No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Recueil du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Échantillonnage PPT systématique randomisé fondé sur la simulation en cas de substitution d'unités

Mary E. Thompson et Changbao Wu[1]

Résumé

L'enquête réalisée par la Chine dans le cadre du programme de lutte internationale contre le tabagisme (ITC pour *International Tobacco Control*) comprend un plan d'échantillonnage à plusieurs degrés avec probabilités inégales où les grappes du niveau supérieur sont sélectionnées par la méthode d'échantillonnage PPT systématique randomisé. Durant l'exécution de l'enquête, il faut résoudre le problème suivant : plusieurs grappes de niveau supérieur échantillonnées refusent de participer et doivent être remplacées par des unités de substitution sélectionnées parmi les unités non incluses dans l'échantillon initial, de nouveau par la méthode d'échantillonnage PPT systématique randomisé. Dans de telles conditions, les probabilités d'inclusion de premier ordre des unités finales sélectionnées sont très difficiles à calculer et la détermination des probabilités d'inclusion de deuxième ordre devient virtuellement impossible. Dans le présent article, nous élaborons une méthode fondée sur la simulation pour calculer les probabilités d'inclusion de premier et de deuxième ordre lorsque le calcul direct est prohibitif ou impossible. Nous démontrons l'efficacité de la méthode que nous proposons en nous appuyant sur des considérations théoriques et des exemples numériques. Nous incluons plusieurs fonctions et codes R/S-PLUS pour la procédure proposée. La méthode peut être étendue à des situations de refus/substitution plus complexes susceptibles de survenir en pratique.

MOTS CLÉS : probabilité d'inclusion; estimateur d'Horvitz-Thompson; méthode de Rao-Sampford; biais relatif; échantillonnage avec probabilités inégales sans remise.

1. Introduction

Dans l'analyse de données d'enquête complexe, la construction des poids de sondage est la première étape critique. Elle débute par le calcul des probabilités d'inclusion de premier ordre, qui est souvent simple si le plan d'échantillonnage original est bien exécuté, sans aucune altération ni modification. Par exemple, si les unités d'échantillonnage sont sélectionnées avec la probabilité d'inclusion (□) proportionnelle à la taille (PPT ou □pt), alors les probabilités d'inclusion s'obtiennent facilement par un simple rééchelonnage de la variable de taille. Parmi les méthodes d'échantillonnage PPT avec probabilités inégales sans remise utilisables lorsque la taille d'échantillon est fixée arbitrairement, la méthode d'échantillonnage PPT systématique randomisé est la plus simple à appliquer. Elle a été décrite pour la première fois dans Goodman et Kish (1950) à titre de méthode de sélection contrôlée, et a été perfectionnée par Hartley et Rao (1962) qui ont étudié le problème important, mais difficile, du calcul des probabilités d'inclusion de deuxième ordre. Soit xi, i = 1, 2, · · · , N les valeurs connues de la variable de taille, où N est le nombre total d'unités dans la population. Soit zi = xi/X où X = PN i=1 xi et supposons que zi < 1 pour tout i. La méthode d'échantillonnage PPT systématique randomisé procède comme il suit. Disposer les N unités de la population dans un ordre aléatoire et poser que A0 = 0 et Aj = Pj i=1(nzi) sont les totaux cumulatifs dans l'ordre en question, de sorte que 0 = A0 < A1 < · · · < AN = n. Soit u un nombre aléatoire uniforme dans l'intervalle [0, 1]. Les n unités qu'il faut inclure dans l'échantillon sont celles dont l'indice j satisfait Aj−1 □ u + k < Aj pour k = 0, 1, · · ·, n − 1. Soit s l'ensemble des n unités échantillonnées et □i = P(i 2 s), les probabilités d'inclusion de premier ordre. La méthode d'échantillonnage PPT systématique randomisé satisfait la condition

$$\pi_i = nz_i, \ \ i = 1, 2, \cdots, N. \tag{1.1}$$

Plusieurs autres méthodes d'échantillonnage sans remise qui satisfont (1.1) pour une taille d'échantillon fixe arbitraire n ont également été proposées dans la littérature, y compris la méthode d'échantillonnage avec probabilités inégales bien connue de Rao-Sampford (Rao, 1965; Sampford 1967) et celles de Chao (1982), Chen et coll. (1994), Tillé (1996), ainsi que Deville et Tillé (1998). Les importants travaux de recherche sur les méthodes d'échantillonnage PPT ont été stimulés en grande partie par l'utilisation de l'estimateur d'Horvitz-Thompson (HT) $\hat{T} = \sum_{i \in s} y_i / \pi_i$ pour le chiffre de population $T = \sum_{i=1}^{N} y_i$ d'une variable

[1] Mary E. Thompson, Département de statistique et de science actuarielle, Université de Waterloo, methomps@uwaterloo.ca; Changbao Wu, Département de statistique et de science actuarielle, Université de Waterloo, cbwu@uwaterloo.ca

d'intérêt y. L'estimateur HT est très efficace quand y est fortement corrélé à la variable de taille x et que la méthode d'échantillonnage satisfait (1.1). Il s'agit de l'unique estimateur sans biais par rapport au plan parmi la classe d'estimateurs linéaires $\sum_{i\in s} w_i y_i$ pour T si les poids w_i dépendent uniquement de i. Bien que le choix d'une méthode d'échantillonnage PPT puisse être souhaitable d'un point de vue théorique, l'exécution est souvent compliquée et parfois même impossible, en raison de contraintes et de limites pratiques. Certains compromis et modifications sont nécessaires. Toutefois, le plan modifié ne satisfait plus la condition (1.1). Le calcul direct des probabilités finales d'inclusion devient souvent difficile, voire impossible. Parmi les problèmes courants qui, dans la pratique des enquêtes, nécessitent l'altération du plan d'échantillonnage original, les unités qui refusent de participer et la substitution d'unités sont les plus fréquents. L'exemple qui suit illustre cette situation. L'enquête réalisée par la Chine dans le cadre du projet d'évaluation de la politique *International Tobacco Control* (ITC) (Enquête ITC–Chine) repose sur un plan d'échantillonnage à plusieurs degrés avec probabilités inégales pour la sélection de fumeurs et de non-fumeurs adultes dans sept villes. Chaque ville possède une structure administrative hiérarchique naturelle

