

STRATIFICATION EN ENQUÊTES ENTREPRISES : UNE REVUE ET QUELQUES AVANCÉES

Sophie Baillargeon, Louis-Paul Rivest¹ et Michel Ferland²

RÉSUMÉ

Cet article présente une méthode de stratification se basant sur une variable auxiliaire X connue pour toutes les unités de la population. La méthode permet d'inclure une strate recensement et une strate à tirage nul dans le plan d'échantillonnage. Elle permet également de tenir compte d'une non-réponse anticipée ainsi que de l'écart entre la variable de stratification X et la variable à l'étude Y . À l'image de la méthode de Lavallée et Hidiroglou qu'elle généralise, la méthode proposée est optimale : les bornes de stratification sont déterminées en minimisant la taille d'échantillon requise pour atteindre un certain niveau de précision. Nous présentons la généralisation d'un algorithme développé par Sethi pour résoudre le problème d'optimisation sous-jacent à cette méthode.

MOTS CLÉS : Stratification optimale; méthode généralisée de Lavallée-Hidiroglou; algorithme d'optimisation de Sethi.

ABSTRACT

This paper presents a stratification method based on an auxiliary variable X known for all the units in the population. With this method, a take-all stratum and a take-none stratum can be included in the stratification scheme. Moreover, the method takes into account an anticipated non-response and the discrepancy between the stratification variable X and the study variable Y . The proposed method generalizes the Lavallée and Hidiroglou stratification method; it determines the stratum boundaries by minimizing the sample size required to reach a certain level of precision. An algorithm, first presented by Sethi, is generalized to carry out the numerical optimization underlying the proposed stratification method.

KEY WORDS : Optimal stratification; Generalized Lavallée-Hidiroglou method; Sethi's optimization algorithm.

1. INTRODUCTION

Lors de la planification d'une enquête en entreprises, une mesure de taille X est parfois disponible pour toutes les unités de la population, c'est-à-dire pour toutes les entreprises visées. Si la mesure de taille X est fortement reliée à la variable à l'étude Y , la stratification de la population en fonction de X aidera à minimiser la taille d'échantillon nécessaire pour atteindre un certain niveau de précision dans l'estimation du Y total. La stratification consiste à diviser la population en L strates tel que la strate h se compose de toutes les unités pour lesquelles $b_{h-1} \leq X < b_h$, où $b_1 < b_2 < \dots < b_{L-1}$, $b_0 = 0$ et $b_L = \max\{X_i\} + 1$.

Plusieurs techniques ont été proposées pour sélectionner les bornes de stratification b_1 à b_{L-1} . Une méthode encore très populaire est la règle de la racine carrée cumulée de la fréquence ($\text{cum}\sqrt{f}$) de Dalenius et Hodges (1959). Gunning et Horgan (2004) ont aussi proposé une méthode simple, appelée méthode géométrique, assurant des précisions par strate approximativement égales. Pour ces deux méthodes, les bornes sont d'abord choisies et, dans un deuxième temps, la taille d'échantillon n requise pour atteindre un certain niveau de précision est

¹Sophie Baillargeon et Louis-Paul Rivest, Département de mathématiques et de statistique, Pavillon Vachon, 1045 avenue de la Médecine, Université Laval, Québec (QC), Canada, G1V 0A6. Sophie.Baillargeon@mat.ulaval.ca, Louis-Paul.Rivest@mat.ulaval.ca

²Michel Ferland, Statistique Canada, 150 Tunney's Pasture Driveway, Ottawa (ON), Canada, K1A 0T6. Michel.Ferland@statcan.ca

obtenue puis distribuée entre les strates selon une règle de répartition. Les bornes ainsi obtenues sont au mieux approximativement optimales pour la méthode $\text{cum}\sqrt{f}$ (Cochran, 1977). Lavallée et Hidiroglou (1988) ont pourvu au manque d’approches optimales en proposant une méthode qui détermine les bornes b_1 à b_{L-1} pour lesquelles n est minimum tout en respectant le niveau de précision cible. Ils ont aussi suggéré d’échantillonner toutes les unités de la strate contenant les plus grandes unités. L’ajout de cette strate, appelée strate recensement, est particulièrement intéressant dans le contexte d’enquêtes en entreprises où les populations sont souvent très asymétriques. Les quelques très grandes entreprises d’une population ont une influence capitale sur le total de la variable à l’étude.

