6. Récupérée le 21 novembre 2012 à partir du http://jcmc.indiana.edu/vol6/issue3/boznjak.html.

Bradburn, N.M. (1978). Respondent Burden. *Proceedings of the Section of Survey Research Methods*, American Statistical Association.

Brewer, P.B., Graf, J. et Willnat, L. (2003). Priming or Framing. Media Influence on Attitudes toward Foreign Countries. *Gazette: The International Journal for Communication Studies*, 65, 493-508.

Cook, C., Heath, F. et Thompson, R.L. (2000). A meta-analysis of response rates in web- or Internet-based surveys. *Educational and Psychological Measurement*, 60, 821-836.

Couper, M.P. (2000). Web surveys. A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P. (2008). *Designing Effective Web Surveys*. Cambridge: Cambridge University Press.

Couper, M.P., Traugott, M.W. et Lamais, M.J. (2001). Web survey design and administration. *Public Opinion Quarterly*, 65, 230-253.

Couper, M.P., et Miller, P.V. (2008). Web survey methods. Introduction. *Public Opinion Quarterly*, 72, 831-835.

Crawford, S.D., Couper, M.P. et Lamias, M.J. (2001). Web surveys. Percepsintions of Burden. *Social Science Computer Review*, 19, 146-62.

Dillman, D.A., Conradt, J. et Bowker, D. (1998). *Influence of Plain vs. Fancy Design on Response Rates for Web Surveys.* Document présenté à la réunion annuelle de l'American Statistical Association, Dallas, TX.

Dillman, D.A., Gertseva, A. et Mahon-Haft, T. (2005). Achieving usability in establishment survey through the application of visual design principles. *Journal of Official Statistics*, 21, 183-214.

Dillman, D.A. (2007). *Mail and Internet Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons, Inc.

Deutskens, E., De Ruyter, K., Wetzels, M. et Oosterveld, P. (2004). Response rate and response quality of Internet-based surveys: An experimental study. *Marketing Letters*, 15, 21-36.

Edwards, P.J., Roberts, I., Clarke, M.J., Diguiseppi, C., Wentz, R., Kwan, I., Cooper, R., Felix, L.M. et Pratap, S. (2009). Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*, 8, 3. Récupérée le 21 novembre 2012 à partir du http://www.ncbi.nlm.nih.gov/pubmed/19588449.

Etter, J., Cucherat, M. et Perneger, T.V. (2002). Questionnaire color and response rates to mailed surveys: A randomized trial and a meta-analysis. *Evaluation & The Health Professions*, 25, 185-99.

Faubert, J. (1994). Seeing depth in colour: More than just what meets the eyes. *Vision Research*, 34, 1165-1186.

Fay, R.E., Bates, N. et Moore, J. (1991). Lower mail response in the 1990 Census: A preliminary interpretation. Dans *Proceedings of the Annual Research Conference of the U.S. Census Bureau*, 3-32. Washington DC: Census Bureau. Récupérée le 27 novembre 2012 à partir du https://www.census.gov/srd/papers/pdf/rsm2010-13.pdf.

Galesic, M. (2006). Dropouts on the web: Effects of interest and Burden experienced during an online survey. *Journal of Official Statistics*, 22, 313-328.

Galesic, M., et Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73, 349-360.

Ganassali, S. (2008). The influence of the design of web survey questionnaires on the quality of responses. *Survey Research Methods*, 2, 21-32.

Gorn, G.J., Chattopadhyay, A., Yi, T. et Dahl, D.W. (1997). Effects of color as an executional cue in advertising: They're in the shade. *Management Science*, 43, 1387-1400.

Groves, R.M., Fowler Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E. et Tourangeau, R. (2004). *Survey Methodology*. Hoboken: John Wiley & Sons, Inc.

Hall, R.H., et Hanna, P. (2004). The impact of web page text-background color combinations on readability, retention, aesthetics, and behavioral intention. *Behaviour & Information Technology*, 23, 183-195.

Heerwegh, D. (2004). Using Progress Indicators in Web Surveys. Paper prepared for the 59th AAPOR conference (Phoenix, Arizona 13 au 16 mai 2004). Récupérée le 27 novembre 2012 à partir du https://perswww.kuleuven.be/~u0034437/public/Files/Heerwegh%20Using%20Progress%20Indicators.pdf.

Hillygus, S., Nie, N., Prewitt, K. et Pals, G. (2006). Civic Mobilization and Privacy Concerns in the 2000 Census. New York: Russell Sage Foundation.