$$\text{Ville} \longrightarrow \text{District de voirie} \longrightarrow \text{Îlot résidentiel} \longrightarrow \text{Ménage} \longrightarrow \text{Individu}$$

qui a été intégrée dans le plan d'échantillonnage. Aux niveaux supérieurs, la méthode d'échantillonnage PPT systématique randomisé est appliquée pour sélectionner dix districts de voirie dans chaque ville, avec probabilité proportionnelle à la taille de la population du district, puis deux îlots résidentiels sont tirés dans chaque district sélectionné, de nouveau par échantillonnage PPT systématique randomisé avec probabilité proportionnelle à la taille de la population de l'îlot. Les ménages et les individus dans les ménages sont ensuite sélectionnés en utilisant une méthode d'échantillonnage aléatoire simple modifié. Le plan original consistait à sélectionner 40 fumeurs adultes et 10 non-fumeurs adultes dans chacun des 20 îlots résidentiels, afin d'obtenir un échantillon final de 800 fumeurs et de 200 non-fumeurs dans chaque ville. Cependant, un problème s'est posé durant l'exécution de l'enquête : plusieurs grappes de niveau supérieur sélectionnées (d'abord les districts de voirie, puis les îlots résidentiels) ont refusé de participer à l'enquête, en raison de conflits avec d'autres activités ou de la non-disponibilité des ressources humaines. Ces grappes qui refusent doivent être remplacées par des unités de substitution sélectionnées parmi les unités non incluses dans l'échantillon initial; une possibilité consiste à recourir de nouveau à l'échantillonnage PPT systématique randomisé pour obtenir la taille d'échantillon global cible. Selon les plans d'échantillonnage à plusieurs degrés tels que celui utilisé pour l'enquête ITC–Chine, les probabilités d'inclusion de premier ordre des individus sélectionnés dans l'échantillon final peuvent être calculées en multipliant les probabilités d'inclusion des unités aux diverses étapes. Lorsque la méthode d'échantillonnage PPT systématique randomisé est modifiée en raison de la substitution d'unités à une certaine étape, la condition (1.1) n'est plus vérifiée pour l'échantillon final à cette étape. Le calcul des probabilités d'inclusion de premier ordre devient alors très difficile et celui des probabilités d'inclusion de deuxième ordre, virtuellement impossible. À l'annexe A, nous donnons une méthode de calcul direct (5.2) de π_i quand l'échantillon initial et l'échantillon de substitution sont sélectionnés tous deux par échantillonnage PPT systématique randomisé, en supposant que le refus de participer est aléatoire dans l'échantillon initial et qu'aucun refus n'a lieu dans l'échantillon de substitution. L'expression est valide conditionnellement au nombre de refus et à l'ordre de population utilisé (après randomisation) pour la sélection de l'échantillon initial. Il est évident que, même selon des conditions et des hypothèses aussi contraignantes, le traitement de l'expression proprement dite devient difficile, même pour une taille d'échantillon qui n'est pas tellement grande. Dans le présent article, nous démontrons, en nous appuyant sur des arguments théoriques et des exemples numériques, que les probabilités d'inclusion de premier et de deuxième ordre peuvent être estimées exactement par des simulations de Monte Carlo lorsque l'on dispose de renseignements complets sur le plan de sondage. Nos exemples numériques sont motivés par l'enquête ITC–Chine pour laquelle l'échantillonnage PPT systématique randomisé sert de méthode de référence, mais nos résultats théoriques et la méthodologie générale s'appliquent aussi à d'autres méthodes d'échantillonnage avec probabilités inégales sans remise. À la section 2, nous présentons les résultats relatifs à l'exactitude des méthodes fondées sur la simulation. À la section 3, nous donnons des exemples numériques et des comparaisons. À l'annexe C, nous présentons plusieurs fonctions et codes R/S-PLUS pour la procédure proposée, qui ont été élaborés au départ pour l'enquête ITC-Chine. À la section 4, nous formulons certaines remarques supplémentaires.

2. Propriétés des méthodes fondées sur la simulation

Si le calcul des probabilités d'inclusion exactes est impossible ou prohibitif, mais que l'on dispose de renseignements complets sur le plan de sondage, il est facile d'utiliser des méthodes de simulation de Monte Carlo pour estimer les probabilités d'inclusion. Désignons le plan d'échantillonnage probabiliste

entièrement spécifié par p . La méthode fondée sur la simulation est simple, à savoir tirer K échantillons indépendants, tous selon le même plan d'échantillonnage p ; soit M_i le nombre d'échantillons qui contiennent l'unité i . Alors, la probabilité d'inclusion de premier ordre $\pi_i = P(i \in s)$ peut être estimée par $\pi_i^* = M_i/K$. Pour un i particulier, les M_i suivent une loi binomiale et les π_i^* satisfont $E(\pi_i^*) = \pi_i$ et $Var(\pi_i^*) \le (4K)^{-1}$. Supposons, par exemple, que nous pouvons nous permettre de prendre K aussi grand que 25×10^6; alors, $P(|\pi_i^* - \pi_i| < 0,001) \ge 0,99$ pour toute probabilité donnée π_i . Une mesure plus pertinente de l'exactitude des méthodes fondées sur la simulation est la performance de l'estimateur d'Horvitz-Thompson lorsqu'on utilise les probabilités d'inclusion simulées. Soit $\hat{T} = \sum_{i \in s} y_i/\pi_i$ et $\tilde{T} = \sum_{i \in s} y_i/\pi_i^*$. Pour un échantillon donné, le biais relatif dû à l'utilisation de $\tilde{T}$ au lieu de $\hat{T}$ est défini par $(\hat{T} - \tilde{T})/\hat{T}$. Sans perte de généralité, nous supposons que $y_i \ge 0$ pour tout i . Nous montrons à l'annexe B que, pour tout $\varepsilon > 0$ et l'échantillon donné s ,

$$P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} < \varepsilon\right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i \in s}\frac{1}{\pi_i} - n\right). \tag{2.1}$$

