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

Techniques d'enquête

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

par David G. Steel et Robert Graham Clark

Date de diffusion : 19 décembre 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Programme des services de dépôt

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

Comment accéder à ce produit

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

Normes de service à la clientèle

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

Publication autorisée par le ministre responsable de Statistique Canada

This publication is also available in English.

Note de reconnaissance

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

Signes conventionnels

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

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

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

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

David G. Steel et Robert Graham Clark[1]

Résumé

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

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

1 Introduction

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

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

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

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

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

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

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

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

2 Notation et critère objectif

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\sum_{i \in U} u_i v_i = N \bar{u}\,\bar{v}\,(1 + C_{u,v}). \tag{2.5}$$

3 Plan optimal avec paramètres de coût et de variance connus

3.1 Plan assisté par modèle optimal

Les valeurs de $(\pi_i : i \in U)$ qui minimisent (2.3) sous la contrainte (2.4) sont

$$\pi_i = C_f \frac{z_i^{1/2} c_i^{-1/2}}{\sum_{j \in U} z_j^{1/2} c_j^{1/2}} \propto z_i^{1/2} c_i^{-1/2} \tag{3.1}$$

et la variance anticipée résultante est

$$AV_{opt} = E_M var_p\left[\hat{t}_y\right] = \sigma^2 C_f^{-1} \left(\sum_{i \in U} c_i^{1/2} z_i^{1/2} \right)^2 - \sigma^2 \sum_{i \in U} z_i. \tag{3.2}$$

Cette expression peut être obtenue facilement en utilisant les multiplicateurs de Lagrange ou l'inégalité de Cauchy-Schwarz, et généralise Särndal et coll. (1992, résultat 12.2.1, p. 452) pour tenir compte des coûts inégaux. Une plus grande probabilité de sélection est attribuée aux unités dont la variance est plus élevée ou dont le coût est plus faible. Toutefois, dans (3.1), les racines carrées de z_i et c_i signifient que, dans de nombreuses enquêtes, les probabilités de sélection ne varient pas spectaculairement.

Dans le cas particulier de l'échantillonnage stratifié où $c_i = \bar{c}_h$ et $z_i = \bar{z}_h$ pour les unités i dans la strate h, (3.1) devient la répartition stratifiée optimale habituelle avec $\pi_i \propto \sqrt{\bar{z}_h / \bar{c}_h}$, de sorte que $n_h \propto N_h \sqrt{\bar{z}_h / \bar{c}_h}$.

Nous supposons que le dernier terme de (3.2), qui représente la correction pour population finie, est négligeable. L'application de (2.5) donne :

$$AV_{opt} \approx \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}\left(1+C_{\sqrt{c},\sqrt{z}}\right)^2}{\left(1+C_{\sqrt{c}}^2\right)\left(1+C_{\sqrt{z}}^2\right)} \tag{3.3}$$

où $C_{\sqrt{c}}$ et $C_{\sqrt{z}}$ désignent les coefficients de variation de population de $\sqrt{c_i}$ et $\sqrt{z_i}$, respectivement. Pour que nos résultats puissent être interprétés, nous supposons que les coûts par unité c_i et les variances σz_i ne sont pas reliés, de sorte que $C_{\sqrt{c},\sqrt{z}} = 0$. Cette hypothèse n'est pas toujours satisfaite en pratique, mais toute relation entre c_i et z_i sera propre à un échantillon particulier et pourrait être positive ou négative. Afin de dégager des principes généraux, il est logique d'ignorer ce genre de relation. En pratique, il est souvent raisonnable de supposer également que $C_{\sqrt{c}}$ et $C_{\sqrt{z}}$ sont faibles. Un développement en série de Taylor montre alors que $C_c^2 \approx 4C_{\sqrt{c}}^2$ et $C_z^2 \approx 4C_{\sqrt{z}}^2$. En regroupant ces approximations, (3.3) devient

$$AV_{opt} = \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}}{\left(1+\frac{1}{4}C_c^2\right)\left(1+\frac{1}{4}C_z^2\right)}. \tag{3.4}$$

Voir l'annexe pour les détails de ces calculs.

