Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Une approche arborescente de la formation de strates dans les enquêtes-entreprises polyvalentes

par Roberto Benedetti, Giuseppe Espa et Giovanni Lafratta

Décembre 2008

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 217-226
Statistique Canada, N° 12-001-X au catalogue

Une approche arborescente de la formation de strates dans les enquêtes-entreprises polyvalentes

Roberto Benedetti, Giuseppe Espa et Giovanni Lafratta [1]

Résumé

Pour concevoir un échantillon aléatoire simple stratifié sans remise à partir d'une population finie, il faut résoudre deux grandes questions : définir une règle de partition de la population en strates et répartir les unités d'échantillonnage entre les strates sélectionnées. Dans le présent article, nous examinons une stratégie arborescente en vue d'aborder conjointement ces deux questions quand l'enquête est polyvalente et que de l'information multivariée, quantitative ou qualitative, est disponible. Nous formons les strates à l'aide d'un algorithme divisif hiérarchique qui sélectionne des partitions de plus en plus fines en minimisant, à chaque étape, la répartition d'échantillon requise pour atteindre les niveaux de précision établis pour chaque variable étudiée. De cette façon, nous pouvons satisfaire un grand nombre de contraintes sans augmenter fortement la taille globale d'échantillon et sans écarter certaines variables sélectionnées pour la stratification ni diminuer le nombre de leurs intervalles de classe. En outre, l'algorithme a tendance à ne pas définir de strate vide ou presque vide, ce qui évite de devoir regrouper certaines strates. Nous avons appliqué la méthode au remaniement de l'Enquête sur la structure des exploitations agricoles en Italie. Les résultats indiquent que le gain d'efficacité réalisé en utilisant notre stratégie n'est pas trivial. Pour une taille d'échantillon donnée, cette méthode permet d'obtenir la précision requise en exploitant un nombre de strates qui est habituellement égal à une fraction très faible du nombre de strates disponibles quand on combine toutes les classes possibles provenant de n'importe quelle covariable.

Mots clés : Stratification multivariée ; répartition polyvalente optimale de l'échantillon ; enquête sur la structure des exploitations agricoles ; plan d'échantillonnage.

1. Introduction

De nombreuses enquêtes-entreprises s'appuient sur des méthodes d'échantillonnage stratifiées dans lesquelles un échantillonnage aléatoire simple sans remise est exécuté dans chaque strate (voir, par exemple, Sigman et Monsour 1995 et, pour les enquêtes sur les fermes, Vogel 1995). Habituellement, la base de sondage à partir de laquelle sont sélectionnées les unités est établie en utilisant des données administratives ou des données de recensement, représentées par une riche base de données sur des variables auxiliaires, qui sont chacune susceptibles d'être exploitées pour former les strates. De surcroît, les enquêtes de ce genre sont souvent polyvalentes et les estimations des multiples variables étudiées doivent atteindre des niveaux de précision donnés.

Satisfaire un aussi grand nombre de contraintes sans accroître considérablement la taille d'échantillon est un objectif généralement considéré comme strictement relié au choix du nombre de variables de stratification et de leurs intervalles de classe (Kish et Anderson 1978). Cette façon de voir tient au fait bien connu que des partitions plus fines de la population introduisent une plus grande quantité d'information utile pour réduire les variances des estimations ; par contre, leur application accroît le risque que les unités deviennent des « sauteuses » de strate.

Dénommons stratification *atomisée* celle obtenue en formant les strates par combinaison de toutes les classes possibles pour n'importe laquelle des covariables utilisées. Si le nombre correspondant de ces strates de base, ou atomes, excède un seuil donné imposé par des restrictions d'ordre pratique, il semble inévitable de devoir reconcevoir l'enquête en sélectionnant un nombre plus faible de variables de stratification ou en créant un moins grand nombre de classes pour chacune d'elles. Néanmoins, il convient de noter qu'un autre moyen d'obvier à cette situation indésirable est fondé sur l'argument suivant : nous pouvons vraiment interpréter la stratification atomisée comme une solution extrême du problème de formation des strates, puisque, entre le cas de l'absence de stratification et celui de l'utilisation de la stratification atomisée, il existe toute une gamme de possibilités de sélectionner une stratification dont les sous-populations peuvent être formées par des unions d'atomes.

Nous proposons, pour réaliser cette sélection, de définir un plan d'échantillonnage stratifié à structure arborescente. Nous formons les strates à l'aide d'un algorithme divisif hiérarchique qui sélectionne des partitions de plus en plus fines en minimisant, à chaque étape, la répartition d'échantillon requise pour atteindre les niveaux de précision fixés pour chaque variable étudiée. La procédure est séquentielle et détermine un chemin allant de la stratification nulle,

1. Roberto Benedetti, professeur, Département des sciences commerciales, statistiques, technologiques et environnementales (DASTA), Université « G. d'Annunzio », Viale Pindaro 42, Pescara, IT-65127. Courriel : benedett@unich.it ; Giuseppe Espa, professeur, Département d'économie, Université de Trente, Via Inama 5, Trente, IT-38100. Courriel : giuseppe.espa@economia.unitn.it ; Giovanni Lafratta, professeur adjoint, Département des sciences commerciales, statistiques, technologiques et environnementales (DASTA), Université « G. d'Annunzio », Viale Pindaro 42, Pescara, IT-65127. Courriel : giovanni.lafratta@unich.it.

c'est-à-dire celle dont la strate unique correspond à la population, à la stratification atomisée. À chaque étape, nous sélectionnons la variable qui doit être utilisée pour définir la nouvelle partition plus désagrégée : chaque strate de la partition courante est divisée sur une covariable, en utilisant tour à tour toutes les classes de cette dernière disponibles, et la covariable qui réduit le mieux la taille globale allouée est celle qui est choisie.

Bloch et Segal (1989) ont discuté de l'application des méthodes arborescentes de classification (voir, par exemple, Breiman, Friedman, Olshen et Stone 1984) à la formation des strates, mais ils s'intéressaient principalement à l'interprétation au niveau des strates des relations entre les covariables et une seule variable de résultat. Au contraire, nos règles de partition de la population sont axées directement sur la répartition optimale des unités d'échantillonnage dans les strates sélectionnées. Les méthodes classiques traitant le cas univarié (Dalenius et Hodges 1959 ; Singh 1971 ; Lavallée et Hidiroglou 1988 ; Hedlin 2000 ; Lu et Sitter 2002 ; Gunning et Horgan 2004 ; pour une revue voir Horgan 2006) ne peuvent pas être étendues facilement au cas où l'on cherche à exploiter plusieurs covariables pour la stratification. Par conséquent, les solutions proposées par ces auteurs ont peu d'utilité pratique si l'enquête est polyvalente et que l'on dispose d'information sur de multiples covariables. Dans un tel contexte, des méthodes visant à satisfaire un grand nombre de contraintes sur les erreurs lorsqu'on minimise la taille d'échantillon ont été proposées par Bethel (1985, 1989) et Chromy (1987). Valliant et Gentle (1997) ont également abordé le problème dans le cas de cadres d'échantillonnage à deux degrés. Pour une stratification donnée, nous choisissons d'appliquer la règle de répartition de Bethel et, par conséquent, la procédure consiste à sélectionner les partitions subséquentes en minimisant la fonction de coût de l'enquête correspondant aux stratifications constituées par les strates non partitionnées au moment de la minimisation et les sous-strates partitionnées disponibles.