Si ε est petit et que $n^{-1}\sum_{i \in s}(1/\pi_i) \approx 1/(N^{-1}\sum_{i=1}^{N}\pi_i) \approx N/n$, une borne inférieure pratique pour $P(|\hat{T} - \tilde{T}|/\hat{T} < \varepsilon)$ sera

$$\Delta = 1 - \frac{2(N-n)}{K\varepsilon^2}. \tag{2.2}$$

Si l'on veut que $\varepsilon = 0,01$ et $\Delta = 0,98$, alors pour $N - n = 100$ le nombre (théorique) d'échantillons indépendants requis pour la simulation est $K = 10^8$. Puisque la borne inférieure donnée par (2.1) est prudente, et valide pour toute variable réponse, on s'attendrait à ce qu'un K plus petit ayant une valeur autour de 10^7 , voire même 10^6 , donne de bons résultats pour la plupart des scénarios pratiques où $N \le 100$, ce que corroborent les exemples numériques présentés à la section 3. L'estimation des probabilités d'inclusion de deuxième ordre $\pi_{ij} = P(i, j \in s)$ n'impose aucune difficulté supplémentaire, excepté que le nombre total d'échantillons simulés, K , requis pour obtenir le même niveau d'exactitude relative que pour les probabilités de premier ordre est plus élevé. Soit M_{ij} le nombre d'échantillons simulés parmi les K échantillons indépendants qui contiennent à la fois i et j . Soit $\pi_{ij}^* = M_{ij}/K$ l'estimation pour π_{ij} . Supposons que le but soit d'estimer une grandeur de population quadratique

$$Q = \sum_{i=1}^{N}\sum_{j=1}^{N} q(y_i, y_j).$$

Les estimateurs de type Horvitz-Thompson de Q en utilisant π_{ij} ou π_{ij}^* sont donnés, respectivement, par

$$\hat{Q} = \sum_{i \in s}\sum_{j \in s}\frac{q(y_i, y_j)}{\pi_{ij}} \quad \text{et} \quad \tilde{Q} = \sum_{i \in s}\sum_{j \in s}\frac{q(y_i, y_j)}{\pi_{ij}^*}.$$

En suivant le même argument que celui menant à (2.1), nous pouvons montrer que

$$P\left(\frac{|\hat{Q} - \tilde{Q}|}{\hat{Q}} < \varepsilon\right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i \in s}\sum_{j \in s}\frac{1}{\pi_{ij}} - n^2\right). \tag{2.3}$$

À titre d'indication approximative, supposons que $\pi_{ij} = n(n-1)/(N(N-1))$ pour tous (i, j) , $i \ne j$, et que ε est petit; alors, une borne inférieure pour $P(|\hat{Q} - \tilde{Q}|/\hat{Q} < \varepsilon)$ est approximativement $1 - 2(N+n)(N-n)/(K\varepsilon^2)$. Si nous la comparons à Δ donné par (2.2), il est évident que nous avons besoin d'un K beaucoup plus grand pour obtenir la même borne inférieure, quoique, dans les deux cas, la borne inférieure est prudente et la valeur de K effectivement requise peut être plus faible. Par

ailleurs, les probabilités d'inclusion de deuxième ordre sont utilisées pour estimer les paramètres de deuxième ordre, tels que la variance de population ou la variance d'un estimateur linéaire. L'exactitude souhaitée de l'estimation est moins cruciale que dans le cas des paramètres de premier ordre, comme le total ou la moyenne de population et, par conséquent, un nombre compris entre 10^6 et 10^7 devrait être acceptable pour K dans de nombreuses situations pratiques. L'aspect le plus critique pour les méthodes fondées sur la simulation est manifestement la faisabilité des calculs. Entre autres, elle dépend en grande partie de la valeur de K choisie, de la complexité du plan d'échantillonnage et de la puissance de calcul disponible. Si $K = 10^6$ et que l'on souhaite obtenir les résultats fondés sur la simulation en dix heures, il faut tirer 28 échantillons simulés par seconde. L'échantillonnage PPT systématique randomisé est la méthode d'échantillonnage avec probabilités inégales sans remise la plus efficace en ce qui concerne l'exécution des calculs. Elle ne comporte qu'un ordonnancement aléatoire simple et la sélection d'un point de départ aléatoire. La plupart des méthodes concurrentes sont des méthodes d'échantillonnage réjectif ou des sélections séquentielles compliquées pour lesquelles la sélection d'échantillons simulés est beaucoup plus longue. Nous donnons à la section 4 une comparaison des temps de processeur requis pour calculer les π_i simulées pour l'échantillonnage PPT systématique randomisé et pour l'échantillonnage avec probabilités inégales de Rao-Sampford.

3. Exemples numériques

L'information sur le plan de sondage utilisée à la présente section est adaptée d'après l'enquête ITC-Chine. Le nombre de districts de voirie (grappes de niveau supérieur) dans chacune des sept villes faisant partie du champ d'observation de l'enquête varie de $N = 20$ à $N = 120$. Dans chaque ville, $n = 10$ districts sont sélectionnés par la méthode d'échantillonnage PPT systématique randomisé. En cas de refus, des districts de substitution sont sélectionnés parmi ceux non inclus dans l'échantillon initial, de nouveau par échantillonnage PPT systématique randomisé. Pour les besoins de l'illustration, nous utilisons l'information sur le plan de sondage provenant de la plus petite ville (c'est-à-dire $N = 20$). Des commentaires supplémentaires sur les cas pour lesquels N est grand sont formulés à la section 4.

3.1 Probabilités d'inclusion de premier ordre

Nous démontrons d'abord l'exactitude des valeurs simulées de π_i lorsque les valeurs exactes sont connues. Puis, nous examinons l'effet des substitutions d'unités sur la valeur finale de π_i et les propriétés de l'estimateur d'Horvitz-Thompson pour un total de population en utilisant les valeurs simulées de π_i. Les probabilités d'inclusion simulées en présence de substitution d'unités sont comparées à celles obtenues en supposant que le plan modifié est encore un échantillonnage PPT.

