L'ÉCHANTILLONNAGE ÉQUILIBRÉ PAR LA MÉTHODE DU CUBE ET LA MÉTHODE RÉJECTIVE

Ibrahima Ousmane Ida[1] *et Louis-Paul Rivest*[2]

[1]*Étudiant au deuxième cycle en statistique à l'Université Laval.*
[2]*Professeur titulaire à l'Université Laval.*

Résumé :
Au cours de ces dernières années, les techniques d'échantillonnage équilibré ont connu un regain d'intérêt. En effet, elles permettent d'estimer exactement ou approximativement les totaux des variables auxiliaires avec l'estimateur d'Horvitz-Thompson, afin d'améliorer l'efficacité des estimations . Encore récemment, des nouvelles procédures ont été proposées. Il s'agit notamment de la méthode du cube, une méthode exacte présentée par Deville and Tillé (2004), et une méthode approximative, l'algorithme réjectif de Fuller (2009). Après une brève présentation de ces deux méthodes dans le cadre d'un inventaire de pêcheurs, nous comparons à l'aide de simulations Monte Carlo, les plans de sondages produits par ces deux méthodes d'échantillonnage.
Mots clés : Méthode du cube, méthode réjective, simulations Monte Carlo.

Abstract :
In recent years, balanced sampling techniques have experienced a renewed interest. They allow to reproduce the structure of the population in samples in order to improve the efficiency of survey estimates. New procedures have been proposed. These include the cube method, an exact method presented by Deville and Tillé (2004), and an approximate method, the Fuller (2009) rejective algorithm. After a brief presentation of these methods as part of an angler survey, we compare using Monte Carlo simulations, the survey designs produced by these two sampling algorithms.
Key words : Cube method, rejective algorithm, Monte Carlo simulations.

1 Présentation du problème

En 2014 et 2015, des enquêtes sur la pêche sportive au bar rayé en Gaspésie ont été réalisées en vue de rendre compte de l'ampleur de cette activité récréative dans la région. Sollicité pour la conception du plan de sondage, le service de consultation statistique de l'Université Laval est amené à élaborer un algorithme complexe pour la sélection des échantillons, car le plan devait tenir compte de nombreuses contraintes dont certaines sont d'ordre spatial et d'autres temporel.

D'une part, la zone d'enquête est constituée de sites de pêche qui n'ont pas la même importance en termes d'attraction de pêcheurs. Aussi, ces sites sont localisés dans des secteurs qui représentent le premier niveau de subdivision de la zone. D'autre part, une journée d'enquête comprend différentes périodes durant lesquelles l'affluence des personnes sur les sites pourraient varier considérablement. Les périodes sont encore composées de sous-périodes. Tous ces facteurs doivent être intégrés dans le plan de façon que les échantillons fournissent des fréquences de visites par secteurs et par sites proportionnelles à l'importance de ceux-ci et qui soient réparties équitablement entre les périodes et les sous-périodes.

Pour tenter d'apporter des solutions au problème posé, nous avons emprunté une voie alternative qui consiste à utiliser l'échantillonnage équilibré. Il s'agit en effet d'une technique qui permet d'inclure divers types de contraintes dans des plans de sondage. Ainsi, à partir de plans hautement stratifiés, nous avons employé deux variantes de cette technique qui sont la méthode du cube (Deville and Tillé, 2004; Tillé, 2011b,a) et la méthode réjective (Fuller, 2009). Cependant, les préoccupations ne sont pas seulement d'intégrer des contraintes, elles consistent aussi à vérifier si ces méthodes ne modifient pas les probabilités d'inclusion des unités.

2 Notation

Soient une population U de taille N et p variables auxiliaires $x_1, x_2, \cdots, x_p$ dont les valeurs sont connues pour toutes les unités de U. Si $x_i = (x_{1i}, x_{2i}, \cdots, x_{pi})'$ est le vecteur de valeurs prises par les p variables auxiliaires pour l'unité i, le vecteur des totaux de ces variables dans la population s'écrit :

$$T_x = \sum_{i \in U} x_i \tag{1}$$

où i est une unité de la population U et x_i le vecteur des valeurs des variables auxiliaires pour cette unité.
Soient $p(s)$ un plan de sondage et Z_i $(i = 1, \cdots, N)$ des variables indicatrices telles que :

$$Z_i = \begin{cases} 1, & si\ i \in s; \\ 0, & sinon. \end{cases} \qquad et \qquad P[Z_i = z] = \begin{cases} \pi_i, & si\ z = 1; \\ 1 - \pi_i, & si\ z = 0. \end{cases}$$

où s représente l'échantillon tiré de la population U en utilisant un plan $p(s)$ et π_i la probabilité d'inclusion de l'unité i de U.