Comme nous l'avons dit auparavant, dans le cadre général de la stratification multivariée, notre but est, à l'instar de Kish et Anderson (1978), d'utiliser certains résultats obtenus dans le domaine de l'échantillonnage stratifié en vue de répondre aux besoins des praticiens d'enquête. Ces derniers exécutent quotidiennement des enquêtes multivariées (plusieurs variables disponibles pour la stratification) et polyvalentes (plusieurs variables et de nombreuses autres statistiques sont les objectifs principaux des activités de l'enquête). Donc, l'objectif de notre approche est de permettre de combiner la stratification et la répartition de l'échantillon. Autrement dit, nous nous soucions du choix du nombre de variables de stratification, du nombre d'intervalles de classe pour chaque variable et de la répartition optimale de Bethel dans les strates. Étant donné ce choix, notre méthodologie ne peut être réduite à la solution classique du problème de stratification multivariée, c'est-à-dire le recours à des techniques multivariées, telles que l'analyse en grappes et l'analyse en composantes principales (voir, par exemple, Mulvey 1983, Pla 1991 et Jarque 1981). En fait, dans cette branche de la littérature, les auteurs n'utilisent pas (ou n'utilisent qu'indirectement) les variables d'intérêt et se servent uniquement des variables auxiliaires, et ils n'abordent pas la question de la répartition. Il serait encore plus difficile de justifier de réduire notre approche à celles passées en revue par Särndal, Swensson et Wretman (1992, sections 12.6 et 12.7) : les techniques présentées à la section 12.6 combinent la stratification et la répartition multivariée de l'échantillon, mais elles ne sont pas polyvalentes, tandis que les méthodes de la section 12.7 sont polyvalentes, mais sont basées sur des strates prédéterminées.

La présentation de l'article est la suivante. À la section 2, nous présentons la procédure que nous proposons pour le calcul des arbres de stratification. Nous décrivons en détail l'algorithme utilisé pour générer la suite de stratifications et nous montrons comment il peut être représenté sous forme d'un arbre de classification. Nous discutons aussi des critères d'arrêt, afin de déterminer leur incidence sur le nombre optimal de strates. À la section 3, nous examinons les moyens d'exploiter un arbre de stratification pour concevoir l'Enquête sur la structure des exploitations agricoles (ESEA) de l'Union européenne. Nous illustrons notre technique de stratification en produisant un ensemble arborescent de strates et de répartitions d'échantillon en nous servant d'un ensemble d'atomes de base définis au moyen de l'information multivariée recueillie durant le cinquième Recensement général de l'agriculture réalisé en Italie en l'an 2000. Enfin, à la section 4, nous présentons certaines conclusions, en nous concentrant sur les questions d'ordre pratique ayant trait à la formation de strates à l'aide d'arborescences et en exposant comment la procédure peut être utilisée pour mieux gérer les enquêtes polyvalentes basées sur des plans de sondage stratifiés.

2. Une méthode de création d'arbres de stratification multivariés

Considérons une population finie P de N unités sur lesquelles les variables $y_1, ..., y_g, ..., y_G$ doivent être étudiées pour estimer leurs totaux en se servant d'une stratification sur P, c'est-à-dire un ensemble F de sous-populations non vides H_F, appelées strates, qui partitionnent P. Notre problème est de savoir comment sélectionner F afin de minimiser la taille globale allouée de l'échantillon correspondant n_F de telle façon que, pour $g = 1, ..., G$, le coefficient de variation (CV_g) correspondant à la

g^{e} variable d'intérêt ne soit pas supérieur au niveau de précision souhaité, disons $\varepsilon_g > 0$.

Pour un F donné, cette minimisation est exécutée en calculant la règle de répartition de l'échantillon de Bethel (1985). Plus précisément, désignons par n_h, $h = 1, ..., H_F$, la part de l'échantillon dans la strate h. Le coût global d'enquête correspondant à F peut alors être donné par

$$f(n_1, ..., n_{H_F}) = c_F + \sum_{h=1}^{H_F} c_h n_h,$$

où c_F est un coût fixe indépendant de $\mathbf{n}_F = (n_1, ..., n_{H_F})'$, et c_h représente le coût d'échantillonnage d'une unité dans la strate h. De surcroît, soit Y_g le total dans P de la g^{e} variable de réponse, N_h la taille de la h^{e} strate de F et $S_{h,g}^2$ la variance de y_g dans la strate h. Alors, la g^{e} contrainte sur la précision requise peut s'exprimer :

$$(\text{CV}_g)^2 \le \varepsilon_g^2 \equiv \sum_{h=1}^{H_F} \frac{N_h^2 S_{h,g}^2}{n_h} - \sum_{h=1}^{H_F} N_h S_{h,g}^2 \le Y_g^2 \varepsilon_g^2$$

$$\equiv \sum_{h=1}^{H_F} \frac{N_h^2 S_{h,g}^2}{\left(Y_g^2 \varepsilon_g^2 + \sum_{h=1}^{H_F} N_h S_{h,g}^2\right) n_h} \le 1,$$

de sorte que, si nous considérons les quantités suivantes, que nous dénommons unités normalisées de précision,

$$\xi_{h,g} = N_h^2 S_{h,g}^2 \Big/ \left(Y_g^2 \varepsilon_g^2 + \sum_{h=1}^{H_F} N_h S_{h,g}^2\right),$$

le problème de la répartition optimale de F peut s'exprimer de la façon suivante :

$$\begin{array}{lll} \min & f(\mathbf{n}_F) & \\ \text{sous les contraintes} & \sum_{h=1}^{H_F} \xi_{h,g} / n_h \le 1, & g = 1, ..., G, \\ & 1 / n_h > 0, & h = 1, ..., H_F. \end{array}$$

Bethel (1989) a obtenu la solution de ce genre de problème, disons n_h^*, $h = 1, ..., H_F$, comme il suit :

$$1 / n_h^* =$$

$$\begin{cases} \sqrt{c_h} \Big/ \left(\sqrt{\sum_{g=1}^{G} \alpha_g^* \xi_{h,g}} \sum_{l=1}^{H_F} \sqrt{c_l \sum_{g=1}^{G} \alpha_g^* \xi_{l,g}} \right) & \text{si } \sum_{g=1}^{G} \alpha_g^* \xi_{h,g} > 0, \\ +\infty & \text{autrement,} \end{cases}$$

où $\alpha_g^* = \lambda_g / \sum_{g=1}^{G} \lambda_g$, et λ_g est le multiplicateur de Lagrange de la contrainte sur l'erreur maximale permise dans l'estimation de la g^{e} variable étudiée, et indique si la g^{e} contrainte est « active » dans la solution du problème de répartition (à savoir, si $\alpha_g^* = 0$, alors la contrainte n'est pas active). La répartition optimale globale correspondante est donc donnée en fixant $n_F = \sum_{h=1}^{H_F} n_h^*$.

