UNE PROCÉDURE ADAPTATIVE POUR LA PRIORISATION DES APPELS TÉLÉPHONIQUES LORS DE LA COLLECTE DES DONNÉES

Jean-François Beaumont[1], Cynthia Bocci[2] & David Haziza[3]

[1] *Statistique Canada, Immeuble R.-H.-Coats, 16e étage, 100 promenade Tunney's Pasture, Canada, K1A 0T6 (jean-francois.beaumont@statcan.gc.ca)*
[2] *Statistique Canada, Immeuble R.-H.-Coats, 18e étage, 100 promenade Tunney's Pasture, Canada, K1A 0T6 (cynthia.bocci@statcan.gc.ca)*
[3] *Université de Montréal, Département de mathématiques et statistiques, Montréal, Canada, H3C 3J7, (David.Haziza@umontreal.ca)*

Résumé. Nous proposons une procédure adaptative pour la priorisation des appels dans le contexte d'enquêtes téléphoniques dont la collecte des données est assistée par ordinateurs. Notre procédure est adaptative au sens où l'effort assigné à une unité échantillonnée peut varier d'une unité à l'autre et peut aussi varier au cours de la collecte. Le but d'utiliser une procédure adaptative est habituellement d'accroître la qualité des estimations pour un coût de collecte donné ou, alternativement, de réduire le coût pour une qualité donnée. Un critère de qualité souvent considéré dans la littérature est le biais d'un estimateur qui n'est pas ajusté pour la non-réponse. Bien que la réduction du biais dû à la non-réponse soit un objectif souhaitable, nous soutenons que ce n'est pas un critère utile à utiliser lors de la collecte des données parce que le biais qu'on peut enlever à cette étape au moyen d'une procédure adaptative de collecte peut aussi être enlevé à l'étape d'estimation au moyen d'ajustements de poids. Par conséquent, nous proposons plutôt une procédure de priorisation des appels qui, étant donné un échantillon donné, tente de minimiser la variance d'un estimateur ajusté pour la non-réponse sous une contrainte de budget de collecte. Beaumont, Bocci et Haziza (2014) ont montré l'efficacité de cette procédure dans une étude par simulations.

Mots-clés. Ajustement des poids pour la non-réponse, biais dû à la non-réponse, paradonnées, plan de collecte adaptatif et variance due à la non-réponse.

1 Introduction

Dans cet article, nous nous concentrons sur le problème de la priorisation des appels dans le contexte d'enquêtes téléphoniques dont la collecte des données est assistée par ordinateurs. Nous proposons une procédure adaptative de collecte des données qui tente de maximiser la qualité des estimations étant donné un budget fixe. Notre procédure est adaptative au sens où l'effort assigné à une unité échantillonnée peut varier d'une unité à l'autre et peut aussi varier au cours de la collecte des données. Dans ce dernier cas, l'utilisation de paradonnées est requise. Les paradonnées sont des données au sujet du processus de collecte des données telles que des taux de réponse à différents moments de la collecte pour différents sous-groupes de l'échantillon.

Dans la littérature, on considère souvent comme critère de qualité le biais d'un estimateur qui n'est pas ajusté pour la non-réponse. Bien que la réduction du biais dû à la non-réponse soit un objectif souhaitable, nous croyons que ce n'est pas un critère utile à utiliser lors de la collecte des données parce que le biais qu'on peut enlever à cette étape au moyen d'une procédure adaptative de collecte peut aussi être enlevé à l'étape d'estimation au moyen d'ajustements de poids. Par exemple, on pourrait considérer une procédure de collecte qui priorise les appels de telle sorte que les taux de réponse pour différents domaines d'intérêt sont similaires et ensuite utiliser un estimateur qui n'est

pas ajusté pour la non-réponse. On s'attend à ce que le biais dû à la non-réponse de cette stratégie soit équivalent à celui d'une stratégie dans laquelle les appels sont effectués complètement aléatoirement mais où on utilise un estimateur qui ajuste les poids de sondage par l'inverse des taux de réponse à l'intérieur des domaines d'intérêt. Cela peut s'expliquer par le fait que la réduction du biais dû à la non-réponse nécessite de l'information auxiliaire disponible autant pour les répondants que les non-répondants (information sur le domaine d'intérêt dans l'exemple ci-dessus). Que cette information soit utilisée à l'étape de collecte des données ou non ne devrait pas faire de différence en termes de biais en autant que cette information soit utilisée à l'étape d'estimation. Beaumont, Haziza et Bocci (2014) confirment ce point. Par conséquent, le critère de qualité que nous suggérons de minimiser est la variance d'un estimateur ajusté pour la non-réponse conditionnellement à l'échantillon sélectionné.

