INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2008-08

Echantillonage Equilibré Stratifié

G. CHAUVET[1]

[1] Laboratoire de Statistique d'Enquête, CREST-ENSAI, rue Blaise Pascal, Campus de Ker Lann, 35170 BRUZ, France, (chauvet@ensai.fr)

Echantillonnage équilibré stratifié

Guillaume Chauvet[1]
24 mars 2008

Résumé

Lors de la sélection d'un échantillon, une pratique courante consiste à définir un plan de sondage stratifié sur des sous-populations. La variance est alors réduite par rapport à un tirage direct si les strates sont bien homogènes au regard de la variable d'intérêt. Si des variables auxiliaires sont disponibles pour chaque individu, l'échantillonnage peut être amélioré par tirage équilibré mais il n'est pas possible d'assurer un équilibrage correct au niveau de chaque strate si l'allocation d'échantillon est faible. Nous proposons ici une méthode de tirage permettant de sélectionner un échantillon équilibré sur l'ensemble de la population, en respectant une allocation fixée au sein de chaque strate. Des applications au cas d'un tirage stratifié de taille 2 dans chaque strate et au cas d'un échantillonnage rotatif sont également proposées.

Mots-Clés : Echantillonnage équilibré ; Echantillonnage rotatif ; Méthode du Cube ; Stratification ; Tirage à probabilités inégales.

Abstract

Populations are currently stratified into subpopulations before drawing samples, with a gain in accuracy if the strata are homogeneous for the interest variable. If auxiliary variables are available for each unit, the sampling may be enhanced by balanced sampling if the sampling sizes per stratum are large. We propose a sampling technique for the selection of balanced samples, with respect to given allocation per stratum. Applications to the important case of two units selected per stratum and to rotating samples are proposed.

Keywords : Balanced Sampling ; Cube Method ; Rotating Samples ; Sampling with unequal probabilities ; Stratification.

[1]Laboratoire de Statistique d'Enquête, CREST/ENSAI, rue Blaise Pascal, Campus de Ker Lann, 35 170 Bruz, France, chauvet@ensai.fr

1. INTRODUCTION

Dans le cas d'un tirage stratifié, la population est partitionnée en sous-populations appelées strates dans lesquelles des échantillons sont sélectionnés de façon indépendante. La variance est réduite par rapport à un tirage de même taille réalisé directement dans l'ensemble de la population si les strates ainsi constituées sont bien homogènes au regard de la variable d'intérêt. A l'intérieur de chaque strate, chaque échantillon est généralement sélectionné selon un plan de taille fixe. Si une information auxiliaire est disponible, cet échantillonnage peut être amélioré en utilisant la méthode du Cube (Deville et Tillé, 2004, Chauvet et Tillé, 2006) d'échantillonnage équilibré. L'échantillonnage équilibré assure que l'estimateur de Horvitz-Thompson du total de variables de contrôle est égal au total exact de ces variables. La méthode du Cube permet de sélectionner des échantillons exactement équilibrés, ou approximativement équilibrés si l'équilibrage exact est impossible. Au sein de chaque strate, l'estimateur de Horvitz-Thompson sera plus précis si les variables d'équilibrage sont bien corrélées à la variable d'intérêt.

L'équilibrage sera bien respecté au sein de chaque strate si le nombre de variables de contrôle reste faible devant la taille d'échantillon. Mais dans certains cas, l'allocation par strate est trop faible pour permettre l'équilibrage : si la population est stratifiée de façon très fine, une pratique courante

consiste à sélectionner un échantillon de taille 2 dans chaque strate. Il n'est alors pas possible d'imposer une condition autre que la contrainte de taille fixe d'échantillon dans chaque strate.