Supposons maintenant que les totaux estimés et leurs variances sont connus pour un ensemble donné de $M > 1$ strates de base $A_1, ..., A_m, ..., A_M$, de sorte que nous pouvons nous appuyer sur deux matrices $M \times G$ de totaux $\mathbf{T} = (Y_{m,g})$ et de variances des estimations $\mathbf{V} = (S_{m,g}^2)$, respectivement, et sur les tailles N_m, $m = 1, ..., M$. La définition de ces strates, que nous nommerons dans la suite *atomes*, est basée sur un ensemble de covariables $X_1, ..., X_k, ..., X_K$ de la façon suivante. Soit $x_{i,k}$ la valeur de X_k mesurée sur l'unité $i \in P$, et considérons l'ensemble de valeurs observées distinctes pour X_k dans P, $\Xi_k = \{x \in \mathbb{R} : \exists i \in P : x = x_{i,k}\}$. Nous construisons $M = \prod_{k=1}^{K} |\Xi_k|$ atomes, un pour chaque vecteur $(a_{m,1}, ..., a_{m,K})$ dans le produit cartésien $\Xi = \otimes_{k=1}^{K} \Xi_k$ en fixant pour $m = 1, ..., M$

$$A_m = \bigcap_{k=1}^{K} A_{m,k},$$

où $A_{m,k} = \{i \in P : x_{i,k} = a_{m,k}\}$. Dans le cas où les covariables X_k sont continues, l'ensemble Ξ_k contiendra N atomes non vides. Comme l'algorithme est divisif hiérarchique, le nombre d'atomes finaux n'a aucune incidence sur les étapes de l'algorithme. Seule la phase initiale de construction de la statistique agrégée et, au plus, la répartition de la mémoire sont affectées. Notre expérience empirique donne à penser que le temps de calcul varie peu si la taille des atomes est égale à 1. Au contraire, les variables continues ou ordonnées accélèrent l'algorithme, car le nombre de partitions binaires possible est, si le nombre de valeurs est le même, beaucoup plus faible que dans le cas des variables catégoriques. Nous vérifions que cette construction produit une stratification : chaque unité de la population figure dans un atome, et un atome seulement. Afin d'illustrer nos définitions, référons-nous aux données présentées au tableau 1, où est décrit un simple exemple dans lequel il est supposé qu'un ensemble de $M = 9$ atomes (obtenus en exploitant $K = 2$ covariables possédant toutes deux trois valeurs distinctes, à savoir 1, 2 et 3) constitue la stratification de base pour étudier $G = 2$ variables, dont les totaux et les variances des estimations sont également donnés, ainsi que les tailles d'atome. Dans ce contexte, nous avons $\Xi_1 = \Xi_2 = \{1, 2, 3\}$ et, par exemple, A_8 est la sous-population dont les éléments i sont tels que $x_{i,1} = a_{8,1} \equiv 3$ et $x_{i,2} = a_{8,2} \equiv 2$.

Tableau 1
Exemple de données pour 9 atomes et 2 variables étudiés

Atomes				Variables étudiées			
Id	Définition		Tailles	Totaux		Variances	
m	$a_{m,1}$	$a_{m,2}$	N_m	$Y_{m,1}$	$Y_{m,2}$	$S_{m,1}^2$	$S_{m,2}^2$
1	1	1	1 000	10	10	16	25
2	1	2	1 000	10	10	16	4
3	1	3	1 000	10	10	16	4
4	2	1	1 000	10	10	16	25
5	2	2	1 000	10	10	16	4
6	2	3	1 000	10	10	16	4
7	3	1	1 000	10	10	4	25
8	3	2	1 000	10	10	4	16
9	3	3	1 000	10	10	4	16

La méthode que nous proposons produit une suite de stratifications qui peuvent être représentées comme un arbre de classification. Définissons le niveau l d'un nœud donné ν dans l'arbre comme étant le nombre d'arcs dans la chaîne (unique) raccordant le nœud ν au nœud racine, et désignons par r_l le nombre de nœuds partageant le même niveau l. Puisqu'un seul nœud sera divisé à chaque niveau, nous avons $r_l = l + 1$ pour chaque l. À chaque niveau $l \geq 0$, la méthode détermine une classe F_l de r_l sous-populations non vides en lesquelles P peut être partitionnée, en les plaçant dans une correspondance bijective avec les nœuds de niveau l. Les strates contenues dans F_l sont toutes des candidates à la partition sur toute covariable donnée X_k et, en suivant Bethel (1989), nous calculons la répartition d'échantillon qui minimise de manière optimale la fonction de coût d'enquête pour la stratification constituée des strates non divisées dans F_l et des deux sous-strates qui définissent la partition en question. La meilleure partition au niveau l est celle qui est la plus favorable en ce sens qu'elle réduit l'échantillon alloué, par rapport à celui caractérisant F_l, plus que toute autre partition possible sur n'importe laquelle des covariables utilisées. La répartition optimale correspondant à la stratification définie par cette partition la meilleure, indiquée par $n_{b,\,l+1}$, est adoptée comme taille d'échantillon optimale au niveau $l + 1$ et considérée comme une borne supérieure contraignant les répartitions au niveau suivant de classification. À l'étape de l'initialisation, nous fixons $F_0 = \{P\}$, dont la strate unique est donc équivalente à l'ensemble de la population, et la meilleure taille d'échantillon $n_{b,0}$ est calculée comme étant le maximum parmi les tailles optimales obtenues en tenant compte, séparément, de chaque niveau de précision individuel ε_g fixé au sujet de la g^{e} variable étudiée :

$$n_{b,0} = \max_{g = 1, \ldots, G} \frac{N^2 S_g^2}{Y_g^2 \varepsilon_g^2 + N S_g^2},$$

où Y_g est l'estimation du total pour y_g sur P et S_g^2 est la variance correspondante (voir, pour la répartition optimale avec un seul item, Cochran 1977, pages 97-106, et Särndal et coll. 1992, pages 104-109).

