Techniques d'enquête, juin 1995
Vol. 21, n° 1, pp. 73–80
Statistique Canada

Plans à chaînes de Markov pour l'échantillonnage à une unité par strate

F. JAY BREIDT[1]

RÉSUMÉ

Dans le domaine de l'échantillonnage de populations finies, les résultats classiques indiquent que l'échantillonnage systématique est le plan le plus efficace pour l'échantillonnage à une unité par strate avec probabilités égales dans le cas de certains types de superpopulations autocorrélées, mais que l'échantillonnage aléatoire simple stratifié est peut-être nettement supérieur à l'échantillonnage systématique si la superpopulation correspond à une tendance avec erreurs non corrélées. Que faire si la superpopulation est la somme d'une tendance et d'erreurs autocorrélées? Intuitivement, on peut penser qu'un "compromis" entre les deux plans serait la meilleure solution. Dans le présent article, nous construisons de tels plans mixtes, et nous montrons qu'il s'agit d'exemples de plans à chaînes de Markov, une vaste classe de méthodes de sélection d'une unité par strate dans une population finie. Ces plans comprennent, comme cas spéciaux, l'échantillonnage systématique, l'échantillonnage systématique compensé et l'échantillonnage aléatoire simple stratifié avec une unité d'échantillonnage par strate. Les probabilités d'inclusion de premier et de second ordre sont obtenues pour les plans à chaînes de Markov, ce qui donne l'estimateur de Horvitz-Thompson et sa variance. L'efficacité de l'estimateur de Horvitz-Thompson est évaluée à l'aide de modèles de superpopulation. Des exemples numériques montrent que les nouveaux plans examinés ici peuvent être plus efficaces que les plans habituels dans le cas de superpopulations qui sont la somme d'une tendance et d'erreurs autocorrélées. Nous présentons un exemple d'application de plans à chaînes de Markov, pour les fins du National Resources Inventory de 1992 en Alaska.

MOTS CLÉS: Échantillonnage systématique compensé; National Resources Inventory; échantillonnage systématique.

1. INTRODUCTION

L'échantillonnage stratifié, qui consiste à diviser une population finie en strates non chevauchantes et à tirer des échantillons de chaque strate, est une technique courante et efficace pour réduire l'erreur d'échantillonnage. En pratique, les plans à échantillonnage stratifié dans lesquels on prélève une seule unité d'échantillonnage par strate sont très répandus. Mentionnons par exemple l'échantillonnage aléatoire simple stratifié, ainsi que l'échantillonnage systématique et ses variantes (p. ex. Murthy et Rao 1988).

Les échantillons systématiques sont vulnérables aux erreurs systématiques. Dans les échantillonnages géographiques à grande échelle, par exemple, les routes, les lignes électriques, les systèmes d'irrigation, *etc.* peuvent être la cause d'erreurs systématiques. Pour illustrer ce danger, on cite souvent le cas des routes délimitant des sections ("section roads"), qui existent dans certaines régions des États-Unis couvertes par l'inventaire public des terres. Ce système à quadrillage comprend des parcelles carrées appelées sections, qui ont chacune un mille de côté et qui, souvent, sont bornées par des routes dans les régions agricoles du Midwest. Le prélèvement d'un échantillon systématique avec intervalle d'un mille, si l'origine choisie au hasard tombait au mauvais endroit, pourrait laisser croire que l'Iowa est entièrement couvert de routes de gravier!

L'échantillonnage systématique offre l'avantage de l'efficacité quand la population échantillonnée est positivement autocorrélée, ce qui est souvent le cas dans les problèmes d'échantillonnage temporel ou géographique, car il force les observations à être le plus possible éloignées les unes des autres, et donc le moins corrélées possible.

L'autocorrélation et l'erreur systématique sont deux sujets de préoccupation dans le National Resources Inventory (NRI), un échantillonnage aréolaire des terres non fédérales aux États-Unis, réalisé tous les cinq ans par le Soil Conservation Service du United States Department of Agriculture. Les éléments de données du NRI, recueillis à la fois par télédétection et par des observations au sol, comprennent les caractéristiques du sol, l'utilisation des terres, les pratiques agricoles, les mesures d'érosion, *etc.*

Le plan d'échantillonnage du NRI de 1992 pour la région nord-ouest de l'État de l'Alaska est une version contrôlée d'un échantillonnage à une unité par strate. La région a été divisée en bandes de latitude de vingt minutes. Chaque bande a été divisée en strates de 500,000 acres, et chacune de ces dernières a été divisée à son tour en une grille 10×10 de cellules indexées selon la latitude et la longitude. Une cellule par strate a été incluse dans l'échantillon. Dans chaque bande de vingt minutes, les strates étaient parcourues d'est en ouest pour la sélection des cellules. Les nombres aléatoires qui déterminaient les cellules sélectionnées dans le sens de la longitude, ainsi que les nombres aléatoires qui les déterminaient dans le sens de la latitude, évoluaient comme deux chaînes de Markov indépendantes. (Les résultats de base de la théorie des chaînes de Markov utilisés dans le présent article sont exposés dans des ouvrages d'introduction aux processus stochastiques, comme celui de Taylor et Karlin 1984). Les détails du plan sont présentés à la section 2.

[1] F. Jay Breidt, Iowa State University, Department of Statistics, Ames, IA 50011-1210, U.S.A.