Nous proposons l'utilisation d'un algorithme d'échantillonnage adapté de la méthode du Cube, assurant un équilibrage sur l'ensemble de la population pour des variables de contrôle choisies et permettant le respect strict de l'allocation souhaitée au sein de chaque strate. Les échantillons ne sont alors plus sélectionnés indépendamment dans chaque strate. La précision est améliorée par rapport à un sondage stratifié avec tirage de taille fixe dans chaque strate si sur l'ensemble de la population les variables d'équilibrage sont bien corrélées à la variable d'intérêt. Une méthode analogue a été utilisée pour la sélection des unités primaires de l'Echantillon-Maître 1999 par l'Insee. L'Echantillon-Maître est un échantillon de logements, sélectionné dans le Recensement de 1999, et servant de base de sondage pour les enquêtes sur les ménages. On trouvera une description détaillée du plan de sondage de l'Echantillon-Maître dans Bourdalle et al. (2000). Les logements sont à l'origine regroupés au sein d'unités urbaines ou d'unités rurales. Dans la sous-population des unités les moins urbanisées, un échantillon d'environ 6 % de ces unités est sélectionné par tirage stratifié selon la région. Les tailles d'échantillon par région étant trop faibles pour permettre un équilibrage régional, les tirages dans les différentes régions ont été coordonnés afin d'assurer

un équilibrage sur des super-régions (correspondant à des regroupements de régions).

Le papier est organisé de la façon suivante. La section 2 introduit nos notations sur l'échantillonnage stratifié et l'échantillonnage équilibré. L'algorithme de tirage est présenté en section 3. La méthode est évaluée en section 4 à l'aide de quelques simulations.

2. NOTATIONS

2.1 Echantillonnage stratifié

On considère une population finie U constituée de N individus. L'objectif est d'estimer le total $t_y = \sum_{k \in U} y_k$ d'une variable d'intérêt y prenant les valeurs y_k sur les unités k de U. On suppose que l'on dispose d'une information permettant de partitionner la population en H strates $U_1, \ldots, U_H$ de tailles respectives $N_1, \ldots, N_H$. On suppose également que q variables d'équilibrage $x_1, \ldots, x_q$ sont disponibles, ou plus précisément que le vecteur $\mathbf{x}_k = (x_{1k}, \ldots, x_{qk})'$ donnant les valeurs prises par ces variables d'équilibrage est connu pour chaque individu de la population. On suppose sans perte de généralité que les q vecteurs $(x_{j1}, \ldots, x_{jk}, \ldots, x_{jN})', j = 1, \ldots, q$ sont linéairement indépendants.

Dans la suite de cette section, on suppose qu'un échantillon aléatoire S est sélectionné par tirage stratifié. Dans chaque strate U_h, un échantillon S_h est sélectionné indépendamment selon un plan de sondage $p_h(.)$. L'échantillon S est donné par la réunion des $S_h, h = 1, \ldots, H$. La probabilité d'inclusion de l'unité k est la probabilité π_k que l'unité k soit sélectionnée dans l'échantillon, et la probabilité d'inclusion jointe est la probabilité π_{kl} que deux unités distinctes k et l soient sélectionnées conjointement dans l'échantillon. On notera $\boldsymbol{\pi} = (\pi_k)_{k\in U}$ et $\boldsymbol{\pi}^{\mathbf{h}} = (\pi_k)_{k\in U_h}$. On suppose qu'au sein de chaque strate U_h, le plan $p_h(.)$ est de taille fixe. On a en particulier $\sum_{k\in U_h} \pi_k = n_h, h = 1, \ldots, H$ où n_h désigne l'allocation dans la strate U_h. L'estimateur de Horvitz-Thompson $\hat{t}_{y\pi} = \sum_{k\in S} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \hat{t}^h_{y\pi}$, où $\hat{t}^h_{y\pi} = \sum_{k\in S_h} \frac{y_k}{\pi_k}$, est un estimateur sans biais de $t_y = \sum_{k=1}^{H} t^h_y$, où $t^h_y = \sum_{k\in U_h} y_k$ désigne le total sur U_h de la variable d'intérêt y. Le vecteur $\hat{t}_{\mathbf{x}\pi} = \sum_{k\in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{h=1}^{H} \hat{t}^h_{\mathbf{x}\pi}$, où $\hat{t}^h_{\mathbf{x}\pi} = \sum_{k\in S_h} \frac{\mathbf{x}_k}{\pi_k}$, est un estimateur sans biais de $t_{\mathbf{x}} = \sum_{k\in U_h} t^h_{\mathbf{x}}$, où $t^h_{\mathbf{x}} = \sum_{k\in U_h} \mathbf{x}_k$ donne le vecteur des totaux des variables auxiliaires $\mathbf{x}$ sur U_h.