Quand $l > 0$, l'ensemble de strates F_{l-1} optimal à l'étape $l - 1$ est celui qui est analysé. La meilleure répartition d'échantillon à l'étape l, $n_{b,l}$ est fixée au départ égale à $n_{b,\,l-1}$ et, pour chaque strate $U \in F_{l-1}$ et chaque variable auxiliaire X_k, l'algorithme qui suit est exécuté. Soit A_U l'ensemble d'atomes contenus dans la strate courante U, de sorte que l'égalité $U = \bigcup A_U$ est vérifiée et soit $m(A)$ une fonction donnant l'indice attribué à tout atome $A (m(A) = m_0$ si, et uniquement si $A = A_{m_0})$; nous pouvons alors exprimer l'ensemble de valeurs pris par X_k pour les unités contenues dans tout atome de A_U comme il suit :

$$Q_k = \{q \in \mathbb{R} \colon \exists A \in A_U \colon\ q = a_{m(A),\,k}\}.$$

Si X_k est une variable ordonnée, pour chaque q dans Q_k autre que $\max(Q_k)$, la strate U est partitionnée en deux ensembles $U_1 = U_{q,1}$ et $U_2 = U_{q,2}$ de la façon suivante :

$$U_{q,1} = \bigcup \{A \in A_U \colon\ a_{m(A),\,k} \leq q\}$$

et $U_{q,2}$ est le complément relatif de $U_{q,1}$ dans U, c'est-à-dire l'ensemble de tous les $i \in U$ qui ne sont pas compris dans $U_{q,1}$:

$$U_{q,2} = U \setminus U_{q,1}.$$

Dans notre exemple, pour une strate U définie comme étant $A_1 \cup A_2 \cup A_8$, nous avons $A_U = \{A_1, A_2, A_8\}$, $Q_1 = \{1, 3\}$ et $Q_2 = \{1, 2\}$ (voir le tableau 1), de sorte que notre algorithme essaiera de diviser U en $U_1 = A_1 \cup A_2$ et $U_2 = A_8$ en utilisant X_1, et en $U_1 = A_1$ et $U_2 = A_2 \cup A_8$ en utilisant X_2. Si, au contraire, X_k n'est pas ordonnée, U est alors partitionné en les ensembles U_1 et U_2 pour chaque sous-ensemble approprié U_1 de U, avec $U_2 = U \setminus U_1$.

Nous avons donc une stratification candidate correspondante, à savoir

$$C = (F_{l-1} \setminus \{U\}) \cup \{U_1\} \cup \{U_2\},$$

qui inclut toutes les strates de F_{l-1} autres que U, et, en plus, U_1 et U_2. Pour chaque strate C dans l'ensemble C, les totaux estimés de Y_g, $g = 1, \ldots, G$,

$$Y_{C,g} = \sum_{A \in A_C} Y_{m(A),\,g},$$

et leurs variances correspondantes

$$S_{C,g}^2 = (N_C - 1)^{-1} \Big(\sum_{A \in A_C} (N_A - 1) S_{m(A),\,g}^2 + \sum_{A \in A_C} N_A (N_A^{-1} Y_{m(A),\,g} - N_C^{-1} Y_{C,g})^2 \Big),$$

sont calculés et la répartition d'échantillon n_C est donc obtenue en appliquant la règle de Bethel. Si $n_C < n_{b,l}$, alors la partition (U_1, U_2) devient la partition courante la meilleure, la meilleure stratification candidate C^* devient C et $n_{b,l}$ est mise à jour pour devenir n_C. De cette façon, la méthode divisive qui produit le meilleur résultat, c'est-à-dire la plus petite taille d'échantillon, est sélectionnée pour créer les strates optimales suivantes :

$$F_l = C^*.$$

Dans le cadre de notre exemple, pour les niveaux de précision $\varepsilon_1 = \varepsilon_2 = 0{,}1$, décrivons la partition optimale au

niveau $l = 1$, c'est-à-dire celle qui divise la population entière en deux strates. Si nous utilisons les données décrites au tableau 1, l'algorithme indique que la meilleure partition de $U = P$ est basée sur la variable X_2 et est obtenue en fixant

$$U_1 = \bigcup \{A \in A_P : a_{m(A),k} \leq 2\}$$

$$= A_1 \cup A_2 \cup A_4 \cup A_5 \cup A_7 \cup A_8$$

et, en conséquence, $U_2 = P \setminus U_1 = A_3 \cup A_6 \cup A_9$. Cette partition optimale est représentée à la figure 1, où sont indiquées en détail, pour chaque strate, la taille, la définition en termes d'atomes inclus, la répartition courante et les statistiques d'estimation.

Les questions concernant le nombre optimal de strates sont prises en compte en définissant le critères d'arrêt de la procédure de création de l'arborescence. Nous décidons d'arrêter l'algorithme si la différence relative entre la taille optimale d'échantillon au niveau courant et celle au niveau précédent est plus faible qu'un paramètre donné $\delta > 0$:

$$\delta > (n_{b,l-1} - n_{b,l}) / n_{b,l-1}. \qquad (1)$$

Comme l'algorithme de Bethel converge vers un vecteur dont l'étendue est $(0, +\infty)^{l+1}$, ses données d'entrée doivent être arrondies aux nombres entiers correspondants les plus proches vers l'infini ; par conséquent, surtout en présence d'un grand nombre de petites strates, une répartition donnée produira vraisemblablement une taille d'échantillon plus grande que la précédente. En outre, dans ce cas, nous avons décidé d'arrêter notre procédure. Pour éviter d'avoir des strates trop petites et, par conséquent, statistiquement instables, des règles supplémentaires peuvent être établies pour empêcher des désagrégations plus poussées des strates courantes si les sous-strates correspondantes ont une cardinalité plus faible qu'une taille de strate minimale prédéfinie. Il est également possible d'atténuer les complexités de la gestion de l'enquête en imposant un nombre maximal de strates.

Cette approche, qui consiste à exécuter une recherche exhaustive dans chaque partition, garantit que la stratification et la répartition correspondantes sont optimales, mais seulement conditionnellement aux partitions exécutées antérieurement. Nous savons que la monotonicité des solutions et l'optimalité conditionnelle de chaque sous-arborescence obtenue par partition récursive de chaque nœud sont des conditions nécessaires, mais non suffisantes pour qu'un arbre binaire soit optimal. Afin de s'assurer de l'optimalité globale, nous devons ajouter à ces conditions l'exigence qu'une stratification optimale en, disons, H strates ne puisse être obtenue que par partition d'un des nœuds de la stratification optimale en $H - 1$ strates. Autrement dit, nous devons supposer qu'une partition optimale en H strates est un sous-espace de la partition optimale en $H - 1$ strates, ce qui implique que partitionner une strate donnée ne modifiera pas la fonction objectif – c'est-à-dire la répartition – dans les $H - 1$ strates restantes. Cependant, cette hypothèse est rarement vérifiée dans les applications pratiques d'enquête, puisque la partition d'une strate induit habituellement une modification des répartitions optimales dans toutes les strates restantes non partitionnées.