Ignorer les coûts

Si l'on fait abstraction des coûts, alors (3.1) fait penser que $\pi_i \propto z_i^{1/2}$. Pour faire des comparaisons pour le même coût prévu, C_f,

$$\pi_i = C_f \frac{z_i^{1/2}}{\sum_{j\in U} z_j^{1/2} c_j} \tag{3.5}$$

et la variance anticipée résultante est

$$AV_{nocosts} = \sigma^2 C_f^{-1}\left(\sum_{i\in U} z_i^{1/2}\right)\left(\sum_{i\in U} c_i z_i^{1/2}\right) - \sigma^2 \sum_{i\in U} z_i. \tag{3.6}$$

En effectuant des calculs similaires à ceux utilisés à la section 3.1, nous obtenons

$$AV_{nocosts} \approx \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}}{\left(1+\frac{1}{4}C_z^2\right)}. \tag{3.7}$$

Voir l'annexe pour des renseignements détaillés. La comparaison de (3.7) et de (3.4) montre que tenir compte des coûts dans le plan de sondage donne lieu à la division de la variance anticipée par $\left(1+(1/4)C_c^2\right)$.

4 Effet de l'utilisation de paramètres de coût estimés

En pratique, les coûts c_i ne sont pas connus précisément. Supposons qu'ils sont remplacés par les estimations $\hat{c}_i = b_i c_i$. L'utilisation de la variable auxiliaire et des coûts estimés dans les probabilités optimales implique que $\pi_i \propto z_i^{1/2}\hat{c}_i^{-1/2}$. Pour faire des comparaisons pour les mêmes coûts prévus,

$$\pi_i = C_f \frac{z_i^{1/2}\hat{c}_i^{-1/2}}{\sum_{j\in U} z_j^{1/2}\hat{c}_j^{-1/2}c_j}.$$

La variance anticipée résultante est

$$AV_{ests} = \sigma^2 C_f^{-1}\left(\sum_{i\in U}\hat{c}_i^{1/2}z_i^{1/2}\right)\left(\sum_{j\in U} z_j^{1/2}\hat{c}_j^{-1/2}c_j\right) - \sigma^2\sum_{i\in U} z_i. \tag{4.1}$$

Si nous supposons que les valeurs de b_i ne sont pas reliées aux valeurs de c_i et z_i, alors

$$AV_{ests} = \sigma^2 C_f^{-1}\left(\sum_{i\in U} c_i^{1/2}z_i^{1/2}\right)^2 N^{-2}\left(\sum_{i\in U} b_i^{-1/2}\right)\left(\sum_{i\in U} b_i^{1/2}\right) - \sigma^2\sum_{i\in U} z_i. \tag{4.2}$$

Voir l'annexe pour des renseignements détaillés. Si le coefficient de variation de b_i est faible, l'approximation par développement en série de Taylor donne $N^{-2}\sum b_i^{-1/2}\sum b_i^{1/2} \approx 1+(1/4)C_b^2$. En appliquant cela, ainsi que les mêmes approximations qu'à la sous-section 3.1, (4.2) devient

$$AV_{ests} = \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}\left(1+\frac{1}{4}C_b^2\right)}{\left(1+\frac{1}{4}C_c^2\right)\left(1+\frac{1}{4}C_z^2\right)}. \tag{4.3}$$

Voir l'annexe pour des renseignements détaillés.

La comparaison de (4.3) et (3.7) montre qu'utiliser des paramètres de coûts estimés au lieu de faire totalement abstraction des coûts a pour effet de multiplier la variance anticipée par $\left[1+(1/4)C_b^2\right]/\left[1+(1/4)C_c^2\right]$. Par conséquent, l'utilisation de l'information sur les coûts est valable à condition que $C_b < C_c$. Le coefficient de variation des facteurs d'erreur doit être plus faible que celui des coûts par unité réels sur l'ensemble de la population.

5 Exemples de modèles de coût

Les principales quantités qui déterminent l'utilité des données sur les coûts par unité sont C_b et C_c. Les plans optimaux établis en utilisant l'information sur l'inégalité des coûts n'étant pas très fréquents, la littérature sur les valeurs types de ces mesures est peu abondante. Divers facteurs peuvent donner lieu à des coûts inégaux, y compris les effets de mode de collecte, la géographie et la propension à répondre, et les publications traitant de ces questions donnent une vague idée des modèles de coût qui peuvent être appliqués en pratique.

