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

Actes du symposium de Statistique Canada, 2006
Enjeux méthodologiques reliés à la mesure de la santé des populations

Risque de divulgation et estimation de la variance

W. Wilson Lu et Randy R. Sitter[1]

Résumé

La protection contre la divulgation de l'identité des répondants dans les données d'enquête publiées constitue un enjeu d'ordre pratique pour de nombreux organismes gouvernementaux. Parmi les méthodes de protection figurent la suppression des identificateurs de grappe et de strate, de même que la modification des données ou la permutation des valeurs entre les enregistrements des répondants. Malheureusement, les identificateurs de grappe et de strate sont généralement nécessaires à l'estimation de la variance axée sur la linéarisation ainsi qu'aux méthodes de répétition, dans la mesure où le rééchantillonnage porte habituellement sur les unités de sondage du premier degré dans les strates. On pourrait penser que la diffusion d'un ensemble de poids de rééchantillonnage duquel les identificateurs de strate et de grappe auraient été supprimés permettrait de régler une partie du problème, particulièrement si l'on fait appel à une méthode de rééchantillonnage aléatoire, comme celle du *bootstrap*. Dans le présent article, nous démontrons dans un premier temps que, en considérant les poids de rééchantillonnage comme des observations dans un espace dimensionnel de haut niveau, on peut facilement utiliser un algorithme de mise en grappes pour reconstruire les identificateurs de grappe, peu importe la méthode de rééchantillonnage, même si les poids de rééchantillonnage ont été modifiés aléatoirement. Nous proposons ensuite un algorithme rapide qui permet de permuter les identificateurs de grappe et de strate des unités finales avant la création des poids de rééchantillonnage, sans influer de façon significative sur les estimations de la variance des caractéristiques visées qui en résultent. Ces méthodes sont illustrées par leur application aux données publiées issues des *National Health and Nutrition Examination Surveys*, enquêtes pour lesquelles les questions de divulgation sont extrêmement importantes.

MOTS CLÉS : Répliques répétées équilibrées (BRR), *bootstrap*, confidentialité, *jackknife*.

1. Introduction

La protection contre la divulgation involontaire de l'identité des répondants dans les fichiers de données d'enquêtes complexes faisant l'objet d'une diffusion publique s'impose comme un enjeu de plus en plus important, à mesure que s'élargit l'accès Internet aux données du recensement et à d'autres données auxiliaires et que s'accroissent la vitesse et la convivialité des ressources informatiques. La majeure partie de la documentation dans ce domaine porte sur le masquage des données diffusées. Une mesure simple et très fréquente que l'on utilise pour éviter la divulgation de l'identité consiste à supprimer les identificateurs des strates et des unités primaires d'échantillonnage (UPE) (identificateurs des grappes de premier degré dans une enquête fondée sur un échantillon à plusieurs degrés). Cette suppression peut faire en sorte qu'il est plus difficile pour le « pirate » d'apparier une grappe échantillonnée et une grappe de population et donc de préciser ses recherches en vue de retracer l'identité des répondants. Mais elle peut rendre aussi difficile l'estimation directe de la variance puisque les identificateurs de strate et d'UPE sont généralement requis pour l'obtention d'estimateurs asymptotiquement sans biais de la variance.

On pourrait croire alors que la diffusion, avec les données, d'un ensemble de poids de rééchantillonnage duquel les identificateurs de strate et d'UPE auraient aussi été supprimés permettrait de régler le problème. Selon Yung (1997), tel n'est pas nécessairement le cas. Il propose un *bootstrap* répété pour tenter de créer un ensemble de poids de rééchantillonnage à partir duquel il est plus difficile de reconstruire les identificateurs d'UPE et/ou de strate. Comme nous le montrerons, cette méthode n'atteint pas son objectif et, en fait, il est possible de reconstituer rapidement et aisément les identificateurs d'UPE à partir d'à peu près n'importe quel ensemble de poids de rééchantillonnage.

[1]W. Wilson Lu, Département de mathématiques et de statistiques, Université Acadia, 12, av. University, Wolfville (N.-É.), Canada, B4P 2R6; Randy R. Sitter, Département de statistiques et d'actuariat, Université Simon Fraser, 8888, ch. University, Burnaby (C.-B.), Canada, V5A 1S6.