Comment ce plan spécial se compare-t-il à des plans plus courants d'échantillonnage à une unité par strate? On constate, comme nous le décrivons à la section 2, qu'une application simple des chaînes de Markov permet de décrire une vaste classe de plans d'échantillonnage avec probabilités égales pour la sélection d'une unité par strate dans une population finie. Cette classe comprend des techniques courantes comme l'échantillonnage aléatoire simple stratifié, l'échantillonnage systématique et l'échantillonnage systématique compensé, ainsi que les plans appliqués à l'Alaska dont nous venons de parler. Il est facile, par ailleurs, de créer de nouveaux plans appartenant à cette classe. Ce traitement unifié des plans à une unité par strate permet de faire des comparaisons d'efficacité.

Les probabilités d'inclusion de premier et de second ordre pour tous ces plans sont obtenues à la section 3, ce qui donne l'estimateur de Horvitz-Thompson et sa variance. Comme dans bon nombre des publications sur le sujet (Madow et Madow 1944; Cochran 1946; Sedransk 1969; Bellhouse et Rao 1975; Wolter 1985; Bellhouse 1988; *etc.*), la moyenne de la variance liée au plan de l'estimateur de Horvitz-Thompson est évaluée pour divers modèles de superpopulation. Des expressions compactes des moyennes, pour un modèle, des variances liées au plan sont obtenues. Des exemples numériques présentés à la section 4 montrent que les plans proposés dans cet article peuvent être plus efficaces que les plans d'échantillonnage habituels à une unité par strate pour des superpopulations qui sont la somme d'une tendance et d'erreurs autocorrélées. Un bilan est présenté à la section 5.

Bien que l'exemple qui nous inspire soit bidimensionnel, ce sont des plans unidimensionnels qui seront examinés tout au long de l'article. La plupart des démonstrations et déductions sont directes et sont omises par souci de concision.

2. PLANS À CHAÎNES DE MARKOV

Considérons le problèmes du prélèvement d'un échantillon dans une population finie de $N = na$ unités étiquetées, désignées par

$$\begin{aligned} U &= \{1, \ldots, N\} \\ &= \{1, \ldots, a, a+1, \ldots, 2a, \ldots, \\ &\qquad (n-1)a+1, \ldots, na\}. \end{aligned}$$

La valeur d'une variable étudiée $y_k = y_{(i-1)a+j} = y_{ij}$ est associée à chaque étiquette k; la notation y_k ou y_{ij} sera utilisée aussi bien pour les variables aléatoires que pour les réalisations des variables aléatoires.

Ici, n est la taille de l'échantillon et a est l'*intervalle d'échantillonnage*. Les n sous-ensembles

$$\{(i-1)a+1, \ldots, (i-1)a+a\} \quad (i = 1, \ldots, n)$$

seront appelés les *strates*. Le but est de prélever une unité par strate. Souvent, on définit un plan d'échantillonnage stratifié comme un plan dans lequel des échantillons probabilistes indépendants sont prélevés dans chaque strate, mais la contrainte d'indépendance n'est pas utilisée ici.

Soit une matrice de probabilités de transition doublement stochastique P. On définit un *échantillon à chaîne de Markov* comme

$$s = \{R_1, a + R_2, \ldots, (n-1)a + R_n\},$$

où $R_1, \ldots, R_n$ est la chaîne de Markov définie par P et $R_1 \sim$ uniforme $(1, \ldots, a)$. Formellement, donc, un *plan à chaîne de Markov* (MC) est une fonction $p(\cdot\,;P)$ telle que

$$\begin{aligned} p(s;P) &= \Pr\{s = \{r_1, a + r_2, \ldots, (n-1)a + r_n\}\} \\ &= \Pr\{R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n\} \\ &= \begin{cases} P_{r_{n-1},r_n} P_{r_{n-2},r_{n-1}} \cdots P_{r_1,r_2}/a, \\ \qquad \text{pour} \quad r_1, \ldots, r_n \in \{1, \ldots, a\}, \\ 0, \quad \text{autrement.} \end{cases} \end{aligned}$$

Les plans MC définis dans le présent article sont reliés aux plans présentés dans Chandra, Sampath et Balasubramani (1992), dans lesquels un vecteur $1 \times N$ de probabilités initiales de sélection et une matrice $N \times N$ de probabilités de transition de périodicité n déterminent un plan d'échantillonnage sans remise. Chandra et coll. se concentrent sur la production de plans comportant des probabilités d'inclusion de second ordre strictement positives. Ils ne se penchent pas explicitement sur les plans à une unité par strate examinés ici, mais on peut directement intégrer ces derniers à leur structure, en construisant le vecteur de probabilités initiales et la matrice de probabilités de transition qui conviennent.

Le résultat suivant est utile à la détermination des caractéristiques probabilistes des plans MC.

Résultat 1 Considérons une chaîne de Markov pour laquelle la matrice de probabilités de transition P est doublement stochastique (c.-à-d. que la somme des entrées de toutes les lignes et de toutes les colonnes est égale à un) et R_1 a une distribution uniforme discrète, avec masse $1/a$ sur chacun des états $1, \ldots, a$. Alors, R_i a une distribution uniforme discrète sur les états $1, \ldots, a$ pour tout i. En particulier, R_i a comme moyenne $(a+1)/2$ et comme variance $V(R_i) = (a^2 - 1)/12$.

Certains cas spéciaux de plans MC sont intéressants.

Échantillonnage aléatoire simple stratifié. Si la matrice de probabilités de transition est

$$H = [1/a]_{j,j'=1}^{a},$$

alors

$$\Pr\{R_{i'} = j' \mid R_i = j\} = 1/a = \Pr\{R_{i'} = j'\}$$
$$(j, j' = 1, \ldots, a;\ i < i'),$$

ce qui, conjugué à la propriété de Markov, entraîne que $R_1, \ldots, R_n$ sont indépendants du point de vue probabiliste. Dans ce cas, le plan MC correspond à l'échantillonnage aléatoire simple stratifié à une unité par strate (ST).