Toutes ces méthodes supposent que la variable à l’étude Y est égale à la variable de stratification X . En réalité, Y et X sont habituellement fortement reliées mais non égales. Utiliser X pour construire les strates et calculer n risque de sous-estimer la taille d’échantillon nécessaire pour atteindre le niveau de précision cible. Certains auteurs proposent de prédire Y à partir de X puis de construire les strates à partir du Y prédit (Hidiroglou et Laniel, 2001). Un autre moyen proposé afin de tenir compte d’un modèle entre Y et X est l’utilisation de variances *anticipées* de Y étant donné X . Cette façon de faire a l’avantage, contrairement à l’utilisation du Y prédit, de tenir compte de l’incertitude provenant de la prédiction de Y par X . Rivest (2002) a généralisé la méthode de Lavallée-Hidiroglou en remplaçant les variances de X par les variances anticipées de Y étant donné X dans le critère à optimiser.

Rivest (2002) utilise un modèle loglinéaire pour représenter les écarts possibles entre Y et X . Nous généralisons ici ce modèle en y incluant de la mortalité tel que proposé par Ferland et Batten (2007). Elle permet de tenir compte, dans l’optimisation de n , du fait que certaines entreprises risquent de fermer leurs portes entre la collecte des valeurs de X et la tenue de l’enquête ($X > 0$ mais $Y = 0$). De plus, nous tenons compte d’une non-réponse anticipée dans la construction du plan d’échantillonnage stratifié et nous utilisons une règle générale de répartition (Hidiroglou et Srinath, 1993). Cette règle comporte comme cas particuliers les répartitions de Neyman, de puissance et proportionnelle. Avec la méthode de Lavallée-Hidiroglou, la strate recensement est automatiquement ajoutée au plan d’échantillonnage. Nous la mettons ici en option et nous ajoutons même la possibilité d’inclure une strate à tirage nul pour les plus petites unités. Une telle strate vide biaise légèrement l’estimateur de la somme des Y . Par contre, elle peut permettre de réduire la taille d’échantillon requise pour atteindre un certain niveau de précision. Dans leur article, Lavallée et Hidiroglou (1988) utilisent l’algorithme de Sethi pour minimiser n . Dans les sections suivantes, nous présentons les formules nécessaires à la mise en oeuvre de cet algorithme dans le cas généralisé que nous proposons. Ensuite, nous présentons un exemple numérique afin d’illustrer l’utilité de la méthode proposée.

2. FORMULE GÉNÉRALE POUR LA DÉTERMINATION DE LA TAILLE D’ÉCHANTILLON DANS UN PLAN D’ÉCHANTILLONNAGE STRATIFIÉ

Quelques-unes des notations standards en échantillonnage stratifié utilisées dans cet article sont les suivantes :

- L est le nombre de strates, l’indice h représente une strate et $h = L$ réfère à la strate des grandes unités ;
- N_h est, pour $h = 1, \dots, L$, la taille de la strate h et $N = \sum_{h=1}^L N_h$ est la taille totale de la population ;
- n_h est la taille de l’échantillon dans la strate h et $n = \sum_{h=1}^L n_h$ est la taille totale de l’échantillon ;
- \bar{Y}_h et \bar{y}_h sont les moyennes de Y dans la strate h pour la population et l’échantillon respectivement ;
- S_{yh} est l’écart type de Y dans la strate h pour la population ;
- r_h est le taux de réponse dans la strate h , la non-réponse est supposée aléatoire à l’intérieur des strates.

La première strate contient les plus petites unités. Sa taille d’échantillon est $n_1 = 0$ si une strate à tirage nul est incluse dans le plan d’échantillonnage. Lorsque $n_1 = 0$, nous définissons $\bar{y}_1 = 0$. La $L^{\text{ième}}$ strate contient les plus grandes unités. Sa taille d’échantillon est $n_L = N_L$ si une strate recensement est incluse dans le plan d’échantillonnage. Une telle strate contribue à la variance de l’estimateur du total de Y lorsque le taux de réponse dans cette strate est inférieur à 1 ($r_L < 1$). Dans ce contexte, la présence de plus d’une strate recensement peut être envisagée. En conséquence, les strates sont divisées en trois groupes : \mathcal{A} pour la strate à tirage nul, \mathcal{B} pour les strates à tirage partiel (échantillonnées partiellement) et \mathcal{C} pour les strates recensement. Ce dernier ensemble

peut être l'ensemble vide \emptyset , le singleton $\{L\}$, ou, en de rares occasions, un ensemble contenant plus d'une strate. Soulignons que $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$ forme l'ensemble des L strates.