Une autre méthode couramment utilisée pour masquer les données issues d'enquêtes complexes consiste à modifier les valeurs ou à permuter les valeurs de données entre les cas. Dans notre application, il s'agirait de permuter les identificateurs de strate et/ou d'UPE avant de créer l'ensemble de poids de rééchantillonnage. Cette opération n'aurait d'incidence que sur l'estimation de la variance et pourrait être effectuée de manière à limiter cette incidence sur les estimations de la variance des caractéristiques mesurées par l'enquête qui en résultent. Cette idée est proposée par Lu (2004), et une méthode semi-manuelle de mise en œuvre dans le cadre de l'édition 2001-2002 des *National Health and Nutrition Examination Surveys* (NHANES) est présentée dans Dohrmann et coll. (2006). Il s'agit essentiellement d'apparier les grappes du second degré pour certaines variables démographiques au moyen d'un algorithme de couplage des enregistrements, puis d'examiner les appariements afin de repérer ceux qui feront l'objet d'une permutation des identificateurs d'UPE pour toutes les unités finales de la grappe du second degré.

Le présent article vise deux objectifs : 1) illustrer à quel point il est facile de repérer les UPE à partir des poids de rééchantillonnage; et 2) proposer une solution axée sur un algorithme rapide permettant de permuter les identificateurs d'UPE dans les poids de rééchantillonnage.

2. Risques de divulgation à partir des poids de rééchantillonnage

2.1 Enquêtes complexes, poids déterminés par le plan d'échantillonnage et rééchantillonnage

Pour comprendre les méthodes d'estimation de la variance par rééchantillonnage, examinons un plan de sondage à plusieurs degrés dans lequel les UPE (grappes) sont sélectionnées par tirage avec remise, ou sont traitées ainsi pour les fins de l'estimation de la variance, et dans lequel des sous-échantillons indépendants sont prélevés dans des grappes qui ont été sélectionnées plus d'une fois. Supposons que n_h UPE sont sélectionnées avec des probabilités p_{hi} avec remise ou avec des probabilités d'inclusion $\pi_{hi} = n_h p_{hi}$ indépendamment dans chaque strate. Supposons que $\hat{Y}_{hi}$ est un estimateur linéaire sans biais du vecteur des totaux pour la i^{e} UPE de la strate h pour l'échantillonnage du second degré et des degrés subséquents, de sorte que $\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} / (n_h p_{hi})$ est un estimateur linéaire sans biais du vecteur des totaux de la strate Y_h. Un estimateur linéaire sans biais du total $Y = \sum_h Y_h$ est ensuite donné par $\hat{Y} = \sum_h \hat{Y}_h$. On peut le représenter par $\hat{Y} = \sum_{(hik)\in s} w_{hik} y_{hik}$, où s est l'échantillon total des éléments, et w_{hik} et $y_{hik} = (y_{1hik}, y_{2hik}, \ldots, y_{phik})'$ désignent respectivement le poids d'échantillonnage (ou déterminé par le plan d'échantillonnage) et le vecteur des valeurs rattachées au hik^{e} élément échantillonné $(k = 1, \ldots, n_{hi}; i = 1, \ldots, n_h; h = 1, \ldots, H)$.

Souvent, un estimateur d'enquête peut s'exprimer sous la forme d'une fonction du vecteur des totaux estimés, c.-à-d. $\hat{\theta} = g(\hat{Y})$. La fonction de la répartition de la population peut être estimée par $\hat{F}_n(t) = \sum_s w_{hik} I_{[y_{hik} \le t]} / \sum_s w_{hik}$, où $I_{[\bullet]}$ est la fonction de l'indicateur, les p^{e} quantiles d'échantillon peuvent être estimés par $\hat{F}^{-1}(p)$, et $\hat{F}^{-1}$ est la fonction inverse de $\hat{F}$.

2.2 Poids de rééchantillonnage et identificateurs de strate/UPE

Dans cette section, nous supprimons le triple indice et posons $\hat{Y} = \sum_{j\in s} w_j y_j$, où $j = (hik)$. Les données d'intérêt public desquelles les identificateurs de strate et d'UPE ont été supprimés seraient diffusées ainsi, dans un ordre déterminé aléatoirement.