L'utilisation d'un mode mixte d'interview est l'une des raisons pour lesquelles les coûts par unité peuvent être inégaux. Différents modes de collecte, comme l'interview sur place ou par téléphone assistée par ordinateur, l'envoi de questionnaires par la poste ou l'utilisation de questionnaires en ligne, ou l'interview en face à face, peuvent être utilisés pour obtenir les réponses auprès de différents répondants (Dillman, Smyth et Christian 2009). Cette approche peut être adoptée pour réduire les coûts ou pour améliorer le taux de réponse, mais il faut veiller à ce qu'elle n'introduise pas de biais dû aux effets de mode. Les effets de mode peuvent comprendre des effets de sélection (qui ne posent généralement pas problème) et des effets de mesure (qui entraînement habituellement un biais), et ces deux types d'effets sont souvent difficiles à isoler l'un de l'autre (Vannieuwenhuyze, Loosveldt et Molenbergs 2012). Les économies résultant de l'utilisation de modes mixtes pourraient éventuellement être amplifiées en incorporant les coûts selon le mode de collecte dans le plan de sondage, comme il est décrit dans le présent article. Groves (1989, p. 538) compare les coûts par répondant de l'interview téléphonique (38,00 $) et de l'interview sur place (84,90 $) de la population générale. Si l'on connaissait la préférence de toutes les unités figurant dans une base de sondage et que chaque mode était préféré par une moitié d'entre elle, cela impliquerait que $C_c = 0,38.$ Greenlaw et Brown-Welty (2009) ont comparé l'utilisation de questionnaires papier et de questionnaires en ligne, et ont constaté des coûts par répondant de 4,78 $ et de 0,64 $, respectivement, dans le cadre d'une enquête auprès des membres d'une association professionnelle. Dans le cas d'une option à mode mixte, les deux tiers des répondants ont opté pour l'option de réponse en ligne. Si l'on connaît d'avance les préférences, alors $C_c = 0,76.$

Une autre raison de l'existence de coûts variables est que certains répondants sont plus difficiles à recruter que d'autres, et nécessitent un plus grand nombre de visites ou de rappels. Groves et Heeringa (2006, section 2.2) ont réalisé un essai d'enquête où les intervieweurs classaient les non-répondants au moment du premier contact comme étant susceptibles ou non de répondre. Lors du suivi, le taux de réponse a été de 73,7 % pour le premier groupe comparativement à 38,5 % pour le deuxième. Cela donne à penser que le coût par répondant serait au moins 1,9 fois plus élevé pour le deuxième groupe que pour le premier. (En fait, le ratio serait plus élevé, parce qu'un plus grand nombre de tentatives de suivi serait faites pour le groupe difficile.). Si les deux groupes contiennent chacun 50 % des répondants, alors $C_c = 0,31.$

La géographie est une autre source de différences de coût dans les enquêtes avec intervieweur. Dans le cas de l'Enquête sur la population active de l'Australie, les coûts ont été modélisés en utilisant une composante par îlot et une composante par logement (Hicks 2001, tableau 4.2.1 à la section 4.2) selon le type de région (15 types ont été définis). En supposant un échantillonnage constant de 10 logements par îlot, le coût par logement net varie de 4,98 $ dans le centre-ville de Sydney et de Melbourne à 6,71 $ dans les régions peu peuplées et autochtones. Bien qu'il s'agisse d'une différence de coût significative entre les types de régions, la grande majorité de la population se retrouve dans trois de ces types (zone habitée, croissance extérieure et grande ville) où les coûts par logement ne varient qu'entre 5,71 $ et 6,07 $. Par contre l'estimation de C_c est très faible, soit 0,054.

Le tableau 5.1 montre l'amélioration en pourcentage approximative de la variance anticipée lorsqu'on utilise l'information estimée sur les coûts pour différentes valeurs de C_c et C_b, certaines d'entre elles

