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

Article

Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation

par Yves Tillé

Décembre 2011

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2011
Vol. 37, N° 2, pp. 233-246
Statistique Canada, N° 12-001-X au catalogue

Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation

Yves Tillé [1]

Résumé

Le présent article propose un examen et une évaluation de l'échantillonnage équilibré par la méthode du cube. Il débute par une définition de la notion d'échantillon équilibré et d'échantillonnage équilibré, suivie par un court historique du concept d'équilibrage. Après un exposé succinct de la théorie de la méthode du cube, l'accent est mis sur les aspects pratiques de l'échantillonnage équilibré, c'est-à-dire l'intérêt de la méthode comparativement à d'autres méthodes d'échantillonnage et au calage, le domaine d'application, la précision de l'équilibrage, le choix des variables auxiliaires et les moyens de mettre la méthode en œuvre.

Mots clés : Échantillonnage ; équilibrage ; estimateur de Horvitz-Thompson.

1. Introduction

Bien que le concept d'échantillonnage équilibré existe depuis les touts débuts de la statistique d'enquête, son application a été difficile, parce que presque toutes les méthodes proposées étaient énumératives ou réjectives et que le temps de calcul était considérable. L'algorithme de la méthode du cube a été proposé en 1998 par Deville et Tillé, et une première implémentation a été écrite par trois étudiants de l'École Nationale de la Statistique et de l'Analyse de l'Information de Rennes, en France (voir Bousabaa, Lieber et Sirolli 1999). Finalement, la méthode a été publiée dans Tillé (2001) et dans Deville et Tillé (2004). Depuis, plusieurs implémentations de la méthode du cube ont été proposées et plusieurs gestionnaires d'enquête l'ont utilisée pour sélectionner des échantillons, les applications les plus importantes étant le nouveau recensement français et l'échantillon-maître français.

Notre objectif est d'évaluer le développement et l'utilisation de l'échantillonnage équilibré aux cours des dix dernières années afin de mieux déterminer quand et comment la méthode du cube peut être appliquée pour sélectionner les échantillons de ménages ou d'établissements. À la section 2, nous discutons du concept d'échantillon équilibré et d'échantillonnage équilibré. À la section 3, nous présentons une liste de cas particuliers. À la section 4, nous faisons succinctement l'historique de ce concept pour le cadre basé sur un modèle et celui basé sur le plan. Ensuite, à la section 5, nous donnons un bref aperçu de la méthode du cube, qui représente une classe d'algorithmes permettant de sélectionner aléatoirement des échantillons équilibrés en fixant les probabilités d'inclusion (voir Deville et Tillé 2004 ; Tillé 2001, 2006b). Nous tentons de présenter les grands principes de cet algorithme sans nous attarder à la description détaillée des aspects purement techniques de la méthode. La section 6 est consacrée aux principes d'estimation de la variance sous échantillonnage équilibré. Enfin, à la section 7, nous discutons de l'intérêt pratique de l'échantillonnage équilibré et comparons ce dernier à d'autres méthodes d'échantillonnage et au calage. Nous donnons également une liste d'applications récentes. Cette section traite aussi de l'exactitude de l'équilibrage, du choix des variables auxiliaires et des moyens de mettre en œuvre l'échantillonnage équilibré. L'article se termine par une bibliographie exhaustive sur l'échantillonnage équilibré et ses applications.

2. Échantillonnage équilibré

2.1 Définition d'un échantillon équilibré

Considérons un échantillon s de taille n qui est un sous-ensemble d'une population finie U de taille N. Un échantillon est dit équilibré si, pour un vecteur de variables auxiliaires $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kp})'$,

$$\frac{1}{n}\sum_{k\in S}\mathbf{x}_k = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k, \qquad (1)$$

ce qui signifie que les moyennes d'échantillon des variables x concordent avec leurs moyennes de population.

Brewer (1999) fait une distinction entre la sélection équilibrée d'échantillons et la sélection aléatoire d'échantillons. Cependant, un échantillon équilibré peut être sélectionné aléatoirement. Si un échantillon aléatoire S est sélectionné aléatoirement, chaque unité de la population a une probabilité π_k d'inclusion dans l'échantillon. Dans ce cas, un échantillon aléatoire doit satisfaire les équations d'équilibrage suivantes :

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U}\mathbf{x}_k. \qquad (2)$$

1. Yves Tillé, Université de Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Suisse. Courriel : yves.tille@unine.ch.

Autrement dit, dans un échantillon équilibré, le total des variables x est estimé sans erreur. Plusieurs auteurs, dont Cumberland et Royall (1981) et Kott (1986), donneraient à un échantillon qui satisfait l'équation (2) le nom d'« échantillon équilibré sur π », par opposition à l'« échantillon équilibré sur la moyenne » défini par l'équation (1). Néanmoins, dans le présent article, nous considérerons que (1) est simplement un cas particulier de (2) qui se produit quand $\pi_k = n/N$ ou quand l'échantillon n'est pas tiré aléatoirement. Dans les deux cas, nous parlons d'un échantillon équilibré.

2.2 Plan de sondage équilibré

Soit $p(s)$ le plan de sondage, c'est-à-dire la probabilité que l'échantillon s soit sélectionné, telle que $p(s) = \Pr(S = s)$, où S est l'échantillon aléatoire et $n(S)$, la taille de l'échantillon S. Selon la définition de Deville et Tillé (2004), un plan de sondage $p(\cdot)$ est dit *équilibré* sur les variables auxiliaires $x_1, \ldots, x_p$ si l'estimateur de Horvitz-Thompson satisfait l'équation (2). Dans un plan de sondage équilibré, les probabilités d'inclusion sont fixées avant le tirage de l'échantillon. Un échantillonnage équilibré peut être considéré comme une sorte de calage qui est directement intégré dans le plan de sondage. Le principal problème est que les équations d'équilibrage (2) peuvent rarement être satisfaites exactement. Nous donnons à cette difficulté le nom de « problème d'arrondi ».

Exemple 1. Si $N = 4$, $n = 2$, $\pi_k = 1/2$, pour tout $k \in U$ et $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 4$, les équations d'équilibrage données en (2) deviennent

$$\frac{1}{n}\sum_{k\in s} x_k = \frac{1}{N}\sum_{k\in U} x_k,$$

qui équivaut à

$$\sum_{k\in s} x_k = \frac{n}{N}\sum_{k\in U} x_k. \qquad (3)$$

Puisque

$$\frac{n}{N}\sum_{k\in U} x_k = \frac{2}{4}(0 + 1 + 2 + 4) = 3.5,$$

et que le premier membre de (3) est toujours un nombre entier, un échantillon exactement équilibré n'existe pas.

Effectivement, la sélection de l'échantillon est un problème en nombres entiers. La méthode du cube a donc pour objectif de tirer un échantillon qui satisfait exactement les probabilités d'inclusion π_k tout en restant aussi équilibré que possible.

3. Cas particuliers d'échantillonnage équilibré

3.1 Échantillonnage avec probabilités inégales et stratification

Certains plans d'échantillonnage bien connus sont des cas particuliers d'échantillonnage équilibré :

1. L'échantillonnage avec une taille d'échantillon fixe est un cas particulier de l'échantillonnage équilibré. Dans ce cas, la seule variable d'équilibrage est π_k. Les équations d'équilibrage données en (2) deviennent

$$\sum_{k\in S}\frac{\pi_k}{\pi_k} = \sum_{k\in S} 1 = \sum_{k\in U}\pi_k,$$

ce qui signifie que la taille d'échantillon doit être fixe.

2. La stratification est un cas particulier d'échantillonnage équilibré. Supposons que la population est partitionnée en H strates U_h, $h = 1, \ldots, H$, de tailles N_h, $h = 1, \ldots, H$, et qu'un échantillon est tiré dans chaque strate par échantillonnage aléatoire simple sans remise avec taille d'échantillon fixe n_h, $h = 1, \ldots, H$. Dans ce cas, les variables d'équilibrage sont les variables indicatrices des strates

$$\delta_{kh} = \begin{cases} 1 & \text{si } k \in U_h \\ 0 & \text{autrement.} \end{cases}$$

Sous un plan de sondage stratifié, les estimateurs de Horvitz-Thompson des tailles des strates sont exactement égaux aux tailles des strates, ce qui est une propriété de l'équilibrage sur les variables indicatrices des strates. En effet, puisque les probabilités d'inclusion dans la strate h sont $\pi_k = n_h/N_h$, $k \in U_h$, les équations d'équilibrage deviennent

$$\sum_{k\in S}\frac{N_h\delta_{kh}}{n_h} = \sum_{k\in U}\delta_{kh} = N_h,\ h = 1, \ldots, H,$$

et sont exactement satisfaites.

Ces deux plans sont bien connus et sont appliqués fréquemment en statistique officielle afin de réduire la variance. Le concept plus général d'équilibrage donne plus de liberté en vue de choisir les meilleures variables d'équilibrage qui augmenteront l'exactitude des estimateurs.