Tableau 1. Représentation matricielle des poids déterminés par le plan d'échantillonnage et des poids de rééchantillonnage

ÉCHANTILLLON	*CARACTÉRISTIQUES*	*POIDS DÉTERMINÉS PAR LE PLAN D'ÉCHANTILLONNAGE*	*POIDS DE RÉÉCHANTILLONNAGE*
1	y_1	w_1	$w_{1(1)}\ w_{1(2)}\ \cdots\ w_{1(R)}$
2	y_2	w_2	$w_{2(1)} w_{2(2)}\ \cdots\ w_{2(R)}$
…	…	…	… … … … … …
m	y_m	w_m	$w_{m(1)} w_{m(2)}\ \cdots w_{m(R)}$

Tous les estimateurs habituels de la variance par rééchantillonnage (*jackknife*, *bootstrap*, répliques répétées équilibrées de Fay) peuvent être réécrits en fonction de la sélection de *R* sous-ensembles de l'échantillon complet $y_1, y_2, \ldots, y_m$, où *m* est le nombre total d'unités finales dans l'échantillon, selon un mécanisme quelconque de rééchantillonnage permettant de produire les estimations de la répétition représentées par $\hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(R)}$. La r^e estimation de la répétition, $\hat{\theta}_{(r)}$, est calculée de la même manière que $\hat{\theta}$ mais à partir du r^e ensemble de poids de rééchantillonnage $w_{j(r)}, j = 1, \ldots, m$. La variation entre les estimations de la répétition, $v(\hat{\theta}) = \sum_{r=1}^{R} c_r (\hat{\theta}_{(r)} - \hat{\theta})^2$, sert ensuite à estimer $V(\hat{\theta})$, où c_r représente les constantes propres à la méthode de répétition.

La forme habituelle des données diffusées est présentée au tableau 1. L'utilisateur final peut ensuite employer le programme qui calcule $\hat{\theta}$ à partir de $y_1, y_2, \ldots, y_m$ et $w_1, w_2, \ldots, w_m$ pour obtenir $\hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(R)}$ en l'appliquant à $y_1, y_2, \ldots, y_m$ et $w_{1(r)}, w_{2(r)}, \ldots, w_{m(r)}$ pour $r = 1, \ldots, R$. Afin de comprendre comment on peut utiliser les poids de rééchantillonnage du tableau 1 pour reconstituer les identificateurs d'UPE et/ou de strate même s'ils sont présentés dans un ordre établi aléatoirement, supposons que $\delta_{j(r)} = w_{j(r)} / w_j$, les ratios des poids de rééchantillonnage et des poids déterminés par le plan d'échantillonnage, où $j = 1, \ldots, m$ et $r = 1, \ldots, R$ présentés au tableau 2.

Nous voulons maintenant démontrer à quel point il est facile de reconstituer les identificateurs d'UPE à partir du tableau 2. Nous appliquons ensuite la démarche à des données réelles et effectuons une brève étude numérique pour illustrer la simplicité et l'efficacité de la méthode proposée, même lorsque les poids ont été perturbés de façon aléatoire ou ont fait l'objet d'un rajustement quelconque.

Pour ce faire, considérons la i^e ligne du tableau 2 comme un objet à *R* dimensions avec les entrées $\delta_{j(1)}, \ldots, \delta_{j(R)}$ et définissons la distance, $d(j,l)$ de toute paire d'éléments de l'échantillon j et l. On peut aisément voir que si l'on applique un algorithme de mise en grappes à ces lignes, les unités de la même UPE se retrouveront dans la même grappe.

Gardant à l'esprit la mise en grappes des unités échantillonnées, nous n'avons qu'à trouver un algorithme de mise en grappes, de préférence un qui soit facile d'accès et convivial, et effectuer un test pour déterminer si l'algorithme peut ou non repérer les unités des UPE avec un degré élevé de précision. Un examen rapide des progiciels courants R et SAS nous a permis de trouver dans ceux-ci la fonction « hclust » et la procédure « Proc FASTCLUS », respectivement. Toutes les deux acceptent des données multivariées et forment des arbres de grappes. Bien que,

selon nous, les deux donnent de bons résultats, la procédure de SAS est nettement plus rapide et peut traiter facilement des ensembles de données très importants, de sorte que nous limiterons nos discussions à la procédure « Proc FASTCLUS ». Nous n'avons pas effectué d'opérations plus complexes pour bien montrer qu'il est possible d'obtenir ces résultats même avec des compétences et des outils rudimentaires.

