Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Procédures d'échantillonnage pour évaluer l'exactitude du couplage d'enregistrements

Paul A. Smith, Shelley Gammon, Sarah Cummins, Christos Chatzoglou et Dick Heasman[1]

Résumé

Les ensembles de données administratives servent de plus en plus couramment de sources de données en statistique officielle dans le cadre d'efforts visant à produire plus efficacement un plus grand nombre de produits. De nombreux produits résultent du couplage de deux ensembles de données ou plus, souvent réalisé en plusieurs phases en appliquant différentes méthodes et règles. Dans ces situations, nous aimerions pouvoir évaluer la qualité du couplage, ce qui comprend une certaine réévaluation des appariements ainsi que des non-appariements. Dans le présent article, nous discutons de méthodes d'échantillonnage en vue d'obtenir des estimations des nombres de faux négatifs et de faux positifs en exerçant un contrôle raisonnable sur l'exactitude des estimations ainsi que sur les coûts. Des approches de stratification des appariements (non-appariements) pour l'échantillonnage sont évaluées en utilisant des données provenant du recensement de la population de l'Angleterre et du Pays de Galles de 2011.

Mots-clés : Échantillonnage inverse, échantillonnage contrôlé, erreur de couplage.

1. Introduction

Le couplage (appariement) d'enregistrements joue un rôle de plus en plus important dans la production des statistiques officielles en raison des pressions accrues concernant l'utilisation d'ensembles de données administratives comme composante de leur production. Par conséquent, il importe de pouvoir évaluer la qualité des processus d'appariement. Il est généralement reconnu (p. ex., Ferrante et Boyd 2012, Randall *et coll.* 2013, Vatsalan *et coll.* 2014) que les mesures clés sont la précision du processus d'appariement, exprimée sous la forme $P = \frac{VP}{L} = \frac{VP}{VP + FP}$ où *VP* est le nombre d'appariements corrects (« vrais positifs »), et *FP* est le nombre de faux positifs parmi *L* appariements faits durant le processus d'appariement, et le rappel, exprimé sous la forme $R = \frac{VP}{VP + FN}$ où *FN* est le nombre d'appariements manqués (« faux négatifs »). Quand une mesure globale est nécessaire, la *f*-mesure $f = 2\frac{PR}{P + R}$ (la moyenne harmonique de la précision et du rappel) quantifie le compromis entre les deux mesures.

Afin d'estimer ces mesures de manière raisonnable pour des ensembles de données susceptibles de contenir de nombreux appariements, nous avons besoin d'estimations pour *VP*, *FP* et *FN*. Si l'objectif est de procéder à des évaluations qui constituent une « norme de référence » reposant sur un examen détaillé (résolution manuelle) des appariements possibles, une procédure d'échantillonnage est nécessaire pour définir les nombres appropriés d'appariements et de non-appariements qu'il convient d'examiner afin de faire des estimations de la précision et du rappel suffisamment exactes. Dans le présent article, nous exposons comment une procédure d'échantillonnage stratifiée peut être utilisée pour ce processus, et indiquons les données d'entrée et les sources de données nécessaires pour affiner l'échantillonnage.

[1] Paul A. Smith, S3RI, University of Southampton, Highfield, Southampton, SO17 1BJ, Royaume-Uni, p.a.smith@soton.ac.uk; Shelley Gammon, Sarah Cummins, Christos Chatzoglou, et Dick Heasman, Office for National Statistics, Segensworth Road, Fareham, PO15 5RR, Royaume-Uni

Un processus complet de couplage d'enregistrements comprend habituellement plusieurs étapes (dénotées ici par *h* car nous considérerons des stratifications fondées sur ces étapes) – par exemple, d'abord un processus d'appariement exact, puis diverses procédures fondées sur des règles, et enfin des méthodes d'appariement probabilistes, qui peuvent être répétées en adoptant une gamme de stratégies de groupement (mise en bloc) (Herzog et coll., section 12.1). Dans le présent article, nous utilisons l'appariement des enregistrements du recensement de l'Angleterre et du Pays de Galles de 2011 et de l'Enquête sur la couverture du recensement (ECR) à titre d'exemple; dans ce cas, le résultat de l'appariement manuel découlant du traitement des données du recensement est considéré comme étant l'état de « vrai » appariement, de sorte que celui-ci est connu. Nous évaluons ensuite un processus d'appariement automatisé, d'abord en utilisant sept clés d'appariement (une phase déterministe), puis une régression logistique (une phase probabiliste). Nous supposons que la précision, p, varie en fonction de la méthode d'appariement (Winkler 2004 indique, par exemple, que la variation des appariements issus de l'application de différentes définitions de clés de groupement (*blocking passes*) peut être considérable).