Comme l'échantillonnage est de taille fixe dans chaque strate U_h, la variance de l'estimateur de Horvitz-Thompson s'obtient à l'aide de la formule de va-

riance de Sen-Yates-Grundy :

$$Var\left(\hat{t}_{y\pi}\right) = \frac{1}{2}\sum_{h=1}^{H}\sum_{k\neq l\in U_h}\left(\pi_k\pi_l - \pi_{kl}\right)\left(\frac{y_k}{\pi_k} - \frac{y_k}{\pi_k}\right)^2. \quad (1)$$

Cette variance sera faible si $\frac{y_k}{\pi_k}$ est approximativement constant au sein de chaque strate.

2.2 Echantillonnage équilibré stratifié

Au sein de chaque strate, la précision de l'estimateur de Horvitz-Thompson peut être améliorée en utilisant les variables auxiliaires $\mathbf{x}$ et l'algorithme du Cube (Deville et Tillé, 2004) permettant de sélectionner des échantillons équilibrés. Le plan de sondage $p_h(.)$ est dit équilibré sur les variables $\mathbf{x}$ si les équations

$$\hat{t}^h_{\mathbf{x}\pi} = t^h_{\mathbf{x}} \quad (2)$$

sont exactement respectées. Dans la plupart des cas, un échantillon exactement équilibré ne peut être trouvé. Supposons par exemple que la population U_h contienne 100 individus sur lesquels est définie une variable x à deux modalités, 0 et 1, telle que 53 individus de la population présentent la modalité 0. Sélectionner un échantillon de taille 10, à probabilités égales, équilibré sur la variable x, supposerait de sélectionner un échantillon contenant 5.3 indi-

vidus présentant la modalité $x = 0$ et 4.7 individus présentant la modalité $x = 1$, ce qui est impossible. L'objectif est donc généralement de sélectionner un échantillon approximativement équilibré, c'est à dire tel que

$$\hat{t}^h_{\mathbf{x}\pi} \simeq t^h_{\mathbf{x}}. \tag{3}$$

La méthode du Cube (Deville et Tillé, 2004) permet de répondre à cet objectif. Elle se décompose en deux phases appelées phase de vol et phase d'atterrissage. A chaque étape de la phase de vol, on décide aléatoirement de sélectionner ou d'écarter définitivement l'une des unités de la population. A l'issue de la phase de vol, on obtient dans chaque strate U_h un vecteur $\boldsymbol{\pi}^{\mathbf{h}*} = (\pi^*_k)_{k\in U_h} \in [0,1]^N$ vérifiant les conditions suivantes :

$$E(\boldsymbol{\pi}^{\mathbf{h}*}) = \boldsymbol{\pi}^{\mathbf{h}}, \tag{4}$$

$$\sum_{k\in U_h} \frac{\mathbf{x}_k}{\pi_k}\pi^*_k = \sum_{k\in U_h} \mathbf{x}_k, \tag{5}$$

$$Card\{k \in U_h \ ; \ 0 < \pi^*_k < 1\} \leq q. \tag{6}$$

Le vecteur $\boldsymbol{\pi}^{\mathbf{h}*}$ donne le résultat de la phase de vol : π^*_k vaut 1 si l'unité k est sélectionnée, 0 si elle est rejetée et est comprise entre 0 et 1 strictement si la décision n'est pas encore prise pour l'unité k après la phase de

vol. Les équations (4) et (5) assurent que les probabilités d'inclusion et les contraintes d'équilibrage sont parfaitement respectées à l'issue de la phase de vol. L'équation (6) assure qu'il reste à trancher pour au plus q individus, où q désigne le nombre de variables d'équilibrage. La phase de vol s'arrête quand les contraintes d'équilibrage ne peuvent plus être exactement respectées. Durant la phase d'atterrissage, les contraintes d'équilibrage sont légèrement relâchées afin de terminer l'échantillonnage, mais les probabilités d'inclusion sont toujours exactement respectées.