Nous avons utilisé un ensemble de données publiées tirées des NHANES assorti de 42 ensembles de poids de rééchantillonnage BRR de Fay. Pour déterminer l'incidence de l'ajout de bruit aux poids de rééchantillonnage dans le but de limiter la capacité qu'a le pirate d'obtenir les identificateurs d'UPE, nous introduisons un bruit aléatoire $\varepsilon_{j(r)}$ aux poids de rééchantillonnage, ces poids de rééchantillonnage perturbés étant représentés par $w^*_{j(r)} = w_{j(r)}(1+\varepsilon_{j(r)}) = \delta_{j(r)}(1+\varepsilon_{j(r)})w_j = \delta^*_{j(r)}w_j$, où $\delta_{j(r)} = w_{j(r)}/w_j$ sont les éléments apparaissant au tableau 2, $\delta^*_{j(r)}$, en sont la version perturbée, et $\varepsilon_{j(r)}$ sont les $U(-\delta,\delta)$ indépendants et pareillement distribués. Pour diverses valeurs de δ, nous appliquons la procédure « Proc FASTCLUS » (un simple code de SAS disponible auprès des auteurs) comme nous l'avons décrit plus haut pour obtenir les unités des UPE. Si nous supposons que le nombre d'UPE est connu, ce qui est souvent le cas, la méthode permet, dans tous les cas, d'attribuer correctement les unités aux UPE. Si nous établissons que le nombre maximum de grappes est supérieur au nombre réel de grappes, les résultats obtenus sont, dans tous les cas, une partition de l'ensemble réel des UPE, c'est-à-dire que « Proc FASTCLUS » produit plus de grappes qu'il n'y en a réellement, mais uniquement par la subdivision d'UPE.

Pour montrer que la méthode *bootstrap* ou *M-bootstrap* de Yung (1997) offre encore moins de protection, nous avons conçu l'étude numérique suivante : 1) créer 100 ensembles de poids *bootstrap*, chacun d'eux ayant une moyenne de 20, conformément à la méthode décrite dans l'article de Yung (1997); 2) appliquer ensuite la méthode en utilisant seulement 2 000 des 9 965 poids de rééchantillonnage et uniquement les répétitions 2, 3, 4 et 5. Nous obtenons, pour l'attribution des unités aux UPE originales, un taux d'erreur de 2,5 % pour deux répétitions et de 0 pour trois répétitions ou plus, respectivement. Nous répétons la simulation sans faire la moyenne des poids, c'est-à-dire selon la méthode classique du *bootstrap* avec $m^*_h = n_h - 1$. Les taux d'erreur s'établissent alors à 47,5 %, 28 %, 5,5 % et 1,5 % pour 2 à 5 répétitions, respectivement. L'algorithme de mise en grappes donne de très bons résultats quant à la reconstitution des UPE originales dans les deux cas, même lorsqu'on n'utilise qu'un petit nombre d'ensembles de poids de rééchantillonnage.

En résumé, il est évident qu'un utilisateur, et même un utilisateur non chevronné, peut utiliser les poids de rééchantillonnage, quelle que soit la façon dont ils sont créés, avec ou sans ajustement, pour reconstituer les identificateurs des UPE originales assez facilement. De plus, la perturbation aléatoire des poids de rééchantillonnage n'offre qu'une protection minimale.

3. Algorithme de permutation proposé

3.1 Méthode de la permutation séquentielle

La méthode proposée dans la présente section consiste à permuter les identificateurs d'UPE des unités finales, avant la création des poids de rééchantillonnage ou la création des pseudo-UPE aux fins de l'estimation de la variance. Nous désignons parfois cette méthode par le terme « permutation d'unités entre les UPE » puisqu'il s'agit essentiellement de la même chose. Cette opération doit être effectuée de manière à réduire au minimum les effets sur les estimations de la variance des principales caractéristiques. Idéalement, l'opération devrait être exécutée au moyen d'un algorithme rapide automatique. Cet algorithme de permutation devrait respecter les critères suivants.

1. Puisque l'un de nos principaux objectifs est d'empêcher les utilisateurs finals d'avoir accès aux identificateurs des UPE originales, nous devons permuter une partie considérable des unités dans chaque UPE originale pour qu'il soit impossible de repérer celle-ci dans toute analyse par grappes des configurations des poids de rééchantillonnage. De plus, dans toute pseudo-UPE constituée, le nombre d'unités de l'UPE originale, quelle qu'elle soit, ne devrait pas être indûment élevé.

2. L'algorithme doit limiter l'incidence de l'opération sur l'estimation de la variance.

Dohrmann et coll. (2006) emploient une méthode en deux étapes. Il s'agit, dans un premier temps, d'apparier des unités de différentes UPE puis, à la deuxième étape, de permuter une proportion, déterminée par l'utilisateur, de ces unités appariées entre les UPE. Nous proposons plutôt une méthode de permutation séquentielle, qui vise à éviter la permutation simultanée d'une partie des unités appariées par paires. Nous établissons au lieu de cela une règle permettant de déterminer la paire optimale d'unités aux fins de la permutation en fonction d'un critère d'optimalité à l'étape en cours, puis nous effectuons la permutation et nous répétons l'opération jusqu'à ce que suffisamment d'unités aient été permutées.