étant suggérées par les exemples susmentionnés. Les valeurs négatives indiquent que le plan de sondage est moins efficace que si l'on fait complètement abstraction des coûts. Le tableau donne à penser que l'utilisation de l'information sur les coûts ne vaut la peine qu'en cas de variation respectable des coûts par unité; sinon, l'avantage est très faible et peut disparaître en cas d'une imprécision, même faible, des coûts estimés. Les enquêtes à mode de collecte mixte sont celles offrant le plus de possibilités d'exploitation des coûts par unité variables dans le plan de sondage, mais le risque d'un biais de mesure doit être évalué méticuleusement, en utilisant des méthodes telles que celles décrites dans Vannieuwenhuyze, Loosveldt et Molenberghs (2010), Vannieuwenhuyze et coll. (2012), Vannieuwenhuyze et Loosveldt (2013) et Schouten, Brakel, Buelens, Laan et Klaus (2013). Il serait peut-être même possible d'incorporer les effets de mode (ou l'incertitude au sujet des effets de mode) dans le plan optimal au moyen du modèle de variance, une approche qui pourrait être le sujet de futures études. Les constatations faites dans la présente étude font penser qu'une telle approche mérite d'être envisagée.

Tableau 5.1
Amélioration en pourcentage de la variance anticipée découlant de l'utilisation de l'information estimée sur les coûts comparativement à l'absence d'information sur les coûts.

Coefficient de variation des coûts par unité (C_c) (%)		Coefficient de variation du facteur d'erreur (C_b) (%)			
	Scénario possible	0	10	25	50
5		0,1	-0,2	-1,5	-6,2
10	Déplacement de l'intervieweur dû à l'éloignement	0,2	0,0	-1,3	-6,0
20		1,0	0,7	-0,6	-5,2
30	Propension à répondre	2,2	2,0	0,7	-3,9
40	Mode mixte (interview par téléphone/sur place)	3,8	3,6	2,3	-2,2
50		5,9	5,6	4,4	0,0
75	Mode mixte (papier/en ligne autoadministré)	12,3	12,1	11,0	6,8

6 Discussion

L'utilisation de coûts par unité inégaux peut améliorer l'efficacité des plans de sondage. Pour que les gains d'efficacité soit appréciables, les coûts par unité doivent varier considérablement. Même en l'absence d'erreur d'estimation, un coefficient de variation de 50 % peut n'entraîner qu'une amélioration de 6 % de la variance anticipée. Si ce coefficient de variation est de 75 %, comme cela peut se produire dans une enquête à mode de collecte mixte, la réduction de la variance anticipée (ou de la taille de l'échantillon pour une précision fixe) peut être supérieure à 12 %. Les coûts sont estimés avec une certaine erreur, ce qui réduit l'amélioration d'un facteur déterminé par la variation relative des erreurs relatives d'estimation des coûts au niveau individuel.

Annexe

A.1 Calculs détaillés

Lemme 1 : Soit u_i défini pour $i \in U$. Soit $u_i = \bar{u} + \theta e_i$, où $\sum_{i\in U} e_i = 0$ et θ est petit. Alors :

a. $\overline{\sqrt{u}} = \sqrt{\bar{u}} - \frac{1}{8}\theta^2 \bar{u}^{-3/2} S_e^2 + o\left(\theta^2\right)$.

b. $S_{\sqrt{u}}^2 = \frac{1}{4}\theta^2 \bar{u}^{-1} S_e^2 + o\left(\theta^2\right) = \frac{1}{4}\bar{u}^{-1} S_u^2 + o\left(\theta^2\right)$.

c. $N^{-2}\left(\sum_{i\in U} u_i^{1/2}\right)\left(\sum_{i\in U} u_i^{-1/2}\right) = 1 + \frac{1}{4}\theta^2 \bar{u}^{-2} S_e^2 + o\left(\theta^2\right) = 1 + \frac{1}{4} C_u^2 + o\left(\theta^2\right)$.

d. $C_{\sqrt{u}}^2 = \frac{1}{4}\theta^2 \bar{u}^{-2} S_e^2 + o\left(\theta^2\right) = \frac{1}{4} C_u^2 + o\left(\theta^2\right)$.

La notation $o\left(C_u^2\right)$ peut être utilisée à la place de $o\left(\theta^2\right)$, puisque $C_u^2 = \theta^2 C_e^2$, ce qui est fait dans la suite de l'annexe.

