De nouvelles macros SAS d'échantillonnage équilibré

Guillaume CHAUVET (*), Yves TILLE (**)

(*) CREST-ENSAI, Rennes, France, (**) Université de Neuchâtel, Suisse

1. Un algorithme rapide d'échantillonnage équilibré

Cette partie est la traduction de l'article [2] co-écrit avec Yves Tillé.

1.1. Introduction

La méthode du Cube, qui permet la sélection d'échantillons équilibrés, a été développée à l'ENSAI France) (cf. [4], [5] et [8]) et des étudiants ont à l'origine écrit le programme. Le programme actuellement utilisé à l'Insee a été écrit par Frédéric Tardieu, et finalisé par Bernard Weytens sur les recommandations de l'Unité de Méthodes Statistiques et de la Maîtrise d'œuvre Méthodologique du Nouveau Recensement.

La méthode a été d'abord consacrée à la sélection d'unités primaires dans un sondage à deux degrés, car le temps d'exécution était proportionnel au carré de la taille de la population. Cette méthode a ensuite été appliquée à plusieurs problèmes statistiques importants. Par exemple, les groupes de rotation de communes et d'adresses du Nouveau Recensement français ont été sélectionnés à l'aide de la méthode du Cube ([1] ; [6]). Cette méthode est en fait une famille d'algorithmes dont l'implémentation peut prendre différentes formes. Nous proposons ici une implémentation très rapide. L'originalité consiste à appliquer l'étape de base à un sous-ensemble d'unités et non à la population entière restante. Ce sous-ensemble évolue à chaque étape de l'algorithme, et le temps d'exécution ne dépend plus du carré de la taille de la population.

Dans la section 1.2, nous définissons les notations. Dans la section 1.3, nous donnons un bref rappel de la méthode du Cube. Le nouvel algorithme est proposé dans la section 1.4. Ensuite, en section 1.5, nous discutons de l'implémentation de l'algorithme rapide et nous présentons des résultats numériques. Finalement, nous montrons en section 1.6 que cet algorithme peut être appliqué au problème d'échantillonnage à probabilités inégales, et en section 1.7 qu'il peut être vu comme une généralisation du tirage poissonnien.

1.2. Notations et échantillonnage équilibré

Considérons une population finie U de taille N dont les unités peuvent être identifiées par des numéros $k \in \{1,...,N\}$. Le but est d'estimer le total $Y = \sum_{k \in U} y_k$ d'une variable d'intérêt y qui prend les valeurs $y_k, k \in U$ sur les unités de la population. On suppose aussi que les vecteurs de valeurs $\mathbf{x_k} = \left(x_{k1},...,x_{kj},...,x_{kp}\right)'$ prises par p variables auxiliaires sont connues pour toutes les unités de la population. On suppose sans perte de généralité que les p vecteurs $\left(x_{1j},...,x_{kj},...,x_{Nj}\right)', j = 1,...,p$ sont linéairement indépendants.

Un échantillon est représenté par un vecteur $\mathbf{s} = \left(s_1,...,s_k,...,s_N\right)'$, où s_k prend la valeur 1 si k est dans l'échantillon et 0 sinon. Un plan de sondage $p(.)$ est une distribution de probabilité sur l'ensemble $\mathsf{S} = \{0,1\}^N$ de tous les échantillons possibles. L'échantillon aléatoire $\mathbf{S}$ prend la valeur $\mathbf{s}$ avec la probabilité $\Pr(\mathbf{S} = \mathbf{s}) = p(\mathbf{s})$. La probabilité d'inclusion de l'unité k est la probabilité $\pi_k = \Pr(S_k = 1)$ que k soit dans l'échantillon et la probabilité d'inclusion d'ordre 2 est la probabilité $\pi_{kl} = \Pr(S_k = 1 \text{ and } S_l = 1)$ que deux unités distinctes appartiennent conjointement à l'échantillon. L'estimateur de Horvitz-Thompson donné par $\hat{Y} = \sum_{k \in U} \frac{S_k y_k}{\pi_k}$ est un estimateur sans biais de Y. L'estimateur de Horvitz-Thompson du $j^{ème}$ total auxiliaire $X_j = \sum_{k \in U} x_{kj}$ est $\hat{X}_j = \sum_{k \in U} \frac{S_k x_{kj}}{\pi_k}$. Le vecteur d'estimateurs de Horvitz-Thompson, $\hat{\mathbf{X}} = \sum_{k \in U} \frac{S_k \mathbf{x_k}}{\pi_k}$, estime sans biais le vecteur des totaux des variables auxiliaires $\mathbf{X} = \sum_{k \in U} \mathbf{x_k}$.

Le but est de construire un plan de sondage équilibré, c'est à dire un plan de sondage tel que $\hat{\mathbf{X}} = \mathbf{X}$. Néanmoins, un plan de sondage exact n'existe pas dans la plupart des cas. L'objectif est donc de trouver un plan de sondage approximativement équilibré au sens où $\hat{\mathbf{X}} \approx \mathbf{X}$. Si $\mathbf{a_k} = \frac{\mathbf{x_k}}{\pi_k}$, et $\mathbf{A} = \left(\mathbf{a_1},...,\mathbf{a}_k,...,\mathbf{a}_N\right)$, alors un plan de sondage équilibré est tel que

$$\mathbf{AS} = \mathbf{A\pi} \qquad (1)$$

Les équations définies par le système (1) seront appelées équations d'équilibrage.

1.3. La méthode du Cube

La méthode du Cube est constituée de deux phases appelées phase de vol et phase d'atterrissage. Pendant la phase de vol, les contraintes sont toujours exactement respectées. L'objectif est d'arrondir aléatoirement presque toutes les probabilités d'inclusion à 0 ou 1. La phase d'atterrissage consiste à solutionner au mieux le fait que les équations d'équilibrage (1) ne puissent être exactement respectées. La phase de vol est décrite dans l'Algorithme 1.

Algorithme 1 : Procédure générale d'équilibrage : phase de vol

Initialiser d'abord à $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Ensuite, au temps $t = 0,...,T$ répéter les trois étapes suivantes.