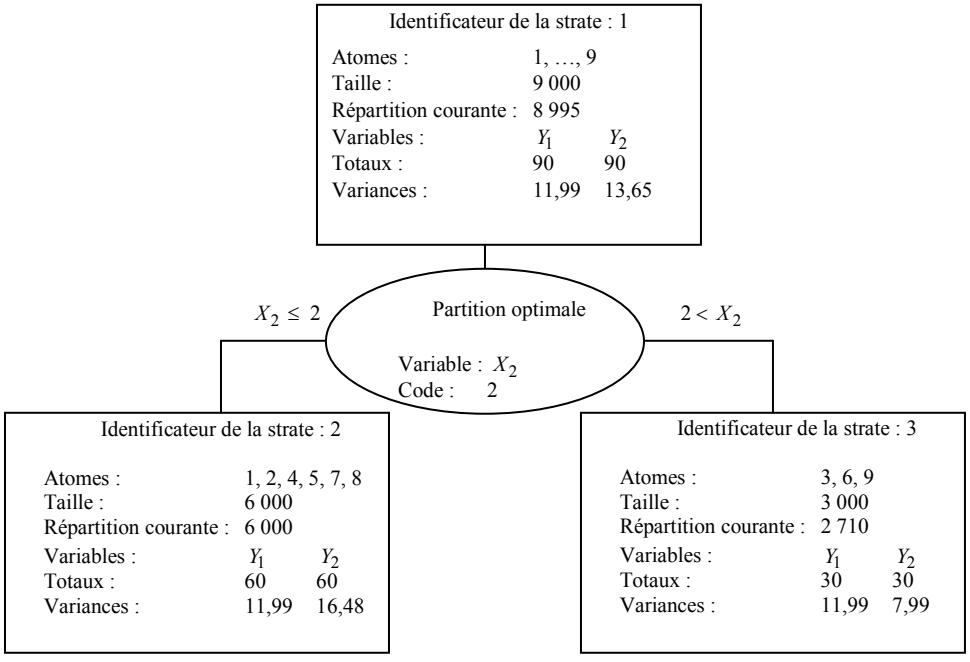

Figure 1 La première partition optimale pour les données utilisées comme exemple

L'algorithme que nous proposons, inspiré par la nature séquentielle et récursive des arbres binaires, peut être considéré comme une approche heuristique du problème de la stratification multivariée qui nous permet de déceler de bonnes strates, presque optimales, au prix d'un fardeau de calcul raisonnable. Par conséquent, cette méthode est efficace pour partitionner des populations en utilisant de grands ensembles de variables auxiliaires de stratification continues et qualitatives. En outre, la structure simple des arbres binaires offre une grande souplesse en ce qui concerne l'introduction de tout nombre de contraintes supplémentaires, telles que des bornes inférieures sur le nombre d'unités dans chaque strate.

3. Formation de strate pour l'Enquête sur la structure des exploitations agricoles en Italie

Conformément aux exigences des politiques agricoles de l'Union européenne, l'Enquête sur la structure des exploitations agricoles (ESEA) est exécutée tous les deux ans (règlement du Conseil (CEE) n° 70/66) en vue de recueillir des données sur des variables technoéconomiques caractéristiques des entreprises agricoles de l'UE. Elle représente la source principale d'information pour le projet EUROFARM (règlement du Conseil (CEE) n° 571/88), un ensemble de banques de données à utiliser pour traiter les enquêtes de l'UE sur la structure des exploitations agricoles. Les États membres sont tenus de prendre toutes les mesures appropriées pour exécuter l'ESEA sur leur territoire et sont libres de choisir un critère d'échantillonnage, mais le questionnaire et la précision requise, au niveau national, pour les estimations des variables étudiées sont déterminés par les règlements de la Communauté (voir les règlements CE n° 837/90 et n° 959/93, et les décisions 1998/377/CE et 2000/115/CE subséquentes de la Commission).

Afin d'illustrer notre méthode de stratification, nous exécutons l'algorithme décrit à la section 2 pour établir le plan de sondage de l'ESEA italienne et définir un ensemble arborescent de strates et de répartitions à l'aide d'information multivariée. Tous les algorithmes ont été implémentés par l'un des auteurs en langage MATLAB ; une application console Win32 a également été développée en C++ pour permettre des traitements par lot. Le plan de sondage exploite la liste des exploitations agricoles établies durant le cinquième Recensement général de l'agriculture tenu en Italie à l'automne 2000. ISTAT, l'Institut national de statistique de l'Italie, est responsable des mises à jour de cette liste par intégration des enregistrements administratifs, mais ces mises à jour n'étaient pas disponibles au moment de la rédaction de l'article. Pour initialiser la procédure, nous avons besoin d'un ensemble d'atomes dans lesquels la population d'exploitations agricoles italiennes doit être partitionnée. Nous obtenons cet ensemble de strates de base par agrégation des exploitations agricoles ayant en commun les mêmes classes pour sept covariables. Nous choisissons quatre variables associées à l'utilisation des terres et aux cheptels, à savoir la superficie agricole utile (SAU), le nombre de bovins (NB), le nombre de porcins (NP) et le nombre d'ovins et de caprins (NOC). Afin de tenir compte des caractéristiques géographiques des exploitations agricoles, nous avons également ajouté, comme variable de stratification, l'altitude de l'exploitation (ALT). Enfin, nous avons recueilli des renseignements au sujet de l'administration et de l'organisation de l'exploitation au moyen de deux variables nommées personnalité juridique de l'exploitant (PJ) et mode de tenure de l'exploitation (MT).

Les étendues des covariables concernant la structure de l'exploitation agricole sont réparties en quatre classes pour le nombre de bovins ($NB = 0$, $1 \le NB < 10$, $10 \le NB < 50$, $50 \le NB$), le nombre de porcins ($NP = 0$, $1 \le NP < 500$, $500 \le NP < 1{,}000$, $1{,}000 \le NP$) et le nombre d'ovins et de caprins ($NOC = 0$, $1 \le NOC < 250$, $250 \le NOC < 500$, $500 \le NOC$) et en sept classes pour la superficie agricole utile ($SAU = 0$, $0 < SAU < 1$, $1 \le SAU < 5$, $5 \le SAU < 10$, $10 \le SAU < 50$, $50 \le SAU < 100$, $SAU \ge 100$ ha). L'étendue des valeurs de l'altitude est divisée en cinq classes : montagnes intérieures, montagnes côtières, collines intérieures, collines côtières et plaines. Les classes pour la personnalité juridique de l'exploitant sont définies de manière à faire la distinction entre les titulaires exclusifs, les personnes morales (sociétés) et les groupes de personnes physiques (partenariat) dans un groupe, les entreprises coopératives, les associations d'exploitants, les institutions publiques et, enfin, les personnalités juridiques autres que celles qui précèdent (par exemple, consortia), que nous nommeront personnalités juridiques résiduelles. Les exploitations agricoles sont également stratifiées en tenant compte du mode de tenure, en faisant la distinction entre les propriétaires-exploitants (avec des sous-classes supplémentaires basées sur les catégories de travailleurs agricoles : membres de la famille, principalement membres de la famille, principalement non-membres de la famille), les locataires-exploitants, le partage des superficies agricoles exploitées et les modes de tenure autres que ceux qui précèdent. La combinaison de toutes les classes possibles pour chacune des covariables sélectionnées aboutit à 2 964 atomes non vides, qui constituent le point de départ de la procédure.