3.2 Strates chevauchantes

La construction d'un plan de sondage stratifié est souvent un exercice difficile. Les statisticiens essaient fréquemment d'effectuer la stratification en utilisant plusieurs variables qualitatives. Cependant, dans la plupart des cas, le croisement de toutes les strates de toutes les variables rend les cellules trop petites pour qu'un échantillon puisse être

sélectionné dans chacune d'elles. Dans le contexte du calage, les statisticiens effectuent généralement le calage sur les totaux marginaux et non sur toutes les cellules contenues dans un tableau de contingence. Puisqu'un échantillonnage équilibré peut être considéré comme une sorte de calage qui est intégré directement dans le plan de sondage, on souhaiterait également effectuer l'équilibrage en utilisant uniquement les totaux marginaux. Néanmoins, la théorie habituelle de la stratification ne permet pas le chevauchement des strates, puisque la stratification doit être une partition de la population. Maintenant, la méthode du cube permet d'équilibrer directement sur les totaux des strates chevauchantes en utilisant simplement les indicateurs des strates comme variables d'équilibrage.

3.3 Équilibrage sur une constante

Un autre cas particulier intéressant de l'échantillonnage équilibré est celui où une constante est utilisée comme variable d'équilibrage. Si $\mathbf{x}_k = 1$ pour tout $k \in U$, les équations d'équilibrage deviennent

$$\sum_{k \in S} \frac{1}{\pi_k} = \sum_{k \in U} 1 = N.$$

En fait,

$$\sum_{k \in S} \frac{1}{\pi_k}$$

est l'estimateur de Horvitz-Thompson de N. Cela signifie que, si la variable d'équilibrage utilisée est une constante, la taille de population estimée concorde avec la taille connue N, ce qui est loin d'être un fait acquis quand les unités statistiques sont sélectionnées avec probabilités d'inclusion inégales.

4. Historique du concept d'équilibrage et méthodes existantes

L'idée de l'échantillonnage équilibré est très ancienne et reliée au vague concept de représentativité qui était déjà employé par Kiaer (1896, 1899, 1903, 1905). Le premier article consacré au tirage d'un échantillon équilibré est dû à Gini (1928) et à Gini et Galvani (1929) qui ont tiré un échantillon de 29 parmi 214 districts italiens afin d'égaler plusieurs totaux de population. Neyman (1952) et Yates (1960) ont tous deux condamné l'article de Gini et Galvani, essentiellement parce que la sélection de l'échantillon n'était pas aléatoire (voir Langel et Tillé 2010). Les premières méthodes de tirage d'un échantillon équilibré aléatoire ont été proposées par Yates (1946) et par Thionet (1953), mais elles étaient réjectives en ce sens qu'elles comportaient le tirage aléatoire d'échantillons ou le remplacement aléatoire d'unités dans l'échantillon jusqu'à ce qu'un échantillon suffisamment équilibré soit obtenu.

Dans le cadre basé sur un modèle, Royall (1976a, 1976b) a préconisé d'utiliser l'échantillonnage équilibré afin d'atteindre la stratégie optimale et de se protéger contre l'erreur de spécification du modèle (voir aussi Royall et Pfeffermann 1982 ; Kott 1986 ; Cumberland et Royall 1988 ; Royall 1988 ; Tirari 2006 ; Nedyalkova et Tillé 2009). Bien que plusieurs méthodes de sélection d'un échantillonnage équilibré soient présentées dans le livre de Valliant, Dorfman et Royall (2000), ces méthodes ne spécifient pas nécessairement les probabilités d'inclusion de l'échantillon. Dans le cadre basé sur un modèle, il est important que l'échantillon soit équilibré. Cependant, cet échantillon ne doit pas toujours être sélectionné aléatoirement.

Hájek (1981) a également recommandé d'utiliser l'échantillonnage équilibré. Pour Hájek, un échantillonnage équilibré est un cas particulier d'une stratégie représentative, une stratégie consistant en un couple formé d'un plan de sondage et d'un estimateur. Une stratégie représentative est une stratégie qui permet d'estimer les totaux des variables auxiliaires sans erreur. En ce sens, le plan de sondage équilibré avec l'estimateur de Horvitz-Thompson est une stratégie représentative. Hájek (1981) propose une procédure réjective qui consiste à tirer une série d'échantillons jusqu'à ce que l'on obtienne un échantillon équilibré. Les procédures réjectives ont deux inconvénients : si plusieurs variables d'équilibrage sont utilisées, la procédure peut être très lente ; deuxièmement, les probabilités d'inclusion des plan réjectifs ne sont pas les mêmes que celle du plan original. Les probabilités d'inclusion des unités statistiques qui sont proches des moyennes de population sont augmentées au détriment de celles des unités qui sont éloignées du centre (voir par exemple les simulations de Legg et Yu 2010).

Une autre méthode de sélection consiste à énumérer tous les échantillons possibles, puis à construire un plan de sondage uniquement pour tirer les échantillons qui sont adéquatement équilibrés. Un plan de ce genre peut être construit en recourant à la programmation linéaire. Cette technique a été utilisée par Ardilly (1991) pour sélectionner les unités primaires de l'échantillon-maître français. Néanmoins, cette méthode ne s'applique qu'à de petites tailles de population en raison de l'explosion combinatoire du nombre d'échantillons quand la taille de la population est grande.

Deville, Grosbras et Roth (1988) et Deville (1992) ont proposé des méthodes multivariées d'échantillonnage équilibré avec probabilités d'inclusion égales. Hedayat et Majumdar (1995) ont proposé l'adaptation d'une technique basée sur un plan expérimental qui permettrait de créer un plan de sondage équilibré. De nouveau, cette technique est limitée aux probabilités d'inclusion égales. Enfin, la méthode du cube a été proposée par Deville et Tillé (2004).

Cette méthode est générale en ce sens que les probabilités d'inclusion sont exactement satisfaites, que ces probabilités peuvent être égales ou inégales, et que l'échantillon est aussi équilibré que possible.

Fuller (2009) a étudié une procédure réjective en fixant un intervalle de tolérance en dehors duquel l'échantillon est rejeté et a proposé un estimateur de variance. Même si les probabilités d'inclusion sont modifiées par une procédure réjective, Fuller (2009) montre que des estimations efficaces sont obtenues en utilisant les probabilités d'inclusion du plan original. En utilisant un ensemble de simulations, Legg et Yu (2010) ont comparé cette procédure réjective à la méthode du cube et montré que les deux méthodes donnent des résultats équivalents. Enfin, Dudoignon et Vanheuverzwyn (2006) ont proposé une méthode rapide d'échantillonnage équilibré pour les totaux marginaux, tandis que Périé (2008) a proposé une méthode basée sur des nombres aléatoires permanents qui fournissent un échantillon équilibré. Dans le cas de la méthode de Périé (2008), les probabilités d'inclusion ne sont qu'approximativement satisfaites.

5. La méthode du cube

5.1 Idées principales

La méthode du cube (voir Deville et Tillé 2004 ; Tillé 2001, 2006a, 2006b ; Ardilly 2006) représente une classe d'algorithmes d'échantillonnage qui réalise le tirage d'un échantillon équilibré et satisfait exactement un ensemble de probabilités d'inclusion données. La méthode du cube est une extension de la méthode de scission élaborée par Deville et Tillé (1998). Elle est basée sur une transformation aléatoire du vecteur des probabilités d'inclusion jusqu'à l'obtention d'un échantillon tel que :

(i) les probabilités d'inclusion soient exactement satisfaites ;
(ii) les équations d'équilibrage soient satisfaites autant qu'il est possible.

La méthode doit son nom à la représentation géométrique d'un plan de sondage. En effet, un échantillon peut être représenté par un vecteur d'indicateurs d'échantillons :

$$\mathbf{s} = (I[1 \in s] \dots I[k \in s] \dots I[N \in s])',$$

où $I[k \in s]$ prend la valeur 1 si $k \in s$ et 0 autrement. Un échantillon peut donc être vu comme l'un des sommets d'un hypercube de dimension N comme l'illustre la figure 1.

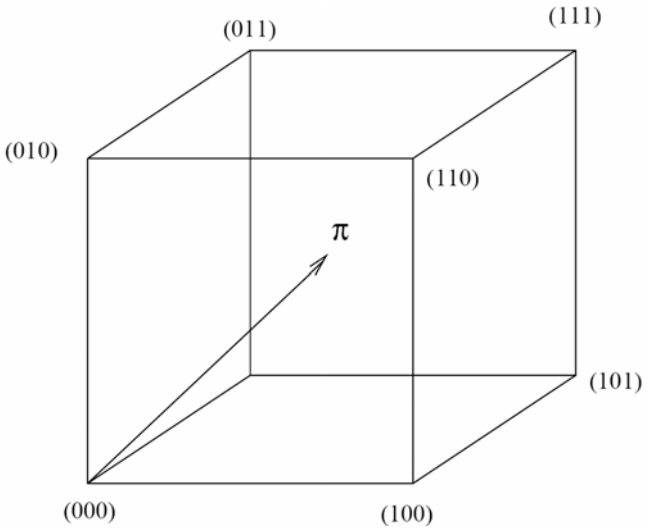

Figure 1 Échantillons possibles dans une population de taille $N = 3$

Définissons aussi

$$\mathrm{E}(\mathbf{s}) = \sum_{s \in S} p(\mathbf{s})\mathbf{s} = \boldsymbol{\pi},$$

où $\boldsymbol{\pi} = [\pi_k]$ est le vecteur des probabilités d'inclusion. Les équations d'équilibrage

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k,$$

peuvent aussi s'écrire

$$\sum_{k \in U} \breve{\mathbf{x}}_k \, s_k = \sum_{k \in U} \breve{\mathbf{x}}_k \, \pi_k, \qquad (4)$$

où $s_k \in \{0, 1\}$ et $\breve{\mathbf{x}}_k = \mathbf{x}_k / \pi_k$, $k \in U$. L'expression (4) est un système d'équations contenant des valeurs inconnues s_k qui définissent un sous-espace affine dans $\mathbb{R}^N$ de dimension $N - p$ désigné par Q, où

$$Q = \left\{ \mathbf{u} \in \mathbb{R}^N \mid \sum_{k \in U} \breve{\mathbf{x}}_k \, u_k = \sum_{k \in U} \mathbf{x}_k \right\}.$$

Le problème du tirage d'un échantillon équilibré peut donc être reformulé. Un plan de sondage équilibré consiste à choisir un sommet de l'hypercube de dimension N (un échantillon) qui demeure dans le sous-espace linéaire Q. Les figures 2 et 3 montrent, respectivement, deux exemples : le premier correspond à une contrainte de taille d'échantillon fixe et le deuxième, à une contrainte donnant lieu à un problème d'arrondi.