Etape 1 : Générer un vecteur quelconque $\mathbf{u}(t) = \{u_k(t)\} \neq \mathbf{0}$, aléatoire ou non, tel que $\mathbf{u}(t)$ soit dans le noyau de la matrice $\mathbf{A}$, et que $u_k(t) = 0$ si $\pi_k(t)$ est un entier

Etape 2 : Calculer $\lambda_1^*(t)$ et $\lambda_2^*(t)$, les plus grandes valeurs de $\lambda_1(t)$ et $\lambda_2(t)$ telles que $0 \leq \boldsymbol{\pi}(t) + \lambda_1(t) \leq 1$, et $0 \leq \boldsymbol{\pi}(t) - \lambda_2(t) \leq 1$. Notons que $\lambda_1^*(t) > 0$ et $\lambda_2^*(t) > 0$.

Etape 3 : Sélectionner

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\mathbf{u}(t) \text{ avec la probabilité } q(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\mathbf{u}(t) \text{ avec la probabilité } 1 - q(t) \end{cases} \quad (2)$$

où $q(t) = \dfrac{\lambda_2^*(t)}{\lambda_1^*(t) + \lambda_2^*(t)}$.

Cette procédure générale est répétée jusqu'à ce qu'il ne soit plus possible de réaliser l'étape 1.

Si T est la dernière étape de l'algorithme et $\boldsymbol{\pi}^* = \boldsymbol{\pi}(T)$, alors Deville et Tillé ([4]) montrent que :

a) $E(\boldsymbol{\pi}^*) = \boldsymbol{\pi}$

b) $\mathbf{A}\boldsymbol{\pi}^* = \mathbf{A}\boldsymbol{\pi}$

c) Si $q = Card\{k; 0 < \pi_k^* < 1\}$ alors $q \leq p$, où p est le nombre de variables auxiliaires

Le vecteur $\boldsymbol{\pi}^*$ peut être un échantillon, mais dans la plupart des cas il y a au plus q éléments non entiers dans $\boldsymbol{\pi}^*$. Si $q > 0$, le problème d'arrondi est réglé par la phase d'atterrissage. La première solution consiste à relâcher une contrainte et à relancer à nouveau la phase de vol, jusqu'à ce qu'il ne soit plus possible de "bouger" à nouveau à l'intérieur de l'hyperplan des contraintes. Les contraintes sont donc relâchées successivement. La seconde solution utilise un programme linéaire pour obtenir le meilleur plan de sondage équilibré approché (cf. Deville et Tillé [4]). Le temps d'exécution est donné essentiellement par la phase de vol. La nouvelle implémentation ne concerne que la phase de vol, et la phase d'atterrissage reste inchangée.

1.4. Une implémentation très rapide

Le but de cette nouvelle implémentation est d'obtenir une réduction du temps d'exécution. Dans l'algorithme général, la recherche d'un vecteur $\mathbf{u}$ de $Ker\mathbf{A}$ est très coûteuse. L'idée de base est d'utiliser une sous-matrice $\mathbf{B}$ contenant uniquement $p+1$ colonnes de $\mathbf{A}$. Notons que le nombre de variables p est plus petit que la taille de la population N , et que $rang(\mathbf{B}) \leq p$. La dimension du noyau de $\mathbf{B}$ est donc supérieure ou égale à 1.

Un vecteur $\mathbf{v}$ de $Ker\mathbf{B}$ peut alors être utilisé pour construire un vecteur $\mathbf{u}$ de $Ker\mathbf{A}$ en complétant $\mathbf{v}$ avec des zéros pour les colonnes de $\mathbf{A}$ qui ne sont pas dans $\mathbf{B}$. Avec cette idée, tous les calculs peuvent être faits uniquement sur $\mathbf{B}$. Cette méthode est décrite dans l'Algorithme 2.

Si $\tilde{T}$ est la dernière étape de l'algorithme et $\tilde{\boldsymbol{\pi}} = \boldsymbol{\pi}(\tilde{T})$, nous avons :

a) $E(\tilde{\boldsymbol{\pi}}) = \boldsymbol{\pi}$

b) $\mathbf{A}\tilde{\boldsymbol{\pi}} = \mathbf{A}\boldsymbol{\pi}$

c) Si $\tilde{q} = Card\{k; 0 < \tilde{\pi}_k < 1\}$ alors $\tilde{q} \leq p$, où p est le nombre de variables auxiliaires.

Dans le cas où certaines contraintes peuvent être exactement satisfaites, on peut poursuivre la phase de vol. Supposons que $\mathbf{C}$ désigne la matrice contenant les colonnes de $\mathbf{A}$ qui correspondent à des valeurs non entières de $\tilde{\boldsymbol{\pi}}$, et que $\boldsymbol{\varphi}$ est le vecteur des valeurs non entières de $\tilde{\boldsymbol{\pi}}$. Si $\mathbf{C}$ n'est pas de plein rang, une ou plusieurs étapes de l'algorithme général peuvent encore être appliquées à $\mathbf{C}$ et $\boldsymbol{\varphi}$. Un retour à l'Algorithme 1 est donc nécessaire pour les dernières étapes

1.5. Implémentation et résultats numériques

L'implémentation de l'algorithme rapide est assez simple. La matrice $\mathbf{A}$ ne doit jamais être chargée entièrement, et peut donc rester dans un fichier qui peut être lu séquentiellement. C'est pour cette raison qu'il n'existe plus aucune restriction sur la taille de la population. La recherche d'un vecteur $\mathbf{u}$ dans le noyau de la sous-matrice $\mathbf{B}$ limite le choix de la direction $\mathbf{u}$. Dans la plupart des cas, une seule direction est possible. Afin d'augmenter l'entropie du plan de sondage, on peut éventuellement réaliser un tri aléatoire sur les unités avant d'appliquer l'algorithme.

L'Algorithme 2 a été implémenté sous forme d'une macro SAS-IML. Nous avons testé les capacités de l'algorithme à sélectionner des échantillons dans de grandes populations avec de nombreuses variables d'équilibrage. Nous avons utilisé une population de 313,702 unités, correspondant aux adresses des grandes communes (comptant 10 000 habitants ou plus) de la région Rhône-Alpes. Toutes les unités sont sélectionnées avec les mêmes probabilités d'inclusion égales à $\frac{1}{5}$. Les variables d'équilibrage sont :

- Une constante afin d'obtenir une taille fixe d'échantillon
- 18 variables socio-démographiques, qui sont présentées en Table 1
- 81 variables qui sont les produits d'une variable indicatrice de la présence d'adresses dans les 81 communes et du nombre de ménages dans les adresses

Algorithme 2 : Algorithme rapide pour la phase de vol

(a) *Initialisation*

(i) Les unités avec des probabilités d'inclusion égales à 0 ou 1 sont retirées de la population avant d'appliquer l'algorithme, de sorte que toutes les unités restantes vérifient $0 < \pi_k < 1$.