2. Stratification

L'échantillonnage stratifié est le plus efficace quand les caractéristiques d'intérêt sont homogènes dans les strates, et qu'elles sont hétérogènes entre les strates (Cochran 1977, section 5.7). Par conséquent, afin de choisir une stratification efficace, nous recherchons des variables qui ont une bonne puissance explicative pour la caractéristique à étudier, et les utilisons de préférence pour définir les strates.

Nous présumons qu'une stratégie d'appariement de haute qualité comprend un certain nombre d'étapes, et que les valeurs de précision varient selon l'étape à laquelle l'appariement a été déclaré (qui devrait donc être consignée). La population de faux négatifs est constituée de toutes les paires qui n'ont été reconnues comme un appariement à aucune étape. Par conséquent, elles ne se distinguent pas par l'étape d'appariement, qui n'est pas utilisée pour estimer le rappel. Nous aimerions à la fois tirer parti des différences entre les étapes pour concevoir l'échantillonnage pour l'évaluation de la précision et obtenir des estimations par étape pour fournir des renseignements en retour sur l'utilité des étapes. Nous nous attendons donc à ce que la stratification par la méthode ou procédure de groupement qui a permis de déterminer l'appariement soit utile. Il se peut que les caractéristiques des paires d'enregistrements (autres que la méthode de couplage) aient aussi une incidence sur la probabilité d'un appariement vrai, et ce genre d'information peut également être utilisé dans la stratification pour estimer le rappel.

Au départ, nous ne disposons d'aucune information permettant de déterminer quelles variables sont importantes pour la stratification. Cependant, une fois qu'une évaluation de la qualité a été faite, nous possédons un échantillon de paires d'enregistrements avec des résultats « vrais » qui peuvent être analysés pour orienter des études plus approfondies. Une évaluation doit donc nous fournir des données contenant la paire d'enregistrements, y compris les variables d'intérêt pour la stratification (p. ex., emplacement, caractéristiques démographiques, paradonnées (par exemple, temps écoulé entre les dates d'extraction/de collecte de l'information dans les enregistrements à apparier)). Nous avons également besoin du résultat de l'évaluation de la qualité, c'est-à-dire si l'appariement a été jugé correct ou incorrect. Nous utilisons ces données pour élaborer un modèle de régression logistique destiné à cerner les meilleurs prédicteurs de l'indicateur de vrais/faux appariements.

Pour estimer FN (ce que nous faisons en estimant $Q = FN/(FN + VN)$ et en multipliant par le nombre de non-appariements), nous regardons parmi les paires d'enregistrements correspondant à des non-appariements et examinons l'évaluation de la qualité pour voir si ces paires auraient dû ou non être des appariements. Bien que toutes les paires de cas aient passé toutes les étapes sans qu'il y ait un appariement, la probabilité d'un appariement correct à l'issue de l'étape d'appariement probabiliste finale sera vraisemblablement un prédicteur important. Le nombre de paires qui sont des non-appariements est, en général, très grand, et bon nombre de ces paires sont des vrais non-appariements (VN). Afin que l'évaluation soit pratique, nous supposerons généralement qu'un grand nombre de ces paires (p. ex., appariements homme-femme) ont une probabilité tellement faible d'être un faux non-appariement (c.-à-d. d'être un vrai appariement) que l'on peut les ignorer. Par conséquent, nous aurons une strate qui n'est jamais prise en considération pour l'échantillonnage (un échantillon avec seuil d'inclusion, Haziza et coll. 2010, Smith 2013 section 5.3.4). Pour les paires qui sont des non-appariements, dont est exclue cette strate avec seuil d'inclusion, nous pouvons construire un modèle de régression logistique pour le résultat « vrai » de façon similaire.

2.1 Prédicteurs des cas faussement positifs en utilisant le couplage recensement-ECR

Pour étudier les facteurs associés aux erreurs sous forme de résultats faussement positifs, nous avons d'abord utilisé une approche entièrement automatisée pour refaire l'appariement des enregistrements du Recensement de 2011 et de l'Enquête de couverture du recensement (ECR) par couplage déterministe et probabiliste (groupement par code postal). Les paires d'enregistrements appariés ont ensuite été analysées en utilisant le couplage des enregistrements du Recensement et de l'ECR de 2011 comme ensemble de données appariées de référence pour indiquer l'état d'appariement « vrai ».

Un total de 619 458 appariements ont été effectués, dont 0,22 % étaient des faux positifs (appariements incorrects). Vu que la construction d'un modèle en partant des données complètes engloutirait les faux appariements, 500 paires d'enregistrements étant des vrais positifs et 500 paires d'enregistrements étant des faux positifs ont été échantillonnées (par EAS) comme données d'apprentissage pour le modèle. Les échantillons ont été tirés dans les strates définies par les méthodes de couplage, proportionnellement à la taille des strates, afin d'être certain que toutes les méthodes de couplage soient représentées.