Les tailles d'échantillon dans les strates sont

$$n_h = \begin{cases} 0 & \text{pour } h \in \mathcal{A} \\ (n - \sum_{h \in \mathcal{C}} N_h) a_h & \text{pour } h \in \mathcal{B} \\ N_h & \text{pour } h \in \mathcal{C} \end{cases}$$

où a_h est la règle générale de répartition proposée par Hidiroglou et Srinath (1993). Elle est définie par

$$a_h = \frac{\gamma_h}{\sum_{k \in \mathcal{B}} \gamma_k} \quad \text{pour } h \in \mathcal{B} \quad \text{où } \gamma_h = N_h^{2q_1} \bar{Y}_h^{2q_2} S_{yh}^{2q_3},$$

et (q_1, q_2, q_3) est spécifié par l'utilisateur. Les valeurs $q_1 = 1/2, q_2 = 0, q_3 = 1/2$ donnent la répartition de Neyman, la répartition proportionnelle correspond à $q_1 = 1/2, q_2 = q_3 = 0$ tandis que la répartition de puissance est obtenue lorsque $q_1 = q_2 = p/2$ et $q_3 = 0$.

L'estimateur de T_y , le total de Y , peut s'écrire $\hat{T}_y = \sum N_h \bar{y}_h$; son erreur quadratique moyenne est donnée par :

$$EQM(\hat{T}_Y) = (N_{\mathcal{A}} \bar{Y}_{\mathcal{A}})^2 + \sum_{h \in \mathcal{B} \cup \mathcal{C}} N_h^2 S_{yh}^2 \left(\frac{1}{n_h r_h} - \frac{1}{N_h} \right). \quad (1)$$

En isolant n dans l'équation 1, on obtient

$$n = \sum_{h \in \mathcal{C}} N_h + \frac{\sum_{h \in \mathcal{B}} N_h^2 S_{yh}^2 / (a_h r_h)}{EQM(\hat{T}_Y) - (N_{\mathcal{A}} \bar{Y}_{\mathcal{A}})^2 + \sum_{h \in \mathcal{B}} N_h S_{yh}^2 + \sum_{h \in \mathcal{C}} N_h S_{yh}^2 (1 - 1/r_h)}.$$

Le niveau de précision à atteindre sur \hat{T}_y s'écrit $EQM(\hat{T}_y) = c^2 T_y^2$ où c est la racine carré de l'erreur quadratique moyenne relative (REQMR) visée. Si le plan d'échantillonnage ne contient pas de strate à tirage nul, aucun biais n'est induit et le REQMR revient au coefficient de variation. L'étendue $c = 1\%$ à 10% est souvent utilisée. Les moments de Y ne sont pas connus donc dans la formule ci-dessus on remplace T_y, S_{yh}^2 et a_h par leurs valeurs anticipées T_{ay}, S_{ayh}^2 et a_{ah} qui dépendent de la variable de stratification X et du modèle conditionnel de Y étant donné X . Les bornes de stratification optimales sont les valeurs de b_1, \dots, b_{L-1} qui minimisent

$$n = \sum_{h \in \mathcal{C}} N_h + \frac{\sum_{h \in \mathcal{B}} N_h^2 S_{ayh}^2 / (a_{ah} r_h)}{c^2 T_{ay}^2 - T_{a_{\mathcal{A}y}}^2 + \sum_{h \in \mathcal{B}} N_h S_{ayh}^2 + \sum_{h \in \mathcal{C}} N_h S_{ayh}^2 (1 - 1/r_h)}, \quad (2)$$

où $T_{a_{\mathcal{A}y}}$ est le total anticipé de Y dans la strate à tirage nul. Cette valeur vaut 0 si $\mathcal{A} = \emptyset$.

Le calcul de l'erreur quadratique moyenne en présence d'une strate à tirage nul qui est considéré ici représente le cas où aucune estimation de biais n'est disponible dans cette strate. Dans plusieurs situations, des données administratives permettent d'estimer ce biais, au moins partiellement. Dans ce cas, il serait intéressant de diminuer dans l'équation (1) la contribution du biais à l'erreur quadratique moyenne. Ceci permettrait sans doute d'envisager des strates à tirage nul de plus grande taille que celle présentée dans l'exemple de la section 5.

3. MODÈLE LOGLINÉAIRE AVEC MORTALITÉ DE Y ÉTANT DONNÉ X

Nous proposons un modèle dans lequel la variable Y peut valoir 0 avec une probabilité non nulle. La probabilité de cet événement décroît en fonction de la taille de l'entreprise. Le modèle considéré ici comporte une probabilité

de survie p_h qui varie entre les strates. Les écarts entre Y et X sont aussi modélisés par de petites perturbations ϵ , sur l'échelle logarithmique, avec $\epsilon \sim N(0, \sigma^2)$. Le modèle conditionnel de Y étant donné X dans la strate h est :

$$Y = \begin{cases} \exp(\alpha + \beta \log(X) + \epsilon) & \text{avec probabilité } p_h \\ 0 & \text{avec probabilité } 1 - p_h \end{cases} . \quad (3)$$