(ii) Les probabilités d'inclusion sont chargées dans le vecteur $\boldsymbol{\pi}$

(iii) Le vecteur $\boldsymbol{\psi}$ est constitué des $p+1$ premiers éléments de $\boldsymbol{\pi}$

(iv) Un vecteur de rangs $\mathbf{r} = (1,2,\ldots,p,p+1)'$ est créé

(v) La matrice $\mathbf{B}$ est constituée des $p+1$ premières colonnes de $\mathbf{A}$

(vi) On initialise $k = p+2$

(b) *Boucle de base*

(i) On choisit un vecteur $\mathbf{u}$ dans le noyau de $\mathbf{B}$

(ii) Seul $\boldsymbol{\psi}$ est modifié (et non le vecteur $\boldsymbol{\pi}$) à l'aide de la technique de base

Calculer λ_1^* et λ_2^*, les plus grandes valeurs de λ_1 et λ_2 telles que $0 \le \boldsymbol{\psi} + \boldsymbol{\lambda}_1 \mathbf{u} \le 1$, et $0 \le \boldsymbol{\psi} - \boldsymbol{\lambda}_2 \mathbf{u} \le 1$. Noter que $\lambda_1^* > 0$ et $\lambda_2^* > 0$.

(iii) Sélectionner

$$\boldsymbol{\psi} = \begin{cases} \boldsymbol{\psi} + \boldsymbol{\lambda}_1^*(t)\mathbf{u} & \textit{avec la probabilité } q \\ \boldsymbol{\psi} - \boldsymbol{\lambda}_2^*(t)\mathbf{u} & \textit{avec la probalbilité } 1-q \end{cases} \qquad (2)$$

où $q = \dfrac{\lambda_2^*}{\lambda_1^* + \lambda_2^*}$.

(iv) (*Les unités correspondant à des $\psi[i]$ entiers sont retirées de $\mathbf{B}$ et sont remplacés par les probabilités d'inclusion de nouvelles unités. L'algorithme s'arrête à la fin du fichier.)*

Pour $i = 1,\ldots,p+1,$

Si $\psi[i] = 0$ ou $\psi[i] = 1$ alors

$\boldsymbol{\pi}[\mathbf{r}[i]] = \boldsymbol{\psi}[i]$

$\mathbf{r}[i] = k$

Si $k \le N$ alors $\boldsymbol{\psi}[i] = \boldsymbol{\pi}[k]$

Pour $j = 1,\ldots,p, \mathbf{B}[i,j] = \mathbf{A}[k,j]$

K=k+1

Sinon aller à l'étape (c) (i)

(v) Aller à l'étape (b) (i)

(c) *Fin de la première partie de la phase de vol*

(i) Pour $i = 1,\ldots,p+1,\ \boldsymbol{\pi}[\mathbf{r}[i]] = \boldsymbol{\psi}[i]$

Table 1 : Liste des variables socio-démographiques

NLOG	Nombre de ménages
NLOGCO	Nombre de ménages dans les adresses collectives
H0019	Nombre d'hommes de moins de 20 ans
H2039	Nombre d'hommes de 20 à 39 ans
H4059	Nombre d'hommes de 40 à 59 ans
H6074	Nombre d'hommes de 60 à 74 ans
H7599	Nombre d'hommes de 75 ans et plus
F0019	Nombre de femmes de moins de 20 ans
F2039	Nombre de femmes de 20 à 39 ans
F4059	Nombre de femmes de 40 à 59 ans
F6074	Nombre de femmes de 60 à 74 ans
F7599	Nombre de femmes de 75 ans et plus
ACTIFS	Nombre d'actifs
INACTIFS	Nombre d'inactifs
NATFN	Nombre de personnes françaises de naissance
NATFA	Nombre de personnes françaises par acquisition
NATHE	Nombre de personnes étrangères hors Union Européenne
NATUE	Nombre de personnes étangères de l'Union Européenne

Les 81 dernières variables assurent que le nombre de ménages est équilibré dans chaque commune. Cent variables d'équilibre sont donc utilisées. Un échantillon de 62,741 adresses a été sélectionné, à l'aide d'un ordinateur personnel (Penthium 3, 1 Gh). La population a été triée par taille d'adresse décroissante. La sélection a été effectuée en 1 heure et 50 minutes environ. La condition de taille fixe est parfaitement réalisée. Nous avons ensuite calculé les ratios du carré de la différence entre les estimateurs de Horvitz-Thompson et les totaux et les variances de l'estimateur de Horvitz-Thompson pour un sondage aléatoire simple :

$$R = \frac{\left(\hat{X}_j - X\right)^2}{Var_{simple}(\hat{X}_j)}$$

Ces ratios sont présentés dans le Tableau 2, et montrent une amélioration très importante de la précision.

Table 2 : Ratios du carré de la différence entre les estimateurs de Horvitz-Thompson et les totaux estimés et les variances de l'estimateur de Horvitz-Thompson pour un sondage aléatoire simple

Variable	Ratio	Variable	Ratio	Variable	Ratio
NLOG	$2.7\ 10^{-5}$	H7599	$1.2\ 10^{-4}$	ACTIFS	$9.7\ 10^{-7}$
NLOGCO	$2.5\ 10^{-5}$	F0019	$6.0\ 10^{-6}$	INACTIFS	$2.3\ 10^{-5}$
H0019	$1.4\ 10^{-6}$	F2039	$1.9\ 10^{-6}$	NATFN	$2.0\ 10^{-5}$
H2039	$8.2\ 10^{-5}$	F4059	$2.0\ 10^{-6}$	NATFA	$1.3\ 10^{-5}$
H4059	$3.0\ 10^{-7}$	F6074	$8.6\ 10^{-5}$	NATHE	$7.8\ 10^{-5}$
H6074	$3.2\ 10^{-5}$	F7599	$6.7\ 10^{-5}$	NATUE	$5.8\ 10^{-4}$