Supposons que nous avons déterminé une mesure de distance appropriée selon nos préférences et nos besoins (voir la section suivante pour une discussion de cette question). Plus la distance entre deux unités est petite, plus la probabilité de permutation de ces unités est élevée. Nous classons tout d'abord toutes les paires possibles d'unités $n(n-1)/2$ en ordre croissant de distance. Une fois les paires ainsi classées, il nous suffit de sélectionner les paires d'unités admissibles qui se caractérisent par la distance la plus courte à cette étape et de procéder à la permutation. Posons que $u_{hi} = \lfloor \alpha * n_{hi} \rfloor + 1$ représente le nombre minimum d'unités de l'UPE *hi* devant être permutées et $v_{hi} = \lfloor \beta * u_{hi} \rfloor$, le nombre maximum d'unités devant être permutées entre l'UPE *hi* et une autre UPE, $\lfloor \bullet \rfloor$, le plus grand entier inférieur ou égal, et β, un paramètre d'ajustement qui empêche un taux de permutation indûment élevé entre deux UPE quelles qu'elles soient. Représentons l'ensemble de toutes les paires possibles d'unités $n(n-1)/2$ par $A = \{(j,l) : (h_j i_j k_j), (h_l i_l k_l) \in s\}$.

L'algorithme proposé est simple et très rapide parce que, après le tri initial, il reste très peu de calculs à effectuer à l'étape de l'examen et de la permutation. Cette simplicité fait également de l'algorithme un outil très souple qui permet de tenir compte, le cas échéant, de différentes contraintes et exigences.

3.2 Mesure de distance

Cet algorithme de permutation séquentielle nécessite que l'on définisse une mesure de distance permettant de déterminer les unités qui seront permutées. On pourrait utiliser, par exemple, $d(i_1, i_2) = 1$ ou 0 si les unités i_1 et i_2 appartiennent ou non à la même catégorie dans le cas de variables nominales; et $d(i_1, i_2) = | y_{i_1} - y_{i_2} | / fourchette(y_i)$, dans le cas de variables continues, ce qui rééchelonnerait la mesure dans la fourchette [0, 1] pour toutes les variables. L'importance relative des variables peut être traitée par des poids multiplicatifs.

Nous tenons compte de cette méthode dans la prochaine section qui évalue la performance (représentée par D_3). Cependant, dans le contexte de la permutation des identificateurs d'UPE spécifiquement aux fins de la création des poids de rééchantillonnage pour l'estimation de la variance, on peut justifier plus facilement une mesure de distance exprimant la volonté explicite de limiter l'incidence sur les estimations de la variance qui en résultent. Pour le démontrer, examinons tout d'abord l'estimateur linéaire $\hat{Y}$. Tous les estimateurs habituels de la variance par rééchantillonnage se réduisent à (approximativement pour le *bootstrap*) :

$$v(\hat{Y}) = \sum_{h=1}^{H} n_h^{-1} (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(y_{hi} - \bar{y}_h)^T ,$$

où $y_{hi} = \sum_{k=1}^{n_{hi}} w_{hik} y_{hik}$ et $\bar{y}_h = \sum_{i=1}^{n_h} y_{hi} / n_h$.

L'incidence de la permutation entre les identificateurs d'UPE des unités $\{(hik) \in s_0 \subset s\}$ et les identificateurs d'UPE des unités $\{(hik) \in s_1 \subset s \cap s_0^c\}$ sera $v'(\hat{Y}) = v(\hat{Y}) + \Delta_{01}$, où v' est l'estimateur de la variance appliqué après la permutation et Δ_{01} est l'incidence de la permutation des données au lemme 1.