Exemple 1. π_i^* fondées sur la simulation en l'absence de refus. Dans ce cas, les valeurs exactes de π_i sont données par $\pi_i = nz_i$.

i) Valeurs exactes de π_i :

```
0,5840 0,5547 0,6702 0,5331 0,3085 0,2652 0,3930 0,4180 0,6952 0,3471
0,5993 0,5393 0,8240 0,6868 0,4469 0,2191 0,4237 0,4180 0,7567 0,3163
```

ii) π_i^* simulées, $K = 10^5$:

```
0,5828 0,5545 0,6656 0,5339 0,3071 0,2656 0,3929 0,4205 0,6969 0,3474
0,6009 0,5429 0,8227 0,6865 0,4446 0,2186 0,4215 0,4179 0,7569 0,3194
```

iii) π_i^* simulées, $K = 10^6$:

```
0,5836 0,5558 0,6701 0,5336 0,3081 0,2654 0,3931 0,4180 0,6950 0,3469
0,5994 0,5394 0,8242 0,6864 0,4469 0,2186 0,4237 0,4172 0,7569 0,3166
```

Les π_i^* simulées concordent avec les π_i jusqu'à la deuxième décimale pour $K = 10^5$ et jusqu'à la troisième pour $K = 10^6$ dans la plupart des cas.

Exemple 2. Afin d'évaluer les propriétés de l'estimateur d'Horvitz-Thompson (HT) pour un chiffre de population en utilisant les valeurs réelles de π_i et les π_i^* simulées de l'exemple 1, nous avons généré la variable réponse d'après le modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \cdots, N$, où x_i est la variable de taille et les ε_i sont indépendants et suivent la même loi normale de moyenne 0 et de variance σ^2. Nous avons considéré trois populations (trois valeurs de σ^2) pour lesquelles les coefficients de corrélation de population entre x et y sont, respectivement, $0,3$, $0,5$ et $0,8$. Pour chacune des trois populations, nous avons sélectionné $B = 2000$ échantillons répétés de taille $n = 10$ par échantillonnage PPT systématique randomisé, et pour chaque échantillon, nous avons calculé trois estimateurs HT en utilisant les valeurs réelles de π_i, les π_i^* simulées avec $K = 10^5$ et les π_i^* simulées avec $K = 10^6$, respectivement. Les résultats, qui ne sont pas présentés ici faute d'espace, montrent que le biais relatif de chacun des trois estimateurs HT est inférieur à 0,04 % et que leurs erreurs quadratiques moyennes sont presque identiques.

Exemple 3. Lorsque des refus de participation se produisent dans l'échantillon PPT initial et que des unités de substitution sont sélectionnées parmi les unités non incluses dans l'échantillon initial en utilisant la même méthode d'échantillonnage PPT, deux questions se posent : 1) comment calculer les probabilités d'inclusion π_i pour l'échantillon final et 2) dans quelle mesure la méthode de substitution a-t-elle altéré le plan d'échantillonnage PPT original? Nous pouvons calculer les π_i^* simulées et les comparer aux $\tilde{\pi}_i$ obtenues en supposant que l'échantillonnage demeure PPT après que l'on ait retiré de la base de sondage les unités ayant exprimé un refus. En simulant les π_i^*, nous supposons pour simplifier qu'aucun refus n'est possible parmi les unités non comprises dans l'échantillon initial et, par conséquent, qu'il n'existe aucun refus parmi les unités de substitution. Le nombre de répétitions K est fixé à 10^6 pour la simulation. Nous envisageons deux scénarios dans les conditions où la population contient trois unités qui refusent de participer et que toutes ces unités sont sélectionnées dans l'échantillon initial de taille $n = 10$.

i) Refus de trois grandes unités : π_i^* simulées (deux premières lignes) contre $\tilde{\pi}_i$ (deux dernières lignes) sous l'hypothèse d'un échantillonnage PPT

```
0,7231 0,6981 0,7947 0,6773 0,4354 0,3811 0,5339 0,5619 0,0000 0,4815
0,7363 0,6826 0,0000 0,8070 0,5919 0,3210 0,5678 0,5615 0,0000 0,4441

0,7560 0,7182 0,8677 0,6901 0,3994 0,3434 0,5088 0,5412 0,0000 0,4494
0,7759 0,6983 0,0000 0,8892 0,5786 0,2837 0,5486 0,5412 0,0000 0,4096
```

ii) Refus de trois petites unités : π_i^* simulées (deux premières lignes) contre $\tilde{\pi}_i$ (deux dernières lignes) sous l'hypothèse d'un échantillonnage PPT

```
0,6326 0,6049 0,7167 0,5829 0,0000 0,0000 0,4415 0,4668 0,7406 0,3937
0,6482 0,5901 0,8558 0,7330 0,4965 0,0000 0,4728 0,4664 0,7976 0,3590

0,6343 0,6025 0,7280 0,5790 0,0000 0,0000 0,4268 0,4540 0,7550 0,3770
0,6510 0,5858 0,8949 0,7459 0,4854 0,0000 0,4602 0,4540 0,8218 0,3436
```

Il est évident que la taille des unités qui refusent a un effet spectaculaire sur la distribution des probabilités d'inclusion finales. Si l'on omet de tenir compte de l'altération du plan d'échantillonnage résultant de la substitution des unités et que l'on traite ce plan comme s'il s'agissait encore d'un échantillonnage PPT, les probabilités d'inclusion pour les grandes unités sont exagérées et le rôle des petites unités est atténué. Cette tendance est plus prononcée lorsque de grandes unités figurent parmi celles qui refusent de participer, c'est-à-dire le cas (i) où $\pi_{14}^* = 0,8070$ comparativement à $\tilde{\pi}_{14} = 0,8897$ et $\pi_{16}^* = 0,3210$ comparativement à $\tilde{\pi}_{16} = 0,2837$.