Échantillonnage systématique. Si la matrice de probabilités de transition est I, c.-à-d. la matrice identité $a \times a$, on a

$$\Pr\{R_{i'} = j' \mid R_i = j\} = \begin{cases} 1, j = j', \\ 0, j \neq j', \end{cases}$$

de sorte qu'il existe un lien déterministe entre $R_1, \ldots, R_n$. Par conséquent,

$$s = \{R_1, a + R_1, \ldots, (n - 1)a + R_1\},$$

et le plan MC correspond à l'échantillonnage systématique (SY).

Plans mixtes. Intuitivement, on peut concevoir ST et SY comme étant, d'une certaine façon, deux "extrêmes" qui s'opposent. Si $\rho \in [0,1]$,

$$G_\rho = \rho H + (1 - \rho)I$$

est doublement stochastique. Si $\rho = 0$, le plan est SY et si $\rho = 1$, le plan est ST. Tout autre choix de ρ donnera une séquence formée de "bouts" d'échantillons systématiques. Par conséquent, la classe G_ρ comprend les plans ST et SY, ainsi qu'un ensemble continu de plans MC mixtes.

D'autres combinaisons convexes de matrices doublement stochastiques pourraient être envisagées. La classe des matrices doublement stochastiques est également fermée à l'égard de la multiplication et de la transposition des matrices, ainsi que de la permutation des lignes et des colonnes, de sorte qu'il existe de nombreuses façons de créer des plans MC.

Échantillonnage systématique compensé. Murthy (1967, §5.9d) décrit une méthode de prélèvement d'une unité par strate qu'il appelle *échantillonnage systématique compensé* ("balanced systematic sampling") (BA). Cette méthode donne les échantillons

$$s = \{R_1, a + (a + 1 - R_1), \ldots, (n - 2)a + R_1, (n - 1)a + (a + 1 - R_1)\}$$

si n est pair et

$$s = \{R_1, a + (a + 1 - R_1), \ldots, (n - 2)a + (a + 1 - R_1), (n - 1)a + R_1\}$$

si n est impair. Une caractéristique intéressante de ce plan est que si n est pair et que la population est parfaitement linéaire ($y_{ij} = \beta_0 + \beta_1[(i - 1)a + j]$) la moyenne de l'échantillon est égale à la moyenne de la population pour n'importe quel échantillon. Avec la matrice de probabilités de transition

$$J = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \end{pmatrix}_{a \times a},$$

BA est un plan MC.

Plan du NRI en Alaska. Comme il est indiqué à la section 1, le plan d'échantillonnage du NRI de 1992 pour la région nord-ouest de l'Alaska utilisait deux chaînes de Markov indépendantes pour la sélection contrôlée des cellules dans le sens de la latitude et dans le sens de la longitude. La matrice des probabilités de transition pour les cellules dans le sens de la longitude, P_{long}, est donnée au tableau 1. Ce plan, dénoté ci-après AK, est un plan MC, car P_{long} est doublement stochastique. La plupart des probabilités de transition sont voisines de 0.10, de sorte que la plupart des "déplacements" sont à peu près également probables. Notons toutefois que la masse a été réduite sur la diagonale inverse et à proximité, au profit des angles supérieur gauche et inférieur droit, afin que P_{long} défavorise les vastes déplacements est-ouest, par exemple de la cellule un à la cellule dix, ainsi que les courts déplacements, par exemple de la cellule dix à la cellule un. Par contre, P_{long} favorise les déplacements qui sont environ de longueur dix, par exemple de la cellule deux à la cellule un, deux ou trois. L'échantillon réalisé des cellules dans le sens de la longitude est donc bien réparti d'est en ouest, comme le serait un échantillon systématique, mais sa composante aléatoire offre une protection contre l'erreur systématique. De même, la chaîne de Markov régissant le prélèvement des cellules dans le sens de la latitude a été conçue de manière à produire une bonne dispersion géographique du nord au sud.

Tableau 1

Matrice de probabilités de transition pour l'échantillon à chaînes de Markov de cellules dans le sens de la longitude, pour le National Resources Inventory de 1992, en Alaska. Les entrées sont les probabilités conditionnelles de sélection de la cellule j' de la strate $i + 1$ quand la cellule j de la strate i a été sélectionnée.

Cellule j de la strate i	Cellule j' de la strate $i + 1$									
	1	2	3	4	5	6	7	8	9	10
1	0.05	0.15	0.15	0.15	0.15	0.15	0.10	0.10	0	0
2	0.15	0.15	0.15	0.10	0.10	0.10	0.10	0.10	0.05	0
3	0.15	0.15	0.10	0.10	0.10	0.10	0.05	0.05	0.10	0.10
4	0.15	0.10	0.10	0.10	0.10	0.10	0.10	0.05	0.10	0.10
5	0.15	0.10	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.15
6	0.15	0.10	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.15
7	0.10	0.10	0.05	0.10	0.10	0.10	0.10	0.10	0.10	0.15
8	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.10	0.15	0.15
9	0	0.05	0.10	0.10	0.10	0.10	0.10	0.15	0.15	0.15
10	0	0	0.10	0.10	0.15	0.15	0.15	0.15	0.15	0.05

3. ESTIMATION DE HORVITZ-THOMPSON EN VERTU D'UN PLAN MC

Écrivons le total de la population comme suit

$$t = \sum_U y_k = \sum_{i=1}^{n} \sum_{j=1}^{a} y_{(i-1)a+j} = \sum_{i=1}^{n} \sum_{j=1}^{a} y_{ij}.$$