Lemme 1. Posons $A_{01} = \{(j,l) : identificateur\ de\ grappe\ (h_j i_j k_j) \in s_0$ permuté avec $(h_l i_l k_l) \in s_0\}$ et $\Delta_{jl} = w_{h_j i_j k_j} y_{h_j i_j k_j} - w_{h_l i_l k_l} y_{h_l i_l k_l}$ pour la paire ordonnée $(j,l) \in A_{01}$. Alors

$$\Delta_{01} = \sum_{(j,l)\in A_{01}} \{(\frac{n_{h_j}-1}{n_{h_j}} + \frac{n_{h_l}-1}{n_{h_l}})\Delta_{jl}\Delta_{jl}^T + (\frac{y_{h_j i_j} - \bar{y}_{h_j}}{n_{h_j}} - \frac{y_{h_l i_l} - \bar{y}_{h_l}}{n_{h_l}})\Delta_{jl}^T + \Delta_{jl}(\frac{y_{h_j i_j} - \bar{y}_{h_j}}{n_{h_j}} - \frac{y_{h_l i_l} - \bar{y}_{h_l}}{n_{h_l}})^T\},$$

où la somme porte sur les paires (j,l) de sorte que chaque paire n'apparaît qu'une seule fois dans la somme (preuve disponible auprès des auteurs).

Le lemme 1 met en évidence l'importance des poids déterminés par le plan d'échantillonnage. Une méthode simple d'intégration des poids consiste à appliquer les mesures habituelles de distance à $w_{hik} y_{hik}$ plutôt qu'à y_{hik} (représentée par D_1), c'est-à-dire pour tenter de réduire tous les δ_{jl}. Cette façon de procéder contribue à réduire le plus possible l'incidence de la permutation sur l'estimation de la variance pour les variables à l'étude. Une autre méthode consiste simplement à inclure les poids en les considérant comme une variable supplémentaire dans la permutation (représentée par D_2), c'est-à-dire pour réduire la distance entre les poids et entre les *y*'. Ces méthodes sont également abordées à la section suivante.

Peu importe laquelle de ces méthodes est retenue pour déterminer une mesure de distance, on doit aussi modifier ou pénaliser la mesure en fonction des exigences énoncées au critère 1 de la section 3.1, pour s'assurer que les unités ne sont pas permutées au sein des UPE (et, dans une moindre mesure, au sein des strates). Une façon simple de le faire est d'ajouter une pénalité. Supposons que $\gamma_1 + \gamma_2$ est supérieur à la plus grande distance entre deux unités données. Modifions ensuite la mesure de distance ainsi et prenons

$$d^*(i,j) = d(i,j) + \gamma_1 I_{[i,j\in même\ strate]} + \gamma_2 I_{[i,j\in même\ UPE]}.$$

Puisque $I_{[i,j\in même\ UPE]} = 1$ implique que $I_{[i,j\in même\ strate]} = 1$, cette opération pénalise la permutation au sein des UPE davantage qu'au sein des strates.

Dans certains cas, on retrouve des UPE qui présentent des risques de divulgation plus élevés. Cela peut se produire pour les UPE de petite taille et/ou d'une configuration démographique distinctive. Dans de tels cas, il serait préférable d'ajouter à la mesure de distance un terme qui favoriserait la permutation entre les UPE à faible risque de divulgation et les UPE à risque élevé de divulgation. Par exemple, posons que $\delta_i = 1$ ou -1 si la i^e UPE est à risque élevé ou à faible risque, respectivement, et ajoutons le terme $\gamma_3(\delta_i\delta_j + 1)$. Ce terme impose une pénalité de $2\gamma_3$ à la distance si les UPE *i* et *j* sont toutes deux à faible risque ou toutes deux à risque élevé.

4. Évaluation des résultats

Nous appliquons l'algorithme proposé et la méthode d'appariement-permutation de Dohrmann et coll. (2006) aux fichiers-échantillons des examens et des caractéristiques des personnes de la NHANES de 2003-2004 (voir les liens à la NHANES à l'adresse http://www.cdc.gov/nchs/), déjà rendus publics, pour comparer la vitesse d'exécution, la similitude des unités permutées et la souplesse de l'algorithme de permutation séquentielle et de l'algorithme d'appariement-permutation. Pour ce faire, nous faisons comme si les identificateurs des pseudo-UPE et des strates utilisés ici sont les identificateurs réels d'UPE et de strate, puis nous appliquons les diverses stratégies de masquage de ces identificateurs.