Nous avons étudié 12 variables d'utilisation des terres, dont la liste est présentée au tableau 2. Pour chaque variable étudiée, nous avons calculé les totaux et les variances dans chaque atome à l'aide des données de recensement disponibles, ce qui nous a permis d'exécuter l'algorithme de Bethel à chaque étape de notre procédure. Les paramètres supplémentaires requis pour définir nos critères d'arrêt sont

établis comme il suit. Le nombre maximal de strates est fixé à 300 et nous avons décidé d'interdire les strates dont la taille est inférieure à 10. Nous introduisons une tolérance quant à l'écart relatif entre les tailles d'échantillon optimales aux niveaux subséquents en fixant $\delta = 0$ dans l'équation (1), de sorte que l'algorithme s'arrête si $n_{b,l-1} < n_{b,l}$ pour un niveau donné $l \geq 0$.

Tableau 2
Variables étudiées dans l'Enquête sur la structure des exploitations agricoles en Italie et leurs niveaux de précision

Variable étudiée	CV requis Requis par l'ESEA	CV requis Réalisé par	
		stratification atomisée	arbre de stratification
Céréales	1,00	0,98	0,98
Vignes	3,00	1,38	1,38
Oliviers	3,00	1,11	1,11
Racines fourragères et herbacées	3,00	2,39	2,40
Plantes industrielles	3,00	2,22	2,23
Plantes fourragères	3,00	1,37	1,39
Légumes	3,00	3,03	3,03
Jachères	3,00	2,69	2,78
Nombre de bovins	1,00	0,99	1,00
Nombre de porcins	2,00	0,80	0,82
Nombre d'ovins	2,00	1,99	2,01
Nombre de caprins	2,00	1,92	1,98

Nous avons atteint la convergence puisque le nombre maximal de strates a été obtenu et qu'aucune règle d'arrêt n'a été activée pour $l < 300$. La figure 2 représente les répartitions optimales $n_{b,l}$ en fonction du nombre de strates r_l, $l = 1, ..., 300$ sur une échelle logarithmique, c'est-à-dire en fonction de $\log(n_{b,l})$. Il convient de souligner que l'écart relatif entre les répartitions subséquentes diminue rapidement, les dix premières partitions étant les plus importantes en ce qui concerne ce comportement : en fait, en fixant $\delta = 10\,\%$, la procédure atteindrait la convergence à l'étape $l = 7$.

La figure 3 donne un diagramme de l'arbre de stratification créé jusqu'au niveau 7. Afin d'optimiser la répartition globale, notre critère de partition a créé récursivement des strates de plus en plus petites. La première partition est faite sur la personnalité juridique de l'exploitant, PJ, et les atomes ont été inclus dans la strate fille de gauche si la classe de la variable PJ qu'ils supposaient était titulaire exclusif, institution publique ou PJ résiduelle. Cette partition est celle qui est optimale au niveau 1, puisqu'elle correspond à une partition de la population entière, la seule strate disponible au niveau 0, qui décrit le mieux la répartition de l'échantillon. Cette partition indique seulement que les exploitations agricoles dont l'exploitant est le titulaire exclusif se comportent différemment de celles gérées par des personnalités juridiques plus complexes, comme des sociétés, des partenariats, des associations ou des entreprises coopératives. La deuxième partition est faite sur le nombre de bovins, NB. Elle crée de nouvelles sous-strates de la strate 2 (voir la partie inférieure de la figure 3), à savoir les strates 4 et 5, de la façon suivante : la nouvelle strate 4 est définie comme l'union des atomes de la strate 2 pour lesquels la condition $NB > 10$ est vérifiée, tandis que la strate 5 est le complément relatif de la strate 4 dans la strate 2. De cette façon, l'algorithme détecte la meilleure diminution de la taille globale de l'échantillon (passant de 1 570 313 à 689 404 unités échantillonnées, voir la partie droite de la figure 3) en reconnaissant que les exploitations agricoles caractérisées par un cheptel bovin de taille moyenne ou grande doivent être traitées séparément pour celles dont l'exploitant est titulaire exclusif. La troisième partition est faite sur la superficie agricole utile, SAU. Ici, la strate 4 est partitionnée en une strate contenant les atomes pour lesquels la variable SAU est inférieure à 100 ha (strate 6) et une strate contenant les atomes restants (strate 7). Ces deux nouvelles strates sont à leur tour divisées, par étape successive, c'est-à-dire les étapes 4 à 7 (voir la partie gauche de la figure 3) sur les variables NP et NOC : plus précisément, la procédure suggère de faire la distinction entre les entreprises agricoles ne possédant pas de cheptel ovin et caprin ($NOC = 0$), ou caractérisées par un grand cheptel porcin ($NP \geq 500$).

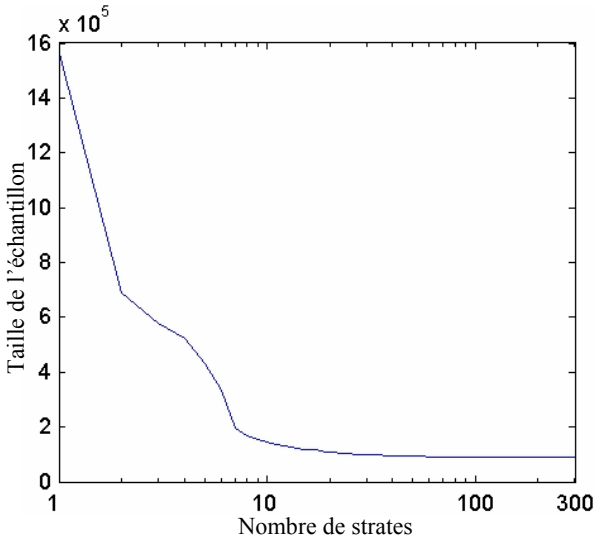

Figure 2 Tailles d'échantillon étape par étape. Les répartitions optimales $n_{b,l}$ sont montrées en fonction du nombre de strates r_l exploitées par le plan d'échantillonnage à structure arborescente aux étapes $l = 0, \ldots, 299$. Une échelle logarithmique est appliquée à l'axe horizontal, de sorte que $n_{b,l}$ est tracé en fonction de $\log(r_l)$. À mesure que le nombre de strates augmente, le plan de stratification à structure arborescente atteint ses objectifs en utilisant une taille globale d'échantillon rapidement décroissante, puisque la procédure améliore considérablement l'efficacité d'échantillonnage au cours de ses dix premières étapes d'exécution