Si le tirage équilibré est à forte entropie, Deville et Tillé (2005) donnent une formule approchée de variance. Dans le cas de l'échantillonnage équilibré stratifié, on obtient

$$Var\left(\hat{t}_{y\pi}\right) \simeq \sum_{h=1}^{H} \sum_{k \in U_h} \frac{b_k}{\pi_k^2} \left(y_k - \beta_{\mathbf{h}}{}' \mathbf{x}_k\right)^2 \tag{7}$$

où $\beta_{\mathbf{h}} = \left(\sum_{l \in U_h} b_l \frac{\mathbf{x}_l}{\pi_l} \frac{\mathbf{x}'_l}{\pi_l}\right)^{-1} \sum_{l \in U_h} b_l \frac{\mathbf{x}_l}{\pi_l} \frac{y_l}{\pi_l}$. Deville et Tillé (2005) proposent plusieurs approximations pour les coefficients b_k. La plus simple consiste à utiliser $b_k = \pi_k(1 - \pi_k)$. La variance sera faible si dans chaque strate la variable d'intérêt y est bien expliquée par les variables d'équilibrage. Notons que le tirage de taille fixe est un cas particulier de plan équilibré, corres-

pondant à un équilibrage sur la variable donnant la probabilité d'inclusion. L'échantillonnage équilibré apparaît donc comme une amélioration du tirage de taille fixe, qui permet d'incorporer une information supplémentaire pour expliquer la variable y afin de réduire la variance.

3 ÉCHANTILLONNAGE ÉQUILIBRÉ STRATIFIÉ AVEC MISE EN COMMUN DES PHASES D'ATTERRISSAGE

Si l'échantillon S est sélectionné selon la procédure d'échantillonnage équilibré stratifié présentée en section 2.2, l'équilibrage sera bien respecté au sein de chaque strate si la phase d'atterrissage porte sur un faible d'individus par rapport à la taille d'échantillon. Plus précisément, l'équation (6) montre que le nombre de variables d'équilibrage doit être faible par rapport à l'allocation d'échantillon dans chaque strate.

Dans certains cas, cette contrainte ne peut être respectée. Il est en effet fréquent d'utiliser un découpage très fin de la population afin d'améliorer sa pertinence, ce qui revient à réduire la taille d'échantillon sélectionnée dans chaque strate en se fixant généralement la limite d'un échantillon de taille 2 pour chacune afin de pouvoir obtenir un estimateur sans biais de variance. Un autre exemple est donné par le plan de sondage de l'Enquête Emploi

(Christine, 2000) : sur l'ensemble du territoire français, les logements sont regroupés selon des critères de contiguïté au sein d'unités secondaires appelées aires, elles-mêmes regroupées au sein d'unités primaires appelées secteurs, chaque secteur contenant 6 aires. Un échantillon de secteurs est sélectionné par tirage stratifié, et partitionné aléatoirement en 6 sous-échantillons selon leur date effective d'introduction dans l'enquête. Au sein de chaque sous-échantillon, chacune des 6 aires de chaque secteur se voit affecter aléatoirement un numéro, de 1 à 6, qui détermine son rang d'interrogation dans le sous-échantillon. Dans chaque sous-échantillon, les aires de chaque rang sont donc sélectionnées selon un tirage à probabilités égales, stratifié par secteur et de taille 1 dans chaque secteur.