Un modèle de régression logistique pour l'état de vrai appariement (faux positif = 0, vrai positif = 1) a été construit en se servant des variables explicatives disponibles :

- indice de dénombrement difficile au recensement (1 à 5) (Hopper, 2011)
- sexe (1, 2, ou données manquantes)
- groupe d'âge (0-17, 18-24, 25-39, 40-64, 65+, données manquantes)
- vit à Londres (1) ou ne vit pas à Londres (0)
- groupe ethnique (blanc, asiatique/asiatique britannique, noir/noir britannique, mixte, autre, données manquantes)
- étape d'appariement – exact (1), fondé sur des règles (2 à 7), probabiliste (8)

Les variables de sexe, groupe d'âge et groupe ethnique ont été vérifiées pour déceler les conflits dans les paires d'enregistrements appariés, en excluant les cas de données manquantes. Puisque la variable de groupe ethnique a été codée différemment pour le Recensement et pour l'ECR, les codes ont été appariés de façon aussi proche que possible. L'examen a révélé 3,84 % de conflits dans les paires d'enregistrements pour le groupe ethnique, 0,29 % de conflits pour le sexe et 0,70 % de conflits pour l'âge. Ces proportions ont été considérées comme peu susceptibles d'avoir un effet sur le modèle.

Une procédure de modélisation par étapes a été utilisée sur huit échantillons différents de 1 000 enregistrements. L'effet de l'étape d'appariement était fortement significatif ($p < 0,0001$) dans tous les échantillons et l'indice de difficulté de dénombrement avait un effet fortement significatif ($p < 0,0001$) dans six des huit échantillons, ce qui indique que ces deux variables sont des prédicteurs puissants des appariements faussement positifs. Comme les catégories définies pour le type d'appariement et l'indice de difficulté de dénombrement n'avaient pas toutes un effet significatif, nous avons procédé à l'analyse des taux de résultats faussement positifs et fait appel au jugement de spécialistes pour simplifier ces catégories : l'indice de difficulté de dénombrement a été recodé (1,2) et (3,4,5), et le type d'appariement a été recodé (1), (2-7) et (8), ce qui correspond aux appariements exact, déterministe et probabiliste.

La modélisation a été effectuée en utilisant le critère d'information bayésien (BIC; ou critère de Schwarz) pour choisir entre les modèles, ce qui offre une bonne protection contre le surajustement. Le modèle simplifié (recodé) possédait un BIC beaucoup plus faible que le modèle original, ce qui indiquait un meilleur ajustement du modèle. Pour le modèle final, l'aire sous la courbe ROC était de 0,8634, indiquant que le modèle est un bon prédicteur des résultats faussement positifs.

2.2 Prédicteurs des cas faussement négatifs en utilisant le couplage Recensement-ECR

Les paires d'enregistrements non appariés décelées durant l'analyse automatisée et les résultats de référence provenant du couplage original des enregistrements du Recensement et de l'ECR de 2011 (voir 2.1) ont été utilisés comme entrées dans un modèle pour les faux négatifs.

Un peu plus de 4 millions de paires d'enregistrements ont été rejetées à l'étape de l'appariement probabiliste, dont 16 000 (0,40 %) étaient de vrais appariements (faux négatifs). Vu que la construction d'un modèle en utilisant les données complètes engloutirait les faux négatifs, 500 paires d'enregistrements étant des vrais négatifs et 500 paires d'enregistrements étant des faux négatifs ont été échantillonnées (par EAS) pour servir de données d'apprentissage pour le modèle. Les échantillons ont été tirés dans les groupes de scores d'appariement probabiliste, proportionnellement à la taille des strates, pour être certain qu'une gamme de probabilités soit représentée.

Un modèle de régression logistique pour l'état de vrai non-appariement (vrai négatif = 0, faux négatif = 1) a été construit en se servant des variables explicatives disponibles :

- indice de dénombrement difficile au recensement (1 à 5)
- probabilité d'appariement à l'étape probabiliste d'appariement (0-0,1, 0,1-0,2, …, 0,9-1)
- vit à Londres (1) ou ne vit pas à Londres (0)
- sexe dans le recensement (1, 2 ou 0 pour données manquantes)
- groupe d'âge dans le recensement (0-17, 18-24, 25-39, 40-64, 65+, données manquantes)
- groupe ethnique dans le recensement (blanc, asiatique/asiatique britannique, noir/noir britannique, mixte, autre, données manquantes)

Les variables du recensement pour le groupe d'âge, le sexe et le groupe ethnique ont été prises en considération dans le modèle; cependant, celles-ci sont fort probablement en conflit avec les valeurs de l'ECR, surtout pour les vrais négatifs. La proportion de conflits était élevée pour le sexe (47 %), le groupe d'âge (42 %) et le groupe ethnique (36 %). Ce fait doit être pris en compte avant d'interpréter les résultats.