1.6. Cas de l'échantillonnage à probabilités inégales

Quand $p = 1$ et que la seule variable auxiliaire est $x_k = \pi_k$, alors l'échantillonnage équilibré revient à réaliser un échantillonnage à probabilités inégales de taille fixe. Dans ce cas, $\mathbf{A} = (1,...,1)$. A chaque étape, la matrice $\mathbf{B}$ vaut $(1,1)$ et $\mathbf{u}$ vaut $(-1,1)$. L'algorithme peut alors être simplifié de la façon suivante :

Algorithme 3 : Méthode du pivot pour des probabilités d'inclusion inégales

- Effectuer éventuellement un tri aléatoire sur les données
- Définition : a, b, u réels et i, j, k : entiers
- $a = \pi_1$; $b = \pi_2$; $i = 1$; $j = 2$
- Pour $k = 1,...,N$: $s_k = 0$
- k=3
- Tant que $k \le n$

  $u =$ variable aléatoire uniforme dans $[0,1]$

  Si $a + b > 1$ alors

  Si $u < \dfrac{1-b}{2-a-b}$: $b = a + b - 1; a = 1$

  Sinon : $a = a + b - 1; a = 1$

  Si $k \le N$ alors

  Si $u < \dfrac{b}{a+b}$: $b = a + b; a = 0$

  Sinon : $a = a + b; b = 0$

  Si a est un entier et $k \le n$ alors $s_i = a; a = \pi_k; i = k; k = k + 1$

  Si b est un entier et $k \le n$ alors $s_j = b; b = \pi_k; j = k; k = k + 1$
- $s_i = a; s_j = b$

L'Algorithme 3 est une méthode très simple d'échantillonnage à probabilités inégales. En fait, c'est une implémentation de la méthode du Pivot proposée par Deville et Tillé ([3]) dans le cadre de la méthode de scission.

1.7. Cas de l'échantillonnage poissonnien

Deville (sans date) remarque qu'avec $p = 0$, i.e. sans équations d'équilibrage, l'algorithme rapide est équivalent à un plan de sondage poissonnien. L'Algorithme 2 peut en effet être alors réécrit de la façon suivante :

Algorithme 4 :

- Définition : ψ réel
- Pour $k = 1,...,N$: $\psi = \pi_k$, u est un scalaire non nul

  Si $u > 0$ alors

  $$\lambda_1^* = \frac{1-\psi}{u};\ \lambda_2^* = \frac{\psi}{u}$$

  Sélectionner

  $$\psi = \begin{cases} \psi + \lambda_1^* u = 1 \text{ avec probabilité } q \\ \psi - \lambda_2^* u = 0 \text{ avec probabilité } 1\text{-}q \end{cases} \quad (2)$$

  où $q = \dfrac{\lambda_2^*}{\lambda_1^* + \lambda_2^*} = \pi_k$.

  Si $u < 0$ alors

  $$\lambda_1^* = -\frac{\psi}{u};\ \lambda_2^* = -\frac{1-\psi}{u}$$

  Sélectionner

  $$\psi = \begin{cases} \psi + \lambda_1^* u = 0 \text{ avec probabilité } q \\ \psi - \lambda_2^* u = 1 \text{ avec probabilité } 1\text{-}q \end{cases} \quad (2)$$

  où $q = \dfrac{\lambda_2^*}{\lambda_1^* + \lambda_2^*} = 1 - \pi_k$.

ce qui définit un plan de Poisson.

2. Un nouveau programme d'échantillonnage équilibré

2.1. Les données en entrée

Les données relatives à la population dans laquelle on veut sélectionner un échantillon équilibré doivent être placées dans une table SAS, contenant toutes les unités de la population et au moins:

- Une variable identifiante
- La variable des probabilités d'inclusion
- Les variables d'équilibrage

Cette table ne doit pas contenir de valeurs manquantes pour les variables indiquées ci-dessus. La variable des probabilités d'inclusion et les variables d'équilibrage doivent être numériques.

2.2. Syntaxe de la macro

2.2.1. Paramètres relatifs à la base de sondage

Tous ces paramètres sont obligatoires.

- **BASE** = nom d'une librairie SAS
  Nom de la librairie SAS contenant la base de sondage sous forme d'une table SAS

- **DATA** = nom d'une table SAS
  Nom de la table SAS contenant les données relatives à la base de sondage

- **ID** = variable
  Nom de la variable qui identifie les unités de la population

- **PI** = variable
  Nom de la variable des probabilités d'inclusion

- **CONTR** = variable(s)
  Noms des variables sur lesquelles l'échantillon est équilibré. Ces noms doivent être séparés avec des blancs.

2.2.2. Paramètres relatifs à l'échantillonnage

Tous ces paramètres sont optionnels.

- **ATTER** = option

  Donne l'option sélectionnée pour la phase d'atterrissage. Les valeurs possibles sont :

  - ATTER = 1
    Les variables d'équilibre sont abandonnées progressivement. La dernière variable du paramètre CONTR est retirée en premier, puis la variable d'avant et ainsi de suite.
  - ATTER = 2
    La phase d'atterrissage est réalisée en considérant tous les échantillons possibles parmi les unités restantes, et en sélectionnant de façon préférentielle ceux qui donnent un écart faible à l'état d'équilibre.

- ATTER = 3
  La phase d'atterrissage est réalisée comme avec l'option précédente, mais en considérant seulement les échantillons dont la taille est égale à la somme des probabilités d'inclusion. Nous obtenons un échantillon de taille fixe. Si cette option est utilisée, la variable des probabilités d'inclusion doit être placée dans le paramètre CONTR.

La valeur par défaut est : ATTER=1. C'est l'option d'atterrissage la plus rapide. Pour que le temps d'exécution soit raisonnable, l'option ATTER=2 ne doit pas être utilisée avec plus de 14 variables d'équilibrage et l'option ATTER=3 ne doit pas être utilisée avec plus de 18 variables d'équilibrage.

❖ **COMPEQ** = option

Vaut 1 si le complémentaire de l'échantillon doit être équilibré sur les mêmes variables également, et 0 sinon.