La méthode du cube (Deville et Tillé 2004) est divisée en deux phases : la phase de vol et la phase d'atterrissage. La phase de vol est une marche aléatoire qui part du vecteur des probabilités d'inclusion et demeure à l'intersection du cube

et du sous-espace de contrainte. Cette marche aléatoire s'arrête à un sommet de l'intersection du cube et du sous-espace des contraintes. À la fin de la phase de vol, si aucun échantillon n'a été obtenu, la phase d'atterrissage se déclenche en vue de sélectionner un échantillon aussi proche que possible du sous-espace des contraintes.

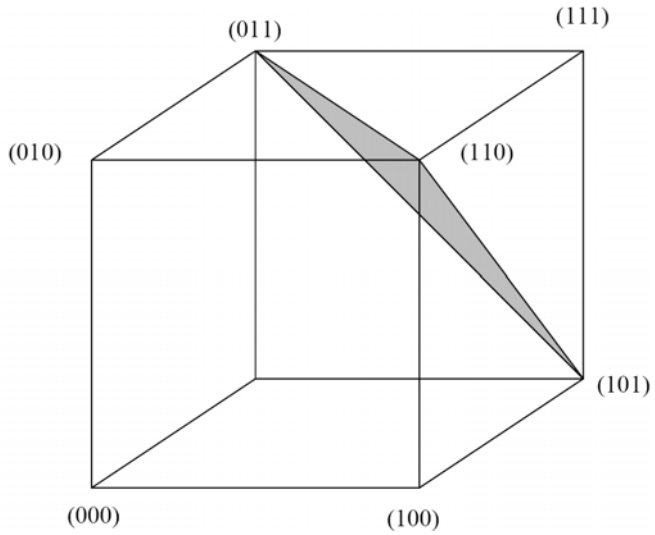

Figure 2 Échantillons possibles dans une population de taille $N = 3$ avec une contrainte de taille d'échantillon fixe $n = 2$

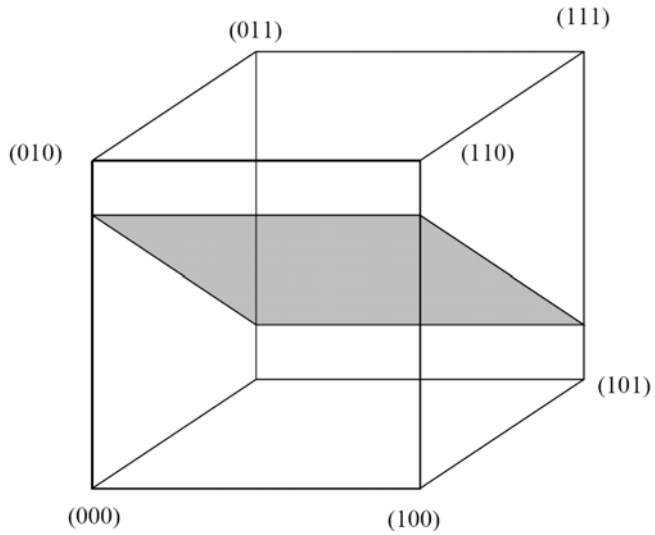

Figure 3 Échantillons possibles dans une population de taille $N = 3$ avec une contrainte et un problème d'arrondi

Exemple 2. Si la contrainte est la taille fixe d'échantillon, la phase de vol transforme aléatoirement un vecteur de probabilités d'inclusion en un vecteur de 0 et de 1. À chaque étape de l'algorithme, le vecteur de probabilités d'inclusion est transformé aléatoirement, mais la somme des probabilités d'inclusion doit demeurer égale à n. Par exemple, avec $\boldsymbol{\pi} = (0{,}5,\ 0{,}5,\ 0{,}5,\ 0{,}5)$ et $n = 2$, nous sommes capables d'obtenir la série suivante de vecteurs :

$$\boldsymbol{\pi} = \begin{pmatrix} 0{,}5 \\ 0{,}5 \\ 0{,}5 \\ 0{,}5 \end{pmatrix} \rightarrow \begin{pmatrix} 0{,}6666 \\ 0{,}6666 \\ 0{,}6666 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0{,}5 \\ 0{,}5 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \mathbf{s}.$$

L'algorithme s'arrête quand toutes les composantes du vecteur sont égales à 0 ou 1.

Exemple 3. Si la contrainte est la taille fixe d'échantillon, un problème d'arrondi a lieu si la somme des probabilités d'inclusion n'est pas un entier. En cas de problème d'arrondi, certaines composantes ne peuvent pas être fixées à zéro. Par exemple, avec $\boldsymbol{\pi} = (0{,}5,\ 0{,}5,\ 0{,}5,\ 0{,}5,\ 0{,}5)$ et

$$\sum_{k \in U} \pi_k = 2{,}5,$$

nous pouvons observer la série suivante de vecteurs :

$$\boldsymbol{\pi} = \begin{pmatrix} 0{,}5 \\ 0{,}5 \\ 0{,}5 \\ 0{,}5 \\ 0{,}5 \end{pmatrix} \rightarrow \begin{pmatrix} 0{,}625 \\ 0 \\ 0{,}625 \\ 0{,}625 \\ 0{,}625 \end{pmatrix} \rightarrow \begin{pmatrix} 0{,}5 \\ 0 \\ 0{,}5 \\ 1 \\ 0{,}5 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0{,}25 \\ 1 \\ 0{,}25 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0{,}5 \\ 1 \\ 0 \end{pmatrix} = \boldsymbol{\pi}^*.$$

Dans ce cas, la phase de vol ne peut pas se terminer par un vecteur de 0 ou de 1 dont la somme est égale à 2,5. Dans ces conditions, la phase de vol se termine par un vecteur contenant une composante non entière.

5.2 La phase de vol

La première étape de la phase de vol est présentée à la figure 4 pour un cas très particulier : la taille de population $N = 3$. La seule contrainte d'équilibrage est la taille fixe d'échantillon $n = 2$. À la première étape, un vecteur $\mathbf{u}(0)$ doit être choisi. Il peut l'être librement, mais doit être tel que $\boldsymbol{\pi} + \mathbf{u}(0)$ demeure dans le sous-espace des contraintes. En fait, la méthode du cube correspond à une famille de méthodes qui dépendent de la façon dont le vecteur $\mathbf{u}(0)$ est choisi. Il peut l'être aléatoirement ou non.

Si, en partant de $\boldsymbol{\pi}$, nous suivons la direction donnée par le vecteur $\mathbf{u}(0)$, nous aboutissons nécessairement à une face du cube. Considérons ce point désigné sur la figure 4 par $\boldsymbol{\pi}(0) + \lambda_1^*(0)\mathbf{u}(0)$. Maintenant si, en partant de $\boldsymbol{\pi}$, nous suivons la direction opposée, c'est-à-dire la direction donnée par le vecteur $-\mathbf{u}(0)$, nous aboutissons aussi à une face du cube. Considérons ce point désigné sur la figure 4 par $\boldsymbol{\pi}(0) - \lambda_2^*(0)\mathbf{u}(0)$. À la première étape, le vecteur $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$ est modifié aléatoirement. Le vecteur $\boldsymbol{\pi}(1)$ sera fixé à $\boldsymbol{\pi}(0) + \lambda_1^*(0)\mathbf{u}(0)$ ou à $\boldsymbol{\pi}(0) - \lambda_2^*(0)\mathbf{u}(0)$. Le choix est

effectué aléatoirement de façon que $E[\boldsymbol{\pi}(1)] = \boldsymbol{\pi}(0)$. À la fin de la première étape de la phase de vol, nous avons donc atteint une face du cube, ce qui signifie qu'au moins une composante de $\boldsymbol{\pi}(1)$ est égale à 0 ou à 1, c'est-à-dire que le problème est réduit d'un problème d'échantillonnage à partir d'une population de taille $N = 3$ à une population de taille $N = 2$. En N étapes au moins, la phase de vol est donc achevée.