Une procédure de modélisation par étapes a été utilisée sur huit échantillons différents de 1 000 enregistrements. L'effet du groupe de probabilité ($p < 0,0001$) et du groupe d'âge ($p < 0,05$) était significatif dans tous les échantillons et celui du sexe ($p < 0,05$) était significatif dans sept des huit échantillons, ce qui indique que ces trois variables sont de puissants prédicteurs des faux négatifs. Comme les catégories définies pour ces variables n'avaient pas toutes un effet significatif, nous avons procédé à une analyse des taux de résultats faussement négatifs et fait appel au jugement de spécialistes pour simplifier les catégories. Le BIC (critère de Schwarz) a de nouveau été utilisé pour évaluer la qualité de l'ajustement du modèle à titre de protection contre le surajustement.

Le taux de faux négatifs pour le groupe de probabilité était fortement affecté par le choix du score seuil dans la phase d'appariement automatisé (dans ce cas, 0,5), les paires d'enregistrements juste en dessous du seuil ayant un taux de faux négatifs beaucoup plus élevé que celles situées au-dessus (excepté pour celles possédant un score très élevé). Les catégories 0-0,1 (1), 0,1-0,3 (2), 0,3-0,5 (3), 0,5-0,7 (4), 0,7-0,9 (5), 0,9-1 (6) ont produit un bon ajustement du modèle. Le groupement des scores de probabilité doit tenir compte de la méthode probabiliste et du seuil de désignation d'un appariement utilisé. Pour le groupe d'âge, la catégorie pour les données manquantes était celle produisant le taux le plus élevé de faux négatifs ainsi que la catégorie d'âge le plus avancé (65+), de sorte que les catégories ont été groupées comme il suit : données manquantes (0), 0-64 (1), 65+ (2). De même, pour la variable de sexe, la catégorie des données manquantes présentait le taux le plus élevé de faux négatifs tandis que les taux pour les catégories hommes et femmes étaient similaires, de sorte que la catégorie des données manquantes a été retenue et les catégories hommes et femmes ont été groupées. Les deux variables groupées ont amélioré l'ajustement du modèle. Le modèle final avec les scores de probabilité, les groupes d'âge et le sexe groupés davantage possédait une aire sous la courbe ROC de 0,8785, ce qui indique que ces facteurs sont de bons prédicteurs des faux négatifs.

3. Détermination et répartition de la taille des échantillons

Après avoir déterminé une stratification appropriée, par une analyse de données antérieures ou par analogie avec des données antérieures, ou par hypothèse, nous devons spécifier le niveau de qualité cible des estimations de la précision et du rappel, et les utiliser pour déterminer les tailles globales d'échantillon requises et comment elles doivent être réparties entre les strates. Dans l'exposé qui suit, nous nous concentrons sur la précision, qui est le cas le plus simple.

Les estimations de la précision et du rappel sont des estimations de proportions, p et r respectivement, et très souvent, l'on s'attend à ce que ces proportions soient assez proches de 1. Il existe donc un risque que la borne supérieure des intervalles de confiance soit supérieure à 1. En outre, quand p devient petite, une contrainte de

variance est insuffisante pour permettre de tester les différences entre des valeurs significatives de *p*. Pour maîtriser ces risques, il est nettement préférable d'atteindre un coefficient de variation cible plutôt qu'une variance cible, particulièrement lorsqu'on ne dispose d'aucune information *a priori* pour donner une estimation initiale de *p*.

3.1 Répartition de Neyman

Cochran (1977, section 5.12) donne une expression pour la taille de l'échantillon dans le cas de l'échantillonnage stratifié de proportions sous une répartition optimale présumée en utilisant une variance cible. Elle nécessite des estimations initiales des proportions de strate $\tilde{p}_h$. La variance peut être remplacée par $c^2\tilde{p}^2$ où *c* est le cv cible et $\tilde{p}$ est une estimation initiale de la précision globale, ce qui donne une expression pour la taille requise d'échantillon pour un cv fixé,

$$n_0 = \frac{\left(\sum_h W_h \sqrt{\tilde{p}_h(1-\tilde{p}_h)}\right)^2}{c^2\tilde{p}^2}, \qquad n = \frac{n_0}{1+\frac{1}{Nc^2\tilde{p}^2}\sum_h W_h\tilde{p}_h(1-\tilde{p}_h)}$$

où n_0 fournit une première approximation si la correction pour population finie est négligeable et *n* tient compte de cette correction. Cette procédure dépend toutefois des estimations initiales $\tilde{p}_h$, qui doivent être réalistes. Cependant, afin d'exercer un meilleur contrôle sur le cv, nous aimerions appliquer une méthode qui ne dépend pas des proportions $\tilde{p}_h$ supposées.