La valeur par défaut est : COMPEQ=0

Nous utilisons ici un résultat de Tillé et Favre ([9]) ; la démonstration est en Annexe 1. Cette option permet de sélectionner plusieurs échantillons disjoints, équilibrés sur les mêmes variables, avec des probabilités d'inclusion fixés. Supposons que nous voulions sélectionner deux échantillons disjoints, équilibrés sur la variable x, avec des probabilités d'inclusion π_k. On sélectionne le premier échantillon équilibré S_1, de la façon habituelle, avec l'option COMPEQ=1. On sélectionne ensuite un échantillon S_2 dans le complémentaire de S_1, avec des probabilités d'inclusion $\frac{\pi_k}{1-\pi_k}$, équilibré sur la variable $(z_k) = \frac{x_k}{1-\pi_k}$. Cette méthode peut être utilisée pour un nombre quelconque de variable d'équilibre. On peut sélectionner jusqu'à $Min_{k\in U}\left[\frac{1}{\pi_k}\right]$ échantillons équilibrés à l'aide de cette méthode (où [ ] désigne la partie entière). C'est le principe de la construction des groupes de rotation du Nouveau Recensement.

Cette option multiplie par 2 le nombre de variables d'équilibrage, et donc par 4 le temps d'exécution.

Si toutes les probabilités d'inclusion sont égales, le complémentaire de l'échantillon est automatiquement équilibré sur les mêmes variables, et l'option est donc inutile. Pour plus de détails, voir l'Annexe 1.

2.2.3. Paramètres relatifs aux sorties

❖ **SORT** = nom d'une table SAS

Nom de la table SAS contenant les données en sortie. Cette table appartient à la librairie mentionnée dans le paramètre BASE. Elle contient toutes les unités de la population, et une variable ECH égale à 1 si l'unité est sélectionnée dans l'échantillon, et 0 sinon.

2.3. Exemples

Nous avons utilisé une population de 26,471 unités correspondant à la ville de Lyon, donnée par le Recensement de 1999. Les échantillons sont sélectionnés à l'aide d'un ordinateur personnel (Penthium 4, 1.8 Gh).

2.3.1. Exemple 1

Nous sélectionnons d'abord un échantillon avec des probabilités égales $\left(\frac{1}{5}\right)$, équilibré sur les variables socio-démographiques mentionnées dans la Table 1 (18 variables) et une constante pour obtenir la taille fixe. Nous utilisons la première option d'atterrissage.

On sélectionne ainsi un échantillon de 5,295 unités, en quelques secondes. Les résultats sont présentés dans la Table 3.

Table 3 : Ecart relatif entre le vrai total et l'estimateur de Horvitz-Thompson du total pour les variables d'équilibre

Variable	Estimateur de Horvitz-Thompson du total	Vrai total	Ecart relatif (%)
NLOG	251 380	251 279	+0,04%
NLOGCO	243 480	243 381	+0,04%
H0019	46 390	46 395	-0,01%
H2039	75 145	75 116	+0,04%
H4059	46 080	46 078	-0,00%
H6074	20 735	20 726	+0,04%
H7599	10 440	10 435	-0,05%
F0019	46 145	46 156	+0,02%
F2039	83 980	83 957	+0,03%
F4059	51 900	51 881	+0,04%
F6074	28 645	28 637	+0,03%
F7599	21 440	21 421	+0,09%
ACTIFS	206 780	206 732	+0,02%
INACTIFS	224 120	224 070	+0,02%
NATFN	376 425	376 326	+0,03%
NATFA	21 815	21 833	-0,08%
NATHE	22 990	22 978	+0,05%
NATUE	9 670	9 665	+0,05%

2.3.2. Exemple 2

Nous voulons sélectionner un échantillon de 1,500 adresses, équilibré sur les variables socio-démographiques mentionnées dans la Table 1 (18 variables), avec des probabilités d'inclusion proportionnelles à la taille de l'adresse (cette taille est donnée par le nombre de ménages). Nous équilibrons aussi sur la variable des probabilités d'inclusion et utilisons la troisième option d'atterrissage pour obtenir un échantillon de taille fixe.

L'échantillon est sélectionné en moins d'une minute. La condition de taille fixe est parfaitement vérifiée. Les résultats sont présentés dans la Table 4.

Table 4 : Ecart relatif entre le vrai total et l'estimateur de Horvitz-Thompson du total pour les variables d'équilibre

Variable	Estimateur de Horvitz-Thompson du total	Vrai total	Ecart relatif (%)
NLOG	251 279	251 279	+0,00%
NLOGCO	243 071	243 381	+0,13%
H0019	46 596	46 395	+0,43%
H2039	75 091	75 116	-0,03%
H4059	46 195	46 078	+0,25%
H6074	20 733	20 726	+0,03%
H7599	10 495	10 435	+0,57%
F0019	46 196	46 156	+0,09%
F2039	83 966	83 957	+0,01%
F4059	51 983	51 881	+0,20%
F6074	28 644	28 637	+0,02%
F7599	21 512	21 421	+0,42%
ACTIFS	206 834	206 732	+0,05%
INACTIFS	224 576	224 070	+0,23%
NATFN	376 919	376 326	+0,16%
NATFA	21 906	21 833	+0,33%
NATHE	22 993	22 978	+0,07%
NATUE	9 591	9 665	-0,76%

2.3.3. Exemple 3

On suppose maintenant que l'on veut sélectionner plusieurs échantillons, équilibrés sur les mêmes variables. On utilise toujours des probabilités proportionnelles à la taille de l'adresse ; nous voulons sélectionner 3 échantillons de 500 adresses. Dans la population, toutes les inverses de probabilités d'inclusion sont supérieures à 3.98 ; l'échantillonnage coordonné est donc possible. Nous équilibrons également sur les probabilités d'inclusion.

Nous utilisons la première option d'atterrissage et l'option COMPEQ=1. En effet, comme le nombre de variables d'équilibrage est important (38, correspondant aux 19 variables d'équilibrage de base, et 19 autres variables générées par l'option COMPEQ=1), l'échantillonnage ne pourrait pas être réalisé dans un temps raisonnable avec les options ATTER=2 ou ATTER=3.

Les résultats sont présentés dans la Table 5. La condition de taille fixe est parfaitement réalisée pour chacun des échantillons.