3.2 Probabilités d'inclusion de deuxième ordre

L'échantillonnage PPT systématique randomisé a suscité de nombreux travaux de recherche, principalement en vue d'obtenir les probabilités d'inclusion de deuxième ordre π_{ij} et les estimateurs de variance. Hartley et Rao (1962) ont établi des formules exactes pour les π_{ij} quand $n=2$ et $N=3$ ou $N=4$; Connor (1966) a étendu les résultats et dérivé la formule exacte pour les cas généraux n et N, et la méthode de calcul connexe a été implémentée plus tard dans le langage Fortran par Hidiroglou and Gray (1980). La procédure est assez lourde, comme en témoignent les 165 lignes de code Fortran. Le résultat le plus curieux est probablement l'approximation asymptotique des π_{ij} calculées par Hartley and Rao (1962). Dans un article récent, Kott (2005) a montré que l'estimateur de variance d'un estimateur d'Horvitz-Thompson fondé sur l'approximation de Hartley-Rao non seulement donne de bons résultats selon le cadre fondé sur le plan de sondage, mais a aussi de bonnes propriétés sous le modèle. L'approximation de Hartley-Rao a été dérivée au départ en partant de l'hypothèse que n est fixe et que N est grand et correct jusqu'à l'ordre $O(N^{-4})$ (Hartley et Rao, 1962: Équation (5.15) à la page 369). Lors d'une conversation privée avec J.N.K. Rao durant le 23e Symposium international sur les questions de méthodologie de Statistique Canada, celui-ci a fait remarquer que l'approximation est encore valide, même si n est grand, à condition que n/N soit faible. Dans les cas où N n'est pas grand et(ou) que n/N n'est pas faible, comme dans l'exemple de l'enquête ITC-Chine considéré ici, aucune donnée n'a été publiée au sujet de la qualité de l'approximation de Hartley-Rao. Quand la méthode d'échantillonnage PPT systématique randomisé est altérée par la substitution d'unités, il est pratiquement impossible de calculer les probabilités d'inclusion de deuxième ordre ou toute approximation. En revanche, selon l'approche fondée sur la simulation, l'obtention d'estimations très fiables des π_{ij} à l'aide d'un grand nombre d'échantillons simulés reste simple, à condition que la procédure d'échantillonnage altérée soit entièrement spécifiée. Nous examinons par la suite les propriétés des estimateurs de variance en utilisant les π_{ij}^* simulées lorsque la méthode d'échantillonnage PPT systématique randomisé n'est pas altérée. Dans ces conditions, $\pi_i = nz_i$ et l'approximation $\tilde{\pi}_{ij}$ de Hartley-Rao pour π_{ij} peut aussi être incluse dans la comparaison.

Exemple 4. Nous commençons par comparer π_{ij}^* à $\tilde{\pi}_{ij}$ pour chacune des entrées individuelles. Faute d'espace, nous présentons uniquement les résultats pour $i=1,\cdots,5$ et $j=1,\cdots,10$, qui suffisent à dépeindre le tableau général. L'approximation de Hartley-Rao $\tilde{\pi}_{ij}$ est très proche des π_{ij}^* simulées, la concordance allant jusqu'à la deuxième décimale pour la majorité des entrées. Il s'agit manifestement d'une observation intéressante, sachant que $N=20$ et $n=10$.

i) π_{ij}^* simulées, $K=10^6$:

```
0,0000 0,3121 0,3821 0,2975 0,1669 0,1442 0,2116 0,2249 0,3975 0,1873
0,3121 0,0000 0,3623 0,2816 0,1590 0,1372 0,2025 0,2141 0,3766 0,1784
0,3821 0,3623 0,0000 0,3469 0,1899 0,1640 0,2483 0,2659 0,4586 0,2153
0,2975 0,2816 0,3469 0,0000 0,1523 0,1312 0,1938 0,2061 0,3606 0,1717
0,1669 0,1590 0,1899 0,1523 0,0000 0,0742 0,1124 0,1197 0,1968 0,0988
```

ii) Approximation de Hartley-Rao $\tilde{\pi}_{ij}$:

```
0,0000 0,3079 0,3769 0,2952 0,1668 0,1427 0,2143 0,2286 0,3921 0,1884
0,3079 0,0000 0,3569 0,2795 0,1579 0,1351 0,2029 0,2164 0,3712 0,1784
0,3769 0,3569 0,0000 0,3421 0,1932 0,1654 0,2484 0,2649 0,4544 0,2183
0,2952 0,2795 0,3421 0,0000 0,1514 0,1296 0,1946 0,2075 0,3559 0,1710
0,1668 0,1579 0,1932 0,1514 0,0000 0,0732 0,1099 0,1172 0,2010 0,0966
```

Exemple 5. Dans le cas des probabilités d'inclusion de deuxième ordre, l'aspect principal est l'estimation de la variance. Pour une taille d'échantillon fixe, un estimateur sans biais de la variance de l'estimateur d'Horvitz-Thompson $\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i$ est donné par la forme bien connue de Yates-Grundy,

$$v(\hat{Y}_{HT}) = \sum_{j=i+1}^{n} \sum_{i=1}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \tag{3.1}$$

Considérons les trois populations synthétiques décrites dans l'*exemple 2*. Nous obtenons la variance réelle $V = Var(\hat{Y}_{HT})$ par simulation en utilisant $B = 10^5$ échantillons simulés et en la calculant d'après la forme $B^{-1} \sum_{b=1}^{B} (\hat{Y}_b - Y)^2$, où Y est le total de population réel et $\hat{Y}_b$ est l'estimateur d'Horvitz-Thompson de Y calculé d'après le b^{e} échantillon simulé. Nous examinons trois estimateurs de la variance de la forme (3.1), désignés respectivement par v_1, v_2 et v_3, en remplaçant les π_{ij} dans (3.1) respectivement par l'approximation de Hartley-Rao $\tilde{\pi}_{ij}$, les π_{ij}^* simulées pour $K = 10^5$ et les π_{ij}^* pour $K = 10^6$. Nous évaluons la performance de ces estimateurs en nous appuyant sur le biais relatif simulé BR $= B^{-1} \sum_{b=1}^{B} (v^{(b)} - V)/V$ et l'instabilité simulée INST $= \{B^{-1} \sum_{b=1}^{B} (v^{(b)} - V)^2\}^{1/2} / V$, où $v^{(b)}$ est l'estimation de la variance calculée d'après le b^{e} échantillon, en utilisant un autre ensemble de $B = 10^5$ échantillons indépendants. Les résultats sont résumés au tableau 1 ci-après. Les trois populations sont indiquées par le coefficient de corrélation ρ entre y et x.