Pour évaluer l'efficacité du plan d'échantillonnage à structure arborescente, nous calculons la meilleure répartition correspondant à la stratification atomisée, qui a déterminé un échantillon de 89 522 unités. En inspectant l'arbre de stratification, nous pouvons constater qu'une répartition globale fort semblable correspond à la meilleure stratification obtenue au niveau $l = 102$: en fait, pour cette partition de 103 strates, la taille d'échantillon est égale à 89 509. Cela signifie que, pour la même taille d'échantillon, notre algorithme atteint la précision requise pour l'enquête en exploitant un nombre de strates, 103, qui ne représente qu'une très petite fraction de 2 964, le nombre d'atomes disponibles, ce qui facilite l'organisation de l'enquête. Un autre avantage appréciable de notre procédure consiste à éviter les strates instables : il mérite d'être souligné que 1 618 des 2 964 atomes ont une taille égale ou inférieure à 5, alors que la taille minimale de n'importe laquelle des strates optimales au niveau 102 est de 16, si bien qu'il n'est pas nécessaire d'introduire des procédures de fusion de strates. Nous pouvons pousser plus loin les comparaisons et examiner les niveaux de précisions atteints en exécutant, respectivement, la stratification atomisée et l'arbre de stratification à l'étape 102. Ces niveaux, qui sont donnés au tableau 2, peuvent être considérés comme fort semblables pour les deux plans d'échantillonnage. En fait, nous constatons que, pour la stratification atomisée, la répartition de Bethel était activement contrainte sur la précision concernant trois variables étudiées, à savoir les céréales, les légumes et le nombre d'ovins. En ce qui concerne les strates correspondant au niveau 102 de l'arbre, les contraintes antérieures se trouvent aussi être actives, même si une autre contrainte, celle sur la variable Nombre de caprins a aussi donné des résultats serrés pour l'optimisation, les niveaux de précision atteints étant passés de 1,92 % à 1,98 %. Ces résultats donnent à penser qu'en ce qui concerne la partition atomisée, l'arbre peut être utilisé pour déceler une stratification plus compacte de la population, tout en préservant les niveaux de précision atteints et la taille globale d'échantillon.

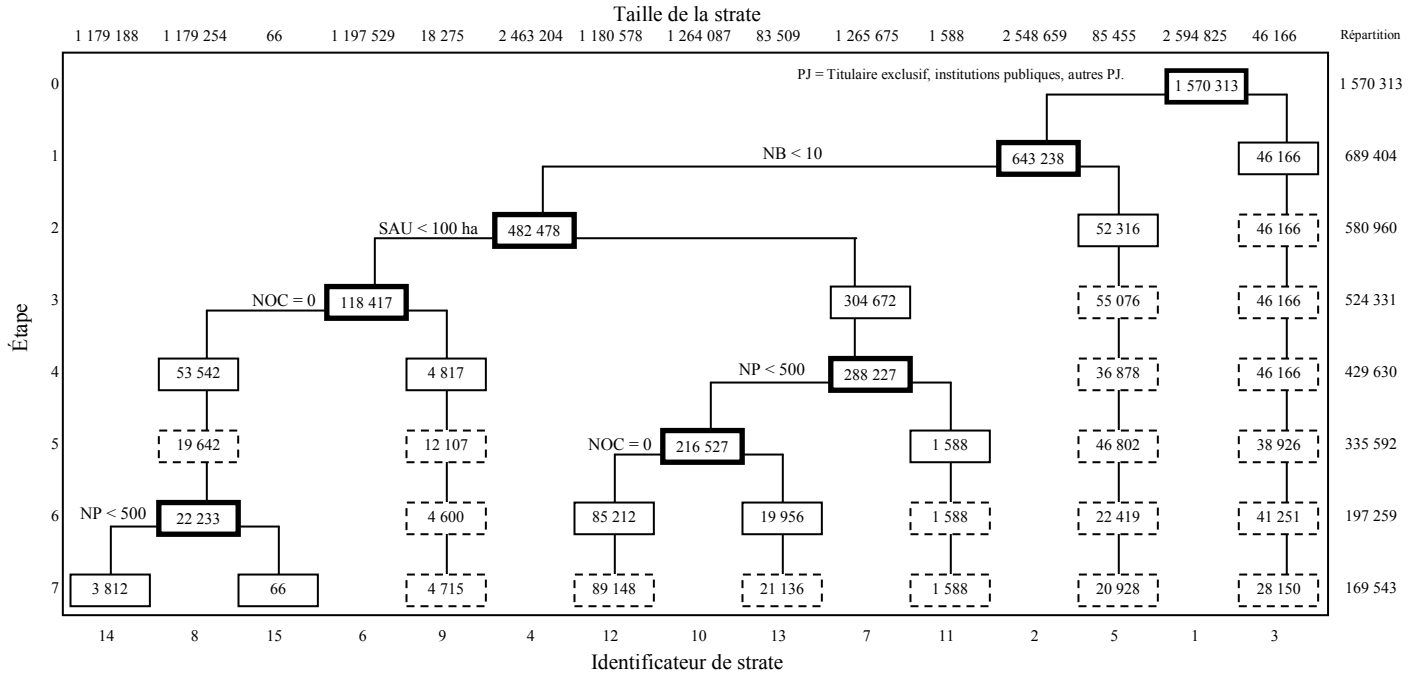

Figure 3 Diagramme de l'arbre de stratification. À la partie inférieure, l'axe horizontal représente l'identificateur de strate, un nombre qui indique de manière unique la sous-population correspondante dans la procédure de stratification. Les tailles de ces strates sont données à la partie supérieure de la figure. À la partie gauche, l'axe vertical représente la suite d'étapes allant de 0 à 7, tandis qu'à la partie droite, il représente la répartition globale optimale correspondant à ces étapes. Les blocs à bordure double représentent des strates partitionnées. Les strates filles sont reliées à leurs parents par des lignes coudées, et, quand elles ne sont plus partitionnées davantage aux étapes subséquentes, elles sont représentées comme des blocs à bordure simple. Pour les strates filles de gauche, la covariable sur laquelle la partition a été faite et la condition qu'elle a satisfait en définissant les sous-strates de gauche sont indiquées au-dessus de la ligne coudée correspondante. Le nombre à l'intérieur d'un bloc donné est la part de l'échantillon que la procédure attribue à la strate correspondante durant l'étape à laquelle le bloc est positionné. Puisqu'une strate peut demeurer non partitionnée durant des étapes qui suivent celle à laquelle elle a été créée, que sa répartition d'échantillon peut varier d'une étape à l'autre, des blocs en traits interrompus sont utilisés pour indiquer les modifications des tailles d'échantillon de strate