2 Mise en contexte

Soit y_i, la valeur de la variable y pour l'unité i de la population U de taille N et $\theta = \sum_{i \in U} y_i$, le total de la population qu'on veut estimer. Soit s, un échantillon de taille n tiré de U selon un certain plan de sondage $p(s)$. Soit s_r, l'ensemble des répondants de taille n_r observé à la fin de la collecte des données et généré selon un certain mécanisme de non-réponse $q(s_r | s)$. On note les cellules d'ajustement de non-réponse par l'indice g, $g = 1,...,G$, où G est le nombre de cellules. Chaque unité de l'échantillon appartient à une et une seule cellule et on suppose qu'on connaît avant le début de l'enquête à quelle cellule chaque unité appartient. Par exemple, les cellules pourraient être formées à partir de domaines d'intérêt importants. Soit s_g, l'ensemble des n_g unités échantillonnées tombant dans la cellule g et s_{rg}, l'ensemble des n_{rg} répondants tombant dans la cellule g à la fin de la collecte des données. On suppose que l'estimateur par dilatation serait utilisé pour estimer θ s'il n'y avait pas de non-réponse. En utilisant notre notation, on peut écrire l'estimateur par dilatation comme suit : $\hat{\theta} = \sum_{g=1}^{G} \sum_{i \in s_g} w_{gi} y_{gi}$, où y_{gi} est la valeur de la variable y pour l'unité i de la cellule g, $w_{gi} = 1/\pi_{gi}$ est son poids de sondage et $\pi_{gi} = \Pr(i \in s_g)$ est sa probabilité de sélection dans l'échantillon. L'estimateur par dilatation $\hat{\theta}$ est sans biais sous le plan pour θ; c'est-à-dire que $E_p(\hat{\theta}) = \theta$. L'indice p indique que l'espérance est évaluée par rapport au plan de sondage.

On note par $\rho_{gi} = \Pr(i \in s_{rg} | s, i \in s_g)$, la probabilité que l'unité échantillonnée i de la cellule g soit un répondant à l'enquête à la fin de la période de collecte des données. Cette probabilité dépend, entre autres, des procédures de collecte et, plus spécifiquement, des ressources utilisées pour obtenir une réponse de l'unité i, $i \in s_g$. Dans ce qui suit, on suppose que la non-réponse est uniforme à l'intérieur des cellules. En d'autres mots, on suppose que toutes les unités de l'échantillon répondent indépendamment les unes des autres et que toutes les unités échantillonnées dans la cellule g ont la même probabilité de répondre à l'enquête; c'est-à-dire que $\rho_{gi} = \rho_g$, pour toute unité $i \in s_g$. Cette hypothèse est fréquente dans la littérature sur la non-réponse dans les enquêtes. En pratique, les probabilités de réponse ρ_g, $g = 1,...,G$, sont inconnues mais elles peuvent être estimées par les taux de réponse $\hat{\rho}_g = n_{rg}/n_g$. Puisque $E_q(\hat{\rho}_g | s) = \rho_g$, la probabilité de réponse ρ_g peut être interprétée comme le taux de réponse espéré dans la cellule g. L'indice q indique que l'espérance est évaluée par rapport au mécanisme de non-réponse. Un estimateur de θ qui est ajusté pour la non-réponse et fréquemment utilisé est :

$$\hat{\theta}_A = \sum_{g=1}^{G} \sum_{i \in s_{rg}} \frac{w_{gi}}{\hat{\rho}_g} y_{gi} \,. \tag{1}$$

Sous certaines conditions, incluant une grande taille d'échantillon et une non-réponse uniforme à l'intérieur des cellules, le carré du biais dû à la non-réponse de $\hat{\theta}_A$, $\left\{E_q(\hat{\theta}_A - \hat{\theta} \middle| s)\right\}^2$, est petit en comparaison à sa variance, $\text{var}_q(\hat{\theta}_A \middle| s)$. Nous faisons cette hypothèse et considérons la variance due à la non-réponse de l'estimateur ajusté (1), $\text{var}_q(\hat{\theta}_A \middle| s)$, comme étant notre indicateur de qualité. Au moyen d'un développement en séries de Taylor au premier degré, cette variance est à peu près égale à :

$$\text{var}_q(\hat{\theta}_A \middle| s) \cong \sum_{g=1}^{G} (\rho_g^{-1} - 1)(n_g - 1) S_{wy,g}^2 \,, \tag{2}$$

où $S_{wy,g}^2 = \frac{1}{n_g - 1} \sum_{i \in s_g} (w_{gi} y_{gi} - \hat{\mu}_g)^2$ et $\hat{\mu}_g = \frac{1}{n_g} \sum_{i \in s_g} w_{gi} y_{gi}$.