Nous nous plaçons à nouveau dans le cas d'une population U découpée en H strates $U_1, \ldots, U_H$, et dans laquelle un vecteur $\mathbf{x}_k = (\pi_k, \mathbf{z}_k)'$ de variables auxiliaires est connu. On suppose que la variable π_k fait partie des contraintes d'équilibrage, afin d'assurer un échantillonnage de taille fixe. Dans le cas où l'allocation par strate est trop faible pour que l'équilibrage permette d'imposer d'autres contraintes que celle de taille fixe dans chaque strate, l'algorithme 1 fournit une méthode alternative d'échantillonnage. Une phase de vol est réalisée indépendamment sur chacune des H strates : on note

$\boldsymbol{\pi}^{\mathbf{h}*} = (\pi_k^*)_{k \in U_h}, h = 1, \dots H$ les vecteurs de probabilités obtenus à l'issue de ces phases de vol et $\boldsymbol{\pi}^* = (\pi_k^*)_{k \in V}$, où V désigne les unités qui n'ont pas encore été échantillonnées ou rejetées. Le vecteur des probabilités obtenues après une dernière phase de vol sur l'ensemble de ces unités restantes est noté $\boldsymbol{\pi}^{**} = (\pi_k^{**})_{k \in V}$. L'ensemble des unités de la strate U_h qui n'ont pas encore été échantillonnées ou rejetées à l'issue de cette nouvelle phase de vol est noté W_h.

Algorithme 1 : Echantillonnage équilibré stratifié avec mise en commun des phases d'atterrissage

---

Etape 1. Réaliser une phase de vol, sur les variables d'équilibrage $\mathbf{x}_k$ et avec les probabilités d'inclusion π_k, indépendamment dans chaque strate U_h.

Etape 2. Réaliser une phase de vol, sur les variables d'équilibrage $\mathbf{x}_k \, \frac{\pi_k^*}{\pi_k}$ et avec les probabilités d'inclusion π_k^*, sur l'ensemble V des unités restantes à l'issue de l'étape 1.

Etape 3. Sélectionner un échantillon de taille fixe dans chaque sous-population W_h, avec des probabilités d'inclusion π_k^{**}.

L'algorithme s'inspire d'une méthode utilisée à l'Insee pour la sélection des unités primaires de l'Echantillon-maître 1999 (Bourdalle et al., 2000), et également proposée par Rousseau et Tardieu (2004) pour la sélection d'échantillons équilibrés dans de grandes bases de sondage en utilisant la macro CUBE disponible sur le site Internet de l'Insee. Le temps d'exécution de cette macro est en effet approximativement proportionnel au carré de la taille de la population. Notons que Chauvet et Tillé (2006) proposent une méthode rapide d'échantillonnage équilibré dont le temps de calcul ne dépend plus que de la taille de la population, et permet de sélectionner directement des échantillons équilibrés sur de très grandes populations.

Utiliser le vecteur de probabilités d'inclusion $\boldsymbol{\pi}^*$ conditionnellement au résultat obtenu à l'issue de l'étape 1, assure que le vecteur $\boldsymbol{\pi}$ des probabilités d'inclusion est respecté en déconditionnant par rapport au résultat de l'étape 1. A l'issue de cette 1ère étape, l'équation (5) implique que

$$\forall h = 1 \dots H \qquad \sum_{k \in U_h / 0 < \pi_k^* < 1} \frac{\mathbf{x}_k}{\pi_k} \pi_k^* = \sum_{k \in U_h} \mathbf{x}_k - \sum_{k \in U_h / \pi_k^* = 1} \frac{\mathbf{x}_k}{\pi_k}$$

et en sommant ces expressions

$$\sum_{k\in V}\frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in U}\mathbf{x}_k - \sum_{k\in U/\pi_k^*=1}\frac{\mathbf{x}_k}{\pi_k}.$$

A l'issue de l'étape 2, on obtient à l'aide de l'équation (5)