Nous choisissons deux groupes de caractéristiques pour les fins de notre simulation. Le premier groupe de quatre caractéristiques démographiques (sexe, âge, ethnicité et revenu familial) et de cinq variables relatives à l'examen médical (poids, taille, indice de masse corporelle, tension artérielle systolique et tension artérielle diastolique) sert à déterminer la distance entre les enregistrements appariés. Ces variables ont été choisies pour qu'il soit possible de les corréler avec une vaste gamme de variables relatives à la santé et à la nutrition. Le second groupe de 26 variables

relatives aux examens en laboratoire (mesurant l'albumine et la créatinine dans l'urine, la formule sanguine et le profil biochimique) et de deux variables de mensurations issues de l'examen médical permet d'évaluer la performance des algorithmes proposés en ce qui a trait à l'estimation de la variance des variables n'ayant pas servi à la procédure de permutation. L'ensemble de données complet de 2003-2004 renferme 10 122 enregistrements, chacun de ceux-ci étant associé à un identificateur d'UPE numéroté de 1 à 30. Toutefois, nous n'utilisons que les 6 217 enregistrements sans valeurs manquantes à l'étape de la permutation. Notre but est d'appliquer les algorithmes proposés afin de permuter les identificateurs d'UPE pour un certain pourcentage d'enregistrements sans modifier de façon significative les estimateurs de la variance de toutes les variables du premier groupe et du deuxième groupe qui en résultent.

Nous avons appliqué l'algorithme d'appariement-permutation et l'algorithme proposé de permutation séquentielle à $\alpha = 10$ %, 20 %, 30 % et 40 %, soit le pourcentage requis d'unités devant être permutées dans une UPE quelle qu'elle soit. L'algorithme proposé de permutation séquentielle nous assure une meilleure maîtrise de la procédure de permutation en ce qui concerne la source des unités permutées dans une UPE donnée. En plus de α , nous pouvons définir la quantité β comme la limite supérieure du pourcentage d'unités faisant l'objet d'une permutation entre toute autre UPE et l'UPE-cible. En introduisant β , nous pouvons suivre la composante de chaque pseudo-UPE formée, de manière à ce que les unités permutées qu'elle renferme proviennent d'une diversité d'UPE originales, ce qui présentera des avantages sur le plan de la confidentialité. Pour évaluer la performance, nous utilisons l'écart relatif absolu (ERA) procentuel moyen des estimateurs de la variance avant et après la permutation, défini ainsi $ERA = \sum_{p=1}^{q} | v'(\hat{Y}_p) - v(\hat{Y}_p) | / [qv(\hat{Y}_p)]$, où la somme porte sur les q variables et où v et v' représentent, respectivement, l'estimateur de la variance avant et après la permutation. Les $q=9$ variables du premier groupe ont servi à la permutation axée sur les trois mesures de distance traitées à la section 3 : D_1) intégration des poids par l'application de la mesure de distance à $w_{hik} y_{hik}$; D_2) intégration des poids par leur inclusion à titre de variable supplémentaire; et D_3) sans intégration de poids.

Tableau 2. *Méthode d'appariement-permutation* : ERA des variables **utilisées** et **non utilisées** dans la permutation

Dist	*K*	*Variables **utilisées** dans la permutation*					*Variables **non utilisées** dans la permutation*			
		Temps machine (en secondes)	*α=10 %*	*20 %*	*30 %*	*40 %*	*α=10 %*	*20 %*	*30 %*	*40 %*
D1	5	481	0,188	0,293	0,763	1,543	3,732	6,530	8,491	9,136
	10	567	0,222	0,363	1,027	1,564	1,848	4,945	6,160	7,113
	20	668	0,173	0,227	1,105	1,664	4,199	5,472	6,028	8,276
	40	949	0,147	0,281	0,762	1,725	3,930	4,317	6,018	8,503
D2	5	290	0,288	0,157	0,369	0,288	1,558	3,056	3,766	3,285
	10	347	0,519	0,439	0,307	0,397	1,769	3,734	6,101	5,137
	20	447	0,399	0,485	0,313	0,581	1,707	3,725	6,093	5,451
	40	682	0,413	0,475	0,322	0,658	1,718	3,725	6,164	5,504
D3	5	285	1,156	1,105	1,875	2,930	2,954	5,720	8,092	8,781
	10	338	1,741	2,906	3,378	4,781	3,546	6,860	9,962	11,83
	20	448	2,191	2,739	1,721	1,317	4,514	7,282	9,352	9,934
	40	696	1,943	2,583	1,554	1,175	4,260	7,135	9,262	9,871

La première partie du tableau 2 donne le temps d'exécution en secondes (sur un ordinateur portatif Dell D810 doté d'un processeur de 2GHz et d'une mémoire vive de 1GO) et l'ERA des variables utilisées dans la permutation pour

divers α par la méthode de l'appariement-permutation selon diverses valeurs *K*, alors que la deuxième partie du tableau donne l'ERA pour les *q=28* variables non utilisées dans la permutation. Comme on peut le voir, D_1 donne de meilleurs résultats que D_2 qui, elle, donne de meilleurs résultats que D_3.