L'estimateur ajusté (1) a un biais négligeable si la taille d'échantillon est grande et si l'hypothèse de non-réponse uniforme à l'intérieur des cellules est vérifiée. Si cette dernière hypothèse n'est pas valide, l'estimateur (1) peut devenir significativement biaisé. Cependant, ce biais de non-réponse ne peut être éliminé à l'étape de collecte des données si aucune information supplémentaire sur les non-répondants n'est utilisée (voir l'étude empirique de Beaumont, Bocci et Haziza, 2014). C'est la raison pour laquelle nous ignorons le biais dû à la non-réponse et nous nous concentrons sur la minimisation de la variance due à la non-réponse.

3 Le coût total et son espérance

Nous supposons que le coût total de la collecte dépend seulement des quantités $C_{NR,g}$, $C_{R,g}$ et m_{gi} qui sont, respectivement, le coût d'une tentative d'appel infructueuse dans la cellule *g*, le coût d'une entrevue dans la cellule *g* et le nombre total de tentatives d'appels à la fin de la période de collecte des données pour l'unité *i* de la cellule *g*. Le coût total peut donc s'exprimer ainsi :

$$C_{TOT} = \sum_{g=1}^{G} C_{TOT,g} \,,$$

où

$$C_{TOT,g} = \sum_{i \in s_{rg}} \left[(m_{gi} - 1) C_{NR,g} + C_{R,g} \right] + \sum_{i \in s_g - s_{rg}} m_{gi} C_{NR,g} \,.$$

Le coût total espéré s'écrit:

$$\tilde{C}_{TOT} = E_q(C_{TOT} \middle| s) = \sum_{g=1}^{G} \tilde{C}_{TOT,g} \,,$$

où

$$\tilde{C}_{TOT,g} = E_q(C_{TOT,g} \middle| s) = (C_{R,g} - C_{NR,g}) n_g \rho_g + C_{NR,g} \sum_{i \in s_g} \tilde{m}_{gi}$$

et $\tilde{m}_{gi} = E_q(m_{gi} \middle| s)$ est le nombre espéré de tentatives d'appels à la fin de la période de collecte des données pour l'unité *i* de la cellule *g*. Supposons que les procédures de collecte sont telles que le nombre total de tentatives d'appels pour l'unité *i* de la cellule *g* est contraint à ne pas être plus grand

qu'une certaine valeur fixée, M_{gi}. La limite M_{gi} peut varier d'une unité échantillonnée à l'autre bien qu'en pratique elle est souvent posée égale à une constante pour toutes les unités de l'échantillon. Le nombre espéré de tentatives d'appels, $\tilde{m}_{gi}$, dépend de la probabilité de répondre à chaque tentative pour l'unité *i* de la cellule *g*, notée par p_{gi}, et du nombre maximum de tentatives, M_{gi}. Le nombre espéré de tentatives d'appels dépend aussi de l'effort fourni pour obtenir une réponse de l'unité *i* de la cellule *g*, lequel est relié au budget total et aux procédures de collecte des données. La dérivation rigoureuse de $\tilde{m}_{gi} = E_q(m_{gi} \mid s)$ n'est pas facile. Pour la simplifier, nous faisons les trois hypothèses suivantes :

i) La probabilité de répondre lors d'une tentative d'appel, p_{gi}, est constante d'une tentative à l'autre.
ii) Pour n'importe quelle unité de l'échantillon, le fait de répondre ou non à l'enquête est indépendant d'une tentative à l'autre.
iii) À la fin de la collecte des données, chaque unité de l'échantillon est soit répondante ou a atteint le nombre maximum de tentatives, M_{gi}.