3.2 Échantillonnage inverse

Haldane (1945) a présenté une technique, connue sous le nom d'échantillonnage inverse, qui contrôle approximativement le coefficient de variation en utilisant un plan séquentiel qui se poursuit jusqu'à ce que *m* événements aient eu lieu, c'est-à-dire, pour estimer la précision, jusqu'à ce que *FP* = *m*. Dans ce cas, la taille d'échantillon finale est une variable aléatoire, ce qui est plus compliqué du point de vue de la planification des ressources, mais qui permet l'accumulation d'information selon une approche séquentielle.

Nous pouvons déduire la valeur requise de *m* pour des contraintes de cv fixes dans une strate unique en utilisant le résultat de Haldane selon lequel $\hat{v}(\hat{p}) \approx \frac{\hat{p}^2(1-\hat{p})}{m-2}$, l'approximation étant bonne pour de faibles valeurs de $\hat{p}$, la probabilité estimée d'un FP. Résoudre l'équation pour trouver *m* donne $m = \frac{(1-\hat{p})}{c^2} + 2$, où *c* est le cv cible, de nouveau avec l'approximation vérifiée pour de faibles valeurs de $\hat{p}$. Par conséquent, *m* est approximativement constant et indépendant de $\hat{p}$ sur des étendues qui sont importantes pour l'évaluation de l'exactitude de l'appariement. Cochran (1977, p77) utilise une approximation différente, mais sa solution est très proche de l'approximation dérivée de l'estimateur de variance de Haldane.

Il est par conséquent simple d'utiliser l'échantillonnage inverse dans une grande population, triée aléatoirement, avec des probabilités variables d'observer un « succès » pour obtenir un estimateur sans biais de la probabilité moyenne dans la population. Dans notre cas, nous aimerions en outre avoir des renseignements sur les probabilités dans chaque strate (à titre d'information sur la qualité de la stratégie de couplage), et nous pourrions espérer obtenir un résultat d'une exactitude similaire en utilisant une plus petite taille d'échantillon si nous étions capables d'utiliser l'échantillonnage inverse dans les strates. Une valeur de *m* approximativement constante ne se traduit toutefois pas en une taille d'échantillon constante, et des tailles d'échantillon prévues très grandes peuvent découler de petites valeurs de $\hat{p}$. Mais pour l'échantillonnage inverse, il n'est pas important de disposer d'estimations exactes (ou même de toute estimation) de $\hat{p}$ pour déterminer l'échantillonnage – il suffit de procéder à l'accumulation de « succès » vers le total cible *r*.

3.3 Échantillonnage inverse stratifié

Il est possible d'écrire une expression pour le cv d'un échantillonnage inverse stratifié en utilisant l'estimateur sans biais de Finney (1949) pour la variance d'un échantillon inverse. Aucune solution analytique au problème de minimisation de $\sum_h n_h$ pour obtenir un cv fixe dans ces conditions n'est encore disponible. Nous adoptons plutôt une approche numérique heuristique. L'algorithme de base (avec certains commentaires concernant sa mise en œuvre) est le suivant :