Pour tout k les probabilités d'inclusion de premier ordre d'un plan MC sont données par

$$\pi_k = \Pr\{k \in s\} = \Pr\{R_i = j\} = 1/a$$

et pour $k \leq l$, les probabilités d'inclusion de second ordre sont données par

$$\pi_{kl} = \begin{cases} 1/a, & \text{pour} \quad i = i', j = j', \\ 0, & \text{pour} \quad i = i', j \neq j', \\ P_{jj'}^{(i'-i)}/a, & \text{pour} \quad i < i'. \end{cases}$$

L'estimateur de Horvitz-Thompson non biaisé selon le plan (Horvitz et Thompson 1952) pour le total de la population est alors

$$\hat{t}_\pi = \sum_s y_k/\pi_k = \sum_{i=1}^{n} \frac{y_{iR_i}}{1/a} = a \sum_{i=1}^{n} \sum_{j=1}^{a} y_{ij} I_{\{R_i=j\}},$$

où

$$I_{\{R_i=j\}} = \begin{cases} 1, & \text{si} \quad R_i = j, \\ 0, & \text{si} \quad R_i \neq j. \end{cases}$$

Les covariances liées au plan des indicateurs $I_{\{R_i=j\}}$ sont données par

$$\begin{aligned} C_{\mathrm{MC}}\left(I_{\{R_i=j\}}, I_{\{R_{i'}=j'\}}\right) &= E_{\mathrm{MC}}\left[I_{\{R_i=j\}} I_{\{R_{i'}=j'\}}\right] - \\ &\quad E_{\mathrm{MC}}\left[I_{\{R_i=j\}}\right] E_{\mathrm{MC}}\left[I_{\{R_{i'}=j'\}}\right] \\ &= \pi_{(i-1)a+j,(i'-1)a+j'} - \\ &\quad \pi_{(i-1)a+j}\, \pi_{(i'-1)a+j'}, \end{aligned}$$

de sorte que la variance liée au plan de $\hat{t}_\pi$ est

$$\begin{aligned} V_{\mathrm{MC}}(\hat{t}_\pi) &= a^2 \sum_{i=1}^{n} \sum_{j=1}^{a} \left(\frac{1}{a} - \frac{1}{a^2}\right) y_{ij} y_{ij} \qquad (1) \\ &+ a^2 \sum_{i=1}^{n} \sum_{j=1}^{a} \sum_{j' \neq j} \left[0 - \frac{1}{a^2}\right] y_{ij} y_{ij'} \\ &+ 2a^2 \sum_{i=1}^{n} \sum_{i'>i} \sum_{j=1}^{a} \left[\frac{P_{jj}^{(i'-i)}}{a} - \frac{1}{a^2}\right] y_{ij} y_{i'j} \\ &+ 2a^2 \sum_{i=1}^{n} \sum_{i'>i} \sum_{j=1}^{a} \sum_{j' \neq j} \left[\frac{P_{jj'}^{(i'-i)}}{a} - \frac{1}{a^2}\right] y_{ij} y_{i'j'}. \end{aligned}$$

Puisque la variance liée au plan dépend de toutes les valeurs de la variable étudiée dans la population finie, il n'est pas facile d'utiliser (1) pour comparer les plans. À l'exemple de Cochran (1946), supposons que les valeurs de la variable étudiée sont produites à partir du modèle de superpopulation

$$\xi : y_{ij} = \mu_{ij} + e_{ij},$$

où les μ_{ij} sont fixes et les e_{ij} sont des variables aléatoires pour lesquelles $E_\xi[e_{ij}] = 0$, $V_\xi(e_{ij}) = \sigma_{ij}^2$ et $C_\xi(e_{ij}, e_{i'j'}) = \sigma_{ij,i'j'}$. On peut alors utiliser, comme base de comparaison des plans, la moyenne pour le modèle de la variance liée au plan.

Proposition 1 En vertu du modèle de superpopulation ξ, la variance moyenne liée au plan de l'estimateur de Horvitz-Thompson est

$$\begin{aligned} E_\xi[V_{\mathrm{MC}}(\hat{t}_\pi)] &= a^2 V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right] + \\ &(a-1) \sum_{i=1}^{n} \sum_{j=1}^{a} \sigma_{ij}^2 - \sum_{i=1}^{n} \sum_{j=1}^{a} \sum_{j' \neq j} \sigma_{ij,ij'} \\ &+ 2a \sum_{i=1}^{n} \sum_{i'>i} \sum_{j=1}^{a} \sum_{j'=1}^{a} \sigma_{ij,i'j'} \left[P_{jj'}^{(i'-i)} - \frac{1}{a}\right] \end{aligned}$$

pour n'importe quel plan MC. Notons que si μ_{ij} est indépendant de j, on a $V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right] = 0$.

La proposition suivante donne une condition suffisante en vertu de laquelle aucun plan MC n'a une variance moyenne liée au plan pire que celle du plan SY.

Proposition 2 Considérons un modèle additif non corrélé,

$$\xi : y_{ij} = \mu_{ij} + e_{ij} = \alpha_i + \beta_j + e_{ij},$$

où $E_\xi[e_{ij}] = 0$, $V_\xi(e_{ij}) = \sigma_{ij}^2$ et $C_\xi(e_{ij}, e_{i'j'}) = 0$. Alors

$$E_\xi[V_{\mathrm{SY}}(\hat{t}_\pi)] \geq E_\xi[V_{\mathrm{MC}}(\hat{t}_\pi)]$$

pour tous les plans MC.