L'hypothèse (i) implique que la probabilité de répondre lors d'une tentative d'appel, p_{gi}, ne dépend pas de caractéristiques qui peuvent varier au fil du temps. L'hypothèse (ii) est plus réaliste si on impose un certain laps de temps entre deux tentatives successives. L'hypothèse (iii) signifie qu'une unité échantillonnée ne peut pas être non répondante à la fin de la collecte des données sans avoir atteint la limite M_{gi}. Cette hypothèse serait satisfaite si le budget total était suffisamment grand et qu'il n'y avait pas de refus. Une conséquence de cette hypothèse est que le nombre espéré de tentatives, $\tilde{m}_{gi}$, est seulement une fonction de p_{gi} et M_{gi}. Bien que nous reconnaissons que ces trois hypothèses puissent ne pas toujours être satisfaites en pratique, nous croyons qu'elles sont utiles pour obtenir une approximation de $E_q(m_{gi} \mid s)$. Cela est confirmé dans l'étude empirique de Beaumont, Bocci et Haziza (2014). En utilisant ces trois hypothèses, nous obtenons :

$$\begin{aligned} \tilde{m}_{gi} &= E_q(m_{gi} \mid s) \\ &= \left(\sum_{t=1}^{M_{gi}-1} t\, p_{gi}(1-p_{gi})^{t-1} \right) + M_{gi}(1-p_{gi})^{M_{gi}-1} \\ &= \frac{1}{p_{gi}}\left(1-(1-p_{gi})^{M_{gi}}\right). \end{aligned} \tag{3}$$

Le développement algébrique pour passer de la seconde à la troisième équation dans (3) est simple mais fastidieux. Il est donc omis.

Puisque $\tilde{m}_{gi}$ dans l'équation (3) n'est une fonction que de p_{gi} et M_{gi}, le coût total espéré peut s'écrire comme une fonction linéaire des taux de réponses espérés, ρ_g, $g = 1,...,G$:

$$\tilde{C}_{TOT} = \lambda_0 + \sum_{g=1}^{G} \lambda_{1g} \rho_g \tag{4}$$

avec $\lambda_0 = \sum_{g=1}^{G} C_{NR,g} \sum_{i \in s_g} \tilde{m}_{gi}$ et $\lambda_{1g} = (C_{R,g} - C_{NR,g}) n_g$.

4 Le problème d'optimisation et sa solution

Notre objectif consiste à déterminer les taux de réponse espérés cibles, ρ_{Tg}, $g = 1,...,G$, qui minimisent la variance (2), où on remplace ρ_g par ρ_{Tg}, tout en satisfaisant la contrainte budgétaire, $\lambda_0 + \sum_{g=1}^{G} \lambda_{1g} \rho_{Tg} = K$, pour une constant *K* qui représente le budget total. La solution est donnée par

$$\rho_{Tg} = \sqrt{\frac{(n_g - 1)S_{wy,g}^2}{\delta \lambda_{1g}}} = \sqrt{\left(\frac{n_g - 1}{n_g}\right) \frac{S_{wy,g}^2}{\delta (C_{R,g} - C_{NR,g})}}, \tag{5}$$

où

$$\delta = \frac{\left(\sum_g \sqrt{\lambda_{1g}(n_g - 1)S_{wy,g}^2}\right)^2}{(K - \lambda_0)^2} \ . \tag{6}$$

La solution (5) n'est en général pas équivalente à maximiser l'indicateur R proposé par Schouten, Cobben et Bethlehem (2009). L'indicateur R est maximisé quand les taux de réponse espérés cibles ρ_{Tg}, $g = 1,...,G$, sont tous égaux. Malheureusement, l'équation (5) n'assure pas que les taux de réponses espérés cibles sont inférieurs à 1. Si on obtient une valeur de ρ_{Tg} qui est supérieure à 1 alors on peut simplement la remplacer par une valeur légèrement inférieure à 1.

La variance due à la non-réponse minimum est obtenue au moyen de (2) lorsqu'on remplace ρ_g par ρ_{Tg}. Cette variance minimum est une fonction du budget total *K*. Il pourrait être utile de faire un graphique de la variance minimum en fonction du budget *K*. On pourrait trouver une valeur du budget au-delà de laquelle la variance minimum ne peut plus être réduite significativement. Il ne serait alors pas justifié de dépenser plus que cette valeur.