Tableau 1. Biais relatif et instabilité des estimateurs de variance

	BR (%)			INST		
Population	v_1	v_2	v_3	v_1	v_2	v_3
$\rho = 0{,}30$	6,1%	1,4 %	-0,3 %	0,66	0,65	0,65
$\rho = 0{,}50$	4,3 %	2,5 %	-1,1 %	0,42	0,44	0,42
$\rho = 0{,}80$	2,6 %	1,2 %	-0,2 %	0,61	0,60	0,60

En ce qui concerne le biais relatif, les trois estimateurs de variance sont acceptables, celui (v_1) fondé sur l'approximation de Hartley-Rao $\tilde{\pi}_{ij}$ ayant le biais le plus important. Dans le cas des estimateurs de variance utilisant les π_{ij}^* simulées, l'accroissement de la valeur de K de 10^5 (c'est-à-dire. v_2) à 10^6 (c'est-à-dire. v_3) rend le biais négligeable, quoique celui obtenu pour $K = 10^5$ soit clairement acceptable en pratique. Pour ce qui est de l'instabilité, les mesures sont comparables pour les trois versions de l'estimateur de la variance.

4. Quelques remarques supplémentaires

En théorie, la méthode de calcul des probabilités d'inclusion fondée sur la simulation s'applique à n'importe quel plan d'échantillonnage, à condition que l'on dispose de renseignements complets sur ce dernier. Elle convient bien au traitement de scénarios de substitution plus complexes ou d'autres types de modification du plan original. Dans le cas de l'enquête ITC-Chine, l'une des unités ayant refusé de participer a dû être remplacée par une unité provenant d'une région particulière de la ville, à cause de contraintes de charge de travail et de restrictions relatives au travail sur le terrain. Dans le cas d'une enquête nationale auprès des jeunes réalisée au Canada, certaines unités ont refusé de participer et des unités de substitution ont été sélectionnées au deuxième, ainsi qu'au troisième degré (écoles) avant d'atteindre la taille d'échantillon cible. Dans ces situations, le calcul des probabilités d'inclusion n'a pas de solution analytique, mais l'approche fondée sur la simulation peut être appliquée sans trop difficulté. Dans un article récent, Fattorini (2006) discute de l'utilisation de la méthode fondée sur la simulation dans le cas de l'échantillonnage spatial où les unités sont sélectionnées séquentiellement. Quand un plan d'échantillonnage PPT est altéré par une ou plusieurs séries de substitutions d'unités, le plan modifié peut également être considéré comme séquentiel. Toutefois, nos résultats théoriques sur

l'exactitude des méthodes fondée sur la simulation diffèrent de ceux de Fattorini. Nous avons utilisé un argument conditionnel et proposé d'évaluer les propriétés de l'estimateur en utilisant les probabilités d'inclusion simulées pour un échantillon donné présentant un intérêt en pratique. La question fondamentale que soulèvent les méthodes fondées sur la simulation est celle de la faisabilité des calculs. L'échantillonnage PPT systématique randomisé présente un énorme avantage pour ce qui est de l'efficacité des calculs. La méthode d'échantillonnage avec probabilités inégales de Rao-Sampford (Rao, 1965; Sampford, 1967), par exemple, est une autre méthode d'échantillonnage PPT populaire. Elle possède plusieurs attributs désirables, dont des expressions explicites pour les probabilités d'inclusion de deuxième ordre, et est plus efficace que l'échantillonnage PPT systématique randomisé (Asok et Sukhatme, 1976). Suit une comparaison des temps de processeur selon l'échantillonnage PPT systématique randomisé et selon échantillonnage PPT de Rao-Sampford pour la simulation des probabilités d'inclusion de premier ordre. La taille d'échantillon est fixée à $n = 10$ et le nombre d'échantillons simulés est $K = 10^6$. Les résultats sont obtenus en utilisant R sur une machine unix à processeur double.

N	PPT systématique	PPT de Rao-Sampford
200	4,7 heures	7,5 heures
100	2,5 heures	5,0 heures
50	1,6 heure	4,4 heures
20	1,2 heure	8,9 heures

Il convient de souligner que, si la méthode de Rao-Sampford demande généralement plus de temps pour obtenir les résultats, elle en prend nettement plus dans le cas où $N = 20$. Cela tient au fait que la méthode de Rao-Sampford s'appuie sur une procédure réjective et qu'il faut habituellement un grand nombre de rejets pour arriver à un échantillon final lorsque la fraction d'échantillonnage n/N est grande. L'échantillonnage PPT systématique randomisé, quant à lui, n'est pas affecté par cette situation, et la méthode fondée sur la simulation peut fournir rapidement des résultats ayant l'exactitude souhaitée pour $N = 400$, voire une valeur plus élevée. Plusieurs fonctions R/S-PLUS et les principaux codes pour l'approche proposée sont présentés à la l'annexe C et sont applicables à d'autres scénarios de substitution moyennant de légères modifications. L'une des raisons de l'utilisation de l'échantillonnage PPT systématique randomisé pour sélectionner les grappes de niveau supérieur dans l'enquête ITC-Chine est que le plan final est autopondéré. Un problème se pose en cas de refus de participer. Comment faut-il sélectionner les unités de substitution de façon que le plan d'échantillonnage modifié final soit encore (approximativement) autopondéré? Dans d'autres circonstances, comme l'échantillonnage avec renouvellement, cela est réalisable; consulter, par exemple, Fellegi (1963). La façon d'atteindre cet objectif dans le cas du plan de l'enquête ITC-Chine est à l'étude.