Dans les calculs suivants, nous supposons que la distribution de X est connue, donnant un poids de $1/N$ aux N valeurs possibles de X . De la fonction génératrice des moments d'une variable aléatoire normale, nous savons que $E(e^{t\epsilon}) = e^{\sigma^2 t^2/2}$. Les moments anticipés de Y étant donné que l'unité appartient à la strate h sont donc

$$\begin{aligned} \bar{Y}_{ah} &= E(Y|b_{h-1} \leq X < b_h) = p_h E(e^\alpha X^\beta e^\epsilon | b_{h-1} \leq X < b_h) \\ &= p_h e^{\alpha + \sigma^2/2} E(X^\beta | b_{h-1} \leq X < b_h) = p_h e^{\alpha + \sigma^2/2} \left\{ \sum_{b_{h-1} \leq X_i < b_h} X_i^\beta / N_h \right\} , \end{aligned}$$

$$\begin{aligned} S_{ayh}^2 &= E(Y^2 | b_{h-1} \leq X < b_h) - \{E(Y | b_{h-1} \leq X < b_h)\}^2 \\ &= p_h e^{2\alpha + 2\sigma^2} E(X^{2\beta} | b_{h-1} \leq X < b_h) - p_h^2 e^{2\alpha + \sigma^2} \{E(X^\beta | b_{h-1} \leq X < b_h)\}^2 \\ &= p_h e^{2\alpha + 2\sigma^2} \left\{ \sum_{b_{h-1} \leq X_i < b_h} X_i^{2\beta} / N_h \right\} - p_h^2 e^{2\alpha + \sigma^2} \left\{ \sum_{b_{h-1} \leq X_i < b_h} X_i^\beta / N_h \right\}^2 . \end{aligned}$$

Notez aussi que la valeur anticipée de T_{ay} est

$$T_{ay} = NE(E(Y|X)) = e^{\alpha + \sigma^2/2} \sum_{h=1}^L p_h N_h E(X^\beta | b_{h-1} \leq X < b_h) = e^{\alpha + \sigma^2/2} \sum_{h=1}^L p_h \left\{ \sum_{b_{h-1} \leq X_i < b_h} X_i^\beta \right\} .$$

De façon similaire, $T_{aAy} = e^{\alpha + \sigma^2/2} p_A N_A E(X^\beta | X < b_1) = e^{\alpha + \sigma^2/2} p_A \{\sum_{X_i < b_1} X_i^\beta\}$ si $\mathcal{A} \neq \emptyset$ et $T_{aAy} = 0$ sinon. Pour un ensemble de bornes $b_1 < b_2 < \dots < b_L$, la taille d'échantillon n requise pour atteindre le REQMR cible c avec le modèle (3) peut être évaluée en remplaçant les expressions pour les moments anticipés calculées ci-dessus dans la formule (2).

4. ALGORITHME DE SETHI

Afin de trouver les bornes de stratification qui minimisent n pour un REQMR cible c , nous allons résoudre $\partial n / \partial b_h = 0$ pour $h = 1, \dots, L-1$. Pour ce faire, nous exprimons d'abord les moments de la section 3 en fonction de $f(x)$, la fonction de densité de la variable de stratification supposée continue. En utilisant la notation $W_h = \int_{b_{h-1}}^{b_h} f(x) dx$, $\phi_h = \int_{b_{h-1}}^{b_h} x^\beta f(x) dx$ et $\psi_h = \int_{b_{h-1}}^{b_h} x^{2\beta} f(x) dx$, nous avons $E(X^\beta | b_{h-1} \leq X < b_h) = \phi_h / W_h$ et $E(X^{2\beta} | b_{h-1} \leq X < b_h) = \psi_h / W_h$. Notons aussi que, dans le modèle loglinéaire, e^α est un facteur multiplicatif qui n'a pas d'impact sur les résultats. On fixe donc $\alpha = 0$ sans perte de généralité. Nous réécrivons donc ici

$$\begin{aligned} \bar{Y}_{ah} &= p_h e^{\sigma^2/2} \phi_h / W_h, & T_{ay} &= N e^{\sigma^2/2} \sum_{h=1}^L p_h \phi_h, \\ S_{ayh}^2 &= p_h e^{2\sigma^2} \psi_h / W_h - p_h^2 e^{\sigma^2} \{\phi_h / W_h\}^2, & T_{aAy} &= N e^{\sigma^2/2} p_A \phi_A \quad \text{si } \mathcal{A} \neq \emptyset. \end{aligned}$$