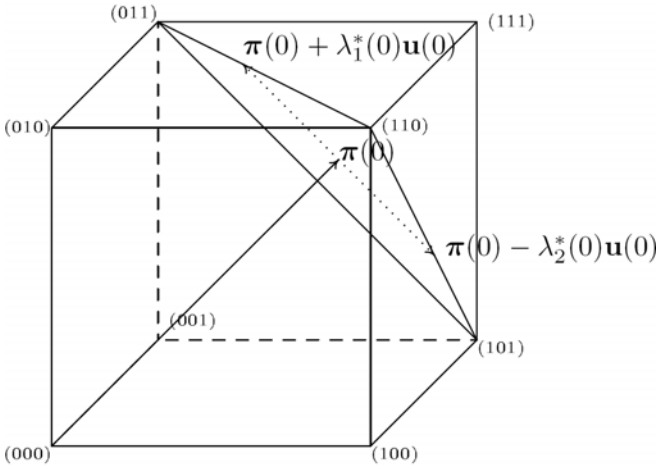

Figure 4 Phase de vol dans une population de taille $N = 3$ avec une contrainte de taille d'échantillon $n = 2$

De façon plus générale, la phase de vol est une marche aléatoire dans l'intersection du sous-espace d'équilibrage et du cube. Cette marche aléatoire s'arrête à un sommet de l'intersection du cube et du sous-espace. La phase de vol est définie par la classe suivante d'algorithmes. Commencer par initialiser à $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Ensuite, au temps $t = 0, ..., T$,

1. Générer un vecteur quelconque $\mathbf{u}(t) = [u_k(t)] \neq 0$ tel que
   i) $\mathbf{u}(t)$ soit dans le noyau de la matrice $\mathbf{A} = (\mathbf{x}_1/\pi_1, \ldots, \mathbf{x}_k/\pi_k, \ldots, \mathbf{x}_N/\pi_N)$, c'est-à-dire $\mathbf{Au}(t) = 0$,
   ii) $u_k(t) = 0$ si $\pi_k(t)$ est un entier.
2. Calculer $\lambda_1^*(t)$ et $\lambda_2^*(t)$, les plus grandes valeurs telles que

$$0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\mathbf{u}(t) \leq 1,$$

$$0 \leq \boldsymbol{\pi}(t) - \lambda_2(t)\mathbf{u}(t) \leq 1.$$

3. Calculer

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\mathbf{u}(t) \text{ avec la probabilité } q_1(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\mathbf{u}(t) \text{ avec la probabilité } q_2(t), \end{cases}$$

où $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ et $q_2(t) = 1 - q_1(t)$.

La phase de vol s'arrête quand il n'est plus possible de trouver un vecteur $\mathbf{u}(t) \neq 0$.

5.3 Phase d'atterrissage

Si, à la fin de la phase de vol, les équations d'équilibrage ne sont pas exactement satisfaites, la phase d'atterrissage est nécessaire. Soit $\boldsymbol{\pi}^* = [\pi_k^*]$ le vecteur obtenu à la dernière étape de la phase de vol. Il est possible de prouver (voir Deville et Tillé 2004) que

$$\text{card}(U^*) \leq p,$$

où

$$U^* = \{k \in U \mid 0 < \pi_k^* < 1\}$$

et p est le nombre de variables d'équilibrage. Le but de la phase d'atterrissage est de trouver un échantillon $\mathbf{s}$ tel que $E(\mathbf{s}|\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$, qui est presque équilibré. Il existe deux moyens de tirer un tel échantillon :

1. *La phase de vol par programmation linéaire* consiste à considérer tous les échantillons possibles de U^*. Un coût est attribué à chaque échantillon. Ce coût est, par exemple, la distance entre l'échantillon et le sous-espace des contraintes. Ensuite, on recherche un plan de sondage de U^* qui minimise le coût prévu et qui satisfait les probabilités d'inclusion $\boldsymbol{\pi}^*$. Ce problème peut être résolu parce que le nombre d'échantillons à considérer est raisonnable étant donné la petite taille de U^*.

2. *La phase de vol par suppression de variables* peut être utilisée quand le nombre de variables d'équilibrage est trop grand pour que le problème de programmation linéaire puisse être résolu par l'algorithme du simplexe ($p > 20$). Si l'on applique cette méthode, une variable auxiliaire est abandonnée à la fin de la phase de vol. Ensuite, on peut retourner à la phase de vol jusqu'à ce qu'il ne soit plus possible de « bouger » dans le sous-espace des contraintes. Les contraintes sont alors relâchées successivement selon un ordre de préférence.

6. Variance et estimation de la variance

6.1 Une technique de résidu

La variance de l'estimateur de Horvitz-Thompson peut être estimée en appliquant une technique de résidus élaborée dans Deville et Tillé (2005). Cette technique est comparable à celle utilisée pour estimer la variance de l'estimateur par calage et a été validée par un ensemble de simulations. La variance estimée de l'estimateur de Horvitz-Thompson est

donc fort semblable à la variance estimée d'un estimateur par la régression généralisée (GREG). Néanmoins, la variance de l'estimateur GREG est généralement sous-estimée, parce qu'elle ne tient pas compte du caractère aléatoire des pondérations. En effet, si la variance habituelle de l'estimateur GREG est calculée pour le cas particulier de la poststratification, nous obtenons la variance d'un plan stratifié avec répartition proportionnelle. La variance de l'estimateur poststratifié est néanmoins plus grande que celle obtenue sous un plan stratifié avec répartition proportionnelle.

6.2 Approximation de la variance

Si le plan de sondage équilibré possède une grande entropie, Hájek (1981) ainsi que Deville et Tillé (2005, méthode 4) ont proposé l'approximation qui suit de la variance sous le plan donnée par :

$$\mathrm{var}_p(\hat{Y}_\pi) \cong \mathrm{var}_{app}(\hat{Y}_\pi) = \sum_{k\in U} d_k \frac{(y_k - \mathbf{x}_k'\mathbf{b})^2}{\pi_k^2}, \quad (5)$$

où l'indice inférieur p désigne le plan de sondage,

$$\mathbf{b} = \left(\sum_{k\in U} d_k \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k^2}\right)^{-1} \sum_{k\in U} d_k \frac{\mathbf{x}_k y_k}{\pi_k^2},$$

et les d_k sont la solution du système non linéaire

$$\pi_k(1-\pi_k) = d_k - \frac{d_k\mathbf{x}_k'}{\pi_k}\left(\sum_{\ell\in U} d_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \frac{d_k\mathbf{x}_k}{\pi_k}, \; k \in U. \quad (6)$$

L'entropie du plan de sondage dépend de la façon dont les vecteurs $\mathbf{u}(t)$ sont choisis durant la phase de vol. Afin d'accroître l'entropie, le vecteur $\mathbf{u}(t)$ peut être choisi aléatoirement ou bien la population peut être triée aléatoirement avant de tirer l'échantillon.

L'expression (5), qui ne contient que les probabilités d'inclusion de premier ordre, a été validée par Deville et Tillé (2005) sous une gamme d'échantillons équilibrés, indépendamment de la façon dont les valeurs de y étaient générées. Une approximation très proche de l'expression (5) a été obtenue par Fuller (2009) ainsi que par Legg et Yu (2010) pour un plan de sondage équilibré obtenu par une méthode réjective dans le cas d'un plan initial utilisant l'échantillonnage de Poisson. Ces approximations ne tiennent pas compte du problème d'arrondi.

6.3 Estimation de la variance

Deville et Tillé (2005) ont proposé une famille d'estimateurs de variance pour l'échantillonnage équilibré de la forme

$$\widehat{\mathrm{var}}(\hat{Y}_\pi) = \sum_{k\in S} c_k \frac{(y_k - \mathbf{x}_k'\hat{\mathbf{b}})^2}{\pi_k^2}, \quad (7)$$

où

$$\hat{\mathbf{b}} = \left(\sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell y_\ell}{\pi_\ell^2}$$

et les c_k sont les solutions du système non linéaire

$$1 - \pi_k = c_k - \frac{c_k\mathbf{x}_k'}{\pi_k}\left(\sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \frac{c_k\mathbf{x}_k}{\pi_k}, \quad (8)$$

qui peut être résolu au moyen d'un algorithme du point fixe.

Dans Deville et Tillé (2005), des variantes plus simples de c_k ont également été proposées. Par exemple, on peut utiliser les valeurs de rechange,

$$\tilde{c}_k \approx \frac{n}{n-p}(1-\pi_k),$$

qui sont très proches de c_k. L'estimateur $\widehat{\mathrm{var}}(\hat{Y}_\pi)$ est approximativement sans biais sous le plan parce qu'il s'agit d'un estimateur par substitution de l'approximation donnée par l'expression (5) (pour plus de renseignements concernant les estimateurs obtenus par substitution, voir Deville 1999), qui est une approximation raisonnable de la variance sous le plan de sondage.