5 La procédure pour la sélection des cas

Une fois qu'on a déterminé les taux de réponses espérés cibles, ρ_{Tg}, il faut trouver l'effort requis pour atteindre ces cibles. Soit e_{gi}, l'effort maximum (en termes de nombre de tentative d'appels) associé à l'unité *i* dans la cellule *g*. Sous les hypothèses (i) et (ii) et en supposant qu'on ne fera pas plus de e_{gi} tentatives pour l'unité *i* dans la cellule *g*, sa probabilité de répondre à l'enquête est $\rho_{gi} = 1 - (1 - p_{gi})^{e_{gi}}$. C'est la probabilité que l'unité *i* dans la cellule *g* réponde en au plus e_{gi} tentatives. Nous pouvons maintenant trouver l'effort e_{gi} qui rend cette probabilité de répondre égale au taux de réponse espéré cible ρ_{Tg} pour chaque unité de l'échantillon. On obtient :

$$e_{gi} = \frac{\ln(1 - \rho_{Tg})}{\ln(1 - p_{gi})}. \tag{7}$$

Notre procédure d'appels consiste à sélectionner les cas avec probabilité proportionnelle à l'effort e_{gi}. Il peut être utile de tronquer les valeurs extrêmes de e_{gi} pour éviter de dépenser une trop grande portion du budget pour de telles unités, particulièrement s'il n'y a pas de limites sur le nombre maximum de tentatives.

6 Autres considérations

En pratique, $S^2_{wy,g}$ et p_{gi} sont inconnus et doivent être estimés. Dans une enquête répétée, un choix naturel est d'utiliser les données recueillies lors d'une période précédente de l'enquête afin d'obtenir des estimations de $S^2_{wy,g}$ et p_{gi}. Il est à noter que l'estimation de p_{gi} requiert un fichier qui contient un enregistrement pour chaque appel effectué et non un enregistrement par unité puisque p_{gi} est la probabilité de répondre à une tentative donnée et non la probabilité de répondre à l'enquête.

La solution du problème d'optimisation de la section 5 peut être obtenue avant la collecte des données. Il pourrait être utile de réviser la solution régulièrement (par exemple, quotidiennement) durant la collecte des données. On peut trouver plus de détails à ce sujet dans Beaumont, Bocci et Haziza (2014).

Notre procédure pourrait être améliorée de plusieurs façons. Par exemple, on pourrait vouloir distinguer deux types de non-répondants à la fin de la collecte : ceux qui ont refusé de répondre à l'enquête et ceux qui n'ont pas été contactés. On pourrait aussi vouloir étendre la notion d'effort au-delà du nombre de tentatives d'appels (par exemple, les incitatifs). Une autre amélioration possible serait de raffiner le problème d'optimisation pour s'assurer que les taux de réponse espérés cibles soient plus petits que 1, ce qui nécessiterait un algorithme itératif afin d'obtenir une solution. Finalement, il serait utile de trouver des hypothèses moins restrictives que les hypothèses (i), (ii) et (iii) de la section 3 pour obtenir le nombre espéré de tentatives $\tilde{m}_{gi}$. Toutes ces améliorations potentielles nécessitent des études plus approfondies.

7 Conclusion

Nous avons décrit une procédure de priorisation des appels qui a pour objectif de minimiser la variance due à la non-réponse d'un estimateur ajusté pour la non-réponse sous une contrainte de budget fixé. L'étude empirique de Beaumont, Bocci et Haziza (2014) a montré que notre procédure est plus efficace sous trois différents mécanismes de non-réponse que d'autres alternatives telles que celle qui consiste à maximiser l'indicateur R. En termes de biais, les différents estimateurs ont montré des résultats similaires. Ceci renforce l'argument que l'utilisation d'information auxiliaire à l'étape de collecte n'a que peu d'effet sur le biais dû à la non-réponse pourvu que cette information soit utilisée à l'étape d'estimation.

Bien que nous nous soyons concentrés sur la réduction de la variance due à la non-réponse, nous croyons que la réduction du biais dû à la non-réponse demeure une considération plus importante en pratique. Une approche permettant de réduire le biais consiste à sélectionner un sous-échantillon aléatoire des non-répondants à un moment pré-déterminé après le début de la collecte. Comme on peut s'attendre à observer de la non-réponse dans le sous-échantillon, notre procédure de priorisation des appels pourrait être utilisée pour traiter cette non-réponse.

Bibliographie

[1] Beaumont, J.-F., Bocci, C. et Haziza, D. (2014), An adaptive data collection procedure for call prioritization, *Journal of Official Statistics* (à paraître).

[2] Schouten, B., Cobben, F. et Bethlehem, J. (2009), Indicators for the representativeness of survey response, *Survey Methodology*, 35, 101-113.