4. Conclusion

La stratégie arborescente pour les enquêtes polyvalentes examinées dans le présent article est planifiée de manière à définir conjointement une règle pour partitionner la population et pour répartir les unités d'échantillonnage entre les strates formées en exploitant l'information multivariée, quantitative ou qualitative. Un algorithme divisif hiérarchique sélectionne des partitions plus fines en minimisant, à chaque étape, la répartition de l'échantillon nécessaire pour atteindre les niveaux de précision requis. De cette façon, de grands nombres de contraintes peuvent être satisfaites sans accroître considérablement le nombre de strates. En outre, les variables choisies pour la stratification ne sont pas écartées simplement sur la base de considérations pratiques et le nombre de leurs intervalles de classe n'est pas diminué. Qui plus est, l'algorithme évite de créer des strates vides ou presque vides, si bien qu'il n'est pas nécessaire de procéder à des agrégations après la stratification en vue de mieux évaluer les variances d'estimation dans les strates.

Néanmoins, notre proposition peut susciter certaines critiques. Théoriquement, notre méthode ne peut pas être considérée comme une généralisation multiréponse de la méthode bien connue de l'arbre de régression et de classification, dont le but est d'exploiter la relation entre les covariables et une variable de résultat unique. En fait, même si nous traitons des enquêtes polyvalentes, notre approche consiste à partitionner l'information disponible de façon à optimiser uniquement une variable, à savoir la répartition d'échantillon dans les strates. De plus, la stratégie d'échantillonnage obtenue par notre méthode ne représente pas nécessairement un optimum global : en fait, la procédure constitue un algorithme de sélection ascendante de strates et, par conséquent, la recherche de l'optimalité à une étape donnée est conditionnée sur la stratification en cours d'utilisation, c'est-à-dire celle basée sur les partitions exécutées antérieurement. Il n'existe aucune garantie que la stratification sélectionnée par la procédure à une certaine étape l sera celle qui est optimale, même uniquement parmi toutes les partitions possibles dans les $l+1$ sous-ensembles de la population. Dans certaines situations, d'autres méthodes, comme la programmation dynamique, peuvent être utilisées pour exécuter une recherche exhaustive efficace de la stratification globalement optimale (voir Bühler et Deutler 1975, et Lavallée 1988).

Nous avons appliqué notre méthode au remaniement de l'Enquête sur la structure des exploitations agricoles en Italie. Les résultats indiquent que notre stratégie donne lieu à des gains d'efficacité : pour une taille donnée d'échantillon, notre méthode permet d'atteindre la précision requise en exploitant un nombre de strates qui est habituellement égal à une très petite fraction du nombre de strates disponibles quand sont combinées toutes les classes possibles pour n'importe laquelle des covariables. En outre, si l'on permet de grands nombres de strates, l'algorithme détecte d'autres stratégies d'échantillonnage pour lesquelles les contraintes sont satisfaites avec des tailles d'échantillon plus faibles que celle qui correspond à la stratification atomisée. Le choix de l'échantillonnage final dépend manifestement de la fonction de coût global de l'enquête. À cette fin, il est possible de recourir aux arbres de stratification pour tenir compte du fait qu'un nombre croissant de strates implique habituellement un accroissement des coûts dû à des problèmes d'organisation de l'enquête, mais qu'il correspond aussi à des tailles d'échantillon plus petites, qui donnent lieu à une réduction des coûts unitaires. La formation des strates par une méthode arborescente peut donc faciliter la gestion de l'enquête, en tant qu'outil d'aide à la sélection du plan d'échantillonnage stratifié qui convient le mieux à la collecte d'information au sujet du phénomène multivarié étudié.

Remerciements

Les auteurs remercient un rédacteur associé et deux examinateurs de leurs commentaires constructifs qui leur ont permis d'améliorer considérablement l'article.

Bibliographie

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. Dans *Proceedings of the Surveys Research Methods Section*, American Statistical Association, 209-212.

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête*, 15, 49-60.

Bloch, D.A., et Segal, M.R. (1989). Empirical comparison of approaches to forming strata – Using classification trees to adjust for covariates. *Journal of the American Statistical Association*, 84, 408, 897-905.

Breiman, L., Friedman, J.H., Olshen, R.A. et Stone, C.J. (1984). *Classification and Regression Trees*, Belmont, CA : Wadsworth International Group.

Bühler, W., et Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika*, 22, 161-175.

Chromy, J. (1987). Design optimization with multiple objectives. Dans *Proceedings of the Surveys Research Methods Section*, American Statistical Association, 194-199.

Cochran, W.G. (1977). *Sampling Techniques*. New York : John Wiley & Sons, Inc.

Dalenius, T., et Hodges, J.L. Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*. 54, 285, 88-101.

Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30, 2, 177-185.

Hedlin, D. (2000). A procedure for stratification by the extended ekman rule. *Journal of Official Statistics*, 16, 1, 15-29.

Horgan, J.M. (2006). Stratification of skewed populations: A review. *Revue Internationale de Statistique*. 74, 1, 67–76.

Jarque, C.M. (1981). A solution to the problem of optimum stratification in multivariate sampling. *Applied Statistics*. 30, 2, 163-169.

Kish, L., et Anderson, D.W. (1978). Multivariate and multipurpose stratification. *Journal of the American Statistical Association*, 73, 361, 24-34.

Lavallée, P. (1988). Two-way optimal stratification using dynamic programming. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Virginie, août 1988, 646-651.

Lavallée, P., et Hidiroglou, M.A. (1988). Sur la stratification de population asymétriques. *Techniques d'enquête*, 14, 35-45.

Lu, W., et Sitter, R.R. (2002). Méthode pratique de stratification multiple par programmation linéaire. *Techniques d'enquête*, 28, 2, 215-224.

Mulvey, J.M. (1983). Multivariate stratified sampling by optimization. *Management Science*, 29, 6, 715-724.

Pla, L. (1991). Determining stratum boundaries with multivariate real data. *Biometrics*, 47, 4, 1409-1422.

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

Sigman, R.S., et Monsour, N.J. (1995). Selecting samples from list frames of businesses. Dans *Business Survey Methods*, (Éds. B.G. Cox, D.A. Binder, B. Nanjamma Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott). New York : John Wiley & Sons, Inc., 133-152.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variable. *Journal of the American Statistical Association*, 66, 336, 829-833.

Valliant, R., et Gentle, J.E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, 25, 337-360.

Vogel, F.A. (1995). The evolution and development of agricultural statistics at the United States department of agriculture. *Journal of Official Statistics*, 11, 2, 161-180.