Après avoir exprimé n en terme de W_h , ϕ_h et ψ_h , les dérivées partielles de n par rapport aux b_h peuvent être évaluées en utilisant la règle de dérivation en chaîne comme suit

$$\frac{\partial n}{\partial b_h} = \frac{\partial n}{\partial W_h} \frac{\partial W_h}{\partial b_h} + \frac{\partial n}{\partial \phi_h} \frac{\partial \phi_h}{\partial b_h} + \frac{\partial n}{\partial \psi_h} \frac{\partial \psi_h}{\partial b_h} + \frac{\partial n}{\partial W_{h+1}} \frac{\partial W_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \phi_{h+1}} \frac{\partial \phi_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \psi_{h+1}} \frac{\partial \psi_{h+1}}{\partial b_h}$$

Notez que

$$\frac{\partial W_h}{\partial b_h} = -\frac{\partial W_{h+1}}{\partial b_h} = f(b_h), \quad \frac{\partial \phi_h}{\partial b_h} = -\frac{\partial \phi_{h+1}}{\partial b_h} = b_h^\beta f(b_h) \quad \text{et} \quad \frac{\partial \psi_h}{\partial b_h} = -\frac{\partial \psi_{h+1}}{\partial b_h} = b_h^{2\beta} f(b_h).$$

En conséquence, nous avons pour $h = 1, \dots, L-1$

$$\frac{\partial n}{\partial b_h} = f(b_h) \left\{ \left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}} \right) + \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}} \right) b_h^\beta + \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) b_h^{2\beta} \right\}.$$

L'algorithme de Sethi propose une façon de résoudre $\partial n / \partial b_h = 0$. Il considère que les dérivées partielles sont proportionnelles à une fonction quadratique en b_h^β . À chaque itération de l'algorithme, la valeur mise à jour de b_h est la plus grande racine de cette fonction quadratique, élevée à la puissance $1/\beta$. Pour $h = 1, \dots, L-1$ on a donc

$$b_h^{\text{nouveau}} = \left\{ \frac{- \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}} \right) + \left\{ \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}} \right)^2 - 4 \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) \left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}} \right) \right\}^{1/2}}{2 \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right)} \right\}^{1/\beta}.$$

4.1 Calcul des dérivées partielles

Pour faciliter le calcul des dérivées partielles, écrivons n de la façon suivante :

$$n = T + \frac{U}{V} = T + \frac{U_1 U_2}{V_1 - V_2 + V_3 + V_4}$$

$$\text{où } T = N \sum_{h \in \mathcal{C}} W_h, \quad V_1 = c^2 T_{ay}^2 / N, \quad V_2 = T_{aAy}^2 / N, \quad V_3 = \sum_{h \in \mathcal{B}} W_h S_{ayh}^2, \quad V_4 = \sum_{h \in \mathcal{C}} W_h S_{ayh}^2 (1 - 1/r_h) \quad \text{et}$$

$$U_1 = \sum_{h \in \mathcal{B}} \frac{W_h^2 S_{ayh}^2}{\gamma_{ah} r_h} = \sum_{h \in \mathcal{B}} \frac{W_h^{(2-2q_1)} S_{ayh}^{2(1-q_3)}}{\bar{Y}_{ah}^{2q_2} r_h}, \quad U_2 = \sum_{h \in \mathcal{B}} \gamma_{ah} = \sum_{h \in \mathcal{B}} W_h^{2q_1} \bar{Y}_{ah}^{2q_2} S_{ayh}^{2q_3}.$$

Pour obtenir les bornes mises à jour b_h^{nouveau} , on doit calculer $\frac{\partial n}{\partial W_h}$, $\frac{\partial n}{\partial \phi_h}$ et $\frac{\partial n}{\partial \psi_h}$. Ci-dessous, nous présentons d'abord des formules générales applicables à W_h , ϕ_h et ψ_h . Nous utilisons θ_h pour représenter W_h , ϕ_h ou ψ_h .

$$\frac{\partial n}{\partial \theta_h} = \frac{\partial T}{\partial \theta_h} + \frac{(\frac{\partial U_1}{\partial \theta_h} U_2 + U_1 \frac{\partial U_2}{\partial \theta_h}) V - U (\frac{\partial V_1}{\partial \theta_h} - \frac{\partial V_2}{\partial \theta_h} + \frac{\partial V_3}{\partial \theta_h} + \frac{\partial V_4}{\partial \theta_h})}{V^2}$$