Il n'est pas facile d'utiliser la méthode du bootstrap pour estimer la variance dans le contexte de l'échantillonnage équilibré. Les échantillons équilibrés avec remise devraient être tirés de l'échantillon original. Une généralisation de la méthode du cube pour l'échantillonnage équilibré avec remise n'a pas encore été décrite. Une solution, proposée par Chauvet (2007), consiste à reconstruire une population artificielle d'après l'échantillon. Ensuite, des échantillons bootstrap sont tirés par une méthode d'échantillonnage équilibré. Une autre solution a été proposée par Fuller (2010) pour l'échantillonnage réjectif équilibré. Breidt et Chauvet (2010a) ont proposé une autre méthode dans laquelle une représentation de la méthode du cube par différence de martingale est utilisée pour approcher les probabilités d'inclusion de deuxième ordre, ce qui permet de construire un estimateur de variance presque sans biais.

7. Échantillonnage équilibré en pratique

7.1 Intérêt de l'échantillonnage équilibré

Dans les cadres assisté par modèle et basé sur un modèle, l'utilisation d'un plan de sondage équilibré avec l'estimateur de Horvitz-Thompson est souvent la stratégie optimale (voir Nedyalkova et Tillé 2009). En effet, quand l'échantillon est

équilibré, les variances des estimateurs de Horvitz-Thompson des variables auxiliaires sont nulles. Sous un modèle linéaire, la variance de l'estimateur de Horvitz-Thompson de la variable d'intérêt dépendra uniquement des résidus du modèle.

Les avantages de l'échantillonnage équilibré sont les suivants :

i) L'échantillonnage équilibré augmente la précision de l'estimateur de Horvitz-Thompson. Ce point a été traité à la section 6. En effet, la variance de l'estimateur de Horvitz-Thompson dépend uniquement des résidus de la régression de la variable d'intérêt en fonction des variables d'équilibrage.
ii) L'échantillonnage équilibré protège contre les grandes erreurs d'échantillonnage. En effet, les échantillons les moins favorables ont une probabilité nulle d'être tirés.
iii) Si la variable d'intérêt est bien expliquée par l'information auxiliaire, dans l'inférence basée sur un modèle, l'échantillonnage équilibré protège contre une erreur de spécification du modèle. Ce point est traité en détail par Royall (1976a, 1976b) et par Valliant et coll. (2000). Une discussion récente de cette question importante est présentée dans Nedyalkova et Tillé (2009, 2010).
iv) L'échantillonnage équilibré permet de s'assurer que les tailles d'échantillon dans les domaines prévus ne soient pas trop faibles ou – situation pire encore – nulles. En effet, si une variable indicatrice du domaine est ajoutée à la liste des variables auxiliaires, la taille du domaine est alors fixée dans l'échantillon.
v) L'échantillonnage équilibré permet d'éviter les pondérations aléatoires. Sous échantillonnage équilibré, nous pouvons utiliser les pondérations de Horvitz-Thompson. Si le plan de sondage ne contient aucune contrainte d'équilibrage (par exemple sous échantillonnage poissonnien), le système de pondération obtenu par une procédure de calage devient très aléatoire, ce qui augmente la variance des estimateurs. Si l'échantillon est équilibré, les pondérations seront moins aléatoires, même si une procédure de calage est utilisée après l'équilibrage.

L'existence de progiciels faciles à utiliser a contribué à l'usage répandu de la méthode du cube dans plusieurs processus statistiques importants. La première grande application de la méthode du cube est la sélection de groupes de rotation pour le recensement français (voir Desplanques 2000 ; Dumais, Bertrand et Kauffmann 2000 ; Durr et Dumais 2001, 2002 ; Dumais et Isnard 2000 ; Bertrand, Christian, Chauvet et Grosbras 2004 ; da Silva, da Silva Borges, Aires Leme et Moura Reis Miceli 2006). Pour les communes de moins de 10 000 habitants, cinq groupes de rotation non chevauchants de communes sont sélectionnés selon un plan de sondage équilibré avec probabilités d'inclusion égales (1/5). Chaque année, un sondage est effectué auprès du cinquième des communes. Donc, après cinq ans, toutes les petites communes ont été sélectionnées. Pour les communes de plus de 10 000 habitants, dans chaque commune, cinq échantillons équilibrés non chevauchants d'adresses sont tirés avec probabilités d'inclusion de 8 %. Donc, après cinq ans, une visite est effectuée à 40 % des adresses. Les variables d'équilibrage sont des variables sociodémographiques tirées du dernier recensement.

Dans l'échantillon-maître français, les unités primaires sont les régions géographiques qui sont sélectionnés selon un plan de sondage équilibré (voir Wilms 2000 ; Christine et Wilms 2003 ; Christine 2006). L'échantillon-maître est obtenu par échantillonnage à plusieurs degrés autopondéré. Donc, les unités primaires sont tirées avec probabilités inégales qui sont proportionnelles à leurs tailles. Les variables d'équilibrage sont des variables sociodémographiques provenant du dernier recensement. Bardaji (2001) et Even (2002) ont également utilisé l'échantillonnage équilibré pour sélectionner un échantillon de bénéficiaires d'emplois subventionnés. Sept populations sont sondées, un échantillon équilibré de bénéficiaires est sélectionné dans chacune des populations en utilisant de deux à cinq variables d'équilibrage selon la population.

La société Électricité de France (EDF) a installé de nouveaux compteurs d'électricité qui permettent de mesurer la consommation d'électricité de chaque ménage sur une base continue. La quantité d'informations recueillies est tellement grande qu'il est impossible d'archiver toutes les données. Dessertaine (2006, 2007) a utilisé l'échantillonnage équilibré pour sélectionner les séries chronologiques de données sur la consommation qui doivent être archivées afin de s'assurer qu'elles représentent aussi exactement que possible la consommation de l'ensemble de la population française. Biggeri et Falorsi (2006) ont utilisé l'échantillonnage équilibré pour améliorer la qualité de l'indice des prix à la consommation en Italie. Gismondi (2007) a testé l'échantillonnage équilibré pour estimer le nombre de nuits que les touristes passent en Italie. D'Alò, Di Consiglio, Falorsi et Solari (2006) ainsi que Falorsi et Righi (2008) ont également proposé d'utiliser un plan de sondage équilibré pour estimer les totaux dans les petits domaines. Des simulations ont été exécutées par Marí, Barbará, Mitas et Passamonti (2007a, 2007b) en Argentine et par Chipperfield (2009) en Australie pour évaluer l'intérêt de l'échantillonnage équilibré pour l'échantillon-maître.

À Statistique Canada, Fecteau et Jocelyn (2006) et Jocelyn (2006) ont testé l'échantillonnage équilibré pour tirer un échantillon d'entreprises. Les entreprises canadiennes non constituées en société produisent leur déclaration de revenus sur papier ou électroniquement. Plus de la moitié des déclarations sont soumises électroniquement. L'échantillonnage équilibré a été utilisé pour sélectionner un échantillon parmi les entreprises ayant produit une déclaration électronique, de manière que, pour certaines variables clés dont la valeur est connue pour l'ensemble de la population, les moyennes d'échantillon concordent avec les moyennes de population connues.

L'échantillonnage équilibré peut également être utilisé pour imputer une valeur manquante en cas de non-réponse partielle. En effet, l'utilisation d'un modèle pour prédire une imputation attribue les valeurs centrales, ce qui donne lieu à une inférence biaisée sur les quantiles. Par contre, une imputation aléatoire augmente généralement les variances des estimateurs. Afin de résoudre ce dilemme, Deville (1998, 2005, 2006), ainsi que Chauvet, Deville et Haziza (2010b, 2010c) ont proposé d'utiliser l'imputation par prédiction et d'ajouter un résidu qui est choisi parmi les résidus des répondants selon un plan de sondage équilibré. Ce faisant, on évite d'ajouter un terme de variance au total de la variable imputée.

7.2 Échantillonnage équilibré versus d'autres techniques d'échantillonnage

L'échantillonnage à probabilités inégales est un cas particulier de la méthode du cube. En effet, quand la probabilité d'inclusion est la seule variable auxiliaire, la taille de l'échantillon est fixe. La méthode du cube est une généralisation de la méthode de scission (voir Deville et Tillé 1998), qui comporte plusieurs algorithmes d'échantillonnage à probabilités inégales (méthode de Brewer, méthode du pivot, méthode corrigée de Sunter, voir Brewer 1975 ; Sunter 1977 ; Deville et Tillé 1998 ; Tillé 2006b). La stratification est aussi un cas particulier d'échantillonnage équilibré. La méthode du cube permet d'effectuer l'équilibrage sur des strates chevauchantes et d'utiliser ensemble des variables qualitatives et quantitatives. Même l'échantillonnage systématique peut être vu comme un échantillonnage équilibré sur la statistique d'ordre reliée à la variable en fonction de laquelle la population est ordonnée.

Presque toutes les autres techniques d'échantillonnage sont des cas particuliers d'échantillonnage équilibré (sauf l'échantillonnage à plusieurs degrés). En fait, l'échantillonnage équilibré est simplement plus général, en ce sens que toutes les autres méthodes d'échantillonnage peuvent être mises en œuvre en se servant de la méthode du cube. Cette dernière permet d'utiliser n'importe quelle variable pour effectuer l'équilibrage avec un temps de calcul raisonnable. Grâce au concept plus général d'équilibrage, les strates peuvent se chevaucher, des variables quantitatives et qualitatives peuvent être utilisées simultanément et les probabilités d'inclusion peuvent être choisies librement.