Preuve Il découle de la proposition 1 que le seul terme qui nous intéresse est $V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right]$ lequel en vertu d'un plan SY est

$$V_{\mathrm{SY}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right] = V_{\mathrm{SY}}\left[\sum_{i=1}^{n} \alpha_i + n\beta_{R_1}\right] = n^2 V(\beta_{R_1}),$$

tandis qu'en vertu d'un plan MC général, ce terme est

$$V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right] = \sum_{i=1}^{n} \sum_{i'=1}^{n} C_{\mathrm{MC}}(\beta_{R_i}, \beta_{R_{i'}}).$$

Puisque $C_{\mathrm{MC}}(\beta_{R_i}, \beta_{R_{i'}}) \leq V(\beta_{R_1})$, la proposition est démontrée. □

Certains modèles particuliers sont examinés dans les cinq sous-sections suivantes.

3.1 Modèle de permutation aléatoire

Dans le cas d'une population en ordre aléatoire, on peut appliquer un modèle de permutation, dans lequel une réalisation des mesures $y_1, \ldots, y_N$ est donnée par l'une des $N!$ permutations également probables de N valeurs fixes. Ce modèle peut s'écrire ainsi

$$\xi_1 : y_{ij} = \bar{y}_U + e_{ij},$$

où $\bar{y}_U = \sum_U y_k/N$. Voir Rao (1975) pour plus de détails. Le résultat suivant est alors une conséquence du théorème 2.1 de Rao et Bellhouse (1978).

Résultat 2 En vertu du modèle de permutation aléatoire,

$$E_{\xi_1}[V_{\mathrm{MC}}(\hat{t}_\pi)] = (N^2/n)(1 - n/N)\sum_U (y_k - \bar{y}_U)^2/(N-1)$$

pour n'importe quel plan MC.

Ainsi, la variance moyenne pour l'ensemble des permutations est exactement $V_{\mathrm{SI}}(\hat{t}_\pi)$, où SI dénote l'échantillonnage aléatoire simple (non stratifié) sans remise. Dans le cas de l'échantillonnage SY, on doit ce résultat originellement à Madow et Madow (1944). Voir aussi Sedransk (1969).

3.2 Modèle d'effets de stratification

Dans le cas d'une population comportant des effets de stratification, on peut appliquer le modèle suivant

$$\xi_2 : y_{ij} = \alpha_i + e_{ij},$$

où les α_i sont des constantes fixes et les e_{ij} sont des variables aléatoires non corrélées de moyenne zéro et de variance σ^2. Notons que si $\alpha_i \equiv \mu$, ξ_2 est une solution de remplacement à ξ_1 comme modèle d'une population en ordre aléatoire.

Résultat 3 En vertu d'un modèle d'effets de stratification,

$$E_{\xi_2}[V_{\mathrm{MC}}(\hat{t}_\pi)] = na(a-1)\sigma^2$$

pour n'importe quel plan MC.

3.3 Modèle de tendance linéaire

Dans le cas d'une population qui présente une tendance linéaire, on peut appliquer le modèle suivant

$$\xi_3 : y_{ij} = \beta_0 + \beta_1[(i-1)a + j] + e_{ij},$$

où β_0 et β_1 sont des constantes fixes et les e_{ij} sont des variables aléatoires non corrélées $(0,\sigma^2)$.

Résultat 4 En vertu du modèle de tendance linéaire ξ_3,

$$E_{\xi_3}[V_{\mathrm{MC}}(\hat{t}_\pi)] = \beta_1^2 a^2 V_{\mathrm{MC}}\left[\sum_{i=1}^{n} R_i\right] + na(a-1)\sigma^2 \quad (2)$$

pour n'importe quel plan MC. Puisque ξ_3 est additif, aucun plan MC n'a une variance probable supérieure à celle du plan SY en vertu d'un modèle de tendance linéaire.

Le seul terme lié au plan dans (2) est $V_{\mathrm{MC}}[\sum_{i=1}^{n} R_i]$. Selon le plan SY, $\sum_{i=1}^{n} R_i = nR_1$, de sorte que

$$V_{\mathrm{SY}}\left[\sum_{i=1}^{n} R_i\right] = n^2 V(R_1),$$

tandis que selon le plan ST,

$$V_{\mathrm{ST}}\left[\sum_{i=1}^{n} R_i\right] = nV(R_1).$$

Selon le plan BA, si n est pair,

$$V_{\mathrm{BA}}\left[\sum_{i=1}^{n} R_i\right] = V_{\mathrm{BA}}\left[\frac{n}{2}R_1 + \frac{n}{2}(a+1-R_1)\right] = 0.$$

Il en découle que si la population est parfaitement linéaire $(\sigma^2 = 0)$,

$$E_{\xi_3}[V_{\mathrm{BA}}(\hat{t}_\pi)] = 0,$$

de sorte que $\hat{t}_\pi = t$ pour tous les échantillons, comme l'a signalé Murthy (1967, p. 165).

Résultat 5 En vertu du modèle de tendance linéaire ξ_3,

$$\begin{aligned} E_{\xi_3}[V_{\mathrm{BA}}(\hat{t}_\pi)] &\leq E_{\xi_3}[V_{\mathrm{ST}}(\hat{t}_\pi)] \\ &\leq E_{\xi_3}[V_{G_\rho}(\hat{t}_\pi)] \\ &\leq E_{\xi_3}[V_{\mathrm{SY}}(\hat{t}_\pi)] = \max_{\mathrm{MC}} E_{\xi_3}[V_{\mathrm{MC}}(\hat{t}_\pi)], \end{aligned} \quad (3)$$

où le terme central est monotone croissant à mesure que $\rho \in [0,1]$ décroît. Si n est pair, le terme de gauche de (3) est égal à $\min_{\mathrm{MC}} E_{\xi_3}[V_{\mathrm{MC}}(\hat{t}_\pi)]$.