Table 5 : Ecart relatif entre le vrai total et l'estimateur de Horvitz-Thompson du total pour les variables d'équilibre

Variable	Vrai total	Estimateur de Horvitz-Thompson du total donné par le 1er échantillon	Ecart relatif (%)	Estimateur de Horvitz-Thompson du total donné par le 2nd échantillon	Ecart relatif (%)	Estimateur de Horvitz-Thompson du total donné par le 3ème échantillon	Ecart relatif (%)
NLOG	251 279	251 279	0,00%	251 279	0,00%	251 279	0,00%
NLOGCO	243 381	243 238	-0,06%	243 238	-0,06%	243 741	-0,13%
H0019	46 395	46 541	+0,31%	46 549	+0,33%	46 408	+0,43%
H2039	75 116	74 940	-0,23%	75 042	-0,10%	75 347	-0,03%
H4059	46 078	46 686	+1,32%	46 229	+0,33%	46 196	+0,25%
H6074	20 726	20 715	-0,05%	20 687	-0,19%	20 754	+0,03%
H7599	10 435	10 093	-3,27%	10 548	+1,08%	10 099	+0,57%
F0019	46 156	46 639	+1,05%	46 433	+0,60%	46 456	+0,09%
F2039	83 957	84 069	+0,13%	84 121	+0,20%	84 187	+0,01%
F4059	51 881	51 753	-0,25%	52 173	+0,56%	52 282	+0,20%
F6074	28 637	28 914	+0,97%	28 479	-0,55%	28 540	+0,02%
F7599	21 421	21 270	-0,71%	21 482	+0,28%	21 044	+0,42%
ACTIFS	206 732	207 197	+0,22%	206 907	+0,08%	207 851	+0,05%
INACTIFS	224 070	224 422	+0,16%	224 835	+0,34%	223 462	+0,23%
NATFN	376 326	376 200	-0,03%	378 181	+0,49%	377 177	+0,16%
NATFA	21 833	22 435	+2,76%	21 260	-2,63%	21 348	+0,33%
NATHE	22 978	23 431	+1,97%	22 820	-0,69%	23 391	+0,07%
NATUE	9 665	9 553	-1,16%	9 482	-1,89%	9 397	-0,76%

3. Equilibrage global et équilibrage stratifié

3.1. Notations

Nous conservons les mêmes notations que dans la première partie. Nous supposons ici que U est divisée en H strates disjointes notées $U_1,...,U_H$. Rappelons que le plan de sondage est dit équilibré sur la variable x si

$$\sum_{k\in U}\frac{S_k x_k}{\pi_k}=\sum_{k\in U}x_k \quad (1)$$

On dira que le plan de sondage est équilibré par strate sur la variable x si

$$\forall h=1...H \quad \sum_{k\in U_h}\frac{S_k x_k}{\pi_k}=\sum_{k\in U_h}x_k \quad (2)$$

Notons que si un plan de sondage est équilibré par strate, il est globalement équilibré sur l'ensemble de la population.

Cette technique a été utilisée par le Nouveau Recensement pour la construction des groupes de rotation de petites communes ; dans chaque région, ces groupes de rotation sont constitués en sélectionnant des échantillons équilibrés globalement sur des variables socio-démographiques, et équilibrés par département sur le nombre de ménages (de façon à assurer que chacun des 5 groupes de rotation contienne un nombre de communes raisonnable de chaque département).

L'échantillonnage équilibré stratifié peut être réalisé en sélectionnant directement un échantillon dans la population globale. En effet, (2) est équivalent à

$$\forall h = 1...H \ \sum_{k \in U} \frac{S_k\left(x_k 1_{k \in U_h}\right)}{\pi_k} = \sum_{k \in U_h} x_k 1_{k \in U_h} \quad (3)$$

Nous avons donc seulement besoin de sélectionner un échantillon dans U, équilibré sur les variables égales au produit des variables d'équilibrage $x_1,...,x_p$ et des variables indicatrices :

$$1_{k \in U_h} = \begin{cases} 1 \text{ si } k \in U_h \\ 0 \text{ sinon} \end{cases}$$

ce qui revient à équilibrer sur $H \times p$ variables. Cette méthode a plusieurs inconvénients :

- Si $H \times p$ est trop grand, on ne peut réaliser la phase d'atterrissage en cherchant l'échantillon qui donne un faible écart à l'état d'équilibre car le nombre d'échantillons à envisager est trop important. La seule option d'atterrissage possible est la première, i.e. de supprimer progressivement les contraintes d'équilibrage
- Toutes les strates n'ont pas la même qualité d'équilibrage. Avec la première option d'atterrissage, l'équilibrage est moins bon pour la strate correspondant aux variables supprimées en premier
- La taille fixe ne peut être atteinte exactement dans chaque strate

Le programme développé ici s'inspire d'une remarque sur le traitement des grosses bases de sondage (cf. [7]). L'idée est la suivante :

- On essaye d'abord d'équilibrer par strate : on réalise une phase de vol indépendamment dans chaque strate, en équilibrant sur les variables auxiliaires
- Quand il n'est plus possible d'équilibrer par strate, on recherche un équilibrage global : on réunit les unités qui n'ont encore été ni échantillonnées ni rejetées pendant les différentes phases de vol, puis on réalise une dernière phase de vol sur ces unités avant d'effectuer une phase d'atterrissage

La justification se trouve en Annexe 2.

3.2. Les données en entrée

Il doit y avoir autant de tables SAS en entrée qu'il y a de strates dans la population : chacune de ces tables contient les données relatives aux unités d'une strate particulière, et au moins :

- La variable des probabilités d'inclusion
- Les variables d'équilibrage

Cette table ne doit pas contenir de valeurs manquantes pour les variables citées ci-dessus. La variables des probabilités d'inclusion, aussi bien que les variables d'équilibrage, doivent être de type numérique.

3.3. Syntaxe de la macro

3.3.1. Paramètres relatifs à la base de données

Tous ces paramètres sont obligatoires.

- **BASE** = nom de librairie SAS
  Nom de la librairie SAS contenant les tables SAS des données en entrée

- **DATA** = table(s) SAS
  Nom(s) des tables SAS contenant les données en entrée. Les noms doivent être séparés avec des blancs. Chaque table contient les unités d'une strate et une seule.
  Par exemple, supposons que la population soit stratifiée en 4 strates U_1, U_2, U_3, U_4. 4 tables sont créées, par exemple STRAT1 pour les unités de la strate U_1, STRAT2 pour les unités de la strate U_2, etc … La syntaxe est alors : DATA= STRAT1 STRAT2 STRAT3 STRAT4.

- **PI** = variable
  Nom de la variable donnant les probabilités d'inclusion

- **ID** = variable
  Nom de la variable qui identifie les unités de la population

- **CONTR** = variable(s)
  Noms des variables sur lesquelles l'échantillon est équilibré. Les noms doivent être séparés avec des blancs.