Il est bien connu que l'estimateur par le ratio et l'estimateur poststratifié sont des cas particuliers de l'estimateur par la régression. Ce dernier est aussi un cas particulier de l'estimateur par calage (qui comprend un ajustement non linéaire). De même, l'échantillonnage équilibré est une méthode d'échantillonnage plus générale qui englobe presque toutes les autres. L'algorithme de la méthode du cube peut paraître compliqué, mais, une fois implémenté, il permet d'exécuter une fonction avec deux arguments, à savoir le vecteur des probabilités d'inclusion et la matrice des variables d'équilibrage.

7.3 Choix de la stratégie d'équilibrage

La principale recommandation est de choisir des variables d'équilibrage qui sont étroitement corrélées aux variables d'intérêt. Comme dans tout problème de régression, les variables d'équilibrage doivent être choisies parcimonieusement : il ne faut pas en choisir un trop grand nombre, parce que la précision n'augmente plus une fois que le nombre de variables est grand et l'instabilité de l'estimateur de variance s'accroît avec chaque variable supplémentaire. En pratique, le but n'est pas d'estimer une variable, mais un ensemble de variables d'intérêt. Donc, l'ensemble de variables auxiliaires doit être corrélé à toutes les variables d'intérêt. De surcroît, les variables auxiliaires ne doivent pas être trop corrélées entre elles.

Lesage (2008) a proposé une méthode pour équilibrer un échantillon sur des statistiques complexes au lieu d'utiliser simplement les totaux de population. L'idée fondamentale consiste à effectuer l'équilibrage sur la valeur linéarisée (ou fonction d'influence) du paramètre d'intérêt. Breidt et Chauvet (2010b) ont proposé de recourir à l'échantillonnage équilibré pénalisé afin de pouvoir éventuellement relâcher certaines contraintes d'équilibrage, ce qui peut être utile par exemple dans l'estimation sur petits domaines.

Dans de nombreux cas, les variables d'équilibrage contiennent des erreurs de mesure. Ainsi, dans la plupart des registres, on peut soupçonner la présence d'erreurs dans les données. Des valeurs manquantes peuvent manifestement exister et les variables auxiliaires sont souvent corrigées par une méthode d'imputation. Pour ce qui est du calage, le fait que les variables auxiliaires contiennent des erreurs n'est pas très important pourvu que le calage soit effectué sur le total des variables auxiliaires du registre. En effet, sous échantillonnage équilibré, on utilise l'estimateur de Horvitz-Thompson qui est sans biais même si les variables

auxiliaires sont fausses. Le gain d'efficacité dépend uniquement de la corrélation entre les variables d'équilibrage et les variables d'intérêt. Cette corrélation est rarement affectée par les erreurs touchant les variables d'équilibrage.

Plusieurs variables peuvent être utilisées pour améliorer les estimations sur petits domaines. Afin de s'assurer qu'un domaine D n'est pas vide, on peut simplement ajouter la variable auxiliaire :

$$x_k = \begin{cases} \pi_k & \text{si } k \in D \\ 0 & \text{autrement,} \end{cases}$$

qui implique que le nombre d'unités échantillonnées qui appartiennent à D est égal à

$$n_D = \sum_{k \in U} x_k = \sum_{k \in D} \pi_k,$$

si n_D est un entier, ou à l'un des deux entiers les plus proches de n_D si n_D n'est pas un entier.

Dans certains cas, il est utile d'équilibrer sur des variables auxiliaires dans des sous-groupes, des domaines ou des strates. Une procédure intéressante décrite dans Chauvet (2009) consiste à exécuter séparément la phase de vol dans chaque strate. Un problème d'arrondi surviendra alors dans chaque strate. Il est possible de fusionner ces problèmes d'arrondi et d'exécuter de nouveau une phase de vol sur l'ensemble de la population. Enfin, la phase d'atterrissage est appliquée uniquement à l'ensemble de la population. Cette procédure permet que les équations d'équilibrage soient approximativement satisfaites dans chaque strate sans cumuler les problèmes d'arrondi.

Les probabilités d'inclusion doivent être calculées avant l'échantillonnage. Si l'on émet l'hypothèse d'un modèle linéaire, ces probabilités doivent, en principe, être proportionnelles aux erreurs du modèle afin de minimiser la variance (voir Tillé et Favre 2005 ; Chauvet, Bonnery et Deville 2010a ; Nedyalkova et Tillé 2009, 2010). Ce choix généralise la méthode d'allocation de Neyman pour l'échantillonnage stratifié (Neyman 1934). Cependant, les probabilités d'inclusion doivent souvent être choisies en fonction d'autres contraintes. Par exemple, afin de construire les groupes de rotation du recensement français, les probabilités d'inclusion doivent être égales à un cinquième.

7.4 Équilibrage versus calage

La stratification est un cas particulier de l'équilibrage, tandis que la poststratification est un cas particulier du calage. Dans le cas de la stratification et de l'équilibrage, les pondérations ne deviennent pas aléatoires. Donc, il s'agit généralement d'une meilleure stratégie. Néanmoins, l'équilibrage requiert une plus grande quantité d'information auxiliaire. En effet, dans l'échantillonnage équilibré, les variables auxiliaires doivent être connues pour toutes les unités de la population, tandis que dans le calage, seuls les totaux de population sont nécessaires. L'équilibrage est une méthode fort intéressante pour les populations de petite taille. Il s'agit donc d'une très bonne méthode pour sélectionner les unités primaires sous un plan de sondage à plusieurs degrés.

Les deux techniques peuvent être utilisées ensemble, car elles ne sont pas contradictoires. La meilleure stratégie consiste à utiliser de concert l'échantillonnage équilibré et le calage, car ce dernier peut résoudre le petit problème d'arrondi susceptible de persister après l'équilibrage. À l'étape de l'estimation, un plus grand nombre de variables auxiliaires sont souvent disponibles car, pour équilibrer un échantillon, l'information auxiliaire doit être connue au niveau individuel, tandis que pour caler l'échantillon, seuls les totaux de population sont nécessaires.

En général, il est recommandé d'effectuer de nouveau le calage sur les variables d'équilibrage à l'étape de l'estimation, même si un plus grand nombre de variables de calage sont disponibles. Si l'on n'utilise que de nouvelles variables pour le calage, l'effet de l'équilibrage risque d'être perdu. Il existe toutefois un cas où le calage peut être utilisé sans effectuer un nouveau calage sur les variables d'équilibrage. Il s'agit de la situation où, conditionnellement aux variables de calage, nous pouvons raisonnablement supposer que les variables d'équilibrage ne sont plus corrélées aux variables d'intérêt. Cela peut se produire quand les variables d'équilibrage et de calage sont les mêmes variables mesurées à des périodes différentes et que les variables de calage sont plus récentes.

Quand le coefficient de détermination entre la variable d'intérêt et les variables auxiliaires est égal à un ou proche de cette valeur, le calage est plus efficace en raison du problème d'arrondi de l'échantillonnage équilibré. Quoi qu'il en soit, la meilleure stratégie consiste toujours à utiliser de concert l'échantillonnage équilibré et le calage (voir la simulation dans Deville et Tillé 2004).

7.5 Précision des équations d'équilibrage

Il est possible de prouver, sous des hypothèses raisonnables (voir Deville et Tillé 2004), qu'avec la méthode du cube,

$$\left| \frac{\widehat{X}_j - X_j}{X_j} \right| < O(p/n),$$

où p est le nombre de variables et $O(x)/x$ est une quantité qui reste bornée quand x tend vers l'infini. Sous échantillonnage aléatoire simple,

$$\left| \frac{\widehat{X}_j - X_j}{X_j} \right| = O_p(\sqrt{1/n}),$$

où $O_p(x)/x$ est une quantité qui demeure bornée en probabilité quand x tend vers l'infini.

Les gains de précision sont par conséquent considérables. Le petit problème d'arrondi peut être résolu par un petit calage. Le problème d'arrondi est dû au fait que le tirage d'un échantillon est un problème en nombres entiers. Il se produit également dans la stratification, qui est un cas particulier de l'équilibrage. Dans le cas de la stratification avec répartition proportionnelle, les sommes des probabilités d'inclusion dans les strates ne sont généralement pas des entiers. Donc, les tailles d'échantillon de strate sont obtenues en arrondissant la somme des probabilités d'inclusion dans les strates. Dans la méthode du cube, cet arrondi est effectué automatiquement et aléatoirement de manière à s'assurer que les probabilités d'inclusion soient exactement satisfaites.