Preuve :

Nous commençons par écrire $\overline{\sqrt{u}}$ comme une fonction de θ :

$$\overline{\sqrt{u}} = N^{-1}\sum_{i\in U}\sqrt{u_i} = N^{-1}\sum_{i\in U}\sqrt{\bar{u} + \theta e_i}.$$

Si nous appelons cette fonction $g\left(\theta\right)$, la différenciation au point $\theta = 0$ donne $g(0) = \sqrt{\bar{u}}$, $g'(0) = 0$ et

$$g''(0) = -\frac{1}{4} N^{-1} \bar{u}^{-3/2} \sum_{i\in U} e_i^2 = -\frac{1}{4}\bar{u}^{-3/2} S_e^2.$$

D'où

$$\overline{\sqrt{u}} = g\left(\theta\right) = g(0) + g'(0)\theta + \frac{1}{2} g''(0)\theta^2 + o\left(\theta^2\right) = \sqrt{\bar{u}} - \frac{1}{8}\theta^2 \bar{u}^{-3/2} S_e^2 + o\left(\theta^2\right)$$

qui est le résultat a.

Le résultat b est prouvé en utilisant le résultat a :

$$\begin{aligned}
S_{\sqrt{u}}^2 &= N^{-1}\sum_{i\in U}\left(\sqrt{u_i}\right)^2 - \left(N^{-1}\sum_{i\in U}\sqrt{u_i}\right)^2 \\
&= \bar{u} - \left(\overline{\sqrt{u}}\right)^2 \\
&= \bar{u} - \left(\sqrt{\bar{u}} - \frac{1}{8}\theta^2 \bar{u}^{-3/2} S_e^2 + o\left(\theta^2\right)\right)^2 \\
&= \bar{u} - \left(\bar{u} + \frac{1}{64}\theta^4 \bar{u}^{-3} S_e^4 - \frac{1}{4}\theta^2 \bar{u}^{-1} S_e^2 + o\left(\theta^2\right)\right) \\
&= \frac{1}{4}\theta^2 \bar{u}^{-1} S_e^2 + o\left(\theta^2\right) = \frac{1}{4}\bar{u}^{-1} S_u^2 + o\left(\theta^2\right).
\end{aligned}$$

Pour obtenir c, nous commençons par écrire $N^{-1}\sum_{i\in U} u_i^{-1/2}$ comme une fonction $g()$ de θ et effectuons un développement en série de Taylor :

$$\begin{aligned} N^{-1}\sum_{i\in U} u_i^{-1/2} &= N^{-1}\sum_{i\in U}\left(\overline{u}+\theta e_i\right)^{-1/2} \\ &= g\left(\theta\right) = g(0)+g'(0)\theta+\frac{1}{2}g''(0)\theta^2+o\left(\theta^2\right) \\ &= \overline{u}^{-1/2}+0\theta+\frac{1}{2}\frac{3}{4}\overline{u}^{-5/2}N^{-1}\sum_{i\in U} e_i^2\theta^2+o\left(\theta^2\right) \\ &= \overline{u}^{-1/2}+\frac{3}{8}\overline{u}^{-5/2}S_e^2\theta^2+o\left(\theta^2\right). \end{aligned} \tag{A.1}$$

Notons que $N^{-1}\sum_{i\in U} u_i^{1/2} = \overline{\sqrt{u}}$. La multiplication de l'expression pour $\overline{\sqrt{u}}$ dans le résultat a par (A.1) donne

$$\begin{aligned} N^{-2}\left(\sum_{i\in U} u_i^{1/2}\right)\left(\sum_{i\in U} u_i^{-1/2}\right) &= \left\{\sqrt{\overline{u}}-\frac{1}{8}\theta^2\overline{u}^{-3/2}S_e^2+o\left(\theta^2\right)\right\}\left\{\overline{u}^{-1/2}+\frac{3}{8}\overline{u}^{-5/2}S_e^2\theta^2+o\left(\theta^2\right)\right\} \\ &= 1+\frac{1}{4}\overline{u}^{-2}S_e^2\theta^2+o\left(\theta^2\right) \\ &= 1+\frac{1}{4}C_u^2+o\left(\theta^2\right) \end{aligned}$$

qui est le résultat c.