3.3.2. Paramètres relatifs aux sorties

- **SORT** = table SAS
  Nom de la table SAS contenant les données en sortie. Cette table appartient à la librairie mentionnée dans le paramètre BASE. Elle contient la variable identifiante mentionnée dans ID, et une variable ECH égale à 1 si l'unité est sélectionnée dans l'échantillon, et 0 sinon.

3.4. Exemple

Nous réutilisons la population des adresses de la commune de Lyon ; la commune est stratifiée en 36 zones appelées Iris (Ilots regroupés selon des indicateurs statistiques). Un 37ème Iris contenant très peu d'adresses a été regroupé avec un des 36 Iris retenu.

Nous sélectionnons un échantillon avec des probabilités égales $\left(\frac{1}{5}\right)$, équilibré sur les variables socio-démographiques mentionnées dans la Table 1 (18 variables). Cet échantillon est sélectionné à l'aide de la macro %ECHANT_STRAT ; on attend donc un échantillon :

- équilibré sur l'ensemble de la commune
- approximativement équilibré dans chaque Iris
- de taille fixe dans chaque Iris

On sélectionne un échantillon de 5,295 unités, en quelques secondes. La table 6 compare les tailles d'échantillon attendues dans chaque strate et la taille d'échantillon obtenue :

Table 6 : Comparaison entre les tailles d'échantillon attendues et les tailles d'échantillon obtenues par strate

Strate	1	2	3	4	5	6	7	8	9	10	11	12
Taille d'échantillon attendue	277,8	222,8	231,4	101,4	34,6	260,4	259,2	160	128,8	20,6	268,8	285
Taille d'échantillon obtenue	278	223	231	102	35	260	259	160	129	21	268	285

Strate	13	14	15	16	17	18	19	20	21	22	23	24
Taille d'échantillon attendue	179,8	50,8	213,6	220,8	199	81,4	24,4	245	213,6	142,8	122,4	113
Taille d'échantillon obtenue	180	51	214	221	199	82	25	245	214	143	122	113

Strate	25	26	27	28	29	30	31	32	33	34	35	36
Taille d'échantillon attendue	134,4	103,6	46,6	153	157,2	114,2	71,4	102,6	55,4	155,2	124,6	18,6
Taille d'échantillon obtenue	134	104	46	153	157	114	71	103	55	156	125	17

Aux arrondis près, la taille fixe est donc parfaitement réalisée dans les strates (à l'exception de la 36, à une unité près). Les estimations sur l'ensemble de la commune sont présentées dans la Table 7.

Table 7 : Ecart relatif entre le vrai total et l'estimateur de Horvitz-Thompson du total pour les variables d'équilibre pour l'ensemble de la commune

Variable	Estimateur de Horvitz-Thompson du total	Vrai total	Ecart relatif (%)
NLOG	251 279	251 279	-0,17%
NLOGCO	243 381	243 381	-0,18%
H0019	46 395	46 395	-0,11%
H2039	75 116	75 116	-0,25%
H4059	46 078	46 078	-0,19%
H6074	20 726	20 726	-0,12%
H7599	10 435	10 435	-0,14%
F0019	46 156	46 156	-0,30%
F2039	83 957	83 957	-0,22%
F4059	51 881	51 881	-0,14%
F6074	28 637	28 637	0,01%
F7599	21 421	21 421	0,00%
ACTIFS	206 732	206 732	-0,17%
INACTIFS	224 070	224 070	-0,18%
NATFN	376 326	376 326	-0,18%
NATFA	21 833	21 833	-0,24%
NATHE	22 978	22 978	-0,10%
NATUE	9 665	9 665	0,10%

L'équilibrage au niveau de la commune entière est donc parfaitement respecté. En ce qui concerne les strates, nous présentons les résultats obtenus de façon synthétique dans la Table 8.

Seules 5 strates sont mal équilibrées (les 5, 10, 19, 33 et 34) ; en dehors de la dernière, elles correspondent toutes à des petites strates dans lesquelles on a tiré une taille faible d'échantillon.

On aurait pu effectuer un échantillonnage similaire avec la macro classique d'échantillonnage équilibré, en s'attaquant directement à la base de sondage constituée par l'ensemble des adresses de ces strates. Il aurait fallu pour cela introduire dans les variables d'équilibrage :

- La probabilité d'inclusion (pour avoir globalement une taille fixe d'échantillon) et les 18 variables mentionnées ci-dessus pour avoir un équilibrage global sur ces variables : en tenant compte des colinéarités, 17 variables d'équilibrage
- Une variable indicatrice d'appartenance aux strates pour requérir une taille fixe dans chaque strate : 35 variables d'équilibrage
- Des variables égales au produit des variables socio-démographiques par les indicatrices d'appartenance aux strates pour requérir un équilibrage par strate sur ces variables socio-démographiques : en tenant compte des colinéarités, 16*35 = 560 variables d'équilibrage

Soit au total 612 variables d'équilibrage, ce qui aurait été beaucoup plus coûteux en temps et aurait donné des qualités d'équilibrage très différentes suivant les strates.

Table 8 : Quelques indicateurs de la qualité de l'équilibrage au niveau des strates

Strate	1	2	3	4	5	6	7	8	9	10	11	12
Ecart relatif maximum (en module) obtenu pour les variables d'équilibrage	4%	19%	6%	4%	44%	2%	12%	13%	12%	48%	7%	3%
Ecart relatif moyen (en module) obtenu pour les variables d'équilibrage	2%	3%	2%	2%	11%	1%	4%	4%	2%	20%	2%	1%

Strate	13	14	15	16	17	18	19	20	21	22	23	24
Ecart relatif maximum (en module) obtenu pour les variables d'équilibrage	4%	19%	7%	6%	9%	9%	24%	3%	17%	10%	22%	23%
Ecart relatif moyen (en module) obtenu pour les variables d'équilibrage	2%	6%	2%	2%	3%	4%	13%	1%	4%	2%	6%	4%

Strate	25	26	27	28	29	30	31	32	33	34	35	36
Ecart relatif maximum (en module) obtenu pour les variables d'équilibrage	29%	16%	13%	13%	9%	27%	14%	12%	27%	33%	29%	16%
Ecart relatif moyen (en module) obtenu pour les variables d'équilibrage	7%	3%	4%	6%	2%	8%	8%	6%	11%	16%	7%	3%