où

$$\frac{\partial T}{\partial \theta_h} = \begin{cases} 0 & \text{pour } h \in \mathcal{A} \cup \mathcal{B} \\ N \frac{\partial W_h}{\partial \theta_h} & \text{pour } h \in \mathcal{C} \end{cases}, \quad \frac{\partial V_1}{\partial \theta_h} = \frac{c^2 2 T_{ay}}{N} \frac{\partial T_{ay}}{\partial \theta_h}, \quad \frac{\partial V_2}{\partial \theta_h} = \begin{cases} \frac{2 T_{aAy}}{N} \frac{\partial T_{aAy}}{\partial \theta_h} & \text{pour } h \in \mathcal{A} \\ 0 & \text{pour } h \in \mathcal{B} \cup \mathcal{C} \end{cases},$$

$$\frac{\partial V_3}{\partial \theta_h} = \begin{cases} \left(\frac{\partial W_h}{\partial \theta_h} S_{ayh}^2 + W_h \frac{\partial S_{ayh}^2}{\partial \theta_h} \right) & \text{pour } h \in \mathcal{B} \\ 0 & \text{pour } h \in \mathcal{A} \cup \mathcal{C} \end{cases}, \quad \frac{\partial V_4}{\partial \theta_h} = \begin{cases} 0 & \text{pour } h \in \mathcal{A} \cup \mathcal{B} \\ \left(1 - \frac{1}{r_h} \right) \left(\frac{\partial W_h}{\partial \theta_h} S_{ayh}^2 + W_h \frac{\partial S_{ayh}^2}{\partial \theta_h} \right) & \text{pour } h \in \mathcal{C} \end{cases},$$

$$\frac{\partial U_1}{\partial \theta_h} = \begin{cases} \frac{1}{r_h} \left\{ (2-2q_1) W_h^{(1-2q_1)} \bar{Y}_{ah}^{(-2q_2)} S_{ayh}^{2(1-q_3)} \frac{\partial W_h}{\partial \theta_h} + \right. & \text{pour } h \in \mathcal{B} \\ \left. W_h^{(2-2q_1)} (-2q_2) \bar{Y}_{ah}^{(-2q_2-1)} S_{ayh}^{2(1-q_3)} \frac{\partial \bar{Y}_{ah}}{\partial \theta_h} + W_h^{(2-2q_1)} \bar{Y}_{ah}^{-2q_2} (1-q_3) S_{ayh}^{2(-q_3)} \frac{\partial S_{ayh}^2}{\partial \theta_h} \right\} & \text{et} \\ 0 & \text{pour } h \in \mathcal{A} \cup \mathcal{C} \end{cases}$$

$$\frac{\partial U_2}{\partial \theta_h} = \begin{cases} 2q_1 W_h^{(2q_1-1)} \bar{Y}_{ah}^{2q_2} S_{ayh}^{2q_3} \frac{\partial W_h}{\partial \theta_h} + W_h^{2q_1} 2q_2 \bar{Y}_{ah}^{(2q_2-1)} S_{ayh}^{2q_3} \frac{\partial \bar{Y}_{ah}}{\partial \theta_h} + W_h^{2q_1} \bar{Y}_{ah}^{2q_2} q_3 S_{ayh}^{2(q_3-1)} \frac{\partial S_{ayh}^2}{\partial \theta_h} & \text{pour } h \in \mathcal{B} \\ 0 & \text{pour } h \in \mathcal{A} \cup \mathcal{C} \end{cases} .$$

Pour le modèle loglinéaire avec mortalité, nous avons dans ces formules

$$\frac{\partial W_h}{\partial \theta_h} = \begin{cases} 1 & \text{pour } \theta_h = W_h \\ 0 & \text{pour } \theta_h = \phi_h \\ 0 & \text{pour } \theta_h = \psi_h \end{cases}, \quad \frac{\partial S_{ayh}^2}{\partial \theta_h} = \begin{cases} p_h e^{\sigma^2} \left(-e^{\sigma^2} \frac{\psi_h}{W_h^2} + 2p_h \frac{\phi_h^2}{W_h^3} \right) & \text{pour } \theta_h = W_h \\ -2p_h^2 e^{\sigma^2} \frac{\phi_h}{W_h^2} & \text{pour } \theta_h = \phi_h \\ \frac{p_h e^{2\sigma^2}}{W_h} & \text{pour } \theta_h = \psi_h \end{cases},$$

$$\frac{\partial \bar{Y}_{ah}}{\partial \theta_h} = \begin{cases} -p_h e^{\sigma^2/2} \frac{\phi_h}{W_h^2} & \text{pour } \theta_h = W_h \\ \frac{p_h e^{\sigma^2/2}}{W_h} & \text{pour } \theta_h = \phi_h \\ 0 & \text{pour } \theta_h = \psi_h \end{cases}, \quad \frac{\partial T_{ay}}{\partial \theta_h} = \begin{cases} 0 & \text{pour } \theta_h = W_h \\ N p_h e^{\sigma^2/2} & \text{pour } \theta_h = \phi_h \\ 0 & \text{pour } \theta_h = \psi_h \end{cases},$$

de plus $\partial T_{aAy}/(\partial \phi_A) = N p_A e^{\sigma^2/2}$ si $\mathcal{A} \neq \emptyset$ et $\partial T_{aAy}/(\partial \phi_A) = 0$ sinon, et $\partial T_{aAy}/(\partial \theta_h) = 0$ pour tout $\theta_h \neq \phi_A$.