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

et en comparant ces deux dernières expressions

$$\sum_{k\in V}\frac{\mathbf{x}_k}{\pi_k}\pi_k^{**} + \sum_{k\in U/\pi_k^*=1}\frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U}\mathbf{x}_k \tag{8}$$

ce qui assure que l'équilibrage sur les variables $\mathbf{x}_k$ est exactement respecté à l'issue de l'étape 2. L'étape 3 permet de terminer l'échantillonnage en respectant la contrainte de taille fixe à l'intérieur de chaque strate U_h, et peut être réalisé au moyen d'un programme linéaire afin de limiter le défaut d'équilibrage (cf Deville et Tillé, 2004).

La variance peut être approchée à l'aide de la formule de variance proposée par Deville et Tillé (2005), si chaque phase de vol de l'algorithme 1 est réalisée avec une entropie forte. L'entropie peut être fortement augmentée

en triant aléatoirement la population concernée préalablement au tirage. Les variables d'équilibrage sont ici d'une part les variables $\mathbf{z}_k$, et d'autre part les variables données par le produit des probabilités d'inclusion et des indicatrices d'appartenance aux strates et assurant une taille fixe d'échantillon dans chaque strate. On a

$$Var\left(\hat{t}_{y\pi}\right) \simeq \sum_{k\in U} \frac{b_k}{\pi_k^2} \left(y_k - \gamma' \mathbf{a}_k\right)^2 \tag{9}$$

avec $\mathbf{a}_k = (\pi_k \ 1_{k\in U_1}, \ldots, \pi_k \ 1_{k\in U_H}, \mathbf{z}_k')'$ et $\gamma = \left(\sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}'_l}{\pi_l}\right)^{-1} \sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}$. On peut l'estimer asymptotiquement sans biais par

$$v\left(\hat{t}_{y\pi}\right) = \sum_{k\in S} \frac{b_k}{\pi_k^3} \left(y_k - \hat{\gamma}' \mathbf{a}_k\right)^2 \tag{10}$$

avec $\hat{\gamma} = \left(\sum_{l\in S} \frac{b_l}{\pi_l} \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}'_l}{\pi_l}\right)^{-1} \sum_{l\in S} \frac{b_l}{\pi_l} \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}$.

4. RÉSULTATS NUMÉRIQUES

Nous réalisons une courte étude par simulations pour tester les performances de notre algorithme d'échantillonnage. Nous générons tout d'abord une population finie de taille 1000, partitionnée en 25 strates de même taille, et contenant 4 variables : 2 variables d'intérêt y_1 et y_2, et 2 variables auxiliaires x_1 et x_2. Tout d'abord, les variables x_1 et x_2 sont générées selon une distri-

bution Gamma de paramètres 4 et 25. La variable y_1 est générée au sein de la strate U_h selon le modèle

$$y_1 = \alpha_{1h} + \epsilon_h. \tag{11}$$

Les ϵ_h sont générés selon une distribution normale de moyenne 0 et de variance σ_h^2. Le modèle utilisé pour générer les valeurs de y_1 est donné par (11), avec $\alpha_{1h} = 20\ h$ et une variance σ_h^2 choisie pour donner un coefficient de détermination R^2 approximativement égal à 0.60 au sein de chaque strate. La variable y_2 est générée selon le modèle

$$y_2 = \alpha_2 + \beta_2\ x_1 + \gamma_2\ x_2 + \eta. \tag{12}$$

Les η sont générés selon une distribution normale de moyenne 0 et de variance ρ^2. Le modèle utilisé pour générer les valeurs de y_2 est donné par (12), avec $\alpha_2 = 500$, $\beta_2 = \gamma_2 = 5$, et une variance ρ^2 choisie pour donner un coefficient de détermination R^2 approximativement égal à 0.60.