Remerciements

Cette étude a été financée en partie par des subventions du Conseil de recherches en sciences naturelles et en génie du Canada. Les auteurs remercient aussi de leur appui les responsables du projet d'évaluation de la politique International Tobacco Control (ITC) et du projet de l'enquête ITC-Chine. Le projet ITC est financé en partie par des subventions du Roswell Park Transdisciplinary Tobacco Use Research Center du National Cancer Institute des États-Unis (P50 CA11236) et des Instituts de recherche en santé du Canada (57897). Le financement du projet ITC-Chine est assuré par le ministère de la Santé et le ministère des Finances de la Chine.

Annexe A. Un calcul direct d'après le refus aléatoire

Sous le plan d'échantillonnage PPT systématique randomisé et en supposant que le refus est aléatoire, il est possible, en principe, de calculer directement les probabilités d'inclusion sous une règle de substitution. Le point de départ consiste à énumérer tous les échantillons initiaux possibles et leurs probabilités en se fondant sur l'ordre de population particulier utilisé pour sélectionner l'échantillon initial. Rappelons que $A_0 = 0$, $A_j = \sum_{i=1}^{j} (nz_i)$ et $A_N = n$. Pour une valeur de départ uniforme choisie $u \in [0,1]$, l'unité j est sélectionnée si

$$A_{j-1} \le u + k < A_j \tag{5.1}$$

pour une valeur de $k = 0,1,\cdots,n-1$. Soit k_j le plus grand nombre entier inférieur à A_j, et soit le reste e_j donné par $e_j = A_j - k_j$. Soit $0 < e_{(1)} \le e_{(2)} \le \cdots \le e_{(N)}$ les statistiques d'ordre des restes, et soit $k_{(1)}, \cdots, k_{(N)}$ les valeurs correspondantes de k. Notons que $e_{(N)} = 1$. Nous pourrions générer N échantillons possibles $s_1, \cdots, s_N$ avec les probabilités respectives

$$e_{(1)},\ e_{(2)} - e_{(1)},\ \cdots,\ e_{(N)} - e_{(N-1)},$$

dont certaines pourraient être égales à 0. Nous commençons par générer s_1. Partant de chaque unité $j = 1,\cdots,N$, plaçons j dans s_1 si $A_{j-1} \le k < A_j$ pour une valeur $k = 0,1,\cdots,n-1$, c.-à-d. que s_1 est sélectionné en utilisant $u = 0$ dans (5.1). À mesure que nous faisons varier u de 0 à 1, nous pouvons identifier séquentiellement différents échantillons possibles. Maintenant, sachant $s_1, \cdots, s_m$, supposons que s_{m+1} est identique à s_m excepté que le $(k_{(m)}+1)^{\text{e}}$ élément est avancé d'une valeur 1. Par exemple, supposons que $n = 4$ et que $s_m = \{1,3,6,9\}$, et supposons que $k_{(m)} = 0$. Alors, $s_{m+1} = \{2,3,6,9\}$. Par ailleurs, si $k_{(m)} = 2$, alors $s_{m+1} = \{1,3,7,9\}$. L'échantillon s_{m+1} aura la probabilité $e_{(m+1)} - e_{(m)}$. Par construction, $\pi_i = nz_i$ pour $i = 1,\cdots,N$. Si l'on souhaite calculer les probabilités d'inclusion de premier et de deuxième ordre seulement, on peut utiliser un algorithme similaire mais plus simple pour calculer directement les probabilités d'inclusion de deuxième ordre, sachant l'ordre initial. Cependant, dans les applications où les probabilités de tous les échantillons sont nécessaires, l'algorithme de génération d'échantillons peut être exécuté. Ainsi, pour les populations de petite taille, il est alors également possible de calculer les probabilités d'inclusion de premier ordre en cas de refus et de substitution. Supposons que nous sélectionnions pour commencer un échantillon de taille n par échantillonnage PPT systématique randomisé. Supposons que n_1 unités de cet échantillon consentent à répondre et qu'un nombre $n_2 = n - n_1$ supplémentaire soit sélectionnées, de nouveau par échantillonnage PPT systématique randomisé à partir des unités n'ayant pas été échantillonnées la première fois. Supposons pour simplifier que les cas de refus dans le premier échantillon surviennent au hasard et qu'il n'y en a aucun dans l'échantillon de substitution. Notons que cette hypothèse diffère de celle utilisée dans l'exemple 3, où l'ensemble des cas de refus est considéré comme étant non aléatoire. La probabilité d'inclusion de l'unité i, sachant l'ordre de population initial supposé est

$$nz_i \times \frac{n_1}{n} + \sum_{m: i \notin s_m} p_1(s_m) \frac{n_2 z_i}{\sum_{j: j \notin s_m} z_j}. \tag{5.2}$$

La somme externe est calculée sur l'ensemble des échantillons s_m de taille n, générés conformément au processus décrit plus haut, mais sans avoir d'unité i, avec les probabilités $p_1(s_m) = e_{(m)} - e_{(m-1)}$. La somme interne figurant au dénominateur est calculée sur l'ensemble des unités j non incluses dans les échantillons s_m de la somme externe. Les probabilités d'inclusion inconditionnelles peuvent être obtenues par calcul de la moyenne approprié sur l'ensemble des ordres de population qui produisent des valeurs distinctes. Cela n'est manifestement faisable que si la population est petite ou que z prend un petit nombre de valeurs.

Annexe B. Calcul de l'expression (2.1)