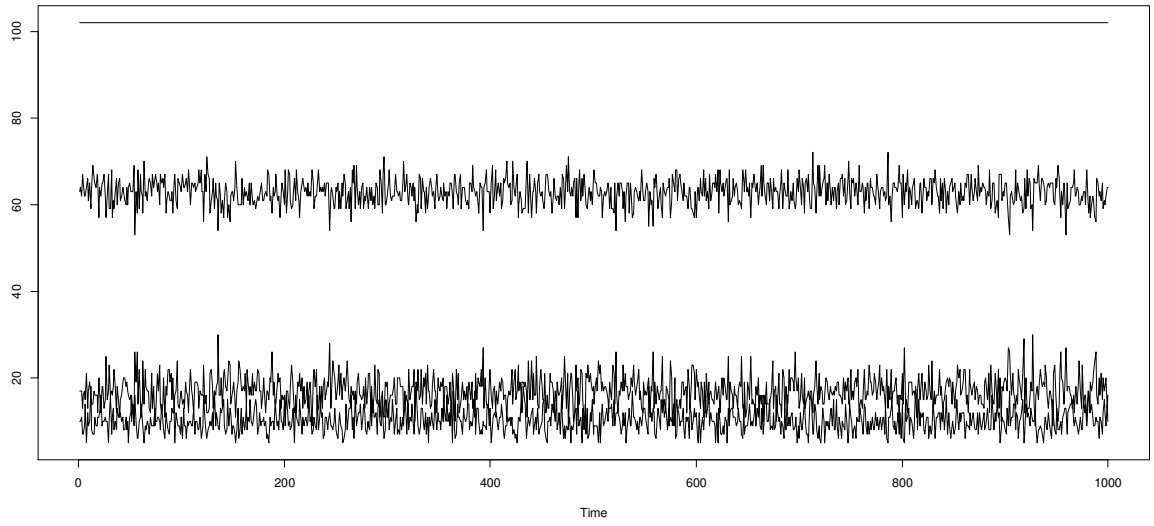
Étape	Commentaires
1. Attribuer une valeur minimale à m_h dans chaque strate.	Le minimum peut être fixé en vue d'améliorer l'exactitude moyennant un faible coût quand les succès sont relativement communs comme dans les plans d'échantillonnage généralisés de Kim et Nachlas (1984).
2. Évaluer le cv en utilisant les tailles d'échantillon de l'étape 1.	Le cv est estimé à la volée en utilisant les résultats des premiers cas.
3. Si la cible de cv n'est pas atteinte, ajouter 1 à la valeur cible de m_h pour chaque strate tour à tour, en laissant les autres inchangées, et recalculer le cv attendu (cv obtenu si le « succès » suivant a lieu après $1/\hat{p}_h$ essai). Garder l'échantillon avec l'unité supplémentaire ayant le cv global le plus faible.	Besoin du cv *espéré* car c'est la seule information disponible séquentiellement. La simulation peut « caler » si une variance domine. La fréquence de cette situation peut être réduite considérablement (mais non éliminée) en limitant le nombre d'augmentations consécutives de m_h dans la même strate (10 est utilisé comme un maximum dans de nombreuses simulations ici); la strate offrant le deuxième meilleur effet sur le cv global est choisie à la place.
4. Répéter l'étape 3 jusqu'à ce que la cible de cv soit atteinte ou que toutes les strates atteignent un maximum.	La taille de l'échantillon peut être plafonnée pour contrôler les coûts avec une pénalité associée en termes de variance, comme dans les plans d'échantillonnage généralisés de Kim et Nachlas (1984).

Tableau 3.3-1
Petit exemple à trois strates. $\tilde{p}_h$ est supposée connue.

Strate	N (millions)	p
1	5	0,01
2	3	0,03
3	2	0,20

Considérons le petit exemple du tableau 3.3-1. (Une valeur de $p = 0{,}2$ dans la strate finale est plutôt grande pour l'hypothèse d'un « petit p » en échantillonnage inverse.) L'application de l'algorithme susmentionné 1 000 fois sur la série générée à partir du tableau 3.3-1 donne les résultats présentés à la figure 3.3-1. La droite supérieure représente la valeur constante $m = 102$ requise pour un cv de 0,1 en échantillonnage inverse. Dans un certain sens, nous nous intéressons davantage à la taille globale d'échantillon, et la sommation de m_h dans l'approche des trois strates donne $m \approx 90$. Donc, du moins en ce qui concerne m, la stratification permet de réaliser une économie. Cependant, compte tenu de la façon dont ces m se traduisent en tailles d'échantillon réalisées n, il n'existe essentiellement aucune différence – pour l'échantillonnage simple, $n = 1\,884$ en moyenne, et pour 3 strates, $n = 1\,905$.

Figure 3.3-1
Nombres de succès requis pour cv = 0,1 en traitant l'ensemble de la population en une fois (droite supérieure), et en utilisant l'algorithme d'échantillonnage inverse stratifié (les trois courbes inférieures, une par strate).



4. Application à un exemple artificiel de neuf strates

L'extension à un exemple de neuf strates plus réaliste (mais toujours artificiel) (tableau 4-1) donne les résultats présentés à la figure 4-1. La répartition de Neyman, effectuée avec les probabilités correctes et, par conséquent, la meilleure qui puisse être réalisée (mais peu susceptible d'être approchée en pratique), semble très similaire en qualité à l'échantillonnage inverse. Cela est corroboré par les boîtes à moustache des cv de la figure 4.1. Même si l'étendue des cv sous la répartition de Neyman est plus grande, il s'agit partiellement d'un artéfact, parce que, dans l'échantillonnage inverse, le critère d'arrêt est le moment où le cv requis est obtenu, ce qui réduit la variation des cv. Il convient aussi de noter que l'échantillonnage inverse stratifié semble produire une estimation de la probabilité globale qui présente un léger biais par défaut. Ce résultat semble aller à l'encontre de l'intuition, puisqu'il existe un estimateur sans biais dans chaque strate et que les pondérations sont connues, mais il se répète d'une simulation à l'autre.