On s'intéresse à l'estimation du total des variables y_1 et y_2. On sélectionne un échantillon de $n = 25$ (respectivement $n = 50$) unités à probabilités égales selon trois plans de sondage :

Plan 1 : tirage stratifié avec sondage aléatoire simple dans chaque strate,

Plan 2 : tirage équilibré sur les variables π, x_1 et x_2,

Plan 3 : tirage équilibré sur les variables π, x_1 et x_2, stratifié avec mise en commun des phases d'atterrissage.

Dans le cas du tirage stratifié, on a donc une allocation de taille 1 (respectivement 2) dans chaque strate. Dans les tirages équilibrés, chaque phase de vol est précédée d'un tri aléatoire de la population concernée. La variance associée au plan 1 est calculée directement. La variance associée aux plans 2 et 3 est approchée sur la base de 10 000 simulations. Le tableau 1 compare les résultats obtenus.

TAB. 1 – Variance associée à l'estimation du total de 2 variables pour les plans de sondage stratifié, équilibré, et stratifié équilibré avec mise en commun des phases d'atterrissage

	n=25		n=50	
Méthode	Var. total y_1 ($\times 10^8$)	Var. total y_2 ($\times 10^9$)	Var. total y_1 ($\times 10^8$)	Var. total y_2 ($\times 10^9$)
Plan 1	6.05	7.13	2.95	3.48
Plan 2	14.31	3.05	7.02	1.40
Plan 3	6.00	3.63	2.98	1.54

Dans chaque cas, le plan de sondage proposé est comparable avec la meilleure des deux stratégies. Si la variable d'intérêt est approximativement constante par strate, l'algorithme proposé donne les mêmes résultats que le plan de

sondage stratifié. Si les variables d'équilibrage sont bien explicatives, les résultats obtenus avec notre algorithme et avec un tirage équilibré direct sont équivalents. La légère perte de précision provient de l'étape d'atterrissage : dans le cas du tirage équilibré direct, on cherche à terminer l'échantillonnage en limitant le défaut d'équilibrage. Avec l'algorithme proposé, la solution retenue est sous-optimale car on ajoute la contrainte supplémentaire d'une taille fixe dans chaque strate.

Dans le cas du tirage stratifié équilibré avec mise en commun des phases d'atterrissage, le tableau 2 compare la variance donnée par 10 000 simulations avec la formule de variance approchée donnée en (9).

TAB. 2 – Comparaison entre la variance donnée par 10 000 simulations et la formule de variance approchée dans le cas de l'estimation de deux totaux pour un plan de sondage stratifié équilibré avec mise en commun des phases d'atterrissage

	n=25		n=50	
	Total y_1 ($\times 10^8$)	Total y_2 ($\times 10^9$)	Total y_1 ($\times 10^8$)	Total y_2 ($\times 10^9$)
Var. simulations	6.0	3.6	3.0	1.5
Var. approchée	5.9	2.7	2.9	1.3

La formule approchée proposée par Deville et Tillé (2005) est proche de la précision exacte si la variance associée à la phase d'atterrissage est faible

devant la variance associée à la phase de vol. Dans le cas de la variable y_2, les variables d'équilibrage sont fortement explicatives. La variance liée à la phase d'atterrissage est alors plus importante, relativement à la phase de vol, et la formule approchée sous-estime la vraie variance. La prise en compte de la variance associée à la phase de vol fera l'objet de travaux ultérieurs.

RÉFÉRENCES

Bourdalle, G, et Christine, M., et Wilms, L. (2000). *Echantillons maître et emploi*. Série Insee Méthodes, 21, pages 139-173, Paris, France.

Chauvet, G., and Tillé, Y. (2006). *A Fast Algorithm of Balanced Sampling*. Computational Statistics, 21, pages 53-61.

Christine, M. (2000). *La construction de l'échantillon de la future Enquête Emploi en continu à partir du Recensement de 1999*. Série Insee Méthodes, 21, pages 175-227, Paris, France.

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

Deville, J-C., and Tillé, Y. (2004). *Variance approximation under balanced sampling*. Journal of Statistical Planning and Inference, 128, pages 569-591.

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