Pour le résultat d, commençons par noter que $\overline{\sqrt{u}} = \sqrt{\overline{u}}+o\left(\theta\right)$ d'après le résultat a, et donc, un développement en série de Taylor d'ordre 1 donne

$$\left(\overline{\sqrt{u}}\right)^{-2} = \left(\sqrt{\overline{u}}\right)^{-2}+o\left(\theta\right) = \overline{u}^{-1}+o\left(\theta\right).$$

En combinant cela avec le résultat b, nous obtenons

$$\begin{aligned} C_{\sqrt{u}}^2 &= S_{\sqrt{u}}^2\left(\overline{\sqrt{u}}\right)^{-2} \\ &= \left\{\frac{1}{4}\theta^2\overline{u}^{-1}S_e^2+o\left(\theta^2\right)\right\}\left\{\overline{u}^{-1}+o\left(\theta\right)\right\} \\ &= \frac{1}{4}\theta^2\overline{u}^{-2}S_e^2+o\left(\theta^2\right) \\ &= \frac{1}{4}C_u^2+o\left(\theta^2\right) \end{aligned}$$

qui donne le résultat d.

Obtention de (3.3)

Pour le cas particulier où $u_i = v_i$, (2.5) devient

$$\sum_{i\in U} u_i^2 = N\overline{u}^2\left(1+C_u^2\right). \tag{A.2}$$

En appliquant (2.5), nous obtenons

$$\sum_{i\in U} c_i^{1/2} z_i^{1/2} = N\overline{\sqrt{c}}\ \overline{\sqrt{z}}\left(1+C_{\sqrt{c},\sqrt{z}}\right) \tag{A.3}$$

où $\overline{\sqrt{c}} = N^{-1}\sum_{i\in U}\sqrt{c_i}$ et $\overline{\sqrt{z}} = N^{-1}\sum_{i\in U}\sqrt{z_i}$. En utilisant (A.2), nous pouvons exprimer $\overline{\sqrt{c}}$ en fonction de $\overline{c}$:

$$\overline{c} = N^{-1}\sum_{i\in U} c_i = N^{-1}\sum_{i\in U}\left(\sqrt{c_i}\right)^2 = \left(\overline{\sqrt{c}}\right)^2\left(1+C_{\sqrt{c}}^2\right). \tag{A.4}$$

De même,

$$\overline{z} = \left(\overline{\sqrt{z}}\right)^2\left(1+C_{\sqrt{z}}^2\right). \tag{A.5}$$

En supposant que le dernier terme de (3.2) est négligeable, l'application de (A.3), (A.4) et (A.5) donne (3.3).

Obtention de (3.4)

Le lemme 1d implique que $C_{\sqrt{c}}^2 = (1/4)C_c^2 + o\left(C_c^2\right) \approx (1/4)C_c^2$ et $C_{\sqrt{z}}^2 = (1/4)C_z^2 + o\left(C_z^2\right) \approx (1/4)C_z^2$. Le résultat (3.4) découle de (3.3) en utilisant ces approximations, et en supposant que $C_{\sqrt{c},\sqrt{z}} = 0$.

Obtention de (3.7)

Premièrement, $\sum_{i\in U} c_i z_i^{1/2} = N\overline{c}\,\overline{\sqrt{z}}\left(1+C_{c,\sqrt{z}}\right)$, d'après (2.5), où $C_{c,\sqrt{z}}$ est la covariance relative dans la population entre les valeurs de $z_i^{1/2}$ et c_i. Nous supposons que les valeurs de c_i et z_i ne sont pas liées, de sorte que $C_{c,\sqrt{z}} = 0$. Nous supposons aussi que le deuxième terme de (3.6) est négligeable, ce qui correspond à une faible fraction d'échantillonnage. Donc, (3.6) devient :

$$AV_{nocosts} = \sigma^2 N^2 C_f^{-1}\overline{c}\left(\overline{\sqrt{z}}\right)^2. \tag{A.6}$$

Partant de (A.5) et du lemme 1d, nous obtenons

$$\left(\overline{\sqrt{z}}\right)^2 = \frac{\overline{z}}{1+C_{\sqrt{z}}^2} \approx \frac{\overline{z}}{1+(1/4)C_z^2}.$$

La substitution dans (A.6) donne (3.7).