3.4 Modèle de population périodique

Dans le cas d'une population affichant une périodicité déterministe (avec période égale à p), on peut appliquer le modèle simple de l'onde sinusoïdale

$$\xi_4 : y_{ij} = \alpha \sin\left\{\frac{2\pi}{p}[(i-1)a + j]\right\} + e_{ij},$$

où les e_{ij} sont des variables aléatoires non corrélées de moyenne zéro et de variance σ^2.

Résultat 6 En vertu du modèle de population périodique ξ_4,

$$E_{\xi_4}[V_{\mathrm{MC}}(\hat{t}_\pi)] = a^2\alpha^2 V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \sin\frac{2\pi}{p}[(i-1)a + R_i]\right] + na(a-1)\sigma^2$$

pour n'importe quel plan MC.

Désignons le modèle de l'onde sinusoïdale ξ_4 avec $p = a$ par ξ_{4a}. En vertu de ξ_{4a},

$$\sin\left\{\frac{2\pi}{p}[(i-1)a + j]\right\} = \sin\frac{2\pi j}{a},$$

de sorte que le modèle est additif et qu'aucun plan MC n'a une variance probable liée au plan supérieure à celle du plan SY. On voit donc que le plan SY ne convient pas à une population affichant une périodicité dont la période est égale à l'intervalle d'échantillonnage (Madow et Madow 1994). Ce résultat peut être généralisé comme suit.

Résultat 7 Si $\mu_{ij} \equiv \beta_j$ dans ξ, alors ξ est un modèle qui s'applique à une population affichant une périodicité déterministe, avec période égale à l'intervalle d'échantillonnage, a. Le modèle ξ est additif, de sorte qu'en vertu de ξ, aucun plan MC n'a une variance probable liée au plan supérieure à celle du plan SY.

3.5 Modèle autocorrélé

De nombreux auteurs, depuis Cochran (1946), ont comparé les plans ST et SY, ainsi que l'échantillonnage aléatoire simple, en vertu d'un modèle de superpopulation autocorrélé. Voir Bellhouse (1988, §4) pour un bilan.

Considérons le modèle d'autocorrélation suivant, dû à Cochran (1946):

$$\xi_5 : y_{ij} = \mu + e_{ij}$$

où $\sigma_{ij,i'j'} = \gamma[(i'-i)a + j' - j]$ pour $i' \geq i$.

Résultat 8 En vertu du modèle autocorrélé ξ_5,

$$E_{\xi_5}[V_{\mathrm{MC}}(\hat{t}_\pi)] = na(a-1)\gamma(0) - 2n\sum_{h=1}^{a-1}\gamma(h)(a-h) + 2a\sum_{h=1}^{n-1}\sum_{j=1}^{a}\sum_{j'=1}^{a}\gamma(ha + j' - j)(n-h)\left(P_{jj'}^{(h)} - \frac{1}{a}\right)$$

pour n'importe quel plan MC.

Résultat 9 Si, pour $h \geq 0$, $\gamma(h)$ est non négatif, non croissant et convexe, c.-à-d.

$$\gamma(h) \geq 0, \gamma(h) \geq \gamma(h+1) \quad \text{et}$$

$$\gamma(h+2) - 2\gamma(h+1) + \gamma(h) \geq 0,$$

alors $E_{\xi_5}[V_{\mathrm{SY}}(\hat{t}_\pi)] = \min_{\mathrm{MC}} E_{\xi_5}[V_{\mathrm{MC}}(\hat{t}_\pi)]$.

Ce résultat est un corollaire d'un théorème dû à Hájek (1959), donné comme théorème 4.1 dans Bellhouse (1988); Bellhouse a clarifié les conditions dans lesquelles le théorème est valable. Le théorème de Hájek généralisait un résultat antérieur de Cochran (1946), qui comparait les plans SY et ST, ainsi que l'échantillonnage aléatoire simple.

4. EFFICACITÉ: QUELQUES EXEMPLES NUMÉRIQUES

Une importante classe de modèles s'appliquant à des processus temporels et spatiaux est décrite ainsi: tendance polynomiale d'ordre peu élevé, plus séquence d'erreurs autocorrélées. On peut donner, comme exemple simple,

$$\xi_{(\beta,\phi)} : y_{ij} = \beta_0 + \beta_1[(i-1)a + j] + e_{ij},$$

où la structure d'autocorrélation est celle du modèle autorégressif (AR) de premier ordre,

$$\sigma_{ij,i'j'} = \gamma[(i'-i)a + j' - j] = \sigma^2\phi^{(i'-i)a+j'-j}$$

pour $i' \geq i$ et $|\phi| < 1$. La variance moyenne liée au plan en vertu de ce modèle est obtenue des résultats 4 et 8. Pour différents choix de β_1 et ϕ, le ratio des variances probables liées au plan,

$$E_\xi[V_{\mathrm{MC}}(\hat{t}_\pi)]/E_\xi[V_{\mathrm{SY}}(\hat{t}_\pi)], \tag{4}$$

est donné au tableau 2 pour divers plans MC. Le tableau comprend également les valeurs pour le plan G_ρ optimal, obtenues en minimisant (4) par rapport à ρ. L'utilisation de ce plan n'est possible que si les paramètres de la superpopulation sont connus; il est donc présenté comme repère plutôt que comme choix possible.

Quand $\beta_1 \neq 0$ et $\phi = 0$, le modèle est ξ_3 et les valeurs du tableau concordent avec le résultat 5: le plan SY est le pire plan MC et le plan BA est le meilleur, tandis que les plans $G_{1/3}$, $G_{2/3}$ et ST se situent entre les deux. Bien que le plan BA soit extrêmement performant pour ce modèle, tout plan MC autre que le plan SY serait un bon choix.