5. EXEMPLE NUMÉRIQUE

Pour illustrer la méthode généralisée de Lavallée-Hidiroglou proposée dans cet article, nous allons présenter un exemple se rapportant à l'Enquête mensuelle sur le commerce de détail (Monthly Retail Trade Survey) de Statistique Canada. Pour mener cette enquête, les entreprises sont stratifiées, à l'intérieur de chaque strate administrative, selon une mesure de taille X . Cette mesure est construite à partir d'informations provenant de la déclaration de revenus des entreprises. La variable à l'étude Y mesure quant à elle les ventes mensuelles annualisées. Pour des questions de confidentialité, l'exemple de cette section à été fait sur des valeurs de X simulées représentant bien la réalité à l'intérieur d'une strate administrative. La population étudiée comprend 2000 entreprises et l'asymétrie des valeurs de X est forte (coefficient d'asymétrie = 28). En fait, deux grandes unités se distinguent nettement du lot, elles représentent à elles seules 8.5% du total de X .

Considérons pour cet exemple que nous voulons atteindre un REQMR de 1% avec 4 strates, dont une strate recensement, en utilisant la répartition de Neyman. La strate recensement s'avère avantageuse ici en raison de la forte asymétrie des données. Nous supposons que les taux de réponse seront de 85%, 90%, 90% et 100% dans les strates $h = 1, 2, 3$ et 4 respectivement, et que les paramètres du modèle loglinéaire avec mortalité entre X et Y sont $\beta = 0.9$, $\sigma^2 = 0.015$ et $p_h = (0.8, 0.9, 0.95, 1)$. Nous allons obtenir des plans d'échantillonnage stratifié avec différentes méthodes : géométrique, $\text{cum}\sqrt{f}$, Lavallée-Hidiroglou original ou généralisé. Nos résultats sont présentés dans le tableau 1 ci-dessous.

Tableau 1 - Résultats

| Méthode | Modèle considéré | N_0 strate à tirage nul | N_1 1 ^{ière} strate à tirage partiel | N_2 2 ^e strate à tirage partiel | N_3 3 ^e strate à tirage partiel | N_4 strate recen- sement | n | REQMR anticipé |
|-----------------------------------|--------------------------|------------------------------------|----------------------------------------------------------|-------------------------------------------------------|-------------------------------------------------------|-------------------------------------|------|-------------------|
| géométrique | aucun | - | 4 | 390 | 1585 | 21 | 1125 | 1.22% |
| $\text{cum}\sqrt{f}$ | aucun | - | 809 | 713 | 375 | 103 | 349 | 1.97% |
| Lavallée-Hidiroglou | aucun | - | 766 | 668 | 413 | 153 | 340 | 2.07% |
| Lavallée-Hidiroglou généralisé | + non-réponse | - | 751 | 674 | 417 | 158 | 365 | 1,93% |
| | + modèle loglinéaire | - | 699 | 717 | 426 | 158 | 434 | 1,61% |
| | + mortalité | - | 244 | 634 | 814 | 308 | 681 | 1,00% |
| | + strate à tirage nul | 31 | 243 | 612 | 807 | 307 | 676 | 1,00% |

La méthode généralisée de Lavallée-Hidiroglou est utilisée 4 fois, en tenant compte à chaque fois de différents paramètres. La première fois, nous avons intégré uniquement la non-réponse à l'optimisation. La deuxième fois, nous avons aussi tenu compte du modèle loglinéaire entre X et Y , sans mortalité pour commencer. C'est à la troisième utilisation de la méthode généralisée de Lavallée-Hidiroglou que la mortalité a été ajoutée. Finalement, en plus de la non-réponse et du modèle loglinéaire avec mortalité, nous avons intégré une strate à tirage nul au plan d'échantillonnage. Le REQMR anticipé apparaissant dans la dernière colonne du tableau est calculé en supposant le modèle complet entre X et Y .