Tableau 3. *Méthode de la permutation séquentielle* : ERA des variables **utilisées** et **non utilisées** dans la permutation

Dist	*β/α*	*Variables **utilisées** dans la permutation*				*Variables **non utilisées** dans la permutation*			
		10 %	*20 %*	*30 %*	*40 %*	*10 %*	*20 %*	*30 %*	*40 %*
D1	10 %	0,052	0,144	0,359	0,468	0,42	1,72	2,34	4,07
	20 %	0,055	0,172	0,284	0,410	0,44	1,78	2,26	4,05
	30 %	0,047	0,173	0,288	0,435	0,38	1,77	2,23	4,01
	40 %	0,049	0,173	0,288	0,435	0,38	1,77	2,23	4,01
D2	10 %	0,406	0,413	0,355	0,665	2,59	3,54	7,59	4,85
	20 %	0,408	0,384	0,474	0,823	2,40	3,50	7,70	4,64
	30 %	0,408	0,384	0,474	0,823	2,40	3,50	7,70	4,64
	40 %	0,408	0,384	0,474	0,823	2,40	3,50	7,70	4,64
D3	10 %	1,560	2,938	2,170	1,145	4,06	9,11	9,59	10,93
	20 %	1,289	2,843	2,183	1,030	4,17	8,88	9,66	11,09
	30 %	1,289	2,843	2,183	1,030	4,17	8,88	9,66	11,09
	40 %	1,289	2,843	2,183	1,030	4,17	8,88	9,66	11,09

Là aussi, le tableau 3 donne l'ERA associé à l'algorithme proposé de permutation séquentielle fondé sur les mêmes α et divers β. L'algorithme proposé de permutation séquentielle semble produire de très bons résultats, nettement supérieurs à ceux de la méthode d'appariement-permutation pour D_1, tandis que les résultats des deux méthodes ne se démarquent pas sensiblement les uns des autres pour D_2 ou D_3. Il convient toutefois de rappeler que nous introduisons, dans l'algorithme proposé de permutation séquentielle, un autre facteur de contrôle β pour restreindre davantage les risques de divulgation. En outre, dans tous les cas, l'application de l'algorithme de permutation séquentielle a exigé de 20 à 36 secondes de temps machine, soit un temps d'exécution nettement inférieur à celui requis par la méthode d'appariement-permutation.

Nous avons également procédé, à des fins de comparaison, à la permutation aléatoire d'unités pour les *q=28* variables qui n'ont pas été utilisées pour la permutation. Les ERA pour α=10 %, 20 %, 30 % et 40 % s'établissent respectivement à 15,72, 29,60, 41,48 et 51,34. La permutation aléatoire a été répétée 1 000 fois, après quoi la moyenne a été calculée. Comme on peut le constater, l'incidence de l'opération sur ces variables est, elle aussi, relativement faible.

5. Conclusions

Après avoir démontré les risques de divulgation des identificateurs d'UPE et de strate au moyen des poids de rééchantillonnage dans les données d'enquête rendues publiques, nous avons proposé l'application d'un algorithme rapide de permutation séquentielle permettant de permuter les identificateurs d'UPE entre les enregistrements avant la création des poids de rééchantillonnage. Cette méthode permet de masquer les identificateurs réels d'UPE sans modifier de façon significative les estimateurs de variance qui en résultent. L'application de cette méthode aux *National Health and Nutrition Examination Surveys* en illustre le potentiel. La rapidité d'exécution de l'algorithme, même lorsqu'il doit traiter des milliers d'enregistrements, permet à l'analyste de l'organisme responsable de la diffusion d'appliquer et d'évaluer la méthode à de nombreuses reprises lors de l'analyse des variables susceptibles

d'être utilisées pour la permutation et de l'incidence de l'opération sur l'estimation de la variance et les risques de divulgation.

Références

Dohrmann, S., Lu, W.W., Park, I., Sitter, R.R., et Curtin, L.R. (2006), "Variance Estimation to Protect Confidentiality in the National Health and Nutrition Examination Surveys", submitted to *Journal of Official Statistics.*

Lu, W.W., (2004), "Confidentiality and Variance estimation in Complex Surveys", thèse de doctorat non publié, Simon Fraser University, Canada.

Yung, W. (1997), "Variance Estimation for Public Use Files Under Confidentiality Constraints", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 434-439.