Obtention de (4.2)

Deux termes dans (4.1) se simplifient en utilisant (2.5). Premièrement,

$$\begin{aligned}\sum_{i\in U}\hat{c}_i^{1/2}z_i^{1/2} &= \sum_{i\in U}b_i^{1/2}c_i^{1/2}z_i^{1/2}\\ &= N\left(N^{-1}\sum_{i\in U}b_i^{1/2}\right)\left(N^{-1}\sum_{i\in U}c_i^{1/2}z_i^{1/2}\right)+C_{\sqrt{b},\sqrt{cz}}\end{aligned} \tag{A.7}$$

où $C_{\sqrt{b},\sqrt{cz}}$ est la covariance entre les valeurs de population de $b_i^{1/2}$ et $c_i^{1/2}z_i^{1/2}$. Deuxièmement,

$$\begin{aligned}\sum_{i\in U}z_i^{1/2}\hat{c}_i^{-1/2}c_i &= \sum_{i\in U}b_i^{-1/2}c_i^{1/2}z_i^{1/2}\\ &= N\left(N^{-1}\sum_{i\in U}b_i^{-1/2}\right)\left(N^{-1}\sum_{i\in U}c_i^{1/2}z_i^{1/2}\right)+C_{1/\sqrt{b},\sqrt{cz}}\end{aligned} \tag{A.8}$$

où $C_{1/\sqrt{b},\sqrt{cz}}$ est la covariance entre les valeurs de population de $b_i^{-1/2}$ et $c_i^{1/2}z_i^{1/2}$.

Si nous supposons que les valeurs de population de b_i ne sont pas reliées aux valeurs de c_i et z_i, de sorte que $C_{\sqrt{b},\sqrt{cz}} = C_{1/\sqrt{b},\sqrt{cz}} = 0,$ et que nous introduisons (A.7) et (A.8) par substitution dans (4.1), alors nous obtenons (4.2).

Obtention de (4.3)

Nous pouvons exprimer (4.2) en fonction de AV_{opt} qui est défini dans (3.2), en supposant que le dernier terme de (3.2) est négligeable, ce qui correspond à une faible fraction d'échantillonnage :

$$AV_{ests} \approx AV_{opt}N^{-2}\sum_{i\in U}b_i^{-1/2}\sum_{i\in U}b_i^{1/2} \tag{A.9}$$

Le lemme 1c implique que

$$N^{-2}\sum_{i\in U}b_i^{-1/2}\sum b_i^{1/2} = 1+\frac{1}{4}C_b^2+o\left(C_b^2\right)\approx 1+\frac{1}{4}C_b^2.$$

L'introduction de cette expression et de (3.3) par substitution dans (A.9) donne (4.3).

Bibliographie

Brewer, K. et Gregoire, T.G. (2009). Introduction to survey sampling. Dans *Handbook of Statistics 29A: Sample Surveys: Design, Methods and Applications*, eds. Pfeffermann, D. and Rao, C.R., Amsterdam: Elsevier/North-Holland, pp. 9-37.

Clark, R.G. (2009). Sampling of subpopulations in two-stage surveys. *Statistics in Medicine*, 28, 3697-3717.

Clark, R.G. et Steel, D.G. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 63-82.

Cochran, W. (1977). *Sampling Techniques*. New York: Wiley, 3rd ed.

Dillman, D., Smyth, J. et Christian, L. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* John Wiley and Sons, 3rd ed.

Greenlaw, C. et Brown-Welty, S. (2009). A comparison of web-based and paper-based survey methods testing assumptions of survey mode and response cost. *Evaluation Review*, 33, 464-480.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

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

Hansen, M., Hurwitz, W. et Madow, W. (1953). *Sample Survey Methods and Theory Volume 1: Methods and Applications*. New York: John Wiley and Sons.

Hicks, K. (2001). Cost and variance modelling for the 2001 redesign of the Monthly Population Survey. www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.037, *Australian Bureau of Statistics Methodology Advisory Committee Paper*.

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

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

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. et Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, 80, 382-399.

Schouten, B., Brakel, J.v.d., Buelens, B., Laan, J.v.d. et Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35(1), 107-121.

Schouten, B., Shlomo, N. et Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 231-253.

Vannieuwenhuyze, J.T. et Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: Three methods to disentangle selection and measurement effects. *Sociological Methods and Research*, 42, 82-104.

Vannieuwenhuyze, J.T., Loosveldt, G. et Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74, 1027-1045.

Vannieuwenhuyze, J. T., Loosveldt, G., et Molenberghs, G. (2012). A method to evaluate mode effects on the mean and variance of a continuous variable in mixed-mode surveys. *International Statistical Review*, 80, 306-322.