7.6 Échantillonnage équilibré dans les enquêtes répétées

L'échantillonnage répété pose un problème important. Celui-ci tient au fait que, si un échantillon équilibré est obtenu par tirage à probabilités d'inclusion inégales, l'échantillon complémentaire n'est pas nécessairement équilibré. En effet, l'égalité

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}=\sum_{k\in U}\mathbf{x}_k$$

n'implique pas que

$$\sum_{k\in U\setminus S}\frac{\mathbf{x}_k}{1-\pi_k}=\sum_{k\in U}\mathbf{x}_k.$$

Ce problème s'est produit dans l'échantillon-maître français. Dans ce plan de sondage, les unités primaires, qui sont des secteurs géographiques, sont sélectionnées avec probabilités inégales qui sont proportionnelles à la taille. Après le tirage de l'échantillon, certaines régions ont demandé des échantillons complémentaires de secteurs qui n'avaient pas été sélectionnés. Cette question est complexe, parce que l'échantillon complémentaire d'un échantillon équilibré n'est plus équilibré et que le but est donc de tirer un échantillon équilibré dans une partie de la population qui n'est plus équilibrée. Tillé et Favre (2004) ont donné quelques méthodes pour coordonner des échantillons équilibrés qui ont été sélectionnés avec probabilités d'inclusion inégales. De manière plus générale, la coordination (au sens de gestion du chevauchement) des échantillons équilibrés peut être difficile quand le plan de sondage est équilibré.

Bien que cela ne soit pas facile, il est possible d'organiser des rotations si tous les échantillons sont sélectionnés ensemble selon un tirage à probabilités d'inclusion égales. En effet, dans ce cas, le complément $\overline{S}=U\setminus S$ des échantillons S est également un échantillon équilibré. Un deuxième échantillon équilibré peut être tiré directement de $\overline{S}$ et ainsi de suite. Cette méthode a été utilisée pour créer cinq groupes de rotation dans l'échantillon-maître français. Les cinq groupes correspondent à cinq échantillons équilibrés de communes.

Si les échantillons sont sélectionnés avec probabilités d'inclusion inégales, certaines solutions sont décrites dans Tillé et Favre (2004). Un cas particulier intéressant peut être résolu facilement, c'est-à-dire quand deux échantillons non chevauchants doivent être sélectionnés avec les mêmes probabilités d'inclusion inégales $\pi_k < 0,5$ pour la même population. D'abord, il faut sélectionner un échantillon S_A équilibré sur $\mathbf{x}_k$ avec les probabilités d'inclusion $\pi_{kA}=2\pi_k$ telles que

$$\sum_{k\in S_A}\frac{\mathbf{x}_k}{2\pi_k}=\sum_{k\in U}\mathbf{x}_k.$$

Ensuite, il faut sélectionner un échantillon S_1 à partir de S_A. Cet échantillon doit être sélectionné avec la probabilité d'inclusion $\pi_{kB}=0,5$ et doit être équilibré sur $\mathbf{x}_k/2\pi_k$, ce qui donne les équations d'équilibrage suivantes :

$$\sum_{k\in S_2}\frac{\mathbf{x}_k/(2\pi_k)}{1/2}=\sum_{k\in S_A}\frac{\mathbf{x}_k}{2\pi_k}=\sum_{k\in U}\mathbf{x}_k.$$

L'échantillon $S_2=S_A\setminus S_1$ est également équilibré.

Si la population évolue au cours du temps (décès et naissances), l'organisation d'une rotation devient beaucoup plus compliquée. La difficulté se manifeste déjà pour les échantillons stratifiés. Néanmoins, dans le cas de la stratification, plusieurs solutions raisonnables existent (voir, entre autres, De Ree 1999 ; Hesse 1998 ; Rivière 1999 ; Nedyalkova, Péa et Tillé 2006).

7.7 Principales implémentations de l'échantillonnage équilibré

Une application SAS/IML® a d'abord été programmée par trois étudiants de l'École Nationale de la Statistique et de l'Analyse de l'Information (ENSAI) (Bousabaa et coll. 1999). La version officielle de l'Institut national de la statistique et des études économiques (Insee), produite par Tardieu (2001) et par Rousseau et Tardieu (2004), est maintenant accessible sur le site Web de l'Insee. Une autre version SAS/IML® produite par Chauvet et Tillé (2005a, 2005b, 2006) est également disponible sur le site Web de l'Université de Neuchâtel. En langage R, le progiciel d'échantillonnage (Tillé et Matei 2007) permet d'utiliser la méthode du cube. Ces progiciels sont gratuits, accessibles sur Internet et faciles à utiliser.

Les programmes existants écrits au moyen du langage R ou SAS/IML® ne présentent aucune limite en ce qui concerne la taille de population. Il est possible d'exécuter une

application comptant 40 variables équilibrées. Afin de sélectionner l'échantillon, les temps de calcul augmentent proportionnellement à $N \times p^2$, où N est la taille de population et p est le nombre de variables d'équilibrage. Il est donc possible de tirer un échantillon dans une population de plusieurs millions d'unités statistiques.

Remerciements

Le présent article a été rédigé à la suite d'une invitation à présenter une communication à la conférence de la Demographic Statistical Methods Division du U.S. Census Bureau, tenue en juin 2008. L'auteur remercie le U.S. Census Bureau, en particulier Patrick Flanagan, sans lequel le présent article n'aurait jamais été rédigé. L'auteur remercie également un rédacteur associé et deux examinateurs anonymes de leurs commentaires et corrections fort utiles qui l'ont aidé à améliorer le manuscrit.

Bibliographie

Ardilly, P. (1991). Échantillonnage représentatif optimum à probabilités inégales. *Annales d'Économie et de Statistique*, 23, 91-113.

Ardilly, P. (2006). *Les Techniques de Sondage*. Technip, Paris.

Bardaji, J. (2001). Un an après la sortie d'un contrat emploi consolidé : près de six chances sur dix d'avoir un emploi. *Premières Informations Synthèses, Direction de l'Animation de la Recherche des Etudes et des Statistiques* (*DARES*) *du Ministère du Travail des relations sociales et de la solidarité*, 43, 3, 1-8.

Bertrand, P., Christian, B., Chauvet, G. et Grosbras, J.-M. (2004). Plans de sondage pour le recensement rénové de la population. Dans *Séries Insee Méthodes* : *Actes des Journées de Méthodologie Statistique*, Paris. Insee.

Biggeri, L., et Falorsi, P.D. (2006). A probability sample strategy for improving the quality of the consumer price index survey using the information of the business register. Dans *Proceedings of the Conference of European Statisticians Group of Experts on Consumer Price Indices*, huitième réunion, Genève, 10-12 mai 2006.

Bousabaa, A., Lieber, J. et Sirolli, R. (1999). La macro cube. Rapport technique, Ensai, Rennes.

Breidt, F.J., et Chauvet, G. (2010a). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 479-487.

Breidt, F.J., et Chauvet, G. (2010b). Penalized balanced sampling. Document de travail, Ensai.

Brewer, K.R.W. (1975). A simple procedure for πpswor. *Australian Journal of Statistics*, 17, 166-172.

Brewer, K.R.W. (1999). Design-based or prediction-based inference? Stratified random *vs* stratified balanced sampling. *Revue Internationale de Statistique*, 67, 35-47.

Chauvet, G. (2007). *Méthodes de Bootstrap en Population Finie*. Thèse de doctorat, Université Rennes 2.

Chauvet, G. (2009). Échantillonnage équilibré stratifié. *Techniques d'enquête*, 35, 123-127.

Chauvet, G., Bonnery, D. et Deville, J.-C. (2010a). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141, 2, 984-994.

Chauvet, G., Deville, J. et Haziza, D. (2010b). Adapting the cube algorithm for balanced random imputation in surveys. Rapport technique, Ensai, Rennes.

Chauvet, G., Deville, J. et Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*.

Chauvet, G., et Tillé, Y. (2005a). *Fast SAS Macros for balancing Samples*: *user's guide*. Software Manual, Université de Neuchâtel, http://www2.unine.ch/statistics/page10890.html.

Chauvet, G., et Tillé, Y. (2005b). New SAS macros for balanced sampling. Dans *Journées de Méthodologie Statistique*, Insee, Paris.

Chauvet, G., et Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21, 9-31.

Chipperfield, J. (2009). An evaluation of cube sampling for ABS household surveys. Rapport technique, Australian Bureau of Statistics.

Christine, M. (2006). Use of balanced sampling in the framework of the master sample for french household surveys. Dans *Joint Statistical Meeting of the American Statistical Association*, Seattle août 2006.

Christine, M., et Wilms, L. (2003). Problèmes théoriques et pratiques de la construction de l'« EMEX » : comment améliorer la précision des extensions régionales des enquêtes nationales grâce à un échantillonnage additionnel ? Dans le *Recueil* : *Symposium 2003*, *Défis Reliés à la Réalisation d'Enquêtes pour la Prochaine Décennie*, Statistique Canada, Ottawa.

Cumberland, W.G., et Royall, R.M. (1981). Prediction models in unequal probability sampling. *Journal of the Royal Statistical Society*, B, 43, 353-367.

Cumberland, W.G., et Royall, R.M. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society*, B, 50, 118-124.

da Silva, A.D., da Silva Borges, A., Aires Leme, R. et Moura Reis Miceli, A.P. (2006). Modalidades alternativas de censo demográfico: o cenário internacional a partir das experiências dos estados unidos, França, holanda, israel e alemanha. Rapport technique, Instituto Brasileiro de Geografia e Estatística.

D'Alò, M., Di Consiglio, L., Falorsi, S. et Solari, F. (2006). Small area estimation of the italian poverty rate. *Statistics in Transition*, 7, 771-784.

De Ree, S.J.M. (1999). Co-ordination of business samples using measured response burden. Dans *Contributed paper*, *52th Session of the ISI Helsinki*.