Hogg, A., et Miller, J. (2003). Watch out for Dropouts. Récupérée le 18 avril 2011 à partir du http://www.quirks.com.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I. et Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, 50, 79-104.

Mahon-Haft, T.A., et Dillman, D.A. (2010). Does visual appeal matter? Effects of web survey aesthetics on survey quality. *Survey Research Methods*, 4, 43-59.

Marcus, B, Bosnjak, M., Linder, S., Pilischenko, S. et Schütz, A. (2007). Compensating for low topic interest and long surveys. A field experiment on nonresponse in web surveys. *Social Science Computer Review*, 25, 372-383.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, 73, 74-97.

Pope, D., et Baker, R.P. (2005). Experiments in Color for Web-Based Surveys. Document présenté à l'atelier de la FedCASIC, Washington, D.C.

Redline, C., et Dillman, D.A. (2002). The influence of alternative visual designs of respondent's performance with branching instructions in self-administered questionnaire. Dans *Survey Response*, (Éds., R. Groves, D.A. Dillman, E. Eltinge et R. Little), 179-196. New York: John Wiley & Sons, Inc.

Singer, E., Hippler, H.J. et Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research*, 4, 256-268.

Singer, E., Mathiowetz, N. et Couper, M.P. (1993). The role of privacy and confidentiality as factors in response to the 1990 census. *Public Opinion Quarterly*, 57, 465-82.

Singer, E., Von Thurn, D.R. et Miller, E.R. (1995). Confidentiality assurances and response: A quantitative review of the experimental literature. *Public Opinion Quarterly*, 59, 66-77.

Singer, E., Van Hoewyk, J. et Neugebauer, R.J. (2003). Attitudes and Behavior - the impact of privacy and confidentiality concerns on participation in the 2000 census. *Public Opinion Quarterly*, 67, 368-384.

Terhanian, G. (2005). How to Produce Credible, Trustworthy Information through Internet-Based Survey Research. Document présenté à la conférence annuelle de l'American Association for Public Opinion Research, Portland, OR.

Tourangeau, R., Couper, M.P. et Conrad, F. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, 71, 91-112.

Toepoel, V., Das, M. et Van Soest, A. (2008). Effects of design in web surveys. Comparing trained and fresh respondents. *Public Opinion Quarterly*, 72, 985-1007.

Vicente, P., et Reis, E. (2010). Using questionnaire design to fight nonresponse bias in web surveys. *Social Science Computer Review*, 28, 251-267.

Weller, L., et Livingston, R. (1988). Effect of color of questionnaire on emotional responses. *The Journal of General Psychology*, 115, 433-440.

Yan, T., Conrad, F.G., Tourangeau, R. et Couper, M.P. (2010). Should I stay or should I go: The effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. *International Journal of Public Opinion Research,* 23, 131-147.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2013.