Tableau 2

Le ratio de la variance probable liée au plan en vertu d'un plan MC et de la variance probable liée au plan en vertu du plan SY pour une superpopulation constituée d'une tendance (droite de pente β_1 ou onde sinusoïdale avec période a et amplitude α) plus des erreurs autorégressives (AR) ($N = 1{,}000$, $\sigma^2 = 100$, $a = 10$). Ici, G_{ρ^*} est le plan mixte optimal, où ρ^* est une fonction des paramètres de la superpopulation. Le ratio du meilleur plan réalisable dans chaque ligne (si ce n'est pas SY) est en italique.

Modèle	ϕ	Plans à chaînes de Markov						
		$G_{1/3}$	$G_{2/3}$	ST	BA	AK	G_{ρ^*}	(ρ^*)
Droite + AR $\beta_1 = 0.7$	−0.5	0.2322	0.2085	0.2001	*0.1666*	0.2056	0.2001	(1.0000)
	0.0	0.2220	0.1983	0.1903	*0.1821*	0.1957	0.1903	(1.0000)
	0.1	0.2187	0.1950	0.1871	*0.1825*	0.1921	0.1871	(1.0000)
	0.5	0.1922	0.1702	*0.1645*	0.1754	0.1659	0.1645	(1.0000)
	0.9	0.0980	0.0778	*0.0742*	0.0768	0.0762	0.0742	(1.0000)
Droite + AR $\beta_1 = 0.4$	−0.5	0.4504	0.4328	0.4262	*0.3647*	0.4304	0.4262	(1.0000)
	0.0	0.4344	0.4172	0.4114	*0.4054*	0.4153	0.4114	(1.0000)
	0.1	0.4291	0.4121	*0.4065*	0.4085	0.4094	0.4065	(1.0000)
	0.5	0.3853	0.3727	0.3724	0.4116	*0.3667*	0.3719	(0.8320)
	0.9	0.1876	*0.1835*	0.1914	0.2170	0.1848	0.1821	(0.5223)
Droite + AR $\beta_1 = 0.1$	−0.5	0.9233	0.9190	0.9163	*0.7941*	0.9175	0.9163	(1.0000)
	0.0	0.9201	0.9177	0.9169	*0.9160*	0.9174	0.9169	(1.0000)
	0.1	0.9191	0.9175	0.9175	0.9349	*0.9156*	0.9174	(0.8156)
	0.5	*0.9160*	0.9289	0.9439	1.0606	0.9185	0.9135	(0.1997)
	0.9	*0.8621*	0.9787	1.0725	1.2710	1.0017	0.7888	(0.0981)
AR pur	−0.5	0.9978	0.9956	0.9935	*0.8617*	0.9942	0.9935	(1.0000)
	0.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	(---)
	0.1	1.0009	1.0019	1.0028	1.0228	1.0001	1.0000	(0.0000)
	0.5	1.0179	1.0357	1.0536	1.1852	1.0245	1.0000	(0.0000)
	0.9	1.2517	1.4380	1.5814	1.8798	1.4734	1.0000	(0.0000)
Sinus + AR $\alpha = 0.1$	−0.5	0.9929	0.9906	0.9884	*0.8578*	0.9892	0.9884	(1.0000)
	0.0	0.9947	0.9946	*0.9945*	0.9950	0.9946	0.9945	(1.0000)
	0.1	0.9955	0.9963	0.9972	1.0175	*0.9945*	0.9954	(0.1925)
	0.5	1.0110	1.0285	1.0462	1.1775	1.0173	0.9977	(0.0364)
	0.9	1.2178	1.3980	1.5371	1.8294	1.4322	0.9999	(0.0018)
Sinus + AR $\alpha = 1.0$	−0.5	0.6747	0.6634	0.6586	*0.6008*	0.6604	0.6586	(1.0000)
	0.0	0.6603	0.6499	*0.6464*	0.6770	0.6477	0.6464	(1.0000)
	0.1	0.6554	0.6455	0.6425	0.6863	*0.6421*	0.6425	(1.0000)
	0.5	0.6149	0.6133	0.6196	0.7320	*0.6041*	0.6121	(0.5079)
	0.9	*0.3570*	0.3832	0.4126	0.5527	0.3877	0.3560	(0.2852)
Sinus + AR $\alpha = 10.0$	−0.5	0.0668	0.0384	*0.0287*	0.1101	0.0323	0.0287	(1.0000)
	0.0	0.0656	0.0372	*0.0275*	0.1115	0.0311	0.0275	(1.0000)
	0.1	0.0652	0.0368	*0.0271*	0.1115	0.0307	0.0271	(1.0000)
	0.5	0.0622	0.0339	*0.0245*	0.1106	0.0277	0.0245	(1.0000)
	0.9	0.0529	0.0247	*0.0154*	0.1016	0.0187	0.0154	(1.0000)

Quand $\beta_1 = 0$ et $\phi \neq 0$, $\xi_{(\beta,\phi)}$ est un cas spécial du modèle ξ_5. Pour $\phi > 0$, on peut conclure aussi bien du résultat 9 que des données du tableau que le plan SY est le plus efficace, car il produit l'échantillon le plus "étalé" possible; toutefois, dans le cas d'une autocorrélation faible, les autres plans MC sont des solutions valables. Le plan BA est très médiocre pour ce modèle, car le plan assure que les éléments d'une paire sur deux R_i, R_{i+1} ne seront pas éloignés de plus de a unités. (Pour la même raison, BA convient bien à une population à autocorrélation négative.) Les plans AK, $G_{1/3}$ et $G_{2/3}$ l'emportent sur le plan ST, car chacun de ces plans favorise des transitions entre états de longueur approximativement égale à a.