Annexe 1 : Equilibrer un échantillon et son complémentaire

Soit U une population finie. Un échantillon s est dit équilibré sur la variable x si

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

Soit $\bar{s}$ un autre échantillon, défini comme étant le complémentaire de s dans U. Les probabilités d'inclusion dans cet échantillon sont alors $\bar{\pi}_k = P(k\in\bar{s}) = 1-\pi_k$, aussi l'échantillon $\bar{s}$ est dit équilibré sur la variable x si

$$\sum_{k\in\bar{s}}\frac{x_k}{1-\pi_k}=\sum_{k\in U}x_k$$

L'équilibrage d'un échantillon s et de son complémentaire $\bar{s}$ sur la variable x peut être réalisé en sélectionnant un échantillon s équilibré sur les variables (x_k) et $\left(\frac{x_k}{1-\pi_k}\right)$. En effet, on a alors :

$$\sum_{k\in s}\frac{x_k}{\pi_k}=\sum_{k\in U}x_k$$

par définition, et :

$$\begin{aligned}\sum_{k\in\bar{s}}\frac{x_k}{1-\pi_k} &= \sum_{k\in U}\frac{x_k}{1-\pi_k}-\sum_{k\in s}\frac{x_k}{1-\pi_k}\\ &= \sum_{k\in s}\frac{x_k}{\pi_k(1-\pi_k)}-\sum_{k\in s}\frac{x_k}{1-\pi_k}\\ &= \sum_{k\in s}\frac{x_k}{1-\pi_k}\left(\frac{1}{\pi_k}-1\right)=\sum_{k\in s}\frac{x_k}{\pi_k}\\ &= \sum_{k\in U}x_k\end{aligned}$$

$\bar{s}$ est donc équilibré lui aussi.

Annexe 2 : Equilibrage stratifié

Soit U une population finie, découpée en H strates disjointes $U_1,...,U_H$. Soit π_k la probabilité d'inclusion de l'unité k, et $\mathbf{x}_k$ le vecteur des variables d'équilibrage.

Nous suivons le procédé décrit en section 3.1, et commençons par réaliser une phase de vol indépendamment dans chaque strate (Phase 1). Avec les mêmes notations que pour l'Algorithme 1, on obtient à la fin de la phase1 :

$$\forall h = 1...H \quad \sum_{k\in U_h} \mathbf{x}_k = \sum_{k\in U_h} \frac{\mathbf{x}_k}{\pi_k}\pi_k$$

$$= \sum_{k\in U_h} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in S_h^*} \frac{\mathbf{x}_k}{\pi_k} + \sum_{k\in U_h^*} \frac{\mathbf{x}_k}{\pi_k}\pi_k^*$$

où S_h^* représente les unités échantillonnées dans la strate U_h et U_h^* les unités restantes (ni rejetées ni sélectionnées, i.e. avec $0 < \pi_k^* < 1$).

Pour la phase 2, on réunit les unités restantes et on sélectionne parmi ces unités un échantillon avec des probabilités d'inclusion π_k^*, équilibré sur les variables $\frac{\mathbf{x}_k}{\pi_k}\pi_k^*$. Soit $U^* = \bigcup_{h=1}^{H} U_h^*$, $S^{**} = \bigcup_{h=1}^{H} S_h^{**}$ l'échantillon sélectionné dans U^* où S_h^{**} désigne l'ensemble des unités de U_h^* sélectionnées lors de la phase 2. L'équilibrage implique que :

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

On note S l'échantillon final, réunion des unités sélectionnées en Phase 1 et de celles sélectionnées en phase 2, i.e. $S = S^* \cup S^{**}$. On note également S_h l'échantillon final des unités sélectionnées dans U_h : $S_h = S_h^* \cup S_h^{**}$.

$$\sum_{k\in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in S^*} \frac{\mathbf{x}_k}{\pi_k} + \sum_{h=1}^{H} \sum_{k\in S_h^*} \frac{\mathbf{x}_k}{\pi_k}$$

$$= \sum_{k\in U^*} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* + \sum_{h=1}^{H} \sum_{k\in S_h^*} \frac{\mathbf{x}_k}{\pi_k}$$

$$= \sum_{h=1}^{H} \left[\sum_{k\in S_h^*} \frac{\mathbf{x}_k}{\pi_k} + \sum_{k\in U_h^*} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* \right] = \sum_{h=1}^{H} \sum_{k\in U_h} \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k$$

L'échantillon S est donc globalement équilibré. D'autre part, on a pour chaque strate h :

$$\sum_{k\in S_h} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in S_h^{**}} \frac{\mathbf{x}_k}{\pi_k} + \sum_{k\in S_h^*} \frac{\mathbf{x}_k}{\pi_k}$$

$$= \sum_{k\in S_h^{**}} \frac{\mathbf{x}_k}{\pi_k} + \sum_{k\in U_h} \mathbf{x}_k - \sum_{k\in U_h^*} \frac{\mathbf{x}_k}{\pi_k}\pi_k^*$$

$$= \sum_{k\in U_h} \mathbf{x}_k + \sum_{k\in U_h^*} \frac{\mathbf{x}_k}{\pi_k}\left[1_{k\in S_h^{**}} - \pi_k^* \right] \approx \sum_{k\in U_h} \mathbf{x}_k$$

Nous avons donc également un équilibrage approché pour chaque strate.

Bibliographie

[1] 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.

[2] Chauvet, G., Tillé, Y. (2005) : A fast algorithm of balance sampling. A paraître dans *Computational Statistics*

[3] Deville, J.-C., Tillé, Y. (1998) : Unequal probability sampling without replacement though a splitting method. *Biometrika*, 85:89-101

[4] Deville, J.-C., Tillé, Y. (2004) : Efficient Balanced Sampling : the Cube method. *Biometrika*, 91:893-912

[5] Deville, J.-C., Tillé, Y. (2005) : Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128 :411-425

[6] 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*, volume 100, pages 37-76, Paris

[7] Rousseau, S., Tardieu, F. (2004) : La macro SAS CUBE d'échantillonnage équilibré – *Documentation de l'utilisateur*

[8] Tillé, Y. (2001). Théorie des Sondages : échantillonnage et estimation en populations finies. Dunod, Paris

[9] Tillé, Y., Favre, A.-C. (2004) : Coordination, combination and extension of balanced samples. *Biometrika*, 91:913-928