S.R. Amer, *RTI International, US*
R. Andridge, *Ohio State University*
N. Bates, *U.S. Census Bureau*
R. Bautista, *NORC at the University of Chicago*
J.-F. Beaumont, *Statistique Canada*
W. Bell, *U.S. Census Bureau*
D. Bellhouse, *University of Western Ontario*
E. Benhin, *Statistique Canada*
Y.G. Berger, *University of Southampton*
P. Biemer, *RTI*
H.J. Boonstra, *Statistics Netherlands*
J. Breidt, *Colorado State University*
B. Buelens, *Statistics Netherlands*
M. Callegaro, *Google Ltd, London*
P. Cantwell, *U.S. Census Bureau*
I.A. Carrillo, *RTI*
R. Chambers, *NIASRA*
R. Chambers, *University of Wollongong, Australia*
G. Chauvet, *ENSAI, France*
J. Chipperfield, *Australian Bureau of Statistics*
G. Datta, *University of Georgia*
T. DeMaio, *U.S. Census Bureau*
S. Dipko, *Westat*
G.B. Durrant, *University of Southampton*
S. Er, *Istanbul University School of Business*
V. Estevao, *Statistique Canada*
R. Fecso, *Ernst & Young LLP*
O. Fischer, *U.S. Census Bureau*
T.I. Garner, *U.S. Bureau of Labour Statistics*
C. Girard, *Statistique Canada*
Y. He, *National Center for Health Statistics*
R. Janicki, *U.S. Census Bureau*
J. Jiang, *U. of California Davis*
D. Judkins, *Abt Associates*
M.G.M. Khan, *University of the South Pacific, Fiji*
J.-K. Kim, *Iowa State University*
B. Klemens, *U.S. Census Bureau*
P. Kokic, *CSIRO, Australia*
S. Kolenikov, *Abt SRBI*
P.S. Kott, *RTI*
P. Lavallée, *Statistique Canada*
H. Lee, *Westat*
L. Lee, *NORC at the University of Chicago*
D. Liao, *RTI*
M. Link, *Neilson*
Y. Lu, *University of New Mexico*
P. Lynn, *University of Essex*
H. Mantel, *Statistique Canada*
D. Marker, *Westat*
A. Matei, *Université de Neuchâtel*
T. Merkouris, *Statistique Canada*
A. Natei, *University of Neuchatel*
J. Opsomer, *Colorado State University*
N. Prasad, *University of Alberta*
G. Ranalli, *University of Perugia, Italy*
A.P. Rao, *Statistical Consultant*
J.N.K. Rao, *Carleton University*
J. Ridenhour, *RTI*
L.-P. Rivest, *Université Laval*
L. Rizzo, *Westat*
E. Robison, *U.S. Bureau of Labor Statistics*
A. Scott, *University of Aukland*
H.-C. Shin, *National Center for Health Statistics*
N. Shlomo, *U of Manchester*
Y. Si, *Columbia University*
J. Siddique, *Northwestern University Feinberg School of Medicine*
R. Sigman, *Westat*
E.V. Slud, *U.S. Census Bureau*
P. Smith, *Office for National Statistics*
M. Sverchkov, *U.S. Bureau of Labor Statistics*
A. Théberge, *Statistique Canada*
Y. Tillé, *University of Neuchatel*
R. Thomas, *Carleton University*
M. Thompson, *University of Waterloo*
M. de Toledo Vieira, *Federal University of Juiz de For a*
R. Varriale, *ISTAT*
M. Williams, *National Agricultural Statistics Service, USDA*
J. Wood, *Office for National Statistics*
C. Yu, *Iowa State University*
G. Zhang, *National Center for Health Statistics*
L.-C. Zhang, *University of Southampton*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2013 : Céline Ethier de la Division de la recherche et de l'innovation en statistique, Joana Bérubé de la Division des méthodes auprès des entreprises, l'équipe de la Division de la diffusion, en particulier Daniel Piché, Jasvinder Jassal, Martin Lachance, Jacqueline Luffman, Kathy Charbonneau et Lucie Gauthier, de même que Julie Dion de la Division des systèmes administratifs et de diffusion.


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

ANNONCES

Demande de candidatures pour le prix Waksberg 2015

La revue *Techniques d'enquête* a mis sur pied une série annuelle de communications sollicitées en l'honneur de Joseph Waksberg, en reconnaissance des contributions qu'il a faites à la méthodologie d'enquête. Chaque année, un éminent statisticien d'enquête est choisi pour rédiger un article où il examine l'évolution et l'état actuel d'un thème important du domaine de la méthodologie d'enquête. L'article reflète le mélange de théorie et de pratique caractéristique des travaux de Joe Waksberg.

Le lauréat du prix Waksberg recevra une prime en argent et présentera la communication sollicitée Waksberg 2015 au Symposium de Statistique Canada qui se tiendra à l'automne de 2015. L'article paraîtra dans un numéro de *Techniques d'enquête* (publication prévue pour décembre 2015).

L'auteur de l'article Waksberg 2015 sera choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*. Les candidatures ou les suggestions de thèmes doivent être envoyées avant le 28 février 2014 à la présidente du comité, Cynthia Clark (cynthia_clark@nass.usda.gov).

Les gagnants et articles précédents du prix Waksberg sont

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

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

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

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

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

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

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

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

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

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

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

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

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

2014 Connie **Citro**, Sujet de l'article à l'étude.

Membres du comité de sélection de l'article Waksberg (2013-2014)

Cynthia Clark, *USDA* (Présidente)
Louis-Paul Rivest, *Université de Laval*
Tommy Wright, *U.S. Bureau of the Census*
J.N.K. Rao, *Carleton University*

Présidents précédents :

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)
Elizabeth A. Martin (2010 - 2011)
Mary E. Thompson (2011 - 2012)
Steve Heeringa (2012 - 2013)

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 29, No. 2, 2013

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 29, No. 3, 2013

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 41, No. 2, June/juin 2013

CONTENTS TABLE DES MATIÈRES

Volume 41, No. 3, September/septembre 2013