De ces résultats, on remarque premièrement que la méthode géométrique ne fonctionne pas bien, et ce, à cause de 4 unités très petites. De plus, la méthode de Lavallée-Hidiroglou propose un n un peu plus petit que $\text{cum}\sqrt{f}$. Si on fait une correction a posteriori pour la non-réponse à partir des résultats de Lavallée-Hidiroglou original (n_h devient n_h/r_h), on obtient $n = 367$. Ce résultat se rapproche du $n = 365$ de Lavallée-Hidiroglou généralisé tenant compte de la non-réponse. Cependant, Lavallée-Hidiroglou généralisé tenant compte de la non-réponse modifie légèrement les strates en déplaçant quelques unités vers la strate recensement pour laquelle aucune non-réponse n'a été supposée. Si on ajoute à cela le modèle loglinéaire, n augmente d'environ 20% sans que les strates soient trop changées. Cela permet au REQMR anticipé de se rapprocher du REQMR cible. Lorsque finalement on ajoute la mortalité, on peut enfin atteindre le REQMR cible, mais pour cela le n doit être augmenté de près de 50% par rapport au n de Lavallée-Hidiroglou original. On note une augmentation importante de la taille des deux dernières strates pour les grandes unités car la mortalité est supposée moins importante dans ces strates. Finalement, en ajoutant une strate à tirage nul, on diminue le n de 5 en induisant un biais négligeable (biais relatif de 0.08%).

Notons que pour la dernière méthode, avec la strate à tirage nul, l'algorithme de Sethi ne converge pas. On a dû utiliser un autre algorithme, attribuable à Kozak (2004), afin de trouver les bornes optimale. Cet algorithme applique à chaque itération un changement aléatoire à une strate choisie elle aussi aléatoirement et accepte le changement seulement s'il permet de diminuer n . Cette action est répétée jusqu'à ce que le n reste inchangé pendant un certain nombre d'itérations consécutives. L'algorithme de Kozak respecte bien le caractère discret de la minimisation de n .

6. CONCLUSION

Dans la littérature, l'intégration des différences entre la variable à l'étude Y et la variable de stratification X dans la confection d'un plan d'échantillonnage stratifié optimal a reçu peu d'attention. Pourtant, l'écart entre les deux variables a un impact important sur la précision des estimations échantillonales. Dans cet article, nous avons suggéré une méthode généralisée de Lavallée-Hidiroglou pour construire des plans d'échantillonnage stratifié en tenant compte d'un modèle loglinéaire avec mortalité entre Y et X . Cette méthode permet aussi de tenir compte d'une non-réponse anticipée et d'intégrer une strate à tirage nul dans le plan d'échantillonnage.

Les calculs de l'exemple numérique ont tous été faits en R à l'aide d'un package nommé *stratification* actuellement en développement. Dans ce package, les méthodes $\text{cum}\sqrt{f}$, géométrique, Lavallée-Hidiroglou original et généralisé sont toutes implantées. La méthode de Lavallée-Hidiroglou, originale ou généralisée, peut même être mise en oeuvre avec l'algorithme de Sethi ou encore celui de Kozak qui souffre moins de problèmes numériques que l'algorithme de Sethi. Le package R *stratification* sera sous peu rendu disponible gratuitement sur internet.

REMERCIEMENTS

Nous désirons remercier Pierre Lavallée, à qui nous devons l'idée d'inclure une strate à tirage nul dans le plan d'échantillonnage. Nous remercions également le Conseil de recherches en sciences naturelles et en génie du Canada et le Fonds québécois de la recherche sur la nature et les technologies pour leur soutien financier.

RÉFÉRENCES

- COCHRAN, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York, 3e édition.
- DALENIUS, T. et HODGES, J. L. (1959). Minimum Variance Stratification (Corr : V58 p1161). *Journal of the American Statistical Association*, 54:88–101.
- FERLAND, M. et BATTEN, D. (2007). Sampling for the MWRTS Redesign, Appendix A - Enhanced Lavallée-Hidiroglou Method. Cahier de travail de la direction de la méthodologie DMEE-2007-002E, Statistique Canada, Ottawa. 27–33.
- GUNNING, P. et HORGAN, J. M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology*, 30(2):159–166.
- HIDIROGLOU, M. A. et LANIEL, N. (2001). Sampling and Estimation Issues for Annual and Sub-annual Canadian Business Surveys. *International Statistical Review*, 69:487–504.
- HIDIROGLOU, M. A. et SRINATH, K. P. (1993). Problems Associated with Designing Subannual Business Surveys. *Journal of Business & Economic Statistics*, 11:397–405.
- KOZAK, M. (2004). Optimal Stratification Using Random Search Method in Agricultural Surveys. *Statistics in Transition*, 6(5):797–806.
- LAVALLÉE, P. et HIDIROGLOU, M. A. (1988). On the Stratification of Skewed Populations (Corr : V14 p347). *Survey Methodology*, 14:33–43.
- RIVEST, L.-P. (2002). A Generalization of the Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys. *Survey Methodology*, 28(2):191–198.