Desplanques, G. (2000). La rénovation du recensement de la population. Dans *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFDS-Insee sur la rénovation du recensement*, 2-5.

Dessertaine, A. (2006). Sondages et séries temporelles : une application pour la prévision de la consommation électrique. Dans *Actes des journées Françaises de Statistique 2006*, Clamart, France.

Dessertaine, A. (2007). Sampling and data-stream: Some ideas to built balanced sampling using auxiliary Hilbertian informations. Dans *Proceedings of 56th the International Statistical Institute Conference*: *IPM56 - New methods of sampling*, Lisboa, Portugal.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. Dans *Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*, Örebro (La Suède).

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. Dans *Recueil de la Section des méthodes d'enquêtes des communications présentées au 26ème congrès de la Société Statistique du Canada*, 103-110, Sherbrooke.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

Deville, J.-C. (2005). Imputation stochastique et échantillonnage équilibré. Rapport technique, École Nationale de la Statistique et de l'Analyse de l'Information.

Deville, J.-C. (2006). Stochastic imputation using balanced sampling. Dans *Joint Statistical Meeting of the American Statistical Association*, Seattle août 2006.

Deville, J.-C., Grosbras, J.-M. et Roth, N. (1988). Efficient sampling algorithms and balanced sample. Dans *COMPSTAT*, *Proceedings in Computational Statistics*, Heidelberg. Physica Verlag, 255-266.

Deville, J.-C., et Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Dudoignon, L., et Vanheuverzwyn, A. (2006). Tirage d'un échantillon à probabilités inégales : application au panel Médiamat. Dans *Actes de des Journées de Méthodologie Statistique*, 1-10.

Dumais, J., Bertrand, P. et Kauffmann, B. (2000). Sondage, estimation et précision dans la rénovation du recensement de la population. Dans *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFDS-Insee sur la rénovation du recensement*, 6-26.

Dumais, J., et Isnard, M. (2000). Le sondage de logements dans les grandes communes dans le cadre du recensement rénové de la population. Dans *Séries Insee Méthodes* : *Actes des Journées de Méthodologie Statistique*, Paris. Insee, 100, 37-76.

Durr, J.-M., et Dumais, J. (2001). La rénovation du recensement français. Dans le *Recueil* : *Symposium 2001*, *La Qualité des Données d'un Organisme Statistique : Une Perspective Méthodologique*, Statistique Canada, Ottawa.

Durr, J.-M., et Dumais, J. (2002). La rénovation du recensement français. *Techniques d'enquête*, 28, 47-53.

Even, K. (2002). Improved tool for evaluating employment and vocational training policy: Panel of beneficiaries. *Premières Informations Synthèses*, *Direction de l'Animation de la Recherche des Études et des Statistiques (DARES) du Ministère du Travail des relations sociales et de la solidarité*, 33, 1, 1-7.

Falorsi, P.D., et Righi, P. (2008). Une approche d'échantillonnage équilibré pour des plans de sondage à stratification multidimensionnelle pour l'estimation pour petits domaines. *Techniques d'enquête*, 34, 247-259.

Fecteau, S., et Jocelyn, W. (2006). Une application de l'échantillonnage équilibré : le plan de sondage des entreprises non incoprporées. Dans *Méthodes d'enquêtes et sondages* : *pratiques européenne et nord-américaine*, (Éds., P. Lavallée et L.-P. Rivest), Paris. Dunod, 405-410.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Fuller, W.A. (2010). Replication variance estimation for rejective sampling. Dans *Seminar of Statistics Canada*, June 2010, Ottawa.

Gini, C. (1928). Une application de la méthode représentative aux matériaux du dernier recensement de la population italienne (1er décembre 1921). *Bulletin of the International Statistical Institute*, 23, 2, 198-215.

Gini, C., et Galvani, L. (1929). Di una applicazione del metodo rappresentative all'ultimo censimento Italiano della popolazione (1° decembri, 1921). *Annali di Statistica*, Séries 6, 4, 1-107.

Gismondi, R. (2007). Quick estimation of tourist nights spent in italy. *Statistical Methods and Applications*, 16, 141-168.

Hájek, J. (1981). *Sampling from a Finite Population*. New York : Marcel Dekker.

Hedayat, A.S., et Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, 44, 237-247.

Hesse, C. (1998). Sampling co-ordination: A review by country. Rapport technique E9908, Direction des Statistique d'Entreprises, Insee, Paris.

Jocelyn, W. (2006). Sampling and estimation strategies for the canadian unincorporated business population. Dans *Joint Statistical Meeting of the American Statistical Association*, Seattle août 2006.

Kiaer, A. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9, 2, 176-183.

Kiaer, A. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 11, 1, 180-185.

Kiaer, A. (1903). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 13, 1, 66-78.

Kiaer, A. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique*, 14, 1, 119-134.

Kott, P.S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *The American Statistician*, 40, 202-204.

Langel, M., et Tillé, Y. (2010). Corrado Gini, a pioneer in balanced sampling and inequality theory. Rapport technique, University of Neuchatel.

Legg, J.C., et Yu, C.L. (2010). Comparaison de méthodes de restriction de l'ensemble d'échantillons. *Techniques d'enquête*, 36, 75-87.

Lesage, E. (2008). Contraintes d'équilibrage non linéraires. Dans *Méthodes d'enquêtes* : *applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales*, (Éds., P. Guilbert, D. Haziza, A. Ruiz-Gazen et Y. Tillé), Paris. Dunod, 285-289.

Marí, G., Barbará, G., Mitas, G. et Passamonti, S. (2007a). Construcción de un estimador de variancia para muestras balanceadas estratificadas. Dans *XXXV Coloquio Argentino de Estadística. Mar del Plata*, L'Argentine. 22, 23 y 24 de Octubre de 2007.

Marí, G., Barbará, G., Mitas, G. et Passamonti, S. (2007b). Muestras equilibradas en poblaciones finitas: un estudio comparativo en muestras de explotaciones agropecuarias. Dans *Undécimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística, noviembre de 2007*, Universidad Nacional de Rosario, L'Argentine.

Nedyalkova, D., Péa, J. et Tillé, Y. (2006). A review of some current methods of coordination of stratified samples. introduction and comparison of new methods based on microstrata. Rapport technique, Université de Neuchâtel.

Nedyalkova, D., and Tillé, Y. (2009). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95, 521-537.

Nedyalkova, D., and Tillé, Y. (2010). Bias robustness and efficiency in model-based inference. Rapport technique, Université de Neuchâtel.

Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. Graduate School; U.S. Department of Agriculture, Washington.

Périé, P. (2008). Échantillonnage à entropie maximale sous contraintes : un algorithme rapide basé sur l'optimisation linéaire en nombres binaires. In *Méthodes d'enquêtes* : *applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales*, (Éds., P. Guilbert, D. Haziza, A. Ruiz-Gazen et Y. Tillé), Paris. Dunod, 294-299.

Rivière, P. (1999). Coordination of samples: The microstrata methodology. Dans *13th International Roundtable on Business Survey Frames*, Paris. Insee.

Rousseau, S., et Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Rapport technique, Insee, Paris.

Royall, R.M. (1976a). Likelihood functions in finite population sampling theory. *Biometrika*, 63, 605-614.

Royall, R.M. (1976b). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Royall, R.M. (1988). The prediction approach to sampling theory. In *Handbook of Statistics Volume 6*: *Sampling*, (Éds., P.R. Krishnaiah et C.R. Rao), Amsterdam. Elsevier/North-Holland, 399-413.

Royall, R.M., et Pfeffermann, D. (1982). Balanced samples and robust bayesian inference in finite population sampling. *Biometrika*, 69, 401-409.

Sunter, A. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

Tardieu, F. (2001). Échantillonnage équilibré: de la théorie à la pratique. Rapport technique, Insee, Paris.

Thionet, P. (1953). *La théorie des sondages*. Insee, Imprimerie nationale, Paris.

Tillé, Y. (2001). *Théorie des sondages* : *échantillonnage et estimation en populations finies*. Dunod, Paris.

Tillé, Y. (2006a). Balanced sampling by means of the cube method. Dans *Joint Statistical Meeting of the American Statistical Association*, Seattle août 2006.

Tillé, Y. (2006b). *Sampling Algorithms*. New York : Springer.

Tillé, Y., et Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika*, 91, 913-927.

Tillé, Y., et Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74, 31-37.

Tillé, Y., et Matei, A. (2007). *The R Package Sampling*. The Comprehensive R Archive Network, http://cran.r-project.org/, Manual of the Contributed Packages.

Tirari, M. (2006). Le plan de sondage équilibré et l'estimation du total d'une population finie. Dans *Méthodes d'enquêtes et sondages* : *pratiques européenne et nord-américaine*, (Éds., P. Lavallée et L.-P. Rivest), Paris, Dunod, 411-416.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. New York : John Wiley & Sons, Inc.

Wilms, L. (2000). Présentation de l'échantillon-maître en 1999 et application au tirage des unités primaires par la macro cube. Dans *Séries Insee Méthodes* : *Actes des Journées de Méthodologie Statistique*, Paris. Insee.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, A109, 12-43.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. Charles Griffin, Londres, L'Angleterre, troisième édition.