Nous montrons ici que, pour tout $\varepsilon > 0$ et un échantillon donné s,

$$P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} < \varepsilon \right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2} \left(\sum_{i \in s} \frac{1}{\pi_i} - n \right),$$

où $\hat{T} = \sum_{i \in s} y_i / \pi_i$, $\tilde{T} = \sum_{i \in s} y_i / \pi_i^*$, et π_i^* sont les probabilités d'inclusion de premier ordre simulées à partir de K échantillons indépendants. En notant que $E(\pi_i^*) = \pi_i$ et $Var(\pi_i^*) = \pi_i(1-\pi_i)/K$, en vertu de l'inégalité de Chebyshev, nous avons $P(|\pi_i^* - \pi_i| > c) \le \pi_i(1-\pi_i)/(Kc^2)$ pour tout $c > 0$. Il s'ensuit que :

$$P\left(\frac{|\pi_i^*-\pi_i|}{\pi_i^*}>\varepsilon\right)=P(\pi_i^*-\pi_i>\pi_i^*\varepsilon)+P(\pi_i^*-\pi_i<-\pi_i^*\varepsilon)$$
$$=P(\pi_i^*-\pi_i>\varepsilon\pi_i/(1-\varepsilon))+P(\pi_i^*-\pi_i<-\varepsilon\pi_i/(1+\varepsilon))$$
$$\le P(|\pi_i^*-\pi_i|>\varepsilon\pi_i/(1-\varepsilon))+P(|\pi_i^*-\pi_i|>\varepsilon\pi_i/(1+\varepsilon))$$
$$\le\frac{(1-\varepsilon)^2\pi_i(1-\pi_i)}{K\varepsilon^2\pi_i^2}+\frac{(1+\varepsilon)^2\pi_i(1-\pi_i)}{K\varepsilon^2\pi_i^2}$$
$$=\frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\frac{1}{\pi_i}-1\right).$$

Si $y_i\ge 0$ pour tout i, alors

$$|\hat{T}-\tilde{T}|\le\sum_{i\in s}\frac{y_i}{\pi_i}\frac{|\pi_i^*-\pi_i|}{\pi_i^*}\le\max_{i\in s}\left\{\frac{|\pi_i^*-\pi_i|}{\pi_i^*}\right\}\hat{T}.$$

Pour tout $\varepsilon>0$ et l'échantillon donné s,

$$P\left(\frac{|\hat{T}-\tilde{T}|}{\hat{T}}<\varepsilon\right)\ge P\left(\max_{i\in s}\left\{\frac{|\pi_i^*-\pi_i|}{\pi_i^*}\right\}<\varepsilon\right)$$
$$\ge 1-\sum_{i\in s}P\left(\frac{|\pi_i^*-\pi_i|}{\pi_i^*}\ge\varepsilon\right)$$
$$\ge 1-\frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i\in s}\frac{1}{\pi_i}-n\right).$$

Annexe C. Mise en œuvre en R/S-PLUS

C1. Une fonction R pour l'échantillonnage PPT systématique randomisé
Les variables d'entrée de la fonction sont `x:` le vecteur de population de variables de taille et `n:` la taille de l'échantillon. La fonction syspps donne l'ensemble de n unités sélectionnées.

```
syspps<-function(x,n){
N<-length(x)
U<-sample(N,N)
xx<-x[U]
z<-rep(0,N)
for(i in 1:N) z[i]<-n*sum(xx[1:i])/sum(x)
r<-runif(1)
s<-numeric()
for(i in 1:N){
if(z[i]>=r)
 s<-c(s,U[i])
  r<-r+1
            }
            }
return(s[order(s)])
                    }
```

C2. Une fonction R pour la simulation des probabilités d'inclusion de deuxième ordre
Les variables d'entrée de la fonction sont `x:` le vecteur de population des variables de taille et `s:` l'ensemble d'étiquettes des unités dans l'échantillon. La méthode d'échantillonnage par défaut est l'échantillonnage PPT systématique randomisé et le nombre d'échantillons répétés est $K=10^6$. La fonction `piij` donne une matrice de dimensions $n\times n$ où la $(ij)^{\text{e}}$ entrée est la probabilité simulée π_{ij}^*, $i,j\in s$.

```
piij<-function(x,s){
N<-length(x)
n<-length(s)
p<-matrix(0,n,n)
for(k in 1:1000000){
ss<-syspps(x,n)
for(i in 1:(n-1)){
for(j in (i+1):n){
if(min(abs(ss-s[i]))+min(abs(ss-s[j]))==0) p[i,j]<-p[i,j]+1
                  }
                  }
                   }
p<-(p+t(p))/1000000
return(p)
                     }
```

C3. Une fonction R pour l'échantillonnage PPT sous substitution d'unités

La méthode par défaut pour la sélection de l'échantillon initial et de l'échantillon de substitution est l'échantillonnage PPT systématique randomisé. La fonction R suivante `sysppssub` est utilisée pour simuler les probabilités d'inclusion sous substitution d'unités. Les variables d'entrée sont `x:` le vecteur de population des variables de taille, `n:` la taille de l'échantillon et `refus:` l'ensemble d'unités de l'échantillon initial qui refusent de répondre. La fonction donne un ensemble d'unités pour l'échantillon final.

```
sysppssub<-function(x,n,refus){
s<-syspps(x,n)
sub<-numeric()
for (i in 1:n){
if(min(abs(s[i]-refus))==0) sub<-c(sub,i)
               }
m<-length(sub)
if(m>0){
s<-s[-sub]
U1<-(1:length(x))[-c(refus,s)]
x1<-x[-c(refus,s)]
s1<-syspps(x1,m)
s<-c(s,U1[s1])
               }
return(s[order(s)])
                              }
```

C4. Codes R pour la simulation des π_i sous substitution d'unités

```
pi<-rep(0,N)
for(i in 1:1000000){
s<-sysppssub(x,n,refus)
for(j in 1:N){
if(min(abs(s-j))==0)
pi[j]<-pi[j]+1
             }
                    }
pi<-pi/1000000
```

Références

Asok, C., et Sukhatme, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *J. Amer. Statist. Assoc.* , 912-918.
estimator. *Survey Methodology* , 189-196.

Chao, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika* , 653-656.

Chen, X.H., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* , 457-469.

Connor, W.S. (1966). An exact formula for the probability that two specified sampling units occur in a sample drawn with unequal probability and without replacement. *J. Amer. Statist. Assoc.* , 384-390.

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

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika* , 269-278.

Fellegi, I. P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *J. Amer. Statist. Assoc.* , 183-201.

Goodman, R. et Kish, L. (1950). Controlled selection – A technique in probability sampling. *J. Amer. Statist. Assoc.* , 350-372.

Hartley, H.O. et Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Ann. Math. Statist.* , 350-374.

Hidiroglou, M.A. et Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics* , 107-112.

Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics* , 433-439. Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *J. Indian Statist. Assoc.* , 173-180.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* , 499-513.

Tillé, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika* , 238-241.