L'estimateur de Horvitz-Thompson du vecteur des totaux des variables auxiliaires est :

$$\widehat{T}_{xHT} = \sum_{i \in s} \frac{x_i}{\pi_i} \tag{2}$$

3 Échantillonnage équilibré

Le plan de sondage $p(s)$ est dit équilibré sur les p variables $x_1, x_2, \cdots, x_p$ si et seulement si il permet de sélectionner des échantillons tels que :

$$\widehat{T}_{xHT} = T_x \tag{3}$$

Remarque 1 *Qu'il s'agisse de plan de sondage simple ou de plan complexe, le principe d'échantillonnage équilibré reste le même. De plus, l'équilibrage peut se faire par rapport aux moyennes des variables auxiliaires au lieu des totaux.*

4 Méthode du cube

La méthode dite du cube doit son nom à la représentation géométrique d'un plan de sondage, car un échantillon s peut être représenté par un vecteur d'indicatrices montrant l'appartenance ou non des unités de la population à l'échantillon. De ce fait, un échantillon s, tiré dans une population U de taille N, se traduit par :

$$s = (Z_1, Z_2, \cdots, Z_N)'$$

où l'indicatrice $Z_i = 1$ si $i \in s$ et $Z_i = 0$ si $i \notin s$.

Alors, un échantillon s s'interprète comme l'un des sommets d'un hypercube $\mathcal{C}$ de dimension N défini ainsi :

$$\mathcal{C} = [0, 1]^N$$

En outre, les équations d'équilibrage (3) peuvent s'écrire en ces termes :

$$\begin{cases} \sum_{i \in U} \frac{x_i}{\pi_i} Z_i = \sum_{i \in U} x_i \\ Z_i \in \{0, 1\}, i \in U \end{cases} \tag{4}$$

où x_i est le vecteur des valeurs prises par les p variables d'équilibrage pour l'unité i de la population U et π_i la probabilité d'inclusion de cette unité.

En fait, ces équations caractérisent un sous-espace de contraintes de dimension N-p, appelé $\mathcal{Q}$, qui appartient à $\mathbb{R}^N$.

La méthode du cube permet donc de tirer un échantillon appartenant à $\mathcal{C} \cap \mathcal{Q}$. Si cet échantillon est un sommet de $\mathcal{C}$, alors il s'agit d'un échantillon exactement équilibré. Par contre, si un tel échantillon n'existe pas, la méthode trouvera une solution approchée.

Remarque 2 *Dans nos applications faites pour répondre aux préoccupations exposées, nous avons conçu plusieurs plans de sondage stratifiés dont certains sont à deux degrés et d'autres à trois degrés. Ainsi, pour tirer les échantillons par la méthode du cube, nous*

avons utilisé la méthode de Hasler and Tillé (2014), adaptée pour une population hautement stratifiée. Cette méthode a été également utilisée par Vallée et al. (2015). L'avantage de cette nouvelle version, c'est qu'elle permet de réduire les contraintes lorsque leur nombre est trop important pour trouver des échantillons équilibrés.

5 Méthode réjective de Fuller

La méthode réjective de Fuller se définit comme une procédure qui permet de sélectionner, parmi les échantillons tirés dans un population finie selon le plan de sondage $p(s)$, le premier échantillon qui satisfait la relation :

$$\left(\widehat{T}_{xHT} - T_x\right)' \left[V\left(\widehat{T}_{xHT}\right)\right]^{-1} \left(\widehat{T}_{xHT} - T_x\right) < \gamma^2 \tag{5}$$

Dans cette inéquation, le terme de droite γ^2 est une valeur de tolérance fixée et définie à partir d'une constante γ strictement positive et $V\left(\widehat{T}_{xHT}\right)$ est la matrice de variances covariances de l'estimateur $\widehat{T}_{xHT}$.

Remarque 3 *Avec les plans hautement stratifiés définis dans nos applications, nous avons utilisé la méthode réjective pour tirer des échantillons. Cependant, nous avons aussi tenté de vérifier si cette méthode ne modifie pas les probabilités d'inclusion des unités, c'est-à-dire si :*

$$P[Z_i = 1 \,|contraintes\ (5)] = \pi_i$$

6 Simulations

Les applications ont été faites à travers des simulations Monte Carlo par les deux méthodes d'échantillonnage. D'une part, grâce à ces simulations, nous avons pu examiner le respect des contraintes après l'emploi de la méthode du cube dans le contexte de l'enquête sur la pêche sportive en Gaspésie. D'autre part, elles nous ont également permis de confronter la méthode du cube à la méthode réjective par une analyse des biais et des erreurs quadratiques moyennes des estimateurs calculés à partir des échantillons simulés.

Bibliographie

Jean-Claude Deville and Yves Tillé. Efficient balanced sampling: the cube method. *Biometrika*, 91(4):893–912, 2004.

Wayne A Fuller. Some design properties of a rejective sampling procedure. *Biometrika*, 96 (4):933–944, 2009.

Caren Hasler and Yves Tillé. Fast balanced sampling for highly stratified population. *Computational Statistics & Data Analysis*, 74:81–94, 2014.

Yves Tillé. Dix années d'échantillonnage équilibré par la méthode du cube: une évaluation. *Techniques d'enquête*, 37:233–246, 2011a.

Yves Tillé. *Sampling algorithms*. Springer, New York, 2011b.

Audrey-Anne Vallée, Bastien Ferland-Raymond, Louis-Paul Rivest, and Yves Tillé. Incorporating spatial and operational constraints in the sampling designs for forest inventories. *Environmetrics*, 26(8):557–570, 2015.