Tableau 4-1
Exemple à neuf strates

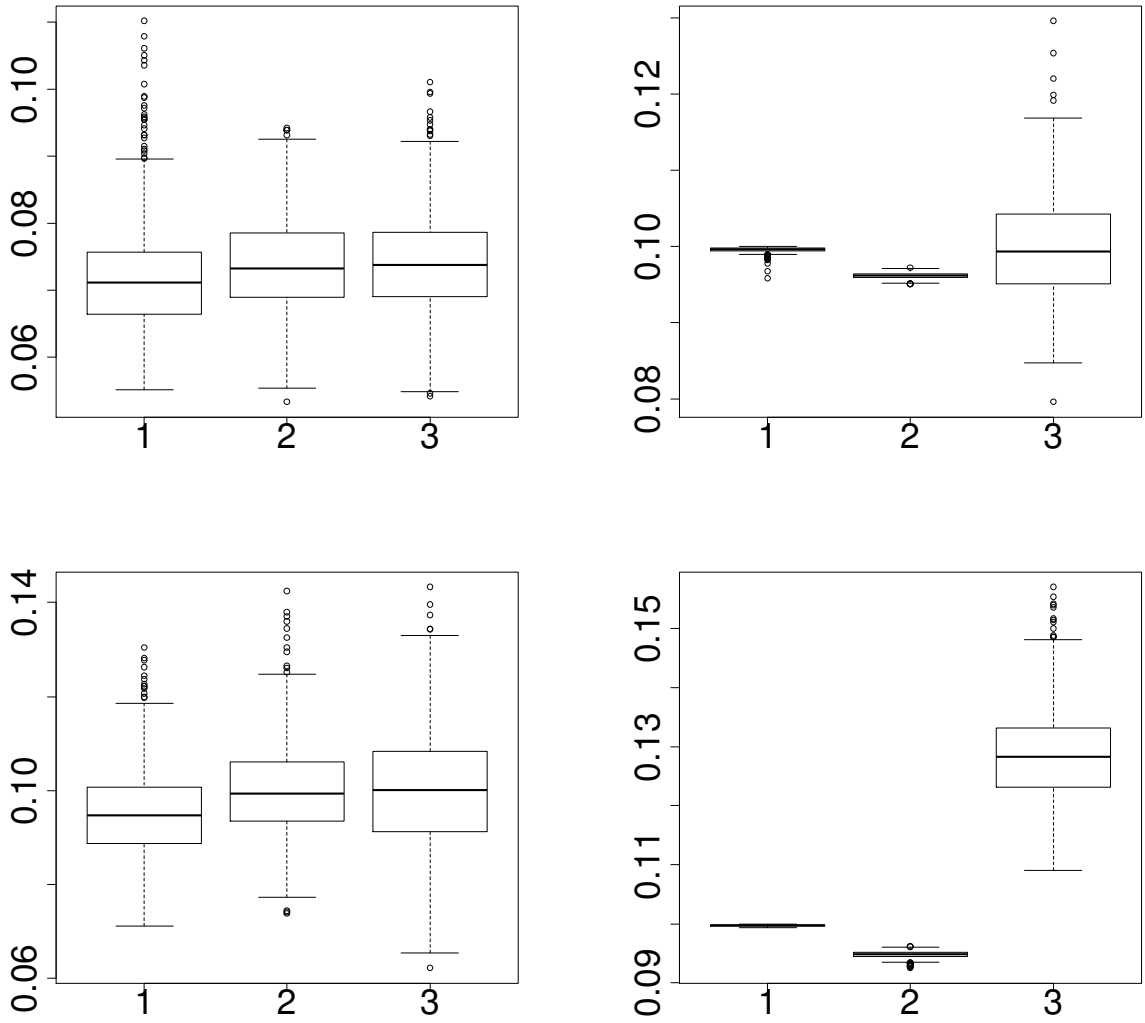
Strate	*N* (de paires, millions)	Pr(*FP*) utilisée dans la répartition de Neyman	vraie Pr(*FP*)	
			Ensemble A	Ensemble B
1	20	0,005	0,005	0,1
2	20	0,01	0,01	0,1
3	16	0,05	0,05	0,1
4	10	0,06	0,06	0,1
5	10	0,07	0,07	0,1
6	90	0,1	0,1	0,1
7	7	0,2	0,2	0,1
8	5	0,3	0,3	0,1
9	3	0,4	0,4	0,1

Par contre, si l'on compare les tailles globales d'échantillon, celles découlant de la répartition de Neyman en vue d'atteindre un cv cible sont généralement plus petites que la taille totale d'échantillon sous échantillonnage inverse

pour le même cv. Dans ce cas, où le coût de la résolution manuelle des appariements est assez élevé, une stratégie générale d'échantillonnage inverse pourrait aboutir à des coûts élevés.