On obtient des résultats semblables pour le modèle de superpopulation

$$\xi_{(\alpha,\phi)} : y_{ij} = \alpha \sin \frac{2\pi j}{a} + e_{ij},$$

où $\sigma_{ij,i'j'}$ est défini comme ci-dessus. Le tableau 2 donne le ratio des variances probables du plan (4) en vertu de ce modèle, qui découle des résultats 6 et 8.

Quand $\alpha \neq 0$ et $\phi = 0$, le modèle est ξ_{4a} et le plan SY n'est pas efficace, comme l'indique le résultat 7. Même pour $\phi \neq 0$, le plan SY montre une bonne efficacité seulement lorsque la périodicité est submergée par un bruit hautement corrélé.

Notons que le tableau 2 ne montre aucun plan dominant: chacun des plans SY, $G_{1/3}$, $G_{2/3}$, ST, BA et AK l'emporte au moins une fois sur les autres. Dans le cas d'une tendance modérée et d'une autocorrélation élevée, les plans AK, $G_{1/3}$ et $G_{2/3}$ peuvent être supérieurs aux plans MC courants. Dans l'ensemble, le tableau 2 laisse croire que certains plans MC non courants, comme $G_{2/3}$ et AK, conviennent assez bien à toute une gamme de populations: ils conservent une bonne part de l'efficacité de SY pour une population autocorrélée, tout en étant prémunis contre les effets systématiques pour d'autres types de populations.

5. DISCUSSION

La classe des plans à chaînes de Markov a été définie, et nous avons vu qu'elle incluait, comme cas spéciaux, l'échantillonnage systématique, l'échantillonnage aléatoire simple stratifié et l'échantillonnage systématique compensé. Certains nouveaux plans ont été présentés (G_ρ, AK), et nous avons montré que leur efficacité se comparait à celle des plans courants d'échantillonnage à une unité par strate en vertu de divers modèles de superpopulation. Les nouveaux plans fonctionnent bien, notamment, dans des exemples numériques reposant sur des superpopulations constituées d'une tendance et d'erreurs autocorrélées. C'est le genre de population auquel on s'intéresse dans de nombreux problèmes d'échantillonnage géographique, comme le National Resources Inventory de 1992 en Alaska. Un plan MC bidimensionnel mis en oeuvre pour cette enquête montre que les plans MC unidimensionnels pourraient être utilement étendus au contexte de l'échantillonnage géographique, bien que des recherches additionnelles s'imposent dans ce domaine.

D'autres travaux sur l'estimation de la variance liée aux plans MC sont également nécessaires. Comme il s'agit de plans d'échantillonnage à une unité par strate, l'estimation sans biais lié au plan de la variance de l'estimateur de Horvitz-Thompson n'est pas possible. Le problème de l'estimation de la variance pour les plans à une unité par strate, en particulier le plan SY, a fait l'objet d'abondants travaux. Par exemple, Wolter (1985) examine en détail huit estimateurs biaisés de la variance pour SY et évalue leurs biais en vertu de modèles de superpopulation. Des travaux semblables, portant sur l'estimateur de la variance de strates groupés (p. ex. Cochran 1977, p. 139) en vertu de plans MC généraux, sont en cours.

REMERCIEMENTS

Cette recherche a reçu l'appui du Soil Conservation Service du U.S. Department of Agriculture, en vertu du SCS Cooperative Agreement No. 68-3A75-2-64. Jeff Goebel (Soil Conservation Service) et Wayne Fuller (Iowa State University) ont élaboré le plan MC pour l'État de l'Alaska. L'auteur tient à remercier les arbitres anonymes pour les commentaires constructifs formulés à l'égard d'une version antérieure de cet article.

BIBLIOGRAPHIE

BELLHOUSE, D.R. (1988). Systematic sampling. Dans *Handbook of Statistics*. (Éds. P.R. Krishnaiah et C.R. Rao), Vol. 6. Amsterdam: North-Holland, 125-145.

BELLHOUSE, D.R., et RAO, J.N.K. (1975). Systematic sampling in the presence of a trend. *Biometrika*, 62, 694-697.

CHANDRA, K.S., SAMPATH, S., et BALASUBRAMANI, G.K. (1992). Markov sampling for finite populations. *Biometrika*, 79, 210-213.

COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3ième Éd. New York: Wiley.

HÁJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Časopis Pro Pěstování Matematiky*, 84, 387-423.

HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

MADOW, W.G., et MADOW, L.H. (1944). On the theory of systematic sampling, I. *Annals of Mathematical Statistics*, 15, 1-24.

MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

MURTHY, M.N., et RAO, T.J. (1988). Systematic sampling with illustrative examples. Dans *Handbook of Statistics*. (Éds. P.R. Krishnaiah et C.R. Rao), (Vol. 6). Amsterdam: North-Holland, 147-185.

RAO, J.N.K. (1975). On the foundations of survey sampling. Dans *A Survey of Statistical Design and Linear Models*. (Éd. J.N. Srivastava). Amsterdam: North-Holland, 489-505.

RAO, J.N.K., et BELLHOUSE, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*, 2,125-141.

SEDRANSK, J. (1969). Some elementary properties of systematic sampling. *Skandinavisk Aktuarietidskrift*, 1-2, 39-47.

TAYLOR, H.M., et KARLIN, S. (1984). *An Introduction to Stochastic Modeling*. Orlando: Academic Press.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.