Figure 4-1

Estimations de $\hat{p}$ (à gauche) et cv estimés (à droite) sous 1 – échantillonnage inverse stratifié, 2 – échantillonnage inverse sur l'échantillon complet (non stratifié), 3 – échantillonnage stratifié avec répartition de Neyman. Les graphiques supérieurs correspondent à la répartition de Neyman calculée à partir des probabilités correctes (ensemble A dans le tableau 4-1), tandis que les graphiques inférieurs s'appuient sur la même répartition, mais avec les probabilités réelles constantes et fixées à 0,1 (ensemble B).



5. Discussion

De façon inattendue, l'échantillonnage inverse stratifié semble ne pas offrir beaucoup d'avantages par rapport à l'échantillonnage inverse ordinaire (mais ne semble pas non plus être pénalisé). Son principal avantage est qu'il permet d'utiliser l'échantillonnage inverse pour fournir des indicateurs de la qualité des appariements issus de chaque étape du processus d'appariement. La taille d'échantillon requise pour l'échantillonnage inverse est en général plus grande que pour une répartition de Neyman en utilisant le même cv cible. Les cv sous la répartition de Neyman sont généralement plus variables, car la procédure numérique heuristique pour l'échantillonnage inverse stratifié s'arrête quand (et uniquement quand) la cible est atteinte.

Cela suggère la stratégie d'échantillonnage qui suit pour évaluer la qualité du couplage :

1. Si des estimations raisonnables de $\tilde{p}_h$ sont disponibles, les utiliser dans une répartition de Neyman dans un plan de sondage stratifié. Cela donnera la plus petite taille d'échantillon associée à une chance raisonnable d'obtenir le cv requis.
2. Si ces estimations ne sont pas disponibles et qu'une estimation globale seulement est nécessaire, utiliser l'échantillonnage inverse sur des données triées aléatoirement.
3. Si des estimations distinctes dans les strates sont souhaitées, suivre l'algorithme pour l'échantillonnage inverse stratifié.

Il pourrait exister d'autres options où l'échantillonnage inverse est utilisé pour obtenir des estimations grossières de $\tilde{p}_h$ qui sont ensuite introduites dans la répartition de Neyman, ce qui pourrait offrir un meilleur contrôle de la taille globale de l'échantillon quand les $\tilde{p}_h$ sont inconnues au départ. Ces options doivent faire l'objet d'études plus approfondies.

Bibliographie

Cochran, W.G. (1977), *Sampling techniques*. New York: Wiley.

Ferrante, A. & J. Boyd (2012), "A Transparent and Transportable Methodology for Evaluating Data Linkage Software", *Journal of Biomedical Informatics*, 45, pp. 165-172.

Haldane, J.B.S. (1945), "On a Method of Estimating Frequencies", *Biometrika*, 33, pp. 222-225.

Haziza, D., G. Chauvet, and J.C. Deville (2010), "A Note on Sampling and Estimation in the Presence of Cut-off Sampling", *Australian and New Zealand Journal of Statistics*, 52, pp. 303-319.

Herzog, T. N., F. J..Scheuren, and W. E. Winkler (2007), *Data quality and record linkage techniques*. New York: Springer Science & Business Media.

Hopper, N.A. (2011), "Predicting patterns of household non-response in the 2011 Census", *Survey Methodology Bulletin*, 69, pp. 9-22.

Kim, S., and J.A. Nachlas (1984), "Estimation in Bernoulli Trials Under a Generalized Sampling Plan", *Technometrics*, 26, pp. 379-387.

Randall, S.M., A.M. Ferrante, J.H. Boyd, and J.B. Semmens (2013), "The Effect of Data Cleaning on Record Linkage Quality", *BMC Medical Informatics and Decision Making*, 13:64, pp. 1-10.

Smith, P. (2013), "Sampling and Estimation for Business Surveys", in G. Snijkers et al. *Designing and conducting business surveys*. Hoboken, New Jersey: Wiley, pp. 165-218.

Vatsalan, D., P. Christen, C. O'Keefe, and V.S. Verykios (2014), "An Evaluation Framework for Privacy-Preserving Record Linkage", *Journal of Privacy and Confidentiality*, 6, pp. 35-75.

Winkler, W. E. (2004), "Approximate string comparator search strategies for very large administrative lists", *Proceedings of the Section on Survey Research Methods*, 2004, pp. 4595-4602.